14. Regression

- A. Introduction to Simple Linear Regression
- B. Partitioning Sums of Squares
- C. Standard Error of the Estimate
- D. Inferential Statistics for b and r
- E. Influential Observations
- F. Regression Toward the Mean
- G. Introduction to Multiple Regression
- H. Exercises

This chapter is about prediction. Statisticians are often called upon to develop methods to predict one variable from other variables. For example, one might want to predict college grade point average from high school grade point average. Or, one might want to predict income from the number of years of education.

Introduction to Linear Regression

by David M. Lane

Prerequisites

• Chapter 3: Measures of Variability

• Chapter 4: Describing Bivariate Data

Learning Objectives

1. Define linear regression

2. Identify errors of prediction in a scatter plot with a regression line

In simple linear regression, we predict scores on one variable from the scores on a second variable. The variable we are predicting is called the *criterion variable* and is referred to as Y. The variable we are basing our predictions on is called the *predictor variable* and is referred to as X. When there is only one predictor variable, the prediction method is called *simple regression*. In simple linear regression, the topic of this section, the predictions of Y when plotted as a function of X form a straight line.

The example data in Table 1 are plotted in Figure 1. You can see that there is a positive relationship between X and Y. If you were going to predict Y from X, the higher the value of X, the higher your prediction of Y.

Table 1. Example data.

х	Y
1.00	1.00
2.00	2.00
3.00	1.30
4.00	3.75
5.00	2.25

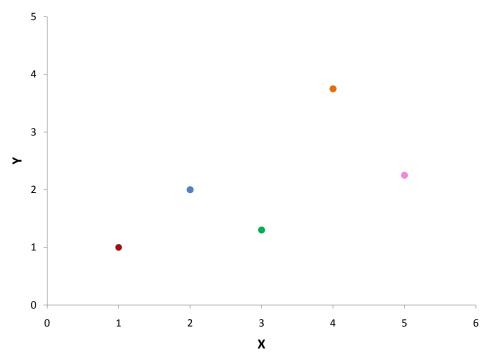


Figure 1. A scatter plot of the example data.

Linear regression consists of finding the best-fitting straight line through the points. The best-fitting line is called a regression line. The black diagonal line in Figure 2 is the regression line and consists of the predicted score on Y for each possible value of X. The vertical lines from the points to the regression line represent the errors of prediction. As you can see, the red point is very near the regression line; its error of prediction is small. By contrast, the yellow point is much higher than the regression line and therefore its error of prediction is large.

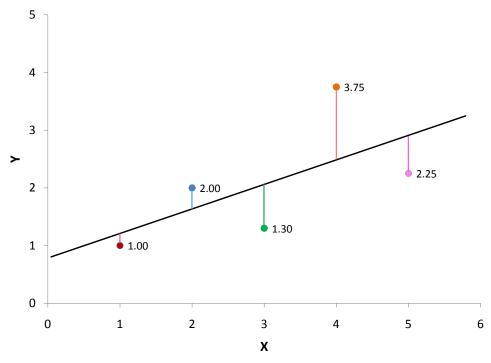


Figure 2. A scatter plot of the example data. The black line consists of the predictions, the points are the actual data, and the vertical lines between the points and the black line represent errors of prediction.

The error of prediction for a point is the value of the point minus the predicted value (the value on the line). Table 2 shows the predicted values (Y') and the errors of prediction (Y-Y'). For example, the first point has a Y of 1.00 and a predicted Y of 1.21. Therefore, its error of prediction is -0.21.

Table 2. Example data.

Х	Y	Y'	Y-Y'	(Y-Y') ²
1.00	1.00	1.210	-0.210	0.044
2.00	2.00	1.635	0.365	0.133
3.00	1.30	2.060	-0.760	0.578
4.00	3.75	2.485	1.265	1.600
5.00	2.25	2.910	-0.660	0.436

You may have noticed that we did not specify what is meant by "best-fitting line." By far the most commonly used criterion for the best-fitting line is the line that minimizes the sum of the squared errors of prediction. That is the criterion that was

used to find the line in Figure 2. The last column in Table 2 shows the squared errors of prediction. The sum of the squared errors of prediction shown in Table 2 is lower than it would be for any other regression line.

The formula for a regression line is

$$Y' = bX + A$$

where Y' is the predicted score, b is the slope of the line, and A is the Y intercept. The equation for the line in Figure 2 is

$$Y' = 0.425X + 0.785$$

For X = 1,

$$Y' = (0.425)(1) + 0.785 = 1.21.$$

For X = 2,

$$Y' = (0.425)(2) + 0.785 = 1.64.$$

Computing the Regression Line

In the age of computers, the regression line is typically computed with statistical software. However, the calculations are relatively easy are given here for anyone who is interested. The calculations are based on the statistics shown in Table 3. M_X is the mean of X, M_Y is the mean of Y, S_X is the standard deviation of Y, and Y is the correlation between Y and Y.

Table 3. Statistics for computing the regression line

Mx	M _Y	Sχ	Sy	r
3	2.06	1.581	1.072	0.627

The slope (b) can be calculated as follows:

$$b = r \frac{s_y}{s_r}$$

and the intercept (A) can be calculated as

$$A = M_Y - bM_X$$
.

For these data,

$$b = (0.627) \frac{1.072}{1.581} = 0.425$$

$$A = 2.06 - (0.425)(3) = 0.785$$

Note that the calculations have all been shown in terms of sample statistics rather than population parameters. The formulas are the same; simply use the parameter values for means, standard deviations, and the correlation.

Standardized Variables

The regression equation is simpler if variables are *standardized* so that their means are equal to 0 and standard deviations are equal to 1, for then b = r and A = 0. This makes the regression line:

$$Z_Y' = (r)(Z_X)$$

where $Z_{Y'}$ is the predicted standard score for Y, r is the correlation, and Z_X is the standardized score for X. Note that the slope of the regression equation for standardized variables is r.

Figure 3 shows a scatterplot with the regression line predicting the standardized Verbal SAT from the standardized Math SAT.

A Real Example

The case study, "SAT and College GPA" contains high school and university grades for 105 computer science majors at a local state school. We now consider how we could predict a student's university GPA if we knew his or her high school GPA.

Figure 3 shows a scatter plot of University GPA as a function of High School GPA. You can see from the figure that there is a strong positive relationship. The correlation is 0.78. The regression equation is

Univ GPA' =
$$(0.675)$$
 (High School GPA) + 1.097

Therefore, a student with a high school GPA of 3 would be predicted to have a university GPA of

University GPA' = (0.675)(3) + 1.097 = 3.12.

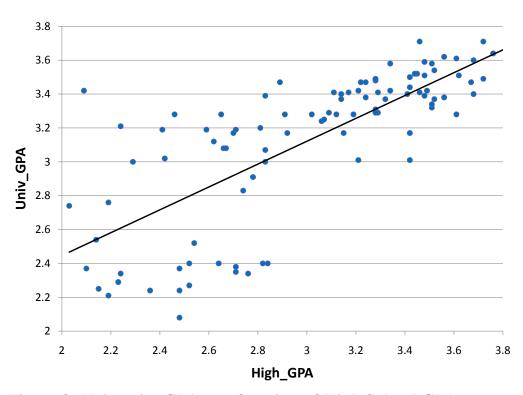


Figure 3. University GPA as a function of High School GPA.

Assumptions

It may surprise you, but the calculations shown in this section are assumption free. Of course, if the relationship between X and Y is not linear, a different shaped function could fit the data better. *Inferential statistics* in regression are based on several assumptions, and these assumptions are presented in a later section of this chapter.

