RESIDUALS AND INFLUENCE IN REGRESSION

EDMUND S. SCANLON

Abstract

The purpose of this paper is to cover some techniques in statistics that are important for testing the appropriateness of a fitted regression equation. These techniques, which are often used by statisticians, are not completely covered in the Proceedings. *Specifically, the areas discussed are:*

- *Elimination of the Constant in the Regression Equation*
- *Regression Diagnostics*
- *Analysis of Residuals*

1. INTRODUCTION

Estimating the parameters of a regression equation entails more than simply fitting a line to a set of data. During the estimation process, it is important to determine if the underlying assumptions are met and whether the equation accurately models the studied process.

The purpose of this paper is to discuss some aspects of regression that are important in testing the appropriateness of the fitted regression equation. These aspects, some of which are briefly covered in the present Syllabus are: the constant term in the multiple regression equation, regression diagnostics, and the analysis of residuals. It should be noted that the material described is contained in the references listed at the end of this paper.

2. THE CONSTANT TERM

At least two papers in the *Proceedings* [(1),(5)] suggest removing the constant term in the regression equation under some circumstances. The circumstances described in the papers seem to be reasonable causes for removal of the constant term. For example, one reason outlined is that the constant does not *explain* any of the change in the

dependent variable. Thus, the constant term should be carefully scrutinized and perhaps removed. Another suggested reason to remove the constant is to achieve a regression model which is intuitively sensible. Unfortunately, the removal of the constant, especially when it is statistically significant, tends to impair the accuracy of the model. Hence, something should be added to compensate for the removal of the constant.

Both papers seem to imply that the constant should be eliminated if the corresponding t-statistic is insignificant (i.e., $|t| < 2$). Or, if the t-statistic is significant, one should try to search for another independent variable in an attempt to reduce the significance of the constant term.

These procedures are unreliable for three reasons:

1. Although it is sometimes not clearly stated in statistics texts, the *objective* in the traditional statistical test is to decide whether or not to reject the null hypothesis. Acceptance of the null hypothesis is not the issue. For example, in analysis-of-variance (ANOVA), the null hypothesis (H_0) is that $\mu_1 = \mu_2 = \ldots = \mu_n$. If the F-statistic is significant, one may reject the null hypothesis. Even in the absence of a high F-statistic, the null hypothesis is difficult to believe; all one can say is that the hypothesis cannot be proved false (i.e., fail to reject H_0).

2. Eliminating the constant gives the origin (which can be considered one observation) an undue amount of leverage on the fitted regression equation (the subject of leverage is discussed in more detail in the next section). As a result of this elimination, the regression line is forced through a particular point, the origin. In other words, the origin, as an observation, is given special treatment because it is not subject to the least squares constraint (i.e., minimize the sum of squares). Thus, the origin has more influence on the regression model than a typical observation.

3. An additional independent variable may not be easily found.

This is not to say that one should never eliminate the constant in the regression equation. The point is that this elimination should not be considered lightly. We will illustrate this with a numerical example in the next section.

3. REGRESSION DIAGNOSTICS

When analyzing the appropriateness of a regression equation, most statisticians review the data to see which observations are "influencing" the estimation of the regression equation coefficients. More generally, the statistician wants to identify subsets of the data that have disproportional influence on the estimated regression model. As discussed by Belsley et al. [2], these influential subsets can come from a number of sources:

1. improperly recorded data,
2. errors that are inherent in the data, and
3. outliers that are legitimately extreme observations.

Belsley et al. indicate some interesting situations that might be subject to detection by diagnostics. Exhibit 1 summarizes these situations. Part 1 displays the ideal situation: All the data is essentially grouped together. In Part 2, the point labeled z is an aberration or outlier; but, since it is near $\bar{x}$ there is no adverse effect on the slope. However, the estimate of the intercept will obviously be influenced. Part 3 also displays a data set with an outlier. However, the outlier in this example is consistent with the slope indicated by the remaining data. Because of this consistency, adding the outlier to the regression calculation reduces the variance of the parameter estimates (i.e., improves the quality of the regression). Generally, if the variance of the independent variable is small, slope estimates will be unreliable. Part 4 is a problem situation, since the outlier essentially defines the slope. In the absence of the outlier, the slope might be anything. The outlier has extreme influence on the slope. Part 5 is a case where there are

two outliers that are both influential and whose effects are complementary. Such a situation may call for use of one of the following procedures:

1. deleting the observations,
2. downweighting (i.e., giving less weight to the observations),
3. reformulating the model (e.g., adding or deleting independent variables); or
4. where possible, using more observations.

Part 6 displays a situation where deletion of either outlier has little effect on the regression outcome because neither outlier exerts much influence upon the regression parameters. Parts 5 and 6 highlight the need to examine the effects of general subsets of data.

To demonstrate some of these situations more clearly, we consider the numerical data and regression results on Exhibit 2, Part 1. This data set, which is plotted on Part 2, is similar to the pattern displayed on Exhibit 1, Part 1. The regression results, as expected from the uniformity of the plot, indicate a very good fit.

In order to illustrate the Exhibit 1, Part 2 situation, the point (7, 14.3) was added to the base data set. The regression results and the plot of the data are on Exhibit 2, Parts 3 and 4, respectively. It is interesting to note how the additional observation influenced the parameter estimates. The constant changed from .702 to 1.171. However, the slope estimate change was negligible (.808 to .784).

The Exhibit 1, Part 3 case can be demonstrated by adding the point (17, 14.3) in lieu of (7, 14.3). This new outlier is consistent with the remaining data (i.e., it lies on the path of the line indicated by the base data set). The regression results and the plot for this revised data set are displayed on Exhibit 2, Parts 5 and 6, respectively. The results indicate minimal change in the parameter estimates. Hence, (17, 14.3) does not have significant influence on the regression model. However, adding the point (17, 14.3) decreased the variance of the parameter estimates. It should also be noted that the standard error of

the residual associated with this outlier is relatively smaller than the standard errors associated with all the other observations. The magnitude of the standard error for a residual relative to the other standard errors is an indication of the leverage of a point (i.e., the *potential* of the point to influence the calculation of the regression equation). Leverage depends on whether an observation is an outlier with respect to the x axis.

