CAS MONOGRAPH SERIES
NUMBER 14

PRACTICAL MIXED MODELS FOR ACTUARIES

Ernesto Schirmacher, PhD, FSA, CSPA

CASUALTY ACTUARIAL SOCIETY

PRACTICAL MIXED MODELS FOR ACTUARIES

Ernesto Schirmacher, PhD, FSA, CSPA

Casualty Actuarial Society
4350 North Fairfax Drive, Suite 250
Arlington, VA 22203
www.casact.org
(703) 276-3100

This monograph provides a practical introduction to an area of actuarial practice that is at the intersection of the theories of credibility and mixed models. Several credibility models are shown to be special cases of the linear mixed model, and thus we may apply all the statistical machinery to assess, refine, and expand them. The text then introduces generalized linear mixed models, removing some of the constraints of the linear mixed model, and thus allowing for applications in the insurance industry. The focus is on the practical application of the theory rather than its development. Therefore, text and computer code are integrated as we discuss examples from property/casualty and health.

Practical Mixed Models for Actuaries
By Ernesto Schirmacher

Practical Mixed Models for Actuaries / Ernesto Schirmacher
ISBN 978-1-7333294-5-3 (print edition)
ISBN 978-1-7333294-6-0 (electronic edition)

Table of Contents

Preface

We all learn better if we engage with the material and connect new ideas or concepts with things we already know; therefore, we encourage readers to replicate the results presented in the text:

- exploratory graphs and tables,
- model fitting, and
- diagnostic graphs and tables.

Readers should attempt to write the code necessary to accomplish the various tasks before consulting the source code for this monograph.

We used the Quarto (Allaire et al. 2024) publishing system to integrate text, computations, tables, and graphs. The source code for this monograph is on GitHub in the repository pmmfa. All computations are done in **R** (R Core Team 2024), and nearly all datasets used are in various **R** packages that the reader can easily install from the Comprehensive R Archive Network (CRAN).

We made use of several packages, and readers wishing to replicate the work should install them. Tables 1, 2, and 3 list the packages, available on CRAN, that we used for computations, graphics/tables, and datasets, respectively.

Table 1. R packages used in computations.

Package Name	Citation
`actuar`	Dutang et al. (2008)
`dhglm`	Lee and Noh (2018)
`gamlss`	Rigby and Stasinopoulos (2005)
`hglm`	Rönnegård et al. (2010)
`lme4`	Bates et al. (2015)
`MASS`	Venables and Ripley (2002)
`mvtnorm`	Genz and Bretz (2009)
`statmod`	Dunn and Smyth (1996)
`tidyverse`	Wickham et al. (2019)

Table 2. R packages used for graphs and tables.

Package Name	Citation
`GGally`	Schloerke et al. (2024)
`kableExtra`	Zhu (2024)
`patchwork`	Pedersen (2024)

Table 3. R packages with datasets used in the text.

Package Name	Citation
`GLMsData`	Dunn and Smyth (2022)
`insuranceData`	Wolny–Dominiak and Trzesiok (2014)
`lars`	Hastie and Efron (2022)
`mdhglm`	Lee et al. (2018)

The `CASdatasets` package, from which we used a couple of datasets, is not available on CRAN due to its size. Please visit the CASdatasets page on GitHub for further information and instructions on how to install it.

The datasets `bus-case.csv` and `medpar` are not available from CRAN. Both can be found from the websites of their corresponding books:

- *Non-Life Insurance Pricing with Generalized Linear Models* (Ohlsson and Johansson 2010) is the source of the dataset `bus-case.csv`.
- *Negative Binomial Regression* (Hilbe 2007) provides the `medpar` dataset.

CAS Monograph Editorial Board

Acknowledgments

I want to thank my colleagues at Bentley University for their support, encouragement, and helpful advice as I navigated the roller-coaster process of putting thoughts to paper and refining them.

The CAS Publications staff and the team at Managed Editing, Inc., did a superb job of catching inconsistencies and polishing the manuscript. My sincere appreciation for all the behind-the-scenes work to get this over the finish line.

I would also like to thank the review team: Kenneth Hsu, Brian Ironside, Clifton Lancaster, Anthony Salis, and Betty Zhu. Their copious comments and suggestions have improved the presentation of the material and kept me from making some blunders. If you come across a passage that helps you understand an idea or concept, please thank them the next time you see them at a conference or event for all the time they spent reviewing it. If, on the other hand, you come across a section that does not make sense, keep in mind that it was not their fault that they had to deal with me.

A special thanks to Kenneth for providing constant positive feedback on the many raw drafts that came his way and for gently keeping me moving forward.

Finally, I would like to thank my family for their patience and understanding as I worked on this project.

About the Author

Ernesto Schirmacher studied undergraduate mathematics at the University of Rochester. He earned a PhD in algebraic combinatorics from the University of Minnesota and began his career as a technical underwriter for Agrippina Ruckversicherung in Cologne, Germany. He then transitioned to the actuarial department, focusing on pricing life and health reinsurance contracts for clients in Latin America. In 2001, he moved to Boston to work in the corporate actuarial department of Liberty Mutual on various projects, including asbestos reserving, economic scenario generation, asset modeling, and the application of statistical models in pricing and reserving. In 2019, he left the industry and began teaching mathematics, statistics, and actuarial science at Bentley University. He is a fellow of the Society of Actuaries and a Certified Specialist in Predictive Analytics. He volunteers with various groups from the Casualty Actuarial Society and The CAS Institute. He can be contacted at eschirmacher@bentley.edu.

1. Introduction

Generalized linear models (GLMs) made their appearance in 1972 with the publication of Nelder and Wedderburn's paper "Generalized Linear Models," and software implementing these models, known as GLIM (Generalized Linear Interactive Modelling), developed by the Working Party on Statistical Computing of the Royal Statistical Society (Nelder 1975) appeared three years later. Nearly 10 years after Nelder and Wedderburn's paper, McCullagh and Nelder published their monograph *Generalized Linear Models*, and a second edition followed in 1989 (McCullagh and Nelder 1989). That publication has been the go-to reference for the subject.

In the second half of the '70s and early '80s, several publications applied GLMs to the premium calculation in motor insurance (Coutts 1975, 1983; Baxter and Coutts 1977; Baxter et al. 1980). But despite these early contributions, there was no widespread adoption within the insurance industry.

Three years after the publication of the second edition of McCullagh and Nelder's book, Brockman and Wright (1992) published a paper that launched the adoption of GLMs in the UK motor insurance market, and 10 years thereafter American actuaries embraced GLMs in the ratemaking process for auto insurance.

Given the events just described, one might think that statisticians developed the theory and computational procedures and then actuaries, slowly, adopted the tools and techniques and put them to practical use. But such a sequence of events is not quite right.

In 1963, Robert A. Bailey published a paper in *Proceedings of the Casualty Actuarial Society* with the title "Insurance Rates with Minimum Bias" (Bailey 1963). In the introduction he writes that the techniques he is about to describe are "methods for obtaining insurance rates that are as accurate as possible for each class and territory and so on." Moreover, he mentions that "many of the techniques presented in the paper are already in use by the various bureaus and other ratemakers in one form or another."

The *minimum bias techniques* that Bailey described are now known to be special cases of a GLM as shown by Mildenhall (1999). These techniques were not developed within a statistical framework backing them, and thus they do not come with some of the standard diagnostic measures, such as residuals and deviance, that are used to check the model development process. Rather, they were created to solve the practical problems that actuaries were facing in managing their books of business.

Note that Bailey's paper predates Nelder and Wedderburn's introduction of GLMs by about 10 years, and thus one might argue that actuaries had developed the proto-idea of GLMs before statisticians. I wonder what might have happened if actuaries in the

'60s and '70s had been in closer contact with their fellow statisticians as they developed the techniques, tools, and computational procedures needed for their jobs. Would the insurance industry have embraced GLMs much earlier?

I believe that a close working relationship between actuaries and statisticians can be fruitful for both parties. In this monograph, we want to bring together two seemingly unrelated areas, credibility and mixed models, at a level accessible to practicing actuaries. Therefore, we will not fully develop the theory, but rather present enough that the main concepts can be grasped and focus on showing how one would implement the ideas through some examples.

The story of the development of credibility theory and mixed models has some parallels to the events described above for GLMs.

Credibility theory is a cornerstone of actuarial science (Hickman and Heacox 1999), and it comes in several flavors—limited fluctuation, greatest accuracy, hierarchical, and multidimensional, to name a few (Bühlmann and Gisler 2005). Greatest accuracy credibility is also known as Bühlmann credibility and was developed in the late 1960s (Bühlmann 1967) and further extended by Bühlmann and Straub (1970). Simply put, credibility is the combination of different estimates to come up with a single estimate (Venter 1996), and though it seems somewhat trivial to combine two estimates by linear interpolation, the method has far-reaching consequences and applications.

One application of credibility theory is concerned with the estimation of a policyholder's next year's premium in a book of business where we have some historical loss information for each insured. Whereas some policyholders may have a large volume of data, others may have very little. Credibility theory allows us to combine each policyholder's own experience and the experience of the whole portfolio.

About 15 years after Bühlmann credibility and 10 years after GLMs were introduced, Laird and Ware (1982) published their seminal paper on the linear mixed-effects model (also called the linear mixed model, or LMM). Up until that point, the linear model and the GLM were used to analyze a sample of data where the observations were independent and identically distributed. Researchers and practitioners were keenly aware that not all of the samples they wanted to analyze obeyed that restriction. In fact, in many situations statistical and actuarial practitioners had a sample of samples—that is, observations came in clusters and the number of clusters could be quite large.

One can view the LMM and the generalized linear mixed model as the next step in the evolution of the linear and generalized linear models, respectively. These models can handle data where some of the observations are no longer independent of each other.

It seems that credibility and mixed models do not have much in common, and for many years statisticians worked on mixed models and actuaries worked on credibility and they did not talk to each other very much. Both areas flourished and both extended their tools and techniques significantly. Then, Frees et al. (1999) made the connection that some credibility models can seen as special cases of the longitudinal data model that can be analyzed with LMMs. This connection allows actuaries to use the full power of mixed models in developing, fitting, and assessing some credibility models.

We begin our exploration in Chapter 2 with a review of GLMs. As most practicing actuaries are acquainted with the theory, we will base our review on working through a non-insurance example. This choice of dataset is deliberate and meant to break any preconceived relationships the reader might have from prior work with insurance applications. The data analyzed relates to the pulmonary function of children and teenagers exposed to cigarette smoke.

Chapter 3 introduces credibility theory, and we focus on the work of Bühlmann (1967) and Bühlmann and Straub (1970). This area is also known as *greatest accuracy credibility*. The expected value of the process variance (EVPV) and the variance of the hypothetical means (VHM) are important concepts that we will later see, under different names, in connection with mixed models. We end this chapter with the work of Hachemeister (1975), who applied the ideas of credibility theory to the linear regression model. His formulation gives us a random intercept and random slope regression model. Hachemeister applied his model to a set of insurance data and noticed that some of the credibility estimates obtained did not line up with some sensible practical considerations. Perhaps these initial counterintuitive results stifled the adoption of these ideas by other actuaries. We will retrace the steps Hachemeister took, see the counterintuitive results, and then apply some insight gained along the way to resolve the issue.

Next, in Chapter 4, we jump onto the statistical bandwagon and explore the ways in which statisticians evolved the standard linear model into the LMM. Instead of starting with the theory, we begin by reformulating the balanced Bühlmann and the Bühlmann–Straub models in the language of the LMM and note that we get the same estimates for the balanced Bühlmann model and nearly the same estimates for the Bühlmann–Straub model. Then, we present the very basics of the theory of the mixed model and apply them to a previous example to show a concrete application. We conclude the chapter by revisiting Hachemeister's data and applying these new tools.

In Chapter 5 we take the LMM and transform it into the *generalized linear mixed model* (GLMM), where we introduce link functions and expand the distribution of the response variable from a normal distribution to the family of linear exponential distributions. With these models we can not only model the response variable but also include explanatory variables for the dispersion parameter. Modeling the dispersion parameter does not require "mixed model" theory. We can achieve this by interlocking two GLMs (Nelder et al. 1998). But that joint model fits well with the approach we undertake in this chapter.

Mixed models, either the linear or the generalized version, are more complex and more difficult to estimate. We require more information from our data, and the computational procedures to estimate the parameters have more potential points of failure. The standard approach to compute the parameters of such models is to use maximum likelihood estimation. Maximizing the likelihood is a nontrivial task often involving analytically intractable integrals. Thus we must resort to numerical optimization techniques. One such technique is Monte Carlo simulation, which depending on the complexity of our model, may require a significant amount of time and the assessment that convergence has been achieved.

We will not use Monte Carlo simulation but rather focus on a different development, namely, the use of the theory of *h*-likelihood, which sits somewhere between the Bayesian and the frequentist approaches. The computational resources needed for this approach, while not small, are reasonable.

In the final chapter, the discussion is focused on the application of the GLMM to three datasets: automobile bodily injury, hospital length of stay, and fleet insurance. For all three datasets, we present an analysis starting with data exploration, moving to model building, and ending with diagnostics.

2. Generalized Linear Models

2.1. Introduction

Actuaries are well acquainted with GLMs, and in this chapter we provide a quick review of the main ideas and concepts as we work through a non-insurance example.

In the late 1970s and early 1980s, researchers in Boston were interested in understanding the effects of maternal smoking on the pulmonary function of children through a seven-year longitudinal study (Tager et al. 1979, 1983). The study subjects and their families were interviewed multiple times. For children 10 years or younger, the parents answered all questions except those regarding their smoking history. All other children answered all questions on their own. During pulmonary testing—a time when parents were not present—researchers asked children about their smoking history.

The longitudinal analysis showed that after adjusting for explanatory variables, such as age, height, change in height, and the smoking status of the child, maternal smoking hurts the development of the child's pulmonary function. A cross-sectional dataset from that investigation is available in the `GLMsData` package under the name `lungcap`.

```
data(lungcap, package = "GLMsData")
```

The dataset has 654 observations and five variables. The pulmonary function of the subjects was assessed through their lung capacity, which was measured via their forced expiratory volume. The forced expiratory volume is the amount of air a subject can expel from their lungs in the first second of a forceful exhalation. A larger volume of exhaled air signals better pulmonary function.

Table 2.1 shows the name, type, and description of each variable in the dataset. Forced expiratory volume, `FEV`, is our response variable, and the indicator variable for smoking, `Smoke`, is the principal variable of interest. The age (`Age`), gender (`Gender`), and height (`Ht`) of each child are variables that may be related to the response, and we want to control for them.

2.2. Exploratory Data Analysis

To work effectively with data we first need to understand what we have available to work with. Exploratory data analysis employs a set of techniques to help us understand and uncover what the data we have may be saying. It is not about confirming that a perceived pattern is true. For that there are other techniques. It is about looking closely

Table 2.1. Names, types, and descriptions of the variables available in the dataset.

Item	Variable	Type	Description
1	FEV	Continuous	The forced expiratory volume in liters.
2	Age	Integer	Age of subject in completed years.
3	Ht	Continuous	Height of the subject in inches.
4	Gender	Binary	The gender of the subject.
5	Smoke	Binary	The smoking status of the subject: Nonsmokers are coded with 0 and smokers are coded with 1.

at the data to find out what we can do with it. It is about learning and finding insights from the data and being able to describe them as easily as possible. In this section, we explore the lung capacity data. The response variable is forced expiratory volume, `FEV`, and the remaining variables may help us explain it.

The `Age` variable ranges from 3 to 19 years old, and the height variable is between 46 and 74 inches. Thus we have a broad spectrum of body sizes, and we should expect `FEV` to vary significantly as age and height varies. Table 2.2 shows summary statistics for the numeric variables. Note that the difference from the median (Q2) down to the first quartile (Q1) and up to the third quartile (Q3) is similar for each variable. This shows us that the bulk of the data, in each case, is fairly symmetric.

During these ages, children grow significantly, and `Age` and `Ht` should be strongly related to each other; in fact, their linear correlation coefficient is equal to 0.79. Thus, including both of these variables in a linear model may pose some estimation problems (multicollinearity).

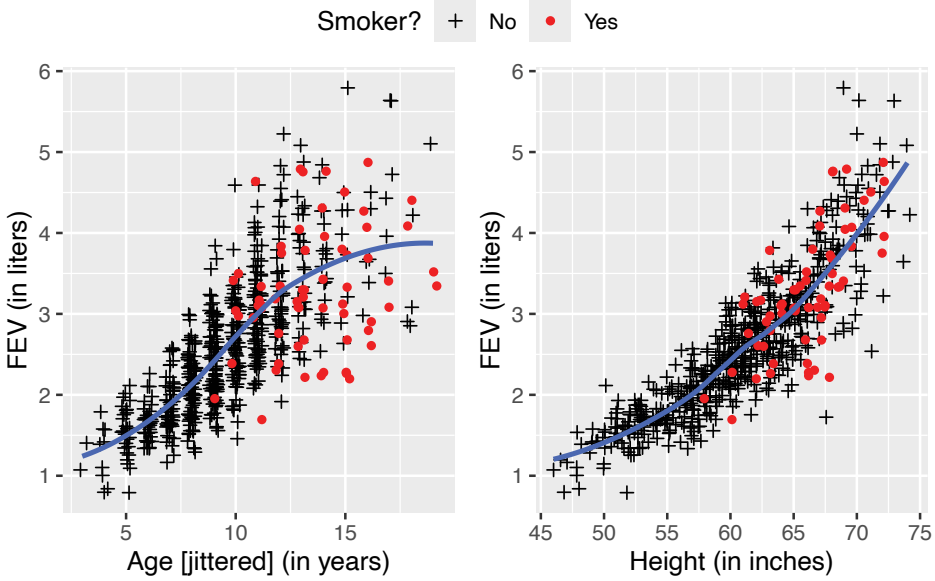
The remaining predictor variables are smoking status (`Smoke`) and gender (`Gender`). Both are binary variables. `Smoke` is an indicator variable where a value of 1 tells us that the child smokes and a value of zero says they do not smoke. There are 65 children who smoke in our dataset (about 10%). For the variable `Gender`, the split between female and male is 49% and 51%, respectively.

In Figure 2.1 we see that both `Age` (left-hand panel) and `Ht` (right-hand panel) have a strong nonlinear relationship with the response variable `FEV`. The nonlinear smooth curves ignore the information about which subjects smoke and which do not.

Table 2.2. Summary statistics for the numeric variables in the lung capacity dataset. Q1 is the first quartile, Q2 is the median, and Q3 is the third quartile.

	Min	Q1	Q2	Mean	Q3	Max
FEV (in liters)	0.79	1.98	2.55	2.64	3.12	5.79
Age (in years)	3.00	8.00	10.00	9.93	12.00	19.00
Height (in inches)	46.00	57.00	61.50	61.14	65.50	74.00

Figure 2.1. Age and height versus FEV. The red circles denote smoking subjects and the plus signs represent nonsmoking subjects. The smooth trend curves, which ignore smoking status, suggest nonlinear relationships with the response variable.



The left-hand panel shows that the relationship between Age and FEV resembles an elongated S curve. Also, we can see that as age increases, the cloud of points shows more dispersion as we move from the lower-left corner to the upper-right corner.

Switching to the right-hand panel, we see that the relationship between height (Ht) and FEV is also nonlinear, but the nonlinear pattern is simpler. In this case, it resembles part of a quadratic or exponential curve where increases in height lead to larger lung volumes. The cloud of points in this case is also more compact than the one, based on age, in the left-hand panel. These observations lead us to favor a model that uses height over one that uses age.

Also note that as the mean value of FEV increases in both scatterplots, the variability in FEV also increases. In other words, both plots show a *fanning out* of FEV as FEV increases. This relationship between the mean of the response and its variance, known as the **mean–variance** relationship, is extremely important in GLMs, as it determines the member of the exponential family of distributions that we should use for our response variable.

The relationship between the mean and the variance of the response variable for many members of the exponential family is given by

$$\mathrm{Var}\left[y\right] = \phi\mu^{b}, \tag{2.1}$$

where ϕ is the dispersion parameter, μ is the mean of the distribution, and b is a non-negative number. Well-known distributions correspond to different values of the exponent b.

For example, if $b = 0$, then we have the normal, or Gaussian, distribution. If $b = 1$ and $\phi = 1$, then the response variable is Poisson distributed, and if $b = 2$, then it is gamma distributed. Other values of b are possible, and not all members of the exponential family have a mean–variance relationship given by Equation 2.1 (e.g., the binomial and negative binomial distributions).

Note that by applying a logarithm to both sides of Equation 2.1 we get the following equation:

$$\ln\left(Var\left[y\right]\right) = \ln\left(\phi\right) + b\ln\left(\mu\right). \tag{2.2}$$

Therefore, we can use our data and ordinary least squares (OLS) to estimate the value of b.

For example, we can proceed as follows. First, create seven bins of approximately equal size for the height variable.

```
lungcap$Ht.bin <- cut_number(lungcap$Ht, n = 7)
```

Using the height bins, summarize the value of `FEV` for each bin by calculating the size of the bin, the mean, and the variance. Store the values in the object `mv` (mean–variance).

```
mv <- lungcap |>
  group_by(Ht.bin) |>
  summarize(sz = n(),
            mn = mean(FEV),
            vr = var(FEV))
```

Now estimate a linear regression equation where the response variable is the logarithm of the variance and the predictor variable is the logarithm of the mean.

```
fm <- lm(log(vr) ~ log(mn),
         data = mv,
         weights = sz)
sfm <- summary(fm)
round(sfm$coef[,1:2], 3)
```

```
             Estimate  Std. Error
(Intercept)    -3.740       0.253
log(mn)         2.015       0.260
```

The coefficient of the logarithm of the mean is our estimate of the value of b. In this case, $b = 2.02$, and since it is close to 2 in value, this suggests that we should model `FEV` as a gamma-distributed random variable. An approximate 95% confidence interval for the value of b is equal to $2.015 \pm 2 \times 0.260 = (1.495, 2.535)$. Clearly, this confidence

interval does not include zero, and therefore using a normal distribution for the response variable would not be a reasonable choice—that is, the normal distribution is not supported by the data.

But what about other choices? Perhaps a Poisson distribution ($b = 1$) or an inverse Gaussian distribution ($b = 3$) would be an appropriate choice. Well, the endpoints of the confidence interval are closer to these distributions, and so one could try them out.

Exercise 2.1 Redo the mean–variance analysis with a different number of bins. Try various choices, maybe $n = 3, 5, 10, 20$. Would you arrive at a similar conclusion?

Solution 2.1 Let $n = 20$ and create this many bins for the height variable

```
lungcap$Ht.bin <- cut_number(lungcap$Ht, n = 20)
```

Next, we group the data by each bin and compute the mean 'mn' and the variance 'vr' for each group. We also add the number of observations in each group 'sz'.

```
mv <- lungcap |>
  group_by(Ht.bin) |>
  summarize(sz = n(),
            mn = mean(FEV),
            vr = var(FEV))
```

Finally, we estimate a OLS regression line

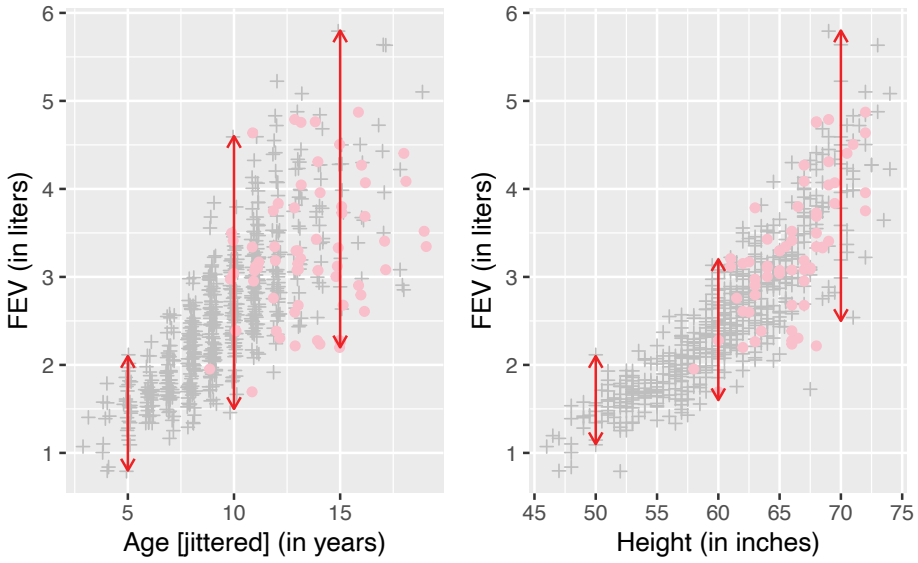
```
fm <- lm(log(vr) ~ log(mn),
         data = mv,
         weights = sz)
round(summary(fm)$coef[,1:2], 3)
```

```
               Estimate   Std. Error
(Intercept)      -3.955        0.226
log(mn)           2.110        0.232
```

The estimated value of b in this case is 2.110, and so we arrive at the same conclusion.

We also could have anticipated that our response variable `FEV` is not normally distributed by carefully inspecting the scatterplots shown in Figure 2.1. Looking at `FEV` versus `Age` (left panel) we see that as we move from the lower-left corner, where the response variable is small, to the upper-right corner, the variability in `FEV` values increases. A similar phenomenon appears in the right-hand panel in the scatterplot of `FEV` versus height (`Ht`). Figure 2.2 shows the increase in variability as the mean of `FEV` increases.

Figure 2.2. Age and height versus FEV. The data has been rendered in muted gray and pink. The arrows on both panels depict the variability of FEV for small, medium, and large values of FEV. The increase in variability is evident.



If our response variable FEV was normally distributed, then all the arrows in Figure 2.2 would have the same length regardless of their position along the horizontal axis. In other words, we would have seen *constant variance* for the response variable.

Exercise 2.2 Looking at Figure 2.1, we can see that the left-hand panel shows a cloud of points centered around the blue trend line that are not as compactly arranged as the points shown in the right-hand panel.

The left panel shows FEV vs. Age and the right panel is FEV vs. Ht.

Why do you think we see this phenomenon?

Solution 2.2 The response variable is FEV, and so it measures size of the lungs. The relationship between the size of the lungs and height is much better defined than the size of lungs and age.

We all know children of the same age but who have very different heights. Some are shorter, and others are taller. The taller ones have more space for larger lungs.

We also know that most children of the same height have similar builds and thus the variability in their lung size is smaller.

We have not yet explored how gender and smoking status may be related to the response variable. We can summarize our data by gender and smoker status and compute mean age, height, and FEV. Table 2.3 shows the results. Note that lung capacity

Table 2.3. Mean age, height, and FEV by gender and smoker status. The number of observations (Obs.) in each cell is also given.

	Nonsmoker				Smoker			
		Mean				Mean		
Gender	Obs.	Age	Height	FEV	Obs.	Age	Height	FEV
Female	279	9.4	59.6	2.38	39	13.3	64.6	2.97
Male	310	9.7	61.5	2.73	26	13.9	68.1	3.74

(FEV) for both genders is higher for smokers than for nonsmokers. Based on this alone, one might conclude that smoking would lead to higher lung capacity. But this would be an erroneous conclusion. The difference arises because the smoker and nonsmoker subjects have different age and height characteristics. We can see that gender does play a role in lung capacity. For both gender groups, female participants have a smaller height and a smaller lung capacity.

From our exploratory analysis we have learned that both age and height are strongly related to our response variable FEV. As FEV increases in mean value, its variability also increases, and thus using a normal distribution would not be supported by the data. In fact, the data suggests that a gamma distribution is appropriate. Also, gender and smoker status seem to play a role in influencing the response.

2.3. Modeling and Diagnostics

From our exploratory data analysis, we propose an initial model with the following specification: the response variable is FEV, and we will model it as a gamma distribution. The explanatory variable height (Ht) should be included in the linear predictor, perhaps entering as a linear term. But the right-hand panel of Figure 2.1 shows that the relationship is not perfectly linear, and thus we may need to use a transformation to obtain a better model. Gender and smoker status should also be included in the model, but we will build the model in stages, adding one variable at a time.

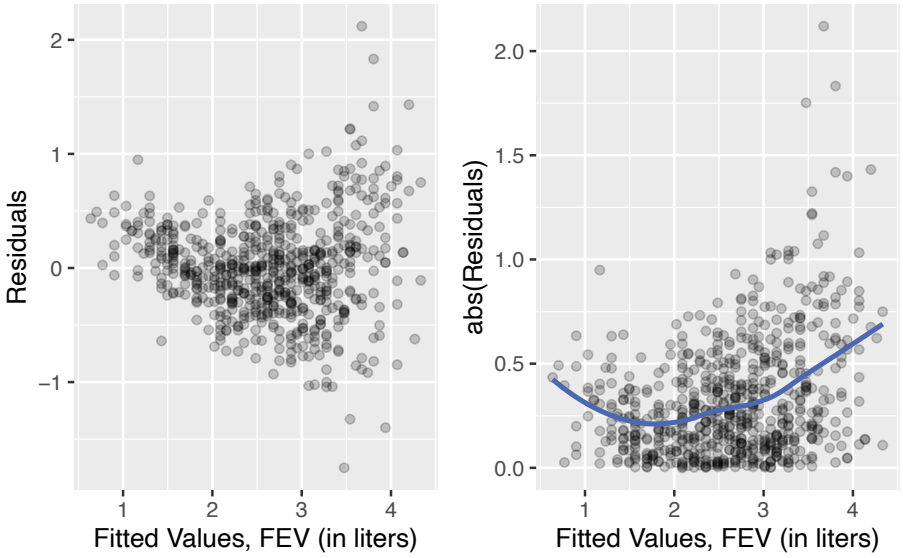
Before embarking on our gamma model we will illustrate how using OLS regression (that is, assuming that our response variable is normally distributed) would not be optimal.

Let us fit an OLS regression model to FEV and Ht and display a plot of residuals versus fitted values.

```
ols.fit <- glm(FEV ~ Ht,
               data = lungcap,
               family = gaussian(link = "identity"))
lungcap <- lungcap |>
  mutate(olsfit.mu = predict(ols.fit, type = "response"),
         olsfit.rD = resid(ols.fit, type = "deviance"))
```

Notice how, in the left panel of Figure 2.3, as the fitted values increase the vertical spread of the residuals also increases. This indicates that the constant variance assumption

Figure 2.3. Fitted values versus deviance residuals for an OLS model for FEV that includes height as an explanatory variable. Note the strong pattern in both panels telling us that our model is not adequate.



of the residuals is not met. The right-hand panel is a modification of the left-hand panel where we plot the absolute value of the residuals and include a smoothing trend line. This small alteration gives us a more nuanced view into the changes of spread as fitted values increase. For this plot, we see that the spread of residuals first decreases and then increases substantially.

Having seen that the normal distribution does not capture the true nature of our data, let us use what we learned during our exploratory data analysis and switch over to using a gamma distribution. We will fit several models and encode the key characteristics in the name. For example, a gamma model with an identity link function and having as main effects height and gender would be written as `gi.HG.fit`.

The general scheme is as follows:

- The first letter identifies the distribution: (n: normal, p: Poisson, g: gamma, i: inverse Gaussian, b: binomial, v: negative binomial, t: Tweedie).
- The second letter stands for the link function: (i: identity $g(x) = x$, l: logarithmic $g(x) = log(x)$, r: reciprocal or inverse $g(x) = 1/x$, s: square root $g(x) = \sqrt{x}$, o: logit $g(x) = log(x/(1-x))$).[1]
- The next group of letters indicates which variables are in the linear predictor: (A: age, H: height, G: gender, S: smoking).

For a numeric variable we may add a number, like `2`, to show that we have a polynomial of second degree in that variable as part of the linear predictor. And, finally, we add the word `fit` to signal that we have a fitted GLM.

[1] Other link functions are possible, such as probit, complementary log-log, and inverse squared.

To illustrate, the OLS model that we fitted above, `ols.fit`, would be named `ni.H.fit` using the proposed scheme (normal distribution, identity link, and main effect height).

Next, we will fit a gamma GLM to `FEV` using height as an explanatory variable and keeping the link function as the identity. The model name is `gi.H.fit`. Before performing this fit, we should develop an idea of what the sign and size of the estimated coefficient should be.

Based on Figure 2.1, we expect the coefficient for height, `Ht`, to be positive and roughly equal to $4/30 \approx 0.13$ (the line connecting the points (45, 1) and (75, 5) seems a reasonable approximation).

```
gi.H.fit <- glm(FEV ~ Ht,
                data = lungcap,
                family = Gamma(link = "identity"))
(sgi.H.fit <- summary(gi.H.fit))
```

```
Call:
glm(formula = FEV ~ Ht, family = Gamma(link = "identity"),
    data = lungcap)

Coefficients:
             Estimate  Std. Error  t value  Pr(>|t|)
(Intercept) -4.530471    0.132547   -34.18    <2e-16  ***
Ht           0.117013    0.002296    50.95    <2e-16  ***
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for Gamma family taken to be 0.02452508)

    Null deviance: 70.791 on 653 degrees of freedom
Residual deviance: 16.159 on 652 degrees of freedom
AIC: 641.83

Number of Fisher Scoring iterations: 5
```

The coefficient for `Ht` is roughly in line with our expectations, and note that the standard errors for both estimates are quite small compared to the size of the estimate.

Is our model a reasonable representation of the data? One way to try to answer this question is to use a technique known as *predictive simulation* (Gelman and Hill 2007). The basic idea is to fit a model to the data, then replicate the data from that fitted model, and finally compare the actual data with the replicates. If we can distinguish the actual data from the replicates, then our fitted model is not a very good representation of the actual data. And, if the actual data and the replicates are indistinguishable, then we have a good model.

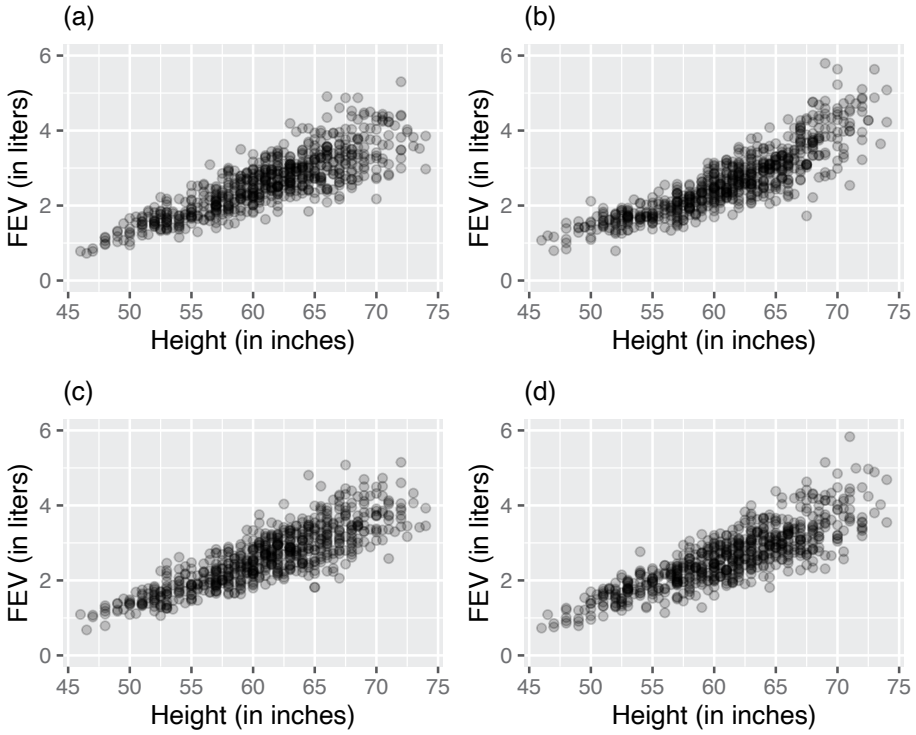
Using our gamma, identity link, with variable height as a main effect—that is, model `gi.H.fit`—we can simulate new datasets and compare them against our actual data.

Using the heights in our dataset, the following code simulates three sets of the response variable, FEV, from the appropriate gamma distribution.

```
set.seed(19390349)
n <- nrow(lungcap)
disp <- sgi.H.fit$dispersion
lungcap <- lungcap |>
  mutate(giHfit.mu = predict(gi.H.fit, type = "response"),
         giHfit.rp1 = rgamma(n,
                             shape = 1/disp,
                             scale = giHfit.mu * disp),
         giHfit.rp2 = rgamma(n,
                             shape = 1/disp,
                             scale = giHfit.mu * disp),
         giHfit.rp3 = rgamma(n,
                             shape = 1/disp,
                             scale = giHfit.mu * disp))
```

In Figure 2.4 we have four panels showing FEV versus height. Three of the panels have the simulated data from our model, and one panel has the actual data. Can you tell which panel has the actual data?

Figure 2.4. One panel contains the actual data, and the other panels have simulated data from a fitted model. Can you identify the panel with the actual data?



When comparing simulated data against actual data we should leverage everything we learned during our exploratory data analysis. We saw the following two key characteristics in the previous section:

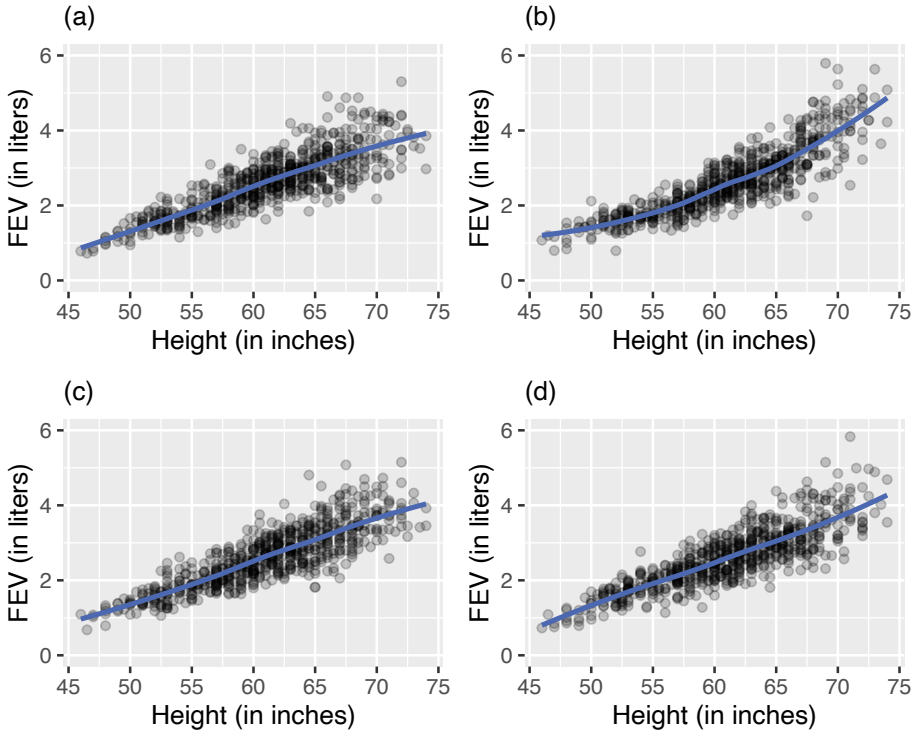
1. the variability in `FEV` increases as height increases, and
2. the relationship between `FEV` and height is convex.

Which panel in Figure 2.4 corresponds to the actual data?

In all four panels, we see that the variability in the response variable `FEV` increases as height increases. Thus, characteristic 1 above holds for all four panels. Can you see which panel violates characteristic 2?

Exercise 2.3 Recreate the four graphs but include a smoothing trend line for each graph to help you see the underlying relationship between `FEV` and `Ht`.

Solution 2.3 Adding a smooth trend line to a scatterplot can be done with a locally weighted regression procedure such as loess or lowess (Cleveland 1979). These methods are implemented in the `geom_smooth()`, which we add to each of the plots.



Note that the panel on the upper right is the only panel where the trend line is convex. All other trend lines are essentially straight lines.

Therefore, model `gi.H.fit` does not capture the convexity between the response variable `FEV` and the predictor variable height `Ht`.

The curvature that we observe between FEV and Ht could be modeled via a quadratic polynomial in Ht or by using a log-link function. Let us fit both models and apply some diagnostics. The quadratic model in height will be named gi.H2.fit (gamma distribution, identity link function, height and height squared as predictors), and the log-link model with height is gl.H.fit (gamma distribution, log-link function, and height as main effect).

```
lungcap$Ht.sq <- lungcap$Ht^2
gi.H2.fit <- glm(FEV ~ Ht + Ht.sq,
                 data = lungcap,
                 family = Gamma(link = "identity"))
gl.H.fit <- glm(FEV ~ Ht,
                data = lungcap,
                family = Gamma(link = "log"))
```

The coefficients for the quadratic model gi.H2.fit are

```
round(coef(gi.H2.fit), 5)
```

```
(Intercept)          Ht       Ht.sq
    5.34181   -0.22664     0.00296
```

and so we can calculate that the minimum value for the curve of predictions from this model occurs at a subject's height equal to

$$\frac{0.22664}{2 \cdot 0.00296} \approx 38.28$$

inches. This value is outside the range of our data but is reasonably close and very plausible by continuing the smooth trend shown in the right-hand panel of Figure 2.1.

The coefficients for the log-link model gl.H.fit are

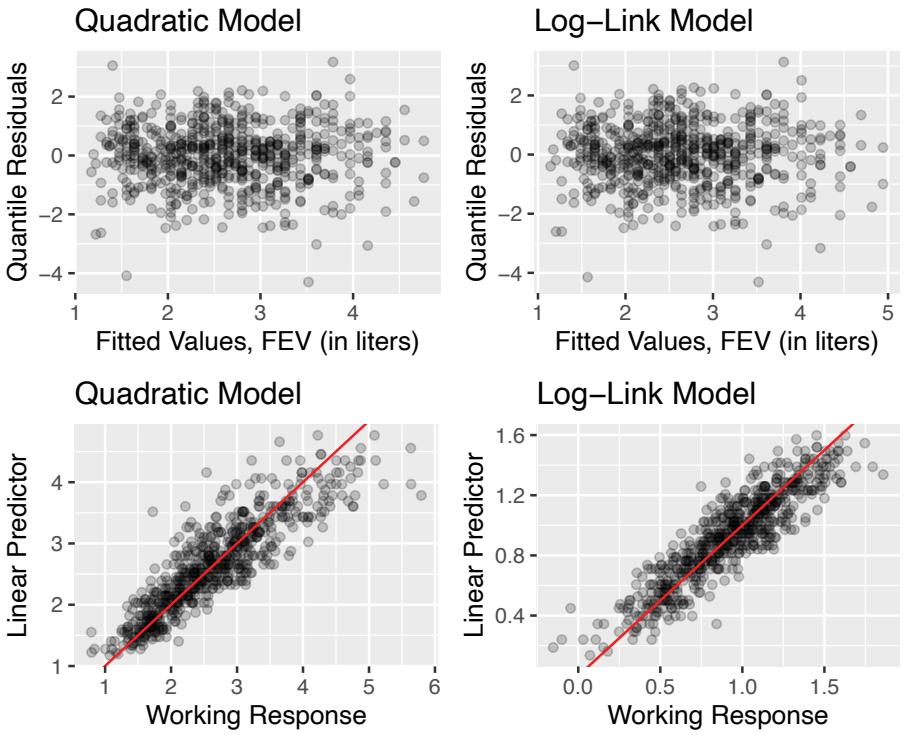
```
round(coef(gl.H.fit), 5)
```

```
(Intercept)          Ht
   -2.26794   0.05224
```

Thus we can infer that as the height for a child increases by 1 inch, the child's FEV will increase by approximately 5.4% ($e^{0.0522} - 1$).

The left-hand panels of Figure 2.5 show the fitted values versus quantile residuals and the linear predictor versus working residuals for the quadratic model in height Ht. The right-hand panels show the same plots for the log-link model.

Figure 2.5. Fitted values versus quantile residuals and working response versus linear predictor for the quadratic as well as the log-link model. All four plots display the desired null pattern, and in the top panels we may have two outlying observations (residuals below the horizontal line at y = –4).



Quantile residuals were introduced in Dunn and Smyth (1996), and an excellent overview of them appears in Dunn and Smyth (2018). Pearson and deviance residuals are the standard choices when analyzing the adequacy of fits for GLMs. Both are approximately normal with deviance residuals being a bit more so, but for discrete distributions the approximation to normality can be particularly bad. Quantile residuals overcome these issues and are strongly recommended for discrete models, and we can use them just as we would use deviance or Pearson residuals for diagnostic purposes.

The top panels of Figure 2.5 display fitted values versus quantile residuals, and we can see in both plots a nice random cloud of points centered about the line $y = 0$. There appear to be two outlying points in both plots below the line $y = -4$. The bottom panels are an *informal* diagnostic on the link function. The plot shows the linear predictor $\hat{\eta}_i$ on the y-axis and the working response

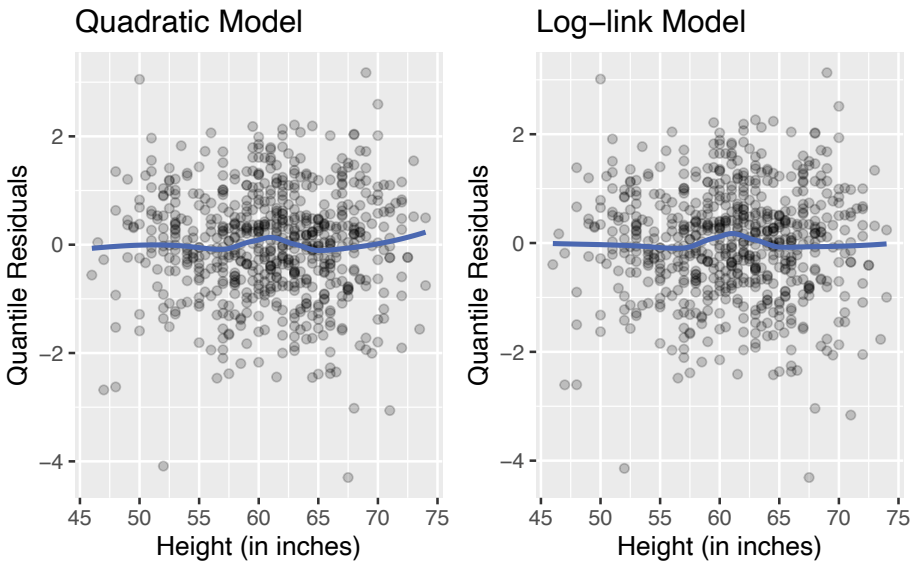
$$z_i = \hat{\eta}_i + \hat{e}_i$$

on the x-axis (Dunn and Smyth 2018, 308). Note that the working response z_i is the sum of the linear predictor and the *working residuals* $\hat{e}_i$. The working residuals are the residuals from the last iteration of the Fisher scoring algorithm used to compute estimates for the coefficients of the model.

If the link function is correct and we have the appropriate explanatory variables in our model and on the right scale, then we expect the points to cluster around the line $y = x$. Major deviations from this null pattern are a red flag that something is wrong with our model. In our example, both panels show the appropriate behavior, but the log-link model has a more cohesive cloud of points around the reference line compared to the identity-link quadratic polynomial in height, which shows a pattern deviating from the line $y = x$ in the upper-right corner.

Exercise 2.4 Plot the quantile residuals versus height for both models. Are there any concerning patterns in the plots?

Solution 2.4 The following figure shows the quantile residuals for both models:



Both panels show a random cloud of points that are centered about the line $y = 0$. Near 60 inches in height we see an upward bump on the scatterplot smoother (blue line) indicating that our models tend to underpredict in that region. But the bump is fairly small.

Currently, our best model uses a gamma distribution and a log-link function. The quadratic polynomial in height with an identity-link function does not show a strong linear relationship between the linear predictor and the working residuals (check the upper-right corner). The model equation for the current best model is

$$\log\left(\mathbb{E}\left[\text{FEV}\right]\right) = \beta_0 + \beta_1 \text{Ht}.$$

Next, we incorporate the `Gender` categorical variable and the indicator variable for `Smoke`. This indicator variable has a value of 1 for subjects who smoke and zero otherwise. We would expect the coefficient for `Smoke` to be negative because we think

that smoking would have a detrimental effect on our lungs and thus diminish lung capacity. The absolute value of the coefficient will tell us how much lung capacity will be affected by smoking. As for `Gender`, the size and direction of the effect is not clear.

```
gl.HGS.fit <- glm(FEV ~ Smoke + Gender + Ht,
                  data = lungcap,
                  family = Gamma(link = "log"))
sgl.HGS.fit <- summary(gl.HGS.fit)
round(coef(sgl.HGS.fit), 4)
```

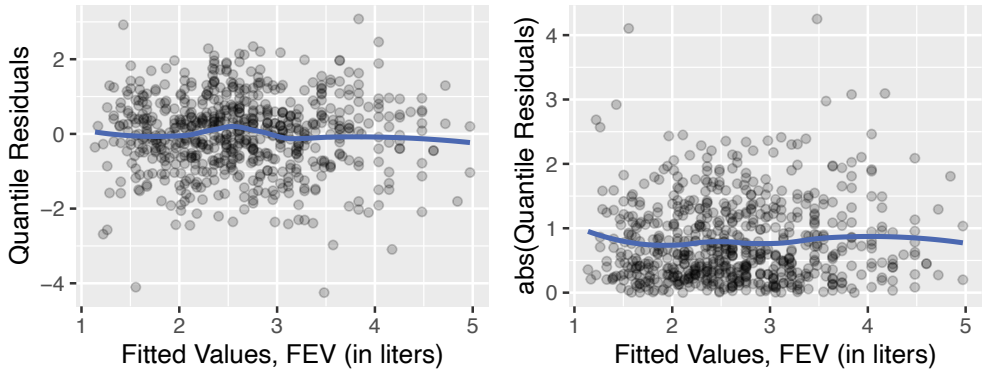
```
             Estimate  Std. Error   t value  Pr(>|t|)
(Intercept)   -2.2615      0.0639  -35.3758    0.0000
Smoke          0.0001      0.0201    0.0055    0.9956
GenderM        0.0187      0.0117    1.5994    0.1102
Ht             0.0520      0.0011   48.8380    0.0000
```

Even though the coefficient for `Smoke` has a positive sign, there is no evidence in the data to suggest that it is different from zero. Similarly, the effect for `Gender` is not statistically significant.

These results **do not show** that children who smoke do not have impaired lung capacity.

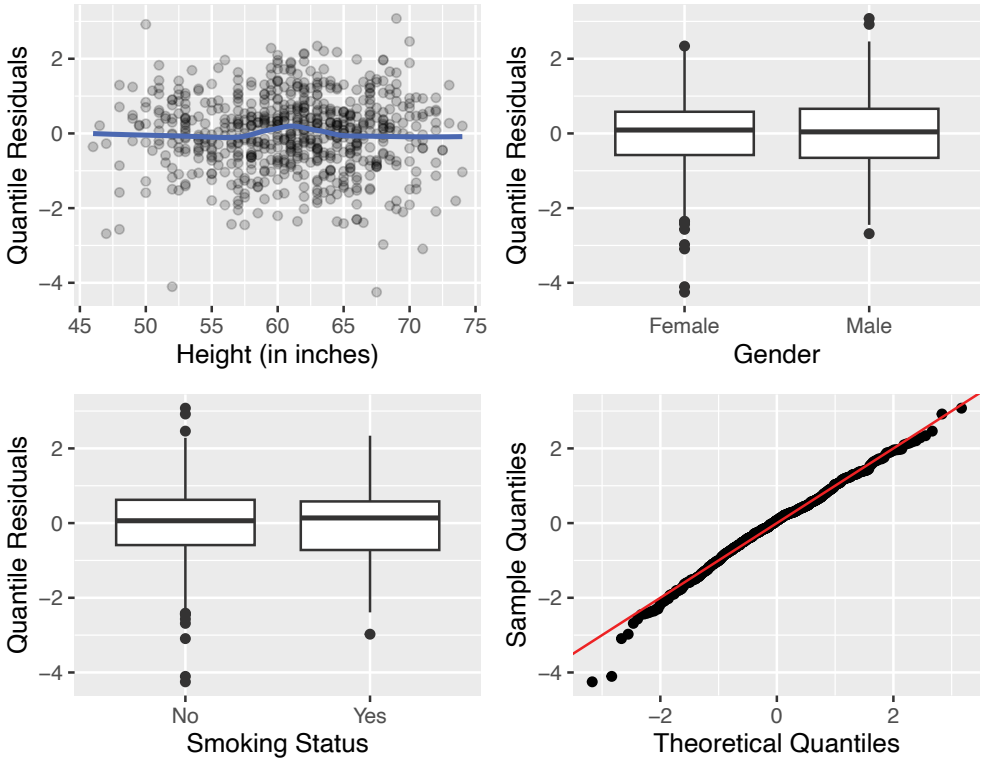
Figure 2.6 shows some diagnostic plots for our final model `gl.HGS.fit`—gamma distribution, log-link, and main effects for height, gender, and smoking status. All six

Figure 2.6. Diagnostic plots for our final model. The two left-hand panels show a random cloud of points centered about the line $y = 0$. The trend line exhibits a small positive bump in the center of the display. The upper-right panel shows that our model captures the increasing variability in the response variable well. The QQ plot in the lower-right panel shows that the lower tail of our data is thicker than it should be.



(continued on next page)

Figure 2.6. (*Continued*)



plots show that our model fits the data well. In the bottom-right panel we can see that the bulk of the data follows the line $y = x$, but we also see a slight deviation from the null pattern in the lower tail of the distribution. In this case, our quantile residuals have a slightly thicker lower tail than a standard normal distribution, but the number of points exhibiting this behavior is very small.

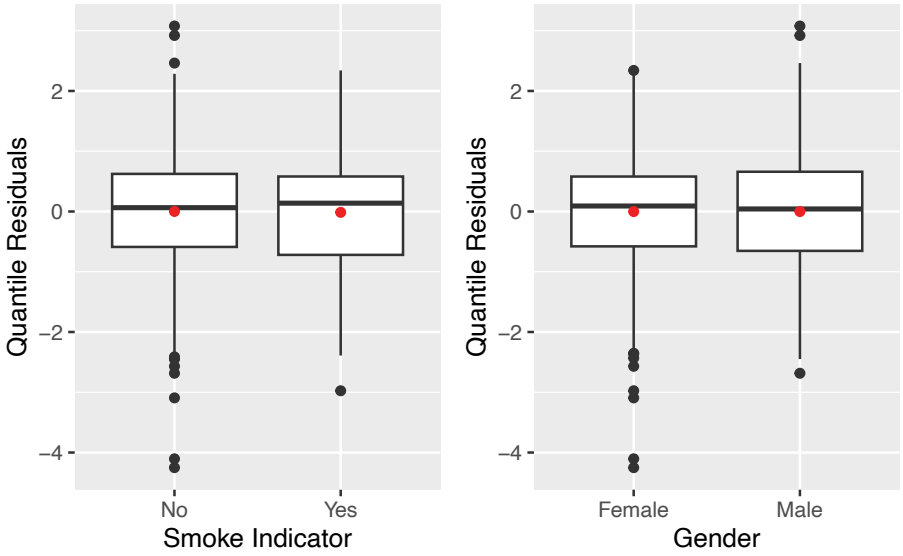
Exercise 2.5 In Figure 2.6 we included diagnostic plots for the variables `Smoke` and `Gender`, but we did not include any commentary. Also the boxplots do not show the mean value.