Partitioning the Sums of Squares

by David M. Lane

Prerequisites

• Chapter 14: Introduction to Linear Regression

Learning Objectives

- 1. Compute the sum of squares Y
- 2. Convert raw scores to deviation scores
- 3. Compute predicted scores from a regression equation
- 4. Partition sum of squares Y into sum of squares predicted and sum of squares error
- 5. Define r² in terms of sum of squares explained and sum of squares Y

One useful aspect of regression is that it can divide the variation in Y into two parts: the variation of the predicted scores and the variation in the errors of prediction. The variation of Y is called the sum of squares Y and is defined as the sum of the squared deviations of Y from the mean of Y. In the population, the formula is

$$SSY = \sum (Y - \mu_Y)^2$$

where SSY is the sum of squares Y, Y is an individual value of Y, and my is the mean of Y. A simple example is given in Table 1. The mean of Y is 2.06 and SSY is the sum of the values in the third column and is equal to 4.597.

Table 1. Example of SSY.

Y	Y-m _y	(Y-m _y) ²
1.00	-1.06	1.1236
2.00	-0.06	0.0036
1.30	-0.76	0.5776
3.75	1.69	2.8561
2.25	0.19	0.0361

When computed in a sample, you should use the sample mean, M, in place of the population mean:

$$SSY = \sum (Y - M_Y)^2$$

It is sometimes convenient to use formulas that use *deviation scores* rather than raw scores. Deviation scores are simply deviations from the mean. By convention, small letters rather than capitals are used for deviation scores. Therefore, the score, y indicates the difference between Y and the mean of Y. Table 2 shows the use of this notation. The numbers are the same as in Table 1.

Table 2. Example of SSY using Deviation Scores.

Y	у	y²
1.00	-1.06	1.1236
2.00	-0.06	0.0036
1.30	-0.76	0.5776
3.75	1.69	2.8561
2.25	0.19	0.0361
10.30	0.00	4.5970

The data in Table 3 are reproduced from the introductory section. The column X has the values of the *predictor variable* and the column Y has the *criterion variable*. The third column, y, contains the differences between the column Y and the mean of Y.

X	Y	у	y ²	Υ'	y'	y' ²	Y-Y'	(Y-Y') ²
1.00	1.00	-1.06	1.1236	1.210	-0.850	0.7225	-0.210	0.044
2.00	2.00	-0.06	0.0036	1.635	-0.425	0.1806	0.365	0.133
3.00	1.30	-0.76	0.5776	2.060	0.000	0.000	-0.760	0.578
4.00	3.75	1.69	2.8561	2.485	0.425	0.1806	1.265	1.600
5.00	2.25	0.19	0.0361	2.910	0.850	0.7225	-0.660	0.436
15.00	10.30	0.00	4.597	10.300	0.000	1.806	0.000	2.791

Table 3. Example data. The last row contains column sums.

The fourth column, y^2 , is simply the square of the y column. The column Y' contains the predicted values of Y. In the introductory section, it was shown that the equation for the regression line for these data is

$$Y' = 0.425X + 0.785.$$

The values of Y' were computed according to this equation. The column y' contains deviations of Y' from the mean of Y' and y'² is the square of this column. The next-to-last column, Y-Y', contains the actual scores (Y) minus the predicted scores (Y'). The last column contains the squares of these errors of prediction.

We are now in a position to see how the SSY is partitioned. Recall that SSY is the sum of the squared deviations from the mean. It is therefore the sum of the y^2 column and is equal to 4.597. SSY can be partitioned into two parts: the sum of squares predicted (SSY') and the sum of squares error (SSE). The sum of squares predicted is the sum of the squared deviations of the predicted scores from the mean predicted score. In other words, it is the sum of the y^2 column and is equal to 1.806. The sum of squares error is the sum of the squared errors of prediction. It is therefore the sum of the $(Y-Y')^2$ column and is equal to 2.791. This can be summed up as:

There are several other notable features about Table 3. First, notice that the sum of y and the sum of y' are both zero. This will always be the case because these variables were created by subtracting their respective means from each value. Also, notice that the mean of Y-Y' is 0. This indicates that although some Y values are higher than their respective predicted Y values and some are lower, the average difference is zero.

The SSY is the total variation, the SSY' is the variation explained, and the SSE is the variation unexplained. Therefore, the proportion of variation explained can be computed as:

$$Proportion\ explained = \frac{SSY'}{SSY}$$

Similarly, the proportion not explained is:

$$Proportion \ not \ explained = \frac{SSE}{SSY}$$

There is an important relationship between the proportion of variation explained and Pearson's correlation: r^2 is the proportion of variation explained. Therefore, if r=1, then, naturally, the proportion of variation explained is 1; if r=0, then the proportion explained is 0. One last example: for r=0.4, the proportion of variation explained is 0.16.

Since the variance is computed by dividing the variation by N (for a population) or N-1 (for a sample), the relationships spelled out above in terms of variation also hold for variance. For example,

$$\sigma_{total}^2 = \sigma_{y'}^2 + \sigma_e^2$$

where the first term is the variance total, the second term is the variance of Y', and the last term is the variance of the errors of prediction (Y-Y'). Similarly, r² is the proportion of variance explained as well as the proportion of variation explained.

Summary Table

It is often convenient to summarize the partitioning of the data in a table such as Table 4. The *degrees of freedom* column (df) shows the degrees of freedom for each source of variation. The degrees of freedom for the sum of squares explained

is equal to the number of predictor variables. This will always be 1 in simple regression. The error degrees of freedom is equal to the total number of observations minus 2. In this example, it is 5 - 2 = 3. The total degrees of freedom is the total number of observations minus 1.

Table 4. Summary Table for Example Data

Source	Sum of Squares	df	Mean Square
Explained	1.806	1	1.806
Error	2.791	3	0.930
Total	4.597	4	

Standard Error of the Estimate

by David M. Lane

Prerequisites

- Chapter 3: Measures of Variability
- Chapter 14: Introduction to Linear Regression
- Chapter 14: Partitioning Sums of Squares

Learning Objectives

- 1. Make judgments about the size of the standard error of the estimate from a scatter plot
- 2. Compute the standard error of the estimate based on errors of prediction
- 3. Compute the standard error using Pearson's correlation
- 4. Estimate the standard error of the estimate based on a sample

Figure 1 shows two regression examples. You can see that in Graph A, the points are closer to the line than they are in Graph B. Therefore, the predictions in Graph A are more accurate than in Graph B.

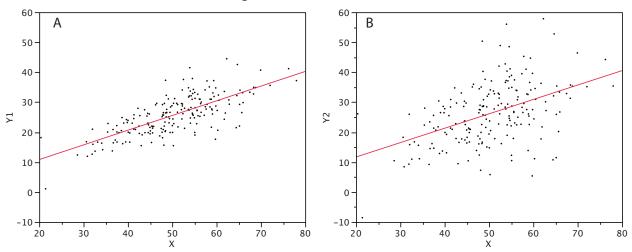


Figure 1. Regressions differing in accuracy of prediction.

The standard error of the estimate is a measure of the accuracy of predictions. Recall that the regression line is the line that minimizes the sum of squared deviations of prediction (also called the sum of squares error). The standard error of the estimate is closely related to this quantity and is defined below:

$$\sigma_{\it est} = \sqrt{rac{\sum{(Y - Y')^2}}{N}}$$

where σ_{est} is the standard error of the estimate, Y is an actual score, Y' is a predicted score, and N is the number of pairs of scores. The numerator is the sum of squared differences between the actual scores and the predicted scores.

Note the similarity of the formula for σ_{est} to the formula for σ :

$$\sigma = \sqrt{\frac{\sum (Y - \mu)^2}{N}}$$

In fact, σ_{est} is the standard deviation of the errors of prediction (each Y - Y' is an error of prediction).

Assume the data in Table 1 are the data from a population of five X, Y pairs.

Table 1. Example data.