To demonstrate another example of leverage, the point (17, 10) was used instead of (17, 14.3) and a new regression equation was calculated. The regression results and the plot can be found on Exhibit 2, Parts 7 and 8, respectively. The contrast between the two outliers, (7, 14.3) and (17, 10), is interesting. The outlier (7, 14.3), an outlier with respect to the y axis, is about 8 units away from where it "should be." The other outlier, (17, 10), an outlier with respect to both the x and y axes, is only about 4 units away from where it "should be." However, the influence of (17, 10) on the parameter estimates is much greater than that of (7, 14.3). As mentioned earlier, the (7, 14.3) outlier influenced only the estimate of the constant. There was a negligible change in the estimate of the slope.

The estimates of the parameters under the varying data sets are summarized in the following table:

	Base Set	Set with (7, 14.3)	Set with (17, 10)
Intercept	.702	1.171	1.394
Slope	.808	.784	.705

As indicated by the table, the point (17, 10) influences both the intercept and slope estimates to a much greater extent than (7, 14.3).

At this time we return to the question of eliminating the constant. It is interesting to note the situation of removing the constant term when fitting a regression line to the base data set (Exhibit 2, Part 1). The regression analysis of the base data set indicates that the t-statistic for the constant term is not "statistically significant." However, removing the constant term from the regression equation influences

the slope estimate considerably. The coefficient of the independent variable is now .883 as compared to .808 .

The preceding examples indicate that when there are two or fewer independent variables, scatter plots such as Exhibit 1 can quickly reveal any outliers. However, when there are more than two independent variables, scatter plots may not reveal multivariate outliers that are separated from the bulk of the data. What follows is a discussion of some diagnostic statistics that are useful in detecting such outliers.

There are a number of different statistics used by statisticians to detect outliers in the data. One such statistic is Cook's D_i (or Cook's Distance) statistic. The statistic is named after the statistician R. D. Cook. Cook's D_i measures the *influence* of the i^{th} observation. It is based on the difference between two estimators (one estimator includes the i^{th} observation in the data; the other excludes the i^{th} observation). Using matrix notation, Cook's D_i is defined as follows:

$$D_i = \{\hat{\beta} - \hat{\beta}(i)\}^T X^T X \{\hat{\beta} - \hat{\beta}(i)\}/ps^2,$$

where:

- X is the n by p matrix that contains the values of the independent variables (i.e., n different values of the $(p-1)$ independent variables together with a first column that is equal to unity, representing the constant); this is the same X that is used in the familiar multiple regression equation $Y = X\beta + e$ (see, for example, Miller and Wichern [6, supplement 5B]);
- X^T is the transpose of X;
- $\hat{\beta}$ is the usual least squares estimator vector of p by 1 dimensions;
- $\hat{\beta}(i)$ is the least squares estimator after the i^{th} data point has been omitted from the data, also p by 1 dimensions;
- p is the number of independent variables plus one;

s^2 is the estimate of variance provided by residual mean square error from using the full data set;

$\{\hat{\beta}-\hat{\beta}(i)\}$ is the difference between the two p by 1 vectors, also p by 1.

A large D_i represents an influential observation; that is, an observation that has more than the average influence on the estimation of the parameters. Presently, there is no formal definition of a "large D_i." However, there are some general rules that statisticians follow. First, if $D_i > 1$, then the observation should probably be considered influential. Second, if all D_is are below 1, a value considerably greater than the other values should be considered influential. Once a point with a large D_i has been identified, the actuary would want to examine the point to be certain that such an observation is typical and not an aberration. With Cook's D_i, the actuary can review all outliers and decide whether or not to eliminate an observation. This process is somewhat analogous to the reserving actuary eliminating high/low loss development link ratios from an average.

The preceding formula for Cook's D_i is rather cumbersome. Fortunately, it is standard output for most statistical software packages.

The so-called *hat statistic*, h_{ii}, is another tool that is helpful in determining which observations have significant *leverage*. Using matrix notation, the hat matrix (which contains the hat statistics) can be derived from the usual regression equations:

$$Y = X\beta + e,$$

$$\hat{Y} = X\hat{\beta};$$

$$\text{Since } \hat{\beta} = (X^T X)^{-1} X^T Y,$$

$$\hat{Y} = X(X^T X)^{-1} X^T Y,$$

$$= HY,$$

where H, the n by n hat matrix, is defined as:

$$H = X(X^T X)^{-1} X^T.$$

H is called the hat matrix because it transforms the vector of observed responses, Y, into the vector of fitted responses, $\hat{Y}$. From this, the vector of residuals can be defined as:

$$\begin{aligned}\hat{e} &= Y - \hat{Y} \\ &= Y - X(X^T X)^{-1} X^T Y \\ &= [I - H]Y.\end{aligned}$$

It is shown in Weisberg [8] that:

$$\mathrm{E}(\hat{e}_i) = 0, \text{ and } \mathrm{Var}(\hat{e}_i) = \sigma^2(1-h_{ii}),$$

where the hat statistic, h_{ii}, is the i^{th} diagonal element of H. This is in contrast to the errors, e_i, for which the variance is constant for all i. Incidentally, the variance for $\hat{y}_i$ is $\sigma^2 h_{ii}$.

Hence, cases with large values of h_{ii} will have small values of $\mathrm{Var}(\hat{e}_i)$. As h_{ii} approaches unity, the variance of the i^{th} residual approaches zero. In other words, as h_{ii} approaches unity, $\hat{y}_i$ (the estimate) approaches the observed value, y_i. This is why h_{ii} is called the leverage of the i^{th} observation. The effect of the i^{th} observation on the regression is more likely to be large if h_{ii} is large. Similar to Cook's D_i, the hat matrix is standard output from common statistical software packages.

How large is a "large" h_{ii} ? This issue is addressed by Belsley et al. They show that, if the explanatory variables are independently distributed as the multivariate Gaussian, it is possible to compute the exact distribution of certain functions of the h_{ii}s. Specifically, $(n-p)[h_{ii}-(1/n)]/(1-h_{ii})(p-1)$ is shown to be distributed as F with $p-1$ and $n-p$ degrees of freedom. For $p>10$ and $n-p>50$, the 95% value for F is less than 2. Hence, $2p/n$ is roughly a good cutoff (twice the balanced average h_{ii}).