Recreate these two boxplots, and add a point showing the mean value of the residuals for each category and comment on what information they contribute to the final model.

Solution 2.5 To add the mean value to the boxplot we use a `stat_summary()` function to compute the mean and add a point to the display.

For both variables, the means of the residuals are centered at zero, they have equal spread, and they have a few outlying points. There are no indications that our model is missing any information from these variables.

Two points have quantile residuals whose absolute value is greater than 4. These points require further investigation. Perhaps the value of the response or predictor variables was recorded incorrectly.



2.4. Summary

This chapter focused on refreshing some of the main concepts for GLMs by working through a concrete example. GLMs provide a richer set of regression models for the analyst to draw on. Many GLMs exhibit a mean–variance relationship of the form $\text{Var}[y] = \phi\mu^b$, where ϕ is the dispersion parameter, μ is the mean of the distribution, and the value of b determines the distribution. Specific values are as follows: $b = 0$ normal (that is, we are back to OLS), $b = 1$, $\phi = 1$ Poisson, $1 < b < 2$ Tweedie distribution with a probability mass at zero, $b = 2$ gamma, and $b = 3$ inverse Gaussian.

In many situations we can use the data to estimate the mean–variance relationship and select an appropriate distribution. We also showed several diagnostic plots, such as the standard residuals versus fitted values, absolute value of residuals against fitted values, an informal check on the link function by plotting the linear predictor against the working responses, and a QQ plot of residuals.

3. Credibility Theory

3.1. Introduction

Consider the following scenario: you are preparing the renewal offer for a policyholder who has been insured for a number of years. While there are many approaches to setting next year's premium, consider the following two positions:

1. Base it entirely on the historical claims experience of this policyholder—that is, use their average claims experience.
2. Completely ignore this policyholder's own claims experience and use the company's average claims experience for all similar policyholders or the industry loss history.

Both positions are extreme. In the first one, you are in essence saying that your policyholder's experience is completely trustworthy for setting next year's premium. Maybe your policyholder is so large and their claims experience so stable that, barring any extraordinary events, next year's claims will be spot on with their historical record. In adopting the second position, you acknowledge that this policyholder's claims experience is not trustworthy (whether good, bad, or mixed), and therefore you'll look for the overall average claims for the entire portfolio of policies to which this policyholder belongs or to industry loss experience.

These two extreme positions are not the only alternatives. There is some middle ground, where we can blend some of the policyholder's historical experience together with the experience of the block of business to which the policyholder belongs. Credibility theory is the body of knowledge, tools, and techniques that allows us to blend the two extreme positions into a far better estimate for our policyholder's next year's premium.

Venter (1996) puts it as follows:

> *Credibility*, simply put, is the weighting together of different estimates to come up with a combined estimate. For instance, an insured's own experience might suggest a different premium from that in the manual. These are two different estimates of the needed premium, which can be combined using credibility concepts to yield an adjusted premium.

And we can summarize it in a formula as

$$\mathrm{AP} = Z \times \mathrm{EP} + \left(1 - Z\right) \times \mathrm{MP}, \tag{3.1}$$

where AP is the adjusted premium, EP is the policyholder's own experience premium, MP is the manual premium (also known as the *complement of credibility*), and $Z \in [0,1]$ is the credibility factor. The adjusted premium is also known as the *credibility premium*. Even though Equation 3.1 is a deceptively simple interpolation formula between EP and MP, it has far-reaching consequences and applications.

Note that as Z approaches 1, the adjusted premium gets closer to the policyholder's own experience premium. And as Z approaches zero, the adjusted premium converges to the manual premium. Thus, if the insured's own experience is highly credible (Z close to 1), we would assign an adjusted premium close to their own experience. If the experience is not credible, then we would assign a premium close to the premium suggested by the manual.

The key question is how to calculate the credibility factor based on observed data. An intuitive understanding is that the **more extensive** the observed data is and the **less it fluctuates**, then the closer the credibility factor will be to 1.

A. H. Mowbray (1914) introduced credibility theory a little over 100 years ago in his paper "How Extensive a Payroll Is Necessary to Give a Dependable Pure Premium?" The title succinctly encapsulated one of the main problems facing casualty actuaries at that time. In the intervening time, credibility theory has developed tremendously, and today there are many approaches and directions. Practicing actuaries are most familiar with two main methods of calculating credibility:

1. Limited fluctuation, or classical, credibility
2. Greatest accuracy, or Bühlmann, credibility

We will not present any results regarding limited fluctuation credibility as it is not connected with LMMs. Readers wanting a review of that branch of credibility can consult Chapter 5 of Herzog (2010). In the next section, we begin our exploration of Bühlmann credibility.

3.2. Greatest Accuracy Credibility

Greatest accuracy credibility, also known as *Bühlmann credibility*, was developed by Bühlmann (1967), who derived the optimal credibility factors by minimizing a squared error in the context of a Bayesian statistical model.

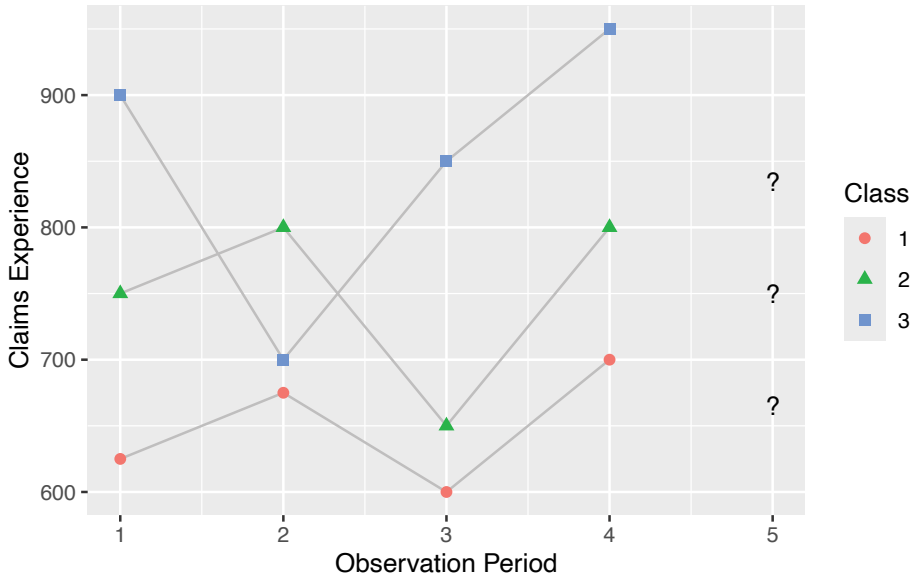
We will start with a basic model and expand to a more complex treatment. While the basic model is too simple to be effectively used in practice, it is important for understanding how more complex models work. Our development closely follows the presentations in Straub (1997) and Kaas (2009).

To keep things concrete, consider the following example (a slightly modified version of Problem 5.84 from Klugman et al. [1998]). We have a portfolio of policyholders with three, $J = 3$, different risk classes, and we have observed their claims experience over the last four, $T = 4$, years. Let X_{jt} be the experience for risk class $j = 1, 2, \ldots, J$ in time period $t = 1, 2, \ldots, T$. Table 3.1 shows the data, and Figure 3.1 provides a graphical representation. We would like to estimate the experience each risk class will have during the next time period $T = 5$.

Table 3.1. Claims experience for a portfolio of three risk classes that have been observed over four time periods.

	Time Period			
Class	1	2	3	4
1	625	675	600	700
2	750	800	650	800
3	900	700	850	950

Figure 3.1. Claims experience for a portfolio of three risk classes that have been observed over four time periods. What should the estimate, for each risk class, be in time period five?



Looking closely at Figure 3.1 and focusing on each risk class at a time, we could say the following: if we continue observing these risk classes for many periods (and assuming that these risks are stable over time), each one of them would fluctuate around a mean claim cost, say, $\bar{X}_j = (\sum_{t=1}^{T} X_{jt})/T$. For example, looking at risk class $j = 3$ (square symbol), which starts in period 1 with a value of 900, it seems plausible that its long-term average cost might be around $\bar{X}_3 = 850$. For risk class $j = 2$ (triangle symbol) with claim cost $X_{21} = 750$ at time $t = 1$, its long-term average might be close to $\bar{X}_2 = 750$; and for risk class $j = 1$ (circle symbol), starting with $X_{21} = 625$, that long-term average may equal $\bar{X}_1 = 650$. The portfolio as a whole (ignoring risk class information) also has a long-term average claim cost that, in this case, would be around $\bar{X} = 750$.

From the experience that we see in Figure 3.1, we might be inclined to say that these three risk classes have **different** long-term claim averages. Is there evidence in this data that this is the case? How might we quantify such evidence?

One way to quantify the evidence for or against different long-term averages would be to use a statistical model for this data. To that end, the experience X_{jt} for risk class $j = 1, 2, \ldots, J$ in time period $t = 1, 2, \ldots, T$ could be decomposed as

$$X_{jt} = m_j + \epsilon_{jt},$$

where m_j is the mean for risk class j and ϵ_{jt} represents an error term. We assume that the error terms are independent and identically distributed with $\epsilon_{jt} \in N(0, \sigma^2)$. Hence, all the X_{jt} are independent and $N(m_j, \sigma^2)$ distributed, with possibly unequal means m_j, but all with equal variance σ^2, across all risk classes. We can test for the equality of all group means via an analysis of variance.

The analysis of variance entails computing two statistics that will be relevant for credibility calculations. The first is the *sum of squares between*,

$$\text{SSB} = \sum_{j=1}^{J} T\left(\bar{X}_j - \bar{X}\right)^2.$$

This statistic has $J - 1$ degrees of freedom. The second statistic is the *sum of squares within*,

$$\text{SSW} = \sum_{j=1}^{J}\sum_{t=1}^{T}\left(X_{jt} - \bar{X}_j\right)^2,$$

and it has $J(T - 1)$ degrees of freedom.

Under the assumption that the group means m_j are equal (this is the null hypothesis), the random variable SSB has a mean equal to $(J - 1)\sigma^2$ and the random variable SSW has a mean equal to $J(T - 1)\sigma^2$. The ratio of these means follows a Fisher distribution with $J - 1$ and $J(T - 1)$ degrees of freedom:

$$F = \frac{\text{MSB}}{\text{MSW}} = \frac{\text{SSB}/\left(J - 1\right)}{\text{SSW}/\left(J\left(T - 1\right)\right)}.$$

For our example, we can calculate the sum of squares between (SSB) and the mean sum of squares between (MSB) as follows:

```
J <- length(levels(dta$class))
Tm <- length(unique(dta$time))
X.jt <- dta$value

Xj.bar <- tapply(X.jt, dta$class, mean)
X.bar <- mean(X.jt)
SSB <- Tm * sum((Xj.bar - X.bar)^2)
MSB <- SSB/(J - 1)
```

The sum of squares within (SSW) and the mean sum of squares within (MSW) are calculated as follows:

```
SSW <- sum((X.jt - rep(Xj.bar, each = Tm))^2)
MSW <- SSW / (J * (Tm - 1))
```

Their values are as follows:

```
c("SSB" = SSB, "MSB" = MSB, "SSW" = SSW, "MSW" = MSW)
```

```
  SSB     MSB    SSW   MSW
80000   40000  56250  6250
```

And so we have that the *F*-statistic, `F.value`, its critical value at 5%, `z.star`, and its *p*-value are as follows:

```
round(c("F.value" = MSB / MSW,
        "z.star" = qf(0.95, J - 1, J * (Tm - 1)),
        "p-value" = pf(MSB/MSW, J - 1, J * (Tm - 1),
                       lower.tail = FALSE)), 4)
```

```
F.value  z.star  p-value
 6.4000  4.2565   0.0187
```

Therefore, in this case, we have evidence that at least two of the means m_1, m_2, m_3 are not equal (we are able to reject the null hypothesis of equal means), and thus we would consider our portfolio to be heterogeneous.

Had the *F*-statistic been below the critical value, then we would not have been able to reject the null hypothesis that all the means are equal. Our data would not have strong evidence of being heterogeneous.

The calculations we just did for the sum of squares between, the sum of squares within, and the *F*-statistic can be done easily by fitting a linear model to the data and creating an analysis of variance table.

```
fm <- lm(value ~ class, data = dta)
anova(fm)
```

```
Analysis of Variance Table

Response: value
           Df  Sum Sq  Mean Sq  F value   Pr(>F)
class       2   80000    40000      6.4  0.01867  *
Residuals   9   56250     6250
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Exercise 3.1 Consider the claims experience for class 3, which has values equal to

```
dta$value[dta$class == 3]
```

```
[1] 900 700 850 950
```

Suppose we subtract the same amount m from each of these values to bring them closer to the values for classes 1 and 2.

What is the smallest value of m such that we would no longer consider our portfolio heterogeneous? In other words, what value of m would yield an F-statistic equal to its critical value at the 5% level?

Solution 3.1 We are looking for the smallest value of m such that the F-statistic for our portfolio is equal to 4.256. To search for the value of m, we can construct a function of one argument, m, so that its minimum value is achieved at the value of m that we are looking for.

For a given value of m, we would need to perform the following steps:

1. Decrease all the values for class 3 by m.
2. Fit a linear model to the data.
3. Compute the F-statistic for this data, `F.value`.
4. Return the square of the difference between `F.value` and the given critical value.

The following function implements these steps. The argument `z.star` is the target critical value we want to achieve, `dt` is the data frame containing our portfolio, and `cls` is the class we want to modify.

```
f <- function(m, z.star, dt, cls) {
  idx <- dt$class == cls
  dt$value[idx] <- dt$value[idx] - m
  fm <- lm(value ~ class, data = dt)
  F.value <- anova(fm)[1,4]
  ans <- (F.value - z.star)^2
  return(ans)
}
```

Now that we have our function, we can search for its minimum value via `optimize()`. The first argument is the function we want to minimize, and the second argument provides an interval to conduct the search. The remaining named arguments are needed by our function `f` to do its calculations. Based on Figure 3.1, it seems reasonable to assume that m should be in the interval from zero to 100.

```
optimize(f, c(0, 100),
         z.star = qf(0.95, 2, 9),
         dt = dta,
         cls = 3)
```

```
$minimum
[1] 38.41004

$objective
[1] 4.28568e-16
```

Therefore, when we reduce the experience for class 3 by $m = 38.41$ we obtain a portfolio that we would consider homogeneous.

Note that in the analysis of variance table shown before the exercise, it is the residual sum of squares, namely, the sum of squares within, that remains constant regardless of the value of m. It is the sum of squares between, that is, the `class` sum of squares, that gets smaller as m increases. Thus, the denominator of the F-statistic is fixed while the numerator gets smaller.

In the preceding analysis, we have treated the risk class means m_j as fixed but unknown. If our portfolio is heterogeneous, we may try to find a way to relate these means to other information we may have about the risk classes. In a practical application, we may have thousands of risk classes. Think about a classification system for automobile insurance with variables such as age, gender, socioeconomic status, years licensed, prior claims, garage location, make and model of car, safety features, engine size, and so forth. Many of these cells would be common and have plenty of data, but many would also be rare and have little data. Our linear model would have to estimate parameters for all these risk categories, and that would present a significant estimation problem. Moreover, as the number of risk classes increases, so does the number of parameters that we need to estimate.

Another way to look at our portfolio would be to assume that the risk class mean m_j is a random draw from a distribution. Thus, we would decompose our data as follows:

$$X_{jt} = \mu + \Xi_j + \epsilon_{jt}, \quad j = 1, 2, \ldots, J, \quad t = 1, 2, \ldots, T, \tag{3.2}$$

where Ξ_j (the capital version of the letter ξ) and ϵ_{jt} are independent random variables with mean zero and

$$Var\left[\Xi_j\right] = \tau^2, \quad Var\left[\epsilon_{jt}\right] = \sigma^2.$$

Note that from Equation 3.2 we have

$$\mathbb{E}\left[X_{jt}\right] = \mu \quad \text{and} \quad \text{Var}\left[X_{jt}\right] = \tau^2 + \sigma^2;$$

that is, the variance of each cell is equal to the sum of the variance for each component in Equation 3.2.

In terms of our portfolio we can interpret the above model as follows: the overall mean is given by μ, and it is the expected value of claim costs for a contract picked at

random from our portfolio. The term Ξ_j is a random variable, and it represents a deviation from the grand mean μ specific to risk class j. The conditional mean of X_{jt} given that $\Xi_j = \xi$ is equal to $\mu + \xi$. This would be the long-term average for risk class j. The variance of Ξ_j is equal to τ^2, and so this parameter controls how spread out individual risk classes are from the overall mean. A large value of τ^2 would lead to a heterogeneous portfolio. A small value of τ^2 suggests a homogeneous portfolio where all risk classes have similar long-term means. The last component, ϵ_{jt}, gives us a deviation for risk class j from its long-term mean $\mu + \xi$ in year t. It represents the fluctuation of the experience around its long-term average.

It is important to note that this model has three parameters: the overall mean μ, the variance component τ^2, and another variance component σ^2. These three parameters are independent of the number of risk classes in our portfolio. We may have just three risk classes, as in the example above, but we could also have hundreds or thousands of risk classes and we would still need to estimate only three parameters.

As Figure 3.1 depicts with the question marks at time $t = 5$, we are interested in estimating the expected value of the unobserved random variables $X_{j,T+1}$. While there may be many ways of estimating that value, we will require it to be a linear combination of the observed data that we have, namely, $X_{11}, X_{12}, \ldots, X_{JT}$. We also want our linear combination to have the same expected value as $X_{j,T+1}$ (we want our estimator to be unbiased) and its squared error to be the smallest among all possible linear combinations.

The following theorem (Kaas 2009, Theorem 8.2.2) tells us that the best estimate for the next period is a credibility-weighted average between $\bar{X}_j$ and $\bar{X}$, where the weight depends on the number of observed periods T and the variance components τ^2 and σ^2.

Theorem 3.1 (Balanced Bühlmann; homogeneous estimator). *Assume that the claim figures X_{jt} for contract j in period t can be written as the sum of stochastically independent components, as follows:*

$$X_{jt} = \mu + \Xi_j + \epsilon_{jt}, \quad j = 1, 2, \ldots, J, \quad t = 1, 2, \ldots, T+1, \tag{3.3}$$

where the random variables Ξ_j are independent and identically distributed with mean $\mathbb{E}[\Xi_j] = 0$ and $Var[\Xi_j] = \tau^2$ and the random variables ϵ_{jt} are also independent and identically distributed with $\mathbb{E}[\epsilon_{jt}] = 0$ and $Var[\epsilon_{jt}] = \sigma^2$ for all j and t. Furthermore, assume that the variables Ξ_j are independent of the variables ϵ_{jt}.

Under these conditions, the homogeneous linear combination $g_{11}X_{11} + \ldots + g_{JT}X_{JT}$ that is the best unbiased predictor of $X_{J,T+1}$ in the sense of minimal mean squared error

$$\mathbb{E}\left[\left(X_{J,T+1} - g_{11}X_{11} - \cdots - g_{JT}X_{JT}\right)^2\right] \tag{3.4}$$

equals the credibility premium

$$z\bar{X}_j + \left(1 - z\right)\bar{X}, \tag{3.5}$$

where

$$z = \frac{\tau^2 T}{\tau^2 T + \sigma^2} = \frac{T}{T + \sigma^2/\tau^2}$$

is the resulting best credibility factor (which in this case is the same for all j);

$$\bar{X} = \frac{1}{JT}\sum_{j=1}^{J}\sum_{t=1}^{T} X_{jt} \tag{3.6}$$

is the collective estimator of μ*; and*

$$\bar{X}_j = \frac{1}{T}\sum_{t=1}^{T} X_{jt} \tag{3.7}$$

is the individual estimator of m_j.

Nonparametric estimators of τ^2, σ^2, and μ are developed in Section 5.5.1 of Klugman et al. (1998). The overall mean μ can be estimated via $\bar{X}$. To estimate σ^2, also known as the expected value of the process variance (EVPV), consider first the following estimate of the variance for risk class *j*:

$$\hat{\sigma}_j^2 = \frac{1}{T-1}\sum_{t=1}^{T}\left(X_{jt} - \bar{X}_j\right)^2. \tag{3.8}$$

We can take the average of these estimates,

$$\hat{\sigma}^2 = \frac{1}{J}\sum_{j=1}^{J}\hat{\sigma}_j^2 = \frac{1}{J(T-1)}\sum_{j=1}^{J}\sum_{t=1}^{T}\left(X_{jt} - \bar{X}_j\right)^2, \tag{3.9}$$

to obtain the EVPV (see Equation 5.75 in Klugman et al. [1998]).

To estimate the variance of the hypothetical means (VHM), we use the relationship (Klugman et al. 1998, 465)

$$\mathrm{Var}\left[\bar{X}_j\right] = \tau^2 + \frac{\sigma^2}{T}. \tag{3.10}$$

The left-hand side is equal to the mean sum of squares between (MSB), and thus our estimator for τ^2 is (Equation 5.76 in Klugman et al. [1998])

$$\hat{\tau}^2 = \frac{1}{J-1}\sum_{j=1}^{J}\left(\bar{X}_j - \bar{X}\right)^2 - \frac{1}{TJ(T-1)}\sum_{j=1}^{J}\sum_{T=1}^{T}\left(X_{jt} - \bar{X}_j\right)^2. \tag{3.11}$$

All three estimators $\hat{\mu}$, $\hat{\tau}^2$, $\hat{\sigma}^2$ are unbiased. Note that the estimator for $\hat{\tau}^2$ is the difference between two expressions, and so in practice it may yield a negative answer.

This is clearly nonsense as we are estimating a variance component. In such cases, it is common to set its value equal to zero and to take the credibility factor as $z = 0$. Intuitively, this makes sense. If the variance of the hypothetical means is zero (or close to zero), then all risk classes have very similar individual means that do not differ from the overall mean.

Using the data for the example, we can implement the preceding formulas. There are many ways to organize the data necessary for these calculations, and we will take an approach that closely corresponds to the preceding formulas even though it may not be the best approach for a large-scale project. Thus, our first step will be to set some key variables, such as J, T, and X_{jt}, and sort the data X_{jt} by class and time period.

```
J <- length(levels(dta$class))
Tm <- length(unique(dta$time))
cls <- dta$class

o <- order(dta$class, dta$time)
Xjt <- dta$value[o]
```

First, we calculate the overall mean, $\bar{X}$, and the mean for each risk class, $\bar{X}_j$, using Equation 3.6 and Equation 3.7.

```
X.bar <- mean(Xjt)
Xj.bar <- tapply(Xjt, cls, mean)
```

Next we calculate $\hat{\sigma}_j^2$ (Equation 3.8) and the EVPV, $\hat{\sigma}^2$, using Equation 3.9.

```
sigmaj.sq <- tapply(
  (Xjt - rep(Xj.bar, each = Tm))^2, cls, sum) / (Tm - 1)
sigma.sq <- mean(sigmaj.sq)
```

Next comes the calculation of $\text{Var}[\bar{X}_j]$ via Equation 3.10.

```
Var.Xj.bar <- sum((Xj.bar - X.bar)^2) / (J - 1)
```

And, finally, the calculation of the variance of the hypothetical means (Equation 3.11):

```
tau.sq <- Var.Xj.bar - sigma.sq / Tm
```

With all of these values, we have that our credibility factor is equal to

$$Z = \frac{T}{T + \hat{\sigma}^2/\hat{\tau}^2} = \frac{4}{4 + 6250/8437.5} = 0.84375, \tag{3.12}$$

and the credibility premiums will be equal to

$$Z\bar{X}_j + (1-Z)\bar{X},$$

yielding the following credibility-weighted premiums:

```
Z <- Tm / (Tm + sigma.sq / tau.sq)
Z * Xj.bar + (1 - Z) * X.bar
```

```
      1         2         3
665.625  750.000  834.375
```

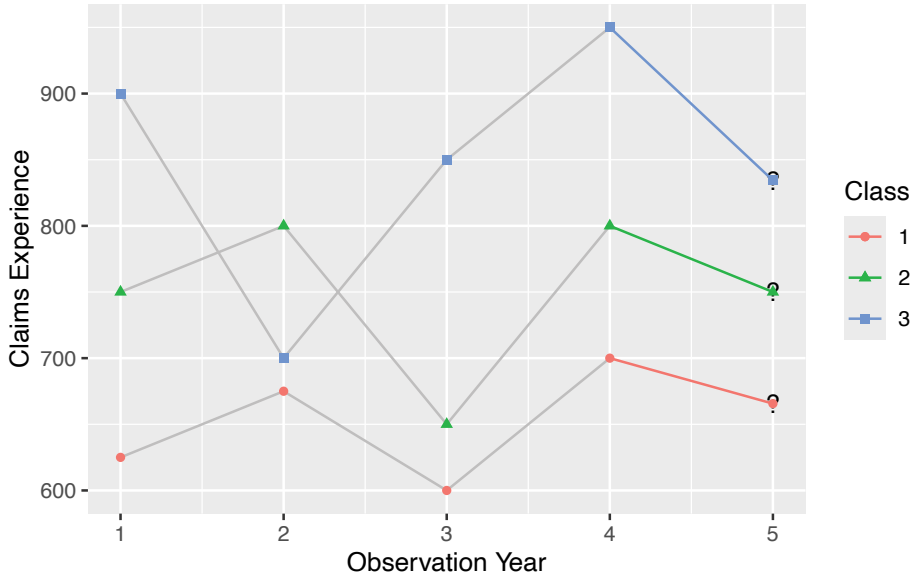
Adding these forecasts to our earlier graph gives us Figure 3.2.

The above calculations for the *balanced* Bühlmann model and other credibility models have been coded into the `cm()` function of the `actuar R` package. We illustrate its use next.

The data is required to be in a different format where the time variable is expressed through different columns in the dataset and the rows represent the different contracts or classes.

```
dtb <- pivot_wider(dta,
                   names_from = time,
                   values_from = value)
dtb
```

Figure 3.2. Claims experience for a portfolio of three risk classes that have been observed over four years. The credibility-weighted estimate for the next year is shown with a colored line segment.



```
# A tibble: 3 x 5
   class    `1´    `2´    `3´    `4´
   <fct>  <dbl>  <dbl>  <dbl>  <dbl>
1  1        625    675    600    700
2  2        750    800    650    800
3  3        900    700    850    950
```

In this case, columns 2 through 5 represent our data. The balanced Bühlmann model, BB, can be fit as follows, and here is a summary of the fitted object:

```
BB <- cm( ~ class,
          data = dtb,
          ratios = 2:5)
summary(BB)
```

```
Call:
cm(formula = ~class, data = dtb, ratios = 2:5)

Structure Parameters Estimators

  Collective premium: 750

  Between class variance: 8437.5
  Within class variance: 6250

Detailed premiums

class  Indiv. mean  Weight  Cred. factor  Cred. premium
1      650          4       0.84375       665.625
2      750          4       0.84375       750.000
3      850          4       0.84375       834.375
```

Here the collective premium is $\bar{X}$; the variance of the hypothetical means τ^2 is labeled "Between class variance"; and the expected value of the process variance σ^2 is the "Within class variance." In the section "Detailed premiums," we see that the individual means $\bar{X}_j$ are in the second column; the next column, "Weight," has the number of observed time periods T; and the credibility factor and the credibility premiums (the last two columns) also match our previous calculations.

The form of the credibility factor may not look particularly nice,

$$Z = \frac{T}{T + \sigma^2/\tau^2},$$

but it has some very appealing and intuitive properties:

1. Since all elements involved are positive, the credibility factor is also positive. Moreover, regardless of the values of T, σ^2, or τ^2, its value is always between zero and 1.

2. As the number of periods of observation T increases, the credibility factor increases toward 1.
3. As the EVPV, σ^2, decreases, the credibility factor increases toward 1.
4. As the VHM, τ^2, increases, the credibility factor increases toward 1.

All these make sense. The more observations you have about your insureds, all else the same, the more confident you should be about their experience. If the process variance is very small, then you should also be more confident about their experience. A small process variance means that the insured's claims experience does not fluctuate too much. And, finally, if the VHM is large, then you know that your insureds have different means, and so you should be more confident about using their own experience versus imposing the overall average experience.

The Bühlmann credibility model has another extremely important property. To calculate the next period's premiums we only need to estimate two parameters: the EVPV, also known as the within-class variance, σ^2, and the VHM, also known as the between-class variance, τ^2. This is always the case regardless of the number of levels the class variable might have.

Unfortunately, the balanced Bühlmann model is not always applicable in practice. One shortcoming is that in the decomposition of the experience X_{jt},

$$X_{jt} = m + \Xi_j + \epsilon_{jt},$$

we have assumed that the deviation Ξ_j from the overall mean m for each risk class j has the same variance, namely, $\text{Var}[\Xi_j] = \sigma^2$. In other words, all risk classes have, in essence, been measured with the same precision. In practice, this may not be a good assumption.

Imagine that the experience X_{jt} is actually the average over individual policyholders that belong in the jth risk class. In this case, the variance across risk classes will not be the same since risk classes will, most likely, have different numbers of policyholders. Another reason for not having equal variances, even if we did have the same number of policyholders, comes about by having policyholders of different sizes within the same risk class. Consider a small supermarket versus a large one.

In these cases, we would want to consider the experience X_{jt} along with a weight w_{jt} such that the bigger the weight, the smaller the variance and vice versa. In the next section, we present the Bühlmann–Straub model, which takes care of these two issues.

3.3. The Bühlmann–Straub Model

In the last section, we saw that the balanced Bühlmann credibility model assumes that each observation in risk class j and time t, X_{jt}, has the same variance, and this may not always reflect reality.

Here, we introduce the Bühlmann–Straub model, which incorporates different weights into the balanced Bühlmann model. We start with the same decomposition of the observations X_{jt} as in the previous section:

$$X_{jt} = m + \Xi_j + \epsilon_{jt}, \quad j = 1, 2, \ldots, J, t = 1, 2, \ldots, T + 1, \tag{3.13}$$

where the unobservable risk components Ξ_j (deviations from the overall mean m for risk class j) are independent and identically distributed with mean zero, and the components ϵ_{jt} (deviations across time from the long-term average of risk class j) are also independent and identically distributed with mean zero. Plus, we assume that Ξ_j and ϵ_{jt} are independent of each other.

Next, we keep the variance of Ξ_j the same as in the previous section,

$$\text{Var}\left[\Xi_j\right] = \tau^2,$$

but we change the assumption for the variance of the components ϵ_{jt} to include the weights w_{jt} to

$$\text{Var}\left[\epsilon_{jt}\right] = \frac{\sigma^2}{w_{jt}}. \tag{3.14}$$

Note that if we set each of the weights w_{jt} equal to 1—that is, let $w_{jt} = 1$ for all j and t—then we are back to the balanced Bühlmann model.

Just as in the balanced Bühlmann model, we would like to find the best homogeneous *unbiased* linear predictor $\sum g_{jt} X_{jt}$ of the risk premium $m + \Xi_j$. The following theorem from Kaas (2009, Theorem 8.4.1, 215) provides the answer using the following quantities and notation. A filled circle, •, in an index location means we are summing across that index. An open circle, ◦, in an index location means we are creating a weighted average over that index with weights provided by the appropriate w_{jt}.

The first expression below sums across all time periods, and so we put a filled circle on the time index. The second expression sums across both indices, time and risk class, and so two filled circles are used.

$$w_{j\bullet} = \sum_{t=1}^{T} w_{jt} \quad \text{and} \quad w_{\bullet\bullet} = \sum_{j=1}^{J}\sum_{t=1}^{T} w_{jt} \tag{3.15}$$

Also note that the first expression above has a value for each value of the index j—therefore we can think of it as a vector of length J. In contrast, the second expression is a single number since we have collapsed along both indices.

The next expression shows us how to compute the J credibility factors. Note that the formula is the same as in the balanced Bühlmann model with T replaced by the sum of the weights across time—that is, $w_{j\bullet}$. Therefore, if all the weights are set equal to 1, then we have that $w_{j\bullet} = T$ for all j, and so this model becomes the balanced Bühlmann model and all J credibility factors Z_j are identical.

$$Z_j = \frac{\tau^2 w_{j\bullet}}{\tau^2 w_{j\bullet} + \sigma^2} = \frac{w_{j\bullet}}{w_{j\bullet} + \sigma^2/\tau^2} \quad \text{and} \quad Z_\bullet = \sum_{j=1}^{J} Z_j \tag{3.16}$$

Finally, the experience X_{jt} can be first summarized by taking a weighted average along the time dimension, leaving us with one average for each risk class. The second

expression then takes the average of the J averages to compute an overall weighted average. The last expression is the weighted average of the individual risk class averages, but using the credibility factors as weights. Keep in mind that the overall average $X_{\circ\circ}$ is, in general, not equal to the credibility-weighted average X_z.

$$X_{j\circ}=\sum_{t=1}^{T}\frac{w_{jt}}{w_{j\bullet}}X_{jt} \quad \text{and} \quad X_{\circ\circ}=\sum_{j=1}^{J}\frac{w_{j\bullet}}{w_{\bullet\bullet}}X_{j\circ} \quad \text{and} \quad X_z=\sum_{j=1}^{J}\frac{Z_j}{Z_\bullet}X_{j\circ} \tag{3.17}$$

Theorem 3.2 (Bühlmann–Straub model). *The mean squared error best homogeneous unbiased predictor $\sum_{it}g_{it}X_{jt}$ of the risk premium $m+\Xi_j$ in model Equation 3.13, that is, the solution to the following restricted minimization problem,*

$$\min_{g_{it}}\mathbb{E}\left[\left(m+\Xi_j-\sum_{it}g_{it}X_{it}\right)^2\right] \tag{3.18}$$

$$\textit{subject to}\ \mathbb{E}\left[m+\Xi_j\right]=\sum_{it}g_{it}\mathbb{E}\left[X_{it}\right], \tag{3.19}$$

is the following credibility estimator:

$$Z_jX_{j\circ}+\left(1-Z_j\right)X_z. \tag{3.20}$$

Theorem 3.2 has the same structure as the balanced Bühlmann theorem. Both results tell us that the next period's pure premium can be estimated as the weighted average of a risk class's average experience $X_{j\circ}$ and the overall average experience for the whole portfolio X_z.

Whereas in the balanced Bühlmann model, the overall average experience is equal to the simple average across all observations, namely,

$$\bar{X}=\frac{\sum_{jt}X_{jt}}{J\bullet T},$$

in the Bühlmann–Straub model, the overall experience should be taken as the credibility-weighted risk class experience X_z, that is,

$$X_z=\sum_{j=1}^{J}\frac{Z_j}{Z_\bullet}X_{j\circ}.$$

Exercise 3.2 In the Bühlmann–Straub model, set all weights w_{jt} equal to 1 and show that you reproduce the results for the balanced Bühlmann model.

Solution 3.2 By letting $w_{jt}=1$ for all $j=1,2,\ldots,J$ and $t=1,2,\ldots,T$, we have that

$$w_{j\bullet}=\sum_{t=1}^{T}w_{jt}=T, \quad \text{and} \quad X_{j\circ}=\sum_{t=1}^{T}\frac{w_{jt}}{w_{j\bullet}}X_{jt}=\frac{1}{T}\sum_{t=1}^{T}X_{jt},$$

and the credibility factors become

$$Z_j = \frac{\tau^2 w_{j\bullet}}{\tau^2 w_{j\bullet} + \sigma^2} = \frac{w_{j\bullet}}{w_{j\bullet} + \sigma^2/\tau^2} = \frac{T}{T + \sigma^2/\tau^2}$$

for all j. Therefore,

$$Z_\bullet = \sum_{j=1}^{J} Z_j = \frac{JT}{T + \sigma^2/\tau^2}.$$

Finally, substituting all the different pieces into X_z gives us

$$X_z = \sum_{j=1}^{J} \frac{Z_j}{Z_\bullet} X_{j\circ} = \sum_{j=1}^{J} \frac{\frac{T}{T + \sigma^2/\tau^2}}{\frac{JT}{T + \sigma^2/\tau^2}} X_{j\circ} = \sum_{j=1}^{J} \frac{1}{J}\left(\frac{1}{T}\sum_{t=1}^{T} X_{jt}\right) = \frac{1}{JT}\sum_{j=1}^{J}\sum_{t=1}^{T} X_{jt} = \bar{X},$$

showing that when the weights in the Bühlmann–Straub model are all equal to 1, we revert back to the balanced Bühlmann model.

For the Bühlmann–Straub model we also need estimators for the model parameters m, σ^2, and τ^2. These estimators are also based on the *sum of squared errors within* and *sum of squared errors between* we have seen before, but incorporating the weights for each observation, namely,

$$\text{SSW} = \sum_{jt} w_{jt}\left(X_{jt} - X_{j\circ}\right)^2,$$

and

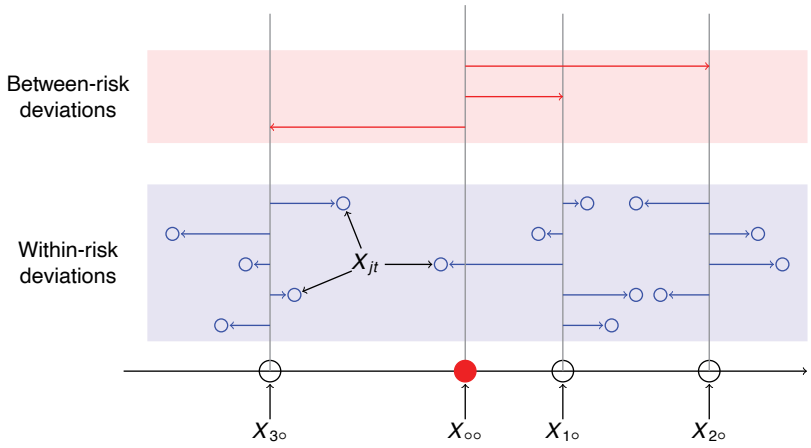
$$\text{SSB} = \sum_{jt} w_{j\bullet}\left(X_{j\circ} - X_{\circ\circ}\right)^2.$$

Figure 3.3 shows one way to think about the between- and within-risk variances. The blue circles represent the actual observations X_{jt} we have available. The large circles on the axis labeled $X_{1\circ}$, $X_{2\circ}$, and $X_{3\circ}$ represent an estimate of the hypothetical means for groups 1, 2, and 3, respectively. For each group, we have the deviations, shown as blue arrows, between the estimated hypothetical mean $X_{j\circ}$ and its actual observations X_{jt}. The magnitude of those differences squared results in the *within* sum of squared errors, SSW.

From the estimated hypothetical means $X_{j\circ}$ we compute an overall mean, shown as a filled red circle, $X_{\circ\circ}$. This represents the collective average for the entire portfolio of risks. The square of the deviations between the overall mean and the hypothetical means, shown in the upper section of Figure 3.3, forms the *between* sum of squared errors, SSB.

Theorem 3.3 (Kaas 2009, Theorem 8.4.2, 218) tells us how to calculate unbiased estimators for m, σ^2, and τ^2.

Figure 3.3. Graphical representation of the within-risk and between-risk deviations. The small open circles represent the actual observations available. The large circles on the horizontal axis are the estimated hypothetical means, and the filled red circle is the overall average. The blue arrows represent the within-risk deviations, and the red arrows are the between-risk deviations.



Theorem 3.3 (Unbiased parameter estimates). *In the Bühlmann–Straub model, the following statistics are unbiased estimators of the corresponding model parameters:*

$$\hat{m} = X_{\circ\circ} \tag{3.21}$$

$$\hat{\sigma}^2 = \frac{1}{J(T-1)} \sum_{jt} w_{jt} \left(X_{jt} - X_{j\circ} \right)^2 \tag{3.22}$$

$$\hat{\tau}^2 = \frac{\sum_j w_{j\bullet} \left(X_{j\circ} - X_{\circ\circ} \right)^2 - (J-1)\hat{\sigma}^2}{w_{\bullet\bullet} - \sum_j w_{j\bullet}^2 / w_{\bullet\bullet}} \tag{3.23}$$

Note that the estimator for τ^2 is the difference between two expressions, and so it is possible that in applying the model the computed value will be negative. In this case, most practitioners will set the value of this parameter to zero.

To illustrate the Bühlmann–Straub model we will generate a synthetic portfolio and compute predictions for the next period. Our discussion follows the development in Kaas (2009, Example 8.4.5, 220).

Let's set up our portfolio with $J = 100$ risk classes and $T = 5$ years of observations that follow the model in Equation 3.13 with the following parameters: $m = 80$, $\tau^2 = 64$, and $\sigma^2 = 100$. We will set up weights ranging from 0.5 to 1.5 and simulate the observations $X_{jt} = m + \Xi_j + \epsilon_{jt}$ with both Ξ_j and ϵ_{jt} as normal random variables with mean

zero and variance τ^2 and σ^2/w_{jt}, respectively. The code to generate the portfolio appears in Listing 3.1.

Before embarking on credibility calculations we should check whether our portfolio is homogeneous. If it is homogeneous, then we do not need to apply credibility and we could estimate the next year's experience by the overall weighted average. To check, we do an analysis of variance.

```
(av <- anova(lm(X.jt ~ j, weights = w.jt)))
```

```
Analysis of Variance Table

Response: X.jt
            Df  Sum Sq   Mean Sq  F value     Pr(>F)
j           99   40395    408.03      3.9  < 2.2e-16  ***
Residuals  400   41850    104.62
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

From the above output we can immediately tell that our portfolio is heterogeneous; the risk classes differ from each other. The *F*-value statistic is too large given the 99 and 400 degrees of freedom (the 5% critical value would be 1.2842).

We can also see immediately from the analysis of variance output that the estimated value of σ^2 is equal to

```
(s2.hat <- av[2,3])
```

```
[1] 104.6239
```

Listing 3.1 Simulation of the Bühlmann–Straub model.

```
J <- 100; Tm <- 5
j <- as.factor(rep(1:J, each = Tm))

m <- 80
t2 <- 64
s2 <- 100

set.seed(12094851)
w.jt <- 0.5 + runif(J * Tm)
X.jt <- m + rep(rnorm(J, 0, sqrt(t2)), each = Tm) +
            rnorm(J * Tm, 0, sqrt(s2/w.jt))

dta <- tibble(risk = j,
              X.jt = X.jt,
              W.jt = w.jt)
write_csv(dta, "BS-simulated-data.csv")
```

The estimator for the overall mean m is the overall weighted average for the data.

```
(m.hat <- X.cc <- sum(w.jt * X.jt) / sum(w.jt))
```

```
[1] 78.43628
```

Finally, we need to implement the calculation for the estimator of τ^2. We start with some preliminary setup where the sum of the weights across time $w_{j\bullet}$ corresponds to `w.jb`, the sum of all weights $w_{\bullet\bullet}$ is w.bb, and the weighted average of the experience across time for each risk class $X_{j\circ}$ is given by `X.jc`.

```
w.jb <- tapply(w.jt, j, sum)
w.bb <- sum(w.jb)
X.jc <- tapply(w.jt * X.jt / w.jb[j], j, sum)
```

Hence, the estimator for τ^2 is given by

```
num <- sum(w.jb * (X.jc - X.cc)^2) - (J - 1) * s2.hat
den <- w.bb - sum(w.jb^2 / w.bb)
(t2.hat <- num / den)
```

```
[1] 60.9652
```

Finally, the credibility factors are

```
Zj.hat <- w.jb / (w.jb + s2.hat / t2.hat)
```

and the collective premium is

```
(X.z <- sum(Zj.hat * X.jc) / sum(Zj.hat))
```

```
[1] 78.53833
```

and putting them together we obtain the following credibility premiums

```
P.hat <- Zj.hat * X.jc + (1 - Zj.hat) * X.z
```

The first 20 estimated credibility premiums are

```
P.hat[1:20]
```

```
       1        2        3        4        5        6
67.41990 74.65462 82.37383 81.77987 75.53623 84.26384
       7        8        9       10       11       12
80.16259 71.72899 95.02067 71.41841 86.42743 80.35881
      13       14       15       16       17       18
90.36646 73.34605 74.71319 78.70976 71.75083 92.52850
      19       20
72.31943 69.29793
```

And as we did for the Bühlmann model, we can use the function `cm()` from package `actuar` to do the calculations necessary for the Bühlmann–Straub model. We first create a data frame with both the observations X_{jt} and the weights w_{jt} along with a column telling us which row of data belongs to which risk class.

First, we create a data frame with the data we have

```
D <- cbind(risk.class = 1:J,
           as.data.frame(matrix(X.jt,
                                nrow = J,
                                ncol = Tm,
                                byrow = TRUE)),
           as.data.frame(matrix(w.jt,
                                nrow = J,
                                ncol = Tm,
                                byrow = TRUE)))
```

and then we estimate the model via

```
(BS <- cm(~ risk.class,
          data = D,
          ratios = 2:6,
          weights = 7:11))
```

```
Call:
cm(formula = ~risk.class, data = D, ratios = 2:6, weights = 7:11)

Structure Parameters Estimators

  Collective premium: 78.53833

  Between risk.class variance: 60.9652
  Within risk.class variance: 104.6239
```

For the first five risk classes, our by-hand calculations match those from the `cm()`

function.

```
rbind("    cm():" = predict(BS)[1:5],
       "by-hand:" = P.hat[1:5])
                1         2         3         4         5
cm():  67.4199  74.65462  82.37383  81.77987  75.53623
by-hand:  67.4199  74.65462  82.37383  81.77987  75.53623
```

And they match across all risk classes:

```
all(round(abs(predict(BS) - P.hat), 10) == 0)
[1] TRUE
```

3.4. Hachemeister Regression

In this chapter we have been working with the basic model

$$X_{jt} = m + \Xi_j + \epsilon_{jt},$$

where j indexes a risk class and t represents time. Both Ξ_j and ϵ_{jt} are random variables, and m is a fixed parameter.

We can extend the basic model in many ways. For example, consider a tree-like structure of an insurance line of business where at the top level we have the entire business. This can be divided into different sectors, and then each sector can again be divided into risk classes. Finally, each risk class has the observed experience. This model is known as Jewell's hierarchical model (Jewell 1975), and the statistical model can be written as

$$X_{sjt} = m + \Xi_s + \Xi_{sj} + \epsilon_{sjt},$$

where $s = 1, 2, \ldots, S$ represents the different sectors, $j = 1, 2, \ldots, J$ corresponds to the different risk classes, and $t = 1, 2, \ldots, T + 1$ indexes the time periods. Extensions to more levels follow the same pattern.

In this case, Ξ_s is the deviation from the overall mean m for sector s, Ξ_{sj} is then the deviation from the sector mean for risk class j, and ϵ_{sjt} represents the fluctuations through time.

In this section, we are interested in a different extension of the basic model that will take us in the direction of the classical linear regression model. This model, known by actuaries as a credibility regression model, was first introduced to the actuarial literature by Hachemeister (1975).

Inspired by the high inflation rates starting in the late 1960s and continuing into the early 1970s, Hachemeister became interested in understanding loss severity trends and used data from private passenger automobile insurance (bodily injury coverage) for five different states to illustrate his ideas. The data is on a quarterly basis from Q3 1970 through Q2 1973 (12 quarters of observations).

Table 3.2 shows the claim severity and the number of claims by state. Note that state 1 has a very large number of claims in each quarter, and state 4 has the least number of claims of all states. The ratio of the number of claims, in each quarter, for state 1 to state 4 is almost always in excess of 20. Figure 3.4 displays a multiple time series plot for severity. The data for each state has been connected with a light gray line, and each state has been labeled on the right-hand side. State 4 has been highlighted with thick connecting segments to illustrate the data for a single state.

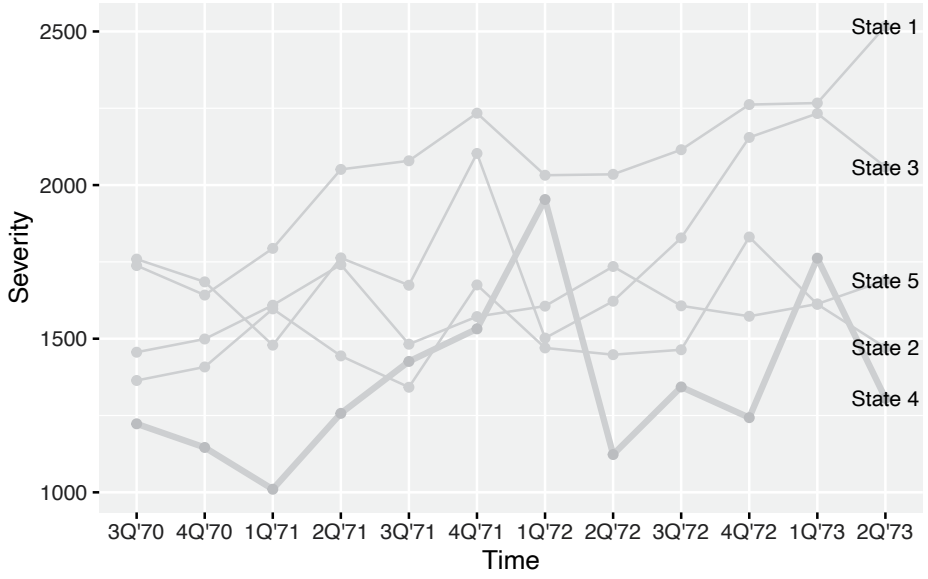
Notice that the variability in severity from quarter to quarter for state 4 is much greater than for other states, and state 1 appears to be the most stable. This feature arises because the volume of claims underlying the severity is quite different between state 1 and state 4.

The severity trend for each state is positive, and we could measure its magnitude by fitting a weighted least squares regression line to each state. Doing so yields the five light purple straight lines shown in Figure 3.5. We have also included a "countrywide" (data combined for all five states) weighted regression line shown in red.

Table 3.2. The Hachemeister data. Number of claims and severity for five different states from private passenger automobile insurance (bodily injury coverage).

	Claim Severity by State					Number of Claims by State				
Period	1	2	3	4	5	1	2	3	4	5
3Q'70	1,738	1,364	1,759	1,223	1,456	7,861	1,622	1,147	407	2,902
4Q'70	1,642	1,408	1,685	1,146	1,499	9,251	1,742	1,357	396	3,172
1Q'71	1,794	1,597	1,479	1,010	1,609	8,706	1,523	1,329	348	3,046
2Q'71	2,051	1,444	1,763	1,257	1,741	8,575	1,515	1,204	341	3,068
3Q'71	2,079	1,342	1,674	1,426	1,482	7,917	1,622	998	315	2,693
4Q'71	2,234	1,675	2,103	1,532	1,572	8,263	1,602	1,077	328	2,910
1Q'72	2,032	1,470	1,502	1,953	1,606	9,456	1,964	1,277	352	3,275
2Q'72	2,035	1,448	1,622	1,123	1,735	8,003	1,515	1,218	331	2,697
3Q'72	2,115	1,464	1,828	1,343	1,607	7,365	1,527	896	287	2,663
4Q'72	2,262	1,831	2,155	1,243	1,573	7,832	1,748	1,003	384	3,017
1Q'73	2,267	1,612	2,233	1,762	1,613	7,849	1,654	1,108	321	3,242
2Q'73	2,517	1,471	2,059	1,306	1,690	9,077	1,861	1,121	342	3,425

Figure 3.4. Hachemeister data showing quarterly experience for private passenger auto (bodily injury coverage) severity for five different states from 3Q 1970 to 2Q 1973 (12 observations). Light gray lines connect the observations for each state to emphasize the state individual trends. State 4 is highlighted with thicker line segments.



Looking at the trend lines for each state shown in Figure 3.5 we can see that each one has a different level of severity and each one has a different positive slope. State 1 has the largest slope, and state 5 has the smallest.

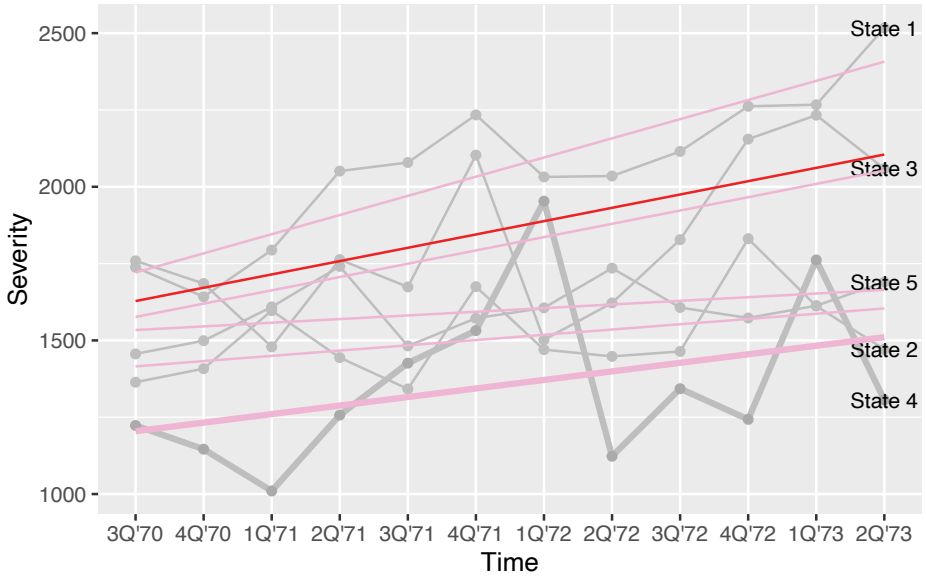
The basic credibility model that we have been working with cannot accommodate this setup, but we can extend it as follows:

$$X_{jt} = \left\{ m^{(1)} + \Xi_j^{(1)} \right\} + \left\{ m^{(2)} + \Xi_j^{(2)} \right\} q_{jt} + \epsilon_{jt}. \tag{3.24}$$

where $m^{(1)}$ and $m^{(2)}$ are unknown fixed parameters, $\Xi_j^{(1)}$ and $\Xi_j^{(2)}$ are random variables with mean zero, and q_{jt} is, in this particular case, the explanatory variable time but in general could be any explanatory variable.

If the random variables $\Xi_j^{(1)}$ and $\Xi_j^{(2)}$ are identically zero (that is, their variance is zero), then the model becomes a classical regression. If only $\Xi_j^{(2)}$ is identically zero, then we have a random intercept model. In our example, every state would have its own intercept, but all states would share the same slope. That is, we would have parallel regression lines. We can estimate such a model via least squares by including an indicator variable for each state (and omitting the intercept). If only $\Xi_j^{(1)}$ is identically zero, then all states have the same intercept but each one has a different slope. This would be a random slope model.

