Case Study --- Pontius Data

The case study given herein was used in Section 4.6.1 to illustrate the construction of a regression model that was used for calibration. We will use some additional tools in analyzing the data and make some important general points that were not brought out in Section 4.6.1 because different statistics were employed there.

Regression analysis, in its various forms, is unquestionably the most frequently used statistical method. One question that the user of regression must address is "How do I know when I have a regression model that is good enough to use?" A simple answer to this question is not possible, as the required "goodness" of the model depends upon the application. For example, if we were trying to predict stock market prices, we would be able to make a considerable amount of money if we could construct a model that explained just the majority of the variation in Y. At the other extreme, we will use a case study in this section to show that accounting for virtually 100% of the variation in Y may not be good enough.

We will illustrate this with some load cell calibration data that are from circa 1975 and were once the data of NIST scientist Paul Pontius, who is now deceased. The dependent variable (Y) is Deflection, and the independent variable (X) is Load. We would like to be able to model Y as a function of X.

The data are given below

Y	(repeat reading)	**X**
0.11019	0.11052	150000
0.21956	0.22018	300000
0.32949	0.32939	450000
0.43899	0.43886	600000
0.54803	0.54798	750000
0.65694	0.65739	900000
0.76562	0.76596	1050000
0.87487	0.87474	1200000
0.98292	0.98300	1350000
1.09146	1.09150	1500000
1.20001	1.20004	1650000
1.30822	1.30818	1800000
1.41599	1.41613	1950000
1.52399	1.52408	2100000
1.63194	1.63159	2250000
1.73947	1.73965	2400000
1.84646	1.84696	2550000
1.95392	1.95445	2700000
2.06128	2.06177	2850000
2.16844	2.16829	3000000

and since the X-values are so large, we will divide them by 10^4, as a matter of convenience. (This does not affect any statistic or graph that is used in assessing the worth of the model.)

Without any prior information to suggest a specific nonlinear model to fit, our logical starting point would be to plot the data and see if a simple linear regression model would likely provide an adequate fit

The scatter plot is as follows:

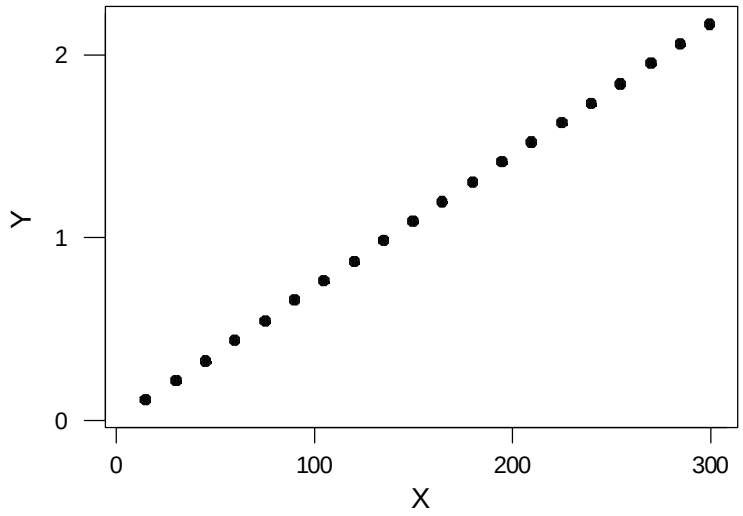


Figure 1

We would be very suspicious of such a scatter plot in almost any field of application as datasets simply do not look this perfect. Since this is calibration data, however, we should not be surprised by a very strong linear relationship between the two variables, and indeed this is what we would hope to see.

Since there are repeat readings, before we attempt to fit a model, we might want to see if there is a discernible pattern in the differences of the repeat readings. A plot of the first reading minus the second reading against X does not exhibit any unusual pattern, but it is somewhat curious that 13 of the 20 differences are negative, and 5 of those differences exceed the largest positive difference in absolute value. Due to the latter, a paired-t test of the first reading versus the second reading has a p-value of .031, so it would have been of interest to determine if there were an explanation for the generally larger second readings.

