

GLM I: Introduction to Generalized Linear Models

By Curtis Gary Dean
Distinguished Professor of Actuarial Science

Ball State University

Classical Multiple Linear Regression

$$\diamond Y_i = a_0 + a_1 X_{i1} + a_2 X_{i2} \dots + a_m X_{im} + e_i$$

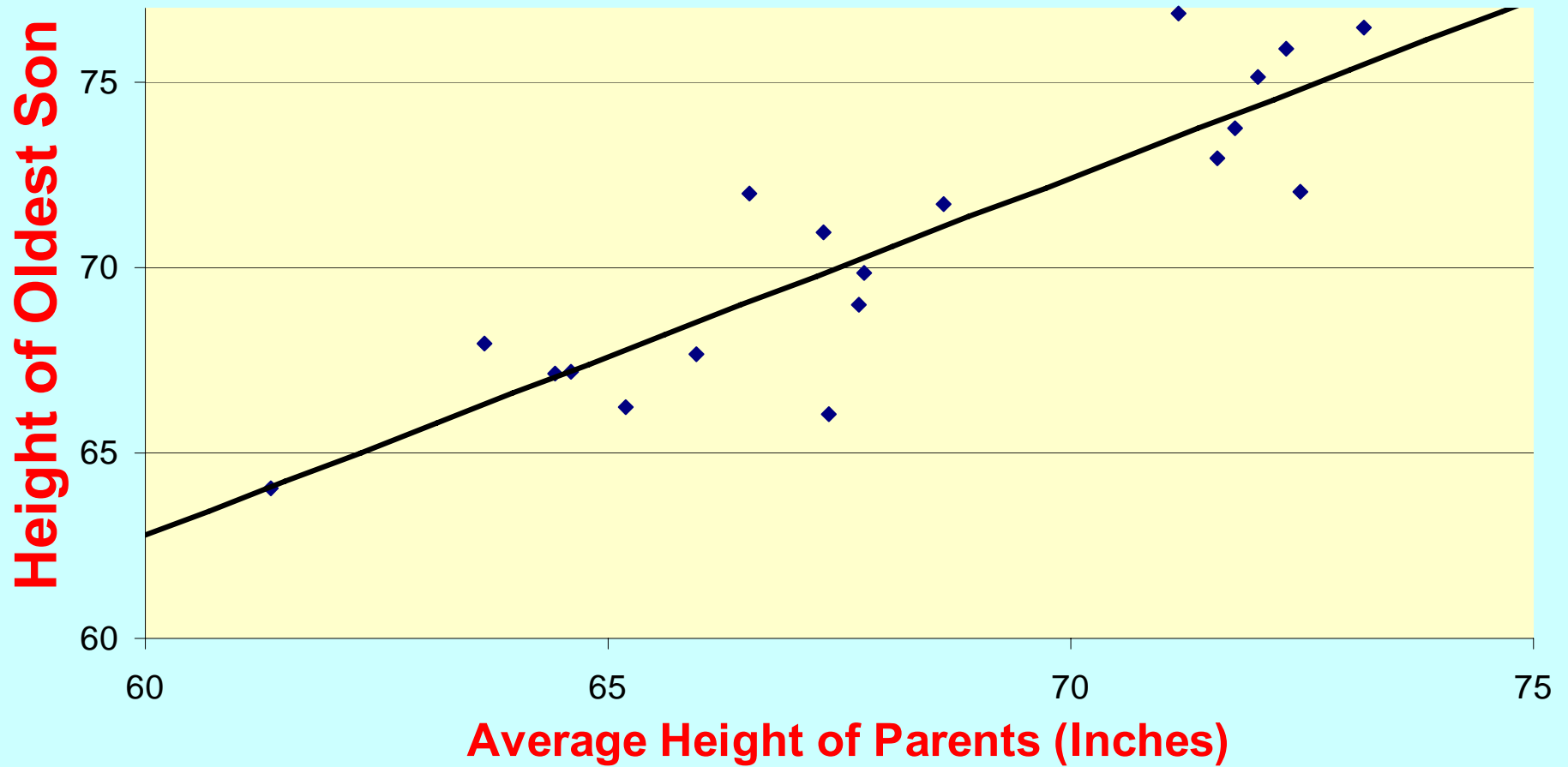
$\diamond Y_i$ are the response variables