Figure 3.5. Hachemeister data including individual regression lines for each state (shown in light purple) and a "countrywide" regression line (shown in red) for the combined data of all five states. The data for state 4 is connected with thick gray lines, and its corresponding individual regression line is also shown with a thick purple line.



The regression lines shown in light purple in Figure 3.5 have been estimated using *only* the information contained in each state. If we think the information in each state is fully credible, these estimated regression lines are appropriate. If, on the other hand, the information in each state is not credible, we would ignore the `state` grouping variable by using all the data together to estimate the collective regression line.

As we have seen with the Bühlmann and the Bühlmann–Straub models, these are two extreme positions, and we can search for a compromise between the two. So Hachemeister set out to develop a credibility regression model. We will follow the discussion in Chapter 8 of Bühlmann and Gisler (2005) closely and focus on the example presented above of linear regression (intercept and slope) even though the ideas clearly apply to more general regression models. Also, we will change our notation slightly and use vectors and matrices to make the connection to linear regression more apparent. To this end, we can restate Equation 3.24 as follows

$$X_j = Q_j \beta\left(\Theta_j\right) + \epsilon_j, \tag{3.25}$$

where X_j is a $T \times 1$ column vector of responses for risk $j = 1, 2, \ldots, J$ and T is the number of time periods we have observed. In Hachemeister's example, this corresponds to the observed severities for state $j = 1, 2, \ldots, 5$ during time periods $t = 1, 2, \ldots, 12$, and therefore we have

$$X_j = \begin{bmatrix} X_{j1} \\ X_{j2} \\ \vdots \\ X_{jT} \end{bmatrix}.$$

The matrix Q_j is our regression design matrix for state j with dimensions $T \times 2$:

$$Q_j = \begin{bmatrix} 1 & t_{j1} \\ 1 & t_{j2} \\ \vdots & \vdots \\ 1 & t_{jN} \end{bmatrix}.$$

The first column of Q_j corresponds to the intercept, and the second column is the time variable. For our example, we have $T_{j1} = 1$, $t_{j2} = 2, \ldots,$ for all states j. The 2×1 vector $\beta(\Theta_j)$ is our regression coefficient. The reason for having the β's depend on Θ_j is that we are thinking of these regression coefficients as dependent on a random variable for each state. Thus, we have

$$B(\Theta_j) = \begin{bmatrix} \beta_0(\Theta_j) \\ \beta_1(\Theta_j) \end{bmatrix},$$

where $\beta_0(\Theta_j)$ is our intercept and $\beta_1(\Theta_j)$ is our slope for state j. Finally, the $T \times 1$ vector ϵ_j of error terms is

$$\epsilon_j = \begin{bmatrix} \epsilon_{j1} \\ \epsilon_{j2} \\ \vdots \\ \epsilon_{jT} \end{bmatrix}.$$

The following definition and theorem come from Chapter 8 of Bühlmann and Gisler (2005, 205, 207).

Definition 3.1 (Hachemeister model assumptions). The risk j is characterized by an individual risk profile ϑ_j, which is itself the realization of a random variable Θ_j. We make the following assumptions:

Conditionally, given Θ_j, the entries X_{jt}, $j = 1, \ldots, J$ are independent, and we have

$$\mathbb{E}\left[X_j|\Theta_j\right]=Q_j\beta\left(\Theta_j\right),$$

where $\beta(\Theta_j)$ is the regression vector and Q_j is the known design matrix, and

$$\mathrm{Cov}\left[X_j, X_j'|\Theta_j\right]=\Sigma_j\left(\Theta_j\right)$$

is the covariance matrix conditional on Θ_j. The pairs (Θ_1, X_1), (Θ_2, X_2), . . . are independent, and also $\Theta_1, \Theta_2, \ldots$ are independent and identically distributed.

Theorem 3.4 (Hachemeister formula). *Under the Hachemeister model assumptions we get that the credibility estimator for* $\beta(\Theta_j)$ *satisfies*

$$\widehat{\beta\left(\Theta_j\right)}=A_jB_j+\left(I-A_j\right)\beta, \tag{3.26}$$

where

$$A_j=U\left(U+\left(Q_j'S_j^{-1}Q_j\right)^{-1}\right)^{-1},$$

$$B_j=\left(Q_j'S_j^{-1}Q_j\right)^{-1}Q_j'S_j^{-1}X_j,$$

$$S_j=\mathbb{E}\left[\Sigma_j\left(\Theta_j\right)\right]=\mathbb{E}\left[Cov\left[X_j, X_j'|\Theta_j\right]\right],$$

$$U=Cov\left[\beta\left(\Theta_j\right),\beta\left(\Theta_j\right)'\right],$$

$$\beta=\mathbb{E}\left[\beta\left(\Theta_j\right)\right].$$

The credibility matrices A_j are for the example we are considering of dimension 2×2. The 2×1 vector B_j is the intercept and slope for each state j. The matrices S_j have dimension $T\times T$ and are of the form

$$S_j=\sigma^2\begin{bmatrix} w_{j1} & 0 & \cdots & 0\\ 0 & w_{j2} & \cdots & 0\\ \vdots & \vdots & \ddots & \vdots\\ 0 & 0 & \cdots & w_{jT}\end{bmatrix}^{-1},$$

where the diagonal entries w_{jt} are known weights. We will use W_j as the diagonal matrix with entries w_{jt} along the main diagonal and zeroes everywhere else. Thus, we can write $S_j=\sigma^2W_j^{-1}$. In our example, the w_{jt} are the number of claims at time t in state j.

The matrix U, of dimension 2×2, is the variance-covariance matrix of the estimated coefficients. Because the matrix U is symmetric, there are at most three distinct entries.

And, finally, the 2×1 vector β is the collective intercept and slope.

Let us apply the Hachemeister formula (Theorem 3.4) to the data we have at hand. In several places we must calculate the product $Q_j' S_j^{-1} Q_j$, with S_j being the diagonal matrix in the earlier paragraph. This product is closely related to $Q_j' W_j Q_j$, which comes up several times, so let us define

$$V_j = Q_j' W_j Q_j,$$

which is a 2×2 matrix. Specializing for the Hachemeister data, we have

$$V_j = \begin{bmatrix} 1 & 1 & \cdots & 1 \\ 1 & 2 & \cdots & T \end{bmatrix} \begin{bmatrix} w_{j1} & 0 & \cdots & 0 \\ 0 & w_{j2} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & w_{jT} \end{bmatrix} \begin{bmatrix} 1 & 1 \\ 1 & 2 \\ \vdots & \vdots \\ 1 & T \end{bmatrix}.$$

Doing the matrix multiplications gives us the following:

$$V_j = \begin{bmatrix} \sum_{t=1}^{T} w_{jt} & \sum_{t=1}^{T} t w_{jt} \\ \sum_{t=1}^{T} t w_{jt} & \sum_{t=1}^{T} t^2 w_{jt} \end{bmatrix}.$$

The entries in the matrix V_j almost look like weighted averages. They are missing a denominator equal to $\sum_{t=1}^{T} w_{jt}$. To keep the notation cleaner, we use the same convention as before: a $\bullet$ symbol in an index position means that we sum all entries along that index. Hence, we have $w_{j\bullet} = \sum_{t=1}^{T} w_{jt}$. Therefore, we can rewrite the above matrix V_j as

$$V_j = w_{j\bullet} \begin{bmatrix} \sum_{t=1}^{T} \frac{w_{jt}}{w_{j\bullet}} & \sum_{t=1}^{T} t \frac{w_{jt}}{w_{j\bullet}} \\ \sum_{t=1}^{T} t \frac{w_{jt}}{w_{j\bullet}} & \sum_{t=1}^{T} t^2 \frac{w_{jt}}{w_{j\bullet}} \end{bmatrix}.$$

To simplify the notation further, note that the off-diagonal entries look like the calculation of the expected value of t because the weights $w_{jt}/w_{j\bullet}$ sum to 1, and so we are thinking of them as a sampling distribution. Similarly, the entry in position (2, 2) looks like the calculation of the second moment.

In view of this, we define the following notations:

$$E_j^{(s)}[t] = \sum_{t=1}^{T} \frac{w_{jt}}{w_{j\bullet}} t \quad \text{and} \quad E_j^{(s)}[X_j] = \sum_{t=1}^{T} \frac{w_{jt}}{w_{j\bullet}} X_{jt},$$

where the superscript (s) signals that we are thinking of the weights $w_{jt}/w_{j\bullet}$ as a sampling distribution. With this notation, we can also write

$$\text{Var}_j^{(s)}[t] = E_j^{(s)}[t^2] - \left(E_j^{(s)}[t]\right)^2.$$

Using all of this, the matrix V_j is now

$$V_j = w_{j\bullet} \begin{bmatrix} 1 & E_j^{(s)}[t] \\ E_j^{(s)}[t] & E_j^{(s)}[t^2] \end{bmatrix},$$

and note that its determinant is equal to $\det(V_j) = w_{j\bullet}^2 \text{Var}^{(s)}[t]$.

Using Theorem 3.4 we can calculate B_j as follows:

$$\begin{aligned} B_j &= V_j^{-1} Q_j' W_j X_j \\ &= \frac{1}{w_{j\bullet} \text{Var}^{(s)}[t]} \begin{bmatrix} E_j^{(s)}[t^2] & -E_j^{(s)}[t] \\ -E_j^{(s)}[t] & 1 \end{bmatrix} \begin{bmatrix} \sum_{t=1}^{T} w_{jt} X_{jt} \\ \sum_{t=1}^{T} w_{jt} t X_{jt} \end{bmatrix} \\ &= \frac{1}{\text{Var}^{(s)}[t]} \begin{bmatrix} E_j^{(s)}[t^2] & -E_j^{(s)}[t] \\ -E_j^{(s)}[t] & 1 \end{bmatrix} \begin{bmatrix} \sum_{t=1}^{T} \frac{w_{jt}}{w_{j\bullet}} X_{jt} \\ \sum_{t=1}^{T} \frac{w_{jt}}{w_{j\bullet}} t X_{jt} \end{bmatrix} \\ &= \frac{1}{\text{Var}^{(s)}[t]} \begin{bmatrix} E_j^{(s)}[t^2] & -E_j^{(s)}[t] \\ -E_j^{(s)}[t] & 1 \end{bmatrix} \begin{bmatrix} E_j^{(s)}[X_{jt}] \\ E_j^{(s)}[tX_{jt}] \end{bmatrix} \\ &= \frac{1}{\text{Var}^{(s)}[t]} \begin{bmatrix} E_j^{(s)}[t^2] E_j^{(s)}[X_{jt}] - E_j^{(s)}[t] E_j^{(s)}[tX_{jt}] \\ E_j^{(s)}[tX_{jt}] - E_j^{(s)}[t] E_j^{(s)}[X_{jt}] \end{bmatrix}. \end{aligned}$$

Let us implement these calculations using Hachemeister's data. First, let's define some quantities: the weights `W.jt`, the time points `T.jt`, the observed severities `X.jt`, and the vector `S`, which tells us which state these observations belong to.

```
W.jt <- db$claims
T.jt <- db$time
```

```
X.jt <- db$severity
S <- db$state
N <- length(unique(T.jt))
J <- length(unique(S))
```

Next, we calculate the summary statistics that we will need. Table 3.3 shows the correspondence between our programming variable names and the written notation used in the text.

```
W.jb <- tapply(W.jt, S, sum)
```

Table 3.3. Correspondence between programming variable and written notation in the text.

Variable	Written Notation	Variable	Written Notation
`W.jb`	$w_{j\bullet}$	`Ej.tX`	$E_j^{(s)}[tX_j]$
`W.bb`	$\sum_{j=1}^{J} w_{j\bullet}$	`Vj.t`	$\text{Var}_j^{(s)}[t]$
`Ej.t`	$E_j^{(s)}[t]$	`Ws.jb`	$\text{Var}_j^{(s)}[t]w_{j\bullet}$
`Ej.t2`	$E_j^{(s)}[t^2]$	`Ws.bb`	$\sum_{j=1}^{J}\text{Var}_j^{(s)}[t]w_{j\bullet}$
`Ej.X`	$E_j^{(s)}[X_j]$		

```
W.bb <- sum(W.jb)
Ej.t <- tapply(W.jt * T.jt, S, sum) / W.jb
Ej.t2 <- tapply(W.jt * T.jt^2, S, sum) / W.jb
Ej.X <- tapply(W.jt * X.jt, S, sum) / W.jb
Ej.tX <- tapply(W.jt * T.jt * X.jt, S, sum) / W.jb
Vj.t <- Ej.t2 - Ej.t^2
Ws.jb <- Vj.t * W.jb
Ws.bb <- sum(Ws.jb)
```

With these definitions, we can calculate the intercept and slope for each state using

```
Bj <- rbind((Ej.t2 * Ej.X - Ej.t * Ej.tX) / Vj.t,
            (Ej.tX - Ej.t * Ej.X) / Vj.t)
dimnames(Bj) <- list(c("Intercept", "Slope"),
                     1:5)
round(Bj, 2)
```

```
                1       2       3       4       5
Intercept 1658.47 1398.30 1533.00 1176.70 1521.90
Slope       62.39   17.14   43.31   27.81   11.87
```

Exercise 3.3 Verify that the intercept and slope we calculated for each state are correct by doing it the easy way; that is, fit a weighted linear regression to the data for each state.

Solution 3.3 For state 4 we would compute as follows:

```
lm.st4 <- lm(severity ~ time,
             data = db,
             subset = state == 4,
             weights = claims)
coef(lm.st4)
```

```
(Intercept)        time
 1176.70407   27.80702
```

And indeed these coefficients match those we computed earlier. We leave similar calculations, for the remaining states, to the reader.

Next, on page 209 of Bühlmann and Gisler (2005), they assume that the 2×2 matrix $U = \text{Cov}[\beta(\Theta_j), \beta(\Theta_j)']$ is diagonal with entries τ_0^2 and τ_1^2, that is,

$$U = \begin{bmatrix} \tau_0^2 & 0 \\ 0 & \tau_1^2 \end{bmatrix}.$$

This implies that the intercept and the slope are independent of each other.

The credibility matrices A_j are given in Theorem 3.4 as

$$A_j = U\left(U + \left(Q_j' S_j^{-1} Q_j\right)^{-1}\right)^{-1},$$

and we can rewrite, by using V_j, as follows

$$A_j = U\left(U + \sigma^2 V_j^{-1}\right)^{-1}.$$

We could substitute expressions for U and V_j and compute, but that requires a lot of calculations. It is best to rewrite as follows:

$$\begin{aligned} A_j &= U\left(U + \sigma^2 V_j^{-1}\right)^{-1} \\ &= \left(I + \sigma^2 V_j^{-1} U^{-1}\right)^{-1} \\ &= \left(V_j + \sigma^2 U^{-1}\right)^{-1} V_j. \end{aligned}$$

The inverse of U is easy because U is a 2×2 diagonal matrix. We have

$$U^{-1} = \frac{1}{\tau_0^2 \tau_1^2} \begin{bmatrix} \tau_1^2 & 0 \\ 0 & \tau_0^2 \end{bmatrix},$$

and noting that we need to multiply by σ^2, we can make the following substitutions. Let $\kappa_0 = \sigma^2/\tau_0^2$ and $\kappa_1 = \sigma^2/\tau_1^2$, then we have

$$U^{-1} = \frac{1}{\sigma^2} \begin{bmatrix} \kappa_0 & 0 \\ 0 & \kappa_1 \end{bmatrix},$$

and so

$$V_j + \sigma^2 U^{-1} = \begin{bmatrix} w_{j\bullet} + \kappa_0 & w_{j\bullet} E_j^{(s)}[t] \\ w_{j\bullet} E_j^{(s)}[t] & w_{j\bullet} E_j^{(s)}[t^2] + \kappa_1 \end{bmatrix}.$$

The inverse of the above matrix is

$$\left(V_j + \sigma^2 U^{-1}\right)^{-1} = \frac{1}{D} \begin{bmatrix} w_{j\bullet} E_j^{(s)}[t^2] + \kappa_1 & -w_{j\bullet} E_j^{(s)}[t] \\ -w_{j\bullet} E_j^{(s)}[t] & w_{j\bullet} + \kappa_0 \end{bmatrix},$$

where D is the determinant given by

$$D = \left(w_{j\bullet} + \kappa_0\right)\left(w_{j\bullet} E_j^{(s)}[t^2] + \kappa_1\right) - \left(w_{j\bullet} E_j^{(s)}[t]\right)^2.$$

Finally, multiplying the above expression by V_j and recalling that $\mathrm{Var}_j^{(s)}[t] = E_j^{(s)}[t^2] - (E_j^{(s)}[t])^2$, we obtain the credibility matrix A_j as

$$A_j = \frac{w_{j\bullet}}{D} \begin{bmatrix} w_{j\bullet} \mathrm{Var}_j^{(s)}[t] + \kappa_1 & \kappa_1 E_j^{(s)}[t] \\ \kappa_0 E_j^{(s)}[t] & w_{j\bullet} \mathrm{Var}_j^{(s)}[t] + \kappa_0 E_j^{(s)}[t^2] \end{bmatrix}. \tag{3.27}$$

To compute the credibility matrices A_j for the Hachemeister data, we will need to find estimators (Bühlmann and Gisler 2005, 216–217) for the three parameters σ^2, τ_0^2, and τ_1^2. They are as follows. Consider an estimator of the variance across time for a single state j:

$$\hat{\sigma}_j^2 = \frac{1}{n-2} \sum_{t=1}^{T} w_{jt} \left(X_{jt} - \hat{\mu}_{jt}\right)^2, \tag{3.28}$$

where $\hat{\mu}_{jt}$ are the fitted values from the regression equation for state j. These we can compute via

```
Ys <- cbind(rep(1, 12), 1:12)
mu.jt <- as.vector(Ys %*% Bj)
sigmaj.sq <- tapply(W.jt * (X.jt - mu.jt)^2, S, sum) / (N - 2)
```

Then, take as an estimator for σ^2 the average of the individual state estimators, that is,

$$\sigma^2 = \frac{1}{J}\sum_{j=1}^{J} \hat{\sigma}_j^2. \tag{3.29}$$

```
sigma.sq <- mean(sigmaj.sq)
```

For the estimators of the variances of the intercept τ_0^2 and slope τ_1^2, we use estimators that are analogous to those in the Bühlmann–Straub model. That is,

$$\hat{\tau}_0^2 = c_0\left\{\frac{J}{J-1}\sum_{j=1}^{J}\frac{w_{j\bullet}}{w_{\bullet\bullet}}\left(B_{0j} - \bar{B}_0\right)^2 - \frac{J\hat{\sigma}^2}{w_{\bullet\bullet}}\right\}, \tag{3.30}$$

where

$$c_0 = \frac{J-1}{J}\left\{\sum_{j=1}^{J}\frac{w_{j\bullet}}{w_{\bullet\bullet}}\left(1 - \frac{w_{j\bullet}}{w_{\bullet\bullet}}\right)\right\}^{-1}, \tag{3.31}$$

and

$$\bar{B}_0 = \sum_{j=1}^{J}\frac{w_{j\bullet}}{w_{\bullet\bullet}}B_{0j}.$$

We can compute these quantities by starting with $\bar{B}_0$.

```
B0.bar <- sum(W.jb / W.bb * Bj[1,])
```

Then we will need c_0 (Equation 3.31), which we will break up into smaller terms,

$$\text{term}_1 = \frac{J-1}{J}, \quad \text{term}_2 = \frac{w_{j\bullet}}{w_{\bullet\bullet}}, \quad \text{term}_3 = 1 - \frac{w_{j\bullet}}{w_{\bullet\bullet}}, \quad \text{term}_4 = \sum_{j=1}^{J}\frac{w_{j\bullet}}{w_{\bullet\bullet}}\left(1 - \frac{w_{j\bullet}}{w_{\bullet\bullet}}\right),$$

and finally calculating c_0 as

$$c_0 = \frac{\text{term}_1}{\text{term}_4}.$$

```
term.1 <- (J - 1) / J
term.2 <- W.jb / W.bb
term.3 <- 1 - term.2
term.4 <- sum(term.2 * term.3)
c0 <- term.1 / term.4
```

And for the calculation of τ_0^2 (Equation 3.30), we also break it up into more manageable pieces:

$$\text{term}_1 = \frac{J}{J-1}, \quad \text{term}_2 = \frac{w_{j\bullet}}{w_{\bullet\bullet}}, \quad \text{term}_3 = \left(B_{0j} - \bar{B}_0\right)^2, \quad \text{term}_4 = \frac{J\sigma^2}{w_{\bullet\bullet}}$$

```
term.1 <- J / (J - 1)
term.2 <- W.jb / W.bb
term.3 <- (Bj[1,] - B0.bar)^2
term.4 <- J * sigma.sq / W.bb
tau0.sq <- c0 * (term.1 * sum(term.2 * term.3) - term.4)
```

And similarly for τ_1^2. The formulas are the same with a small change. Instead of using $w_{j\bullet}$, we use $w^*_{j\bullet}$, where

$$w^*_{j\bullet} = Var_j^{(s)}\left[t\right] \bullet w_{j\bullet}.$$

Therefore, we have

$$\hat{\tau}_1^2 = c_1 \left\{ \frac{J}{J-1} \sum_{j=1}^{J} \frac{w^*_{j\bullet}}{w^*_{\bullet\bullet}} \left(B_{1j} - \bar{B}_1\right)^2 - \frac{J\hat{\sigma}^2}{w^*_{\bullet\bullet}} \right\}, \tag{3.32}$$

where

$$c_1 = \frac{J-1}{J} \left\{ \sum_{j=1}^{J} \frac{w^*_{j\bullet}}{w^*_{\bullet\bullet}} \left(1 - \frac{w^*_{j\bullet}}{w^*_{\bullet\bullet}} \right) \right\}^{-1},$$

and

$$\bar{B}_1 = \sum_{j=1}^{J} \frac{w^*_{j\bullet}}{w^*_{\bullet\bullet}} B_{1j}.$$

The values of B_{0j} and B_{1j} are the intercept and slope, respectively, for each individual state j, and so we have that $\bar{B}_0$ and $\bar{B}_1$ are the weighted averages of the estimated state parameters.

The calculation for τ_1^2 (Equation 3.32) is analogous to τ_0^2. The code to accomplish this follows:

```
B1.bar <- sum(Ws.jb / Ws.bb * Bj[2,])
```

```
term.1 <- (J - 1) / J
term.2 <- Ws.jb / Ws.bb
term.3 <- 1 - term.2
term.4 <- sum(term.2 * term.3)
c1 <- term.1 / term.4
```

```
term.1 <- J / (J - 1)
term.2 <- Ws.jb / Ws.bb
term.3 <- (Bj[2,] - B1.bar)^2
term.4 <- J * sigma.sq / Ws.bb
tau1.sq <- c1 * (term.1 * sum(term.2 * term.3) - term.4)
```

These calculations yield

$$\hat{\sigma}^2 = 4.9870187 \times 10^7, \quad \hat{\tau}_0^2 = 1.8029435 \times 10^4, \quad \hat{\tau}_0^2 = 665.5618.$$

With these parameter estimates we can now calculate the credibility matrices A_j, the collective intercept and slope, and the credibility-weighted estimates for each state. Note that the Hachemeister data had five states, but even if it had data for all 50 states, we would still need to estimate only three parameters. Moreover, these parameters are not specific to these five states. We have treated these states as coming from a population of states, and these parameters estimate features of the population.

Now that we have estimates for σ^2, τ_0^2, and τ_1^2, we can calculate the credibility-weighted estimates of the intercept and slope for our states. The following code sets up the necessary quantities:

```
k0 <- sigma.sq / tau0.sq
k1 <- sigma.sq / tau1.sq

Determ <- (W.jb + k0) * (W.jb * Ej.t2 + k1) - (W.jb * Ej.t)^2
term.11 <- W.jb * Vj.t + k1
term.12 <- k1 * Ej.t
term.21 <- k0 * Ej.t
term.22 <- W.jb * Vj.t + k0 * Ej.t2
```

Using Equation 3.27 we can define a function that will return the credibility matrix A_j as

```
A <- function(j) {
  M <- matrix(c(term.11[j], term.12[j],
                term.21[j], term.22[j]),
              nrow = 2, ncol = 2, byrow = TRUE)
  ans <- W.jb[j] / Determ[j] * M
  dimnames(ans) <- list(c("", ""),
                        c("", ""))
  return(ans)
}
```

We will also need the collective's estimate of the intercept and slope. This estimate is equal to the weighted average of the individual state estimates where the weights are given by the credibility matrices as follows:

$$B_{\text{GLS}} = \left(\sum_{j=1}^{J} A_j\right)^{-1} \sum_{j=1}^{J} A_j B_j,$$

where B_j is the estimate of the intercept and slope for state j. We have labeled the estimate with the subscript "GLS" because it turns out that this estimate is equal to the generalized least squares (GLS) estimate.

```
B.gls <- solve(A(1) + A(2) + A(3) + A(4) + A(5)) %*%
  (A(1) %*% Bj[,1] + A(2) %*% Bj[,2] + A(3) %*% Bj[,3] +
     A(4) %*% Bj[,4] + A(5) %*% Bj[,5])
```

The credibility-weighted estimate for state j, which we label as CW_j, is equal to

$$CW_j = A_j B_j + \left(I - A_j\right) B_{\text{GLS}},$$

where I is a 2×2 identity matrix.

```
CW <- function(j){
  I <- diag(1, nrow = 2, ncol = 2)
  ans <- A(j) %*% Bj[,j] + (I - A(j)) %*% B.gls
  dimnames(ans) <- list(c("Intercept", "Slope"),
                        paste("State", j, sep = " "))
  return(ans)
}
```

For state 1, the credibility matrix A_1 is

```
round(A(1), 4)
```

```
0.8946  0.3389
0.0125  0.9460
```

and the credibility-weighted estimate of the intercept and slope for state 1 are

```
round(CW(1), 2)
```

```
           State 1
Intercept  1652.61
Slope        62.63
```

Table 3.4 assembles the credibility matrices and the standalone, credibility-weighted, and collective estimates.

The last column, labeled "Collective," repeats its entries for every state because we have only a single estimate for the entire portfolio of states. For the "Standalone," "Credibility," and "Collective" columns, we have listed the estimate of the intercept first and of the slope second.

Table 3.4. Hachemeister's credibility matrices and regression estimates. Each pair of rows corresponds to a state. Columns two and three provide the 2 × 2 credibility matrix. The remaining columns give us the intercept (first row) and the slope (second row) for each state.

Credibility Matrix			Intercept and Slope Estimates		
State	Col. 1	Col. 2	Stand-Alone	Credibility	Collective
1	0.8946	0.3389	1,658.47	1,652.61	1,495.75
1	0.0125	0.9460	62.39	62.63	29.09
2	0.6583	1.0286	1,398.30	1,419.30	1,495.75
2	0.0380	0.8222	17.14	15.57	29.09
3	0.6029	1.1851	1,533.00	1,535.05	1,495.75
3	0.0437	0.7740	43.31	41.73	29.09
4	0.3930	1.4753	1,176.70	1,368.48	1,495.75
4	0.0545	0.6122	27.81	10.93	29.09
5	0.7658	0.7245	1,521.90	1,503.30	1,495.75
5	0.0267	0.8812	11.87	14.62	29.09

Carefully inspecting the credibility estimates for each state reveals some peculiarities that Hachemeister (1975, 153) noted in his work as follows:

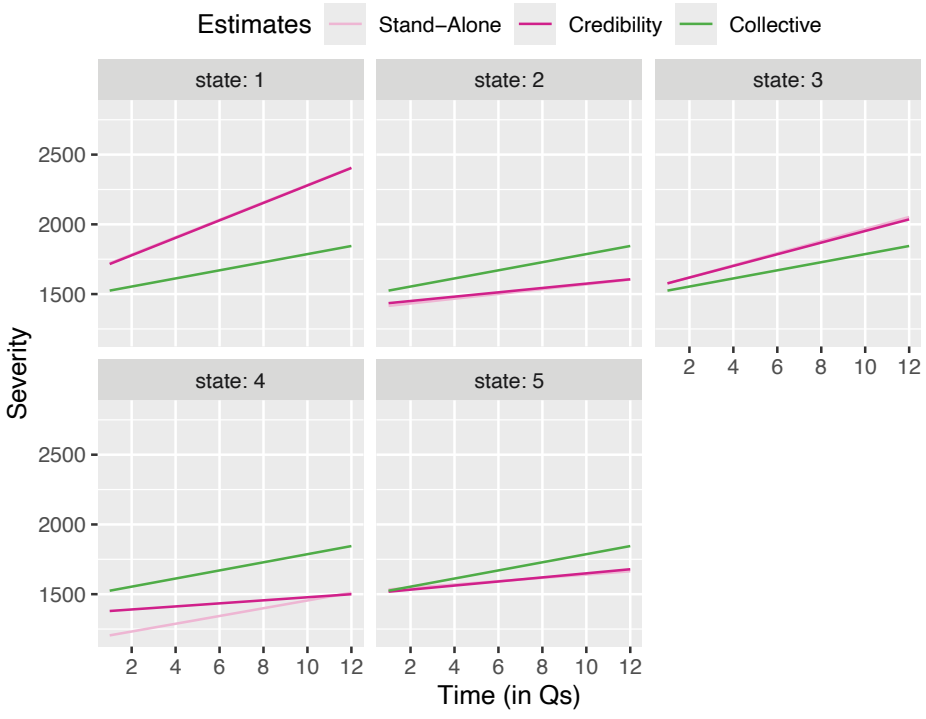
> State #4 trend lines clearly point out a distressing aspect of the credibility adjusted trend line. The credibility adjusted trend line has a lower trend than both the country wide and the state trend line.

This is clearly seen in Figure 3.6, where we have one panel for each state and have included the collective estimate of the trend line (dark green line), the stand-alone state estimate (light purple line), and the credibility-weighted trend estimate (dark purple line).

Exercise 3.4 By carefully inspecting the table of estimates, for both intercept and slope, determine which of the credibility estimates are not between the collective and stand-alone figures.

Solution 3.4 For state 1, both the credibility slope and intercept are outside the intervals defined by the stand-alone and collective estimates, and for state 2, only the slope is not between the stand-alone and collective estimates. State 3 has its intercept outside the collective and the stand-alone figures. The slope for state 4, as we have seen, is outside. Only state 5 has both its intercept and slope within the intervals defined by the stand-alone and collective estimates.

Figure 3.6. Credibility estimates for Hachemeister data. The green line corresponds to the collective estimate. The light purple line is the stand-alone estimate for the state, and the dark purple line is the credibility-weighted estimate. For all the states, except state 4, the stand-alone (light purple) and credibility lines (dark purple) are nearly one on top of the other.



Several authors (De Vylder 1981, 1985; Bühlmann and Gisler 1997) have noted the strange credibility estimates Hachemeister arrived at, and many actuaries would not apply these methods in practice. Some authors (De Vylder 1981, 1985) tried to impose constraints to resolve the issues, and others (Danneburg 1996) have pointed out that those constraints have drawbacks. In 1997, Bühlmann and Gisler (1997) found a simple solution. Recall that the credibility matrix A_j in Equation 3.27 is equal to

$$A_j = \frac{w_{j\bullet}}{D}\begin{bmatrix} w_{j\bullet}\mathrm{Var}_j^{(s)}\left[t\right]+\kappa_1 & \kappa_1 E_j^{(s)}\left[t\right] \\ \kappa_0 E_j^{(s)}\left[t\right] & w_{j\bullet}\mathrm{Var}_j^{(s)}\left[t\right]+\kappa_0 E_j^{(s)}\left[t^2\right] \end{bmatrix},$$

and looking at the off-diagonal elements, namely, $\kappa_1 E_j^{(s)}[t]$ and $\kappa_0 E_j^{(s)}[t]$, we might want to make them equal to zero. If our credibility matrix A_j is diagonal, then the credibility-weighted estimates of the intercept and slope would be split into two individual credibility calculations: one for the intercept and one for the slope. Currently with the credibility matrix we have, the estimate for the intercept involves combining both the intercepts and slopes of the stand-alone and collective estimates. Similarly, the credibility estimate of the slope is a combination of both the intercept and slope estimates of the state and the collective.

So how could those off-diagonal elements be zero? That is, how could we make $E_j^{(s)}[t]$ be zero? Remembering that

$$E_j^{(s)}\left[t\right] = \sum_{t=1}^{T} \frac{w_{jt}}{w_{j\bullet}} t,$$

we could shift our time variable t so that the above expression is equal to zero. In other words, we would like to replace t with $t - t_0$ such that $E_j^{(s)}[t - t_0]$. Namely, we would let

$$t_0 = \sum_{t=1}^{T} \frac{w_{jt}}{w_{j\bullet}} t.$$

Note that t_0 is the weighted average of the time variable for state j. We could also call t_0 the "center of gravity" for state j. With this translation of the time axis we are putting the intercept of our model at time $t = t_0$ instead of at the traditional origin of time $t = 0$.

In this case, the new credibility matrix A_j' becomes

$$A_j' = \frac{w_{j\bullet}}{D'}\begin{bmatrix} w_{j\bullet}\mathrm{Var}_j^{(s)}\left[t-t_0\right]+\kappa_1 & \kappa_1 E_j^{(s)}\left[t-t_0\right] \\ \kappa_0 E_j^{(s)}\left[t-t_0\right] & w_{j\bullet}\mathrm{Var}_j^{(s)}\left[t-t_0\right]+\kappa_0 E_j^{(s)}\left[\left(t-t_0\right)^2\right] \end{bmatrix}$$

$$= \frac{w_{j\bullet}}{D'}\begin{bmatrix} w_{j\bullet}\mathrm{Var}_j^{(s)}\left[t\right]+\kappa_1 & 0 \\ 0 & \left(w_{j\bullet}+\kappa_0\right)\mathrm{Var}_j^{(s)}\left[t\right] \end{bmatrix},$$

where

$$D' = \left(w_{j\bullet} + \kappa_0\right)\left(w_{j\bullet}\mathrm{Var}_j^{(s)}\left[t\right] + \kappa_1\right),$$

and noting that variances are not affected by a linear translation and $E_j^{(s)}[(t - t_0)^2] = \mathrm{Var}_j^{(s)}[t]$. We can simplify to obtain the following credibility matrix:

$$A_j' = \begin{bmatrix} \dfrac{w_{j\bullet}}{w_{j\bullet} + \sigma^2/\tau_0^2} & 0 \\ 0 & \dfrac{w_{j\bullet}\mathrm{Var}_j^{(s)}\left[t\right]}{w_{j\bullet}\mathrm{Var}_j^{(s)}\left[t\right] + \sigma^2/\tau_1^2} \end{bmatrix}.$$

The diagonal entries are of the form of the Bühlmann–Straub credibility factors. Hence, the credibility-weighted intercept and slope will be strictly between the stand-alone and collective estimates, respectively.

We were able to transform the original credibility matrix A_j into a diagonal credibility matrix A_j' by translating the origin of time to the center of gravity. We did all this for a particular state j, and there is no guarantee that the centers of gravity for the states will all coincide with one another. For the Hachemeister data, the individual centers of gravity are

```
CG <- tapply(W.jt * T.jt, S, sum) / W.jb
round(CG, 3)
```

```
    1     2     3     4     5
6.450 6.588 6.300 6.339 6.563
```

and notice that they are all close to each other. The largest difference between any two states is 0.288. The global center of gravity is

```
j0 <- sum(W.jb * CG) / W.bb
round(j0, 3)
```

```
[1] 6.475
```

Bühlmann and Gisler (2005, 214) have noted that in practice the centers of gravity are usually close to each other and that by translating the time variable to the overall center of gravity, the credibility matrices would be nearly diagonal. Table 3.5 shows the credibility estimates when we translate the origin of time to the global center of gravity of 6.475. Note that the credibility matrices are nearly diagonal and the estimated slope for state 4 is now between the stand-alone and collective values. Also, all other

Table 3.5. Credibility matrices and estimated stand-alone, credibility, and collective intercept and slope for the Hachemeister data when the time variable has been centered at the global center of gravity. Note that the off-diagonal elements of the credibility matrices are nearly zero. For the stand-alone, credibility, and collective estimates, the intercept is listed first and the slope second.

Credibility Matrix			Intercept and Slope Estimates		
State	Col.1	Col.2	Stand-Alone	Credibility	Collective
1	0.9731	–0.0014	2,062.46	2,052.54	1,694.98
1	–0.0001	0.9413	62.39	60.69	33.72
2	0.8779	0.0236	1,509.28	1,531.57	1,694.98
2	0.0009	0.7629	17.14	20.91	33.72
3	0.8321	–0.0454	1,813.41	1,793.09	1,694.98
3	–0.0017	0.6881	43.31	40.12	33.72
4	0.6000	–0.0483	1,356.75	1,492.32	1,694.98
4	–0.0018	0.4079	27.81	31.91	33.72
5	0.9288	0.0118	1,598.79	1,605.38	1,694.98
5	0.0004	0.8559	11.87	14.98	33.72

credibility estimates are between the stand-alone and collective values. The credibility calculations we have just completed have been encapsulated in the function `HBG()` (see Appendix 10).

```
sg <- sig.sq(X.jt, T.jt - j0, W.jt, db$state)$sigma.sq
D <- tau(sg, X.jt, T.jt, W.jt, db$state)$D
CW.one.center <- HBG(sg, D, X.jt, T.jt - j0, W.jt,
                     db$state, use.B.gls = TRUE)
```

Just as we translated the origin of time to the global center of gravity, we could do the time translation on a state-by-state basis. That would yield diagonal credibility matrices for each state. In Table 3.6 we have done just that, and comparing all the estimates to those in Table 3.5 we can see that, in this example, the differences are quite small.

```
j0 <- rep(CG, each = 12)
sg <- sig.sq(X.jt, T.jt - j0, W.jt, db$state)$sigma.sq
D <- tau(sg, X.jt, T.jt, W.jt, db$state)$D
CW.many.centers <- HBG(sg, D, X.jt, T.jt - j0, W.jt,
                       db$state, use.B.gls = TRUE)
```

Table 3.6. Credibility matrices and estimated stand-alone, credibility, and collective intercept and slope for the Hachemeister data when the time variable for each state has been centered at its own center of gravity. Note that the off-diagonal elements of the credibility matrices are exactly zero. For the stand-alone, credibility, and collective estimates, the intercept is listed first and the slope second.

Credibility Matrix			Intercept and Slope Estimates		
State	Col.1	Col.2	Stand-Alone	Credibility	Collective
1	0.9731	0.0000	2,060.92	2,051.04	1,693.42
1	0.0000	0.9413	62.39	60.71	33.67
2	0.8779	0.0000	1,511.22	1,533.46	1,693.42
2	0.0000	0.7628	17.14	21.06	33.67
3	0.8324	0.0000	1,805.84	1,787.00	1,693.42
3	0.0000	0.6880	43.31	40.30	33.67
4	0.6002	0.0000	1,352.98	1,489.09	1,693.42
4	0.0000	0.4077	27.81	31.28	33.67
5	0.9288	0.0000	1,599.83	1,606.49	1,693.42
5	0.0000	0.8559	11.87	15.02	33.67

Figure 3.7 shows the credibility regression lines when we translate the time variable to the center of gravity (shown as a vertical gray line) for each state. Note how each intercept, at the center of gravity, is strictly between the state stand-alone estimate and the collective estimate. Also, the credibility-adjusted slopes are between the stand-alone and the collective estimates. In particular, state 4 now has a very plausible regression line. Compare its panel here (Figure 3.7) with its panel in Figure 3.6.

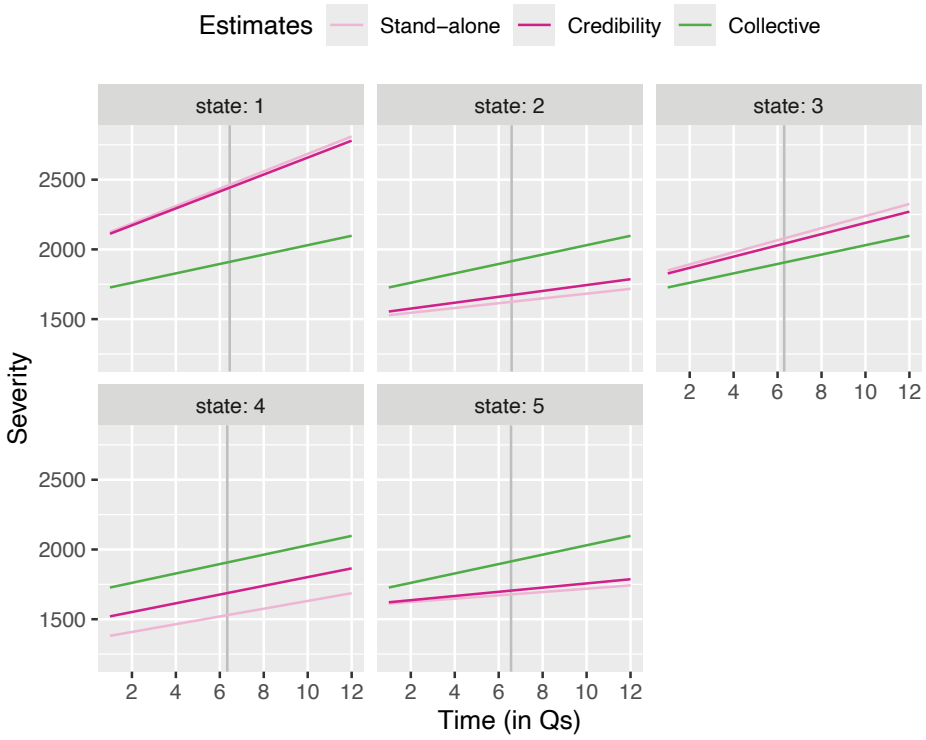
3.5. Summary

In this chapter we discussed the idea that the data we have collected on some risks is more credible than the data on other risks. For those risks whose data is credible, we can use it with confidence to predict next year. But for those risks whose data is not fully credible, we can combine their own data with the data for all risks to come up with a better prediction for next year. Credibility theory tells us how we should put together the collective's data and the own risk data in an optimal way by weighting the two sources together.

We devoted the second section to the balanced Bühlmann model and established that next year's premium should be a weighted average of a risk's own experience and the experience of the collective of all risks. Namely, the credibility premium has the form

$$Z\bar{X}_j + \left(1 - Z\right)\bar{X} \quad \text{with} \quad Z = \frac{T}{T + \sigma^2/\tau^2},$$

Figure 3.7. Credibility estimates for Hachemeister's data with time translation to the center of gravity for each state. Each state center of gravity is shown as a gray vertical line. Note that all the credibility-weighted intercepts (at the center of gravity) and slopes (dark purple) are now strictly between the stand-alone estimates (light purple) and the collective estimates (dark green).



where σ^2 and τ^2 are known, in the actuarial world, as the expected value of the process variance (EVPV) and as the variance of the hypothetical means (VHM), respectively. In the statistical literature, these are known as the *within variance* and the *between variance*, respectively. Note that as the EVPV (within variance), σ^2, increases, the credibility factor Z decreases. Similarly, as the VHM (between variance), τ^2, decreases toward zero, the credibility factor, Z, decreases.

The balanced Bühlmann model is critical to our understanding of credibility procedures, but it is not a very useful model in practice as it assumes that all risks have the same exposure and are observed over the same number of periods.

The third section focused on extending the balanced Bühlmann model to a practically useful model known as the Bühlmann–Straub model. In this model, each risk comes with its own exposure weight, and not all risks need to be observed over the same time period. With these extensions, the credibility premium is of the same form as in the balanced Bühlmann model, namely,

$$Z_j X_{j\bullet} + \left(1 - Z_j\right) X_z \quad \text{with} \quad Z_j = \frac{w_{j\bullet}}{w_{j\bullet} + \sigma^2/\tau^2}.$$

Again, we see that the credibility premium has the same weighted average form as in the balanced Bühlmann model, as do the credibility factors.

In the last section, we explored a credibility regression model first proposed by Hachemeister. Here the basic idea is that we have data on several risks and we would like to estimate a regression line on this data. We could ignore that data came from individual risk classes and fit one regression line to all of the data. But that approach discards important information. We could also fit individual regression lines on each risk class. For some risk classes the volume of information would be large enough to give us a "robust" regression line, but for some of them the volume would be small and we might get some spurious results.

In this situation credibility theory can be applied to estimate the regression coefficients as weighted averages of the individual regression coefficients and the collective regression coefficients. Unfortunately, a naive application of credibility to Hachemeister's data led to implausible results for state 4, where the credibility-weighted trend for that state was both lower than its stand-alone and collective estimates.

This implausible result arises from the fact that in this case the credibility factors are 2×2 matrices with nonzero entries in all four positions, and thus the credibility-weighted intercept and slope are a complex combination of *both* stand-alone and collective intercepts and slopes.

Obtaining plausible estimates required two key insights (Bühlmann and Gisler 1997):

- assume that the variance-covariance matrix of the intercept and slope random coefficients is diagonal, and
- center the time variable at each risk class time center of gravity.

With those insights, the credibility factors are 2×2 diagonal matrices, and so the credibility-weighted intercept and slope are each calculated separately, yielding estimates that are always between the individual risk and the collective values.

4. Linear Mixed Models

4.1. Introduction

In Chapter 3 we looked at the balanced Bühlmann, the Bühlmann–Straub, and Hachemeister's credibility regression models. Those and similar models have also been studied by statisticians under various names—linear mixed models (LMMs), hierarchical models, longitudinal models, and panel models, to name a few. The statistical literature on such models is extensive, and the models are well developed. Actuaries can benefit significantly by using the underlying theory and the tools available to assess such models.

We will introduce LMMs by reframing the credibility models in the standard statistical notation and exploring the tools that have been developed to fit and assess them. This should illustrate how we can apply the LMM theory to practical problems in credibility.

4.2. Balanced Bühlmann Model Revisited

One way to write the balanced Bühlmann model (see Equation 3.2) is

$$X_{jt} = \mu + \Xi_j + \epsilon_{jt}, \quad j = 1, 2, \ldots, J, \quad t = 1, 2, \ldots, T,$$

where X_{jt} is the observation for group j at time t, μ is the overall mean, Ξ_j is a random deviation from the overall mean for group j, and ϵ_{jt} is an error term for the j, t observation. This conforms to the actuarial notation but not to the statistician's. So we will switch the notation to what is commonly used in statistics.

The response variable is typically named Y, and the explanatory variables are usually denoted by X. We can express the balanced Bühlmann model as

$$y_{jt} = \beta + b_j + \epsilon_{jt},$$

where β is the overall mean, b_j is a random variable representing the deviation from the overall mean for the jth group, and ϵ_{jt} is the deviation for the j,t observation. Both b_j and ϵ_{jt} are *deviations*, and therefore we know their means are zero.

Since b_j and ϵ_{jt} are random variables, we need to specify their distributions and how they might be related to each other. For this model, we will have them both be independent, with constant variance, and normally distributed. The variance of b_j will

be denoted by σ_b^2 and denotes the *between*-group variability. Actuaries call this the *variance of the hypothetical means*, or VHM. The hypothetical means are $\beta + b_j$. The variance of ϵ_{jt} is known as *within*-group variability and is denoted by σ^2. In actuarial groups this is known as the *expected value of the process variance*, or EVPV. We can write these specifications as

$$b_j \sim \mathcal{N}\left(0, \sigma_b^2\right) \quad \text{and} \quad \epsilon_{jt} \sim \mathcal{N}\left(0, \sigma^2\right),$$

where $\mathcal{N}(0, \sigma^2)$ represents the normal distribution with mean equal to zero and variance equal to σ^2.

For the balanced Bühlmann model, we did not make the assumption that our random variables were normally distributed. We assumed only that they had finite first and second moments, and we used the method of moments to derive estimates for variances. It turns out that the maximum likelihood estimates of these variances coincide with the method of moments for this simple model. Hence, the balanced Bühlmann model is equivalent to this LMM.

Statisticians would call β a fixed effect and b_i a random effect, and because this model has both fixed and random effects it is called a *mixed-effects* model. The naming of coefficients as either fixed or random is not without controversy. Gelman and Hill (2007, Section 11.4) outline five definitions for these terms, and in their work they avoid using the terms.

Using the same data as in Table 3.1, namely (showing the first few rows of the data frame),

```
# A tibble: 6 x 3
   class   time  value
   <fct>  <int>  <dbl>
1  1          1    625
2  1          2    675
3  1          3    600
4  1          4    700
5  2          1    750
6  2          2    800
```

we will illustrate the fitting of an LMM and show that we arrive at the same estimates. But rather than jumping straight into that mixed model, we want to describe one process of fitting and exploring models to reach that mixed model.

We can start simply by fitting an OLS model that includes only an intercept term. Such a model is usually called a *null* model. It is the most basic model we can have.

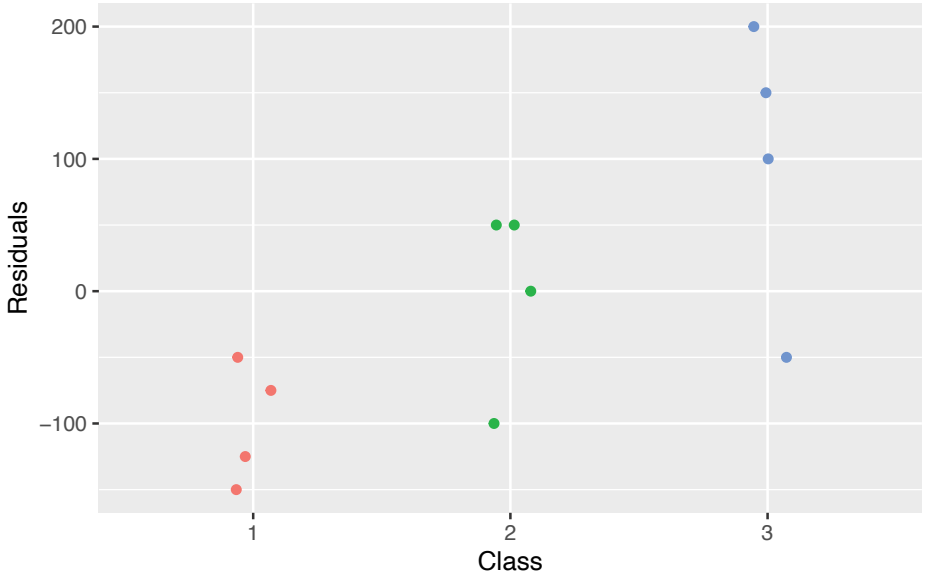
```
BB.null.lm <- lm(value ~ 1,
                 data = dta)
(sBB.null.lm <- summary(BB.null.lm))
```

```
Call:
lm(formula = value ~ 1, data = dta)

Residuals:
    Min      1Q  Median      3Q     Max
-150.00  -81.25  -25.00   62.50  200.00

Coefficients:
            Estimate  Std. Error  t value  Pr(>|t|)
(Intercept)    750.00       32.13    23.34  1.01e-10  ***
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 111.3 on 11 degrees of freedom
```

From the output, we can see that the overall mean is 750 and the residual standard error is 111.29. Thus, we know that $\hat{\beta}_0 = 750$ and $\hat{\sigma}^2 = 12{,}386.36$.

The residuals from this model (see Figure 4.1) show some unsettling patterns. All residuals for class #1 are negative and clustered around –100. Similarly, nearly all the residuals for class #3 are positive and also seem to be clustered around 150. Clearly, an intercept-only model does not fit the data well, and we have an effect from the `class` variable that needs to be incorporated into the model. We can add `class` as a categorical variable to estimate the model.

Figure 4.1. OLS residuals for the balanced Bühlmann example data. Note that the residuals for class #1 all have the same sign, and nearly all points for class #3 also have the same sign. We introduced a slight amount of horizontal jittering to avoid overplotting a pair of residuals.



```
BB.class.lm <- lm(value ~ class - 1,
                  data = dta)
(sBB.class.lm <- summary(BB.class.lm))
```

```
Call:
lm(formula = value ~ class - 1, data = dta)

Residuals:
    Min      1Q  Median      3Q     Max
-150.00  -31.25   12.50   50.00  100.00

Coefficients:
       Estimate  Std. Error  t value  Pr(>|t|)
class1    650.00       39.53    16.44  5.07e-08  ***
class2    750.00       39.53    18.97  1.44e-08  ***
class3    850.00       39.53    21.50  4.79e-09  ***
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 79.06 on 9 degrees of freedom
Multiple R-squared: 0.9918, Adjusted R-squared: 0.9891
F-statistic: 364.3 on 3 and 9 DF, p-value: 1.037e-09
```

In the above model we removed the intercept so that we would estimate a mean value for each class (instead of using one of the classes as an intercept and then estimating deviations from this mean for the other two classes). Note that the residual standard error is now much smaller: 79.06 versus 111.29. This model fits our data more closely.

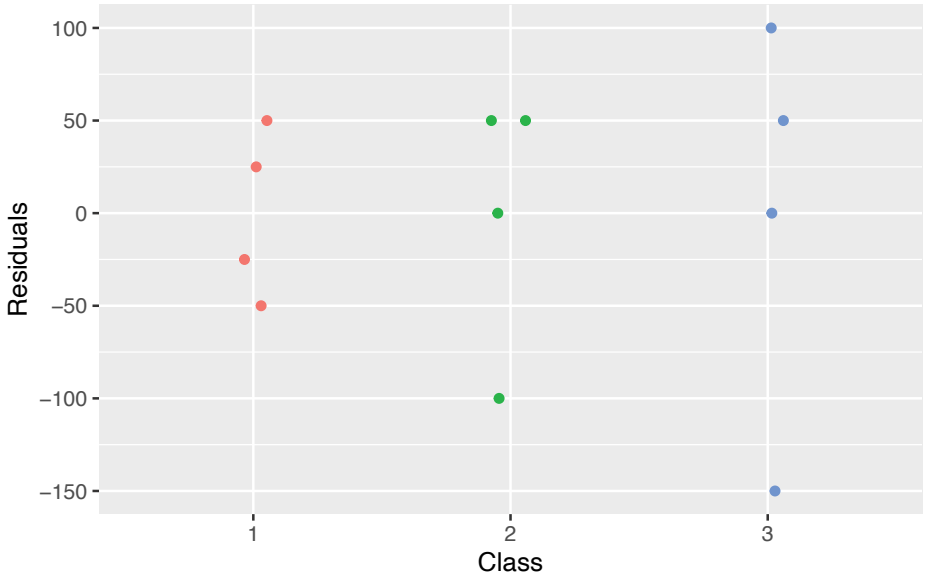
Figure 4.2 shows residuals that are much better behaved. They are all centered around zero with both positive and negative values for each class. We might be quite happy with this model if we were interested in just these three classes. But consider that these three might have been just a sample from hundreds of classes (say, workers compensation occupational classes). If we were to include all possible classes in a model, we may not be able to estimate all of the parameters accurately. Some classes may have lots of data, but others may have very little. More troublesome is the fact that as we add classes, the number of parameters that need to be estimated increases.