	х	Y	Y'	Y-Y'	(Y-Y') ²
	1.00	1.00	1.210	-0.210	0.044
	2.00	2.00	1.635	0.365	0.133
	3.00	1.30	2.060	-0.760	0.578
	4.00	3.75	2.485	1.265	1.600
	5.00	2.25	2.910	-0.660	0.436
Sum	15.00	10.30	10.30	0.000	2.791

The last column shows that the sum of the squared errors of prediction is 2.791. Therefore, the standard error of the estimate is

$$\sigma_{est} = \sqrt{\frac{2.791}{5}} = 0.747$$

There is a version of the formula for the standard error in terms of Pearson's correlation:

$$\sigma_{est} = \sqrt{\frac{(1 - \rho^2)SSY}{N}}$$

where o is the population value of Pearson's correlation and SSY is

$$SSY = \sum (Y - \mu_Y)^2$$

For the data in Table 1, $m_v = 10.30$, SSY = 4.597 and r = 0.6268. Therefore,

$$\sigma_{est} = \sqrt{\frac{(1 - 0.6268^2)(4.597)}{5}} = \sqrt{\frac{2.791}{5}} = 0.747$$

which is the same value computed previously.

Similar formulas are used when the standard error of the estimate is computed from a sample rather than a population. The only difference is that the denominator is N-2 rather than N. The reason N-2 is used rather than N-1 is that two parameters (the slope and the intercept) were estimated in order to estimate the sum of squares. Formulas for a sample comparable to the ones for a population are shown below:

$$s_{est} = \sqrt{\frac{\sum (Y - Y')^2}{N - 2}}$$

$$s_{est} = \sqrt{\frac{2.791}{3}} = 0.964$$

$$s_{est} = \sqrt{\frac{(1 - r^2)SSY}{N - 2}}$$

Inferential Statistics for b and r

by David M. Lane

Prerequisites

- Chapter 9: Sampling Distribution of r
- Chapter 9: Confidence Interval for r

Learning Objectives

- 1. State the assumptions that inferential statistics in regression are based upon
- 2. Identify heteroscedasticity in a scatter plot
- 3. Compute the standard error of a slope
- 4. Test a slope for significance
- 5. Construct a confidence interval on a slope
- 6. Test a correlation for significance
- 7. Construct a confidence interval on a correlation

This section shows how to conduct significance tests and compute confidence intervals for the regression slope and Pearson's correlation. As you will see, if the regression slope is significantly different from zero, then the correlation coefficient is also significantly different from zero.

Assumptions

Although no assumptions were needed to determine the best-fitting straight line, assumptions are made in the calculation of inferential statistics. Naturally, these assumptions refer to the population, not the sample.

- 1. Linearity: The relationship between the two variables is linear.
- 2. Homoscedasticity: The variance around the regression line is the same for all values of X. A clear violation of this assumption is shown in Figure 1. Notice that the predictions for students with high high-school GPAs are very good, whereas the predictions for students with low high-school GPAs are not very good. In other words, the points for students with high high-school GPAs are close to the regression line, whereas the points for low high-school GPA students are not.

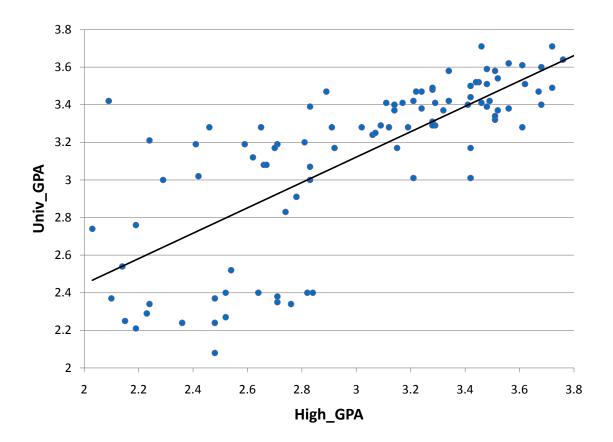


Figure 1. University GPA as a function of High School GPA.

3. The errors of prediction are distributed normally. This means that the distributions of deviations from the regression line are normally distributed. It does not mean that X or Y is normally distributed.

Significance Test for the Slope (b)

Recall the general formula for a t test:

$$t = \frac{statistics - hypothesized\ value}{estimated\ standard\ error\ of\ the\ statistic}$$

As applied here, the statistic is the sample value of the slope (b) and the hypothesized value is 0. The degrees of freedom for this test are:

$$df = N-2$$

where N is the number of pairs of scores.

The estimated standard error of b is computed using the following formula:

$$s_b = \frac{s_{est}}{\sqrt{SSX}}$$

where s_b is the estimated standard error of b, s_{est} is the standard error of the estimate, and SSX is the sum of squared deviations of X from the mean of X. SSX is calculated as

$$SSX = \sum (X - M_{\chi})^2$$

where M_x is the mean of X. As shown previously, the standard error of the estimate can be calculated as

$$s_{est} = \sqrt{\frac{(1 - r^2)SSY}{N - 2}}$$

These formulas are illustrated with the data shown in Table 1. These data are reproduced from the introductory section. The column X has the values of the *predictor variable* and the column Y has the values of the *criterion variable*. The third column, x, contains the differences between the values of column X and the mean of X. The fourth column, x^2 , is the square of the x column. The fifth column, y, contains the differences between the values of column Y and the mean of Y. The last column, y^2 , is simply the square of the y column.

Table 1. Example data.

	х	Y	х	X ²	у	y²
	1.00	1.00	-2.00	4	-1.06	1.1236
	2.00	2.00	-1.00	1	-0.06	0.0036
	3.00	1.30	0.00	0	-0.76	0.5776
	4.00	3.75	1.00	1	1.69	2.8561
	5.00	2.25	2.00	4	0.19	0.0361
Sum	15.00	10.30	0.00	10.00	0.00	4.5970

The computation of the standard error of the estimate (s_{est}) for these data is shown in the section on the standard error of the estimate. It is equal to 0.964.

$$s_{est} = 0.964$$

SSX is the sum of squared deviations from the mean of X. It is, therefore, equal to the sum of the x^2 column and is equal to 10.

$$SSX = 10.00$$

We now have all the information to compute the standard error of b:

$$s_b = \frac{0.964}{\sqrt{10}} = 0.305$$

As shown previously, the slope (b) is 0.425. Therefore,

$$t = \frac{0.425}{0.305} = 1.39$$

$$df = N-2 = 5-2 = 3$$
.

The p value for a two-tailed t test is 0.26. Therefore, the slope is not significantly different from 0.

Confidence Interval for the Slope

The method for computing a confidence interval for the population slope is very similar to methods for computing other confidence intervals. For the 95% confidence interval, the formula is:

```
lower limit: b - (t.95)(s_b)
upper limit: b + (t.95)(s_b)
```

where t_{.95} is the value of t to use for the 95% confidence interval.

The values of t to be used in a confidence interval can be looked up in a table of the t distribution. A small version of such a table is shown in Table 2. The first column, df, stands for degrees of freedom.

T 11 A			
Table 2	Abbre	viated	t table

df	0.95	0.99
2	4.303	9.925
3	3.182	5.841
4	2.776	4.604
5	2.571	4.032
8	2.306	3.355
10	2.228	3.169
20	2.086	2.845
50	2.009	2.678
100	1.984	2.626

You can also use the "inverse t distribution" calculator (<u>external link</u>; requires Java) to find the t values to use in a confidence interval.

Applying these formulas to the example data,

```
lower limit: 0.425 - (3.182)(0.305) = -0.55
upper limit: 0.425 + (3.182)(0.305) = 1.40
```

Significance Test for the Correlation

The formula for a significance test of Pearson's correlation is shown below:

$$t = \frac{r\sqrt{N-2}}{\sqrt{1-r^2}}$$

where N is the number of pairs of scores. For the example data,

$$t = \frac{0.627\sqrt{5 - 2}}{\sqrt{1 - 0.627^2}} = 1.39$$

Notice that this is the same t value obtained in the t test of b. As in that test, the degrees of freedom is N-2 = 5-2 = 3.

Influential Observations

by David M. Lane

Prerequisites

• Chapter 14: Introduction to Linear Regression

Learning Objectives

- 1. Define "influence"
- 2. Describe what makes a point influential
- 3. Define "leverage"
- 4. Define "distance"

It is possible for a single observation to have a great influence on the results of a regression analysis. It is therefore important to be alert to the possibility of influential observations and to take them into consideration when interpreting the results.