At this point, it is appropriate to discuss the difference between *influence* and *leverage*. The leverage of an observation was just defined as h_{ii}. Note that this is independent of the dependent variable. Hence, the definition of leverage ignores the role played by y_i (the observation). Influence, on the other hand, is defined as follows:

> A case is said to be *influential* if appreciable changes in the fitted regression coefficients occur when it is removed from the data [7].

Another way to look at the difference between influence and leverage is as follows. The h_{ii}s are indicating how well the independent data is "spread out." Exhibit 1, Part 4 displays data that contain an observation that has leverage. The amount of leverage would be reduced if there were more observations with larger independent values near that point's independent value.

Cook's D_i actually indicates how much of the leverage is being exerted by the observation on the estimation of the coefficients. Therefore, Cook's D_i is more helpful in analyzing a regression model.

It is interesting to examine the relationship between h_{ii} and D_i to understand the difference between influence and leverage. Weisberg [8] derives the following relationship:

$$D_i = \{e_i / s\,(1-h_{ii})^{1/2}\}^2\, h_{ii} / p\,(1-h_{ii}).$$

This formula is helpful in a number of ways. First, it shows that Cook's D_i can be calculated from data output of the full regression without the need to recompute estimates excluding observations. Second, it displays the relationship of Cook's D_i, the studentized residuals, and the measure of leverage. Third, it shows explicitly that the hat diagonal describes only the potential for influence. D_i will be large only if both h_{ii} and the associated residual are large.

This clarification is important because the two terms (influence and leverage) are sometimes used synonymously. For example, Cook and Weisberg [3] mention authors who interpret h_{ii} as the amount of leverage *or* influence.

Thus far, this discussion has focused on "single row" diagnostics. As mentioned earlier, Exhibit 1, Part 6 indicates the need for multiple row diagnostics; for example, in situations where one outlier masks the effect of another outlier. Such techniques exist and the interested reader is referred to Belsley et al [2].

To illuminate the use of these diagnostics, a multiple regression model is fitted to some pure premium data in Exhibit 3. The model used is similar to some work performed by the Insurance Services Office; namely, the alternative trend models.

In this example, pure premium (PP) is the dependent variable, while the Consumer Price Index (CPI) and the change in the Gross National Product (GNP) are the independent variables. It should be noted that the values of these variables are realistic, but fabricated for the example. Exhibit 3, Part 1 displays results from fitting a regression model to 10 observations. The model fitted is as follows:

$$\text{Pure Premium} = b_0 + b_1\text{CPI} + b_2\text{GNP}.$$

Including all 10 observations in the calculation produces an excellent fit, as indicated by the adjusted R^2. Nevertheless, the residuals do become relatively larger as the pure premium increases. The h_{ii} values indicate that the 1st and the 9th observations have a considerable amount of leverage. The D_1 value indicates that the 1st observation is not influencing the model's coefficients. However, the D_9 value does indicate that the 9th observation has significant influence. In order to improve the model, consideration should be given to removing the 9th observation. Exhibit 3, Part 2 displays output excluding this observation. These results can be summarized as follows:

1. The adjusted R^2 improved slightly.
2. The residuals are now more stable than before.
3. The values of D_i are stable.

4. ANALYSIS OF RESIDUALS

Analysis of residuals is touched upon in some of the present *Syllabus* readings. As indicated in the readings, residuals can be helpful in determining if two required regression assumptions have been violated; namely, the error terms must be independent and the variance must be constant for all observations. Violations of these assumptions are associated with the terms autocorrelation and heteroscedasticity, respectively.

Heteroscedasticity, or non-constant variance, is typically detected by using a so-called residual plot [6]. If the plot of residuals is shaped like a cone (see Exhibit 4), it is likely that heteroscedasticity exists. These residual plots are also helpful in determining whether the regression equation needs an additional independent term.

It is important to note (as mentioned in the previous section) that the variance of $\hat{e}_i$ is not a constant for all i. As a matter of fact, it would be unusual for all the $\hat{e}_i$'s to have the same variance. That is, it is possible to have a pattern similar to Exhibit 4 simply because the h_{ii}s are not constant.

Improved diagnostics can be achieved by dividing the residuals by an estimate of the standard error. Specifically, the residual, $\hat{e}_i$, should be divided by $s(1-h_{ii})^{1/2}$ for all i. These scaled residuals, also known as the studentized residuals, will all have a common variance, if the model is correct. The studentized residuals can then be used, graphically, to test for heteroscedasticity.

An additional point which is not emphasized in many books is the reason the residuals are plotted against $\hat{Y}$ and not Y. The reason is that the residuals and the actual observations are correlated, but the residuals and the fitted values are not.

This can be shown [4] by calculating the sample correlation coefficients between the residuals and the actual and fitted observations. First, the sample correlation coefficient between e and Y, r_{ey}, is calculated as follows:

$$r_{ey} = \{\Sigma(e_i-\bar{e})(Y_i-\bar{Y})\}/\{\Sigma(e_i-\bar{e})^2\,\Sigma(Y_i-\bar{Y})^2\}^{1/2}.$$

The numerator,

$$\Sigma(e_i-\bar{e})(Y_i-\bar{Y}) = \Sigma e_i(Y_i-\bar{Y}) \quad \text{since } \bar{e}=0, \text{ if a constant is in the model}$$

$$= \Sigma e_i Y_i \qquad (\bar{e}=0)$$

$$= e^T Y \qquad \text{in matrix notation}$$

$$= e^T e \qquad \text{because } e^T e = Y^T(I-H)^T(I-H)Y$$

$$= Y^T(I-H)Y$$

$$= e^T Y$$

= Residual sum of squares.

Therefore,

r_{ey} = { Residual sum of squares / Total sum of squares }$^{1/2} = (1-R^2)^{1/2}$.