So we want to think of the three classes we have as being a sample from a population of classes. Thus, we want to estimate the following mixed model:

$$y_{jt} = \beta + b_j + \epsilon_{jt},$$

where $j = 1, 2, 3$ and $t = 1, 2, 3, 4$. This model has a fixed effect β, which is constant across classes, and a random deviation from the overall mean b_j for each class. We can fit this model with the `lmer()` function from the `lme4` package as follows:

Figure 4.2. OLS residuals for the balanced Bühlmann example data with a mean estimate for each class. Note that now all the residuals are centered around zero. We introduced a slight amount of horizontal jittering to avoid overplotting a pair of residuals.




```
BB.mx <- lmer(value ~ 1 + (1 | class),
              data = dta)
```

The response variable is `value`, and the first 1 after the "~" says we want a fixed-effect intercept. We can specify other fixed effects in this part of the formula as we would in fitting a regular regression model. The component in parentheses after the plus sign is for the random effects. Here we have `1 | class` because we want a random intercept for each level of the `class` variable.

The parameters to be estimated for this model are β, the fixed effect; σ_b^2, the between-class variance (VHM); and σ^2 the within-class variance (EVPV), also known as the residual variance.

```
(sBB.mx <- summary(BB.mx))
```

```
Linear mixed model fit by REML ['lmerMod']
Formula: value ~ 1 + (1 | class)
   Data: dta

REML criterion at convergence: 133.6

Scaled residuals:
    Min      1Q  Median      3Q     Max
-1.6997 -0.5929  0.1581  0.6325  1.4626
```

```
Random effects:
 Groups    Name         Variance  Std.Dev.
 class     (Intercept)  8438      91.86
 Residual               6250      79.06
Number of obs: 12, groups: class, 3

Fixed effects:
            Estimate  Std. Error  t value
(Intercept)   750.00       57.74    12.99
```

From the output we have $\hat{\beta} = 750$, $\hat{\sigma}_b^2 = 8{,}438$, and $\hat{\sigma}^2 = 6{,}250$. Thus, we can see that the variance between classes is bigger than the variance within a class. The random effects are

```
round(ranef(BB.mx)$class, 3)
```

```
  (Intercept)
1     -84.375
2       0.000
3      84.375
```

which tells us that our estimated mean is 750 – 84.375 = 665.625 for class #1, 750 + 0 = 750 for class #2, and 750 + 84.375 = 834.375 for class #3. These are the hypothetical means for each class. Note that they are exactly the same estimates as the credibility estimates we calculated in Chapter 3, Equation 3.12. The credibility factor can also be easily derived from the above output:

$$Z = \frac{T}{T + \hat{\sigma}^2/\hat{\sigma}_b^2} = \frac{4}{4 + 6{,}250/8{,}437.5} = 0.84375.$$

The value of T represents the number of observations in each group, and because we are in the *balanced* Bühlmann model, we know that each group has the same number of observations. The above output tells us that there were 12 observations and three groups; hence, $T = 12/3 = 4$.

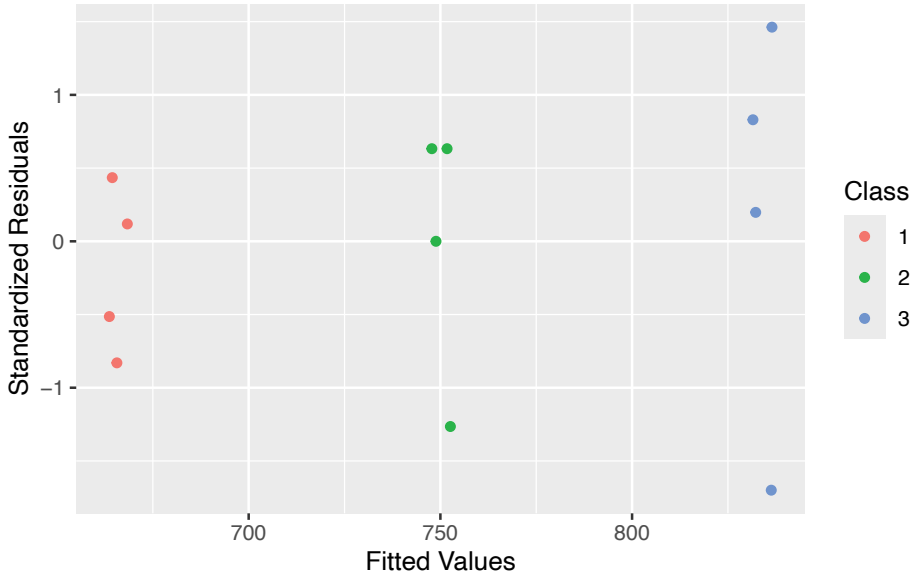
In the mixed model we have assumed that the residual variance σ^2 is constant. By plotting the fitted values against the standardized residuals we can check this assumption. Figure 4.3 shows these residuals, and even though we have a small sample size the residuals are well behaved.

4.3. Bühlmann–Straub Model Revisited

The Bühlmann–Straub model is nearly the same as the balanced Bühlmann model. There are two key differences:

- we do not assume that all risks have been observed for the same number of periods (we no longer have a balanced dataset), and

Figure 4.3. Standardized residuals for the balanced Bühlmann data fitted with the mixed model. Note that all residuals are within 1.5 standard deviations from zero. We added a bit of horizontal jittering to avoid overplotting a pair of residuals.



- we introduce weights so that the residual variance is proportional to the inverse of the weights.

  Therefore, the Bühlmann–Straub model can be written as

$$y_{jt} = \beta + b_j + \epsilon_{jt} \quad \text{with} \quad b_j \sim \mathcal{N}\left(0, \sigma_b^2\right) \quad \text{and} \quad \epsilon_{jt} \sim \mathcal{N}\left(0, \frac{\sigma^2}{w_{jt}}\right),$$

where w_{jt} are the weights associated with the observation from risk j and time t.

In Section 3.3, we illustrated the standard actuarial calculations for this model on a simulated dataset (see Listing 3.1) that had 100 different risk classes and five time periods of observations. The parameters used in the simulation were

$$\beta = 80, \quad \sigma_b^2 = 64, \quad \sigma^2 = 100.$$

Note that in the Bühlmann–Straub model we denoted the between-risk variance (variance of the hypothetical means) by the symbol τ^2, but here we use σ_b^2.

We will use the same simulated data to illustrate how to estimate these parameters via an LMM, and so we load the dataset we created in the previous chapter.

```
bs.dta <- read_csv("BS-simulated-data.csv",
                   col_types = "fdd")
```

We estimate the Bühlmann–Straub model using the linear mixed-effects regression function `lmer()` from the R package `lme4` as follows:

```
BS.mx <- lmer(X.jt ~ 1 + (1 | risk),
              data = bs.dta,
              weights = W.jt)
```

Note that the dataset uses the actuarial notation `X.jt` for the response variable and `risk` for the name of the class variable, where the formula in the first argument, X.jt ~ 1 + (1 | risk), says that we have a fixed-effects intercept, the first 1, and the expression inside parentheses denotes the random component of the model. In this case, the random component is just an intercept that varies by the classification variable `risk`. We can obtain summary information about the fit via the `summary()`, and we have saved the information in an object, `sBS.mx`, to be able to extract some of that information later on.

```
(sBS.mx <- summary(BS.mx))
```

```
Linear mixed model fit by REML ['lmerMod']
Formula: X.jt ~ 1 + (1 | risk)
   Data: bs.dta
Weights: W.jt

REML criterion at convergence: 3901.4

Scaled residuals:
     Min       1Q   Median       3Q      Max
-2.37735  -0.67636  -0.01581  0.64974  2.76651

Random effects:
 Groups   Name        Variance  Std.Dev.
 risk     (Intercept)    61.14     7.819
 Residual               104.64    10.229
Number of obs: 500, groups: risk, 100

Fixed effects:
            Estimate  Std. Error  t value
(Intercept)  78.5384      0.9083    86.46
```

Table 4.1 shows the true value of the model parameters, their estimated values from the mixed model, and the credibility estimates from Section 3.3. Note that the mixed model estimates and the credibility estimates are very close to each other, and both are not far away from true values we used to simulate the data.

Table 4.1. Comparison of actual and estimated model parameters for the Bühlmann–Straub simulated data.

Parameter	True Values	Mixed Model Estimates	Credibility Estimates
β	80	78.5384	78.4363
σ_b^2	64	61.1411	60.9652
σ^2	100	104.6386	104.5239

For the Bühlmann–Straub model, the credibility factors differ by risk group j, and from the above model output they would be equal to

$$\hat{Z}_j = \frac{w_{j\bullet}}{w_{j\bullet} + \hat{\sigma}^2/\hat{\sigma}_b^2} = \frac{w_{j\bullet}}{w_{j\bullet} + 104.64/61.14},$$

where $w_{j\bullet}$ is the sum of all the weights across time for risk j.

From the LMM `BS.mx`, we can also obtain values for the deviations b_j from the overall mean β. These values are $\hat{b}_1, \hat{b}_2, \ldots, \hat{b}_{100}$. The first 20 of them are

```
round(ranef(BS.mx)$risk[1:20,1], 3)
```

```
 [1] -11.126  -3.886   3.838   3.244  -3.004   5.729   1.625
 [8]  -6.815  16.494  -7.125   7.894   1.822  11.837  -5.196
[15]  -3.828   0.172  -6.792  14.001  -6.224  -9.247
```

These values together with the estimate of the fixed effect $\hat{\beta}$ yields the credibility-weighted estimate for each risk—that is, $\hat{\beta} + \hat{b}_j$ is our estimate for risk j. For the first 20 risks we have

```
round(fixef(BS.mx) + ranef(BS.mx)$risk[1:20, 1], 3)
```

```
 [1] 67.412 74.652 82.376 81.782 75.534 84.268 80.164
 [8] 71.724 95.032 71.414 86.433 80.360 90.375 73.343
[15] 74.710 78.710 71.746 92.539 72.314 69.291
```

We add fitted values to our dataset and compute the standardized residuals from our model.

```
bs.dta$mu.mx <- fitted(BS.mx)
bs.dta$sres.mx <- resid(BS.mx, type = "pearson", scaled = TRUE)
```

Figure 4.4. Diagnostic plot for the LMM with random intercepts fitted to the simulated Bühlmann–Straub data. All the standardized residuals are within 2.5 standard deviations from the origin. The overall impression is that of a random cloud of points.

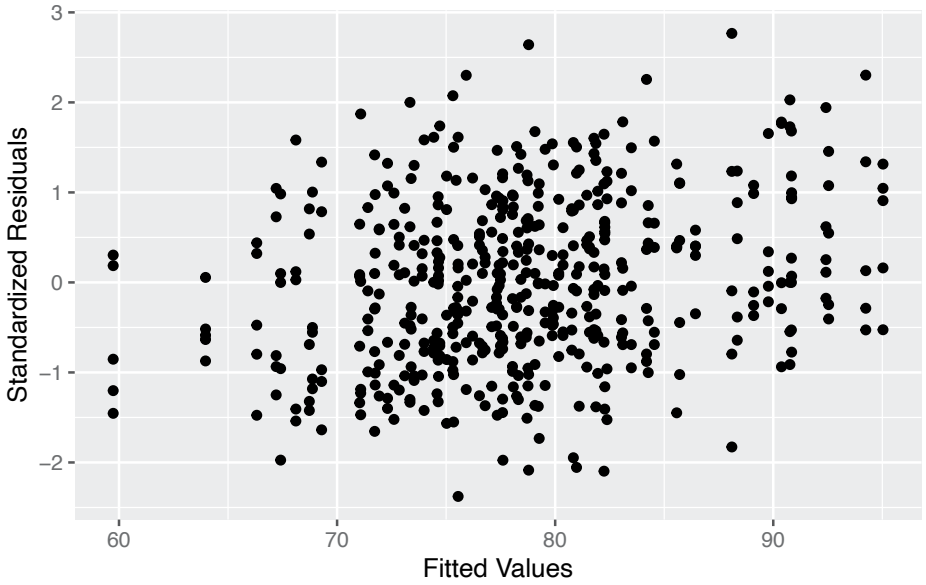


Figure 4.4 displays the fitted values versus the standardized residuals. The scatterplot of points appears like a random cloud of points centered about the line $y = 0$. But if you look closely you might be able to discern an upward-sloping pattern.

Exercise 4.1 Fit a linear regression line to the points shown in Figure 4.4 to show that there is an upward-sloping pattern in the residuals.

Solution 4.1 We fit a linear model to the points shown in Figure 4.4 as follows:

```
BS.lm <- lm(sres.mx ~ mu.mx,
            data = bs.dta)
(sBS.lm <- summary(BS.lm))
```

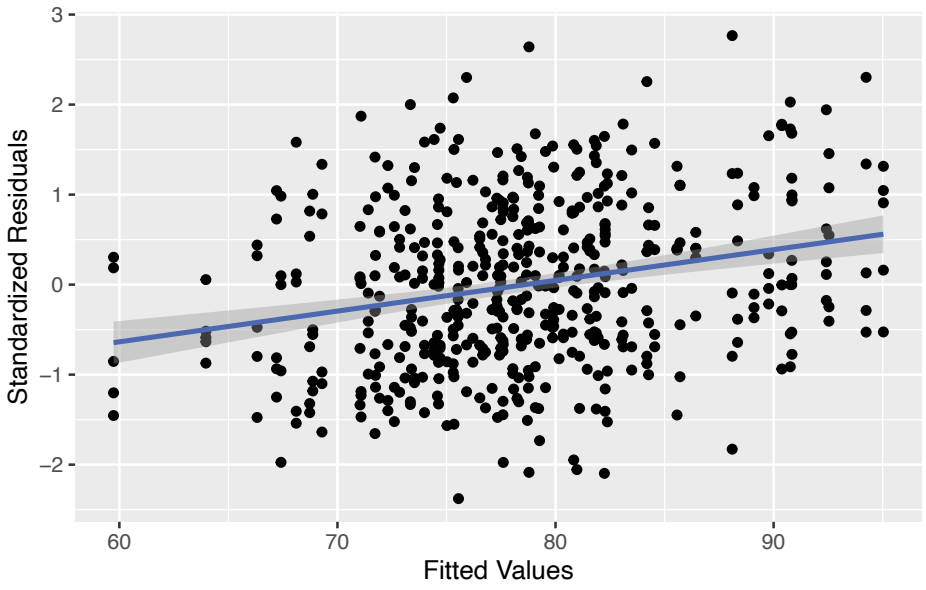
```
Call:
lm(formula = sres.mx ~ mu.mx, data = bs.dta)

Residuals:
     Min       1Q   Median       3Q      Max
-2.27282  -0.66309  -0.05755  0.62719  2.63522

Coefficients:
             Estimate  Std. Error  t value  Pr(>|t|)
(Intercept) -2.683488    0.471425   -5.692  2.14e-08  ***
mu.mx        0.034135    0.005981    5.707  1.97e-08  ***
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Residual standard error: 0.8957 on 498 degrees of freedom
Multiple R-squared: 0.0614, Adjusted R-squared: 0.05951
F-statistic: 32.58 on 1 and 498 DF, p-value: 1.971e-08
```

The value of the coefficient of `mu.mx` is 0.0341, and from the summary information we can see that it is significant.

We can also show the diagnostic plot with the linear regression line.

```
ggplot(data = bs.dta,
       mapping = aes(x = mu.mx,
                     y = sres.mx)) +
  geom_point() +
  geom_smooth(method = "lm") +
  labs(x = "Fitted Values",
       y = "Standardized Residuals")
```

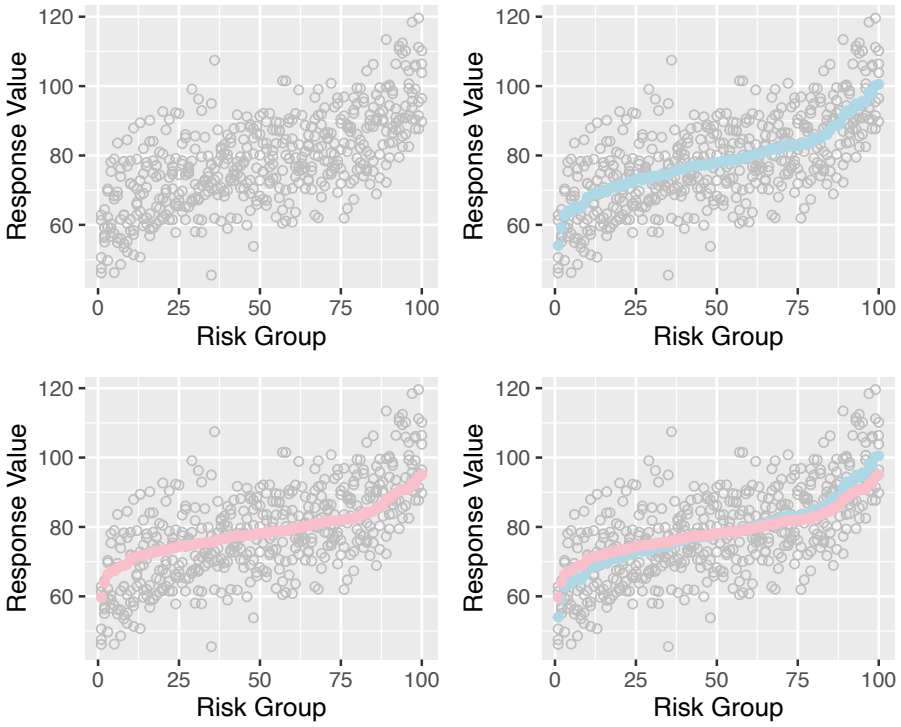


A pattern in the residuals would normally suggest that our model does not fit the data well. In an OLS situation, that would be correct and we would conclude that our model does not capture the underlying pattern in the data. But our situation is more complex than OLS, and the upward-sloping pattern we are seeing in Figure 4.4 is what we should expect.

The pattern we see is the result of shrinking our estimates toward the overall mean. To illustrate the effect, Figure 4.5 shows four panels. The top-left panel displays the response variable on the *y*-axis and the risk group on the *x*-axis. The risk groups have been ordered from the smallest fitted value based on the LMM to the largest. The fitted values from the mixed model are the credibility-weighted values given by

$$\hat{y}_{jt} = Z_j \bar{y}_j + \left(1 - Z_j\right)\bar{y},$$

Figure 4.5. The top-left panel shows the simulated Bühlmann–Straub data where the risk groups have been ordered from the smallest fitted value to the largest. The fitted values come from the LMM and coincide with the credibility-weighted values. The top-right panel shows the average response values (light blue circles) and the bottom-left panel shows the fitted values (pink circles) for each risk group. The bottom-right panel combines all three panels. Note that the fitted values (pink circles) have been shrunk toward the overall mean value of approximately 80.



where $\bar{y}_j$ is the average response value for group j, $\bar{y}$ is the collective average, and Z_j is the credibility factor given by

$$Z_j = \frac{w_{j\bullet}}{w_{j\bullet} + \hat{\sigma}^2 / \hat{\sigma}_b^2}.$$

For this simulated data, each risk group has five observations, and from the plot you can see that the *average for each risk group* ranges from below 60 to a bit more than 100. The top-right panel includes this average $\bar{y}_j$ for each risk group (light blue circles).

The bottom-left panel shows the fitted values from the mixed model (pink circles). These values increase steadily from left to right. Finally, in the bottom-right panel we have superimposed all three panels, and we can clearly see that on both ends of the

graph the fitted values are closer to the collective mean. These fitted values have been shrunk from the individual group average $\bar{y}_j$ to the collective average $\bar{y}$.

When we compute the response residuals from this model, we are taking the difference between actual values (shown as light gray circles) and the risk group's fitted values (shown in pink). These differences are not centered at the mean value for each risk group (shown in light blue), and as we approach both extremes the discrepancy increases. On the left-hand side the differences are more negative, and on the right-hand side they are more positive. Therefore, a residuals-versus-fitted values plot shows a positive trend line.

Exercise 4.2 There are three quantities that affect the amount of shrinkage that will occur—the weights w_{jt}, the between-risk variance σ_b^2 (VHM), and the within-risk variance σ^2 (EVPV).

Take each one in turn, and using the credibility factors Z_j, determine the effect on shrinkage that increasing or decreasing each quantity will have.

Solution 4.2 The amount of shrinkage is controlled by the size of the credibility factor, and for the Bühlmann–Straub model we know that it is given by

$$Z_j = \frac{w_{j\bullet}}{w_{j\bullet} + \sigma^2/\sigma_b^2}.$$

If the value of Z_j is close to 1, there will be very little shrinkage and the fitted values (credibility estimates) will be close to the average of the group $\bar{y}_j$.

If we increase the weights w_{jt}, then Z_j will approach 1. If the within-group variance σ^2 decreases toward zero, then the credibility factor Z_j will approach 1. Also, if the between-group variance σ_b^2 increases toward infinity, then again Z_j will approach 1 and the amount of shrinkage will decrease.

In Appendix A, we write a function to simulate datasets that conform to the Bühlmann–Straub model. We can use that function to explore how different values for the number of risks per group and within-/between-group variances affect the amount of shrinkage.

4.4. Some Linear Mixed-Model Theory

In the previous section, we estimated the LMMs corresponding to the balanced Bühlmann and the Bühlmann–Straub models. We could go straight into Hachemeister's regression model to illustrate that, too, but it would be better to understand some of the key constructions needed for these models to appreciate more complex situations. Therefore, let's start by laying down some standard notation and constructions for the LMM.

For a single level of grouping, we follow the discussion in Frees et al. (1999) closely, and we can write down the LMM as

$$y_j = X_j\beta + Z_j b_j + \epsilon_j, \quad j = 1, 2, \ldots, J$$
$$b_j \sim \mathcal{N}\left(0, D\right), \qquad \epsilon_j \sim \mathcal{N}\left(0, R_j\right),$$

where index j denotes the grouping factor, J is the total number of groups, y_j is a vector of response values for group j with dimension $n_j \times 1$, X_j is the fixed-effects design matrix with dimensions $n_j \times p$, and Z_j is the random-effects design matrix with dimensions $n_j \times q$. The fixed-effects linear predictor consists of p explanatory variables, and so we have $\beta = (\beta_1, \beta_2, \ldots, \beta_p)$ as fixed-effects coefficients. We also have q random-effects explanatory variables.

We assume that the responses between groups are independent, but we allow for serial correlation and weighting by assuming that the variance-covariance matrix for the error terms ϵ_j is an $n_j \times n_j$ matrix, which we write as R_j. We also assume that the expected value of the error terms is zero, that is, $\mathbb{E}[\epsilon_j] = 0$. Moreover, we assume that the group-specific effects b_j are independent and identically distributed with $\mathbb{E}[b_j] = 0$ and variance-covariance matrix D with dimensions $q \times q$. Note that the variance-covariance matrix D does not depend on the group. And we assume that the group-specific effects and the error terms are independent—that is, we have that their covariance $\text{Cov}(b_{ju}, \epsilon_{kv})$ is zero for all combinations of j, u, k, and v. Hence, the variance-covariance matrix for response vector y_j is

$$\begin{aligned}
\text{Var}\left(y_j\right) &= \text{Var}\left(X_j\beta + Z_j b_j + \epsilon_j\right) \\
&= \text{Var}\left(Z_j b_j + \epsilon_j\right) \\
&= \text{Var}\left(Z_j b_j\right) + \text{Var}\left(\epsilon_j\right) + 2\text{Cov}\left(Z_j b_j, \epsilon_j\right) \\
&= Z_j D Z_j^t + R_j \\
&= V_j,
\end{aligned}$$

where a superscript "t" denotes the transpose operation. So we have that the variance-covariance matrix V_j has dimension $n_j \times n_j$ and assume that this matrix is invertible. We also know that this matrix is symmetric, and if we let $N = \max(n_1, n_2, \ldots, n_J)$ be the maximum number of observations we have across all groups, then the matrix V_j has at most $N(N+1)/2$ unknown values. So let τ be the vector of unknown values, and we can denote the dependence of V_j on this vector via $V_j(\tau)$.

For the balanced Bühlmann example we discussed in Section 3.2, we have $J = 3$ groups observed over $N = 4$ periods, and each group had the same number of observations, that is, $n_j = 4$ for all $j = 1, 2, 3$. The vectors of observations were

$$y_1^t = \left(625, 675, 600, 700\right)$$
$$y_2^t = \left(750, 800, 650, 800\right)$$
$$y_3^t = \left(900, 700, 850, 950\right).$$

The design matrices X_j and Z_j are all of dimension 4×1 and have only an intercept as an explanatory variable, namely $p = q = 1$—thus

$$X_j = Z_j = \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix}.$$

The variance-covariance matrix of the group effects D is of dimension 1×1, and we labeled it as σ_b^2 in this chapter and as τ^2 in Section 3.2. We also assumed that error terms ϵ_j were independent of each other, and so we have

$$R_j = \sigma^2 \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}.$$

And the variance-covariance matrix of the responses for group j, $\text{Var}(y_j) = V_j$, is equal to

$$V_j(\tau) = Z_j D Z_j^t + R_j$$
$$= \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix} D \begin{bmatrix} 1 & 1 & 1 & 1 \end{bmatrix} + \sigma^2 \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$
$$= \begin{bmatrix} \sigma_b^2 + \sigma^2 & \sigma_b^2 & \sigma_b^2 & \sigma_b^2 \\ \sigma_b^2 & \sigma_b^2 + \sigma^2 & \sigma_b^2 & \sigma_b^2 \\ \sigma_b^2 & \sigma_b^2 & \sigma_b^2 + \sigma^2 & \sigma_b^2 \\ \sigma_b^2 & \sigma_b^2 & \sigma_b^2 & \sigma_b^2 + \sigma^2 \end{bmatrix};$$

therefore, the vector of variance components is $\tau = (\sigma_b^2, \sigma^2)$.

The entire model for all observations would be assembled by stacking the response vectors y_1, y_2, and y_3 into a single 12×1 column vector. The grand design matrices X and Z are of dimension 12×3, where the first column has four 1s and zeroes after; the second column has four zeroes, then four 1s, and then zeroes; and the final column starts with zeroes and ends with four 1s:

$$\begin{bmatrix} 625 \\ 675 \\ 600 \\ 700 \\ 750 \\ 800 \\ 650 \\ 800 \\ 900 \\ 700 \\ 850 \\ 950 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \end{bmatrix} + \begin{bmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} b_1 \\ b_2 \\ b_3 \end{bmatrix} + \begin{bmatrix} \epsilon_{11} \\ \epsilon_{12} \\ \epsilon_{13} \\ \epsilon_{14} \\ \epsilon_{21} \\ \epsilon_{22} \\ \epsilon_{23} \\ \epsilon_{24} \\ \epsilon_{31} \\ \epsilon_{32} \\ \epsilon_{33} \\ \epsilon_{34} \end{bmatrix},$$

with the variance of the response vector equal to a block diagonal matrix of dimension 12×12 where the first 4×4 block is V_1, the second 4×4 diagonal block is V_2, and the final 4×4 diagonal block is V_3. All other entries are zero.

$$V = \begin{bmatrix} V_1 & & \\ & V_2 & \\ & & V_3 \end{bmatrix}.$$

The GLS estimator of the fixed-effects β assumes that the variance components τ are known and is given by

$$\beta_{GLS} = \left(\sum_{j=1}^{J} X_j^t V_j^{-1} X_j \right)^{-1} \left(\sum_{j=1}^{J} X_j^t V_j^{-1} y_j \right). \tag{4.1}$$

In the balanced Bühlmann model we have $\beta_{GLS} = \bar{y}$, where $\bar{y}$ is the average of all the response values. To see this, first note that the variance-covariance matrix V_j has dimension $n_j \times n_j$ and is of the form (4×4 example)

$$V_j = \begin{bmatrix} a & b & b & b \\ b & a & b & b \\ b & b & a & b \\ b & b & b & a \end{bmatrix},$$

where $a = \sigma_b^2 + \sigma^2$ and $b = \sigma_b^2$. Because this matrix has a lot of structure, its inverse is relatively easy to figure it out by looking at small cases. In our example, we have

$$V_j^{-1} = \frac{1}{a(a+2b) - 3b^2} \begin{bmatrix} a+2b & -b & -b & -b \\ -b & a+2b & -b & -b \\ -b & -b & a+2b & -b \\ -b & -b & -b & a+2b \end{bmatrix},$$

which we can quickly verify by calculating a couple of entries for the matrix product $V_j^{-1} V_j$. For an $n \times n$ matrix, the diagonal in the inverse matrix has the form $a + (n-2)b$ and the off-diagonal elements are all $-b$. The multiplicative constant in front of the matrix is equal to $[a(a + (n-2)b) - (n-1)b^2]^{-1}$.

Exercise 4.3 Verify that the matrix product $V_j^{-1} V_j$ is the identity matrix by calculating the (1, 1) and (2, 1) entries of this matrix and noting that all other entries would be equal to one of these two calculations.

Solution 4.3 For the (1, 1) entry we need to take the dot product of the first row of V_j^{-1} and the first column of V_j. Ignoring the scalar multiplier in front of V_j^{-1}, we have

$$\begin{bmatrix} a+2b & -b & -b & -b \end{bmatrix} \begin{bmatrix} a \\ b \\ b \\ b \end{bmatrix} = a(a+2b) - 3b^2.$$

This is equal to the scalar multiplier in front of V_j^{-1}, and so the (1, 1) entry of the product $V_j^{-1} V_j$ is equal to 1.

For the (2, 1) entry we need to calculate the dot product of the second row of V_j^{-1} and the first column of V_j:

$$\begin{bmatrix} -b & a+2b & -b & -b \end{bmatrix} \begin{bmatrix} a \\ b \\ b \\ b \end{bmatrix} = -ab + ab + 2b^2 - 2b^2 = 0.$$

Hence, the (2, 1) entry is zero.

Since the design matrix X_j is just a column of 1s in the balanced Bühlmann model, the matrix product $X_j^t V_j^{-1}$ is equal to the $1 \times N$ matrix, where each entry is the sum of a column of V_j^{-1}—that is, we have

$$X_j^t V_j^{-1} = \frac{\begin{bmatrix} a-b & a-b & \cdots & a-b \end{bmatrix}}{a\big(a+(N-2)b\big)-(N-1)b^2},$$

where the numerator is a vector of length N and the denominator is a scalar. Multiplying the above vector by X_j on the right, that is, summing up all the entries in the vector, we obtain the 1×1 matrix

$$X_j^t V_j^{-1} X_j = \frac{N(a-b)}{a\big(a+(N-2)b\big)-(N-1)b^2}.$$

Similarly, for the second term in Equation 4.1, we have the 1×1 matrix

$$X_j^t V_j^{-1} y_j = \frac{(a-b)(y_{j1}+y_{j2}+\cdots+y_{jN})}{a\big(a+(N-2)b\big)-(N-1)b^2}.$$

Finally, noting that the denominators are all the same, we can sum these expressions across $j = 1, 2, \ldots, J$, resulting in

$$\left(\sum_{j=1}^{J} X_j^t V_j^{-1} X_j\right)^{-1} \left(\sum_{j=1}^{J} X_j^t V_j^{-1} y_j\right) = \frac{a\big(a+(N-2)b\big)-(N-1)b^2}{JN(a-b)} \cdot \frac{(a-b)(y_{11}+y_{12}+\cdots+y_{JN})}{a\big(a+(N-2)b\big)-(N-1)b^2}$$

$$= \frac{y_{11}+y_{12}+\cdots+y_{JN}}{JN} = \bar{y}.$$

Therefore, we have that the GLS estimator of β in the balanced Bühlmann model is equal to the sample mean.

For the Bühlmann–Straub model, we only have to make some minor changes to the specification in the LMM we used in the balanced Bühlmann case. We allow an unequal number of observations n_j for each group, and we need to incorporate weights w_{jt} for each observation so that larger weights result in a smaller within-group variance; hence, we will set the $n_j \times n_j$ variance-covariance matrix R_j to be equal to

$$R_j = \begin{bmatrix} \frac{\sigma^2}{w_{jt}} & 0 & \cdots & 0 \\ 0 & \frac{\sigma^2}{w_{jt}} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \frac{\sigma^2}{w_{jt}} \end{bmatrix}.$$

This is the same construction we would use when specifying a weighted least squares regression.

The design matrices X_j and Z_j both have dimension $n_j \times 1$, and all entries are equal to 1. The variance-covariance matrix D has dimension 1×1, and we write its single entry as σ_b^2. The parameters to be estimated are β and $\tau = (\sigma_b^2, \sigma^2)$.

To summarize, the LMM with one level of grouping can be written as

$$Y_j = X_j\beta + Z_j b_j + \epsilon_j, \quad j = 1, 2, \ldots, J,$$

where j is the grouping variable and n_j is the number of observations for the jth group. The fixed-effects vector β has p components because the design matrices X_j have p columns representing the explanatory variables. The matrices Z_j have dimension $n_j \times q$ as we have q explanatory variables for the random effects. The random vectors b_j and ϵ_j have dimension n_j and follow normal distributions $\mathcal{N}(0, D)$ and $\mathcal{N}(0, R_j)$, where the matrices D and R_j are the variance-covariance matrices with dimensions $q \times q$ and $n_j \times n_j$, respectively. These matrices must be symmetric and positive definite (otherwise they are not valid variance-covariance matrices).

4.5. Hachemeister's Regression Model Revisited

For the Hachemeister data, we have the severity of claims over 12 quarters from a sample of five states. We have seen from Figure 3.4 that different intercepts and slopes for these states are a reasonable starting point, and we would like to make inferences from the larger population of states that the five came from. Therefore, we will specify an LMM where the linear predictor for the fixed effects has an intercept and the variable

time, and we use the same linear predictor for the random effects. This way we will get a random intercept and a random slope for each state.

```
hm.dta <- read_csv("hachemeister-data.csv",
                   col_types = "fidd")
```

A natural way to specify the LMM in the lmer() function is as follows:

```
hm.mixed.1 <- lmer(severity ~ time + (time | state),
                   data = hm.dta,
                   weights = claims/1000)
```

```
boundary (singular) fit: see help('isSingular')
```

We'll discuss the message displayed regarding the boundary fit after we present the results of the fit.

The first part of the right-hand side of the formula, in this case just time, represents the fixed effects, and the second part, within parentheses, that is, (time | state), is the random component. For both components we did not specify an intercept because **R** includes one by default. The vertical bar within the random component separates the specification for the linear predictor and the grouping variable, for this example, state. For the weights, we have divided the number of claims by 1,000 to keep the numbers in the calculations from getting too large and causing numerical difficulties in the estimation algorithm.

Notice that we have not provided any instructions on what kind of variance-covariance matrices D or R_j we want to use. The matrix D is of dimension 2×2 (the Z_j matrices have two columns: intercept and time), and the matrices R_j are of dimension 12×12 because for each state we have 12 quarterly observations. The default behavior for lmer() is to have R_j be a diagonal matrix with entries equal to σ^2/w_{jt}, since we specified a weights argument in the call. The matrix D will be a general 2×2 variance-covariance matrix. The (1, 1) position is the variance of the intercept, the (2, 2) position is the variance of the slope, and the (2, 1) or (1, 2) positions are the covariance between intercept and slope. Thus for this model we will be estimating two fixed effects, β_0 and β_1, three entries for the matrix D, and the residual variance σ^2.

The summary of the above fitted model is provided below, where you will see two sections—one labeled "Random effects" and the other "Fixed effects."

```
summary(hm.mixed.1)
```

```
Linear mixed model fit by REML ['lmerMod']
Formula: severity ~ time + (time | state)
   Data: hm.dta
Weights: claims/1000
```

```
REML criterion at convergence: 782.4

Scaled residuals:
    Min      1Q  Median      3Q     Max
-1.9751 -0.6266 -0.2336  0.6918  2.6742

Random effects:
 Groups   Name        Variance  Std.Dev.  Corr
 state    (Intercept)  11990.2    109.50
          time           553.2     23.52  1.00
 Residual              47599.0    218.17
Number of obs: 60, groups: state, 5

Fixed effects:
            Estimate  Std. Error  t value
(Intercept)  1501.29       60.76   24.707
time           27.75       11.69    2.374

Correlation of Fixed Effects:
     (Intr)
time  0.541
optimizer (nloptwrap) convergence code: 0 (OK)
boundary (singular) fit: see help('isSingular')
```

The fixed-effects section tells us, for this example, that the population average severity at time $t = 0$ is \$1,501.29 and that, for each additional quarter, average severity will increase by \$27.75. Also we have the standard errors of these coefficients and their t-values, that is, the estimate divided by its standard error. We have evidence in our data that the fixed effects are different from zero.

From the random effects section, we have almost all the information for the remaining parameters. The column labeled `Variance` has the variances for the intercept and for the variable `time`, as well as the residual variance $\hat{\sigma}^2 = 4.7599 \times 10^4$. Note that the column labeled `Std.Dev.` is *not the standard error of the estimated variances*. This column is just the square root of the entries in the `Variance` column. We do not have the estimated covariance between the intercept and `time`, but we can get that information by extracting the entire variance-covariance matrix D and the residual variance too.

```
print(VarCorr(hm.mixed.1), comp = "Variance")
```

```
Groups    Name        Variance  Cov
state     (Intercept) 11990.16
          time          553.16  2575.368
Residual              47598.96
```

The very last line of the summary output shows the message `boundary (singular) fit: see help('isSingular')`. This tells us that during the optimization one of our parameters has reached its boundary.

We can compute the eigenvalues of the variance-covariance matrix D via

```
eigen(VarCorr(hm.mixed.1)$state)$value
```

```
[1]  1.254332e+04  1.136868e-13
```

and see that the second eigenvalue is essentially zero; hence, our matrix D is super close to not being positive definite. A positive definite matrix must have all of its eigenvalues positive. If at least one of them is zero, then the matrix is positive semidefinite.

This is an indication that the default choice of matrix D with three free parameters is not ideal. Therefore, for our next mixed model we will restrict the variance-covariance matrix D to be diagonal. To specify such a model, we need to tell the `lmer()` function that we want independent random components for the intercept and the slope parameters even though both use the same grouping variable. We do this by specifying *two* random effects in the formula for `lmer()`. The first one gives us the random intercept and the second, the random slope. Note that we need to tell **R** explicitly not to include an intercept in the second random effect by using `0 + time`.

```
hm.mixed.2 <- lmer(severity ~ time + (1 | state) +
  (0 + time | state),
                   data = hm.dta,
                   weights = claims/1000)
summary(hm.mixed.2)
```

```
Linear mixed model fit by REML ['lmerMod']
Formula: severity ~ time + (1 | state) + (0 + time | state)
   Data: hm.dta
Weights: claims/1000

REML criterion at convergence: 785.5

Scaled residuals:
    Min      1Q  Median      3Q     Max
-1.8951 -0.6462 -0.2441  0.5449  2.6713

Random effects:
 Groups    Name        Variance  Std.Dev.
 state     (Intercept)  19909.0     141.1
 state.1   time           605.1      24.6
 Residual               48723.8     220.7
Number of obs: 60, groups: state, 5

Fixed effects:
            Estimate  Std. Error  t value
(Intercept)  1491.99       78.46   19.015
time           29.55       12.76    2.315
```

```
Correlation of Fixed Effects:
      (Intr)
time  -0.248
```

From the random-effects section of the summary output, we see only estimates for the diagonal elements of D and the residual variance σ^2. Note that the fixed effects are slightly different than those in our previous model, but for all practical purposes they are the same. The estimated random effects, deviations from the fixed effects, for each state are

```
ranef(hm.mixed.2)$state
```

```
   (Intercept)       time
1   162.868302   32.78440
2   -78.554481  -13.24185
3    44.108233   12.01649
4  -137.982018  -16.88416
5     9.559964  -14.67488
```

and putting these together with the fixed effects we obtain the following credibility-weighted values:

```
round(fixef(hm.mixed.2) +
  t(as.matrix(ranef(hm.mixed.2)$state)), 2)
```

```
                    1        2        3        4        5
(Intercept)  1654.86  1413.44  1536.10  1354.01  1501.55
time           62.34    16.31    41.57    12.67    14.88
```

These are not very different from the estimates we got in Chapter 3 (Table 3.4), and they suffer from the same issues that we noted there. Many of the intercepts and slopes are not between the individual states and collective estimates.

We cannot easily extract the credibility matrices from the fitted model, but Table 1 of Frees et al. (1999) provides an explicit formula by which to calculate these matrices from information we do readily have from the model. The formula is

$$A_j = \frac{\det\left(DW_j\right) I_2 + \hat{\sigma}^2 DW_j}{\det\left(DW_j\right) + \hat{\sigma}^2 \operatorname{trace}\left(DW_j\right) + \hat{\sigma}^4}, \tag{4.2}$$

where I_2 is a 2×2 identity matrix, W_j is given by

$$W_j = \begin{bmatrix} \sum_{k=1}^{n_j} w_{jk} & \sum_{k=1}^{n_j} t_{jk} w_{jk} \\ \sum_{k=1}^{n_j} t_{jk} w_{jk} & \sum_{k=1}^{n_j} t_{jk}^2 w_{jk} \end{bmatrix},$$

"det" is the determinant of a matrix, and "trace" is the sum of the diagonal elements of a square matrix. This formula for the credibility matrix A_j matches the formula we developed in Chapter 3, Equation 3.27. And as we saw in that chapter, the reason for credibility-weighted estimates not to lie between the stand-alone and collective estimates is that the matrix A_j is not diagonal.

Exercise 4.4 Show that Equation 4.2 matches Equation 3.27 in the case where the variance-covariance matrix D is diagonal with entries τ_0^2 and τ_1^2.

Solution 4.4 Before looking at Appendix B for a derivation, try it yourself. Start with the denominator in Equation 4.2 and then work on the numerator. The denominator is just a number, and the numerator is a 2×2 matrix. The calculations are straightforward but tedious.

We also saw in Chapter 3 that to achieve a diagonal credibility matrix we can center the time variable at each group's center of gravity. We can add a centered time variable, `ctime`, to our dataset via

```
CG <- with(hm.dta,
           tapply(time * claims, state, sum) /
             tapply(claims, state, sum))
hm.dta$ctime <- hm.dta$time - CG[hm.dta$state]
rm(CG)
```

and fit the same LMM by replacing `time` with `ctime`.

```
hm.mixed.3 <- lmer(severity ~ ctime + (1|state) +
  (0 + ctime|state),
                   data = hm.dta,
                   weights = claims/1000)
summary(hm.mixed.3)
```

```
Linear mixed model fit by REML ['lmerMod']
Formula: severity ~ ctime + (1 | state) + (0 + ctime | state)
   Data: hm.dta
Weights: claims/1000

REML criterion at convergence: 788.6

Scaled residuals:
    Min      1Q  Median      3Q     Max
-2.0491 -0.7028 -0.1559  0.5252  2.6324
```

```
Random effects:
 Groups    Name         Variance  Std.Dev.
 state     (Intercept)   70838.8    266.16
 state.1   ctime           446.4     21.13
 Residual                49019.8    221.40
Number of obs: 60, groups: state, 5

Fixed effects:
             Estimate  Std. Error  t value
(Intercept)   1674.96      122.12   13.715
ctime           34.09       11.59    2.942

Correlation of Fixed Effects:
      (Intr)
ctime  0.000
```

Note that the fixed effect for the intercept has changed significantly because now we are measuring the mean severity at time approximately $t = 6.5$ instead of at time $t = 0$. Also worth noting is the correlation of the fixed effects shown at the very bottom of the summary output. Now the intercept and the slope have a zero correlation, whereas in our previous model it was –0.248. We expected this result because by centering the time variable we have made the intercept and the centered time variable orthogonal to each other.

The credibility-weighted estimates from this model are

```
round(fixef(hm.mixed.3) +
  t(as.matrix(ranef(hm.mixed.3)$state)), 2)
```

```
                  1        2        3        4        5
(Intercept)  2058.27  1516.73  1799.56  1398.97  1601.24
ctime          60.02    22.45    39.63    32.08    16.27
```

and they now lie between the stand-alone and the collective estimates.

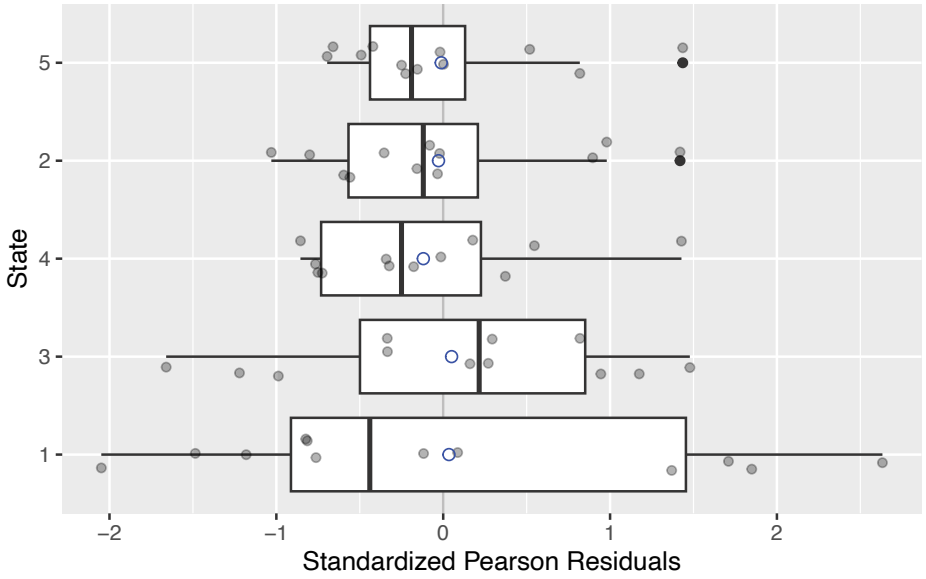
Now that we have an initial model, `hm.mixed.3`, we should check whether our distributional assumptions are valid for the data. There are two basic assumptions:

1. the within-group errors are independent and identically normally distributed with mean zero and variance σ^2/w_{jt} and are independent of the random effects, and
2. the random effects are normally distributed with mean zero and variance-covariance matrix D that does not depend on the group and are independent for different groups.

Checking Within-Group Errors Assumptions

To assess the within-group residuals we can use a boxplot of standardized residuals by state. Figure 4.6 shows such a plot for the `hm.mixed.3` model where we have also included the residual points and a larger blue circle at the mean value of the residuals.

Figure 4.6. Boxplots and underlying data of standardized residuals for the `hm.mixed.3` model. The large blue circle is the average value of the residuals. The states have been ordered by the size of the interquartile range of their residuals.



The states have been arranged in order of the size of their interquartile range, and we can see that there is an increasing spread. Even though all five groups of residuals are centered about zero, their spread is not constant. This suggests that a single within-group variance, σ^2, parameter may not be correct, and we might need to select a more general model where each group has its own parameter.

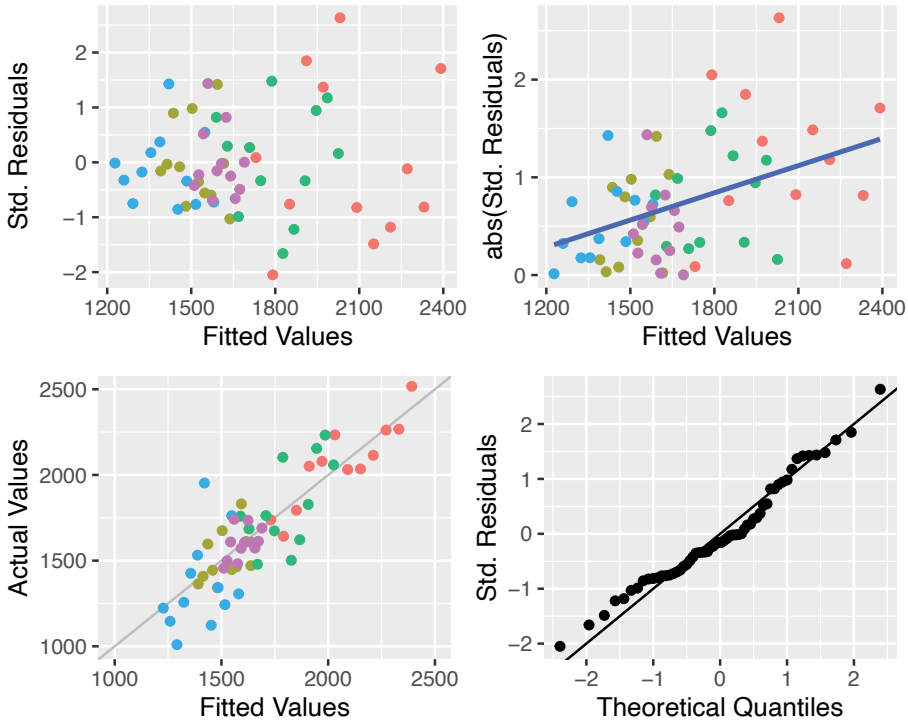
Figure 4.7 shows other diagnostic plots to assess the assumptions about the within-group residuals for the `hm.mixed.3` model. The top-row panels confirm that the residuals do not have constant variability by state. The top-left plot shows a fanning out of the residuals as the fitted values increase in size. The top-right panel displays the absolute value of the residuals against the fitted values, and we have superimposed a least squares estimated trend line that clearly shows our residuals spreading out as the fitted values increase.

The bottom-left panel is an actual-versus-expected plot together with the line $y = x$. It looks like we have most points scattered around the line $y = x$. There is one possible outlier: the point with coordinates close to (1400, 2000). On the bottom right, we have displayed a QQ plot. We would like the points to be on the line $y = x$, and most of them do follow this pattern. But in the lower-left corner, as the theoretical quantiles increase toward negative infinity, all the points are above the line $y = x$. This tells us that our data has a thinner left-hand tail compared with the normal distribution.

Checking Random Effects Assumptions

The second assumption we need to check is the one about the random effects. They should be normally distributed with a mean of zero and variance-covariance matrix D.

Figure 4.7. Diagnostic plots for the `hm.mixed.3` model.



For Hachemeister's data we have only five observations, and so it will be difficult to draw definitive conclusions.

Figure 4.8 displays the QQ plots for the random effects from model `hm.mixed.3`. We expect the points in these plots to lie along a straight line. In this case the assumption of normality seems reasonable. While the points in both panels are not perfectly on a line, their departure is not excessive.

We should also check that the random effects follow a multivariate normal distribution with mean $\mu = (0, 0)$ and variance-covariance matrix D. The probability density function for our two-dimensional example is

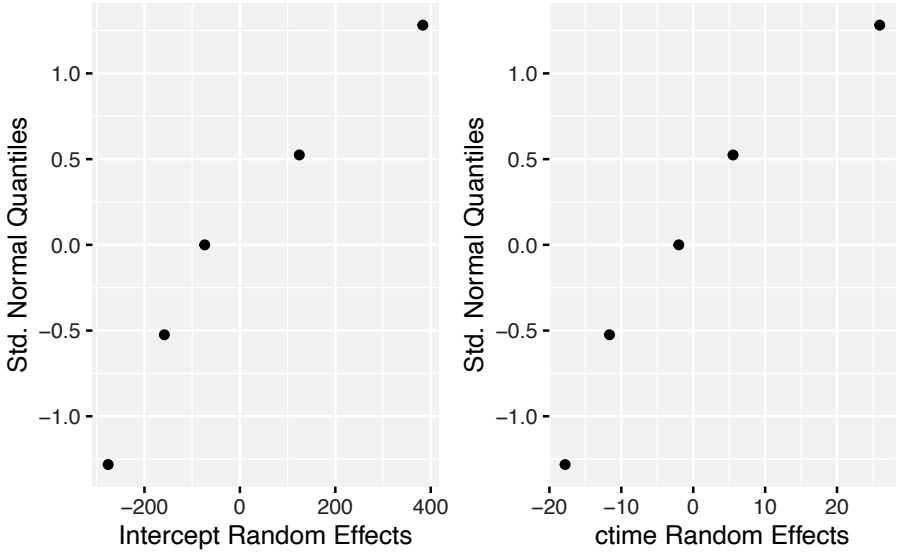
$$f\left(\begin{bmatrix} x_1 \\ x_2 \end{bmatrix}\right) = \frac{1}{2\pi\sqrt{\det(D)}} \exp\left(-\frac{1}{2}\begin{bmatrix} x_1 & x_2 \end{bmatrix} D^{-1} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}\right).$$

Note that the value of the density function f depends on x_1 and x_2 only through the value of the expression inside the exponential function, namely, through what is called the squared Mahalanobis distance:

$$d^2 = \begin{bmatrix} x_1 & x_2 \end{bmatrix} D^{-1} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = ax_1^2 + 2bx_1x_2 + cx_2^2,$$

if D^{-1} has diagonal entries equal to a and c and off-diagonal entries equal to b. The set of points (x_1, x_2) with Mahalanobis distance d^2 all have the same value of the

Figure 4.8. QQ plots of the random effects for the Hachemeister `hm.mixed.3` model.



density function f; that is, these points create a contour line in the three-dimensional $(x_1, x_2, f([x_1, x_2]^t))$ surface of the density function.

In our example, the matrix D has been estimated as

$$D = \begin{bmatrix} 70,838.77 & 0 \\ 0 & 446.39 \end{bmatrix}.$$

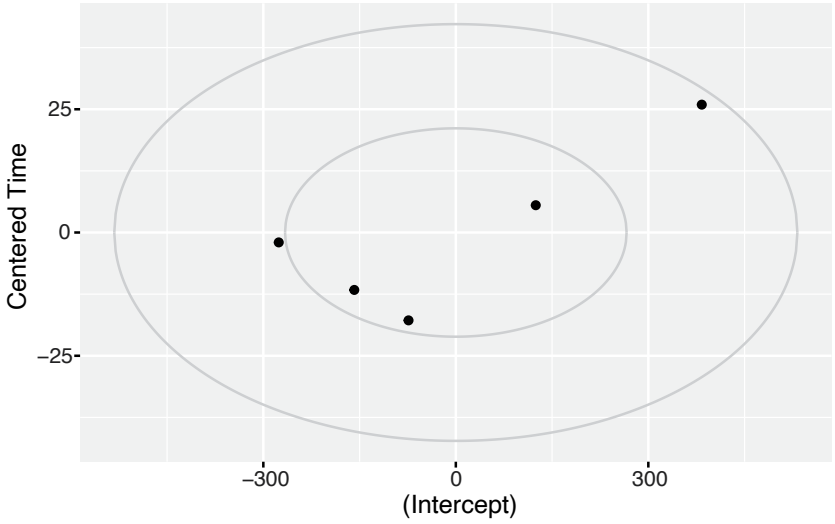
Its inverse, D^{-1}, is also diagonal; hence, $b = 0$, and so we can see that the set of points (x_1, x_2) that have constant Mahalanobis distance form an ellipse whose major and minor axes fall along the x and y axes.

Figure 4.9 shows a scatterplot of the random effects for model `hm.mixed.3` along with ellipses at a Mahalanobis distance of 1 and 2 standard deviations away from the origin. Note that the five points we have available are all within two standard deviations.

If the variance-covariance matrix D had not been diagonal, we would still have elliptical contours, but they would have been rotated around the origin (the mean of our bivariate normal distribution is (0, 0)).