Influence

The influence of an observation can be thought of in terms of how much the predicted scores for other observations would differ if the observation in question were not included. Cook's D is a good measure of the influence of an observation and is proportional to the sum of the squared differences between predictions made with all observations in the analysis and predictions made leaving out the observation in question. If the predictions are the same with or without the observation in question, then the observation has no influence on the regression model. If the predictions differ greatly when the observation is not included in the analysis, then the observation is influential.

A common rule of thumb is that an observation with a value of Cook's D over 1.0 has too much influence. As with all rules of thumb, this rule should be applied judiciously and not thoughtlessly.

An observation's influence is a function of two factors: (1) how much the observation's value on the predictor variable differs from the mean of the predictor variable and (2) the difference between the predicted score for the observation and its actual score. The former factor is called the observation's leverage. The latter factor is called the observation's *distance*.

Calculation of Cook's D (Optional)

The first step in calculating the value of Cook's D for an observation is to predict all the scores in the data once using a regression equation based on all the observations and once using all the observations except the observation in question. The second step is to compute the sum of the squared differences between these two sets of predictions. The final step is to divide this result by 2 times the MSE (see the section on partitioning the variance).

Leverage

The *leverage* of an observation is based on how much the observation's value on the predictor variable differs from the mean of the predictor variable. The greater an observation's leverage, the more potential it has to be an influential observation. For example, an observation with the mean on the predictor variable has no influence on the slope of the regression line regardless of its value on the criterion variable. On the other hand, an observation that is extreme on the predictor variable has, depending on its distance, the potential to affect the slope greatly.

Calculation of Leverage (h)

The first step is to standardize the predictor variable so that it has a mean of 0 and a standard deviation of 1. Then, the leverage (h) is computed by squaring the observation's value on the standardized predictor variable, adding 1, and dividing by the number of observations.

Distance

The distance of an observation is based on the error of prediction for the observation: The greater the error of prediction, the greater the distance. The most commonly used measure of distance is the *studentized residual*. The studentized residual for an observation is closely related to the error of prediction for that observation divided by the standard deviation of the errors of prediction. However, the predicted score is derived from a regression equation in which the observation in question is not counted. The details of the computation of a studentized residual are a bit complex and are beyond the scope of this work.

An observation with a large distance will not have that much influence if its leverage is low. It is the combination of an observation's leverage and distance that determines its influence.

Example

Table 1 shows the leverage, studentized residual, and influence for each of the five observations in a small dataset.

Table 1. Example Data.

ID	Х	Υ	h	R	D
А	1	2	0.39	-1.02	0.40
В	2	3	0.27	-0.56	0.06
С	3	5	0.21	0.89	0.11
D	4	6	0.20	1.22	0.19
Е	8	7	0.73	-1.68	8.86
h is the le	verage l	R is the s	tudentized residual	and D is Cook's measur	e of influence

h is the leverage, R is the studentized residual, and D is Cook's measure of influence.

Observation A has fairly high leverage, a relatively high residual, and moderately high influence.

Observation B has small leverage and a relatively small residual. It has very little influence.

Observation C has small leverage and a relatively high residual. The influence is relatively low.

Observation D has the lowest leverage and the second highest residual.

Although its residual is much higher than Observation A, its influence is much less because of its low leverage.

Observation E has by far the largest leverage and the largest residual. This combination of high leverage and high residual makes this observation extremely influential.

Figure 1 shows the regression line for the whole dataset (blue) and the regression line if the observation in question is not included (red) for all observations. The observation in question is circled. Naturally, the regression line for the whole dataset is the same in all panels. The residual is calculated relative to the line for which the observation in question is not included in the analysis. This can be seen most clearly for Observation E which lies very close to the regression line

computed when it is included but very far from the regression line when it is excluded from the calculation of the line.

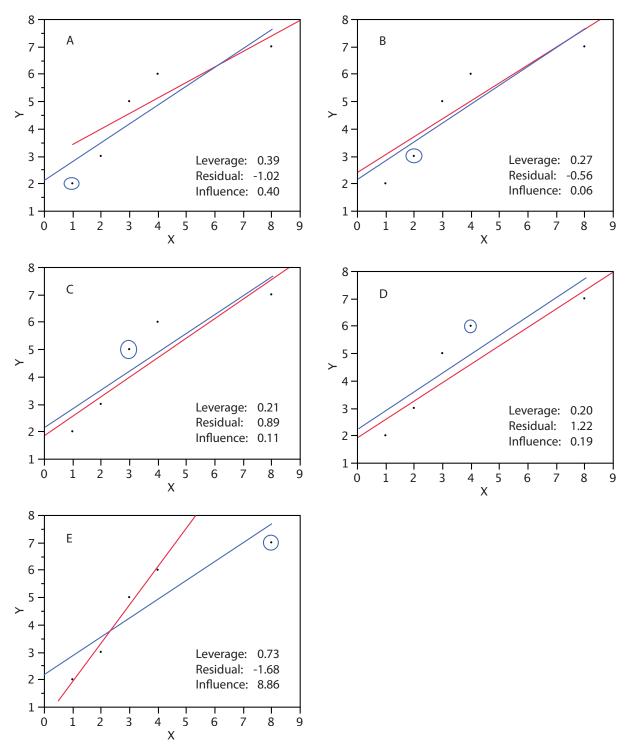


Figure 1. Illustration of leverage, residual, and influence.

The circled points are not included in the calculation of the red regression line. All points are included in the calculation of the blue regression line.

Regression Toward the Mean

by David M. Lane

Prerequisites

• Chapter 14: Regression Introduction

Learning Objectives

- 1. Explain what regression towards the mean is
- 2. State the conditions under which regression toward the mean occurs
- 3. Identify situations in which neglect of regression toward the mean leads to incorrect conclusions
- 4. Explain how regression toward the mean relates to a regression equation.

Regression toward the mean involves outcomes that are at least partly due to chance. We begin with an example of a task that is entirely chance: Imagine an experiment in which a group of 25 people each predicted the outcomes of flips of a fair coin. For each subject in the experiment, a coin is flipped 12 times and the subject predicts the outcome of each flip. Figure 1 shows the results of a simulation of this "experiment." Although most subjects were correct from 5 to 8 times out of 12, one simulated subject was correct 10 times. Clearly, this subject was very lucky and probably would not do as well if he or she performed the task a second time. In fact, the best prediction of the number of times this subject would be correct on the retest is 6 since the probability of being correct on a given trial is 0.5 and there are 12 trials.

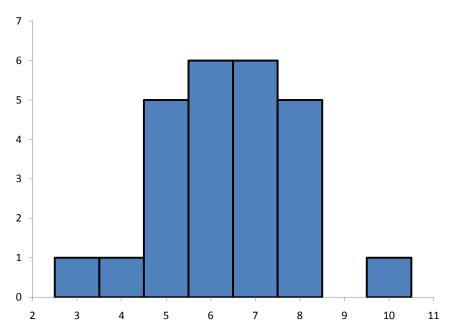


Figure 1. Histogram of results of a simulated experiment.

More technically, the best prediction for the subject's result on the retest is the mean of the binomial distribution with N = 12 and p = 0.50. This distribution is shown in Figure 2 and has a mean of 6.

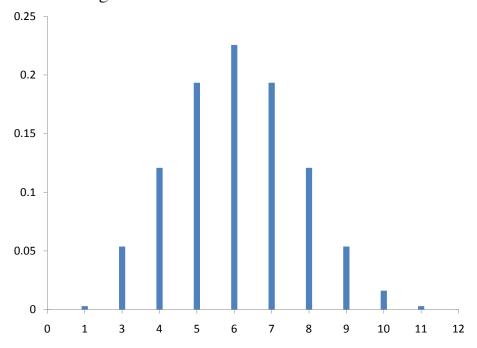


Figure 2. Binomial Distribution for N=12 and p=.50.