The calculation of $r_{e\hat{y}}$ is similar to the above,

$$\Sigma(e_i-\bar{e})(\hat{Y}-\bar{Y}) = \Sigma e_i\hat{Y}_i = e^T\hat{Y} = Y^T(I-H)^T HY = 0.$$

Hence, $r_{e\hat{y}} = 0$.

5. CONCLUSIONS

The field of statistics is a tremendous resource that, except for a theoretical foundation, goes untapped by casualty actuaries. I hope this paper adds modestly to the knowledge of some actuarial practitioners and inspires other such summaries.

REFERENCES

[1] Alff, Gregory N., "A Note Regarding Evaluation of Multiple Regression Models," *PCAS* LXXI, 1984, pp. 84-95.

[2] Belsley, David A., Edwin Kuh, and Roy E. Welsch, *Regression Diagnostics,* New York, John Wiley & Sons, 1980.

[3] Cook, R. D. and Sanford Weisberg, *Residuals and Influence in Regression,* New York, Chapman and Hall, 1982.

[4] Draper, Norman R. and Harry Smith, *Applied Regression Analysis,* New York, John Wiley & Sons, 1981.

[5] Lommele, Jan A. and Robert W. Sturgis, "An Econometric Model of Workmen's Compensation," *PCAS* LXI, 1974, pp. 170-189.

[6] Miller, Robert B. and Dean W. Wichern, *Intermediate Business Statistics,* New York, Holt, Rinehart and Winston, 1977.

[7] Shih, W. J., and Sanford Weisberg, "Assessing Influence in Multiple Linear Regression with Incomplete Data," *Technometrics,* 28, 1986, pp. 231-239.

[8] Weisberg, Sanford, *Applied Linear Regression,* New York, John Wiley & Sons, 1985.