$\diamond X_{ij}$ are predictors

$\diamond i$ subscript denotes i^{th} observation

$\diamond j$ subscript identifies j^{th} predictor

One Predictor: $Y_i = a_0 + a_1X_i$



Classical Multiple Linear Regression: Solving for a_i

Minimize $G(a_0, a_1, \dots, a_m) =$

$$\sum_{i=1}^n (Y_i - a_0 - a_1 X_{i1} - K - a_m X_{im})^2$$

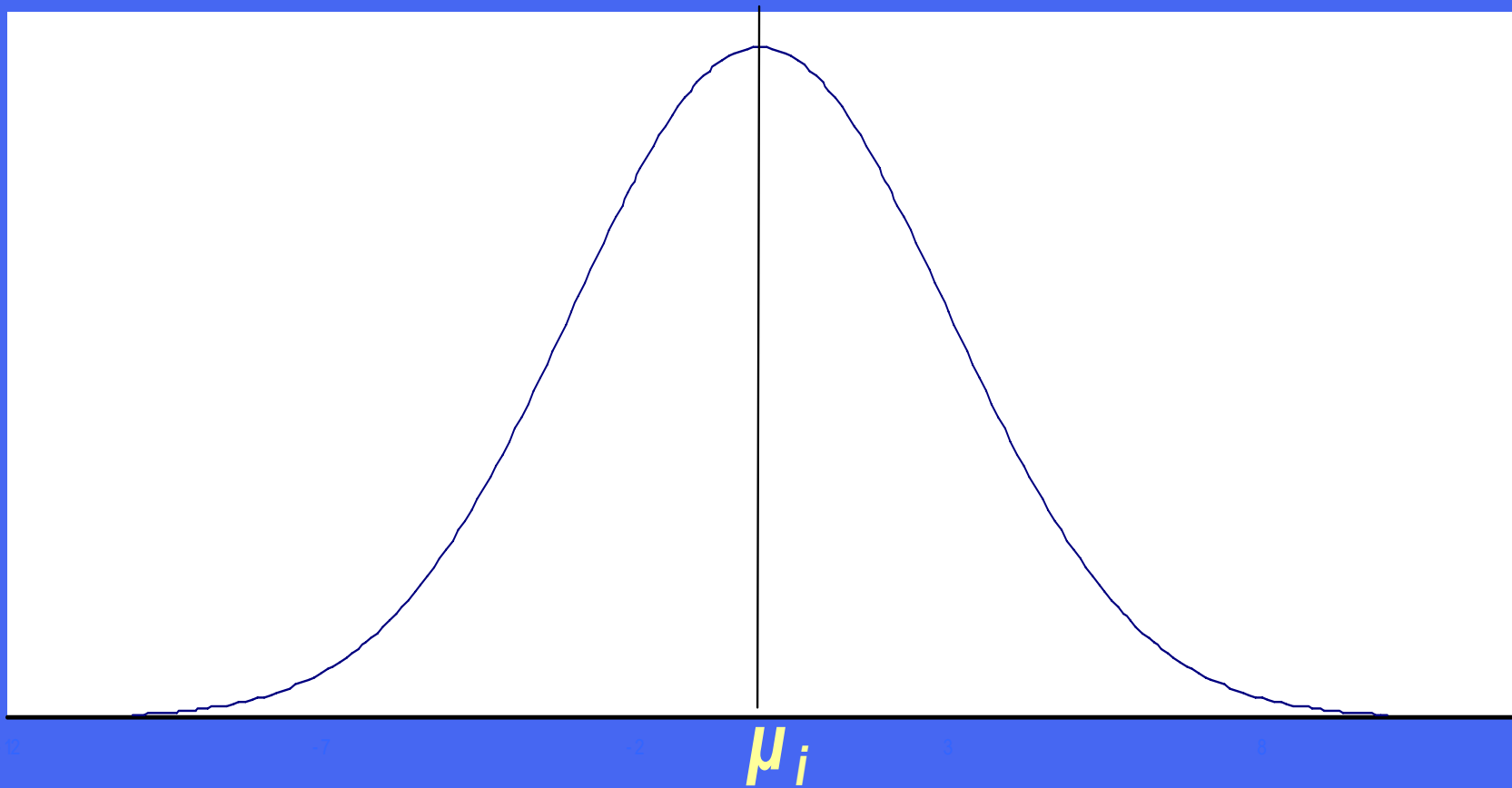
Classical Multiple Linear Regression

❖ $\mu_i = E[Y_i] = a_0 + a_1 X_{i1} + \dots + a_m X_{im}$

❖ Y_i is Normally distributed random variable with constant variance σ^2

❖ Want to estimate $\mu_i = E[Y_i]$ for each i

Response Y_i has Normal Distribution



Problems with Traditional Model

- ❖ Number of claims is discrete
- ❖ Claim sizes are skewed to the right
- ❖ Probability of an event is in $[0,1]$
- ❖ Variance is not constant across data points i
- ❖ Nonlinear relationship between X 's and Y 's

Generalized Linear Models - GLMs

- ❖ Fewer restrictions
- ❖ Y can model number of claims, probability of renewing, loss severity, loss ratio, etc.
- ❖ Large and small policies can be put into one model
- ❖ Y can be nonlinear function of X 's
- ❖ Classical linear regression model is a special case

Generalized Linear Models - GLMs

❖ Same goal

Predict : $\mu_i = E[Y_i]$

Generalized Linear Models - GLMs

❖ $g(\mu_i) = a_0 + a_1 X_{i1} + \dots + a_m X_{im}$

❖ $g(\cdot)$ is the link function

❖ $E[Y_i] = \mu_i = g^{-1}(a_0 + a_1 X_{i1} + \dots + a_m X_{im})$