The point here is that no matter how many coin flips a subject predicted correctly, the best prediction of their score on a retest is 6.

Now we consider a test we will call "Test A" that is partly chance and partly skill: Instead of predicting the outcomes of 12 coin flips, each subject predicts the outcomes of 6 coin flips and answers 6 true/false questions about world history. Assume that the mean score on the 6 history questions is 4. A subject's score on Test A has a large chance component but also depends on history knowledge. If a subject scored very high on this test (such as a score of 10/12), it is likely that they did well on both the history questions and the coin flips. For example, if they only got four of the history questions correct, they would have had to have gotten all six of the coin predictions correct, and this would have required exceptionally good luck. If given a second test (Test B) that also included coin predictions and history questions, their knowledge of history would be helpful and they would again be expected to score above the mean. However, since their high performance on the coin portion of Test A would not be predictive of their coin performance on Test B, they would not be expected to fare as well on Test B as on Test A. Therefore, the best prediction of their score on Test B would be somewhere between their score on Test A and the mean of Test B. This tendency of subjects with high values on a measure that includes chance and skill to score closer to the mean on a retest is called "regression toward the mean."

The essence of the regression-toward-the-mean phenomenon is that people with high scores tend to be above average in skill and in luck, and that only the skill portion is relevant to future performance. Similarly, people with low scores tend to be below average in skill and luck and their bad luck is not relevant to future performance. This does not mean that all people who score high have above average luck. However, on average they do.

Almost every measure of behavior has a chance and a skill component to it. Take a student's grade on a final exam as an example. Certainly, the student's knowledge of the subject will be a major determinant of his or her grade. However, there are aspects of performance that are due to chance. The exam cannot cover everything in the course and therefore must represent a subset of the material. Maybe the student was lucky in that the one aspect of the course the student did not understand well was not well represented on the test. Or, maybe, the student was not sure which of two approaches to a problem would be better but, more or less by chance, chose the right one. Other chance elements come into play as well. Perhaps the student was awakened early in the morning by a random phone call,

resulting in fatigue and lower performance. And, of course, guessing on multiple choice questions is another source of randomness in test scores.

There will be regression toward the mean in a test-retest situation whenever there is less than a perfect (r = 1) relationship between the test and the retest. This follows from the formula for a regression line with standardized variables shown below.

$$Z_{Y'} = (r) (Z_X)$$

From this equation it is clear that if the absolute value of r is less than 1, then the predicted value of Z_Y will be closer to 0, the mean for standardized scores, than is Z_X . Also, note that if the correlation between X and Y is 0, as it would be for a task that is all luck, the predicted standard score for Y is its mean, 0, regardless of the score on X.

Figure 3 shows a scatter plot with the regression line predicting the standardized Verbal SAT from the standardized Math SAT. Note that the slope of the line is equal to the correlation of 0.835 between these variables.

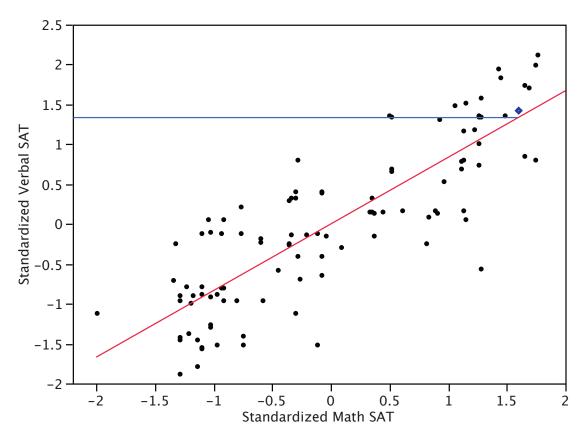


Figure 3. Prediction of Standardized Verbal SAT from Standardized Math SAT.

The point represented by a blue diamond has a value of 1.6 on the standardized Math SAT. This means that this student scored 1.6 standard deviations above the mean on Math SAT. The predicted score is (r)(1.6) = (0.835)(1.6) = 1.34. The horizontal line on the graph shows the value of the predicted score. The key point is that although this student scored 1.6 standard deviations above the mean on Math SAT, he or she is only predicted to score 1.34 standard deviations above the mean on Verbal SAT. Thus, the prediction is that the Verbal SAT score will be closer to the mean of 0 than is the Math SAT score. Similarly, a student scoring far below the mean on Math SAT will be predicted to score higher on Verbal SAT.

Regression toward the mean occurs in any situation in which observations are selected on the basis of performance on a task that has a random component. If you choose people on the basis of their performance on such a task, you will be choosing people partly on the basis of their skill and partly on the basis of their luck on the task. Since their luck cannot be expected to be maintained from trial to trial, the best prediction of a person's performance on a second trial will be

somewhere between their performance on the first trial and the mean performance on the first trial. The degree to which the score is expected to "regress toward the mean" in this manner depends on the relative contributions of chance and skill to the task: the greater the role of chance, the more the regression toward the mean.

Errors Resulting From Failure to Understand Regression Toward the Mean

Failure to appreciate regression toward the mean is common and often leads to incorrect interpretations and conclusions. One of the best examples is provided by Nobel Laureate Daniel Kahneman in his autobiography (external link). Dr. Kahneman was attempting to teach flight instructors that praise is more effective than punishment. He was challenged by one of the instructors who relayed that in his experience praising a cadet for executing a clean maneuver is typically followed by a lesser performance, whereas screaming at a cadet for bad execution is typically followed by improved performance. This, of course, is exactly what would be expected based on regression toward the mean. A pilot's performance, although based on considerable skill, will vary randomly from maneuver to maneuver. When a pilot executes an extremely clean maneuver, it is likely that he or she had a bit of luck in their favor in addition to their considerable skill. After the praise but not because of it, the luck component will probably disappear and the performance will be lower. Similarly, a poor performance is likely to be partly due to bad luck. After the criticism but not because of it, the next performance will likely be better. To drive this point home, Kahneman had each instructor perform a task in which a coin was tossed at a target twice. He demonstrated that the performance of those who had done the best the first time deteriorated, whereas the performance of those who had done the worst improved.

Regression toward the mean is frequently present in sports performance. A good example is provided by Schall and Smith (2000), who analyzed many aspects of baseball statistics including the batting averages of players in 1998. They chose the 10 players with the highest batting averages (BAs) in 1998 and checked to see how well they did in 1999. According to what would be expected based on regression toward the mean, these players should, on average, have lower batting averages in 1999 than they did in 1998. As can be seen in Table 1, 7/10 of the players had lower batting averages in 1999 than they did in 1998. Moreover, those who had higher averages in 1999 were only slightly higher, whereas those who

were lower were much lower. The average decrease from 1998 to 1999 was 33 points. Even so, most of these players had excellent batting averages in 1999 indicating that skill was an important component of their 1998 averages.

Table 1. How the Ten Players with the Highest BAs in 1998 did in 1999.

1998	1999	Difference
363	379	16
354	298	-56
339	342	3
337	281	-56
336	249	-87
331	298	-33
328	297	-31
328	303	-25
327	257	-70
327	332	5

Figure 4 shows the batting averages of the two years. The decline from 1998 to 1999 is clear. Note that although the mean decreased from 1998, some players increased their batting averages. This illustrates that regression toward the mean does not occur for every individual. Although the predicted scores for every individual will be lower, some of the predictions will be wrong.

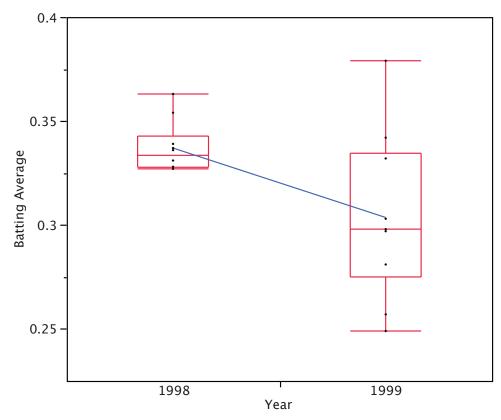


Figure 4. Quantile plots of the batting averages. The line connects the means of the plots.