Note that in Figure 4.9 we chose to display the set of points that are a Mahalanobis distance of 1 and 2 away from the origin because we are all very familiar that, in the one-dimensional standard normal distribution, within these distances we have about 68% and 95% of the total density. For a two-dimensional normal distribution these values do not give us the same proportion of the total density. To find the appropriate values we need to know that the squared Mahalanobis distance d^2 has a chi-squared distribution with p degrees of freedom, where p is the dimension of the multivariate normal distribution.

Figure 4.9. Scatterplot of estimated random effects for model `hm.mixed.3` along with contour lines that are 1 and 2 standard deviations away from the origin.



In our case, $p = 2$, and if we are looking to obtain the same 68% and 95% coverage, we must choose the Mahalanobis distance equal to the following values:

```
crit.points <- sqrt(qchisq(c(0.68, 0.95), df = 2))
names(crit.points) <- c("68%", "95%")
crit.points
```

```
     68%       95%
1.509592  2.447747
```

Exercise 4.5 Using the `mvtnorm` package, generate 2,000 multivariate random points with mean $\mu = (0, 0)$ and variance-covariance matrix

$$D = \begin{bmatrix} 70,838 & 0 \\ 0 & 446 \end{bmatrix}.$$

Plot the points using three different colors depending on whether the points are within Mahalanobis distance 1, between 1 and 2, or beyond 2. Check the proportion of points to discern whether 68% or 95% are within Mahalanobis distance 1 or 2 of the origin.

Solution 4.5 Let us set up the mean vector $\mu = (0, 0)$ and the variance-covariance matrix D.

```
N <- 2000
mu <- c(0,0)
D <- matrix(c(70838, 0, 0, 446),
            nrow = 2, ncol = 2)
```

Using the `rmvnorm()` function we can simulate the points via

```
set.seed(12837)
z <- rmvnorm(N, mean = mu, sigma = D)
```

The object `z` will be a $2{,}000 \times 2$ matrix where each row is one simulated point. Next we want to compute the Mahalanobis distance for each row z_i using the formula

$$\sqrt{\left(z_i - \mu\right)^t D^{-1} \left(z_i - \mu\right)}.$$

We can do this by using the `apply()` function, and we also need to create a categorical variable to distinguish which points are at different Mahalanobis distances. We will put all of this into a data frame

```
tb <- as.data.frame(z)
names(tb) <- c("x", "y")
tb$md <- apply(z, 1, function(z) sqrt(t(z - mu) %*%
                                      solve(D) %*% (z - mu)))
tb$md.bin.1 <- cut(tb$md,
                   breaks = c(-Inf, 1, 2, Inf),
                   labels = c("d < 1", "1 <= d < 2", "d >= 2"))
tb$md.bin.2 <- cut(tb$md,
                   breaks = c(-Inf, crit.points, Inf),
                   labels = c("d < 1.509",
                              "1.509 <= d < 2.448",
                              "d >= 2.448"))
```

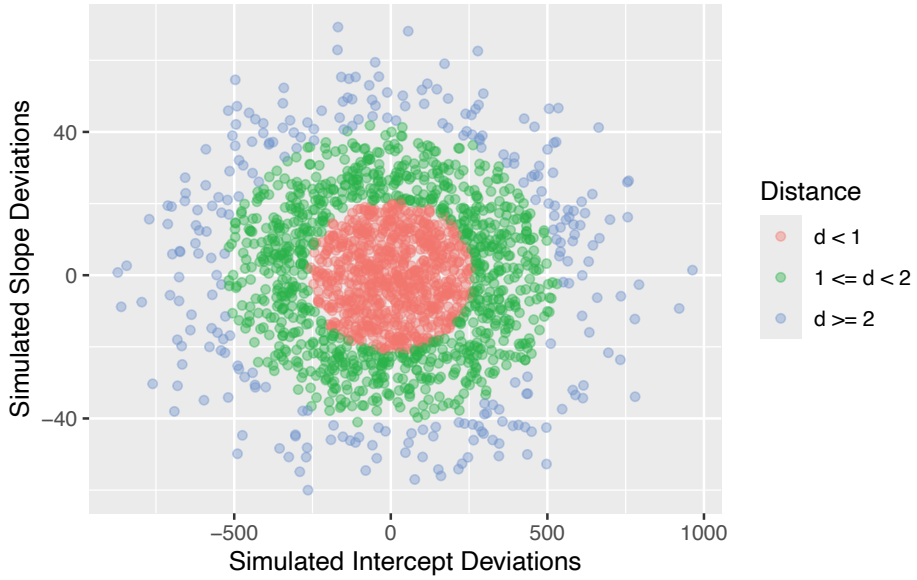
and create the scatterplot.

```
ggplot(data = tb,
       mapping = aes(x = x,
                     y = y,
                     color = md.bin.1)) +
  geom_point(alpha = 0.4) +
  labs(x = "Simulated Intercept Deviations",
       y = "Simulated Slope Deviations",
       color = "Distance")
```

The proportion of points in each colored region is given by

```
xtabs( ~ md.bin.1, data = tb) / N
```

```
md.bin.1
     d < 1   1 <= d < 2    d >= 2
     0.395        0.471     0.134
```



Note how we have only about 40% of the points within a Mahalanobis distance of 1 and about 86.6% within a distance of 2. If we use the correct thresholds given by the chi-squared distribution—namely, 1.509 and 2.448—then we would have the appropriate coverage, as shown next.

```
xtabs( ~ md.bin.2, data = tb) / N
```

```
md.bin.2
     d < 1.509  1.509 <= d < 2.448  d >= 2.448
        0.6940              0.2495       0.0565
```

4.6. Summary

In this chapter we revisited three classic credibility models,

- the balanced Bühlmann model,
- the Bühlmann–Straub model, and
- Hachemeister's credibility regression model

and expressed them in terms of the well-developed branch of statistics known as linear mixed models. For the practicing actuary, embracing LMMs to implement credibility techniques brings substantial benefits. By using this theory, we can bring all the machinery that statisticians have developed to bear on our applications and apply standard software to carry out the necessary computations. We also have at our disposal inference techniques and model-checking procedures. More importantly, LMMs allow us to capture the correlation that exists in many of our datasets, and we have many models to choose from.

We looked closely at LMMs with one level of grouping and introduced the concepts of fixed effects and random effects. One way of thinking about these effects,

but perhaps not a very good one (as pointed out on page 245 of Gelman and Hill [2007]), is that fixed effects estimate features of the population from which our sample was taken and we use random effects for those variables whose values are just a sample of the possible values the population has.

One clear disadvantage of the LMM is that the random effects and the response variable must be normally distributed. This restriction is a serious one for actuarial work, but in the next chapter we will introduce generalized linear mixed models. That class of statistical models expands the well-known framework of GLMs that many actuaries use to include random effects with distributions from the exponential family. With such an expanded set of models, actuaries can significantly increase their modeling capabilities.

5. Generalized Linear Mixed Models

5.1. Introduction

In the previous chapter we introduced LMMs and saw how three classical credibility models are special cases of the linear mixed model theory. LMMs are characterized by having a response variable that is normally distributed and by having random effects that are also normally distributed. These models are an extension of the classical OLS model and applicable in many situations. But we know that real-world data is richer and more complex than the normal distribution can accommodate. The extension of the classical linear model to the generalized linear model, or GLM, where the response distribution is a member of the exponential family, has opened up a new area of techniques and tools well suited to the data and problems that actuaries encounter in practice.

Over the past two decades, actuaries have made good use of GLMs. The next step in expanding such models to more complex data structures is to introduce random effects into the GLM framework and allow those random effects to have other distributions besides the normal.

Also, another important extension focuses on the dispersion parameter. GLM theory keeps the dispersion parameter fixed across all observations. From experience, we know that such a fixed parameter is not always ideal. It would be helpful to link the dispersion parameter to some explanatory variables.

In this chapter we introduce an extension of GLMs known as *hierarchical generalized linear models*, or HGLMs, which will allow us to have random effects whose distributions come from a broader family and to model the dispersion parameter via explanatory variables. Such models are based on the theory of *h*-likelihood, which brings together both Bayesian and frequentist perspectives.

In the next section, we give a brief conceptual introduction to the HGLM without delving too much into the theory. Then, we present several examples of how to use these new models. Our discussion follows the work of Lee et al. (2021) and Lee et al. (2020) closely.

5.2. Hierarchical Generalized Linear Models

HGLMs were introduced in Lee and Nelder (1996). These models use a generalization of the likelihood function called *hierarchical* likelihood, or *h*-likelihood. The maximization of this extended likelihood, under appropriate conditions, gives estimates of both fixed as well as random effects and the dispersion parameter.

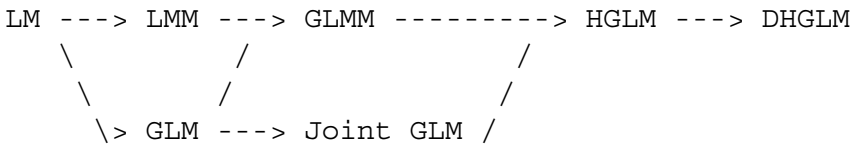
Starting with the linear model, researchers worked in two directions to expand it. The first path created the linear mixed-effects model, or LMM, where the linear predictor can have random terms that are normally distributed. The second path worked on introducing an expanded list of response distributions, giving us the GLM. The combination of the two yields the generalized linear mixed model, or GLMM, with distributions for the response variable from the GLM and random effects in the linear predictor from the LMM.

Also from the GLM framework, practitioners and researchers worked at enhancing the capabilities in modeling the variance of the response variable. GLMs use a single multiplier, ϕ, the dispersion parameter to scale the variance function. In many situations, this single parameter is not adequate to capture the volatility of the response, so a link function and a linear predictor were introduced for the dispersion parameter, giving rise to *joint GLMs*.

Combining GLMMs and joint GLMs and expanding the distributional assumptions for the random effects brings us to the HGLM. This model has a response variable whose distribution comes from the exponential family. The linear predictor has fixed and random effects, and these random effects are not constrained to be normally distributed. The dispersion parameter can be modeled via a separate link function tied to a different linear predictor with fixed effects.

And, finally, we have the *double hierarchical generalized linear model* (DHGLM), where we take an HGLM and allow random effects in the dispersion model and can also introduce explanatory variables via a link function and a linear predictor into the variance of the random effects.

The following diagram is a crude representation of how the various models are interconnected (adapted from Lee et al. 2020, 3).



To help us translate between the mathematical description of a model and the **R** code necessary to implement the model, consider the following mixed model:

$$g\Big(\mathbb{E}\big[y\big]\Big) = X\beta + Zv,$$

where y is the response variable, $g()$ is the link function for the mean, β represents the fixed effects, and v are the random effects. The matrices X and Z are the design matrices for the fixed and random effects, respectively. We also need to specify the distribution of the response variable (a member of the exponential family) and the distribution of the random effects, that is, $v \sim F(\lambda)$, where F just stands for a distribution such as the Gaussian or gamma with parameter vector λ.

So far we have only described a model for the mean. If we are also modeling the dispersion, ϕ, parameter, then we would have

$$h\Big(\mathbb{E}\big[\phi\big]\Big) = W\gamma + Mu,$$

where $h()$ is a link function, γ are fixed effects, u are random effects, and W and M are design matrices. Because the u are random, we also have to specify their distribution, which will come with some parameters.

The full specification of a model can be complex, but using the notation introduced in Chapter 6 of Lee et al. (2020) makes things more manageable. A DHGLM is represented by a pair

$$\Big\{\text{model}\big(\mu\big), \text{model}\big(\phi\big)\Big\},$$

where the first entry is the model for the *mean* and the second entry represents the model for the *dispersion*.

For example, the usual GLM would be written as {GLM(μ), ϕ}, where ϕ is a constant. A joint model would be written as {GLM(μ), GLM (ϕ)}, where we have two regular GLM models that are interlinked to form the joint model. If we want to include a random effect in the model for the mean, we can write it as {HGLM(μ), ϕ}, and if we also want to have the dispersion parameter modeled we would say {HGLM(μ), GLM(ϕ)}.

There are two **R** packages to fit these models: `hglm` and `dhglm`. We'll use the first one briefly when we revisit the Hachemeister data because it allows us to use nearly the same calling code as we did in the previous chapter. But we will mostly use the `dhglm` package.

The `dhglm` package uses two functions to fit a model. The first, `DHGLMMODELING()`, creates the appropriate structures for the mean and dispersion models. The second, `dhglmfit()`, does the actual computations. As an example of their use, a standard log-link Poisson GLM model {GLM(μ), ϕ}, for frequency where the dataset is named `accidents` and the response variable is `count`, the predictor variables are `age` and `gender`, and the amount of exposure to risk is in the variable `exposure` would be specified and fitted as follows:

```
model.mu <- DHGLMMODELING(Model = "mean",
                          Link = "log",
                          LinPred = count ~ age + gender,
                          Offset = log(exposure))
model.phi <- DHGLMMODELING(Model = "dispersion")

fit <- dhglmfit(RespDist = "poisson",
              DataMain = accidents,
              MeanModel = model.mu,
              DispersionModel = model.phi)
```

5.3. Examples

In this section we present four examples to familiarize the reader with specifying and fitting models with the `hglm` and `dhglm` packages:

1. Hachemeister
2. Fabric faults
3. Train accidents
4. Diabetes progression

Quick Revisit with the Hachemeister Data

In the last chapter we fitted several LMMs to the Hachemeister data. Here we will refit the last model, `hm.mixed.3`, using the machinery from HGLMs and compare results. The main function to fit HGLMs is `hglm2()` and can be found in the `hglm` package.

Model `hm.mixed.3` used a centered version of time called `ctime`, that is, a weighted average of time where the weights are the number of claims. Let's load our data and compute `ctime`.

```
hm.dta <- read_csv("hachemeister-data.csv",
                   col_types = "fidd")
CG <- with(hm.dta,
           tapply(time * claims, state, sum) /
             tapply(claims, state, sum))
hm.dta$ctime <- hm.dta$time - CG[hm.dta$state]
```

Using `hglm2()` we specify the model in the same way as before. The response variable is `severity`, and we have fixed effects for the intercept and the time variable `ctime`. We also include uncorrelated random effects for the intercept via `(1|state)` and time `(0 + ctime | state)`. Both of the random effects vary by state. The response distribution is specified to be normally distributed with an identity link function through the `family` parameter. The random effects are also normally distributed with an identity link function via the `rand.family` parameter.

```
hm.hglm.3 <- hglm2(severity ~ ctime + (1 | state) +
                     (0 + ctime | state),
                   data = hm.dta,
                   family = gaussian(link = "identity"),
                   rand.family = gaussian(link = "identity"),
                   weights = claims)
```

The summary output from the fitting process contains two major sections: one for the mean model and the other for the dispersion model. Our model did not specify any structure for the dispersion model, and so it is taken to be a single parameter.

```
summary(hm.hglm.3)
```

```
Call:
hglm2.formula(meanmodel = severity ~ ctime + (1 | state) +
  (0 + ctime|state), data = hm.dta, family =
  gaussian(link = "identity"), rand.family =
  gaussian(link = "identity"), weights = claims)

----------
MEAN MODEL
----------

Summary of the fixed effects estimates:

             Estimate  Std. Error  t-value  Pr(>|t|)
(Intercept)   1674.93      122.28   13.697   < 2e-16  ***
ctime           34.09       11.58    2.944   0.00484  **
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Note: P-values are based on 52 degrees of freedom

Summary of the random effects estimates:

                       Estimate  Std. Error
(Intercept)| state:1   383.3472    123.4313
(Intercept)| state:2  -158.2215    127.8617
(Intercept)| state:3   124.6477    130.2105
(Intercept)| state:4  -276.0770    145.3201
(Intercept)| state:5   -73.6963    125.4177

Summary of the random effects estimates:

                Estimate  Std. Error
ctime| state:1   25.9279     12.2453
ctime| state:2  -11.6410     14.2447
ctime| state:3    5.5344     15.0492
ctime| state:4   -2.0069     17.8084
ctime| state:5  -17.8144     13.2125

----------------
DISPERSION MODEL
----------------

NOTE: h-likelihood estimates through EQL can be biased.

Dispersion parameter for the mean model:
[1] 49017713

Model estimates for the dispersion term:

Link = log
```

```
Effects:
 Estimate  Std. Error
  17.7077      0.1969

Dispersion = 1 is used in Gamma model on deviances to calculate
the standard error(s).

Dispersion parameter for the random effects:
[1] 70873.4  446.5

Dispersion model for the random effects:

Link = log

Effects:
.|Random1
  Estimate  Std. Error
   11.1687      0.7257

.|Random2
  Estimate  Std. Error
    6.1013      0.8774

Dispersion = 1 is used in Gamma model on deviances to calculate
the standard error(s).

EQL estimation converged in 3 iterations.
```

The fixed-effects estimates are shown first, followed by the random effects for the intercept and then the random effects for the predictor variable `ctime`. Putting together the estimated intercepts and slopes (fixed effects plus random effects) from the model above, `hm.hglm.3`, we obtain

```
                    1         2         3         4         5
(Intercept)  2058.280  1516.712  1799.581  1398.856  1601.237
ctime          60.019    22.450    39.625    32.084    16.277
```

The estimates we obtained in the previous chapter based on the GLMM `hm.mixed.3` are

```
                    1         2         3         4         5
(Intercept)  2058.273  1516.728  1799.565  1398.973  1601.241
ctime          60.021    22.445    39.627    32.082    16.272
```

Comparing them, they are virtually identical. In addition, other estimated quantities such as the variances for the random effects are very close to each other. The residual variance for model `hm.mixed.3` is equal to 49,019.8, and for our hierarchical model `hm.hglm.3` it is 49,018.3. In model `hm.mixed.3`, the variance for

the intercept and ctime is 70,838.8 and 446.4, respectively, and for the hierarchical model hm.hglm.3, we have 70,873.4 and 446.5—again close to each other.

Textile Fabric Defects

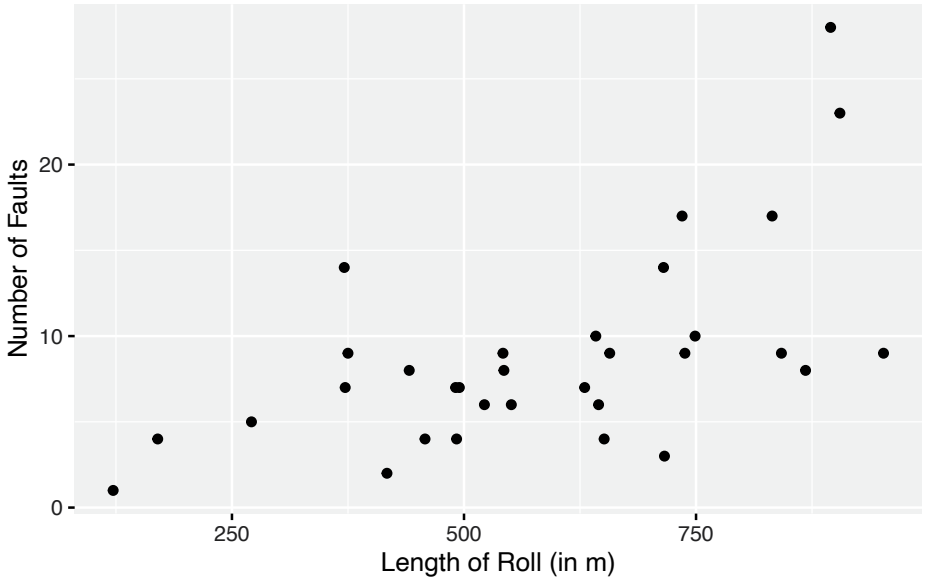
In this example we consider the dataset fabric from Bissell (1972), where we will be investigating the number of faults in rolls of textile fabric. This dataset is available in the mdhglm package, and it has three variables (we have added the logarithm of x as the variable x.lg) and 32 observations. The variable x is the length of the roll, and y is the number of defects. The first few rows of the data are

```
data(fabric, package = "mdhglm")
fabric$x.lg <- log(fabric$x)
head(fabric)
```

```
     x   y  rf      x.lg
1  551   6   1  6.311735
2  651   4   2  6.478510
3  832  17   3  6.723832
4  375   9   4  5.926926
5  715  14   5  6.572283
6  868   8   6  6.766192
```

This dataset has also been analyzed in Lee et al. (2020). The response variable is the number of faults, and so a natural choice would be to use the Poisson distribution. The only predictor variable is the length of the roll of fabric. Figure 5.1 shows that there is a relationship between the response variable and our predictor variable and that the

Figure 5.1. Number of faults in a roll of fabric.

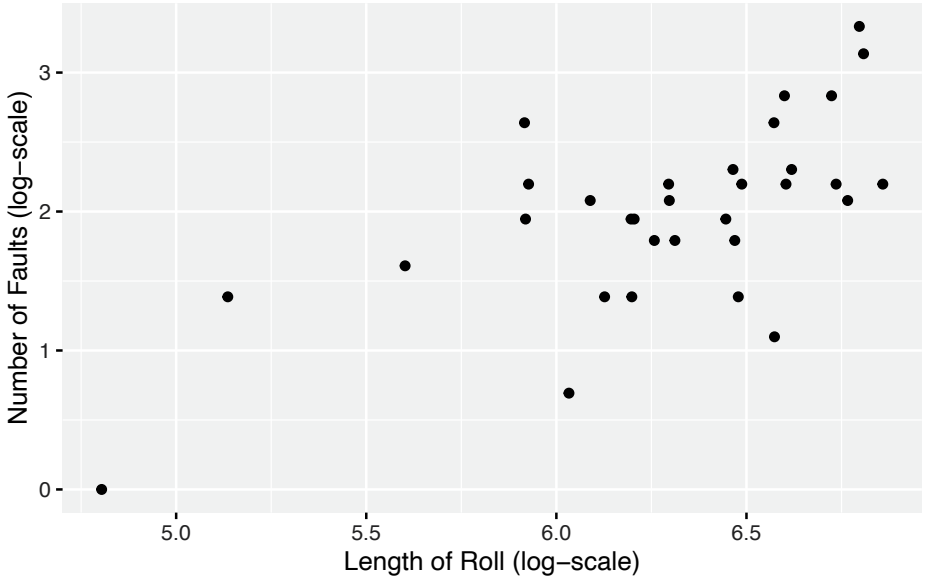


relationship is not linear. As the length of a fabric roll increases, we see an increasing number of defects. If we transform both the response and predictor variables with a logarithmic function (not shown here), the relationship between them seems linear.

Exercise 5.1 Transform both the response and predictor variables by applying a logarithmic function, and plot them. Does the relationship seem linear?

Solution 5.1 While there are several points that do not fall close to a straight line pattern, the overall impression is that these points are indeed closer to a linear pattern than the original data.

```
ggplot(data = fabric,
       mapping = aes(x = log(x),
                     y = log(y))) +
  geom_point() +
  labs(x = "Length of Roll (log-scale)",
       y = "Number of Faults (log-scale)")
```



Therefore, a reasonable starting point would be to use a Poisson GLM with a log-link and the logarithm of the length of a roll (`x.lg`) as our predictor variable—that is, we want to fit the following model:

$$\log\left(\mathbb{E}\left[y_i\right]\right) = \beta_0 + \beta_1 \log\left(x_i\right),$$

where y_i is the number of faults and x_i is the length of the roll of fabric. Fitting such a model yields the following summary:

```
Call:
glm(formula = y ~ x.lg, family = poisson(link = "log"),
  data = fabric)

Coefficients:
            Estimate  Std. Error  z value  Pr(>|z|)
(Intercept)  -4.1730      1.1352   -3.676  0.000237  ***
x.lg          0.9969      0.1759    5.668  1.45e-08  ***
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 103.714 on 31 degrees of freedom
Residual deviance:  64.537 on 30 degrees of freedom
AIC: 191.84

Number of Fisher Scoring iterations: 4
```

Even though both the intercept and the coefficient for the logarithm of the length of a roll of fabric are statistically significant, the model fits very poorly. The residual deviance of 64.5 is extremely large compared with residual degrees of freedom of 30, and we have a clear indication of overdispersion. Perhaps we have misspecified the linear predictor, but given that we have only one variable to work with, there is not much we can do about it. Another reason for the lack of fit could be that our choice of link function (logarithm) is not correct. But we do have some evidence that a log-link function is suitable. Hence, we conclude that the Poisson distribution is not adequate for this data, and we move on to considering the negative binomial distribution.

We know that a negative binomial distribution arises as a mixture of the Poisson and gamma distributions as follows: let u be an unobserved gamma random variable with mean equal to 1 and variance equal to $1/\theta$ and, conditionally on u, let Y be a Poisson random variable with mean equal to λu. Then the marginal distribution of Y will be negative binomial. The standard `glm()` function does not fit negative binomial models, but package `MASS` has the function `glm.nb()` to fit these models. Fitting a negative binomial GLM to the `fabric` data yields the following summary fit information:

```
fab.nb.glm <- glm.nb(y ~ x.lg,
                     data = fabric)
summary(fab.nb.glm)
```

```
Call:
glm.nb(formula = y ~ x.lg, data = fabric,
  init.theta = 8.667407437, link = log)
```

```
Coefficients:
            Estimate  Std. Error  z value  Pr(>|z|)
(Intercept)  -3.7951      1.4577   -2.603   0.00923  **
x.lg          0.9378      0.2280    4.114  3.89e-05  ***
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for Negative Binomial(8.6674) family
 taken to be 1)

    Null deviance: 50.28 on 31 degrees of freedom
Residual deviance: 30.67 on 30 degrees of freedom
AIC: 181.39

Number of Fisher Scoring iterations: 1

              Theta:  8.67
          Std. Err.:  4.17

 2 x log-likelihood:  -175.387
```

Note that the estimated coefficients of this model do not differ significantly from those in the Poisson model, and for this negative binomial model we do not have any evidence of lack of fit. The residual deviance is very close to the residual degrees of freedom. Of course, other diagnostics are needed to fully check the adequacy of this model.

Exercise 5.2 Use the following diagnostic plots to assess the adequacy of the negative binomial model:

1. Quantile residuals vs. fitted values
2. Absolute value of quantile residuals vs. fitted values
3. Quantile residuals vs. predictor variable
4. Linear predictor vs. working responses

Solution 5.2 Let us compute the quantities we need for the diagnostic plots:

```
fabric.res <- fabric |>
  mutate(eta = predict(fab.nb.glm, type = "link"),
         mu = predict(fab.nb.glm, type = "response"),
         rQ = qresid(fab.nb.glm),
         rW = resid(fab.nb.glm, type = "working"),
         wR = rW + eta)
```

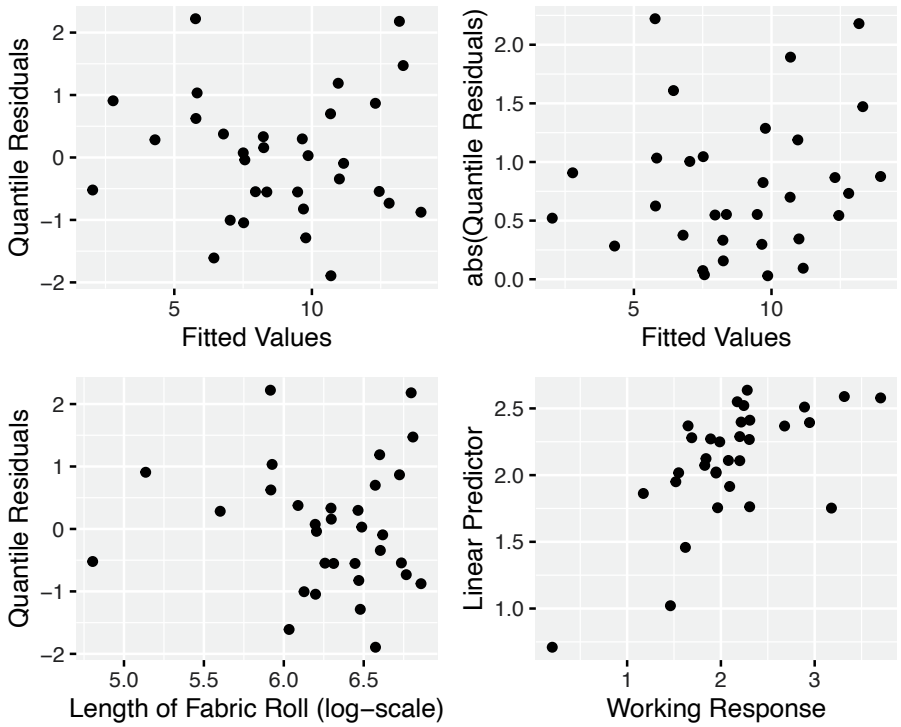
Compute the individual plots.

```
p1 <- ggplot(data = fabric.res,
             mapping = aes(x = mu,
                           y = rQ)) +
```

```
  geom_point() +
  labs(x = "Fitted Values",
       y = "Quantile Residuals")
p2 <- ggplot(data = fabric.res,
             mapping = aes(x = mu,
                           y = abs(rQ))) +
  geom_point() +
  labs(x = "Fitted Values",
       y = "abs(Quantile Residuals)")
p3 <- ggplot(data = fabric.res,
             mapping = aes(x = x.lg,
                           y = rQ)) +
  geom_point() +
  labs(x = "Length of Fabric Roll (log-scale)",
       y = "Quantile Residuals")
p4 <- ggplot(data = fabric.res,
             mapping = aes(x = wR,
                           y = eta)) +
  geom_point() +
  labs(x = "Working Response",
       y = "Linear Predictor")
```

And arrange them in a 2×2 grid.

```
(p1 + p2) / (p3 + p4)
```



Both left-hand panels should display a random cloud of points. Existence of any patterns in these plots would be an indication that our model is not adequate. For the upper-right panel, we would like to see a constant, even spread of points across the y-axis. Any systematic increase or decrease would be an indication that our variance function is not correct. For the last panel in the bottom-right corner, the ideal pattern would be to have all points line up along the line $y = x$. Departures from that pattern would be an informal indication that the link function is not correct (or that we have misspecified the linear predictor).

Since the estimated value of θ in the negative binomial model is 8.67, we know that the variance of the gamma distribution is 1/8.67. Figure 5.2 displays what the estimated density function for the random effect looks like.

We can also view the above model as an LMM. The mean of the Poisson distribution is λu, and we would like to introduce explanatory variables. Hence, using a logarithmic link function we set

$$\log(\lambda u) = \log(\lambda) + \log(u) = X\beta + v,$$

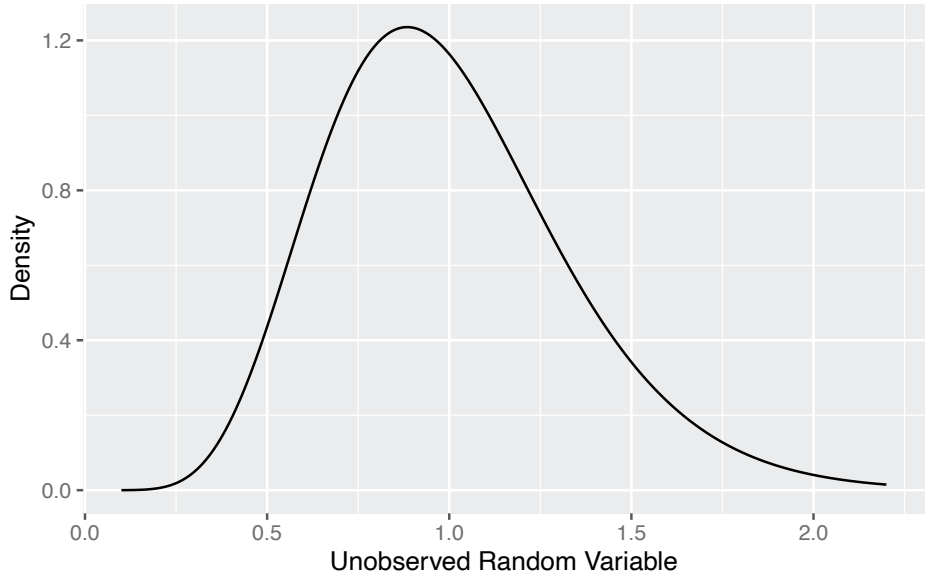
where $X\beta$ incorporates all of our fixed-effects explanatory variables and $v = log\,(u)$ is the unobserved random effect.

For the fabric data we can fit such a model by specifying the structure of

1. the mean model, and
2. the dispersion model.

For now we will keep the dispersion model as a single constant (just like we always do when we fit a GLM).

Figure 5.2. The density function for the random effect *u*.



The mean model is

```
model.mu <- DHGLMMODELING(Model = "mean",
                          Link = "log",
                          LinPred = y ~ x.lg + (1 | rf),
                          RandDist = "gamma")
```

Note that we have chosen a gamma distribution for the random effect.
The dispersion model is just a constant, so we do not specify any components:

```
model.phi <- DHGLMMODELING(Model = "dispersion")
```

We fit this model via

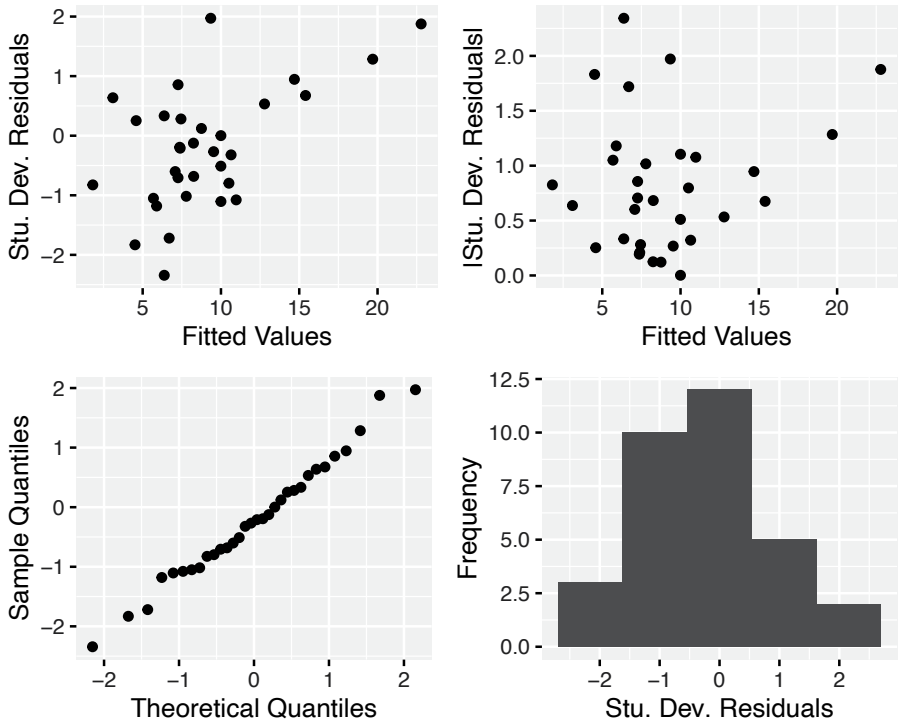
```
fab.hglm.nb <- dhglmfit(RespDist = "poisson",
                        DataMain = fabric,
                        MeanModel = model.mu,
                        DispersionModel = model.phi)
```

```
Distribution of Main Response :
                         "poisson"
[1] "Estimates from the model(mu)"
y ~ x.lg + (1 | rf)
[1] "log"
             Estimate  Std. Error  t-value
(Intercept)   -3.9195      1.4442   -2.714
x.lg           0.9624      0.2259    4.261
[1] "Estimates for logarithm of lambda=var(u_mu)"
[1] "gamma"
    Estimate  Std. Error  t-value
rf    -2.074      0.3623   -5.726
[1] "====== Likelihood Function Values and Condition AIC ======"
                                          [,1]
-2ML (-2 p_v(mu) (h))           :  175.72501
-2RL (-2 p_beta(mu),v(mu) (h)) :  179.88494
cAIC                            :  172.77209
Scaled Deviance                 :   14.28605
df                              :   14.40607
```

Here the estimated fixed effects are close to those reported for the negative binomial model. The variance of the random effect, –2.074, is reported on a logarithmic scale, and exponentiating its value we obtain 0.125682, which is of similar magnitude compared with the variance of the negative binomial model (1/8.6674 = 0.1154).

Figure 5.3 reveals that whereas our model may be adequate, there are some areas of concern. The display shows the studentized deviance residuals. The top-left panel

Figure 5.3. Diagnostic plots from a Poisson-gamma HGLM fitted to the fabric data. The top panels show that the model has some deficiencies because there are discernible patterns. The bottom panels show the expected patterns.



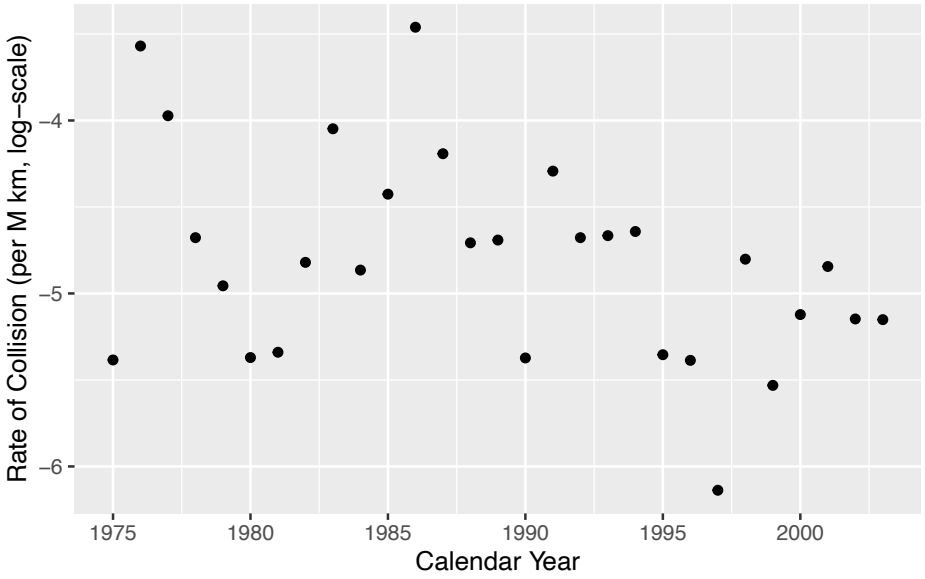
shows that for fitted values greater than about 12.5, we have a group of observations with positive and increasing residuals only. And below 7.5 there seems to be a larger number of observations with negative residuals than positive ones. The top-right panel shows the absolute value of studentized residuals versus fitted values. Ideally, there would be no underlying trend in the residuals, but the graph shows that as fitted values increase, residuals first decrease and then increase. The QQ plot, in the bottom-left panel, shows the expected pattern, and the bottom-right panel also shows a reasonable histogram for the residuals.

Train Accident

In this example we analyze a dataset from Agresti (2002) regarding the number of collisions involving British Rail passenger trains and road vehicles between 1975 and 2003. The available variables are the number of annual collisions (`y`) between trains and road vehicles, the distance traveled per year (`t`) in millions of kilometers, the number of years (`x`) since 1975, and an identification (`id`) label for each row of data. The data is available in the package `mdhglm` under the name `train`, and the first few rows are

```
data(train, package = "mdhglm")
head(train)
```

Figure 5.4. The rate of collisions, per million kilometers traveled (log-scale), between British passenger trains and road vehicles from 1975 to 2003.




```
   x   y    t  id
1  0   2  436   1
2  1  12  426   2
3  2   8  425   3
4  3   4  430   4
5  4   3  426   5
6  5   2  430   6
```

Lee et al. (2020) also analyzed the data, and we follow their discussion closely.

We are interested in understanding the rate of accidents per million kilometers traveled. A scatterplot (not displayed here) shows a nonlinear decreasing trend for the rate of collisions as time increases, but a logarithmic transformation of the response variable shows (see Figure 5.4) a decreasing linear trend with substantial variability around it.

A Poisson GLM might be our first choice for modeling the rate, but again such a model does not fit the data adequately. Overdispersion is clearly present.

Exercise 5.3 Fit a Poisson model to the rate of collisions and show that the fit is not adequate by plotting the absolute value of the quantile residuals against fitted values.

Solution 5.3 To fit a Poisson model with a logarithmic link function to the rate of collisions, we would specify the following model:

$$\log\left(\mathbb{E}\left[\frac{y}{t}\right]\right) = \beta_0 + \beta_1 x.$$

This model can be rewritten as

$$\log\left(\mathbb{E}\left[y\right]\right) = \beta_0 + \beta_1 x + \log\left(t\right),$$

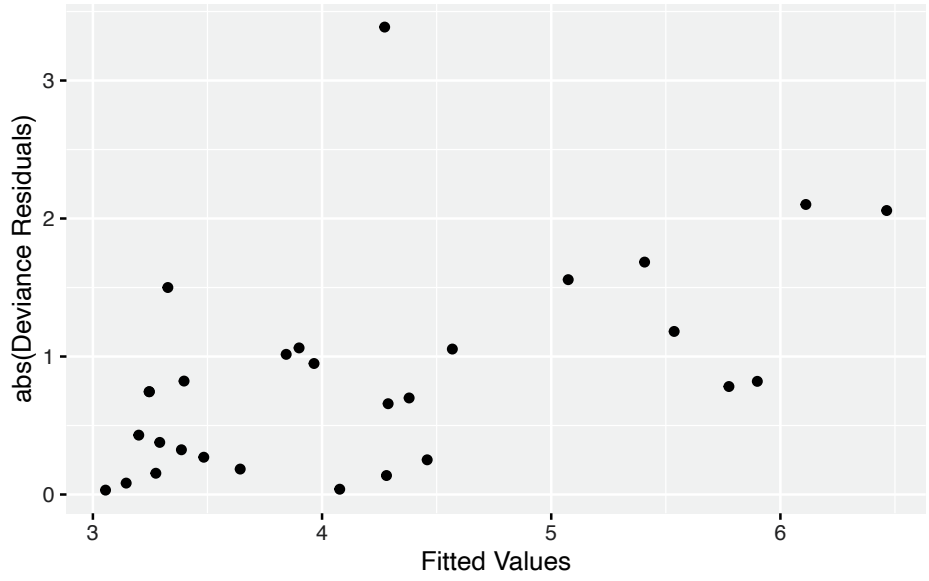
where the last term, $\log(t)$, is an offset term for the amount of exposure. The code to fit the above model is

```
train.poi <- glm(y ~ x + offset(log(t)),
                 data = train,
                 family = poisson(link = "log"))
summary(train.poi)
```

```
Call:
glm(formula = y ~ x + offset(log(t)), family = poisson(link = "log"),
    data = train)

Coefficients:
             Estimate  Std. Error  z value  Pr(>|z|)
(Intercept)  -4.21142     0.15892   -26.50   < 2e-16  ***
x            -0.03292     0.01076    -3.06   0.00222  **
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 47.376 on 28 degrees of freedom
Residual deviance: 37.853 on 27 degrees of freedom
AIC: 133.52

Number of Fisher Scoring iterations: 5
```

Overdispersion is likely since the residual deviance is larger than the degrees of freedom (the mean deviance estimator of the dispersion parameter is equal to 37.853/27 = 1.402). The plot of absolute value quantile residuals shows a clear increasing trend as fitted values increase. This tells us that the variance function we have selected (in this case it is linear because we are using the Poisson distribution) is not increasing fast enough, and so our assumption that the number of collisions is Poisson distributed is not correct.

```
ggplot(data = tibble(mu = predict(train.poi, type = "response"),
                     rD = resid(train.poi, type = "deviance")),
       mapping = aes(x = (mu),
                     y = abs(rD))) +
  geom_point() +
  labs(x = "Fitted Values",
       y = "abs(Deviance Residuals)")
```



Fitting a Poisson-gamma HGLM should be our next choice. We will model the dispersion parameter as a constant. The mean will include a random effect with a gamma distribution, and because our response variable is a rate and we are using a log-link function, we will include an offset in our model.

```
model.mu <- DHGLMMODELING(Model = "mean",
                          Link = "log",
                          LinPred = y ~ x + (1 | id),
                          Offset = log(train$t),
                          RandDist = "gamma")
model.phi <- DHGLMMODELING(Model = "dispersion")

train.hglm <- dhglmfit(RespDist = "poisson",
                       DataMain = train,
                       MeanModel = model.mu,
                       DispersionModel = model.phi)
```

```
Distribution of Main Response :
                        "poisson"
[1] "Estimates from the model(mu)"
y ~ x + (1 | id)
[1] "log"
             Estimate  Std. Error  t-value
(Intercept)  -4.13359     0.22304  -18.533
x            -0.03633     0.01452   -2.502
[1] "Estimates for logarithm of lambda=var(u_mu)"
[1] "gamma"
```

```
    Estimate  Std. Error  t-value
id    -1.752      0.4235   -4.137
[1] "====== Likelihood Function Values and Condition AIC ======"
                                              [,1]
-2ML (-2 p_v(mu) (h))           :  127.79514
-2RL (-2 p_beta(mu),v(mu) (h))  :  136.82822
cAIC                            :  129.86560
Scaled Deviance                 :   11.31993
df                              :   15.56043
```

The fixed effects are both significant at the 5% level. The coefficient for year is negative and shows that as each year goes by we can expect the number of collisions to decrease by about 3.5%. The variance of the random effect (on a log-scale) is –1.752 with a standard error equal to 0.423, and so the variance is statistically different from zero. The density for our random effect is shown in Figure 5.5.

Figure 5.6 displays QQ plots for the Poisson and the Poisson-gamma HGLM models. Note that the Poisson model shows that the distribution of the studentized deviance residuals has a fatter tail than the normal distribution. The two points in the upper-right corner are too large, whereas for the Poisson-gamma HGLM model those two points are much closer to the theoretical line.

Diabetes Progression

For this example, we use a dataset on diabetes patients to illustrate the fitting of a joint GLM. The data, on 442 diabetic patients, was analyzed in Efron et al. (2004) and Antoniadis et al. (2016). Ten baseline variables for the patients were recorded, and

Figure 5.5. Estimated density function for train dataset along with the estimated random effects (points on the *x*-axis).

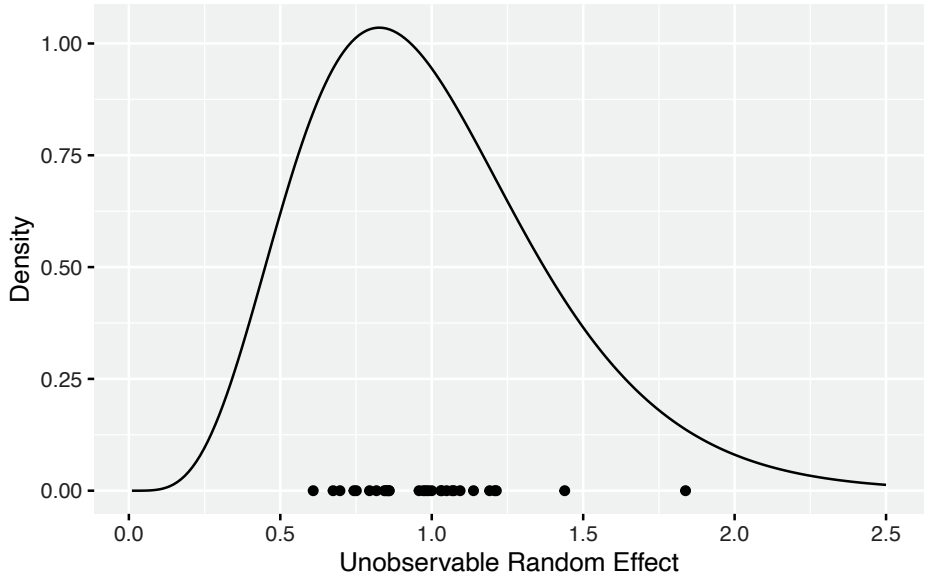
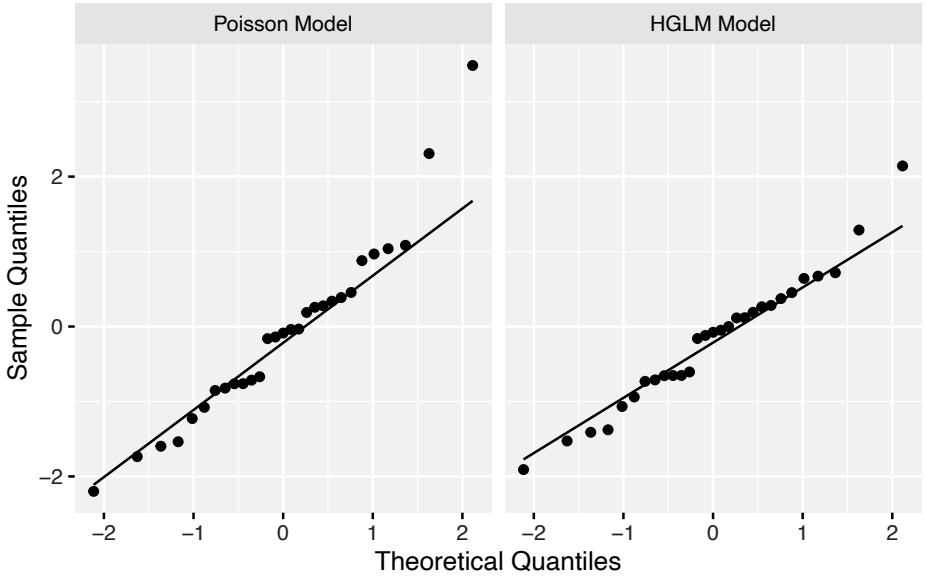


Figure 5.6. QQ plot for the studentized deviance residuals from the Poisson and the Poisson-gamma (HGLM) models.



a year later a measure of disease progression was also collected. A model was sought to predict disease progression based on the baseline variables of age, sex, body mass index, average blood pressure, and six blood serum measurements. Table 5.1 displays the first six rows of the data as shown in Table 1 from Efron et al. (2004). The data is available in the **R** package `lars` under the name `diabetes`. Note that the explanatory variables in the dataset `diabetes` have been scaled to have mean zero and unit variance, but Table 5.1 shows the unscaled values for the first six rows.

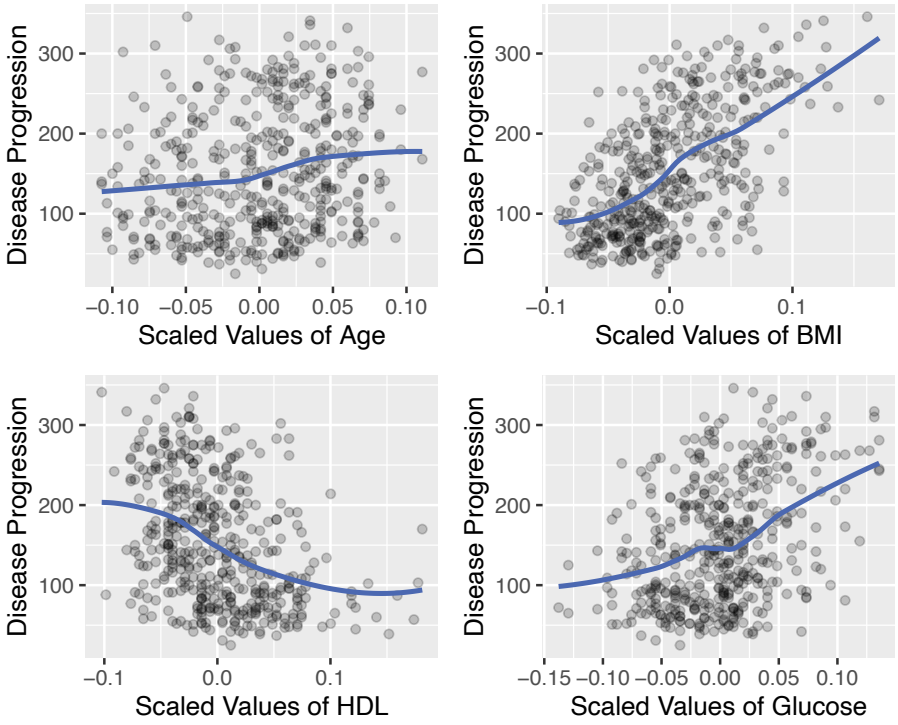
Figure 5.7 displays four exploratory graphs of the response variable, disease progression, versus some explanatory variables. Body mass index (BMI) and glucose

Table 5.1. First six rows of the unscaled diabetes data.

					Serum Measurements						Response
Patient	age	sex	bmi	abp	tc	ldl	hdl	tch	ltg	glu	y
1	59	2	32.1	101	157	93.2	38	4	4.9	87	151
2	48	1	21.6	87	183	103.2	70	3	3.9	69	75
3	72	2	30.5	93	156	93.6	41	4	4.7	85	141
4	24	1	25.3	84	198	131.4	40	5	4.9	89	206
5	50	1	23.0	101	192	125.4	52	4	4.3	80	135
6	23	1	22.6	89	139	64.8	61	2	4.2	68	97

Source: Table 1 in Efron et al. (2004).

Figure 5.7. Exploratory graphs of disease progression versus explanatory variables age, body mass index (BMI), high density lipoprotein cholesterol (HDL), and glucose. Scatterplot smooth lines have been added to aid in detecting the overall pattern. Note that in several of the panels, the variance in disease progression is not constant across the values of the explanatory variables.

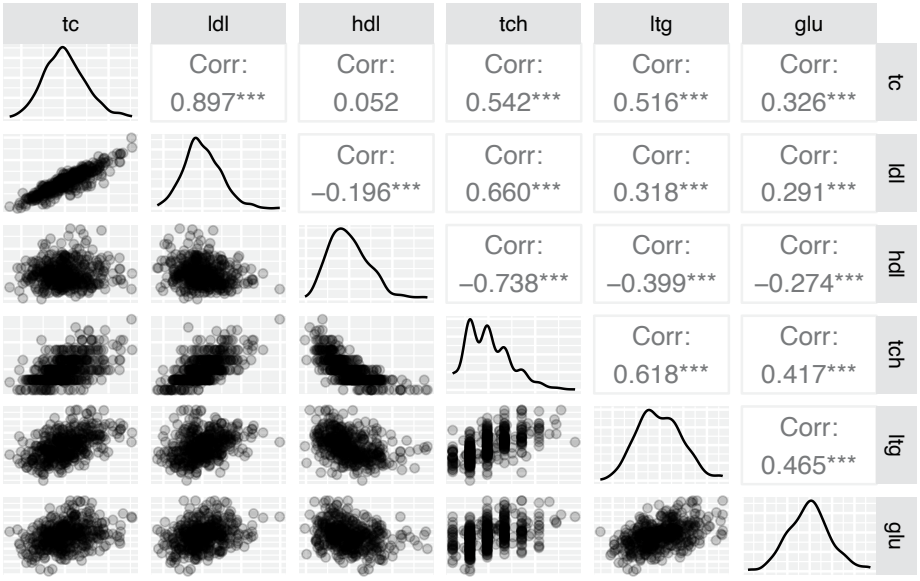


level show a strong relationship to the response, whereas age and high density lipoprotein cholesterol (HDL) show a weaker relationship. We can also see that the variability in the response is not constant across the range of values in the explanatory variables.