▪ Y_i can be Normal, Poisson, Gamma, Binomial, Compound Poisson, ...

▪ Variance can be modeled

GLMs Extend Classical Linear Regression

❖ If link function is identity: $g(\mu_i) = \mu_i$

❖ And Y_i has Normal distribution

→ GLM gives same answer as
Classical Linear Regression*

* Least squares and MLE equivalent for Normal dist.

Exponential Family of Distributions – Canonical Form

$$f(y; \theta, \phi) = \exp \left[\frac{\{\theta \cdot y - b(\theta)\}}{a(\phi)} + c(y, \phi) \right]$$

$$E[Y] = b'(\theta)$$

$$\text{Var}[Y] = b''(\theta)a(\phi)$$

θ is the parameter of interest !

ϕ is often called a nuisance parameter.

Some Math Rules: Refresher

$$(1) \quad \exp(x) = e^x$$

$$(2) \quad x = \exp[\ln x] = \ln[\exp x]$$

$$(3) \quad \ln(xy) = \ln x + \ln y$$

$$(4) \quad \ln(x^r) = r \ln x$$

$$(5) \quad \ln(1/x) = \ln(x^{-1}) = -\ln x$$

$$(6) \quad \ln(x/y) = \ln x - \ln y$$

Normal Distribution in Exponential Family

$$\begin{aligned} f(y; \mu, \sigma^2) &= \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y-\mu)^2}{2\sigma^2}\right) \\ &= \exp\left(\ln\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)\right) \exp\left(-\frac{y^2 - 2\mu y + \mu^2}{2\sigma^2}\right) \\ &= \exp\left(\frac{\mu y - \mu^2 / 2}{\sigma^2} - \frac{y^2}{2\sigma^2} - \ln \sqrt{2\pi\sigma^2}\right) \end{aligned}$$

Normal Distribution in Exponential Family

$$f(y; \mu, \sigma^2) = \exp \left(\frac{\mu y - \mu^2 / 2}{\sigma^2} - \frac{y^2}{2\sigma^2} - \ln \sqrt{2\pi\sigma^2} \right)$$