Basic computer output for simple linear regression does not deviate much across statistical software. Accordingly the output given below is typical:

The regression equation is
Y = 0.00615 + 0.00722 X

Predictor	Coef	SE Coef	T	P
Constant	0.0061497	0.0007132	8.62	0.00
X	0.00722103	0.00000397	1819.29	0.00

S = 0.002171 R-Sq = 100.0% R-Sq(adj) = 100.0%

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	1	15.604	15.604	3.310E+06	0.00
Residual (Error)	38	0.000	0.000		
Total	39	15.604			

Unusual Observations

Obs	X	Y	Fit	SE Fit	Residual	St Resid
1	15	0.11019	0.11447	0.00066	-0.00428	-2.07R
40	300	2.16829	2.17246	0.00066	-0.00417	-2.02R

R denotes an observation with a large standardized residual

This output provides essentially the same message that can be gleaned from the scatter plot; namely that the model provides almost an exact fit to the data. Since there are non-zero residual values, as the last part of the output shows, the fit is not

exact; the R^2 value of 100% is simply rounded off. (The last part of the output shows nothing to be concerned about as under the assumption of normality for the errors there should theoretically be two standardized residuals that exceed 2 in absolute value and somewhat coincidentally that is how many there are.)

There is a commonly used numerical method and some graphical methods that can be used to determine if the model can be improved. The numerical method is a "lack-of-fit" test. With this test the residual sum of squares is broken down into a pure error component and a lack-of-fit component. The former is variability in Y that cannot be fit by any model, and is reflected by points that line up vertically on a scatter plot when there is a single independent variable (X).

As is shown by the list of data, all of the X-values are repeated. Figure 1 does not show vertical scatter, however, simply because the two Y-values at each X-value are very close together. This plus the fact that the residual sum of squares is so small, (0.000) to three decimal places, prevents us from seeing the magnitude of the pure error relative to the magnitude of the lack of fit component.

We can see this numerically when we do a lack-of-fit test, however, and the test result is given below, in an abbreviated table, with more decimal places added so that non-zero tabular entries are displayed.

Analysis of Variance

Source	DF	SS	MS	F	P
Residual Error	38	0.00017915			
Lack of Fit	18	0.00017823	0.0000991	214.75	0.00
Pure Error	20	0.00000092	0.000000046		

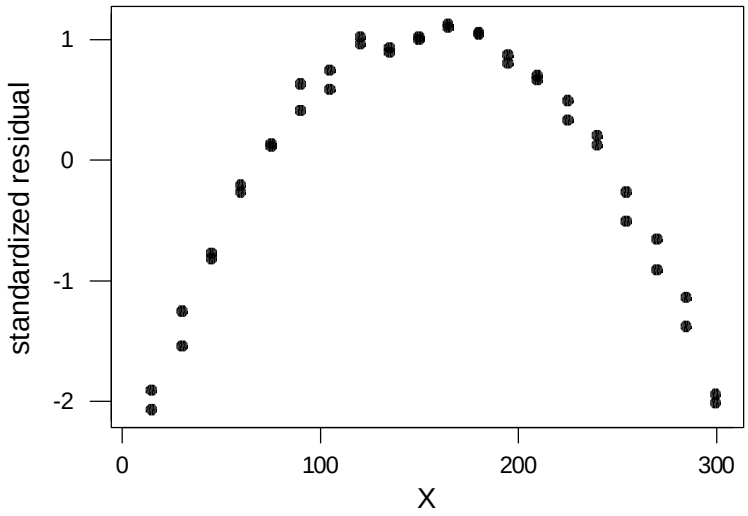
The test shows that the lack of fit is significant, as the F-statistic is very large.