Some of the blood serum measurements (six variables) may be correlated to each other. Figure 5.8 displays a scatterplot matrix of these measurements where we can see that `tc` is highly positively linearly correlated with `ldl` (positions (2, 1) and (1, 2) in the plot matrix). And variable `tch` is highly negatively linearly correlated with `hdl` (positions (4, 3) and (3, 4) in the plot matrix). Variable `ldl` is also linearly correlated with `tch`, as is `ltg` with `tch`.

Based on the observations from Figure 5.7 and Figure 5.8, we suspect that some of the variables will not be significant in predicting the mean response and some will help us model the variance of the response. Hence, we would like to fit a joint GLM model—that is, we want to have a GLM for the response and also a GLM for the dispersion parameter. We specify such a structure as follows:

Figure 5.8. Scatterplot matrix for the blood serum variables. The diagonal entries show nonparametric estimates of the density function for each variable. The upper triangular entries are the pairwise linear correlation coefficients, and the bottom triangular entries are the pairwise scatterplots for the variables.




```
model.mu <- DHGLMMODELING(Model = "mean",
                          Link = "identity",
                          LinPred = y ~ age + sex + bmi +
                            abp + tc + ldl + hdl + tch +
                            ltg + glu)

model.phi <- DHGLMMODELING(Model = "dispersion",
                           Link = "log",
                           LinPred = y ~ age + sex + bmi +
                             abp + tc + ldl + hdl + tch +
                             ltg + glu)
```

Assuming that the response variable, disease progression, is adequately represented as a normal distribution we can fit the joint model via

```
diab.model <- dhglmfit(RespDist = "gaussian",
                       DataMain = diab,
                       MeanModel = model.mu,
                       DispersionModel = model.phi)
```

```
Distribution of Main Response :
                        "gaussian"
[1] "Estimates from the model(mu)"
y ~ age + sex + bmi + abp + tc + ldl + hdl + tch + ltg + glu
[1] "identity"
                          Std.
             Estimate     Error    t-value      p_val         LL      UL
(Intercept)   151.721     2.583  58.732782  0.000e+00     146.66   156.8
age            12.249    54.655   0.224114  8.227e-01     -94.87   119.4
sex          -241.175    56.167  -4.293894  1.756e-05    -351.26  -131.1
bmi           475.325    67.734   7.017552  2.258e-12     342.57   608.1
abp           344.574    62.717   5.494131  3.926e-08     221.65   467.5
tc           -555.372   341.346  -1.627006  1.037e-01   -1224.41   113.7
ldl           277.509   276.958   1.001990  3.163e-01    -265.33   820.3
hdl             1.285   167.371   0.007678  9.939e-01    -326.76   329.3
tch           150.522   142.311   1.057694  2.902e-01    -128.41   429.5
ltg           651.564   135.923   4.793637  1.638e-06     385.16   918.0
glu            60.692    60.032   1.010989  3.120e-01     -56.97   178.4
[1] "Estimates from the model(phi)"
y ~ age + sex + bmi + abp + tc + ldl + hdl + tch + ltg + glu
[1] "log"
             Estimate  Std. Error   t-value
(Intercept)     7.905     0.06813  116.0213
age            -2.710     1.58187   -1.7130
sex            -4.613     1.62120   -2.8454
bmi             1.861     1.76060    1.0571
abp             4.357     1.73308    2.5142
tc            -15.799    11.28435   -1.4001
ldl            16.515     9.20588    1.7939
hdl            -3.058     5.73781   -0.5329
tch            -6.654     4.31590   -1.5417
ltg             8.042     4.62341    1.7393
glu             2.310     1.74794    1.3214
[1] "====== Likelihood Function Values and Condition AIC ======"
                             [,1]
-2ML (-2 h)           : 4737.276
-2RL (-2 p_beta (h)) : 4629.960
cAIC                  : 4759.276
Scaled Deviance       :  431.000
df                    :  431.000
```

The top section of the output gives the estimated coefficients for the model of the response variable, and we can see that variables `sex`, `bmi`, `abp`, and `ltg` are significant at the 5% level. The bottom section shows the estimated coefficients for the dispersion model. Here the coefficients for `sex` and average blood pressure `abp` are significant.

Reestimating the model with only the significant variables gives us the following estimated coefficients:

```
Distribution of Main Response :
                          "gaussian"
[1] "Estimates from the model(mu)"
y ~ sex + bmi + abp + ltg
[1] "identity"
                          Std.
             Estimate   Error  t-value     p_val      LL      UL
(Intercept)     152.2   2.641   57.638 0.000e+00   147.0  157.39
sex            -157.0  56.783   -2.765 5.699e-03  -268.3  -45.69
bmi             585.8  64.444    9.090 9.929e-20   459.5  712.09
abp             312.6  64.841    4.821 1.426e-06   185.5  439.71
ltg             551.0  63.952    8.616 6.930e-18   425.7  676.36
[1] "Estimates from the model(phi)"
y ~ sex + abp
[1] "log"
             Estimate  Std. Error  t-value
(Intercept)     8.020     0.06765  118.543
sex            -3.031     1.46589   -2.068
abp             2.880     1.46782    1.962
[1] "====== Likelihood Function Values and Condition AIC ======"
                              [,1]
-2ML (-2 h)           : 4793.995
-2RL (-2 p_beta (h))  : 4750.245
cAIC                  : 4803.995
Scaled Deviance       :  437.000
df                    :  437.000
```

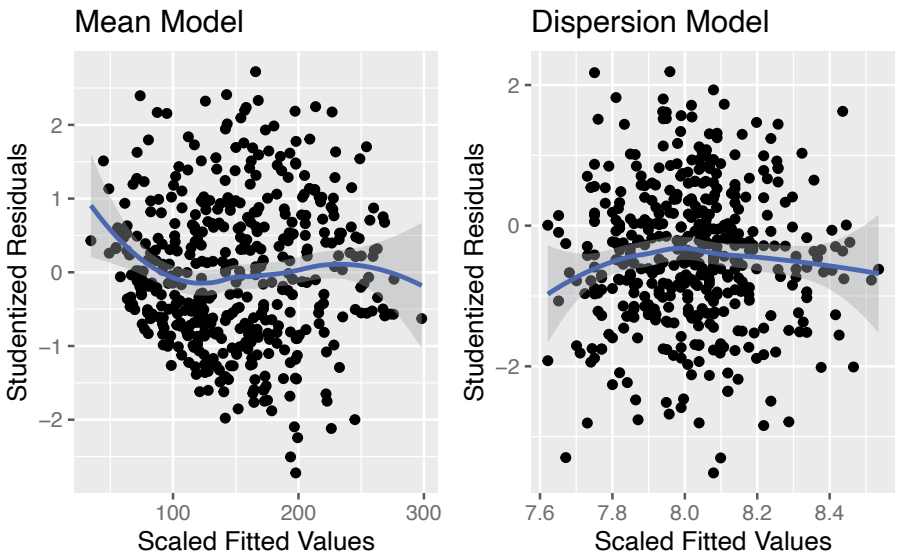
Figure 5.9 shows the studentized deviance residuals against the fitted values for both the mean and dispersion models. For the mean model, the overall shape of the points looks random with a slight increase on the lower end of the fitted values. For the dispersion model, we have slight curvature of the residuals, but it is minimal.

5.4. Summary

In this chapter, we introduced a class of hierarchical generalized linear mixed models (Lee et al. 2021) that extend the GLM by allowing random effects with normal and non-normal distributions and modeling the dispersion parameter via explanatory variables with both fixed and random effects.

We can think of these models as a pair of {mean, dispersion}-models. The standard GLM would be specified as {GLM(μ), constant}, meaning that we have a GLM for the mean of the response variable and a constant dispersion model.

Figure 5.9. Fitted values versus residuals for the mean and dispersion models of the diabetes data.



Other models in the crude diagram

```
LM ---> LMM ---> GLMM ----------> HGLM ---> DHGLM
    \          /                 /
     \        /                 /
      \> GLM ---> Joint GLM /
```

can be specified as follows:

1. **Linear mixed model (LMM).** The mean model would be a hierarchical GLM with normally distributed random effects and a Gaussian distribution for the response, together with the identity link function.
2. **Joint GLM (JGLM).** Both the mean model and the dispersion model are GLMs, and the models are interlinked.
3. **Generalized linear mixed model (GLMM).** The dispersion model is constant. The model for the mean response is a GLM with normally distributed random effects.
4. **Hierarchical generalized linear model (HGLM).** Both the mean and dispersion are modeled. The mean model is a model with random effects that are not restricted to being normally distributed. The dispersion parameter is modeled via a GLM with fixed effects only.
5. **Double hierarchical generalized linear model (DHGLM).** This extends the HGLM model by allowing random effects in the model for the dispersion parameter and allowing the modeling of the variance of the random effects via explanatory variables.

We revisited the Hachemeister dataset to show how the same model (essentially) can be fitted to the data based on the new class of HGLMs. We also presented two new

examples with gamma random effects: fabric fault data and train collisions with road vehicles. For both of those examples, the response variable was a count for which we used a Poisson distribution. But the Poisson model was not adequate for the data because of overdispersion. Hence, we introduced a gamma random effect yielding the negative binomial distribution.

In the final example, we analyzed a dataset quantifying the disease progression of diabetic patients. Here, after noticing that the variance of the response variable was not constant, we decided to introduce explanatory variables to model it. Therefore, we fitted a joint GLM to the data where we specified a linear predictor for the mean disease progression and also introduced another linear predictor for the dispersion model.

In the next chapter, we present several examples that make use of random effects for categorical variables that have a large number of levels. This will bring us back to incorporating credibility into our modeling.

6. Applications

6.1. Massachusetts Auto Bodily Injury Claims

For this example, we use a dataset of automobile bodily injury claims from the Commonwealth of Massachusetts. This data has been used to illustrate several different types of analyses, such as

- modeling hidden exposures (Rempala and Derrig 2005),
- credibility using copulas (Frees and Wang 2005), and
- multivariate credibility (Frees 2003).

The data—available in the **R** package `CASdatasets` (Dutang and Charpentier 2020) under the name `usmassBI2`—is longitudinal and describes the claims experience for 29 randomly selected towns (out of more than 300) in Massachusetts for the years 1993 to 1998. The variables available are `TOWNCODE`, `YEAR`, `AC` (average claims per unit of exposure), `PCI` (per capita income of the town), and `PPSM` (population per square mile).

As described in Frees (2003) and Frees and Wang (2005), the average claim amounts have already been restated in 1991 dollars using the Consumer Price Index (CPI) in order to mitigate any time trends due to inflation. This data has also been analyzed in Chapter 15 of Charpentier (2015), and we follow that discussion.

Data Exploration

Table 6.1 displays the descriptive statistics for average claim size by calendar year. Note that the means and medians look reasonably stable across calendar years. In addition, the standard deviation seems to hover around 35, and the maximums and minimums do not seem to fluctuate heavily; therefore, it seems like the distributions are stable across years.

In Figure 6.1 we have a multiple time series plot (a.k.a. a spaghetti plot) where each line represents the observations, across time, for one town. Two towns have been highlighted: 35 and 53. Town code 35 has large average claims in the first couple of years, and town code 53 has some of the lowest claims across all years. Figure 6.2 displays average claim cost against per capita income and population per square mile. Again, towns 35 and 53 are highlighted: note that town 53 is sparsely populated and has a high per capita income, whereas town 35 has the highest population density and fairly large average claim costs. Also, note that for each town the per capita income and population per square mile do not fluctuate much by year, but the average claims do.

Table 6.1. Descriptive statistics for average claims per unit of exposure for a random sample of 29 towns in Massachusetts. Dollar amounts have been restated to 1991 using the CPI.

	Average Claim Amount					
	1993	1994	1995	1996	1997	1998
Mean	133.00	129.03	143.38	141.17	142.94	134.37
Median	131.57	131.45	138.76	149.00	144.73	131.96
Std. deviation	31.59	32.63	38.28	39.28	36.22	32.85
Minimum	80.03	42.74	61.04	66.20	61.68	74.89
Maximum	212.46	209.52	238.22	201.99	248.75	191.05

Based on this observation, when modeling the average claim cost, an intercept for each town would make sense.

Figure 6.3 displays the histogram of average claim size across all towns and years along with a nonparametric estimate of the density (solid line) and a normal distribution (dashed line) matching the first two moments. Overall, the normal distribution fits this data well.

Figure 6.1. Multiple time series plot of average claims per unit of exposure. Each time series corresponds to one of the 29 towns in the data. The red dots joined by pink lines correspond to town code 35, which has some of the highest average claims during the first couple of years. The blue points joined by light blue lines correspond to town code 53, which has some of the lowest average claims.

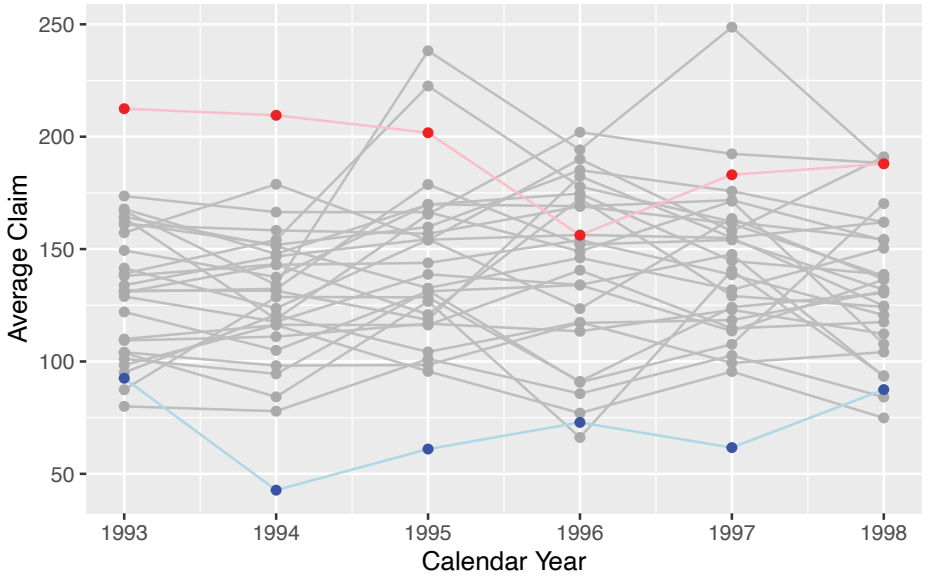


Figure 6.2. Average claim sizes by per capita income (in thousands of dollars) and population per square mile (in log base 10 scale). The red points correspond to town code 35, and the blue points correspond to town code 53.

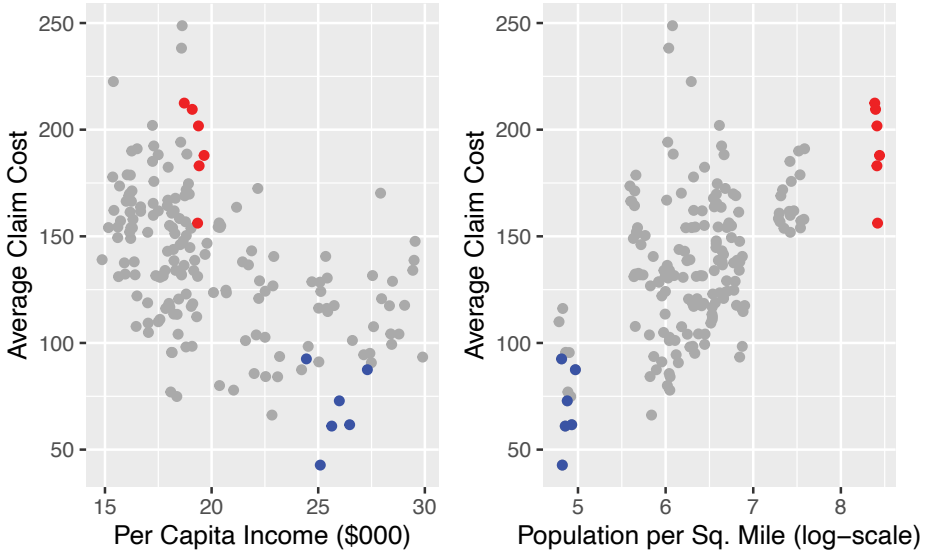
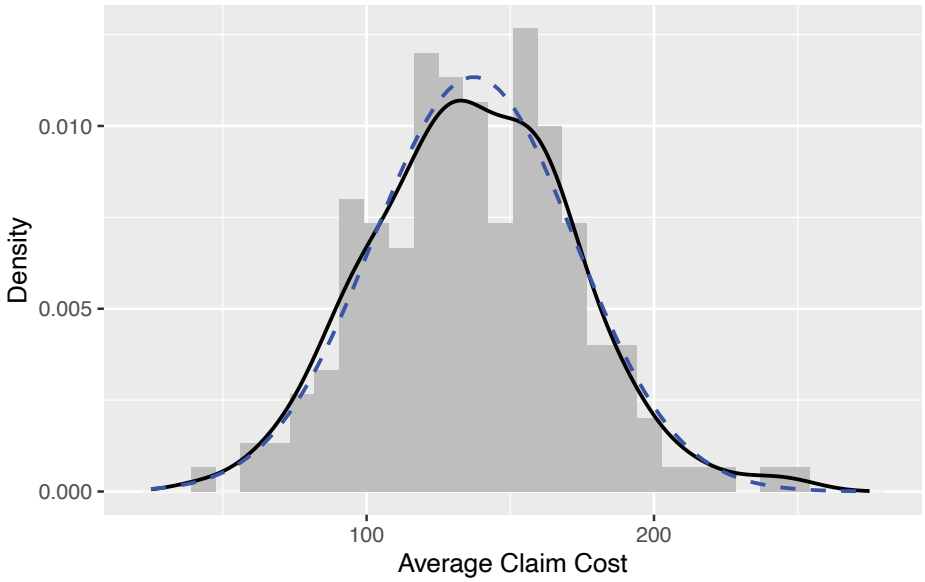


Figure 6.3. Histogram of average claim costs along with a nonparametric estimate of the density function (solid line) and a normal density function (dashed line) chosen to match the empirical mean and standard deviation across all towns and all years.



Modeling Average Claim Size

From our exploratory analysis we can start our modeling of average claim cost by using a normal distribution for the response variable and applying the following transformations to the explanatory variables:

1. Shift the origin for YEAR to 1992 (that is, center this variable at 1992).
2. Scale the per capita income to measure it in thousands of dollars.
3. Compute the logarithm of population per square mile.

Also, we will take calendar years 1993 to 1997 to train our models and keep 1998 for validation purposes.

```
usmassBI2 <- usmassBI2 |>
  mutate(YR = YEAR - 1992,
         lnPPSM = log(PPSM),
         PCI.k = PCI / 1000)

db.train <- usmassBI2 |>
  filter(YEAR < 1998)

db.test <- usmassBI2 |>
  filter(YEAR == 1998)
```

Complete Pooling

First, we fit a model ignoring the TOWNCODE variable—thus we are pooling all of our data together, implicitly assuming that all the towns in Massachusetts form a single homogeneous group. This is clearly not a reasonable assumption, but it is a good starting point.

```
bi.all <- lm(AC ~ PCI.k + lnPPSM + YR,
             data = db.train)
summary(bi.all)
```

```
Call:
lm(formula = AC ~ PCI.k + lnPPSM + YR, data = db.train)

Residuals:
    Min       1Q   Median       3Q      Max
-51.661  -16.846   -0.419   12.680  103.850

Coefficients:
             Estimate  Std. Error  t value  Pr(>|t|)
(Intercept)   68.8695     23.1298    2.978   0.00342  **
PCI.k         -4.2410      0.5604   -7.568  4.47e-12  ***
lnPPSM        22.3442      2.9603    7.548  5.00e-12  ***
YR             3.8353      1.5324    2.503   0.01346  *
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Residual standard error: 25.99 on 141 degrees of freedom
Multiple R-squared: 0.4812, Adjusted R-squared: 0.4701
F-statistic: 43.59 on 3 and 141 DF, p-value: < 2.2e-16
```

The estimated value of the intercept, 68.87, is the average claim cost in 1992 for a town that has a population density of one person per square mile and a per capita income of zero. Such a town does not exist in Massachusetts. The average per capita income (in thousands) and the logarithm of the population per square mile across all towns in our data are 20.06 and 6.38, respectively. Therefore, using our current model, we would estimate the expected claim costs in 1993 to be

$$68.87 - 4.24 \cdot 20.06 + 22.34 \cdot 6.38 + 3.84 \cdot 1 = 130.18,$$

close to the middle of the data for 1993 shown in Figure 6.1.

The coefficient for calendar year of 3.84 tells us that as we move from one year to the next, the average claim cost will increase by this dollar amount. Keeping in mind that the data had already been adjusted to account for inflation, we must attribute this increase to other sources. As per capita income increases by $1,000, we see a decline in the average claim cost of 4.24. And if we had a 10% increase in population density, the average claim cost would increase by about 2.13.

Figure 6.4 and Figure 6.5 show diagnostic plots for the model `bi.all`. All five plots show that the model seems adequate. The left-hand panel of Figure 6.4 shows

Figure 6.4. Diagnostic plots for model `bi.all`. Left-hand panel shows standardized residuals versus fitted values, and the right-hand panel is the absolute value of the standardized residuals against the fitted values. The blue line in each panel is a scatterplot smooth line showing the overall trend of the points.

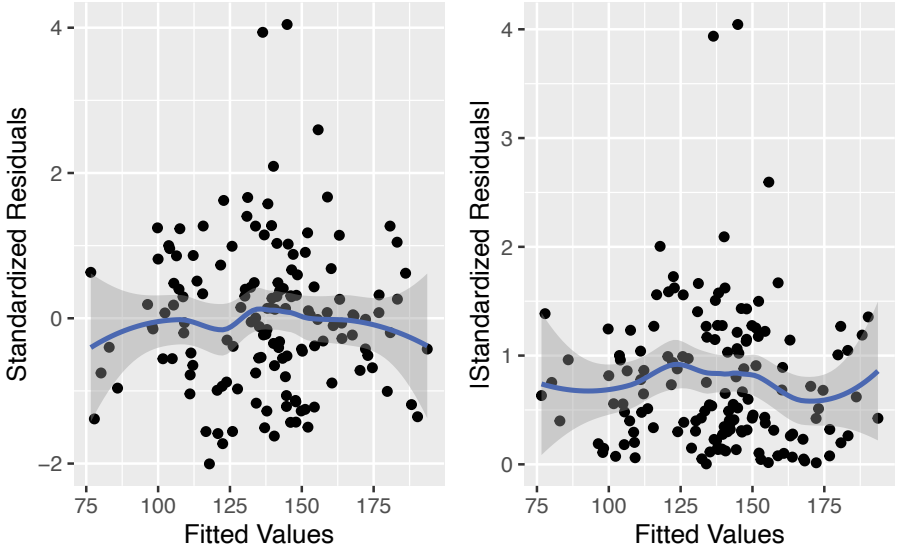
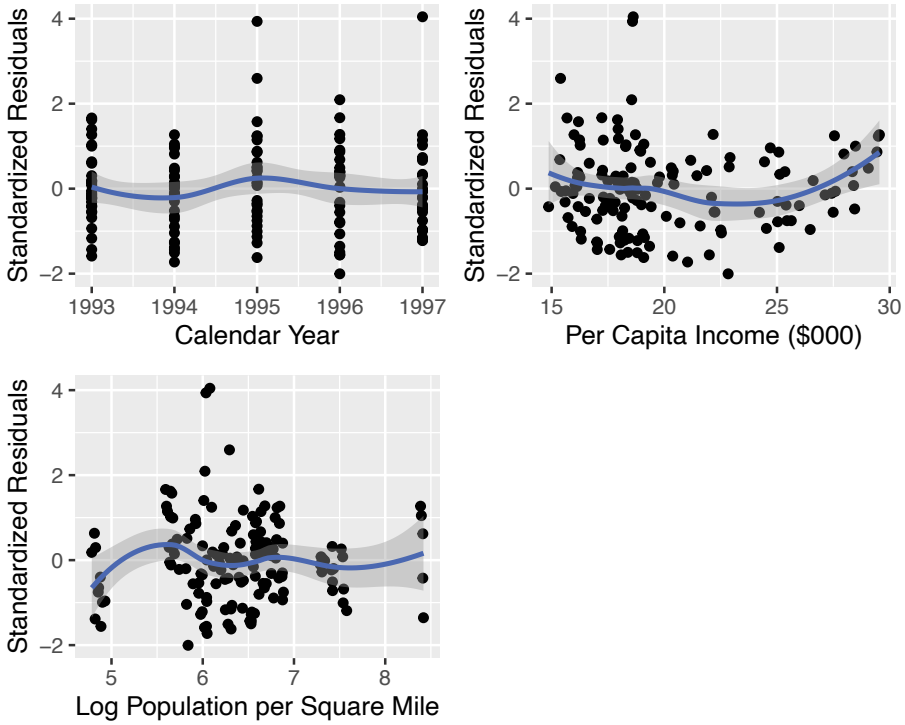


Figure 6.5. Diagnostic plots for model `bi.all`. Each panel shows the standardized residuals against a predictor variable. The blue line in each panel shows the overall trend of points.

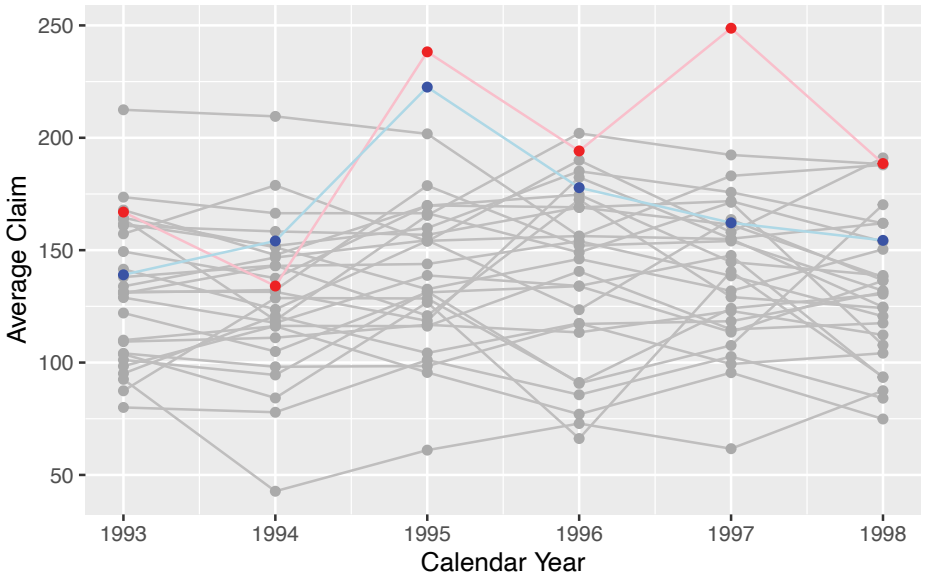


the expected pattern of a random cloud of points centered about $y = 0$. There are four observations with residuals greater than 2. Three of these observations come from `TOWNCODE` 45 and one from `TOWNCODE` 16. Figure 6.6 reproduces Figure 6.1 but highlights town codes 45 and 16. Looking at calendar year 1995 in Figure 6.6, town code 45 corresponds to the red-colored points connected by pink lines. Starting in 1995, this town has some of the highest average claim costs of all towns. Tracing the line for town code 16 (blue-colored points and connecting lines), we can see that whereas in 1995 the town had a large average claim cost, in all other years the average claim cost remained stable.

The right-hand panel of Figure 6.4 shows a fairly uniform spread of points as the fitted values increase; thus, our assumption that the distribution of the response variable, average claim size, is normally distributed seems appropriate. Note that for the largest fitted values we see an upward trend, letting us know that for these values we have more variability in our data than our model provides.

In Figure 6.5, we have plotted the three explanatory variables against the standardized residuals. For calendar year we see no meaningful patterns. For per capita income, we observe that in the range from 20,000 to 25,000, the model tends to overpredict, and above 25,000, the model systematically underpredicts. While the overall pattern is flat for the logarithm of population per square mile, there are a few isolated places where the model tends to overpredict (below 5) and underpredict (slightly above 5.5).

Figure 6.6. Multiple time series plot of average claim per unit of exposure. Each time series corresponds to one of the 29 towns in the data. Here we highlight the two towns (`TOWNCODE` 45 and 16) that have the four highest residuals. `TOWNCODE` 45 is in red with pink connecting lines, and `TOWNCODE` 16 is in blue with light blue connecting lines.



No Pooling

In the previous section, we pooled all of our data, assuming that all 29 sampled towns in Massachusetts would create a homogeneous group, and fitted the linear model

$$\mathbb{E}\left[\mathrm{AC}\right] = \beta_0 + \beta_1 \cdot \mathrm{PCI.k} + \beta_2 \cdot \log\left(\mathrm{PPSM}\right) + \beta_3 \cdot \mathrm{YR.}$$

We can take a diametrically opposite stance and assume that no two towns are similar in any way. With this view, we would fit the above linear model to each town separately, thus creating 29 linear models. Some of them fit well, while others do not. For example, we can collect for each model the R^2 measure as an indicator of model fit (not advocating this is a good measure), yielding

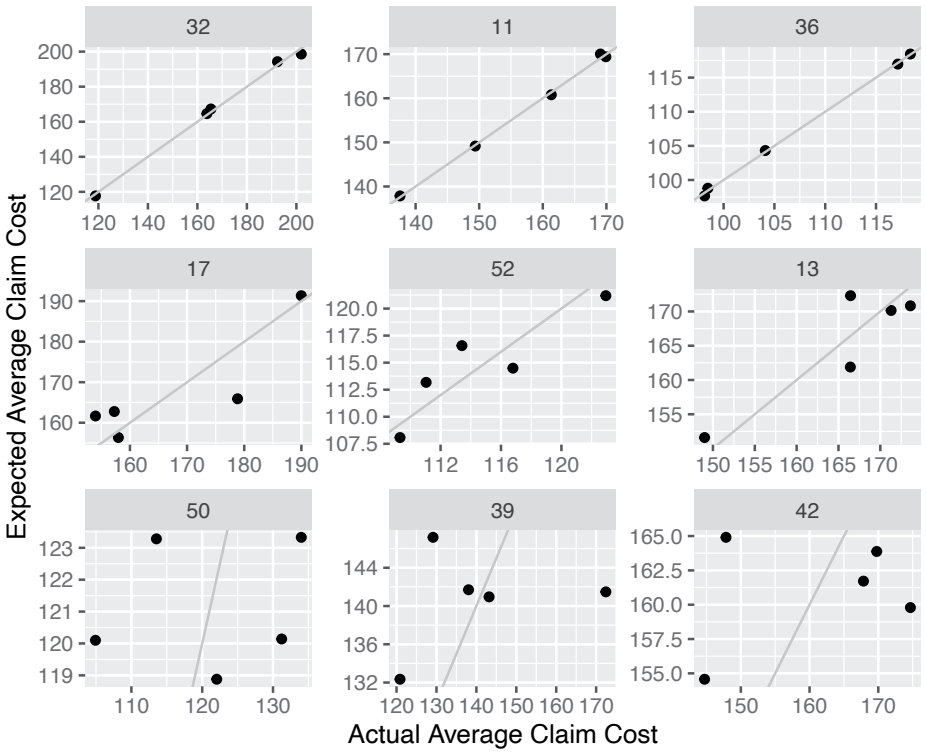
```
   50     39     42     40     31     41     45     33     44
0.028  0.073  0.089  0.424  0.451  0.534  0.552  0.580  0.600
   14     53     30     43     17     52     13     51     12
0.645  0.645  0.720  0.737  0.742  0.790  0.811  0.844  0.857
   38     16     35     37     21     10     34     15     32
0.902  0.915  0.936  0.952  0.959  0.979  0.988  0.993  0.995
   11     36
0.998  0.999
```

For town codes 50, 39, and 42, the adjusted R^2 measure is extremely low (less than 9%), and for town codes 17, 52, and 13 (middle of the list), the R^2 measure is 74.2%, 79.0%, and 81.1%, respectively. For town codes 32, 11, and 36, the measure is above 99.5%. Figure 6.7 shows an actual-versus-expected plot for these nine towns arranged from low R^2 values to high. The gray line represents the line of perfect fit, that is, $y = x$ in each panel.

Clearly for the top three panels in Figure 6.7 the models accurately predict the actual average claims, and we would feel confident in using them to predict the claims experience in the next calendar year. But do we feel similarly about the bottom three models? For town code 50, actual experience in the past five years ranged from about 105 to 135—quite volatile. The model's range of values is from about 119 to 124—a very small range. The probability that our prediction (whatever it might be) reflects actual experience would be quite low.

Thus we have that some towns are highly credible in their experience while others are not. Based on Figure 6.1, we should include a town-specific intercept in our model.

Figure 6.7. Actual versus predicted average claim costs from the regression lines fitted to each town individually. Towns 32, 11, and 36 have the highest R^2 values, and towns 50, 39, and 42 have the lowest values. Towns 17, 52, and 13 are in the middle when R^2 measures are sorted. The panels are arranged from the lowest R^2 in the bottom-left corner to the highest R^2 value in the top-right corner. The gray line in each panel represents the line of perfect fit (y = x).



Fixed-Effects Model

Consider incorporating a town-specific intercept into the regression model. Thus we want to fit the model

$$\mathbb{E}\left[\mathrm{AC}\right] = \alpha_i + \beta_1 \cdot \mathrm{PCI.k} + \beta_2 \cdot \log\left(\mathrm{PPSM}\right) + \beta_3 \cdot \mathrm{YR},$$

where α_i represents the intercept for town i. We can accomplish this by treating TOWNCODE as a categorical variable in our model.

```
bi.fixed <- lm(AC ~ TOWNCODE + PCI.k + lnPPSM + YR,
               data = db.train)
summary(bi.fixed)
```

```
Call:
lm(formula = AC ~ TOWNCODE + PCI.k + lnPPSM + YR, data = db.train)

Residuals:
    Min      1Q  Median      3Q     Max
-55.621  -8.911   0.276   9.058  50.129

Coefficients:
             Estimate  Std. Error  t value  Pr(>|t|)
(Intercept)  1457.753     795.425    1.833    0.0695  .
TOWNCODE11   -101.240      61.491   -1.646    0.1025
TOWNCODE12   -101.223      92.990   -1.089    0.2787
TOWNCODE13   -299.086     182.682   -1.637    0.1044
TOWNCODE14   -215.677     117.945   -1.829    0.0701  .
TOWNCODE15     21.421      19.879    1.078    0.2835
TOWNCODE16   -174.125     112.682   -1.545    0.1251
TOWNCODE17     42.113      32.229    1.307    0.1940
TOWNCODE21   -141.966      76.612   -1.853    0.0665  .
TOWNCODE30   -307.519     178.349   -1.724    0.0874  .
TOWNCODE31   -205.487     117.761   -1.745    0.0837  .
TOWNCODE32   -123.095      78.861   -1.561    0.1213
TOWNCODE33   -121.806      65.728   -1.853    0.0665  .
TOWNCODE34   -175.532      94.975   -1.848    0.0672  .
TOWNCODE35    223.190     117.709    1.896    0.0605  .
TOWNCODE36   -234.455     108.455   -2.162    0.0327  *
TOWNCODE37   -302.709     172.261   -1.757    0.0816  .
TOWNCODE38   -283.961     149.940   -1.894    0.0608  .
TOWNCODE39   -131.217      72.121   -1.819    0.0715  .
TOWNCODE40   -313.326     157.915   -1.984    0.0497  *
TOWNCODE41   -253.651     137.237   -1.848    0.0672  .
TOWNCODE42   -131.500      78.933   -1.666    0.0985  .
```

```
TOWNCODE43   -498.092     262.041   -1.901    0.0599  .
TOWNCODE44   -113.359      66.123   -1.714    0.0892  .
TOWNCODE45   -190.694     135.632   -1.406    0.1625
TOWNCODE50   -277.511     142.634   -1.946    0.0542  .
TOWNCODE51   -293.370     135.513   -2.165    0.0325  *
TOWNCODE52   -186.622      84.174   -2.217    0.0286  *
TOWNCODE53   -518.904     261.780   -1.982    0.0499  *
PCI.k          -1.374       7.595   -0.181    0.8568
lnPPSM       -176.078     106.343   -1.656    0.1005
YR              5.990       2.602    2.302    0.0231  *
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 18.88 on 113 degrees of freedom
Multiple R-squared: 0.7808, Adjusted R-squared: 0.7206
F-statistic: 12.98 on 31 and 113 DF, p-value: < 2.2e-16
```

In our model, the intercept represents the average claims experience for TOWNCODE 10 with a zero per capita income, one person per square mile, and the variable YR set to zero (i.e., 1992). This number does not have practical significance because we know that town code 10 does not have a per capita income of zero or a population of one person per square mile. In town code 10, the per capita income is about $18,500 and the logarithm of the population per square mile is about 7.3. The other TOWNCODE coefficients measure deviations from TOWNCODE 10 with the other variables set as before.

Note that some of the town-specific coefficients are not significant, and many of them are significant at the 10% level but not at the 5% level. Per capita income and population per square mile are no longer significant, and the effect of calendar year has increased to almost 6 and is significant at the 5% level. More importantly, an estimated yearly increase of $6 is of practical significance because it represents about a 4.4% increase from the overall average claim cost over inflation (across all towns and years) of $138.

With this model we have estimated coefficients for the 29 specific towns. And while we can make inferences for them, they are just a sample of the more than 300 towns in the state, and we could not easily justify using them to make inferences about other towns in the state. In the next section, we'll treat these towns as a random sample of the larger population of towns and fit a random-effects model to this data.

Random-Effects Model

Instead of treating the intercepts α_i for each of the 29 sampled towns in Massachusetts as fixed, we can treat them as random variables and fit the following LMM:

$$\mathbb{E}\left[\mathrm{AC}\right] = \alpha_i + \beta_0 + \beta_1 \cdot \mathrm{PCI} + \beta_2 \cdot \log\left(\mathrm{PPSM}\right) + \beta_3 \cdot \mathrm{YR},$$

where α_i is a normal random variable with mean zero and variance σ_α. The index i runs through all the towns, and we fit this model as follows:

```
bi.rnd.int <- lme(AC ~ PCI.k + lnPPSM + YR,
                  data = db.train,
                  random = ~ 1 | TOWNCODE)
summary(bi.rnd.int)
```

```
Linear mixed-effects model fit by REML
  Data: db.train
       AIC       BIC    logLik
  1310.722  1328.415  -649.361

Random effects:
  Formula: ~1 | TOWNCODE
         (Intercept)  Residual
StdDev:    18.44886  19.01929

Fixed effects: AC ~ PCI.k + lnPPSM + YR
                Value  Std.Error   DF    t-value  p-value
(Intercept)  70.25708   39.65931  113   1.771516   0.0792
PCI.k        -4.19414    0.97273  113  -4.311736   0.0000
lnPPSM       21.98209    5.16832  113   4.253237   0.0000
YR            3.82988    1.14148  113   3.355195   0.0011
  Correlation:
         (Intr)   PCI.k  lnPPSM
PCI.k    -0.555
lnPPSM   -0.872   0.096
YR        0.079  -0.197  -0.082

Standardized Within-Group Residuals:
        Min           Q1         Med          Q3         Max
-2.56010534  -0.61512040  0.01209624  0.48774689  2.89511606

Number of Observations: 145
Number of Groups: 29
```

The above output tells us that the random intercepts a_i have a standard deviation equal to $\sigma_\alpha = 18.45$. This is the variability between the towns. Actuaries would call its square the *variance of the hypothetical means*, or VHM, and statisticians would call it the *between-group variability*. The residual standard deviation is equal to $\sigma = 19.02$—statisticians call this the *within-group variability*, and actuaries would say that its square is the *expected value of the process variance*, or EVPV. Note that the *between*-town and the *within*-town standard deviations are quite similar. The ratio

$$\frac{\sigma_\alpha}{\sigma_\alpha + \sigma} = \frac{18.45}{18.45 + 19.02} = 49.2\%$$

is known as the intraclass correlation. For our data this ratio is close to 50%, letting us know that the observations within a town are mildly correlated.

Note that the fixed effects for per capita income, population per square mile, and year are all statistically significant. The coefficient for calendar year is now estimated at $3.83—still a sizable increase beyond the adjustment made to the data (prior to loading it) based on the CPI.

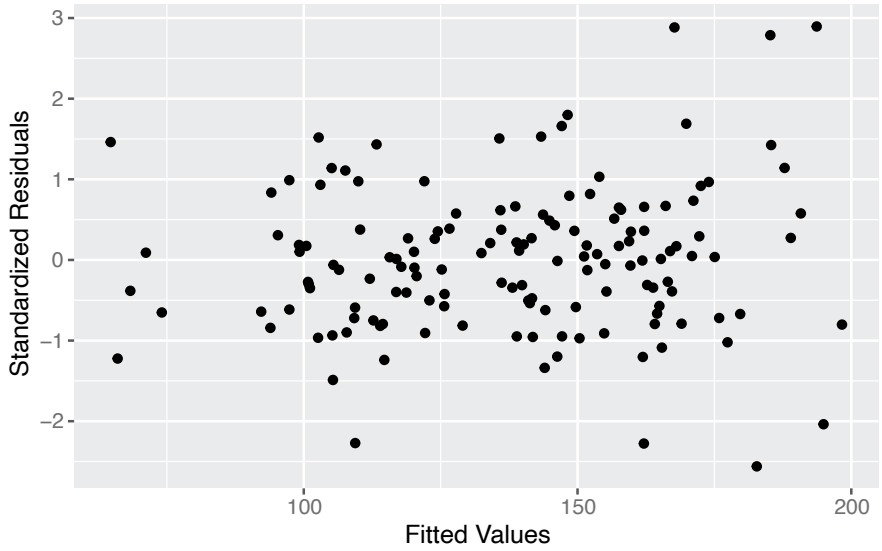
Figure 6.8 shows a diagnostic plot for the model where the *y*-axis has the standardized residuals and on the *x*-axis we have the fitted values. From this plot we can see that the assumption that our response variable has constant variance is reasonable. We do not see any fanning in or out of the residuals. There are no clear outliers in the plot, and so our assumption that the data comes from a single data-generating process (as defined by our model) also seems reasonable.

The between-town variability σ_α certainly seems large enough to be significant, but we should check. The `intervals()` function will display approximate 95% confidence intervals for fixed as well as random effects.

```
intervals(bi.rnd.int)
```

```
Approximate 95% confidence intervals

Fixed effects:
                 lower       est.       upper
(Intercept)  -8.315161  70.257082  148.829325
PCI.k        -6.121279  -4.194135   -2.266991
lnPPSM       11.742719  21.982095   32.221471
YR            1.568409   3.829885    6.091360
```

Figure 6.8. Standardized residuals versus fitted values for the random intercept model fitted to the sample of Massachusetts towns.



```
Random Effects:
Level: TOWNCODE
                    lower      est.     upper
sd((Intercept))  13.25503  18.44886  25.67781

Within-group standard error:
   lower      est.     upper
16.71330  19.01929  21.64345
```

Based on the above output, the between-town variability is clearly significant. Our current model has four fixed parameters and one random parameter even though we have 29 different towns. If we expanded our data to include more towns, this model will still have only five parameters.

Exercise 6.1 From our multiple series plot, Figure 6.1, we might suspect that different towns should have different slope coefficients for the predictor variable YR. Should we add a random component for this variable?

Solution 6.1 If we want to add a random component to the slope of YR, then we want to fit the following model:

$$\mathbb{E}[\mathrm{AC}] = \alpha_i + \beta_0 + \beta_1 \cdot \mathrm{PCI} + \beta_2 \cdot \log(\mathrm{PPSM}) + (\beta_3 + \gamma_i) \cdot \mathrm{YR},$$

where both α_i and γ_i are random variables.

We can fit that model and display approximate 95% confidence intervals for the parameters via

```
bi.rnd.slope <- lme(AC ~ PCI.k + lnPPSM + YR,
                    data = db.train,
                    random = ~ 1 + YR | TOWNCODE)
intervals(bi.rnd.slope)
```

```
Approximate 95% confidence intervals

  Fixed effects:
                lower       est.       upper
(Intercept)  -8.483553  67.376635  143.236823
PCI.k        -5.920253  -4.037949   -2.155646
lnPPSM       12.083632  21.958795   31.833958
YR            1.408192   3.795298    6.182403

  Random Effects:
    Level: TOWNCODE
                           lower        est.       upper
sd((Intercept))       6.6223297  14.3342800  31.0270845
sd(YR)                0.3672731   2.4125204  15.8472126
cor((Intercept),YR)  -0.9935728   0.4066217   0.9988532
```

```
Within-group standard error:
   lower      est.     upper
16.11193  18.64116  21.56742
```

From the output, we can see that the standard deviation for the random slope σ_γ has been estimated at 2.41 and its confidence interval does not include zero; therefore, the model suggests that a random slope for each town is important.

A five-point summary of the estimated random slopes yields

```
summary(ranef(bi.rnd.slope)[["YR"]])
```

```
    Min.   1st Qu.    Median      Mean  3rd Qu.     Max.
-2.35759  -0.82400  -0.07675  0.00000  0.32923  5.83271
```

Model Predictions

So far we have estimated four models:

- `bi.all`: complete pooling of all data,
- `bi.fixed`: fixed effects for `TOWNCODE`,
- `bi.rnd.int`: random effects for `TOWNCODE`, and
- `bi.rnd.slope`: random effects for `TOWNCODE` and `YR`.

From these models we can compute predictions for the training data as well as the validation data and compare them with the actual observations. Table 6.2 shows the following performance measures on the training data:

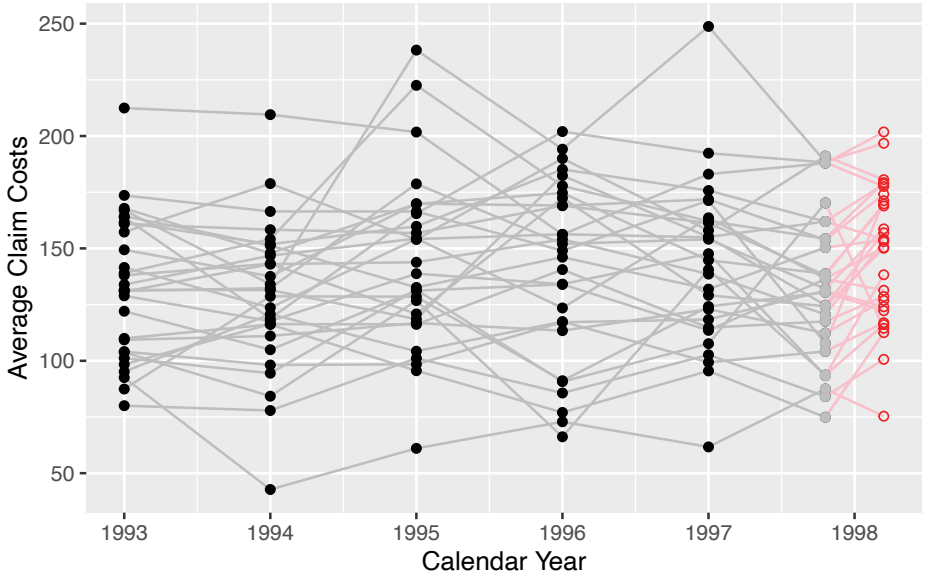
- AIC, Akaike information criterion,
- BIC, Bayesian information criterion,
- MSPE, mean squared prediction error, and
- MAPE, mean absolute prediction error.

For the validation data we use MSPE and MAPE.

Table 6.2. Comparison metrics for all four models using both the training and validation data. AIC is the Akaike information criterion, BIC is the Bayesian information criterion, MSPE is the mean squared prediction error, and MAPE is the mean absolute prediction error.

	Training Data				Validation Data	
Model	AIC	BIC	MSPE	MAPE	MSPE	MAPE
`bi.all`	1,362.21	1,377.09	657.01	19.50	769.90	23.42
`bi.fixed`	1,293.31	1,391.55	277.65	12.43	743.98	22.80
`bi.rnd.int`	1,310.72	1,328.41	298.22	13.06	675.22	21.41
`bi.rnd.slope`	1,312.77	1,336. 36	280.51	12.85	728.66	22.12

Figure 6.9. Time series plot of actual claim costs across calendar years for the 29 randomly selected Massachusetts towns (shown in black) for the training data (1993–1997). The actual 1998 experience is shown in gray along with the predicted values from the random intercept model (shown in red).



The worst model, across all measures, is `bi.all`, where we ignored the variable `TOWNCODE`. Many modelers use AIC/BIC as part of their model selection criteria. Across the above models, AIC would select the `bi.fixed` model and BIC would go with `bi.rnd.int`. Both measures of prediction error, computed on the training data, suggest that the fixed-effects model, `bi.fixed`, is the better choice. We know full well that relying on any measure of performance based on the training data may lead us astray!

We can see that both prediction error measures are larger on the validation data than on the training data and both measures have their minimum for the random intercept model, `bi.rnd.int`. Our selected model is the random intercept model:

$$\mathbb{E}[\mathrm{AC}] = (\beta_0 + \alpha_i) + \beta_1 \cdot \mathrm{PCI} + \beta_2 \cdot \log(\mathrm{PPSM}) + \beta_3 \cdot \mathrm{YR}.$$

Figure 6.9 shows the actual data from 1993 to 1998 together with the predictions for 1998 based on the training data (1993–1997). The predictions are in open red-colored circles and are joined to their actual values by a pink line.

6.2. Hospital Length of Stay

The length of stay at a hospital is a measure health organizations track, and it is important to understand some of the patient characteristics that may influence it. The following example comes from Hilbe (2007), and the data is a random sample of patients

drawn from the state of Arizona Medicare program for a single undisclosed diagnostic group. The data, medpar, is available at the book's website.

The response variable is length of stay, los, a count of the number of days a patient spent in the hospital. The following explanatory variables are available:

1. `provnum`: identifier for the medical provider
2. `hmo`: does the patient belong to a health maintenance organization (HMO)?
3. `white`: does the patient self-identify primarily as a Caucasian?
4. `type1`: was the admission to the hospital **elective**?
5. `type2`: was the admission to the hospital **urgent**?
6. `type3`: was the admission to the hospital **emergency**?
7. `age`: the age group of the patient (1 to 9)
8. `age80`: patient is older than or equal to 80
9. `died`: did the patient die at the hospital?

All variables are indicator variables (1/0) except for `provnum` and `age`. The data has 1,495 observations and 54 unique medical providers. We would like to understand how these explanatory variables could help us predict the length of stay for a newly admitted patient.

The response variable is counting the days that a patient spends in the hospital, and so perhaps we should use a Poisson distribution for the length of stay. But the Poisson distribution would not be entirely appropriate because length of stay can never be zero. What would work is the zero-truncated Poisson distribution defined as follows: we say that N is a zero-truncated Poisson random variable with parameter $\lambda > 0$ if

$$\text{Prob}\left(N = y\right) = \frac{1}{1 - e^{-\lambda}} \frac{e^{-\lambda} \lambda^{y}}{y!} \quad \text{for } y = 1, 2, \ldots$$

Thus a zero-truncated Poisson random variable is a rescaled Poisson random variable where we have removed the possibility of $N = 0$.

Exercise 6.2 Show that the zero-truncated Poisson distribution is a member of the exponential family.

Solution 6.2 To show that a zero-truncated Poisson distribution is a member of the exponential family, we have to rewrite the density function in the form

$$a\left(y, \phi\right) \exp\left[\frac{y\theta - \kappa\left(\theta\right)}{\phi}\right],$$

where ϕ is a dispersion parameter and $a(y, \phi)$ is a normalizing constant.

The mean of the distribution is given by the first derivative of $\kappa(\theta)$, and the variance function is the derivative of the mean with respect to θ, that is, $\kappa''(\theta)$.

We can rewrite the density function as follows:

$$\frac{e^{-\lambda}\lambda^{y}}{y!\left(1-e^{-\lambda}\right)}=\frac{1}{y!}\frac{\lambda^{y}}{\left(e^{\lambda}-1\right)}=\frac{1}{y!}e^{y\log(\lambda)-\log\left(e^{\lambda}-1\right)}.$$

Hence, we have $\theta=\log(\lambda)$, $\kappa(\theta)=\log(e^{e^{\theta}}-1)$, $a(y,\phi)=1/y!$, and $\phi=1$.

Therefore, the zero-truncated Poisson distribution is a member of the exponential family of distributions. The mean of the distribution is

$$\kappa'(\theta)=\frac{e^{\theta}e^{e^{\theta}}}{e^{e^{\theta}}-1}=\frac{\lambda e^{\lambda}}{e^{\lambda}-1}=\mu. \tag{6.1}$$

In an application, we would have an estimate of what the mean of the distribution might be, and so we would like to express the parameter λ in terms of the mean μ. Solving the above equation for λ in terms of the mean μ requires the use of the Lambert W function (see Appendix C), and we have that

$$\lambda=\mu+W_0\left(-\mu e^{-\mu}\right),$$

where W_0 is the principal branch.

The variance function in terms of the mean μ is

$$\kappa''(\theta)=\frac{e^{\theta}e^{e^{\theta}}\left[e^{e^{\theta}}-e^{\theta}-1\right]}{\left(e^{e^{\theta}}-1\right)^{2}}=\frac{\lambda e^{\lambda}\left[e^{\lambda}-\lambda-1\right]}{\left(e^{\lambda}-1\right)^{2}}=\mu\left[1+W_0\left(-\mu e^{-\mu}\right)\right], \tag{6.2}$$

where again W_0 is the principal branch of the Lambert W function.

Exercise 6.3 Compare the Poisson and zero-truncated Poisson distributions with means equal to 1.5, 2.5, and 3.5. Based on this information, what would you conclude in terms of the usefulness of the zero-truncated Poisson distribution?

Solution 6.3 The Poisson distribution with parameter λ has a mean equal to λ. But the zero-truncated Poisson distribution with parameter λ has a mean equal to $\lambda/(1-e^{-\lambda})$, and thus we need to find the appropriate value of λ to give us a zero-truncated Poisson distribution with the correct mean; that is, given the mean μ the correct value of λ (see previous exercise) is

$$\lambda=\mu+W_0\left(-\mu e^{-\mu}\right),$$

where W_0 is the principal branch of the Lambert W function. See Appendix C for more information on Lambert's function.

Figure 6.10 shows the probability mass functions for the Poisson and zero-truncated Poisson random variables with means equal to 1.5, 2.5, and 3.5. The solid dots correspond to the Poisson distribution, and the open circles are the zero-truncated Poisson.

When the mean is small, the probabilities between the two distributions differ significantly. But as the mean increases, the probabilities get closer and closer together. If the mean length of stay is larger than, say, 4 or 5, then using the Poisson distribution instead of the zero-truncated Poisson distribution would yield nearly identical results.