Let $\theta = \mu$ and $a(\phi) = \sigma^2$,

then $b(\theta) = \theta^2 / 2 \rightarrow b'(\theta) = \theta = \mu$

and $Var[Y] = b''(\theta)a(\phi) = 1 \cdot \sigma^2 = \sigma^2$

Poisson Distribution in Exponential Family

$$\Pr[Y = y] = \frac{\mu^y e^{-\mu}}{y!}$$

$$\Pr[Y = y] = \exp\left\{\ln\left(\frac{\mu^y e^{-\mu}}{y!}\right)\right\}$$

θ

$$\Pr[Y = y] = \exp\left\{\frac{(\ln \mu) \cdot y - \mu}{1} - \ln(y!)\right\}$$

Poisson Distribution in Exponential Family

$$\theta = \ln \mu \rightarrow \mu = e^{\theta}$$

$$b(\theta) = \mu = e^{\theta} \quad \text{and} \quad a(\phi) = 1$$

$$E[Y] = b'(\theta) = \frac{d}{d\theta} e^{\theta} = e^{\theta} = \mu$$

$$\text{Var}[Y] = b''(\theta) a(\phi) = \frac{d^2}{d\theta^2} e^{\theta} = \mu$$

Compound Poisson Distribution

- ❖ $Y = C_1 + C_2 + \dots + C_N$
- ❖ N is Poisson random variable
- ❖ C_i are i.i.d. with Gamma distribution
- ❖ This is an example of a Tweedie distribution
- ❖ Y is member of Exponential Family

Members of the Exponential Family

- Normal
- Poisson
- Binomial
- Gamma
- Inverse Gaussian
- Compound Poisson

Variance Structure

❖ $E[Y_i] = \mu_i = b'(\theta_i) \rightarrow \theta_i = b'^{(-1)}(\mu_i)$

❖ $\text{Var}[Y_i] = a(\Phi_i) b''(\theta_i) = a(\Phi_i) V(\mu_i)$



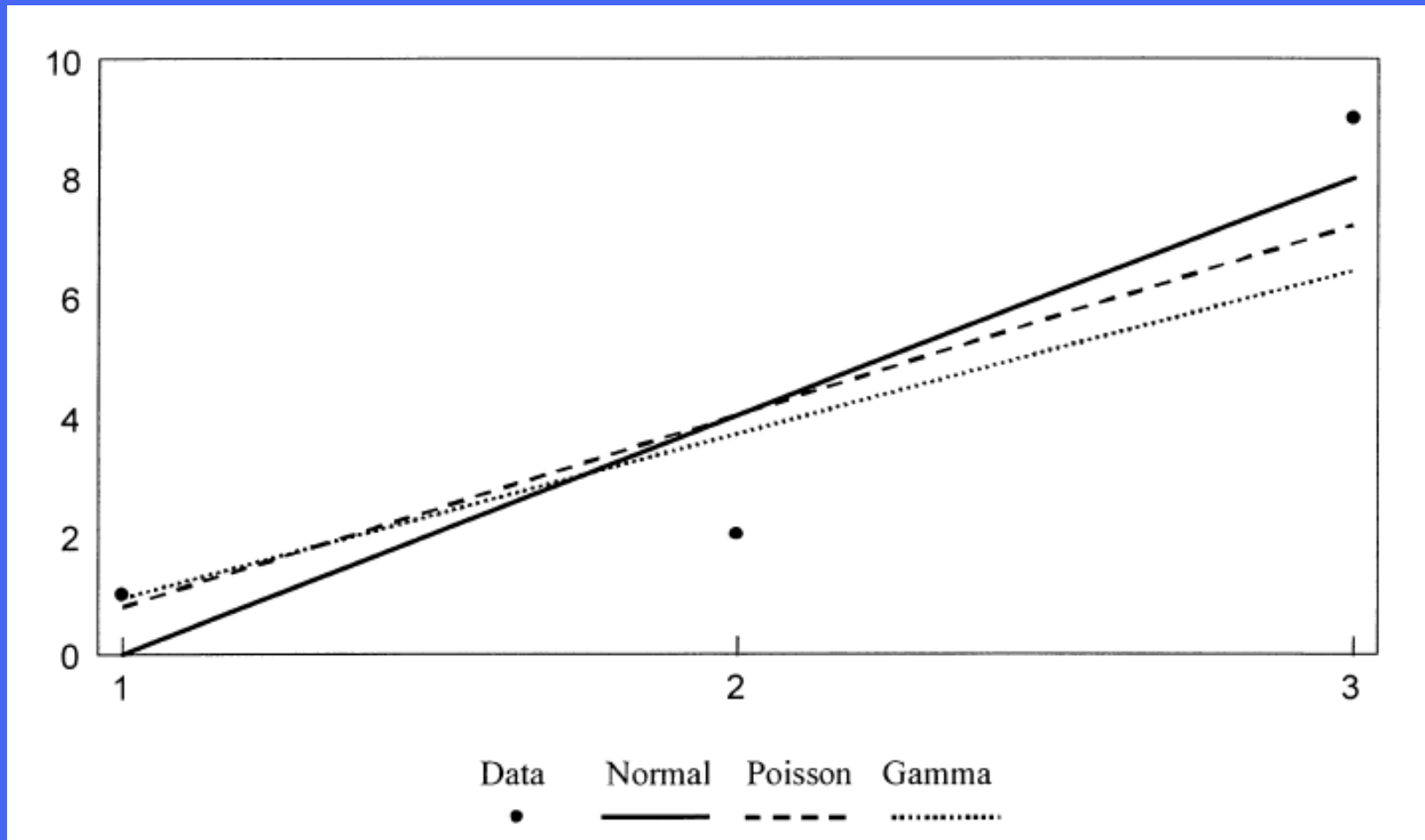
❖ Common form: $\text{Var}[Y_i] = \Phi V(\mu_i)/w_i$

