

Regression Analysis : Review of the Basics

Forgotten Statistics
- Barron's 1996

Simple Regression

Simple regression involves finding the equation of a line that best fits a pattern of observations of two variables: an independent variable X and a dependent variable Y . Assume that the two variables are related by a linear equation of this form:

$$Y = a + bX$$

The slope of the line is b , and a is the y intercept, or constant term. However, the values of a and b are unknown. If you have n observations each for X and Y , you can determine the values of a and b that give the line that best fits the observations.

Let \hat{a} and \hat{b} be the estimated values for a and b that result from your regression calculation. Then, for each value X_i there is a corresponding predicted value of Y (call it \hat{Y}_i):

$$\hat{Y}_i = \hat{a} + \hat{b}X_i$$

The values of \hat{Y} all lie along the regression line. In each case we can determine the difference between the predicted value of Y and the actual value of Y associated with that value of X (call this the error):

$$\text{error}_i = Y_i - \hat{Y}_i$$

By squaring each of these errors and adding, we can find SE_{line} , the total squared error about the regression line:

$$SE_{line} = \sum_{i=1}^n (Y_i - \hat{Y}_i)^2$$

The goal of regression analysis is to choose the values of \hat{a} and \hat{b} that minimize the value of SE_{line} . These values are called the ordinary least squares estimators of a and b , and they are found from these formulas:

$$\text{slope} = \hat{b} = \frac{\bar{xy} - \bar{x}\bar{y}}{\bar{x^2} - \bar{x}^2}$$

$$\text{constant term} = \hat{a} = \bar{y} - \hat{b}\bar{x}$$

The bar over each quantity designates that it is an average. In most practical situations, you will use a computer to perform the calculations.

Here are some sample data:

| X | Y |
|----|---|
| 7 | 6 |
| 5 | 4 |
| 3 | 9 |
| 2 | 8 |
| 10 | 1 |

In reality you would not want to perform these calculations with only five observations; more observations would give you better estimators.

To find the slope and intercept of the regression line, set up a table like this:

| | X | Y | XY | X ² | Y ² |
|----------|-----|-----|----|----------------|----------------|
| | 7 | 6 | 42 | 49 | 36 |
| | 5 | 4 | 20 | 25 | 16 |
| | 3 | 9 | 27 | 9 | 81 |
| | 2 | 8 | 16 | 4 | 64 |
| | 10 | 1 | 10 | 100 | 1 |
| Average: | 5.4 | 5.6 | 23 | 37.4 | 39.6 |

Now use the formulas:

$$\text{slope} = \hat{b} = \frac{23 - 5.4 \times 5.6}{37.4 - 5.4^2} = -0.87864$$

$$\text{constant term} = \hat{a} = 5.6 - (-0.87864 \times 5.4) = 10.3447$$

Figure 9.21 shows the scatterplot for our five observations, with the regression line drawn in.

We also need to be able to measure whether this line fits the data very well. To do this, we calculate the r^2 value:

$$r^2 = 1 - \frac{SE_{line}}{SE_{avg}}$$

The quantity SE_{avg} is the squared error of the Y values about their average. The formula has the result that r^2 is always between 0 and 1, or $0 \leq r^2 \leq 1$. We can calculate r^2 from the following table:

| X | Y | \hat{Y} | error = $Y - \hat{Y}$ | $(Y - \hat{Y})^2$ | $Y - \bar{Y}$ | $(Y - \bar{Y})^2$ | |
|----|---|-----------|-----------------------|-------------------|---------------|-------------------|------|
| 7 | 6 | 4.194 | 1.806 | 3.261 | 0.4 | 0.16 | |
| 5 | 4 | 5.951 | -1.95 | 3.808 | -1.6 | 2.56 | |
| 3 | 9 | 7.709 | 1.291 | 1.667 | 3.4 | 11.56 | |
| 2 | 8 | 8.587 | -0.59 | 0.345 | 2.4 | 5.76 | |
| 10 | 1 | 1.558 | -0.56 | 0.312 | -4.6 | 21.16 | |
| | | | | $SE_{line} =$ | 9.393 | $SE_{avg} =$ | 41.2 |