Exploratory Data Analysis

On average, we should have about 30 patients per provider, but the data shows a lot of variation—we have a provider with a single patient and another with 92. The overall mean length of stay in the hospital is equal to 9.9 days. Table 6.3 shows the top-five and bottom-five providers in terms of their mean length of stay along with the number of patients, their mean age group, how many of them consider themselves Caucasian, and the type of admission to the hospital.

The length of stay is strongly influenced by the type of admission but not the age of the patient. Table 6.4 displays the average hospital stay by age group and type of admission. As we read down the columns, there are no clear upward or downward trends—therefore, age group does not seem to be related to the number of days a patient stays in the hospital. Reading horizontally across, we find that nearly all elective admissions have the shortest stays, emergency admissions have the longest stays, and

Figure 6.10. Poisson and zero-truncated Poisson probabilities for means equal to 1.5, 2.5, and 3.5. As the mean increases, the difference between the Poisson and the zero-truncated Poisson distributions narrows quickly. The closed circles represent the Poisson distribution, and the open circles correspond to the zero-truncated Poisson distribution. Red-colored points correspond to a mean of 1.5, blue corresponds to 2.5, and purple has a mean of 3.5.

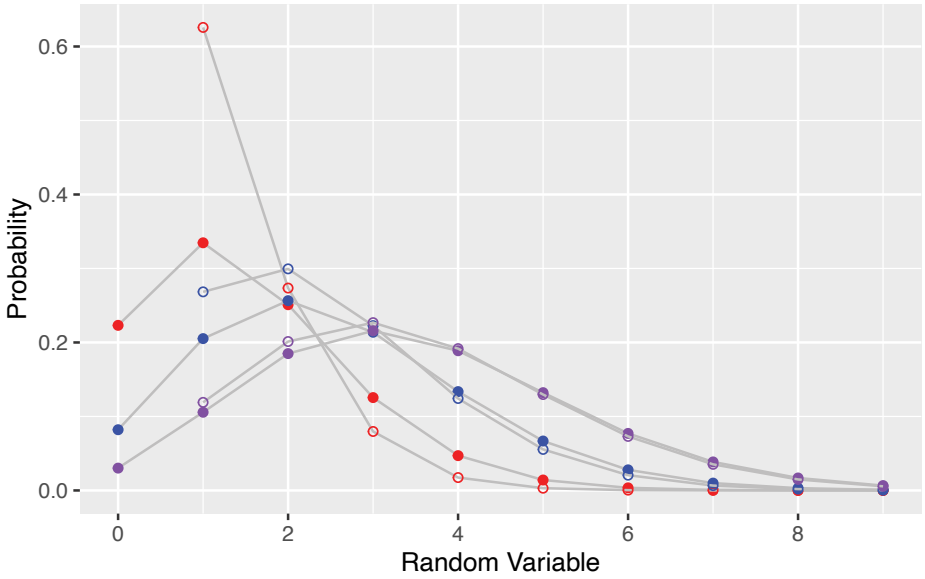


Table 6.3. Top-five and bottom-five providers sorted by mean length of stay in decreasing order. Also showing the number of patients for each provider, their mean age group, the number of patients who consider themselves Caucasian (White), and the number by type of admission to the hospital.

Provider	Number of Patients	Mean Stay	Mean Age	Count White	Type of Admission		
					Elective	Urgent	Emergency
32003	2	47.5	5.0	2	0	2	0
32002	10	28.3	5.3	9	0	0	10
32000	38	26.6	4.8	32	0	0	38
30073	4	21.8	6.5	0	2	2	0
30078	3	18.3	5.0	0	2	1	0
30025	3	4.7	4.3	2	3	0	0
30067	5	4.4	5.4	5	5	0	0
30060	2	3.5	5.5	2	1	1	0
30044	2	3.0	6.5	2	0	2	0
30068	1	2.0	4.0	1	1	0	0

Table 6.4. Average length of stay by age group and type of admission. Note that in nearly all cases elective admissions are the shortest and emergency admissions are the longest. An entry of "NA" means that there is no data for this combination.

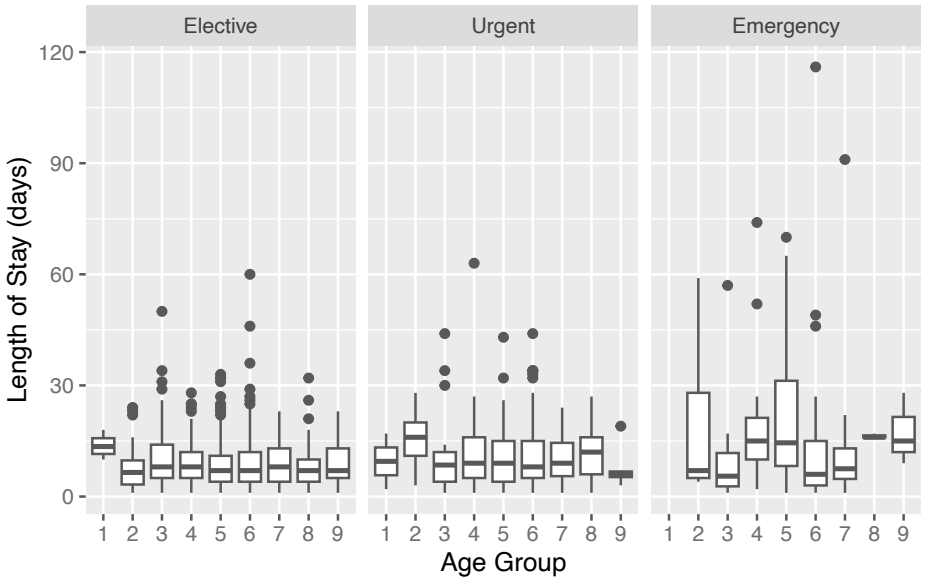
Age Group	Type of Admission		
	Elective	Urgent	Emergency
1	13.8	9.5	NA
2	7.5	16.2	20.6
3	9.7	11.4	12.6
4	9.1	11.0	18.5
5	8.3	10.8	22.7
6	9.1	11.4	16.9
7	8.9	10.6	15.4
8	7.8	12.1	16.3
9	8.7	7.7	17.3

urgent admissions are between. Also note that for age group 1, the average length of stay for elective admissions is very high at 13.8 days. But that high mean value is based on only four observations, and thus we should be careful about using these levels as base levels in our estimation of models.

Exercise 6.4 Further explore the relationship between `age` and `los`. Do not treat age as a numeric variable but do take into account the type of admission. Does the data have many outliers?

Solution 6.4 The following display shows `age` as a categorical variable and length of stay, `los`, using boxplots for each type of admission. Note that most of the outliers are for emergency and urgent admissions.

```
ggplot(data = db,
       mapping = aes(x = factor(age),
                     y = los)) +
  facet_wrap(~ type) +
  geom_boxplot() +
  labs(x = "Age Group",
       y = "Length of Stay (days)")
```



Modeling Length of Stay

Our response variable is length of stay, los, a count of the number of days a patient remained hospitalized. For now we ignore the information supplied by the variable provnum (medical provider) because that variable has a large number of levels, 54.

For our first model, we will use the Poisson distribution and include `hmo`, `white`, `type`, and `age.cat` as explanatory variables. The variable `age.cat` is a categorical

version of age, and we have selected age group 6 as the base level because that level has the largest number of observations. For type of admission, type, we have selected level elective as our base for the same reason.

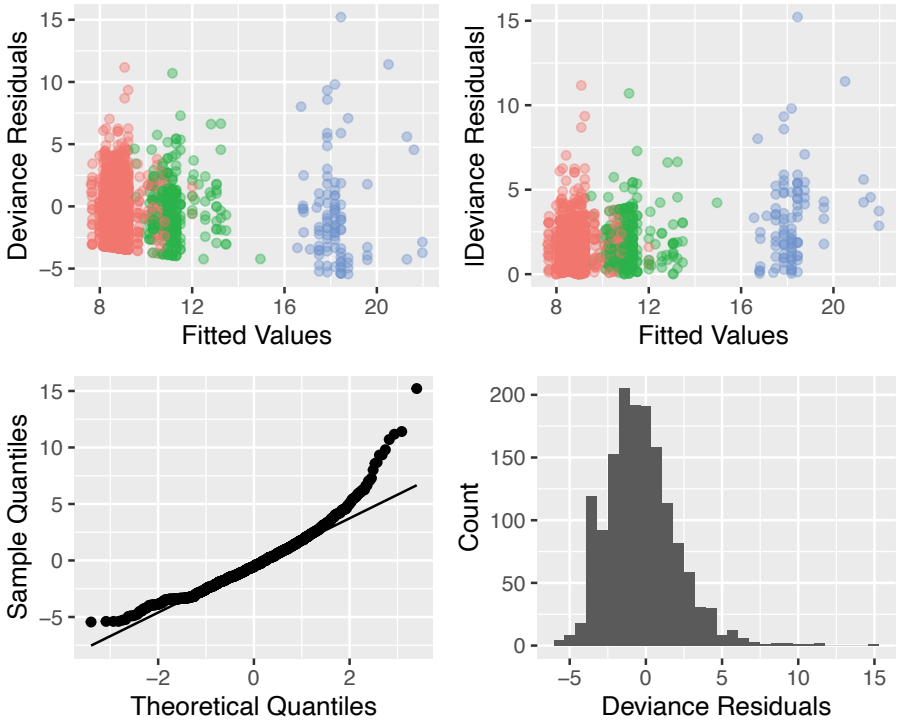
The model fit is summarized below.

```
Call:
glm(formula = los ~ type + white + hmo + age.cat,
    family = poisson(link = "log"), data = db)

Coefficients:
               Estimate  Std. Error  z value  Pr(>|z|)
(Intercept)     2.36401     0.03339   70.795   < 2e-16  ***
typeUrgent      0.21949     0.02113   10.388   < 2e-16  ***
typeEmergency   0.70906     0.02620   27.066   < 2e-16  ***
white          -0.15835     0.02912   -5.437  5.41e-08  ***
hmo            -0.07505     0.02400   -3.128   0.00176  **
age.cat1        0.12164     0.11900    1.022   0.30669
age.cat2       -0.09817     0.04645   -2.113   0.03456  *
age.cat3        0.01670     0.03032    0.551   0.58184
age.cat4       -0.01421     0.02536   -0.560   0.57524
age.cat5       -0.03259     0.02507   -1.300   0.19360
age.cat7       -0.05237     0.02921   -1.793   0.07295  .
age.cat8       -0.09316     0.03895   -2.392   0.01678  *
age.cat9       -0.09152     0.05179   -1.767   0.07716  .
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 8901.1 on 1494 degrees of freedom
Residual deviance: 8125.4 on 1482 degrees of freedom
AIC: 13867

Number of Fisher Scoring iterations: 5
```

Note that it appears that type, white, and hmo are all statistically significant variables. Many of the estimated coefficients for age.cat have a negative sign but do not appear to be significant at the 5% level. Relative to age group 6, every other group appears to have either a similar length of stay or a shorter one (groups 2 and 8).

This model unfortunately does not fit the data well. Figure 6.11 shows several diagnostic plots. The deviance residuals should be approximately normally distributed. Their mean is –0.281, and their standard deviation is 2.315. Clearly we have some very large residuals. The QQ plot in the bottom-left panel shows that the deviance residuals have much thicker tails than the normal distribution, and the bottom-right panel shows a nonsymmetrical distribution for the deviance residuals. Finally, the upper-right panel, which displays the absolute value of the deviance residuals against the fitted values, shows an increasing trend whereby the variance increases as the fitted values increase.

Figure 6.11. Diagnostic plots for the Poisson model predicting length of stay based on type of admission, self-reported race, age category, and whether the patient is a member of an HMO. The clusters, from left to right in the upper panels, correspond to elective, urgent, and emergency admissions to the hospital.



The clusters of observations seen in the upper panels arise because the different types of admission to the hospital (elective, urgent, and emergency) have little overlap in the response variable. We also see a clear decreasing pattern in the upper-left panel. The model overpredicts many of the emergency admissions.

Also, if the model fit had been good, we would expect an estimate of the dispersion parameter to be close to 1. Here both the mean deviance estimate as well as the Pearson estimate of the dispersion parameter are well above 1, indicating overdispersion.

```
Mean Dev. Estimate  Pearson Estimate
          5.482701          6.257832
```

But is the overdispersion real or apparent? Apparent overdispersion can arise when our modeling of the data is deficient. For example, not including an important explanatory variable in our model, using the wrong link function, or the presence of outliers might show that the estimate of the dispersion parameter is greater than 1, leading us to think that the data is overdispersed. But once we fix our model, the estimate of the dispersion parameter falls back close to unity.

Exercise 6.5 Outlier observations violate one of the most important assumptions in regression analysis, namely, that all of the observations come from the same data generation process.

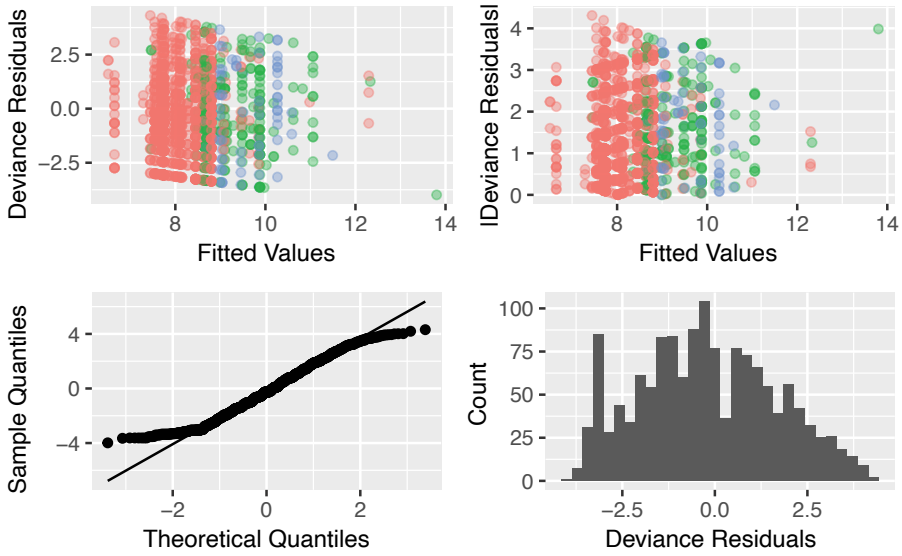
Repeat the above analysis, but first remove the largest 5% of observations in terms of length of stay. The Pearson estimate of the dispersion parameter should now be smaller, but it is still well above 1.

Solution 6.5 Remove the top 5% of observations and fit the Poisson model to the new dataset `dta`.

```
dta <- filter(db,
              db$los < quantile(db$los, probs = 0.95))
los.pois.no <- glm(los ~ type + white + hmo + age.cat,
                   data = dta,
                   family = poisson(link = "log"))
summary(los.pois.no)
```

```
Call:
glm(formula = los ~ type + white + hmo + age.cat,
    family = poisson(link = "log"), data = dta)

Coefficients:
               Estimate  Std. Error  z value  Pr(>|z|)
(Intercept)     2.15992     0.03889   55.541   < 2e-16  ***
typeUrgent      0.11565     0.02388    4.844  1.27e-06  ***
typeEmergency   0.15445     0.03909    3.952  7.76e-05  ***
white          -0.11289     0.03406   -3.314  0.000918  ***
hmo            -0.02104     0.02544   -0.827  0.408236
age.cat1        0.34929     0.11998    2.911  0.003600  **
age.cat2       -0.15257     0.05733   -2.661  0.007781  **
age.cat3        0.04527     0.03467    1.306  0.191613
age.cat4        0.12751     0.02834    4.499  6.84e-06  ***
age.cat5       -0.00572     0.02896   -0.198  0.843411
age.cat7        0.08706     0.03211    2.711  0.006709  **
age.cat8        0.03440     0.04200    0.819  0.412732
age.cat9        0.05025     0.05584    0.900  0.368183
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 4908.2 on 1410 degrees of freedom
Residual deviance: 4801.4 on 1398 degrees of freedom
AIC: 10096

Number of Fisher Scoring iterations: 5
```

Compute fitted values and deviance residuals, and generate the diagnostic plots.



The deviance residuals versus the fitted values (upper-left panel) show that many points seem to be on straight lines with a negative slope. These "patterns" are an artifact indicating that our response variable is an integer, and they arise for count and binomial (with response variable 0 or 1) models. For these models, using quantile residuals is recommended (Dunn and Smyth 1996). You can calculate quantile residuals for GLMs via the function `qresid()`, available in the `statmod` package.

The QQ plot shows that both tails of the deviance residuals are too thin in relation to the normal distribution. The upper-right panel depicts a funnel-type shape fanning inward as the fitted values increase.

The estimates of the dispersion parameter are

```
Mean Dev. Estimate  Pearson Estimate
          3.434447          3.305389
```

Both values are much smaller than in our previous model with all the observations, but they are still much larger than the theoretical value of 1.

The following exercise shows how apparent overdispersion can arise if we do not have the appropriate explanatory variables.

Exercise 6.6 Generate a dataset with a response variable that is Poisson distributed and a categorical variable with four levels. The response variable, `skip`, is the number of classes that students at a university skip in one semester. The categorical variable, `class`, classifies students by how many years they have already been at the university: 0 (freshman), 1 (sophomore), 2 (junior), or 3 (senior).

Each group should have the same number of students, and the mean number of classes skipped for freshmen is 4.5, sophomores is 1.5, juniors is 1.5, and seniors is 6.5.

1. Fit a Poisson GLM to the data ignoring the `class` variable and compute the Pearson estimate of the dispersion parameter. Is overdispersion present in this dataset?
2. Fit a Poisson GLM including the `class` variable and recompute the Pearson estimate of the dispersion parameter. Does the result indicate that the data is overdispersed?

Solution 6.6 Set a seed for the random number generator so we can reproduce our computations and generate our data based on four classes, each Poisson distributed with a different mean.

```
set.seed(12853)
N <- 100
dta <- tibble(skip = c(rpois(N, lambda = 4.5),
                       rpois(N, lambda = 2.5),
                       rpois(N, lambda = 1.5),
                       rpois(N, lambda = 6.5)),
              class = c(rep("freshman", N),
                        rep("sophomore", N),
                        rep("junior", N),
                        rep("senior", N)))
```

Ignoring the class standing of a student, we fit a Poisson model across all observations and compute the Pearson estimate $\hat{\phi}$ of the dispersion parameter.

```
m1 <- glm(skip ~ 1,
          data = dta,
          family = poisson(link = "log"))
(phi.hat <- sum(resid(m1, type = "pearson")^2) / df.residual(m1))
```

```
[1] 2.11863
```

The value of $\hat{\phi}$ is well above its theoretical value of 1, indicating that the data is overdispersed. Next, we include the explanatory variable `class` and recompute the Pearson estimate of the dispersion parameter, yielding a value that is much closer to 1.

```
m2 <- glm(skip ~ class,
          data = dta,
          family = poisson(link = "log"))
(phi.hat <- sum(resid(m2, type = "pearson")^2) / df.residual(m2))
```

```
[1] 0.9422353
```

Therefore, we conclude that the overdispersion we saw earlier is apparent.

As a consequence of the overdispersion, the estimated standard errors for our coefficients are too narrow, and this might lead us to infer that some explanatory variables are significant when in fact they are not. One may compensate for the overdispersion by inflating the standard error with a multiplicative factor equal to the square root of the estimated dispersion parameter. For our data this factor would be equal to approximately 2.5, and applying it to our fitted Poisson model would render all of the estimated coefficients for `age.cat` not significant at the 5% level as well as the indicator for the health maintenance organization, `hmo`.

Exercise 6.7 Adjust the standard errors by fitting a quasi-Poisson model, and verify the claims made in the previous paragraph.

Solution 6.7

```
los.qpoi <- glm(los ~ type + white + hmo + age.cat,
                data = db,
                family = quasipoisson(link = "log"))
summary(los.qpoi)
```

```
Call:
glm(formula = los ~ type + white + hmo + age.cat,
    family = quasipoisson(link = "log"), data = db)

Coefficients:
               Estimate  Std. Error  t value  Pr(>|t|)
(Intercept)     2.36401     0.08353   28.300   < 2e-16  ***
typeUrgent      0.21949     0.05286    4.153  3.48e-05  ***
typeEmergency   0.70906     0.06553   10.820   < 2e-16  ***
white          -0.15835     0.07285   -2.174    0.0299  *
hmo            -0.07505     0.06003   -1.250    0.2114
age.cat1        0.12164     0.29769    0.409    0.6829
age.cat2       -0.09817     0.11620   -0.845    0.3983
age.cat3        0.01670     0.07584    0.220    0.8258
age.cat4       -0.01421     0.06343   -0.224    0.8228
age.cat5       -0.03259     0.06272   -0.520    0.6034
age.cat7       -0.05237     0.07306   -0.717    0.4736
age.cat8       -0.09316     0.09745   -0.956    0.3392
age.cat9       -0.09152     0.12954   -0.707    0.4800
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for quasipoisson family taken to be 6.257848)

    Null deviance: 8901.1 on 1494 degrees of freedom
Residual deviance: 8125.4 on 1482 degrees of freedom
AIC: NA

Number of Fisher Scoring iterations: 5
```

Real overdispersion can also arise because the observations in our data are not independent of each other. That is, there are clusters of observations that are similar to each other and thus would violate the assumption that they are sampled independently. In our current application, patients belonging to a medical provider might have other (unobserved) characteristics in common, creating a cluster that contributes to the overdispersion.

There are 54 unique medical providers in the dataset, many of them (11) with fewer than five data points each. Hence, estimating a Poisson model with `provnum` as an explanatory variable may yield estimated coefficients for the medical providers that are extreme because they are based on a small number of observations. Estimating a Poisson GLM with `provnum` as an explanatory variable yields the following summary:

```
Call:
glm(formula = los ~ type + white + hmo + age.cat + provnum,
    family = poisson(link = "log"), data = db)

Coefficients:
              Estimate  Std. Error  z value  Pr(>|z|)
(Intercept)    2.00488     0.06025   33.277   < 2e-16  ***
typeUrgent     0.23065     0.02506    9.204   < 2e-16  ***
typeEmergency  0.09751     0.04993    1.953  0.050835  .
white         -0.01110     0.03205   -0.346  0.729081
hmo           -0.09637     0.02603   -3.702  0.000214  ***
age.cat1       0.24808     0.12120    2.047  0.040662  *
age.cat2      -0.06257     0.04726   -1.324  0.185514
age.cat3      -0.02743     0.03128   -0.877  0.380545
age.cat4      -0.02967     0.02594   -1.144  0.252772
age.cat5      -0.05041     0.02556   -1.973  0.048538  *
age.cat7      -0.07119     0.02988   -2.383  0.017182  *
age.cat8      -0.13813     0.03965   -3.483  0.000495  ***
age.cat9      -0.07188     0.05272   -1.363  0.172781
provnum30002   0.28683     0.06480    4.426  9.58e-06  ***
provnum30003   0.06364     0.15630    0.407  0.683877
provnum30006   0.31611     0.06218    5.084  3.70e-07  ***
provnum30007  -0.11236     0.12131   -0.926  0.354327
provnum30008   0.10272     0.08681    1.183  0.236687
provnum30009   0.50116     0.08813    5.687  1.29e-08  ***
provnum30010   0.39929     0.06552    6.094  1.10e-09  ***
provnum30011   0.37310     0.07103    5.252  1.50e-07  ***
provnum30012  -0.16181     0.10005   -1.617  0.105826
provnum30013   0.34277     0.06421    5.338  9.40e-08  ***
provnum30014   0.04435     0.06758    0.656  0.511686
provnum30016   0.56156     0.06725    8.350   < 2e-16  ***
provnum30017  -0.37225     0.10082   -3.692  0.000222  ***
provnum30018   0.22986     0.07877    2.918  0.003521  **
provnum30019   0.01020     0.10322    0.099  0.921248
```

```
provnum30022   0.22303     0.06876    3.244  0.001180  **
provnum30023   0.48433     0.15370    3.151  0.001627  **
provnum30024   0.25737     0.07018    3.667  0.000245  ***
provnum30025  -0.41530     0.27244   -1.524  0.127419
provnum30030   0.31769     0.07399    4.294  1.76e-05  ***
provnum30033   0.08566     0.35749    0.240  0.810630
provnum30035  -0.26337     0.17333   -1.519  0.128641
provnum30036   0.38796     0.10186    3.809  0.000140  ***
provnum30037  -0.14532     0.09916   -1.465  0.142793
provnum30038   0.34154     0.06731    5.074  3.89e-07  ***
provnum30043  -0.18059     0.11190   -1.614  0.106569
provnum30044  -1.09085     0.41219   -2.646  0.008134  **
provnum30055   0.18654     0.07591    2.457  0.014002  *
provnum30059   0.16790     0.17885    0.939  0.347845
provnum30060  -0.84098     0.38162   -2.204  0.027547  *
provnum30061   0.38228     0.05985    6.388  1.68e-10  ***
provnum30062  -0.15580     0.10219   -1.525  0.127333
provnum30064   0.37192     0.07454    4.990  6.05e-07  ***
provnum30065   0.38364     0.06843    5.606  2.07e-08  ***
provnum30067  -0.46192     0.21922   -2.107  0.035105  *
provnum30068  -1.27096     0.70906   -1.792  0.073057  .
provnum30069  -0.02031     0.09571   -0.212  0.831949
provnum30073   0.98774     0.12270    8.050  8.27e-16  ***
provnum30078   0.84927     0.14678    5.786  7.22e-09  ***
provnum30080   0.20900     0.08087    2.584  0.009753  **
provnum30083   0.17626     0.09096    1.938  0.052642  .
provnum30084   0.47395     0.16488    2.874  0.004047  **
provnum30085   0.07268     0.08337    0.872  0.383291
provnum30086   0.12834     0.08538    1.503  0.132803
provnum30087   0.37814     0.07492    5.048  4.47e-07  ***
provnum30088   0.24951     0.06398    3.900  9.62e-05  ***
provnum30089   0.18284     0.06584    2.777  0.005484  **
provnum30092  -0.06362     0.10993   -0.579  0.562791
provnum30093   0.28739     0.07169    4.009  6.10e-05  ***
provnum30094   0.12694     0.11596    1.095  0.273659
provnum32000   1.22846     0.07710   15.932   < 2e-16  ***
provnum32002   1.29685     0.09218   14.069   < 2e-16  ***
provnum32003   1.65103     0.11696   14.117   < 2e-16  ***
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 8901.1 on 1494 degrees of freedom
Residual deviance: 7080.2 on 1429 degrees of freedom
AIC: 12927

Number of Fisher Scoring iterations: 5
```

Previously, in Table 6.3, we displayed the top-five and bottom-five medical providers by the average length of stay of their patients. Provider 30068 has the lowest average length of stay, only two days, but has only one patient. The estimated coefficient for this provider is equal to –1.271 with a standard error equal to 0.709. This is the lowest estimated coefficient, and it has the highest standard error. Should we completely trust such an estimate? Most likely not.

Exercise 6.8 Which medical provider has the largest estimated coefficient? How many patients does this provider have, and what is the average length of stay for these patients?

Solution 6.8 Provider 32003 has the largest estimated coefficient, equal to 1.651, with a standard error of 0.117. The number of patients for this provider is equal to only two.

Since the number of observations for each medical provider varies significantly, we may try to address that by applying some weights (based on the number of observations by provider) to the observations in our dataset. Unfortunately, doing a weighted quasi-Poisson regression where the weights are proportional to the number of patients for each medical provider yields estimated coefficients that are close to the unweighted estimates (except for a handful of providers).

```
Call:
glm(formula = los ~ type + white + hmo + age.cat + provnum,
    family = quasipoisson(link = "log"), data = db2,
    weights = n.obs)

Coefficients:
               Estimate  Std. Error  t value  Pr(>|t|)
(Intercept)    1.965719    0.127297   15.442   < 2e-16  ***
typeUrgent     0.233498    0.057253    4.078  4.79e-05  ***
typeEmergency  0.070619    0.109089    0.647  0.517506
white          0.053577    0.075344    0.711  0.477139
hmo           -0.095672    0.052505   -1.822  0.068641  .
age.cat1       0.241992    0.282850    0.856  0.392389
age.cat2      -0.111448    0.113022   -0.986  0.324264
age.cat3      -0.004385    0.068611   -0.064  0.949054
age.cat4      -0.054710    0.058995   -0.927  0.353893
age.cat5      -0.074684    0.056869   -1.313  0.189301
age.cat7      -0.112541    0.066895   -1.682  0.092717  .
age.cat8      -0.166028    0.086961   -1.909  0.056432  .
age.cat9      -0.181743    0.126114   -1.441  0.149774
provnum30002   0.289466    0.129847    2.229  0.025950  *
provnum30003   0.077958    0.926951    0.084  0.932988
provnum30006   0.313415    0.120664    2.597  0.009489  **
```

provnum30007	-0.121399	0.479380	-0.253	0.800119	
provnum30008	0.105866	0.238715	0.443	0.657483	
provnum30009	0.496947	0.293825	1.691	0.090997	.
provnum30010	0.402419	0.133208	3.021	0.002564	**
provnum30011	0.362179	0.159036	2.277	0.022913	*
provnum30012	-0.174294	0.305841	-0.570	0.568846	
provnum30013	0.339376	0.127773	2.656	0.007993	**
provnum30014	0.039314	0.132206	0.297	0.766228	
provnum30016	0.569275	0.150209	3.790	0.000157	***
provnum30017	-0.372642	0.276532	-1.348	0.178015	
provnum30018	0.224550	0.199666	1.125	0.260936	
provnum30019	0.015502	0.348737	0.044	0.964551	
provnum30022	0.236672	0.142651	1.659	0.097316	.
provnum30023	0.480942	1.109426	0.434	0.664713	
provnum30024	0.260151	0.152573	1.705	0.088394	.
provnum30025	-0.421529	2.363310	-0.178	0.858462	
provnum30030	0.310828	0.175149	1.775	0.076169	.
provnum30033	0.060146	5.410482	0.011	0.991132	
provnum30035	-0.254227	1.033011	-0.246	0.805638	
provnum30036	0.379939	0.403487	0.942	0.346536	
provnum30037	-0.141328	0.304890	-0.464	0.643049	
provnum30038	0.338189	0.140090	2.414	0.015900	*
provnum30043	-0.184867	0.402848	-0.459	0.646375	
provnum30044	-1.099493	4.418330	-0.249	0.803514	
provnum30055	0.175451	0.176889	0.992	0.321428	
provnum30059	0.148876	1.315957	0.113	0.909943	
provnum30060	-0.857764	4.090546	-0.210	0.833936	
provnum30061	0.379276	0.113578	3.339	0.000861	***
provnum30062	-0.165666	0.331095	-0.500	0.616901	
provnum30064	0.395085	0.181491	2.177	0.029652	*
provnum30065	0.394153	0.147450	2.673	0.007600	**
provnum30067	-0.456194	1.462445	-0.312	0.755132	
provnum30068	-1.271439	10.819353	-0.118	0.906468	
provnum30069	-0.028841	0.298681	-0.097	0.923087	
provnum30073	1.045035	0.829866	1.259	0.208134	
provnum30078	0.904141	1.197396	0.755	0.450320	
provnum30080	0.204418	0.208851	0.979	0.327859	
provnum30083	0.173088	0.272612	0.635	0.525579	
provnum30084	0.536975	1.353745	0.397	0.691678	
provnum30085	0.078455	0.214062	0.367	0.714042	
provnum30086	0.127506	0.230451	0.553	0.580152	
provnum30087	0.376572	0.183396	2.053	0.040222	*
provnum30088	0.244327	0.124009	1.970	0.049005	*
provnum30089	0.174125	0.129690	1.343	0.179609	
provnum30092	-0.065579	0.398792	-0.164	0.869404	

```
provnum30093    0.289241    0.158450    1.825  0.068143  .
provnum30094    0.120033    0.491714    0.244  0.807179
provnum32000    1.253908    0.167028    7.507  1.06e-13  ***
provnum32002    1.340420    0.323211    4.147  3.56e-05  ***
provnum32003    1.634917    1.116040    1.465  0.143161
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for quasipoisson family taken to be 2.544499)

    Null deviance: 4014.1 on 1494 degrees of freedom
Residual deviance: 3434.7 on 1429 degrees of freedom
AIC: NA

Number of Fisher Scoring iterations: 5
```

This approach does not resolve our issue. Moreover, these medical providers are not the universe of all medical providers. They are only a sample from all possible medical providers, and we would like to be able to infer something about their population. To that end, we fit a GLMM with a random intercept varying by provider.

First, we define the model for the mean where we want a log-link, fixed effects of `type`, `white`, `hmo`, and `age.cat`, and we want to define a random intercept for `provnum`. In addition, we fit the GLMM with a Poisson distribution, a log-link function for the mean, a random effect for medical provider, and a constant dispersion parameter.

```
model.mu <- DHGLMMODELING(Model = "mean",
                          Link = "log",
                          LinPred = los ~ type + white +
                            hmo + age.cat + (1 | provnum),
                          RandDist = "gamma")
model.phi <- DHGLMMODELING(Model = "dispersion")
```

```
los.re <- dhglmfit(RespDist = "poisson",
                   DataMain = db,
                   MeanModel = model.mu,
                   DispersionModel = model.phi)
```

```
Distribution of Main Response :
                        "poisson"
[1] "Estimates from the model(mu)"
los ~ type + white + hmo + age.cat + (1 | provnum)
[1] "log"
```

```
                Estimate  Std. Error  t-value
(Intercept)      2.41205     0.07681  31.4018
typeUrgent       0.23690     0.02488   9.5218
typeEmergency    0.13434     0.04818   2.7885
white           -0.01709     0.03173  -0.5385
hmo             -0.09678     0.02600  -3.7222
age.cat1         0.24307     0.12109   2.0074
age.cat2        -0.06605     0.04717  -1.4003
age.cat3        -0.02877     0.03122  -0.9215
age.cat4        -0.02933     0.02590  -1.1324
age.cat5        -0.05183     0.02552  -2.0312
age.cat7        -0.07229     0.02982  -2.4243
age.cat8        -0.13931     0.03961  -3.5167
age.cat9        -0.07316     0.05266  -1.3894
[1] "Estimates for logarithm of lambda=var(u_mu)"
[1] "gamma"
          Estimate  Std. Error  t-value
provnum     -1.502      0.2001    -7.51
[1] "====== Likelihood Function Values and Condition AIC ======"
                                          [,1]
-2ML (-2 p_v(mu) (h))           : 13045.785
-2RL (-2 p_beta(mu),v(mu) (h)) : 13107.507
cAIC                            : 12929.385
Scaled Deviance                 :  7087.167
df                              :  1431.471
```

The estimated coefficients for our random intercept model do not differ significantly from the fixed effects estimated in the Poisson GLM. The intercept and the coefficient for the emergency type of admission are the only ones where the difference is a bit larger.

Table 6.5 shows the estimated coefficients and their *t*-statistics for the Poisson, weighted Poisson, and the Poisson model with random intercepts. In all three models the type of admission to the hospital is significant. Note that in all three models urgent admissions have a longer average length of stay compared with elective admissions. But even though emergency admissions also have a positive coefficient, the size is smaller than for urgent admissions, which we might feel goes against intuition. Moreover, for the Poisson and weighted quasi-Poisson models this coefficient is not statistically significant, but it is for the random intercepts model.

From these coefficients and their standard errors we can see that the type of admission is important, but the size of these coefficients may not be intuitive. Both urgent and emergency admissions have positive coefficients, indicating that these patients will, on average, stay longer in the hospital compared with elective admissions. But the coefficient for urgent admission is much bigger than that for emergency admission. We might expect emergency admission patients to be in worse health compared with urgent admissions and thus stay longer in the hospital.

Table 6.5. Estimated coefficients and their *t*-values from three models that include `provnum` as an explanatory variable but whose coefficients are not shown. The models are Poisson, weighted quasi-Poisson, and Poisson with random intercepts.

	Poisson			Wtd. quasi-Poisson			Random Intercepts		
Variable	Est.	*t*-stat		Est.	*t*-stat		Est.	*t*-stat	
Intercept	2.005	33.277	*	1.966	15.442	*	2.412	31.402	*
Type Urgent	0.231	9.204	*	0.233	4.078	*	0.237	9.522	*
Type Emergency	0.098	1.953		0.071	0.647		0.134	2.788	*
White	–0.011	–0.346		0.054	0.711		–0.017	–0.539	
HMO	–0.096	–3.702	*	–0.096	–1.822		–0.097	–3.722	*
Age Group 1	0.248	2.047	*	0.242	0.856		0.243	2.007	*
Age Group 2	–0.063	–1.324		–0.111	–0.986		–0.066	–1.400	
Age Group 3	–0.027	–0.877		–0.004	–0.064		–0.029	–0.921	
Age Group 4	–0.030	–1.144		–0.055	–0.927		–0.029	–1.132	
Age Group 5	–0.050	–1.973	*	–0.075	–1.313		–0.052	–2.031	*
Age Group 7	–0.071	–2.383	*	–0.113	–1.682		–0.072	–2.424	*
Age Group 8	–0.138	–3.483	*	–0.166	–1.909		–0.139	–3.517	*
Age Group 9	–0.072	–1.363		–0.182	–1.441		–0.073	–1.389	

The self-identified indicator of race, `white`, is not significant, but the indicator of membership in a health maintenance organization, `hmo`, is significant with a negative coefficient, suggesting that HMO patients' length of stay is about 10% shorter than that of non-HMO patients. Some of the age group coefficients are not statistically significant while others are, and the coefficients do not suggest a linear relationship between increasing age and length of stay. Note that the youngest age group has a positive coefficient that is statistically significant. This suggests that young patients tend to stay longer (about 25% longer) at the hospital compared with patients in age group 6.

Exercise 6.9 Perhaps the reason younger patients stay in the hospital longer than patients in age group 6 is because older patients are more likely to die in the hospital.

One of the variables in our dataset, `died`, tells us whether the patient died at the hospital. We cannot use that variable to predict the length of stay, but we can check whether our intuition that patients dying in the hospital tend to be the older patients is correct.

Based on the data, compute the empirical probability of dying at the hospital split by age group.

Solution 6.9 The variable `dead` is an indicator variable where a 1 means the patient died at the hospital, and otherwise it is zero. To compute the probability of dying, we need to group our data by age and compute the mean value of died for each group. It's important to also know how many patients are in each group.

```
db |>
  group_by(age.cat) |>
  summarize(n.dead = sum(died),
            n.patients = n(),
            prob.death = mean(died)) |>
  arrange(levels(age.cat))
```

```
# A tibble: 9 x 4
  age.cat  n.dead  n.patients  prob.death
  <fct>     <int>       <int>       <dbl>
1 1             1           6       0.167
2 2            14          60       0.233
3 3            38         163       0.233
4 4            84         291       0.289
5 5           104         317       0.328
6 6           120         328       0.366
7 7            87         191       0.455
8 8            46          93       0.495
9 9            19          46       0.413
```

Note that as age increases, the probability of dying also increases, except for age group 9, where it drops.

Figure 6.12 displays some of the standard diagnostic plots for our random intercept model. The panel on the upper left shows the fitted values versus the studentized residuals. The fitted values are on the scale of the linear predictor (as opposed to being on the scale of the response). There is a cluster of 50 claims with a fitted value between 3 and 3.5. These claims come from three medical providers, and nearly all of them are emergency admissions, with a few of them being urgent admissions. None of the patients belong to an HMO. Note that the studentized residuals for these observations have both positive and negative values. Hence, our current model both underpredicts as well as overpredicts these patients.

The panel on the upper right shows the absolute value of the studentized residuals against the fitted values. There is a general upward trend mainly driven by the observations with fitted values in excess of 3.

The panel on the lower left displays a QQ plot showing that the studentized residuals have a thicker right-hand tail than the normal distribution. And the lower-right panel shows that the distribution of the residuals is skewed to the right, in agreement with the QQ plot.

Figure 6.13 shows the estimated density function for the random effects along with the individual estimates for each medical provider.

Figure 6.12. Diagnostic plots for the length of stay random intercept model by medical provider.

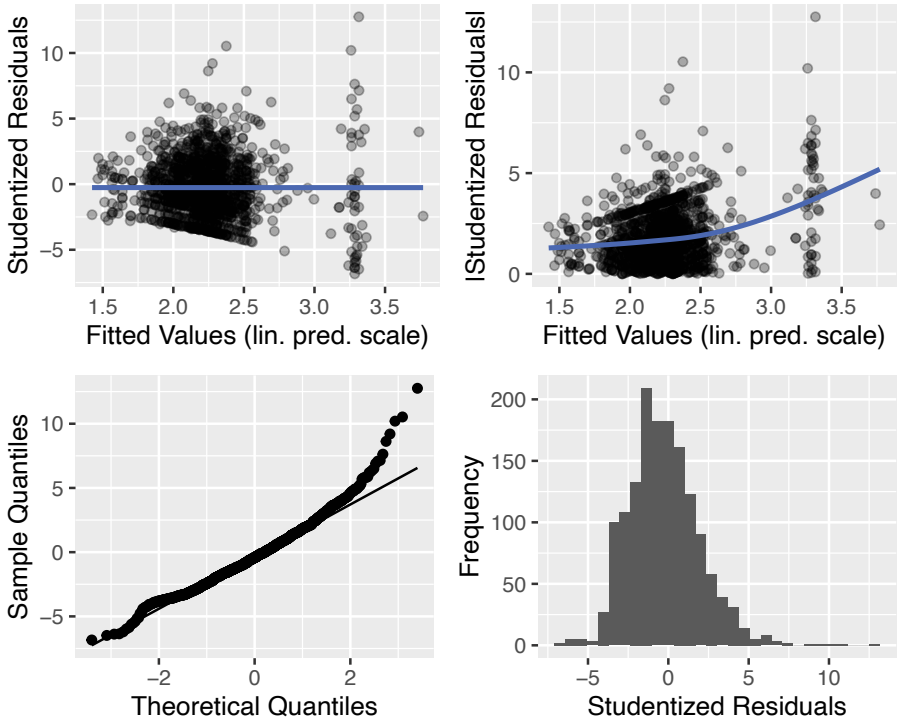
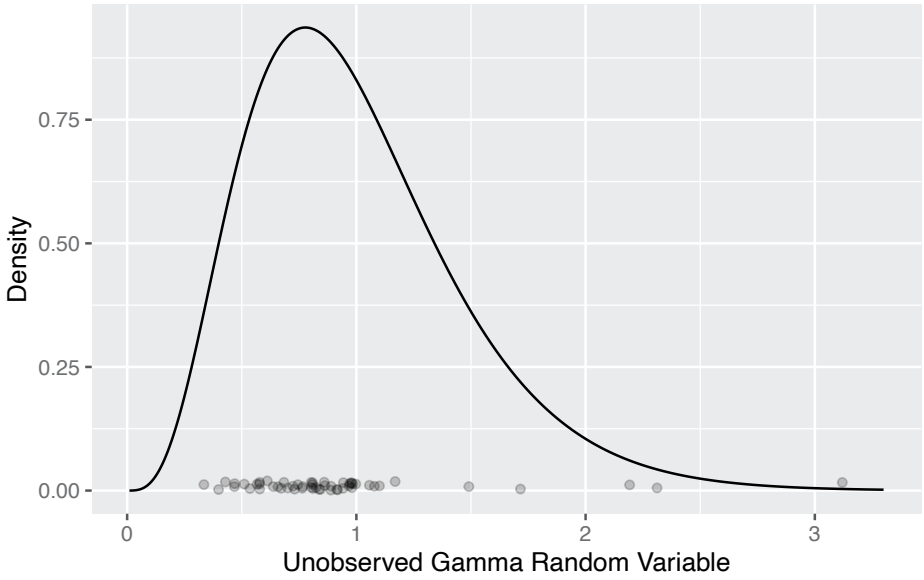


Figure 6.13. The estimated density function for the intercept random effects along with the estimated medical provider random effects.



6.3. Swedish Bus Insurance

In this section we use a bus insurance example from Section 4.5 of Ohlsson and Johansson (2010). The data comes from the former Swedish insurance company Wasa for the years 1990 to 1998. It concerns insuring transportation companies and can be accessed from Ohlsson and Johansson's book. Each company owns one or more buses that are insured for shorter or longer periods of time. At that time, the pricing scheme was rather simple, based on geographic zone and the age class of the bus.

The variables available and their descriptions, taken from Ohlsson and Johansson (2010), are as follows. We use an English abbreviation first and list the original Swedish acronym in parentheses:

1. `zone` (`ZON`): geographic subdivision of Sweden into seven zones, based on parishes and numbered 1 through 7.
2. `bus.age` (`BUSSALD`): the age class of the bus, in the span 0 to 4.
3. `co.id` (`KUNDNR`): an ID for the company, recoded here for confidentiality reasons.
4. `no.obs` (`ANTAVT`): number of observations for the company in a given tariff cell based on the zone and age class. There may be more than one observation per bus, since each renewal is counted as a new observation.
5. `dur` (`DUR`): duration measured in days and aggregated over all observations in the tariff cell.
6. `clm.cnt` (`ANTSKAD`): the corresponding number of claims.
7. `tot.cost` (`SKADKOST`): the corresponding total claim cost in Swedish kronor (unadjusted for inflation).

The variable `dur` is the amount of time a policy is in force, and thus we may use it as a measure of exposure. The premium is based on a *bus-year* unit of exposure, and the premium for a company would be the sum across all buses in its fleet.

Exploratory Data Analysis

The dataset has 1,542 observations and seven variables. The variables geographic zone, `zone`; age class of the bus, `bus.age`; and company identification, `co.id`, are categorical, and the remaining variables—number of observations, `no.obs`; duration (or exposure), `dur`; claim counts, `clm.cnt`; and total claim cost, `tot.cost`—are numeric.

There are 666 unique company IDs in the dataset. Some companies have a large amount of exposure while others have very little, so using this categorical variable when estimating frequency or severity at the level of a single company would be problematic.

Table 6.6 displays summary statistics for the numeric variables. Note that the number of claims, `clm.cnt`, ranges from zero to 402, but the 75th percentile is just one claim. Hence, we suspect that something is not quite right with the number of claims for one or more companies. Similarly, the total claim cost has some negative entries, and we can see that only 616 records have a non-missing entry. The missing entries correspond to having zero claim counts.

Table 6.6. Summary statistics for the numeric variables in the bus dataset. Variable `no.obs` is the number of observations in a particular tariff cell. Note that each renewal counts as one observation. Duration, `dur`, is measured in days. The variable `clm.cnt` is the number of claims, and `tot.cost` is the total loss cost.

				Percentiles				
Variable	Count	Mean	Std. Dev.	25	50	75	Min	Max
no.obs	1,542	13	33	2	4	9	1	392
dur	1,542	2,284	4,970	365	725	1,876	1	66,327
clm.cnt	1,542	2	15	0	0	1	0	402
tot.cost	616	52,871	143,944	0	3,525	39,897	−17,318	1,330,610

Regarding the negative entries for total loss cost, they arise from 60 rows of data. We do not know why those entries have negative loss costs, and we will remove them from our analysis. Also, the top-10 claim counts are

```
[1]  402  377  55  53  38  34  34  29  28  27
```

The two largest, 402 and 377, come from company 145. We suspect that these entries are erroneous, and because we cannot go back to the source systems or other sources of information to correct them, we will delete company 145 from our analysis. This will remove a total of nine rows of data.

In Figure 6.14 we have the histograms of our four numeric variables. Note that all four are highly skewed to the right, and to enhance each of the displays we have omitted some large observations.

In the 1990s, the rating plan might have been relatively simple—perhaps it would have used only two rating factors: zone and age class. Using these two variables, Table 6.7 shows the empirical frequency (per year of exposure) for each of the 35 (7×5) rating cells in the plan. There is considerable variation in the empirical frequencies, and thus we can use `zone` and `bus.age` as part of a rating plan.

Exercise 6.10 Summarize the number of claims in the bus data by `zone` and `bus.age`. How many cells have a small number of claims, say, fewer than five?

You can repeat the exercise with exposure.

Solution 6.10 We summarize the data as follows:

```
db |>
  group_by(zone, bus.age) |>
  summarize(clm.cnt = sum(clm.cnt),
            .groups = "drop") |>
  pivot_wider(names_from = bus.age,
              values_from = clm.cnt)
```

```
# A tibble: 7 x 6
  zone      '0'    '1'    '2'    '3'    '4'
  <fct>   <dbl>  <dbl>  <dbl>  <dbl>  <dbl>
1 1          20      8     14     24     38
2 2          14     26     13     25    151
3 3          11     19     19     11    132
4 4          83    105    150    118    709
5 5           3      8      0      1     51
6 6          29     14     14     16    184
7 7           1      2      1      3      7
```

Note that most of the cells in zone 7 and zone 5 have very few counts. All other cells have many more claim counts. In particular, buses in age category 4 have the most claims, and of the zones, zone 4 has the most claims.

Looking at exposure, we have

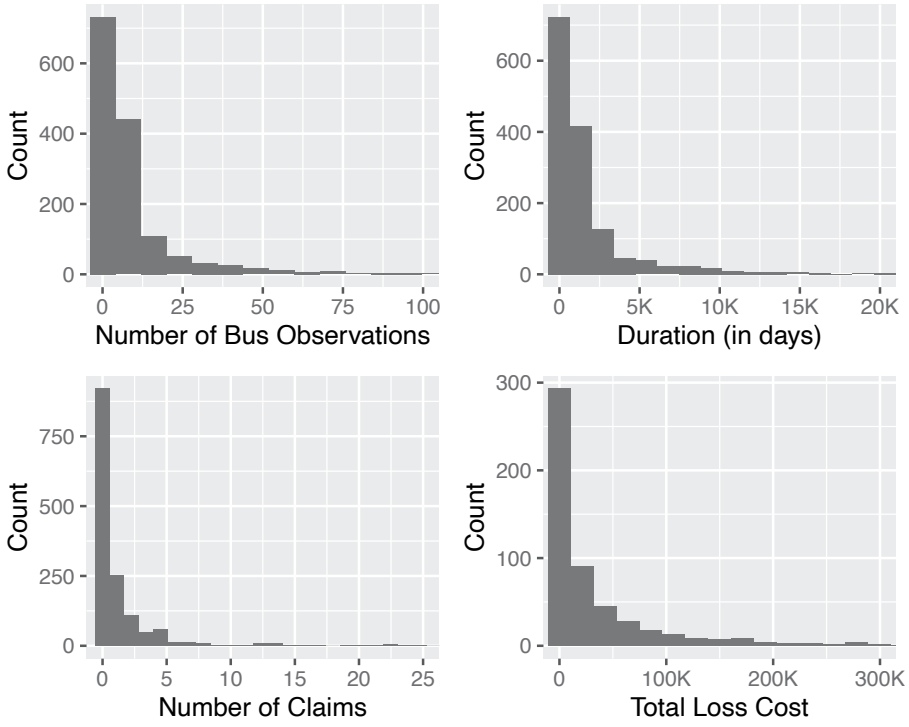
```
db |>
  group_by(zone, bus.age) |>
  summarize(expo = sum(dur),
            .groups = "drop") |>
  pivot_wider(names_from = bus.age,
              values_from = expo)
```

```
# A tibble: 7 x 6
  zone      '0'     '1'     '2'     '3'      '4'
  <fct>   <dbl>   <dbl>   <dbl>   <dbl>    <dbl>
1 1       10061   10823   15073   30924    79106
2 2       22407   19023   18736   24372   168620
3 3       25624   24768   22716   29321   278519
4 4       88665  112912  143189  151654  1434918
5 5        3299    4541    3800    2958    87265
6 6       20791   21966   20740   23695   296228
7 7        2739    2496    3135    3377    31632
```

Again, zone 4 and age category 4 have the most exposure. Zones 5 and 7 have much smaller exposures.

Because both `zone` and `bus.age` have few levels, we can reliably estimate the frequency for most cells, but within each cell we may still have many companies whose experiences are not similar to each other. Hence, we would like to include a company identifier, `co.id`, into the rating plan. Unfortunately, the company identifier has more than 600 unique entries. Hence, adding it as a regular classification variable would give us a rating plan with $7 \times 5 \times 660 = 23{,}100$ individual cells. Because our data only has around 1,500 observations, most of the cells would be empty. We cannot go in this direction.

Figure 6.14. Histograms of the numeric variables in the bus dataset. Note that all four are heavily skewed to the right, and to enhance the display we do not show all data points. For the number of bus observations, 30 observations greater than 100 are not shown. For duration, 26 observations greater than 20,000 are not displayed. Seven claim counts greater than 25 are not shown, and 23 observations for total loss cost greater than 300,000 are also not shown.



What we can do is bring to bear the tools of credibility and mixed-effects models to help us incorporate the information we have regarding the experience of each company.

Modeling Frequency

Let us start with modeling frequency of claims by first looking at the empirical data we have. The overall *yearly frequency*, without regard to any classification variables, is equal to 0.228. If we cross-classify our data by geographical zone and bus age category, then the *yearly frequency* for each combination is displayed in Table 6.7.

Note that zone 1 and bus age category 0 has a very high empirical yearly frequency of 0.726. Looking at the data for this cell, along with the individual companies' annual frequency, given in Table 6.8, we can see that the experience is quite heterogeneous. We have companies with a small amount of exposure (356 and 460 days) along with companies with more exposure (2,191 and 1,949 days).

We can fit a GLM to geographic zone and bus age, ignoring company for the moment, to get an initial sense of how we might model frequency.

Table 6.7. Empirical frequency (per year of exposure) for the bus dataset by geographic zone and age class of the bus.

	Bus Age Class				
Zone	0	1	2	3	4
1	0.726	0.270	0.339	0.283	0.175
2	0.228	0.499	0.253	0.374	0.327
3	0.157	0.280	0.305	0.137	0.173
4	0.342	0.339	0.382	0.284	0.180
5	0.332	0.643	0.000	0.123	0.213
6	0.509	0.233	0.246	0.246	0.227
7	0.133	0.292	0.116	0.324	0.081

Table 6.8. Observations available for zone 1 and bus age category 0 along with the empirical annual frequency for each company. The overall annual frequency for this cell is equal to 0.726.

Zone	Bus Age	Company	Duration	Claim Count	Frequency
1	0	518	1,079	8	2.706
1	0	184	1,213	4	1.204
1	0	226	1,949	5	0.936
1	0	597	457	1	0.799
1	0	231	1,777	1	0.205
1	0	385	2,191	1	0.167
1	0	471	356	0	0.000
1	0	539	460	0	0.000
1	0	15	579	0	0.000

Exercise 6.11 Fit a Poisson model to frequency of claims using `zone` and `bus.age` as classification variables and a log-link function. Does the model fit the data well?