EXHIBIT 1
Part 1

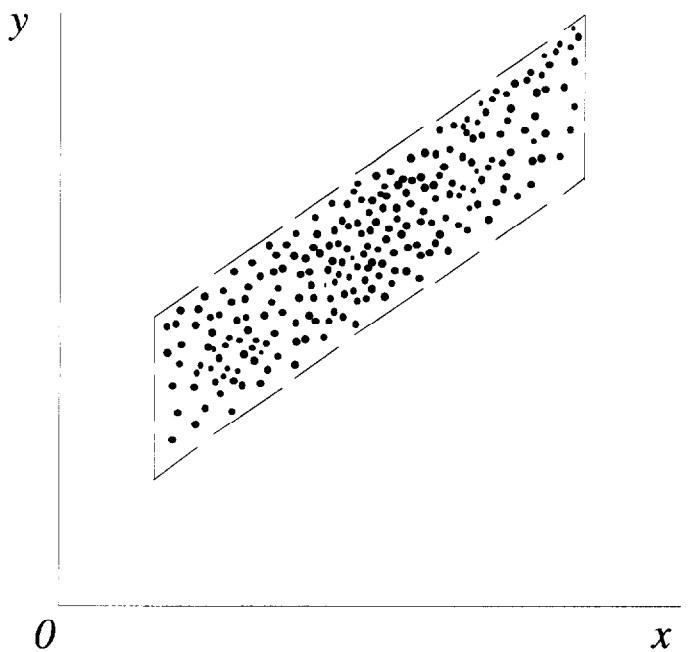

EXHIBIT 1
Part 2

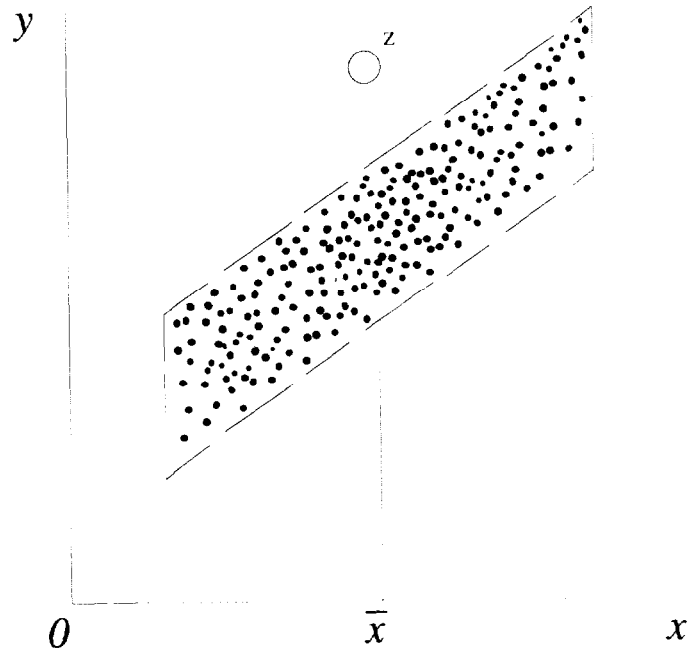

EXHIBIT 1
Part 3

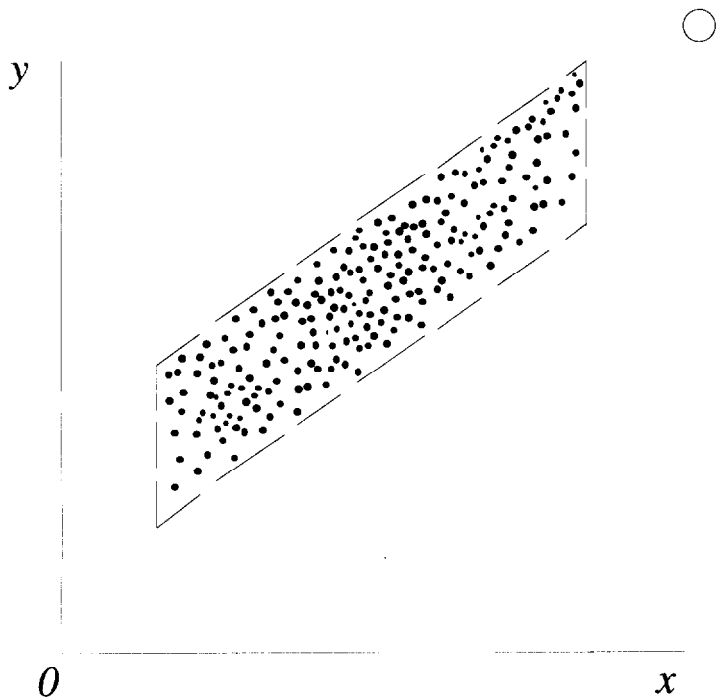

EXHIBIT 1
Part 4

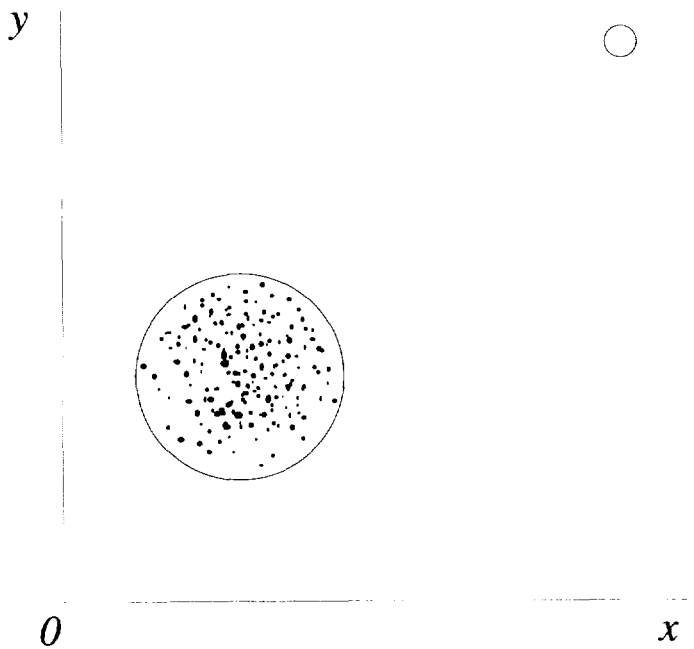

EXHIBIT 1
Part 5

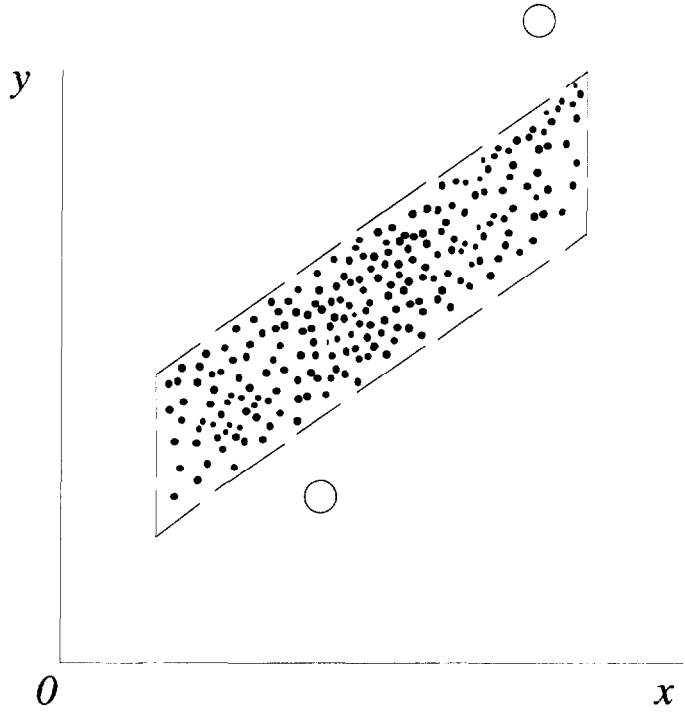

EXHIBIT 1
Part 6

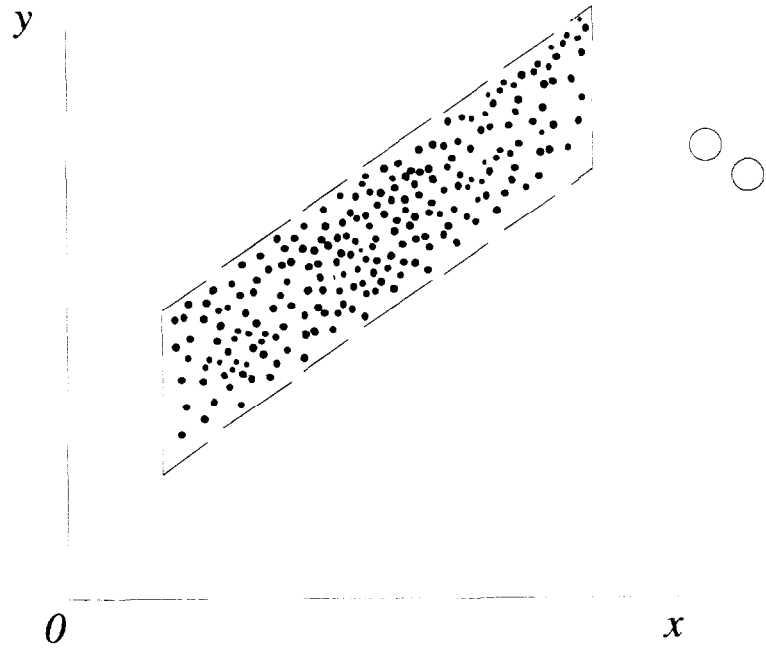

EXHIBIT 2
Part 1

REGRESSION ANALYSIS

	Parameter Estimates	Standard Error	*t* Statistic	Probability > |*t*|
Intercept	0.7016	0.5387	1.302	0.2046
Coefficient	0.8079	0.0628	12.870	0.0001