✓ Φ is constant across data but weights applied to data points

Variance Functions $V(\mu)$

	<u>$V(\mu)$</u>
❖ Normal	μ^0
❖ Poisson	μ
❖ Binomial	$\mu(1-\mu)$
❖ Tweedie	$\mu^p, 1 < p < 2$
❖ Gamma	μ^2
❖ Inverse Gaussian	μ^3
❖ Recall:	$\text{Var}[Y_i] = \Phi V(\mu_i)/w_i$

Variance at Point and Fit



Variance of Y_i and Fit at Data Point i

- ❖ $\text{Var}(Y_i)$ is big \rightarrow looser fit at data point i
- ❖ $\text{Var}(Y_i)$ is small \rightarrow tighter fit at data point i

$$\text{Tightness of fit} \propto \frac{1}{\text{Var}(Y_i)}$$

Why Exponential Family?

- ❖ Distributions in Exponential Family can model a variety of problems
- ❖ Standard algorithm for finding coefficients a_0, a_1, \dots, a_m

Modeling Number of Claims

<u>Policy</u>	<u>Sex</u>	<u>Territory</u>	<u>Number of Claims in 5 Years</u>
1	M	02	0
2	F	01	0
3	F	01	0
4	F	02	1
5	F	01	0
6	F	02	1
7	M	02	2
8	M	02	2
9	M	02	1
10	F	01	1
:	:	:	:

Assume a Multiplicative Model

❖ μ_i = expected number of claims in five years

$$\text{❖ } \mu_i = B_{F,01} \times C_{\text{Sex}(i)} \times C_{\text{Terr}(i)}$$

❖ If i is Female and Terr 01

$$\rightarrow \mu_i = B_{F,01} \times 1.00 \times 1.00$$

Multiplicative Model

❖ $\mu_i = \exp(a_0 + a_S X_{S(i)} + a_T X_{T(i)})$

❖ $\mu_i = \exp(a_0) \times \exp(a_S X_{S(i)}) \times \exp(a_T X_{T(i)})$

❖ i is Female $\rightarrow X_{S(i)} = 0$; Male $\rightarrow X_{S(i)} = 1$

❖ i is Terr 01 $\rightarrow X_{T(i)} = 0$; Terr 02 $\rightarrow X_{T(i)} = 1$

Values of Predictor Variables

<u>Policy</u>	<u>Sex</u>	$X_{S(i)}$	<u>Territory</u>	$X_{T(i)}$
1	M	1	02	1
2	F	0	01	0
3	F	0	01	0
4	F	0	02	1
5	F	0	01	0
6	F	0	02	1
7	M	1	02	1
8	M	1	02	1
9	M	1	02	1
10	F	0	01	0

Natural Log Link Function

❖ $\ln(\mu_i) = a_0 + a_S X_{S(i)} + a_T X_{T(i)}$

❖ μ_i is in $(0, \infty)$

❖ $\ln(\mu_i)$ is in $(-\infty, \infty)$

Poisson Distribution in Exponential Family

$$\Pr[Y = y] = \exp\left\{\frac{\ln \mu \cdot y - \mu}{1} - \ln(y!)\right\}$$

$$\theta = \ln \mu$$

$$b(\theta) = e^{\theta}$$

Natural Log is Canonical Link for Poisson

$$\diamond \theta_i = \ln(\mu_i)$$

$$\diamond \theta_i = a_0 + a_S X_{S(i)} + a_T X_{T(i)}$$

Estimating Coefficients a_1, a_2, \dots, a_m

- ❖ **Classical linear regression uses least squares**
- ❖ **GLMs use Maximum Likelihood Method**
- ❖ **Solution will exist for distributions in exponential family**