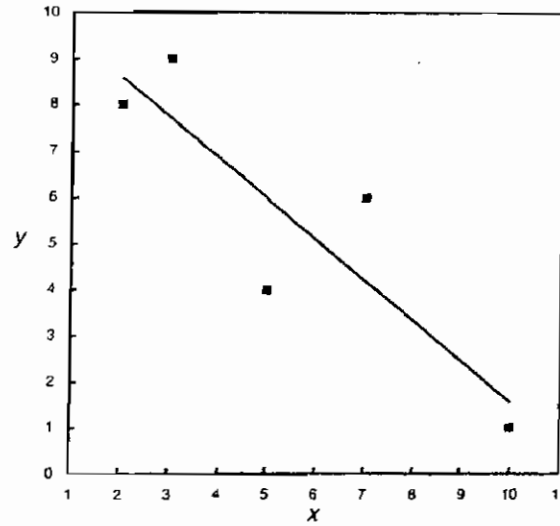


Figure 9.21: Scatterplot with regression line

We calculate \hat{Y} from the equation $10.3447 - .87864X$. The SE_{line} and SE_{avg} values are the sums of their respective columns. Now we can calculate r^2 :

$$r^2 = 1 - \frac{9.393}{41.2} = .772$$

This value means that 77.2% of the variation in Y is accounted for by variations in X . An r^2 value of 1 would mean a perfect fit; an r^2 value of zero would mean that you could predict Y just as well without knowing X as you can by knowing the regression equation.

The value of r^2 can also be calculated from this formula, since it is the square of the correlation coefficient:

$$r^2 = \frac{(\bar{xy} - \bar{x} \bar{y})^2}{(\bar{x}^2 - \bar{x}^2)(\bar{y}^2 - \bar{y}^2)} = \frac{(23 - 5.4 \times 5.6)^2}{(37.4 - 5.4^2)(39.6 - 5.6^2)} = .772$$

For situations where there is more than one independent variable, see **multiple regression**. If the true relation between x and y is of the form $y = ab^x$ or $y = mx^n$, see **logarithm**.

THE R SQUARED VALUE

We need more than just a description of the best line—we need a way to measure whether it is very good. (The best possible line still might not be very good.) The computer will report a value known as the r squared value (r^2) to indicate how well the relationship fits. These are the properties of the r^2 value:

1. r^2 is always between 0 and 1.
2. $r^2 = 1$ if all of the observations fit along a straight line, as when we looked for the relationship between the top number and bottom number on a die (see Figure 7.2).
3. $r^2 = 0$ if the two quantities are completely independent, as when we looked for a relationship between the numbers on two different dice (see Figure 7.1).
4. The r^2 value gives the percent of variation in y that can be accounted for by variations in x .

Suppose a contest is to be held between Rosencrantz and Guildenstern. Both will be trying to guess the value of a variable y . Rosencrantz has no information except for the average value of y . Guildenstern, on the other hand, knows in advance the value of x , and he knows the regression equation connecting y and x . The question is: Will Guildenstern do better than Rosencrantz at guessing y ?

Rosencrantz's best strategy is simply to guess that y will be equal to its average \bar{y} . For any given observation of y (call it y_i), there will be a certain amount of error between the actual value and Rosencrantz's prediction:

$$\text{error}_i = y_i - \bar{y}$$

To get the total error of Rosencrantz's plan, we will add these up (squaring the error first, so all of the errors become converted into positive numbers). Call this quantity SE_{avg} , for squared error about the average:

$$SE_{avg} = \sum_{i=1}^n (y_i - \bar{y})^2$$

This assumes that there are n observations.