Adjusted *R*-Squared: 0.8636

X	*Y*	Predicted Value *Y*	Std. Err. of Pred.	Residual	Std. Err. of Residual
2.50	2.10	2.7214	0.3985	-0.6214	1.0050
3.00	2.30	3.1254	0.3721	-0.8254	1.0151
3.00	3.30	3.1254	0.3721	0.1746	1.0151
3.00	4.30	3.1254	0.3721	1.1746	1.0151
4.00	3.30	3.9333	0.3220	-0.6333	1.0321
4.00	5.40	3.9333	0.3220	1.4667	1.0321
5.00	4.30	4.7413	0.2771	-0.4413	1.0450
5.50	5.60	5.1453	0.2574	0.4547	1.0500
6.00	7.30	5.5492	0.2403	1.7508	1.0541
6.50	4.40	5.9532	0.2262	-1.5532	1.0572
7.00	6.00	6.3572	0.2158	-0.3572	1.0594
7.00	6.50	6.3572	0.2158	0.1428	1.0594
7.50	7.50	6.7611	0.2097	0.7389	1.0606
8.20	6.30	7.3267	0.2088	-1.0267	1.0608
9.00	6.60	7.9731	0.2189	-1.3731	1.0587
9.00	8.50	7.9731	0.2189	0.5269	1.0587
9.00	9.00	7.9731	0.2189	1.0269	1.0587
9.50	6.50	8.3770	0.2306	-1.8770	1.0562
10.00	9.30	8.7810	0.2458	0.5190	1.0528
10.50	8.00	9.1850	0.2639	-1.1850	1.0484
11.00	8.50	9.5890	0.2843	-1.0890	1.0431
11.00	10.60	9.5890	0.2843	1.0110	1.0431
12.00	10.00	10.3969	0.3302	-0.3969	1.0295
12.00	11.50	10.3969	0.3302	1.1031	1.0295
12.50	12.00	10.8009	0.3552	1.1991	1.0211
13.00	10.00	11.2048	0.3810	-1.2048	1.0117
13.00	12.50	11.2048	0.3810	1.2952	1.0117

EXHIBIT 2
Part 2

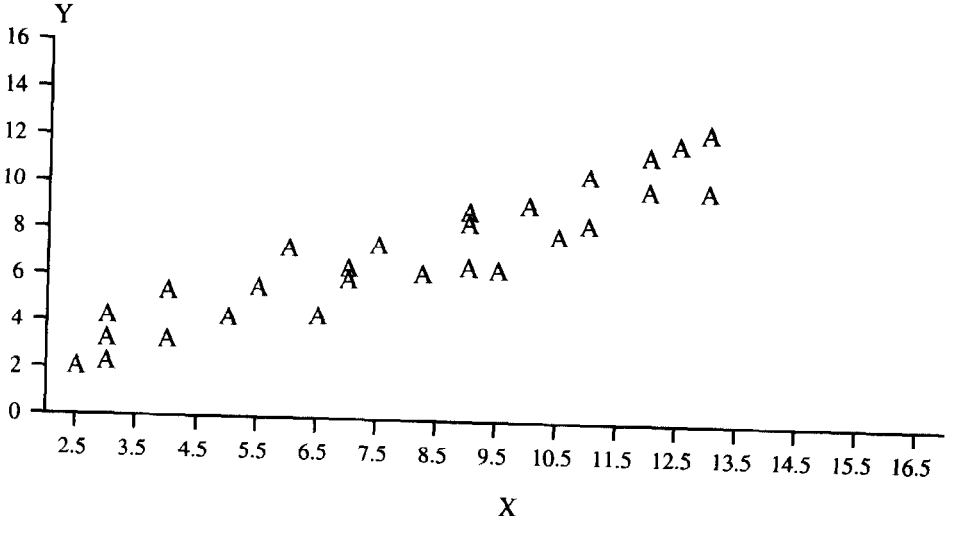

EXHIBIT 2
Part 3

REGRESSION ANALYSIS

	Parameter Estimates	Standard Error	*t* Statistic	Probability > |*t*|
Intercept	1.1709	0.9196	1.273	0.2142
Coefficient	0.7840	0.1078	7.276	0.0001

Adjusted *R*-Squared: 0.6579

X	*Y*	Predicted Value *Y*	Std. Err. of Pred.	Residual	Std. Err. of Residual
2.50	2.10	3.1319	0.6784	-1.0319	1.7312
3.00	2.30	3.5241	0.6329	-1.2241	1.7484
3.00	3.30	3.5241	0.6329	-0.2241	1.7484
3.00	4.30	3.5241	0.6329	0.7759	1.7484
4.00	3.30	4.3085	0.5465	-1.0085	1.7773
4.00	5.40	4.3085	0.5465	1.0915	1.7773
5.00	4.30	5.0929	0.4691	-0.7929	1.7992
5.50	5.60	5.4851	0.4353	0.1149	1.8077
6.00	7.30	5.8772	0.4058	1.4228	1.8146
6.50	4.40	6.2694	0.3817	-1.8694	1.8198
7.00	6.00	6.6616	0.3640	-0.6616	1.8234
7.00	6.50	6.6616	0.3640	-0.1616	1.8234
7.50	7.50	7.0538	0.3538	0.4462	1.8254
8.20	6.30	7.6029	0.3531	-1.3029	1.8256
9.00	6.60	8.2304	0.3715	-1.6304	1.8219
9.00	8.50	8.2304	0.3715	0.2696	1.8219
9.00	9.00	8.2304	0.3715	0.7696	1.8219
9.50	6.50	8.6226	0.3923	-2.1226	1.8175
10.00	9.30	9.0148	0.4191	0.2852	1.8116
10.50	8.00	9.4070	0.4507	-1.4070	1.8039
11.00	8.50	9.7992	0.4863	-1.2992	1.7947
11.00	10.60	9.7992	0.4863	0.8008	1.7947
12.00	10.00	10.5836	0.5662	-0.5836	1.7711
12.00	11.50	10.5836	0.5662	0.9164	1.7711
12.50	12.00	10.9757	0.6094	1.0243	1.7567
13.00	10.00	11.3679	0.6542	-1.3679	1.7405
13.00	12.50	11.3679	0.6542	1.1321	1.7405
7.00	14.30	6.6616	0.3640	7.6384	1.8234