Regression toward the mean plays a role in the so-called "Sophomore Slump," a good example of which is that a player who wins "rookie of the year" typically does less well in his second season. A related phenomenon is called the Sports Illustrated Cover Jinx.

An experiment without a control group can *confound* regression effects with real effects. For example, consider a hypothetical experiment to evaluate a reading-improvement program. All first graders in a school district were given a reading achievement test and the 50 lowest-scoring readers were enrolled in the program. The students were retested following the program and the mean improvement was large. Does this necessarily mean the program was effective? No, it could be that the initial poor performance of the students was due, in part, to bad luck. Their luck would be expected to improve in the retest, which would increase their scores with or without the treatment program.

For a real example, consider an experiment that sought to determine whether the drug propranolol would increase the SAT scores of students thought to have test anxiety (external link). Propranolol was given to 25 high-school students chosen because IQ tests and other academic performance indicated that they had not done as well as expected on the SAT. On a retest taken after receiving propranolol, students improved their SAT scores an average of 120 points. This was a *significantly* greater increase than the 38 points expected simply on the basis of having taken the test before. The problem with the study is that the method of selecting students likely resulted in a disproportionate number of students who had bad luck when they first took the SAT. Consequently, these students would likely have increased their scores on a retest with or without the propranolol. This is not to say that propranolol had no effect. However, since possible propranolol effects and regression effects were confounded, no firm conclusions should be drawn.

Randomly assigning students to either the propranolol group or a control group would have improved the experimental design. Since the regression effects would then not have been systematically different for the two groups, a significant difference would have provided good evidence for a propranolol effect.

Introduction to Multiple Regression

by David M. Lane

Prerequisites

- Chapter 14: Simple Linear Regression
- Chapter 14: Partitioning Sums of Squares
- Chapter 14: Standard Error of the Estimate
- Chapter 14: Inferential Statistics for b and r

Learning Objectives

- 1. State the regression equation
- 2. Define "regression coefficient"
- 3. Define "beta weight"
- 4. Explain what R is and how it is related to r
- 5. Explain why a regression weight is called a "partial slope"
- 6. Explain why the sum of squares explained in a multiple regression model is usually less than the sum of the sums of squares in simple regression
- 7. Define R² in terms of proportion explained
- 8. Test R² for significance
- 9. Test the difference between a complete and reduced model for significance
- 10. State the assumptions of multiple regression and specify which aspects of the analysis require assumptions

In simple linear regression, a criterion variable is predicted from one predictor variable. In multiple regression, the criterion is predicted by two or more variables. For example, in the SAT case study, you might want to predict a student's university grade point average on the basis of their High-School GPA (HSGPA) and their total SAT score (verbal + math). The basic idea is to find a linear combination of HSGPA and SAT that best predicts University GPA (UGPA). That is, the problem is to find the values of b₁ and b₂ in the equation shown below that gives the best predictions of UGPA. As in the case of simple linear regression, we define the best predictions as the predictions that minimize the squared errors of prediction.

$$UGPA' = b_1HSGPA + b_2SAT + A$$

where UGPA' is the predicted value of University GPA and A is a constant. For these data, the best prediction equation is shown below:

$$UGPA' = 0.541 \times HSGPA + 0.008 \times SAT + 0.540$$

In other words, to compute the prediction of a student's University GPA, you add up (a) their High-School GPA multiplied by 0.541, (b) their SAT multiplied by 0.008, and (c) 0.540. Table 1 shows the data and predictions for the first five students in the dataset.

TT 11 1	1 D	1 D	1.	, •
Table 1	ı. Data	and Pr	eaic	tions.

HSGPA	SAT	UGPA'
3.45	1232	3.38
2.78	1070	2.89
2.52	1086	2.76
3.67	1287	3.55
3.24	1130	3.19

The values of b (b1 and b2) are sometimes called "regression coefficients" and sometimes called "regression weights." These two terms are synonymous.

The *multiple correlation* (R) is equal to the correlation between the predicted scores and the actual scores. In this example, it is the correlation between UGPA' and UGPA, which turns out to be 0.79. That is, R = 0.79. Note that R will never be negative since if there are negative correlations between the predictor variables and the criterion, the regression weights will be negative so that the correlation between the predicted and actual scores will be positive.

Interpretation of Regression Coefficients

A regression coefficient in multiple regression is the slope of the linear relationship between the criterion variable and the part of a predictor variable that is independent of all other predictor variables. In this example, the regression coefficient for HSGPA can be computed by first predicting HSGPA from SAT and saving the errors of prediction (the differences between HSGPA and HSGPA'). These errors of prediction are called "residuals" since they are what is left over in HSGPA after the predictions from SAT are subtracted, and they represent the part

of HSGPA that is independent of SAT. These residuals are referred to as HSGPA.SAT, which means they are the residuals in HSGPA after having been predicted by SAT. The correlation between HSGPA.SAT and SAT is necessarily 0.

The final step in computing the regression coefficient is to find the slope of the relationship between these residuals and UGPA. This slope is the regression coefficient for HSGPA. The following equation is used to predict HSGPA from SAT:

$$HSGPA' = -1.314 + 0.0036 \times SAT$$

The residuals are then computed as:

The linear regression equation for the prediction of UGPA by the residuals is

$$UGPA' = 0.541 \times HSGPA.SAT + 3.173$$

Notice that the slope (0.541) is the same value given previously for b_1 in the multiple regression equation.

This means that the regression coefficient for HSGPA is the slope of the relationship between the criterion variable and the part of HSPGA that is *independent* of (uncorrelated with) the other predictor variables. It represents the change in the criterion variable associated with a change of one in the predictor variable when all other predictor variables are held constant. Since the regression coefficient for HSGPA is 0.54, this means that, holding SAT constant, a change of one in HSGPA is associated with a change of 0.54 in UGPA. If two students had the same SAT and differed in HSGPA by 2, then you would predict they would differ in UGPA by (2)(0.54) = 1.08. Similarly, if they differed by 0.5, then you would predict they would differ by (0.50)(0.54) = 0.27.

The slope of the relationship between the part of a predictor variable independent of other predictor variables and the criterion is its partial slope. Thus the regression coefficient of 0.541 for HSGPA and the regression coefficient of 0.008 for SAT are partial slopes. Each partial slope represents the relationship between the predictor variable and the criterion holding constant all of the other predictor variables.

It is difficult to compare the coefficients for different variables directly because they are measured on different scales. A difference of 1 in HSGPA is a fairly large difference, whereas a difference of 1 on the SAT is negligible. Therefore, it can be advantageous to transform the variables so that they are on the same scale. The most straightforward approach is to standardize the variables so that they all have a standard deviation of 1. A regression weight for standardized variables is called a "beta weight" and is designated by the Greek letter β . For these data, the beta weights are 0.625 and 0.198. These values represent the change in the criterion (in standard deviations) associated with a change of one standard deviation on a predictor [holding constant the value(s) on the other predictor(s)]. Clearly, a change of one standard deviation on HSGPA is associated with a larger difference than a change of one standard deviation of SAT. In practical terms, this means that if you know a student's HSGPA, knowing the student's SAT does not aid the prediction of UGPA much. However, if you do not know the student's HSGPA, his or her SAT can aid in the prediction since the β weight in the simple regression predicting UGPA from SAT is 0.68. For comparison purposes, the β weight in the simple regression predicting UGPA from HSGPA is 0.78. As is typically the case, the partial slopes are smaller than the slopes in simple regression.