The test was possible in this instance because the X-values were repeated. (At least one X-value must be repeated for the test to be performed.) When the X-values occur at random, we can't expect to observe repeats, so other methods, such as the method of nearest neighbors, must be employed in grouping the data and performing an approximate test.

Although the lack-of-fit test tells us that the model can be improved, the test result does not tell us how to do so. There are various types of residual plots that can be employed for this purpose, and more than one plot should be used since there is no guarantee that any one plot will give the appropriate message in a given application.

A commonly used plot is a plot of the standardized residuals against either X or $\hat{Y}$. (The plots will have the same general configuration when $\hat{\beta}_1$ is positive, and will be mirror images when $\hat{\beta}_1$ is negative.) Of course one of these two plots will rarely be needed in simple linear regression as the scatter plot will generally be suggestive of the type of (nonlinear) term to add to the model.

This is one of those rare occasions when the scatter plot is uninformative, however. Given below is the standardized residuals plot.



The graph strongly suggests that a quadratic term should be added to the model, and when the term is added the results are as follows.

The regression equation is

$$Y = 0.000674 + 0.00732X - 0.000000X^2$$

Predictor	Coef	SE Coef	T	P
Constant	0.0006736	0.0001079	6.24	0.000
X	0.00732059	0.00000158	4638.65	0.000
X^2	-0.00000032	0.00000000	-64.95	0.000

S = 0.0002052 R-Sq = 100.0% R-Sq(adj) = 100.0%

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	2	15.6040	7.8020	1.853E+08	0.000
Residual Error	37	0.0000	0.0000		

Total	39	15.6040

Source	DF	Seq SS
C3	1	15.6039
C4	1	0.0002

Unusual Observations

Obs	C3	C1	Fit	SE Fit	Residual	St Resid
2	30	0.21956	0.22001	0.00007	-0.00045	-2.33R
17	255	1.84646	1.84687	0.00005	-0.00041	-2.07R
26	90	0.65739	0.65697	0.00004	0.00042	2.11R
39	285	2.06177	2.06137	0.00007	0.00040	2.09R

R denotes an observation with a large standardized residual

The output shows that the quadratic term is statistically significant, although its contribution is quite small. Since statistically significant results won't always have practical significance, it is desirable to look beyond these results. With the linear term only in the model, the average value of $|Y - \hat{Y}|$ is 0.00183, compared with 0.00016 when the quadratic term is additionally in the model. There is thus essentially one decimal place difference and subject-matter specialists say that this degree of improvement is important (and thus practically significant) in calibration work.

As stated, no one residual plot can be expected to always give the correct signal. Certain facts are known, however. In particular, since the raw residuals are orthogonal to the Xs and to $\hat{Y}$, a line fit through the points on a plot of the raw residuals against either one of the Xs or against $\hat{Y}$ will have a slope of zero, and when standardized residuals are used, the slope will be very close to zero. This

means that a standardized residuals plot will generally correctly identify the need for a term for which the configuration of plotted points that would suggest the term would logically have a correlation of zero with the horizontal axis variable. A quadratic term falls into this category, but reciprocal and log terms do not, as a plot that suggests the need for those terms would have standardized residuals that are highly correlated with the horizontal axis variable.

A plot that is more likely to detect the need for such a term is a partial residuals plot, which is a plot of $e + \widehat{\beta}_i X_i$ against X_i. That plot fails for this example, however, because the linear term overwhelms e with the result that the plot is almost exactly a straight line, which shows the strong linear relationship that we observed earlier.

See Chapter 5 of Ryan (1997) and Cook (1993, 1994) for information on other types of residual plots and more information on when certain types of residual plots should be effective.

REFERENCES

Cook, R. D. (1993). Exploring partial residual plots. *Technometrics*, **35**, 351-362

Cook, R. D. (1994). On the interpretation of regression plots. *Journal of the American Statistical Association*, **89**, 177-189.

Ryan, T. P. (1997). *Modern Regression Methods*. New York: Wiley.