EXHIBIT 2
Part 4

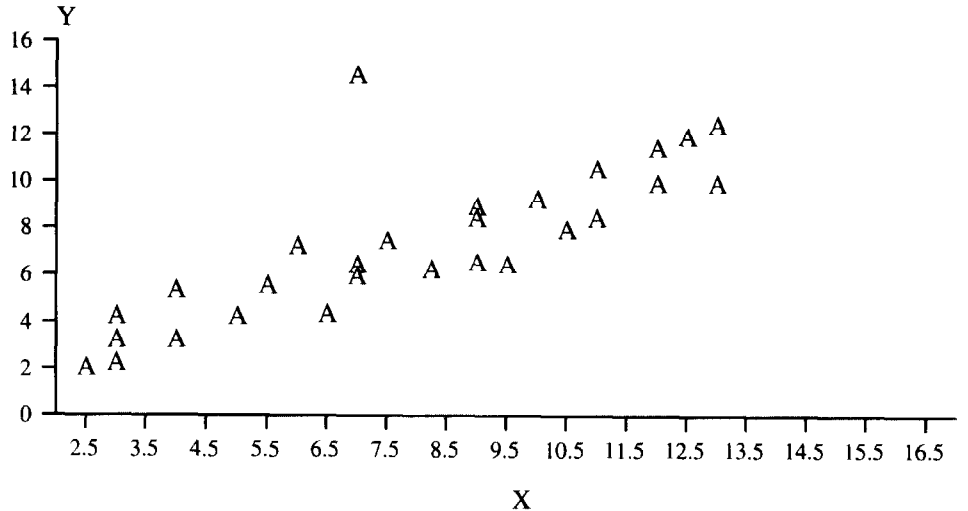

EXHIBIT 2
Part 5

REGRESSION ANALYSIS

	Parameter Estimates	Standard Error	*t* Statistic	Probability > |*t*|
Intercept	0.7229	0.4930	1.466	0.1546
Coefficient	0.8048	0.0547	14.713	0.0001

Adjusted *R*-Squared: 0.8887

X	*Y*	Predicted Value *Y*	Std. Err. of Pred.	Residual	Std. Err. of Residual
2.50	2.10	2.7348	0.3723	-0.6348	0.9929
3.00	2.30	3.1372	0.3496	-0.8372	1.0011
3.00	3.30	3.1372	0.3496	0.1628	1.0011
3.00	4.30	3.1372	0.3496	1.1628	1.0011
4.00	3.30	3.9420	0.3064	-0.6420	1.0152
4.00	5.40	3.9420	0.3064	1.4580	1.0152
5.00	4.30	4.7467	0.2674	-0.4467	1.0261
5.50	5.60	5.1491	0.2502	0.4509	1.0305
6.00	7.30	5.5515	0.2348	1.7485	1.0341
6.50	4.40	5.9539	0.2218	-1.5539	1.0369
7.00	6.00	6.3562	0.2115	-0.3562	1.0391
7.00	6.50	6.3562	0.2115	0.1438	1.0391
7.50	7.50	6.7586	0.2044	0.7414	1.0405
8.20	6.30	7.3220	0.2004	-1.0220	1.0413
9.00	6.60	7.9658	0.2047	-1.3658	1.0404
9.00	8.50	7.9658	0.2047	0.5342	1.0404
9.00	9.00	7.9658	0.2047	1.0342	1.0404
9.50	6.50	8.3681	0.2119	-1.8681	1.0390
10.00	9.30	8.7705	0.2223	0.5295	1.0368
10.50	8.00	9.1729	0.2354	-1.1729	1.0339
11.00	8.50	9.5753	0.2509	-1.0753	1.0303
11.00	10.60	9.5753	0.2509	1.0247	1.0303
12.00	10.00	10.3801	0.2871	-0.3801	1.0208
12.00	11.50	10.3801	0.2871	1.1199	1.0208
12.50	12.00	10.7824	0.3073	1.2176	1.0149
13.00	10.00	11.1848	0.3285	-1.1848	1.0082
13.00	12.50	11.1848	0.3285	1.3152	1.0082
17.00	14.30	14.4039	0.5192	-0.1039	0.9246

EXHIBIT 2
Part 6

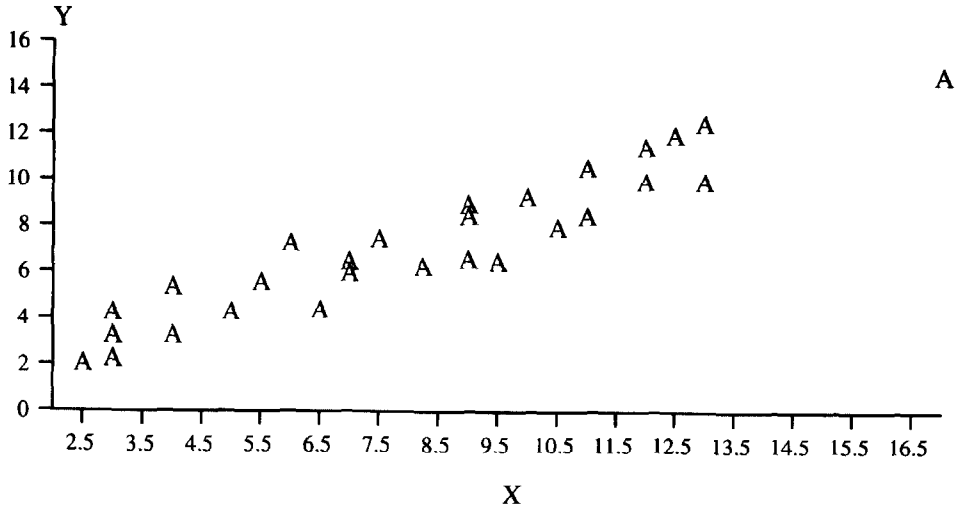

EXHIBIT 2
Part 7

REGRESSION ANALYSIS

	Parameter Estimates	Standard Error	*t* Statistic	Probability > |*t*|
Intercept	1.3944	0.6061	2.3006	0.0297
Coefficient	0.7046	0.0672	10.4851	0.0001