Partitioning the Sums of Squares

Just as in the case of simple linear regression, the sum of squares for the criterion (UGPA in this example) can be partitioned into the sum of squares predicted and the sum of squares error. That is,

$$SSY = SSY' + SSE$$

which for these data:

$$20.798 = 12.961 + 7.837$$

The sum of squares predicted is also referred to as the "sum of squares explained." Again, as in the case of simple regression,

In simple regression, the proportion of variance explained is equal to r^2 ; in multiple regression, the proportion of variance explained is equal to R^2 .

In multiple regression, it is often informative to partition the sums of squares explained among the predictor variables. For example, the sum of squares explained for these data is 12.96. How is this value divided between HSGPA and SAT? One approach that, as will be seen, does not work is to predict UGPA in separate simple regressions for HSGPA and SAT. As can be seen in Table 2, the sum of squares in these separate simple regressions is 12.64 for HSGPA and 9.75 for SAT. If we add these two sums of squares we get 22.39, a value much larger than the sum of squares explained of 12.96 in the multiple regression analysis. The explanation is that HSGPA and SAT are highly correlated (r = .78) and therefore much of the variance in UGPA is confounded between HSGPA or SAT. That is, it could be explained by either HSGPA or SAT and is counted twice if the sums of squares for HSGPA and SAT are simply added.

Table 2. Sums of Squares for Various Predictors

Predictors	Sum of Squares
HSGPA	12.64
SAT	9.75
HSGPA and SAT	12.96

Table 3 shows the partitioning of the sums of squares into the sum of squares uniquely explained by each predictor variable, the sum of squares confounded between the two predictor variables, and the sum of squares error. It is clear from this table that most of the sum of squares explained is confounded between HSGPA and SAT. Note that the sum of squares uniquely explained by a predictor variable is analogous to the partial slope of the variable in that both involve the relationship between the variable and the criterion with the other variable(s) controlled.

Table 3. Partitioning the Sum of Squares

Source	Sum of Squares	Proportion
HSGPA (unique)	3.21	0.15
SAT (unique)	0.32	0.02
HSGPA and SAT (Confounded)	9.43	0.45
Error	7.84	0.38
Total	20.80	1.00

The sum of squares uniquely attributable to a variable is computed by comparing two regression models: the complete model and a reduced model. The complete model is the multiple regression with all the predictor variables included (HSGPA and SAT in this example). A reduced model is a model that leaves out one of the predictor variables. The sum of squares uniquely attributable to a variable is the sum of squares for the complete model minus the sum of squares for the reduced model in which the variable of interest is omitted. As shown in Table 2, the sum of squares for the complete model (HSGPA and SAT) is 12.96. The sum of squares for the reduced model in which HSGPA is omitted is simply the sum of squares explained using SAT as the predictor variable and is 9.75. Therefore, the sum of squares uniquely attributable to HSGPA is 12.96 - 9.75 = 3.21. Similarly, the sum of squares uniquely attributable to SAT is 12.96 - 12.64 = 0.32. The confounded sum of squares in this example is computed by subtracting the sum of squares uniquely attributable to the predictor variables from the sum of squares for the complete model: 12.96 - 3.21 - 0.32 = 9.43. The computation of the confounded sums of squares in analyses with more than two predictors is more complex and beyond the scope of this text.

Since the variance is simply the sum of squares divided by the degrees of freedom, it is possible to refer to the proportion of variance explained in the same way as the proportion of the sum of squares explained. It is slightly more common to refer to the proportion of variance explained than the proportion of the sum of squares explained and, therefore, that terminology will be adopted frequently here.

When variables are highly correlated, the variance explained uniquely by the individual variables can be small even though the variance explained by the variables taken together is large. For example, although the proportions of variance

explained uniquely by HSGPA and SAT are only 0.15 and 0.02 respectively, together these two variables explain 0.62 of the variance. Therefore, you could easily underestimate the importance of variables if only the variance explained uniquely by each variable is considered. Consequently, it is often useful to consider a set of related variables. For example, assume you were interested in predicting job performance from a large number of variables some of which reflect cognitive ability. It is likely that these measures of cognitive ability would be highly correlated among themselves and therefore no one of them would explain much of the variance independent of the other variables. However, you could avoid this problem by determining the proportion of variance explained by all of the cognitive ability variables considered together as a set. The variance explained by the set would include all the variance explained uniquely by the variables in the set as well as all the variance confounded among variables in the set. It would not include variance confounded with variables outside the set. In short, you would be computing the variance explained by the set of variables that is independent of the variables not in the set.

Inferential Statistics

We begin by presenting the formula for testing the significance of the contribution of a set of variables. We will then show how special cases of this formula can be used to test the significance of R^2 as well as to test the significance of the unique contribution of individual variables.

The first step is to compute two regression analyses: (1) an analysis in which all the predictor variables are included and (2) an analysis in which the variables in the set of variables being tested are **excluded**. The former regression model is called the "complete model" and the latter is called the "reduced model." The basic idea is that if the reduced model explains much less than the complete model, then the set of variables excluded from the reduced model is important.

The formula for testing the contribution of a group of variables is:

$$F = \frac{\frac{SSQ_c - SSQ_R}{P_c - P_R}}{\frac{SSQ_T - SSQ_c}{N - P_c - 1}} = \frac{MS_{explained}}{MS_{error}}$$

where:

 SSQ_C is the sum of squares for the complete model,

 SSQ_R is the sum of squares for the reduced model,

 p_C is the number of predictors in the complete model,

 p_R is the number of predictors in the reduced model,

 ${\sf SSQ}_T$ is the sum of squares total (the sum of squared deviations of the criterion variable from its mean), and

N is the total number of observations

The degrees of freedom for the numerator is p_c - p_r and the degrees of freedom for the denominator is N - p_C -1. If the F is significant, then it can be concluded that the variables excluded in the reduced set contribute to the prediction of the criterion variable independently of the other variables.

This formula can be used to test the significance of R^2 by defining the reduced model as having no predictor variables. In this application, SSQ_R and $p_R = 0$. The formula is then simplified as follows:

$$F_{(pc,N-pc-1)} = \frac{\frac{SSQ_C}{p_C}}{\frac{SSQ_T - SSQ_C}{N - p_C - 1}} = \frac{MS_{\text{explained}}}{MS_{error}}$$

which for this example becomes:

$$F = \frac{\frac{12.96}{2}}{\frac{20.80 - 12.96}{105 - 2 - 1}} = \frac{6.48}{0.08} = 84.35.$$

The degrees of freedom are 2 and 102. The F distribution calculator shows that p < 0.001.

The reduced model used to test the variance explained uniquely by a single predictor consists of all the variables except the predictor variable in question. For example, the reduced model for a test of the unique contribution of HSGPA contains only the variable SAT. Therefore, the sum of squares for the reduced model is the sum of squares when UGPA is predicted by SAT. This sum of squares is 9.75. The calculations for F are shown below:

$$F_{(1,102)} = \frac{\frac{12.96 - 9.75}{2 - 1}}{\frac{20.80 - 12.96}{105 - 2 - 1}} = \frac{3.212}{0.077} = 41.80.$$

The degrees of freedom are 1 and 102. The F distribution calculator shows that p < 0.001.

Similarly, the reduced model in the test for the unique contribution of SAT consists of HSGPA.

$$F = \frac{\frac{12.96 - 12.64}{2 - 1}}{\frac{20.80 - 12.96}{105 - 2 - 1}} = \frac{0.322}{0.077} = 4.19.$$

The degrees of freedom are 1 and 102. The F distribution calculator shows that p = 0.0432.

The significance test of the variance explained uniquely by a variable is identical to a significance test of the regression coefficient for that variable. A regression coefficient and the variance explained uniquely by a variable both reflect the relationship between a variable and the criterion independent of the other variables. If the variance explained uniquely by a variable is not zero, then the regression coefficient cannot be zero. Clearly, a variable with a regression coefficient of zero would explain no variance.