Likelihood and Log Likelihood

$$L(y_1, \dots; \theta_1, \dots) = \prod_{i=1}^n f(y_i; \theta_i)$$

$$\lambda(y_1, \dots; \theta_1, \dots) = \ln[L(y_1, \dots; \theta_1, \dots)]$$

$$\lambda(y_1, \dots; \theta_1, \dots) = \sum_{i=1}^n \ln f(y_i; \theta_i)$$

Find a_0 , a_S , and a_T for Poisson

Maximize:

$$\lambda(y_1, \dots; \theta_1, \dots) = \sum_{i=1}^n \theta_i y_i - e^{\theta_i} - \ln y_i!$$

with $\theta_i = a_0 + a_S x_{S(i)} + a_T x_{T(i)}$

Iterative Numerical Procedure to Find a_i 's

- ❖ Use statistical package or actuarial software
- ❖ Specify link function and distribution type
- ❖ “Iterative weighted least squares” is the numerical method used

Solution to Our Example

❖ $a_0 = -.288 \rightarrow \exp(-.288) = .75$

❖ $a_S = .262 \rightarrow \exp(.262) = 1.3$

❖ $a_T = .095 \rightarrow \exp(.095) = 1.1$

❖ $\mu_i = \exp(a_0) \times \exp(a_S X_{S(i)}) \times \exp(a_T X_{T(i)})$

❖ $\mu_i = .75 \times 1.3^{X_{S(i)}} \times 1.1^{X_{T(i)}}$

❖ i is Male, Terr 01 $\rightarrow \mu_i = .75 \times 1.3^1 \times 1.1^0$

Testing New Drug Treatment

X_1	X_2		Y
<u>Dosage</u>	<u>Age</u>	<u>Cure</u>	<u>Value</u>
1.0	30	Yes	1
1.0	43	No	0
1.0	82	No	0
1.5	45	No	0
1.5	67	No	0
1.5	26	Yes	1
2.0	33	Yes	1
2.0	50	Yes	1
2.0	72	No	0
2.5	31	Yes	1
2.5	45	Yes	1
2.5	75	Yes	1

Multiple Linear Regression

	<u>Dependent Variable: Y</u>	
<u>Cure</u>	<u>Actual</u>	<u>Predicted Probability</u>
Yes	1	0.5179
No	0	0.3298
No	0	-0.2345
No	0	0.5366
No	0	0.2183
Yes	1	0.8115
Yes	1	0.9460
Yes	1	0.7000
No	0	0.3817
Yes	1	1.2107
Yes	1	1.0081
Yes	1	0.5740

Logistic Regression Model

❖ $p =$ probability of cure, p in $[0,1]$

❖ odds ratio: $p/(1-p)$ in $[0, +\infty]$

❖ $\ln[p/(1-p)]$ in $[-\infty, +\infty]$

❖ $\ln[p/(1-p)] = a + b_1X_1 + b_2X_2$

 Link function

Logistic Regression Model

X_1	X_2		Dependent Variable Y	
<u>Dosage</u>	<u>Age</u>	<u>Cure</u>	<u>Value</u>	<u>Predicted Probability</u>
1.0	30	Yes	1	0.568
1.0	43	No	0	0.000
1.0	82	No	0	0.000
1.5	45	No	0	0.648
1.5	67	No	0	0.000
1.5	26	Yes	1	1.000
2.0	33	Yes	1	1.000
2.0	50	Yes	1	1.000
2.0	72	No	0	0.000
2.5	31	Yes	1	1.000
2.5	45	Yes	1	1.000
2.5	75	Yes	1	0.784

Which Exponential Family Distribution?

- ❖ **Frequency: Poisson, {Negative Binomial}**
- ❖ **Severity: Gamma**
- ❖ **Loss ratio: Compound Poisson**
- ❖ **Pure Premium: Compound Poisson**
- ❖ **How many policies will renew: Binomial**

What link function?

- ❖ Additive model: identity
- ❖ Multiplicative model: natural log
- ❖ Modeling probability of event: logistic