Adjusted R-Squared: 0.8013

X	*Y*	Predicted Value *Y*	Std. Err. of Pred.	Residual	Std. Err. of Residual
2.50	2.10	3.1560	0.4577	-1.0560	1.2206
3.00	2.30	3.5083	0.4298	-1.2083	1.2307
3.00	3.30	3.5083	0.4298	-0.2083	1.2307
3.00	4.30	3.5083	0.4298	1.2417	1.2307
4.00	3.30	4.2129	0.3767	-0.9129	1.2480
4.00	5.40	4.2129	0.3767	1.1871	1.2480
5.00	4.30	4.9175	0.3288	-0.6175	1.2615
5.50	5.60	5.2698	0.3076	0.3302	1.2668
6.00	7.30	5.6221	0.2887	1.6779	1.2713
6.50	4.40	5.9745	0.2727	-1.5745	1.2748
7.00	6.00	6.3268	0.2601	-0.3268	1.2774
7.00	6.50	6.3268	0.2601	0.1732	1.2774
7.50	7.50	6.6791	0.2513	0.8209	1.2792
8.20	6.30	7.1723	0.2464	-0.8723	1.2801
9.00	6.60	7.7360	0.2516	-1.1360	1.2791
9.00	8.50	7.7360	0.2516	0.7640	1.2791
9.00	9.00	7.7360	0.2516	1.2640	1.2791
9.50	6.50	8.0883	0.2605	-1.5883	1.2773
10.00	9.30	8.4406	0.2733	0.8594	1.2747
10.50	8.00	8.7930	0.2895	-0.7930	1.2711
11.00	8.50	9.1453	0.3084	-0.6453	1.2666
11.00	10.60	9.1453	0.3084	1.4547	1.2666
12.00	10.00	9.8499	0.3530	0.1501	1.2549
12.00	11.50	9.8499	0.3530	1.6501	1.2549
12.50	12.00	10.2022	0.3778	1.7978	1.2477
13.00	10.00	10.5545	0.4038	-0.5545	1.2395
13.00	12.50	10.5545	0.4038	1.9455	1.2395
17.00	10.00	13.3730	0.6383	-3.3730	1.1367

EXHIBIT 2
Part 8

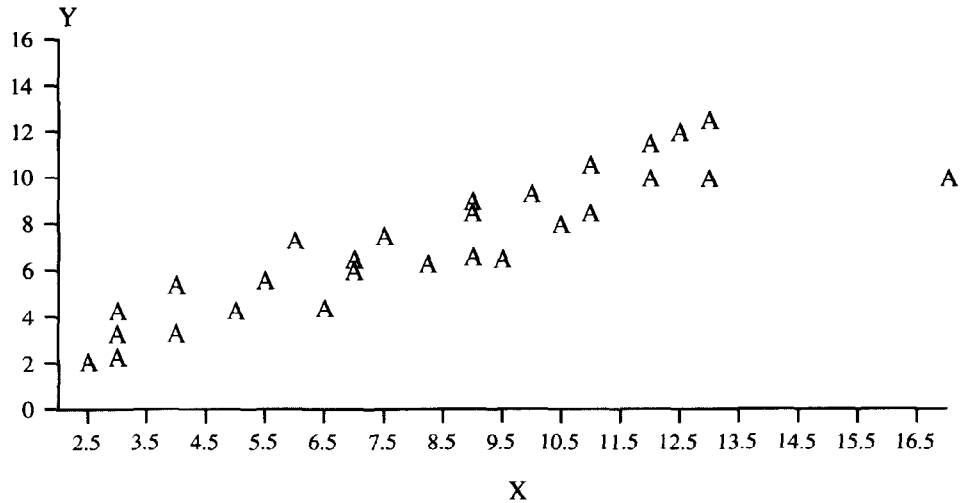

EXHIBIT 3
Part 1

ANALYSIS OF REGRESSION MODEL

	Parameter Estimates	Standard Error	*t* Statistic	Probability > |*t*|
Intercept	-54.262	7.051	-7.696	0.0001
GNP	2.570	0.449	5.724	0.0007
CPI	1.024	0.037	27.676	0.0001

Adjusted *R*-Squared: 0.9916

Obs	PP	GNP	CPI	Fitted PP	Residual
1	105.0	3.00	148	105.0	-0.0446
2	111.0	3.07	153	110.3	0.6540
3	114.0	2.97	157	114.2	-0.1863
4	118.0	4.14	158	118.2	-0.2171
5	126.0	4.00	166	126.1	-0.0518
6	128.0	3.42	170	128.7	-0.6586
7	134.0	2.67	177	133.9	0.0985
8	134.0	1.78	178	132.6	1.3612
9	131.0	1.09	180	132.9	-1.9144
10	138.0	1.50	183	137.0	0.9591

Residual Plot

Obs	-2 -1 0 1 2	Hat Diagonal	Cook's *D*
1		0.4818	0.001
2	*	0.2808	0.073
3		0.1939	0.003
4		0.3069	0.009
5		0.3664	0.001
6	*	0.2345	0.055
7		0.2108	0.001
8	**	0.2147	0.203
9	****	0.3979	1.269
10	**	0.3125	0.192

EXHIBIT 3
Part 2

ANALYSIS OF REGRESSION MODEL

	Parameter Estimates	Standard Error	*t* Statistic	Probability > |*t*|
Intercept	-51.960	3.234	-16.067	0.0001
GNP	1.987	0.232	8.565	0.0001
CPI	1.022	0.017	60.118	0.0001

Adjusted R-Squared: 0.9984

Obs	PP	GNP	CPI	Fitted PP	Residual
1	105.0	3.00	148	105.3	-0.2673
2	111.0	3.07	153	110.5	0.4832
3	114.0	2.97	157	114.4	-0.4064
4	118.0	4.14	158	117.8	0.2471
5	126.0	4.00	166	125.7	0.3487
6	128.0	3.42	170	128.6	-0.5874
7	134.0	2.67	177	134.3	-0.2519
8	134.0	1.78	178	133.5	0.4941
9	138.0	1.50	183	138.1	-0.0600

Obs	Residual Plot (-2 -1 0 1 2)	Hat Diagonal	Cook's D
1		0.4899	0.206
2	*|	0.2856	0.200
3	*|	0.2018	0.080
4	*|	0.3423	0.074
5	*|	0.3928	0.198
6	*|	0.2353	0.212
7	*|	0.2310	0.038
8	*|	0.3382	0.288
9	*|	0.4831	0.010

EXHIBIT 4