Solution 6.11 The Poisson model can be fitted via

```
fq.poi <- glm(clm.cnt ~ zone + bus.age,
              data = db,
              family = poisson(link = "log"),
              offset = log(dur))
summary(fq.poi)
```

```
Call:
glm(formula = clm.cnt ~ zone + bus.age, data = db, offset = log(dur),
    family = poisson(link = "log"))

Coefficients:
             Estimate  Std. Error  z value  Pr(>|z|)
(Intercept)  -6.943244    0.123755  -56.105   < 2e-16  ***
zone2         0.279470    0.118671    2.355    0.0185  *
zone3        -0.271849    0.122356   -2.222    0.0263  *
zone4        -0.084820    0.103045   -0.823    0.4104
zone5        -0.002343    0.160733   -0.015    0.9884
zone6         0.034376    0.117055    0.294    0.7690
zone7        -0.718687    0.284924   -2.522    0.0117  *
bus.age1      0.008051    0.108257    0.074    0.9407
bus.age2      0.014442    0.104842    0.138    0.8904
bus.age3     -0.213333    0.106407   -2.005    0.0450  *
bus.age4     -0.531115    0.083951   -6.327  2.51e-10  ***
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 2142.9 on 1472 degrees of freedom
Residual deviance: 1985.3 on 1462 degrees of freedom
AIC: 3502.9

Number of Fisher Scoring iterations: 5
```

The mean deviance estimate of the dispersion parameter ϕ is equal to

```
[1] 1.358
```

which is clearly larger than the theoretical value of $\phi = 1$, indicating that the data is over-dispersed and the Poisson model does not fit well.

The Pearson estimate of the dispersion parameter yields a value of 1.465 and thus a similar conclusion.

Since the Poisson model does not fit the data well, we proceed to fit a negative binomial model using `zone` and `bus.age` as main effects. Before fitting the model, we select zone 4 and bus age 4 as the base levels for these factors' variables because at these levels they have the most exposure.

```
Call:
glm.nb(formula = clm.cnt ~ zone.f + bus.age.f + offset(log(dur)),
    data = db, init.theta = 2.027840354, link = log)

Coefficients:
             Estimate  Std. Error   z value  Pr(>|z|)
(Intercept)  -7.66330     0.06203  -123.536   < 2e-16  ***
zone.f1       0.17132     0.16878     1.015  0.310083
```

```
zone.f2       0.30441     0.12369     2.461  0.013848  *
zone.f3      -0.07103     0.11665    -0.609  0.542551
zone.f5       0.20278     0.18607     1.090  0.275806
zone.f6       0.16002     0.10442     1.533  0.125394
zone.f7      -0.49637     0.32616    -1.522  0.128046
bus.age.f0    0.59037     0.11936     4.946  7.57e-07  ***
bus.age.f1    0.46727     0.11740     3.980  6.88e-05  ***
bus.age.f2    0.40988     0.11836     3.463  0.000534  ***
bus.age.f3    0.37029     0.11220     3.300  0.000965  ***
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for Negative Binomial(2.0278) family
taken to be 1)

    Null deviance: 1252.1 on 1472 degrees of freedom
Residual deviance: 1196.4 on 1462 degrees of freedom
AIC: 3188.6

Number of Fisher Scoring iterations: 1

              Theta: 2.028
          Std. Err.: 0.252

 2 x log-likelihood: -3164.615
```

Note that the coefficients for `bus.age` are positive and decline steadily as the age category increases. All of the `zone` coefficients have large standard errors compared with their estimated values except for zone 2. Hence, it appears that all zones, except zone 2, have a similar claim frequency as zone 4.

Annual frequency predictions from the negative binomial model are displayed in Table 6.9, and they do not take into account that the claims experience comes from different companies.

Next, we incorporate the information available in the company identification variable, `co.id`, by adding it as a random effect for the intercept. We first define the model structure we want for the mean, and we will keep the dispersion parameter fixed; therefore, we are defining a GLMM where the response distribution is Poisson and the random effect is gamma distributed.

```
model.mu <- DHGLMMODELING(Model = "mean",
                          Link = "log",
                          LinPred = clm.cnt ~ zone.f +
                            bus.age.f + (1 | co.id),
                          RandDist = "gamma",
                          Offset = log(db$dur))
model.phi <- DHGLMMODELING(Model = "dispersion")
```

Table 6.9. Annual mean frequency predictions from the negative binomial model with main effects for geographic zone and bus age class.

	Bus Age Class				
Zone	0	1	2	3	4
1	0.367	0.325	0.307	0.295	0.204
2	0.420	0.371	0.350	0.337	0.232
3	0.288	0.255	0.241	0.231	0.160
4	0.309	0.274	0.258	0.248	0.171
5	0.379	0.335	0.316	0.304	0.210
6	0.363	0.321	0.303	0.291	0.201
7	0.188	0.167	0.157	0.151	0.104

We fit our model via the `dhglmfit()` function as follows:

```
fq.poi.re <- dhglmfit(RespDist = "poisson",
                      DataMain = db,
                      MeanModel = model.mu,
                      DispersionModel = model.phi)
```

```
Distribution of Main Response :
                          "poisson"
[1] "Estimates from the model(mu)"
clm.cnt ~ zone.f + bus.age.f + (1 | co.id)
[1] "log"
             Estimate  Std. Error    t-value
(Intercept)  -7.65055     0.05797  -131.9806
zone.f1       0.22644     0.14761     1.5340
zone.f2       0.44224     0.13197     3.3512
zone.f3       0.03305     0.12813     0.2579
zone.f5       0.28805     0.18542     1.5535
zone.f6       0.20941     0.11306     1.8522
zone.f7      -0.57862     0.36349    -1.5919
bus.age.f0    0.51538     0.08706     5.9201
bus.age.f1    0.50149     0.08194     6.1204
bus.age.f2    0.46698     0.07892     5.9170
bus.age.f3    0.36315     0.08001     4.5390
[1] "Estimates for logarithm of lambda=var(u_mu)"
[1] "gamma"
       Estimate  Std. Error  t-value
co.id    -1.097      0.1033   -10.62
```

```
[1] "====== Likelihood Function Values and Condition AIC ======"
                                              [,1]
-2ML  (-2 p_v(mu)  (h))              : 3171.369
-2RL  (-2 p_beta(mu),v(mu)  (h))  : 3199.679
cAIC                                 : 3107.045
Scaled Deviance                      : 1194.541
df                                   : 1264.553
```

From the above output we can see that the estimated coefficients are not too different from those we obtained for the negative binomial model. We also have the estimated variance for our unobserved gamma random effect. Its value is equal to 0.334. The density function for the random effect, along with company estimated random effects, is shown in Figure 6.15.

Note that we have five points with an estimated random effect greater than 2.25. In Table 6.10 we extract these companies and provide their total claim count, exposure,

Figure 6.15. The estimated density function for the intercept random effects along with the individual estimated company effects.

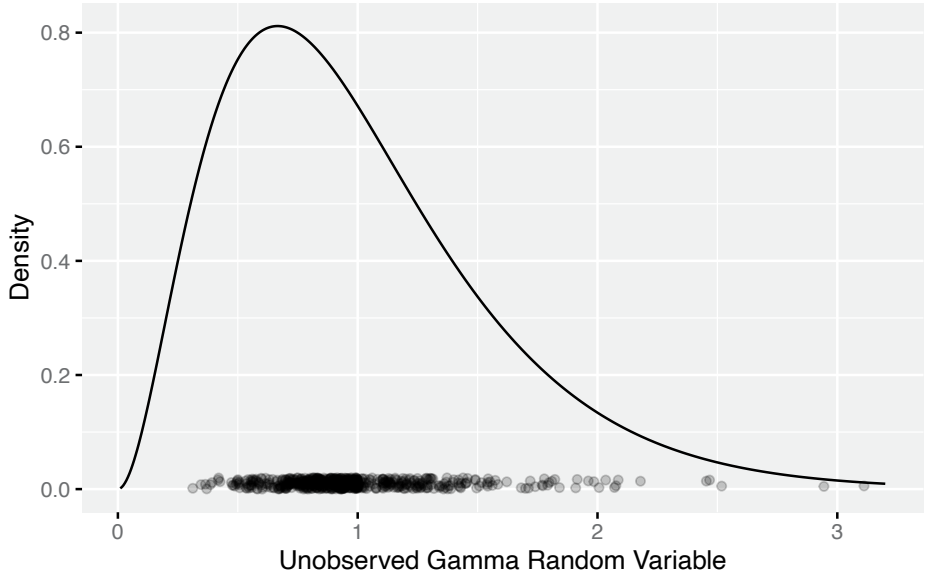


Table 6.10. Empirical annual frequency and estimated random effect for the top-five companies ranked on the size of their random effect.

Company	Number of Claims	Exposure (in days)	Annual Frequency	Random Effect
561	37	13,615	0.992	3.112
559	106	59,833	0.647	2.944
301	26	17,216	0.551	2.518
535	41	29,378	0.509	2.467
406	12	4,709	0.930	2.454

empirical annual frequency, and the estimated random effect. Note that for these companies the empirical frequency is quite large, given that the portfolio annual frequency is 0.228, and so we would expect them to have large estimated random effects.

Based on this GLMM, the *annual frequency* would be calculated via the formula

$$\mu_F = \exp\Big(\text{intercept} + \text{zone} + \text{bus.age} + \log\big(365\big)\Big) \times \big(\text{random effect of company}\big).$$

The first term in the above expression represents all the combinations of the fixed effects. The last term is the random effect for the company. Table 6.11 shows the annual frequency for each combination of geographic zone and bus age category based on the fixed effects, and Table 6.12 shows 30 companies along with their estimated random effects. The companies have been sorted by the magnitude of the random effects, and the table displays the 10 smallest and largest random effects, as well as 10 entries from the middle. Figure 6.16 shows a scatterplot of all companies by their total exposure and estimated random effects along with their total claim count. Companies with large random effects tend to have worse claims experience, that is, lower exposure and large numbers of claims. Companies with random effects below 1 tend to have better claims experience.

Table 6.11. Estimated annual frequency based only on the fixed effects from the GLMM.

	Bus Age Category				
Zone	0	1	2	3	4
1	0.365	0.360	0.347	0.313	0.218
2	0.452	0.446	0.431	0.389	0.270
3	0.301	0.296	0.286	0.258	0.179
4	0.291	0.287	0.277	0.250	0.174
5	0.388	0.382	0.369	0.333	0.232
6	0.358	0.354	0.342	0.308	0.214
7	0.163	0.161	0.155	0.140	0.097

Table 6.12. Estimated company random effects. Companies have been sorted from smallest to largest random effects.

Smallest		Medium		Largest	
Company	Random Effect	Company	Random Effect	Company	Random Effect
524	0.312	478	0.926	271	2.034
136	0.346	587	0.929	169	2.071
289	0.365	154	0.929	297	2.077
368	0.370	314	0.930	548	2.086
549	0.385	91	0.931	226	2.179
285	0.393	665	0.932	406	2.454
279	0.418	526	0.932	535	2.467
183	0.420	5	0.932	301	2.518
522	0.426	437	0.932	559	2.944
373	0.473	184	0.932	561	3.112

Figure 6.16. Total company exposure and claim counts along with the estimated company random effect.

References

Agresti, A. 2002. *Categorical Data Analysis*. 2nd ed. Wiley Series in Probability and Statistics. Wiley-Interscience.

Allaire, J. J., C. Teague, C. Scheidegger, Y. Xie, C. Dervieux, and G. Woodhull. 2024. "Quarto." https://doi.org/10.5281/zenodo.5960048.

Antoniadis, A., I. Gijbels, S. Lambert-Lacroix, and J.-M. Poggi. 2016. "Joint Estimation and Variable Selection for Mean and Dispersion in Proper Dispersion Models." *Electronic Journal of Statistics* 10 (1). https://doi.org/10.1214/16-EJS1152.

Bailey, R. A. 1963. "Insurance Rates with Minimum Bias." *Proceedings of the Casualty Actuarial Society* 50: 4–14.

Bates, D., M. Mächler, B. Bolker, and S. Walker. 2015. "Fitting Linear Mixed-Effects Models Using lme4." *Journal of Statistical Software* 67 (1): 1–48. https://doi.org/10.18637/jss.v067.i01.

Baxter, L. A., and S. M. Coutts. 1977. "An Analysis of Motor Insurance Claims Data." Presented at the 13th ASTIN Colloquium, Washington, DC, June.

Baxter, L. A., S. M. Coutts, and S. A. F. Ross. 1980. "Applications of Linear Models in Motor Insurance." In *Transactions of the 21st International Congress of Actuaries, Zurich and Lausanne, 19th-26, 1980*. Vol. 2. ICA.

Bissell, A. F. 1972. "A Negative Binomial Model with Varying Element Sizes." *Biometrika* 59 (2): 435–41. https://doi.org/10.2307/2334588.

Brockman, M. J., and T. S. Wright. 1992. "Statistical Motor Rating: Making Effective Use of Your Data." *Journal of the Institute of Actuaries (1886–1994)* 119 (3): 457–543. https://www.jstor.org/stable/41141077.

Bühlmann, H. 1967. "Experience Rating and Credibility." *ASTIN Bulletin* 4 (3): 199–207. https://doi.org/10.1017/S0515036100008989.

Bühlmann, H., and A. Gisler. 1997. "Credibility in the Regression Case Revisited (A Late Tribute to Charles A. Hachemeister)." *ASTIN Bulletin* 27 (1): 83–98.

Bühlmann, H., and A. Gisler. 2005. *A Course in Credibility Theory and Its Applications*. Springer.

Bühlmann, H., and E. Straub. 1970. "Glaubwürdigkeit Für Schadensätze." *Bulletin of the Swiss Association of Actuaries* 70 (1): 111–33.

Charpentier, A., ed. 2015. *Computational Actuarial Science with R*. Chapman and Hall/CRC.

Cleveland, W. S. 1979. "Robust Locally Weighted Regression and Smoothing Scatterplots." *Journal of the American Statistical Association* 74 (368): 829–36. https://doi.org/10.1080/01621459.1979.10481038.

Coutts, S. M. 1975. "Some Methods of Predicting the Number of Motor Car Accidents." Master's thesis, University College London.

Coutts, S. M. 1983. "An Actuarial Approach to Motor Insurance Rating." PhD diss., The City University.

Danneburg, D. R. 1996. "Basic Actuarial Credibility Models: Evaluation and Extension." PhD diss., Amsterdam.

De Vylder, F. 1981. "Regression Model with Scalar Credibility Weights." *Bulletin of Swiss Ass. of Act.* 1: 27–39.

De Vylder, F. 1985. "Non-Linear Regression in Credibility Theory." *Insurance: Mathematics and Economics* 4 (3): 163–72. https://doi.org/10.1016/0167-6687(85)90012-5.

Dunn, P. K., and G. K. Smyth. 1996. "Randomized Quantile Residuals." *Journal of Computational and Graphical Statistics* 5 (3): 236–44.

Dunn, P. K., and G. K. Smyth. 2018. *Generalized Linear Models with Examples in R*. Springer. https://doi.org/10.1007/978-1-4419-0118-7.

Dunn, P. K., and G. K. Smyth. 2022. *GLMsData: Generalized Linear Model Data Sets*. Manual. https://CRAN.R-project.org/package=GLMsData.

Dutang, C., and A. Charpentier. 2020. "CASdatasets: Insurance Datasets." https://github.com/dutangc/CASdatasets.

Dutang, C., V. Goulet, and M. Pigeon. 2008. "Actuar: An R Package for Actuarial Science." *Journal of Statistical Software* 25 (7): 1–37. https://doi.org/10.18637/jss.v025.i07.

Efron, B., T. Hastie, I. Johnstone, and R. Tibshirani. 2004. "Least Angle Regression." *Annals of Statistics* 32 (2): 407–51. https://www.jstor.org/stable/3448465.

Frees, E. W. 2003. "Multivariate Credibility for Aggregate Loss Models." *North American Actuarial Journal* 7 (1): 13–37. https://doi.org/10.1080/10920277.2003.10596074.

Frees, E. W., and P. Wang. 2005. "Credibility Using Copulas." *North American Actuarial Journal* 9 (2): 31–48. https://doi.org/10.1080/10920277.2005.10596196.

Frees, E. W., V. R. Young, and Y. Luo. 1999. "A Longitudinal Data Analysis Interpretation of Credibility Models." *Insurance: Mathematics and Economics* 24 (3): 229–47.

Gelman, A., and J. Hill. 2007. *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Analytical Methods for Social Research. Cambridge University Press.

Genz, A., and F. Bretz. 2009. *Computation of Multivariate Normal and t Probabilities*. Lecture Notes in Statistics. Springer-Verlag.

Hachemeister, C. A. 1975. "Credibility for Regression Models with Application to Trend." In *Credibility: Theory and Applications*, edited by P. M. Kahn. Academic Press.

Hastie, T., and B. Efron. 2022. *Lars: Least Angle Regression, Lasso and Forward Stagewise*. Manual. https://CRAN.R-project.org/package=lars.

Herzog, T. N. 2010. *Introduction to Credibility Theory*. 4th ed. ACTEX Academic Series. ACTEX Publications.

Hickman, J. C., and L. Heacox. 1999. "Credibility Theory: The Cornerstone of Actuarial Science." *North American Actuarial Journal* 3 (2): 1–8. https://doi.org/10.1080/10920277.1999.10595793.

Hilbe, J. M. 2007. *Negative Binomial Regression*. Cambridge University Press.

Jewell, W. S. 1975. "The Use of Collateral Data in Credibility Theory: A Hierarchical Model." Research Memorandum RM-75-024. International Institute for Applied Systems Analysis. https://pure.iiasa.ac.at/492.

Kaas, R., ed. 2009. *Modern Actuarial Risk Theory: Using R*. 2nd ed. Springer.

Klugman, S. A., H. H. Panjer, and G. E. Willmot. 1998. *Loss Models: From Data to Decisions*. Wiley Series in Probability and Statistics. Wiley.

Laird, N. M., and J. H. Ware. 1982. "Random-Effects Models for Longitudinal Data." *Biometrics* 38 (4): 963–74. https://doi.org/10.2307/2529876.

Lee, Y., and J. A. Nelder. 1996. "Hierarchical Generalized Linear Models." *Journal of the Royal Statistical Society. Series B (Methodological)* 58 (4): 619–78. https://www.jstor.org/stable/2346105.

Lee, Y., M. Molas, and M. Noh. 2018. *Mdhglm: Multivariate Double Hierarchical Generalized Linear Models*. Manual. https://CRAN.R-project.org/package=mdhglm.

Lee, Y., J. A. Nelder, and Y. Pawitan. 2021. *Generalized Linear Models with Random Effects: Unified Analysis via H-likelihood*. 2nd ed. Monographs on Statistics and Applied Probability 153. CRC Press.

Lee, Y., and M. Noh. 2018. *Dhglm: Double Hierarchical Generalized Linear Models*. Manual. https://CRAN.R-project.org/package=dhglm.

Lee, Y., L. Rönnegård, and M. Noh. 2020. *Data Analysis Using Hierarchical Generalized Linear Models with R*. CRC Press.

McCullagh, P., and J. A. Nelder. 1989. *Generalized Linear Models*. 2nd ed. Monographs on Statistics and Applied Probability 37. Chapman and Hall. https://doi.org/10.1201/9780203753736.

Mildenhall, S. J. 1999. "A Systematic Relationship Between Minimum Bias and Generalized Linear Models." *Proceedings of the Casualty Actuarial Society* 86: 393–487.

Mowbray, A. H. 1914. "How Extensive a Payroll Exposure Is Necessary to Give a Dependable Pure Premium?" *Proceedings of the Casualty Actuarial Society* 1 (1): 36–42. https://www.casact.org/sites/default/files/database/proceed_proceed14_1914.pdf.

Nelder, J. A. 1975. "Announcement by the Working Party on Statistical Computing: GLIM (Generalized Linear Interactive Modelling Program)." *Journal of the Royal Statistical Society. Series C (Applied Statistics)* 24 (2): 259–61. https://www.jstor.org/stable/2346575.

Nelder, J. A., and R. W. M. Wedderburn. 1972. "Generalized Linear Models." *Journal of the Royal Statistical Society. Series A (General)* 135 (3): 370–84. https://doi.org/10.2307/2344614.

Nelder, J. A, Y. Lee, B. Bergman, A. Hynén, A. F. Huele, and J. Engel. 1998. "Joint Modeling of Mean and Dispersion." *Technometrics* 40 (2): 168–75.

Ohlsson, E., and B. Johansson. 2010. *Non-Life Insurance Pricing with Generalized Linear Models*. EAA Lecture Notes. Springer.

Pedersen, T. L. 2024. *Patchwork: The Composer of Plots*. Manual. https://CRAN.R-project.org/package=patchwork.

R Core Team. 2024. *R: A Language and Environment for Statistical Computing*. Manual. R Foundation for Statistical Computing. https://www.R-project.org/.

Rempala, G. A., and R. A. Derrig. 2005. "Modeling Hidden Exposures in Claim Severity via the Em Algorithm." *North American Actuarial Journal* 9 (2): 108–28. https://doi.org/10.1080/10920277.2005.10596206.

Rigby, R. A., and M. D. Stasinopoulos. 2005. "Generalized Additive Models for Location, Scale, and Shape (with Discussion)." *Applied Statistics* 54 (3): 507–54.

Rönnegård, L., X. Shen, and M. Alam. 2010. "Hglm: A Package for Fitting Hierarchical Generalized Linear Models." *The R Journal* 2 (2): 20–28. https://journal.r-project.org/archive/2010-2/RJournal_2010-2_Roennegaard~et~al.pdf.

Schloerke, B., D. Cook, J. Larmarange, F. Briatte, et al. 2024. *GGally: Extension to 'Ggplot2'*. Manual. https://CRAN.R-project.org/package=GGally.

Straub, E. 1997. *Non-Life Insurance Mathematics*. 2nd ed., corr. print. Springer; Swiss Association of Actuaries.

Tager, I. B., S. T. Weiss, A. Muñoz, B. Rosner, and F. E. Speizer. 1983. "Longitudinal Study of the Effects of Maternal Smoking on Pulmonary Function in Children." *New England Journal of Medicine* 309 (12): 699–703.

Tager, I. B., S. T. Weiss, B. Rosner, and F. E. Speizer. 1979. "Effect of Parental Cigarette Smoking on the Pulmonary Function of Children." *American Journal of Epidemiology* 110 (1): 15–26. https://doi.org/10.1093/oxfordjournals.aje.a112783.

Venables, W. N., and B. D. Ripley. 2002. *Modern Applied Statistics with S*. 4th ed. Springer. https://www.stats.ox.ac.uk/pub/MASS4/.

Venter, G. G. 1996. "Credibility." In *Foundations of Casualty Actuarial Science*. 3rd ed. Casualty Actuarial Society.

Wickham, H., M. Averick, J. Bryan, W. Chang, et al. 2019. "Welcome to the tidyverse." *Journal of Open Source Software* 4 (43): 1686. https://doi.org/10.21105/joss.01686.

Wolny–Dominiak, A., and M. Trzesiok. 2014. *insuranceData: A Collection of Insurance Datasets Useful in Risk Classification in Non-Life Insurance.* Manual. https://CRAN.R-project.org/package=insuranceData.

Zhu, H. 2024. *kableExtra: Construct Complex Table with 'Kable' and Pipe Syntax*. Manual. https://CRAN.R-project.org/package=kableExtra.

APPENDICES

A. Bühlmann–Straub Simulation

In the Bühlmann–Straub model, the observation for risk class j in time period t, X_{jt}, can be decomposed as

$$X_{jt} = m + \Xi_j + \epsilon_{jt},$$

where m is the overall average and $j = 1, 2, \ldots, J$ and $t = 1, 2, \ldots, T + 1$. The unobservable component Ξ_j represents deviations from the overall mean m for risk class j, and we assume that they are independent and identically distributed with mean zero and $\text{Var}[\Xi_j] = \tau^2$. The component ϵ_{jt} represents deviations across time from the long-term average of risk class j, that is, $m + \Xi_j$. We also assume that they are independent and identically distributed with variance given by

$$\text{Var}\left[\epsilon_{jt}\right] = \frac{\sigma^2}{w_{jt}},$$

where w_{jt} are weights.

The Bühlmann–Straub model assumes that only the first and second moments exist. To simulate data from their model, we will assume normal distributions for the two random components: Ξ_j and ϵ_{jt}. We will also assume that the weights w_{jt} are uniformly distributed in an interval.

The code below uses the abbreviations in Table A.1.

The following function, `sim.BS()`, simulates a Bühlmann–Straub dataset with given parameter values and returns a data frame.

```
sim.BS <- function(
    sim.label = "A",
    J = 100,
    N = 5,
    beta = 80,
    sigma.b.sq = 64,
    sigma.sq = 100,
```

Table A.1. Code constructs and corresponding mathematical notation.

Code	Notation	Comments
`J`	J	Number of risk classes
`N`	$T+1$	Number of time periods
`m`	m	Overall average
`sigma.b.sq`	τ^2	Between-risk variance
`sigma.sq`	σ^2	Within-risk variance
`weight.min`		Minimum weight
`weight.spread`		Length of interval for weights
`risk.dev`	Ξ_j	Risk deviation from m
`time.dev`	ϵ_{jt}	Risk deviation from $m+\Xi_j$
`weight`	w_{jt}	Weights

```
    weight.min = 0.5,
    weight.spread = 1) {
  risk.dev <- rep(rnorm(J, mean = 0, sd = sqrt(sigma.b.sq)),
                  each = N)
  time.dev <- rnorm(J * N, mean = 0, sd = sqrt(sigma.sq))
  weight <- weight.min + runif(J * N, min = 0,
                               max = weight.spread)
  tb <- tibble(
    sim.label = rep(sim.label, J * N),
    risk = factor(rep(1:J, each = N)),
    Wt = weight,
    rsk.dev = risk.dev,
    tme.dev = time.dev,
    Y = beta + rsk.dev + tme.dev
  )
  return(tb)
}
```

With this function we can generate four different datasets with a different number of observations per risk, $N = 5, 10, 20, 40$, and store them in a list.

```
set.seed(398845)
BS.data <- list(
  BS.5  = sim.BS(sim.label = "5 Obs. per Risk", N = 5),
  BS.10 = sim.BS(sim.label = "10 Obs. per Risk", N = 10),
  BS.20 = sim.BS(sim.label = "20 Obs. per Risk", N = 20),
  BS.40 = sim.BS(sim.label = "40 Obs. per Risk", N = 40))
```

To each dataset we will fit a linear mixed-effects model and store the fitted models in a list.

```
BS.models <- map(BS.data,
                 \(d) lmer(Y ~ 1 + (1 | risk),
                           data = d,
                           weights = Wt))
```

Next, we compute the fitted values and standardized residuals for each model and append them to the dataset. We also calculate the slopes of the fitted linear regression where the response variable is the standardized residual and the predictor variable is fitted value.

```
BS.FV.Res <- map(BS.models,
                 \(m) {
                   tb <- getData(m)
                   tb$mu <- fitted(m)
                   tb$rsP <- resid(m, type = "pearson") /
                                   sigma(m)
                   return(tb)})

BS.slopes <- map_dbl(BS.FV.Res,
                     \(tb) {
                       fm <- lm(rsP ~ mu,
                                data = tb)
                       sfm <- summary(fm)
                       return(coef(sfm)[2,1])})

BS.pvals <- map_dbl(BS.FV.Res,
                    \(tb) {
                      fm <- lm(rsP ~ mu,
                               data = tb)
                      sfm <- summary(fm)
                      return(coef(sfm)[2,4])})
```

We collect all the information into a single data frame and create a categorical variable to identify the simulation.

```
BS.results <- reduce(BS.FV.Res, bind_rows)
BS.results$sim.label <- factor(BS.results$sim.label)
BS.results$sim.label <- fct_relevel(
  BS.results$sim.label,
  str_c(str_sub(names(BS.slopes), 4),
        " Obs. per Risk")[order(BS.slopes,
                                decreasing = TRUE)])
```

Figure A.1. Results of fitting a linear regression model to simulated data from the Bühlmann–Straub model. The panels have been arranged from the largest slope in the upper left (5 obs. per risk) to the smallest slope in the lower right (40 obs. per risk).

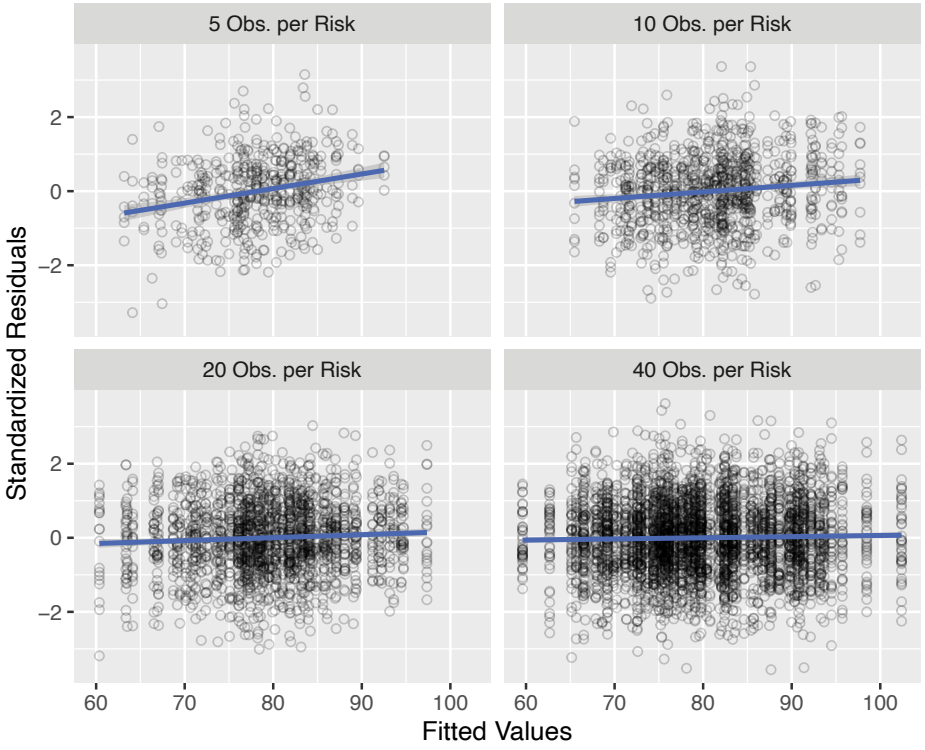


Figure A.1 shows the results, where we can see that as the number of observations increases, the slope of the line decreases. For this particular set of simulated data, the slopes for the 5, 10, and 20 observations per risk sets are statistically different from zero but not for the 40 observations per risk set.

```
tb <- round(rbind(BS.slopes,
      BS.pvals), 4)
dimnames(tb) <- list(c("Slope", "P-Value"),
                     c("5 Obs.", "10 Obs.", "20 Obs.",
                       "40 Obs."))
tb
```

```
         5 Obs.  10 Obs.  20 Obs.  40 Obs.
Slope    0.0393   0.0177    0.008   0.0031
P-Value  0.0000   0.0001    0.007   0.0889
```

B. Equivalence of Credibility Matrices

In Chapter 3 we found an expression for the credibility matrix A_j as shown in Equation 3.27, and in Section 4.4, using linear mixed model theory, we found another expression, Equation 4.2. We want to show here that these two seemingly different expressions are equivalent.

Equation 4.2 says that, assuming the matrices D and W_j have the forms

$$D=\begin{bmatrix}\tau_0^2 & 0\\ 0 & \tau_1^2\end{bmatrix} \quad \text{and} \quad W_j=\begin{bmatrix}w_{j\bullet} & \sum_t t w_{jt}\\ \sum_t t w_{jt} & \sum_t t^2 w_{jt}\end{bmatrix},$$

then the credibility matrix A_j is given by

$$A_j=\frac{\det\left(DW_j\right)I_2+\hat{\sigma}^2 DW_j}{\det\left(DW_j\right)+\hat{\sigma}^2\,\text{trace}\left(DW_j\right)+\hat{\sigma}^4}. \tag{B.1}$$

We want to show that this is equivalent to Equation 3.27, that is,

$$A_j=\frac{w_{j\bullet}}{d}\begin{bmatrix}w_{j\bullet}\text{Var}_j^{(s)}\left[t\right]+\kappa_1 & \kappa_1 E_j^{(s)}\left[t\right]\\ \kappa_0 E_j^{(s)}\left[t\right] & w_{j\bullet}\text{Var}_j^{(s)}\left[t\right]+\kappa_0 E_j^{(s)}\left[t^2\right]\end{bmatrix}, \tag{B.2}$$

where

$$d=\left(w_{j\bullet}+\kappa_0\right)\left(w_{j\bullet}E_j^{(s)}\left[t^2\right]+\kappa_1\right)-\left(w_{j\bullet}E_j^{(s)}\left[t\right]\right)^2$$

and $\kappa_0=\sigma^2/\tau_0^2$ and $\kappa_1=\sigma^2/\tau_1^2$.

The notations $E_j^{(s)}[t]$ and $\text{Var}_j^{(s)}[t]$ are shorthand for

$$E_j^{(s)}\left[t\right]=\sum_t \frac{w_{jt}}{w_{j\bullet}}t \quad \text{and} \quad \text{Var}_j^{(s)}\left[t\right]=E_j^{(s)}\left[t^2\right]-\left(E_j^{(s)}\left[t\right]\right)^2.$$

Equivalence of the Denominator

We will start by showing that the denominator of Equation B.1 is equal to d in Equation B.2.

The matrix DW_j is equal to

$$\begin{bmatrix} \tau_0^2 w_{j\bullet} & \tau_0^2 \sum_t t w_{jt} \\ \tau_1^2 \sum_t t w_{jt} & \tau_1^2 \sum_t t^2 w_{jt} \end{bmatrix}$$

and hence its determinant is

$$\det\left(DW_j\right) = \tau_0^2 \tau_1^2 w_{j\bullet} \sum_t t^2 w_{jt} - \tau_0^2 \tau_1^2 \left(\sum_t t w_{jt}\right)^2,$$

and its trace is

$$\mathrm{tr}\left(DW_j\right) = \tau_0^2 w_{j\bullet} + \tau_1^2 \sum_t t^2 w_{jt}.$$

Therefore, the denominator is equal to

$$\tau_0^2 \tau_1^2 w_{j\bullet} \sum_t t^2 w_{jt} - \tau_0^2 \tau_1^2 \left(\sum_t t w_{jt}\right)^2 + \sigma^2 \tau_0^2 w_{j\bullet} + \sigma^2 \tau_1^2 \sum_t t^2 w_{jt} + \sigma^4.$$

In the above expression, we can replace the sums with the $E_j^{(s)}$ notation to obtain

$$\tau_0^2 \tau_1^2 w_{j\bullet}^2 E_j^{(s)}\left[t^2\right] - \tau_0^2 \tau_1^2 w_{j\bullet}^2 \left(E_j^{(s)}\left[t\right]\right)^2 + \sigma^2 \tau_0^2 w_{j\bullet} + \sigma^2 \tau_1^2 w_{j\bullet} E_j^{(s)}\left[t^2\right] + \sigma^4.$$

Next, we factor out $\tau_0^2\tau_1^2$, yielding

$$\tau_0^2 \tau_1^2 \left[w_{j\bullet}^2 E_j^{(s)}\left[t^2\right] - \left(w_{j\bullet} E_j^{(s)}\left[t\right]\right)^2 + \kappa_1 w_{j\bullet} + \kappa_0 w_{j\bullet} E_j^{(s)}\left[t^2\right] + \kappa_0 \kappa_1 \right],$$

and now we just rearrange like terms to get

$$\tau_0^2 \tau_1^2 \left[\left(w_{j\bullet} + \kappa_0\right) \left\{ w_{j\bullet} E_j^{(s)}\left[t^2\right] + \kappa_1 \right\} - \left(w_{j\bullet} E_j^{(s)}\left[t\right]\right)^2 \right].$$

Apart from the factor $\tau_0^2\tau_1^2$, this is the same expression for the denominator in Equation B.2 that we wanted to show. The factor $\tau_0^2\tau_1^2$ will also appear in our calculations of the numerator, and thus it will cancel out.

Equivalence of the Numerator

The numerator in Equation B.1 and Equation B.2 is a 2×2 matrix, and so we will show the equivalence of the (1, 1), (1, 2), and (2, 2) entries. The entry at (2, 1) is similar to the (1, 2) entry.

The (1, 1) entry in Equation B.1 is equal to

$$\tau_0^2\tau_1^2 w_{j\bullet}\sum_t t^2 w_{jt} - \tau_0^2\tau_1^2\left(\sum_t t w_{jt}\right)^2 + \sigma^2\tau_0^2 w_{j\bullet}.$$

Again we factor out $\tau_0^2\tau_1^2$ to get

$$\tau_0^2\tau_1^2\left[w_{j\bullet}\sum_t t^2 w_{jt} - \left(\sum_t t w_{jt}\right)^2 + \kappa_1 w_{j\bullet}\right].$$

Next, we multiply the first factor inside the square brackets by $w_{j\bullet}/w_{j\bullet}$ and the second factor by $w_{j\bullet}^2/w_{j\bullet}^2$. This will allow us to rewrite these two items using $\mathrm{Var}_j^{(s)}[t]$ as follows

$$\tau_0^2\tau_1^2\left[w_{j\bullet}^2\,\mathrm{Var}_j^{(s)}[t] + \kappa_1 w_{j\bullet}\right].$$

Now we factor out $w_{j\bullet}$ to obtain our final result

$$\tau_0^2\tau_1^2 w_{j\bullet}\left[w_{j\bullet}\,\mathrm{Var}_j^{(s)}[t] + \kappa_1\right],$$

which apart from the $\tau_0^2\tau_1^2$ factor matches the (1, 1) entry in Equation B.2.

For the (1, 2) entry we start with

$$\sigma^2\tau_0^2\sum_t t w_{jt}.$$

Multiply this expression by $w_{j\bullet}/w_{j\bullet}$ to obtain

$$\sigma^2\tau_0^2 w_{j\bullet}\sum_j t\frac{w_{jt}}{w_{j\bullet}} = \sigma^2\tau_0^2 w_{j\bullet}E_j^{(s)}[t] = \tau_0^2\tau_1^2 w_{j\bullet}\left[\kappa_1 E_j^{(s)}[t]\right].$$

Again, apart from the factor $\tau_0^2\tau_1^2$, this is the expression we wanted to arrive at.

Finally, for the (2, 2) entry we begin with

$$\tau_0^2\tau_1^2 w_{j\bullet}\sum_t t^2 w_{jt} - \tau_0^2\tau_1^2\left(\sum_t t w_{jt}\right)^2 + \sigma^2\tau_1^2\sum_t t^2 w_{jt}.$$

Multiplying the first and the last terms by $w_{j\bullet}/w_{j\bullet}$ and the second term by $w_{j\bullet}^2/w_{j\bullet}^2$ and simplifying by using the $E_j^{(s)}[\,]$ notation, we get

$$\tau_0^2\tau_1^2 w_{j\bullet}^2 E_j^{(s)}[t^2] - \tau_0^2\tau_1^2 w_{j\bullet}^2\left(E_j^{(s)}[t]\right)^2 + \sigma^2\tau_1^2 w_{j\bullet}^2 E_j^{(s)}[t^2].$$

Now, factor out $\tau_0^2\tau_1^2 w_{j\bullet}$ to arrive at

$$\tau_0^2\tau_1^2 w_{j\bullet}\left[w_{j\bullet}\mathrm{Var}_j^{(s)}[t] + \kappa_0 E_j^{(s)}[t^2]\right].$$

The factor $\tau_0^2\tau_1^2$ will cancel with the same factor we have in the denominator, and so we have shown the equivalence of both expressions for the credibility matrix A_j.

C. Lambert *W* Function

The Lambert W function is also known as the omega function or the product logarithm function. Consider the function $f(x) = xe^x$ with domain $x \in \mathbb{R}$ and range $f(x) \in [-e^{-1}, \infty)$. Figure C.1 shows a plot of $f(x)$ against x for $x \in [-4, 1]$, and note that as x moves toward negative infinity the value of $f(x)$ approaches zero from below. The function $f(x)$ has a minimum value at the point $(-1, -e^{-1})$.

The Lambert W function solves the equation $y = f(x)$ for x in terms of y—that is,

$$y = xe^x \quad \text{if and only if} \quad W(y) = x.$$

But in the interval $[-e^{-1}, 0)$ Lambert's function could take on two values, and we must choose which solution we want to use (see Figure C.2). The red-colored curve (with y values greater than or equal to -1) is known as the principal branch and is denoted by W_0. The blue-colored curve represents another possible solution. This branch is denoted by W_{-1}.

C.1. Mean for Zero-Truncated Poisson

In the hospital length of stay application described in Section 6.2, we determined that the mean of the zero-truncated distribution can be expressed as (see Equation 6.1)

$$\mu = \frac{\lambda e^{\lambda}}{e^{\lambda} - 1},$$

and we would like to solve this equation for λ in terms of μ. To that end, we rewrite the above equation as

$$\mu e^{\lambda} - \mu = \lambda e^{\lambda}$$

and now move the term μe^{λ} to the other side of the equation to obtain

$$-\mu = (\lambda - \mu) e^{\lambda}.$$

Multiply both sides by a factor of $e^{-\mu}$, yielding

$$-\mu e^{-\mu} = (\lambda - \mu) e^{\lambda - \mu}.$$

Figure C.1. The function $f(x) = xe^x$ in the interval from –4 to 1.

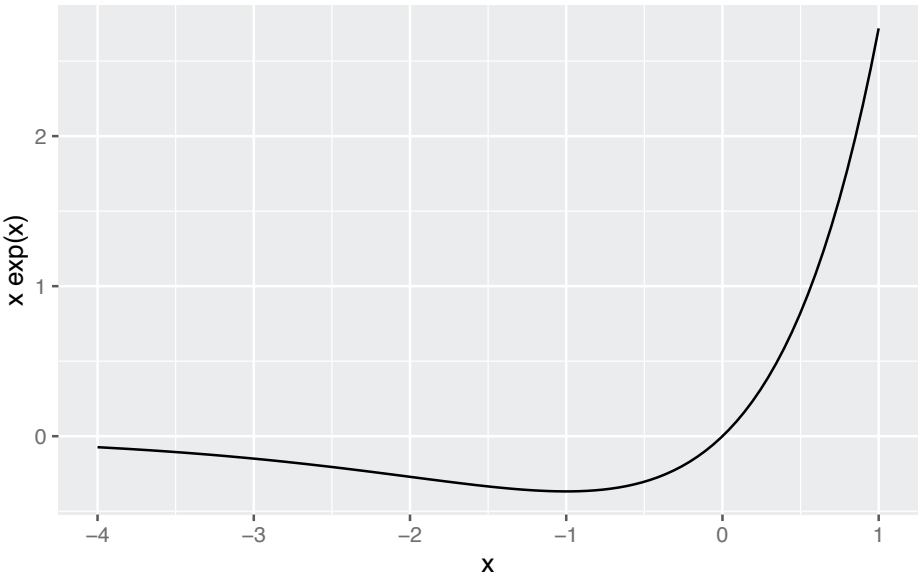
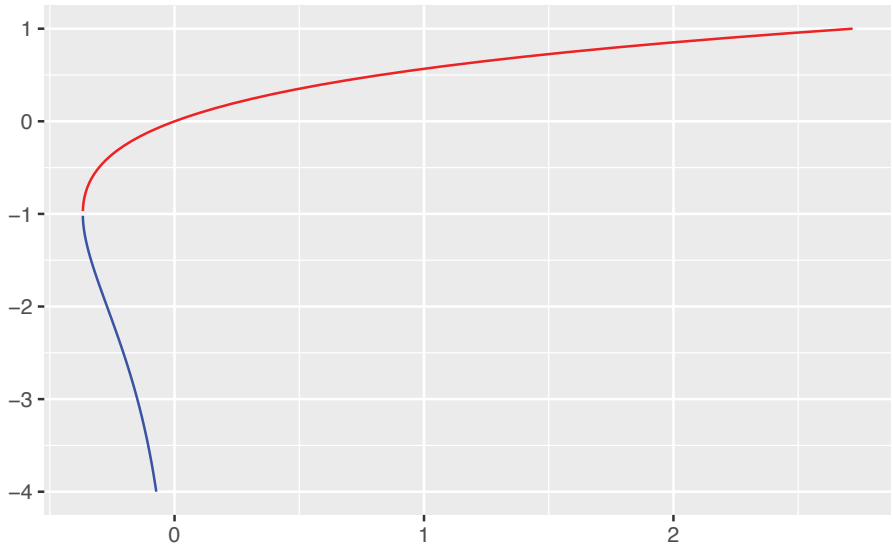


Figure C.2. The two branches of the Lambert W function, that is, the inverse of the function $f(x) = xe^x$. The red-colored curve is known as the principal branch and denoted as W_0, and the blue-colored curve is denoted as W_{-1}.



Notice that this last equation is of the form $z = we^w$, and we can apply Lambert's W_0 to both sides to obtain

$$W_0\left(-\mu e^{-\mu}\right)=\lambda-\mu,$$

and so we have that

$$\lambda=\mu+W_0\left(-\mu e^{-\mu}\right),$$

giving us the parameter λ of the distribution in terms of the mean of the distribution.

C.2. Variance Function for Zero-Truncated Poisson Distribution

To express the variance function of the zero-truncated Poisson distribution in terms of the mean, we start with the following expression for the variance (see Equation 6.2) in terms of λ:

$$\frac{\lambda e^{\lambda}\left[e^{\lambda}-\lambda-1\right]}{\left(e^{\lambda}-1\right)^2}.$$

We can rewrite this expression as

$$\frac{\lambda e^{\lambda}}{e^{\lambda}-1}\cdot\frac{e^{\lambda}-1-\lambda}{e^{\lambda}-1}=\mu\left[1-\frac{\lambda}{e^{\lambda}-1}\right],$$

where we have used the fact that $\mu = \lambda e^{\lambda}/(e^{\lambda} - 1)$. The second term inside the square brackets can be rewritten as

$$\frac{\lambda}{e^{\lambda}-1}=\frac{\mu}{e^{\lambda}}$$

by using the same relationship between μ and λ that we used in the previous equation. Therefore, our expression for the variance so far looks like

$$\mu\left[1-\frac{\mu}{e^{\lambda}}\right].$$

Next, we substitute $\lambda = \mu + W_0(-\mu e^{-\mu})$ and use the identity $e^{W_0(x)} = x/W_0(x)$ to get

$$\mu\left[1-\frac{\mu}{e^{\lambda}}\right]=\mu\left[1-\frac{\mu}{e^{\mu+W_0\left(-\mu e^{-\mu}\right)}}\right]=\mu\left[1-\frac{\mu W_0\left(-\mu e^{-\mu}\right)}{e^{\mu}\left(-\mu e^{-\mu}\right)}\right],$$

which simplifies to

$$\mu\left[1+W_0\left(-\mu e^{-\mu}\right)\right],$$

which is what we wanted to show, that is, the variance function expressed in terms of the mean of the distribution.

D. Bühlmann–Gisler Estimators

The functions `HBG()`, `sig.sq()`, and `tau()` defined below implement the estimation procedures provided in Chapter 8 of Bühlmann and Gisler (2005). These functions are used to implement Theorem 3.4.

Function `HBG()` (Hachemeister Bühlmann–Gisler) is the main function used to calculate credibility estimates. The function `sig.sq()` codes the formula for the estimator of $\hat{\sigma}$ (Equation 3.29), and the function `tau()` codes the estimator for $\hat{\tau}$ that combines the variance estimators for the intercept (Equation 3.30) and for the slope (Equation 3.32) into a diagonal matrix.

```
HBG <- function(sigma.sq,
                D,
                X.jt,
                T.jt,
                W.jt,
                state,
                use.B.gls = TRUE) {
  if (!is.factor(state)) {
    state <- factor(state)
  }
  J <- length(levels(state))
  W.jb <- tapply(W.jt, state, sum)
  Fj.t <- tapply(W.jt * T.jt, state, sum)
  Fj.t2 <- tapply(W.jt * T.jt^2, state, sum)
  Fj.X <- tapply(W.jt * X.jt, state, sum)
  Fj.tX <- tapply(W.jt * T.jt * X.jt, state, sum)
  I <- diag(1, nrow = 2, ncol = 2)

  W <- map(1:J, function(i) {
    ans <- matrix(c(W.jb[i], Fj.t[i],
                    Fj.t[i], Fj.t2[i]),
                  nrow = 2, ncol = 2,
                  byrow = TRUE)
    return(ans)
  })
```

```
sW <- reduce(W, '+')
M <- map(1:J, function(i) {
  ans <- matrix(c(Fj.X[i], Fj.tX[i]),
                nrow = 2, ncol = 1)
  return(ans)
})
sM <- reduce(M, '+')
xi <- map(1:J, function(i) {
  DW <- D %*% W[[i]]
  dt <- det(DW)
  tr <- sum(diag(DW))
  den <- dt + sigma.sq * tr + sigma.sq^2
  ans <- (dt * I + sigma.sq * DW) / den
  return(ans)
})
sxi <- reduce(xi, '+')
B <- map(1:J, function(i) {
  ans <- solve(W[[i]]) %*% M[[i]]
  return(ans)
})

WB <- pmap(list(xi, W, M), ~ ..1 %*% solve(..2) %*% ..3)
sWB <- reduce(WB, '+')

B.gls <- solve(sxi) %*% sWB
B.all <- solve(sW) %*% sM
B.col <- if (use.B.gls) B.gls else B.all
CW <- map2(xi, B, ~ .x %*% .y + (I - .x) %*% B.col)

tb <- cbind(as.matrix(rep(1:J, each = 2), ncol = 1),
            reduce(xi, rbind),
            reduce(B, rbind),
            reduce(CW, rbind),
            matrix(B.col, nrow = 2 * J))
dimnames(tb) <- list(NULL,
                     c("state", "CM.1", "CM.2",
                       "Standalone", "Credibility",
                       "Collective"))
tb <- as_tibble(tb)

ans <- list(sigma.sq = sigma.sq,
            D = D,
            dta = data.frame(X.jt = X.jt,
                             T.jt = T.jt,
                             W.jt, W.jt,
                             state = state),
```

```
                use.B.gls = use.B.gls,
                W = W,
                M = M,
                xi = xi,
                B = B,
                B.gls = B.gls,
                B.all = B.all,
                B.col = B.col,
                CW = CW,
                tb = tb)
  return(ans)
}
```

```
sig.sq <- function(X.jt, T.jt, W.jt, state) {
  if (!is.factor(state)) {
    state <- factor(state)
  }
  J <- length(levels(state))
  W.jb <- tapply(W.jt, state, sum)
  W.bb <- sum(W.jb)
  Fj.t <- tapply(W.jt * T.jt, state, sum)
  Fj.t2 <- tapply(W.jt * T.jt^2, state, sum)
  Fj.X <- tapply(W.jt * X.jt, state, sum)
  Fj.tX <- tapply(W.jt * T.jt * X.jt, state, sum)

  W <- map(1:J, function(i) {
    ans <- matrix(c(W.jb[i], Fj.t[i],
                    Fj.t[i], Fj.t2[i]),
                  nrow = 2, ncol = 2,
                  byrow = TRUE)
    return(ans)
  })
  M <- map(1:J, function(i) {
    ans <- matrix(c(Fj.X[i], Fj.tX[i]),
                  nrow = 2, ncol = 1)
    return(ans)
  })
  B <- map(1:J, function(i) {
    ans <- solve(W[[i]]) %*% M[[i]]
    return(ans)
  })

  Y <- map(tapply(T.jt, state, list),
           function(x) cbind(rep(1, length(x)), x))
  mu <- map2(Y, B, function(x,y) as.vector(x %*% y))
  w <- tapply(W.jt, state, list)
  x <- tapply(X.jt, state, list)
```

```
  sigmaj.sq <- pmap(list(w, x, mu),
                    ~ sum(..1 * (..2 - ..3)^2) /
                          (length(..1) - 2))
  sigma.sq <- reduce(sigmaj.sq, `+`) / length(sigmaj.sq)

  ans <- list(dta = data.frame(state = state,
                               X.jt = X.jt,
                               T.jt = T.jt,
                               W.jt = W.jt),
              W = W,
              M = M,
              B = B,
              Y = Y,
              mu = mu,
              w = w,
              x = x,
              sigmaj.sq = sigmaj.sq,
              sigma.sq = sigma.sq)
  return(ans)
}
```

```
tau <- function(sigma.sq, X.jt, T.jt, W.jt, state) {
  if (!is.factor(state)) {
    state <- factor(state)
  }
  J <- length(levels(state))
  W.jb <- tapply(W.jt, state, sum)
  W.bb <- sum(W.jb)
  Fj.t <- tapply(W.jt * T.jt, state, sum)
  Fj.t2 <- tapply(W.jt * T.jt^2, state, sum)
  Fj.X <- tapply(W.jt * X.jt, state, sum)
  Fj.tX <- tapply(W.jt * T.jt * X.jt, state, sum)
  Vj.t <- Fj.t2 / W.jb - (Fj.t / W.jb)^2
  Ws.jb <- Vj.t * W.jb
  Ws.bb <- sum(Ws.jb)

  W <- map(1:J, function(i) {
    ans <- matrix(c(W.jb[i], Fj.t[i],
                    Fj.t[i], Fj.t2[i]),
                  nrow = 2, ncol = 2,
                  byrow = TRUE)
    return(ans)
  })
```

```
  M <- map(1:J, function(i) {
    ans <- matrix(c(Fj.X[i], Fj.tX[i]),
                  nrow = 2, ncol = 1)
    return(ans)
  })
  B <- map(1:J, function(i) {
    ans <- solve(W[[i]]) %*% M[[i]]
    return(ans)
  })
  B0 <- map_dbl(B, function(x) x[1,1])
  B0.bar <- sum(W.jb * B0 / W.bb)
  c0 <- (J - 1) / (J * sum(W.jb / W.bb * (1 - W.jb / W.bb)))
  tau0.sq <- c0 * (J * sum(W.jb * (B0 - B0.bar)^2 / W.bb)
                   / (J - 1) - J * sigma.sq / W.bb)

  B1 <- map_dbl(B, function(x) x[2,1])
  B1.bar <- sum(Ws.jb * B1 / Ws.bb)
  c1 <- (J - 1) / (J * sum(Ws.jb / Ws.bb * (1 - Ws.jb / Ws.bb)))
  tau1.sq <- c1 * (J * sum(Ws.jb * (B1 - B1.bar)^2 / Ws.bb)
                   / (J - 1) - J * sigma.sq / Ws.bb)

  ans <- list(sigma.sq = sigma.sq,
              dta = data.frame(state = state,
                               X.jt = X.jt,
                               T.jt = T.jt,
                               W.jt = W.jt),
              W = W,
              M = M,
              B = B,
              Bj = rbind(B0, B1),
              B.bar = matrix(c(B0.bar, B1.bar), nrow = 2,
                             ncol = 1),
              c01 = c(c0, c1),
              D = diag(c(tau0.sq, tau1.sq)))
  return(ans)
}
```

ABOUT THE SERIES:

CAS monographs are authoritative, peer-reviewed, in-depth works focusing on important topics within property and casualty actuarial practice. For more information on the CAS Monograph Series, visit the CAS website at www.casact.org.

Expertise. Insight.
Solutions.

casact.org