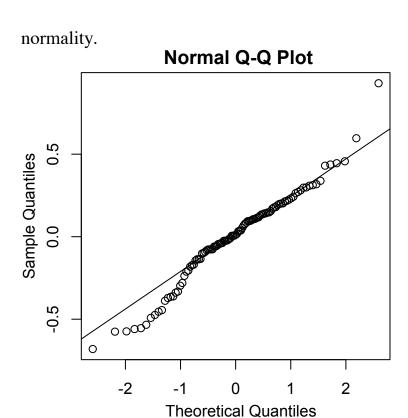
Other inferential statistics associated with multiple regression that are beyond the scope of this text. Two of particular importance are (1) confidence intervals on regression slopes and (2) confidence intervals on predictions for specific observations. These inferential statistics can be computed by standard statistical analysis packages such as R, SPSS, STATA, SAS, and JMP.

Assumptions

No assumptions are necessary for computing the regression coefficients or for partitioning the sums of squares. However, there are several assumptions made when interpreting inferential statistics. Moderate violations of Assumptions 1-3 do not pose a serious problem for testing the significance of predictor variables. However, even small violations of these assumptions pose problems for confidence intervals on predictions for specific observations.

1. Residuals are normally distributed:

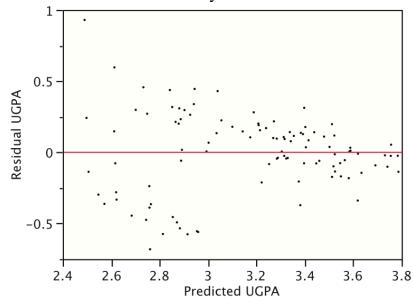
As in the case of simple linear regression, the residuals are the errors of prediction. Specifically, they are the differences between the actual scores on the criterion and the predicted scores. A Q-Q plot for the residuals for the example data is shown below. This plot reveals that the actual data values at the lower end of the distribution do not increase as much as would be expected for a normal distribution. It also reveals that the highest value in the data is higher than would be expected for the highest value in a sample of this size from a normal distribution. Nonetheless, the distribution does not deviate greatly from



2. Homoscedasticity:

It is assumed that the variance of the errors of prediction are the same for all predicted values. As can be seen below, this assumption is violated in the example data because the errors of prediction are much larger for observations with low-to-medium predicted scores than for observations with high predicted scores. Clearly, a confidence interval on a low predicted UGPA would

underestimate the uncertainty.



3. Linearity:

It is assumed that the relationship between each predictor variable and the criterion variable is linear. If this assumption is not met, then the predictions may systematically overestimate the actual values for one range of values on a predictor variable and underestimate them for another.

Statistical Literacy

by David M. Lane

Prerequisites

• Chapter 14: Regression Toward the Mean

In a discussion about the Dallas Cowboy football team, there was a comment that the quarterback threw far more interceptions in the first two games than is typical (there were two interceptions per game). The author correctly pointed out that, because of regression toward the mean, performance in the future is expected to improve. However, the author defined regression toward the mean as, "In nerd land, that basically means that things tend to even out over the long run."

What do you think?

Comment on that definition.

That definition is sort of correct, but it could be stated more precisely. Things don't always tend to even out in the long run. If a great player has an average game, then things wouldn't even out (to the average of all players) but would regress toward that player's high mean performance.

References

Schall, T., & Smith, G. (2000) Do Baseball Players Regress Toward the Mean? The *American Statistician*, 54, 231-235.

Exercises

Prerequisites

All material presented in the Regression chapter

- 1. What is the equation for a regression line? What does each term in the line refer to?
- 2. The formula for a regression equation is Y' = 2X + 9.
 - a. What would be the predicted score for a person scoring 6 on X?
 - b. If someone's predicted score was 14, what was this person's score on X?
- 3. What criterion is used for deciding which regression line fits best?
- 4. What does the standard error of the estimate measure? What is the formula for the standard error of the estimate?

5.

- a. In a regression analysis, the sum of squares for the predicted scores is 100 and the sum of squares error is 200, what is R2?
- b. In a different regression analysis, 40% of the variance was explained. The sum of squares total is 1000. What is the sum of squares of the predicted values?
- 6. For the X,Y data below, compute:
 - a. r and determine if it is significantly different from zero.
 - b. the slope of the regression line and test if it differs significantly from zero.
 - c. the 95% confidence interval for the slope.

X	Υ
4	6
3	7
5	12
11	17
10	9
14	21

- 7. What assumptions are needed to calculate the various inferential statistics of linear regression?
- 8. The correlation between years of education and salary in a sample of 20 people from a certain company is .4. Is this correlation statistically significant at the . 05 level?
- 9. A sample of X and Y scores is taken, and a regression line is used to predict Y from X. If SSY' = 300, SSE = 500, and N = 50, what is:
 - (a) SSY?
 - (b) the standard error of the estimate?
 - (c) R^2 ?
- 10. Using linear regression, find the predicted post-test score for someone with a score of 45 on the pre-test.

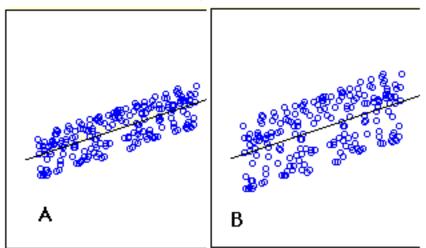
Pre	Post
59	56
52	63
44	55
51	50
42	66
42	48
41	58
45	36
27	13
63	50
54	81
44	56
50	64
47	50
55	63
49	57
45	73
57	63
46	46
60	60
65	47
64	73
50	58
74	85
59	44

11. The equation for a regression line predicting the number of hours of TV watched by children (Y) from the number of hours of TV watched by their parents (X) is Y' = 4 + 1.2X. The sample size is 12.

- a. If the standard error of b is .4, is the slope statistically significant at the .05 level?
- b. If the mean of X is 8, what is the mean of Y?
- 12. Based on the table below, compute the regression line that predicts Y from X.

M _X	M _Y	S _X	S _Y	r
10	12	2.5	3.0	-0.6

13. Does A or B have a larger standard error of the estimate?



- 14. True/false: If the slope of a simple linear regression line is statistically significant, then the correlation will also always be significant.
- 15. True/false: If the slope of the relationship between X an Y is larger for Population 1 than for Population 2, the correlation will necessarily be larger in Population 1 than in Population 2. Why or why not?
- 16. True/false: If the correlation is .8, then 40% of the variance is explained.
- 17. True/false: If the actual Y score was 31, but the predicted score was 28, then the error of prediction is 3.

Questions from Case Studies

Angry Moods (AM) case study

- 18. (AM) Find the regression line for predicting Anger-Out from Control-Out.
 - a. What is the slope?
 - b. What is the intercept?
 - c. Is the relationship at least approximately linear?
 - d. Test to see if the slope is significantly different from 0.
 - e. What is the standard error of the estimate?

SAT and GPA (SG) case study

- 19. (SG) Find the regression line for predicting the overall university GPA from the high school GPA.
 - a. What is the slope?
 - b. What is the y-intercept?
 - c. If someone had a 2.2 GPA in high school, what is the best estimate of his or her college GPA?
 - d. If someone had a 4.0 GPA in high school, what is the best estimate of his or her college GPA?

Driving (D) case study

- 20. (D) What is the correlation between age and how often the person chooses to drive in inclement weather? Is this correlation statistically significant at the .01 level? Are older people more or less likely to report that they drive in inclement weather?
- 21. (D) What is the correlation between how often a person chooses to drive in inclement weather and the percentage of accidents the person believes occur in inclement weather? Is this correlation significantly different from 0?
- 22. (D) Use linear regression to predict how often someone rides public transportation in inclement weather from what percentage of accidents that person thinks occur in inclement weather. (Pubtran by Accident)
 - (a) Create a scatter plot of this data and add a regression line.
 - (b) What is the slope?

- (c) What is the intercept?
- (d) Is the relationship at least approximately linear?
- (e) Test if the slope is significantly different from 0.
- (f) Comment on possible assumption violations for the test of the slope.
- (g) What is the standard error of the estimate?