Guildenstern might have an advantage over Rosencrantz because his guess for y will come from the regression equation:

$$\hat{y}_i = a + bx_i$$

As before, the hat placed over the y indicates that it is an estimator. Let e_i represent the error that Guildenstern will make from his prediction:

$$e_i = y_i - \hat{y}_i = y_i - (a + bx_i)$$

We can calculate the total squared error of Guildenstern's predictions, calling it SE_{line} , for squared error about the regression line:

$$SE_{line} = \sum_{i=1}^n (y_i - \hat{y}_i)^2 = \sum_{i=1}^n [y_i - (a + bx_i)]^2$$

(See Figure 7.8.)

Suppose that Guildenstern's squared error using the regression is just as large as Rosencrantz's, who did not use the regression. In that case the regression is totally worthless. Knowledge of x provides no help in predicting the value of y . On the other hand, suppose that SE_{line} is zero. In that case, Guildenstern is able to make perfect predictions using the regression analysis, and Rosencrantz doesn't stand a chance in the competition.

Therefore, we define the r^2 value as follows:

$$r^2 = 1 - \frac{SE_{line}}{SE_{avg}}$$

So, $r^2 = 1$ if SE_{line} is 0, and $r^2 = 0$ if $SE_{line} = SE_{avg}$.

One other possibility to consider is to suppose that the observations of x and y fit along a perfectly horizontal line. You would be tempted to say there is a relationship, because you can find a line that fits all the points. However, because the slope of this line is zero, it is clear that changing the value of x doesn't affect the value of y . It turns out in this case that you don't need x to predict the value of y , because y never changes. The r^2 value is undefined because it would require division by zero.

Multiple Regression

Consider a situation where one dependent variable (Y) can be expressed as a linear function of several independent variables (X_1, X_2, \dots, X_{m-1}) of this form:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_{m-1} X_{m-1}$$

The true values of the β 's are unknown; however, you do have n observations of each of the variables. Multiple regression analysis is used to estimate the values of the coefficients that provide the best fit to this equation. Note that there are m coefficients to estimate: one for each of the $m - 1$ independent variables, plus the constant term β_0 . (For the case where there is only one independent variable, see **simple regression**.) It is assumed that there is a random variable called the error term that accounts for variations in Y that are not explained by the equation. The error term has zero mean; it is often assumed to have a normal distribution with an unknown variance. That variance should be small if the multiple regression equation is to be very reliable.

The regression calculation reports a set of values B_0, B_1, \dots, B_{m-1} that are used as estimators for the true values $\beta_0, \beta_1, \dots, \beta_{m-1}$. Once values for the coefficients have been found, it is possible to find an estimated value \hat{Y}_i for each set of values of the independent variables:

$$\hat{Y}_i = B_0 + B_1X_{i1} + B_2X_{i2} + B_3X_{i3} + \dots + B_{m-1}X_{i(m-1)}$$

The quantity X_{ij} is the i th observation of the variable X_j . This estimated value can be compared with the true value Y for that observation. The goal of multiple regression is to minimize the sum of the squared deviations between the true value Y_i and the estimated value \hat{Y}_i :

$$\text{squared error} = \sum_{i=1}^n (Y_i - \hat{Y}_i)^2$$

The calculation procedure for the regression coefficients is very complicated, requiring a knowledge of matrix algebra. If the independent variables are arranged into a matrix \mathbf{X} of n rows and m columns (where each column represents the n observations of one of the independent variables, and one column consists solely of 1's, to take account of the constant term), and the dependent variable observations are arranged into an n by 1 matrix \mathbf{Y} , then the m -by-1 coefficient matrix \mathbf{B} comes from this formula:

$$\mathbf{B} = (\mathbf{X}^{\text{tr}}\mathbf{X})^{-1}\mathbf{X}^{\text{tr}}\mathbf{Y}$$

The matrix \mathbf{X}^{tr} is the transpose of \mathbf{X} , and $(\mathbf{X}^{\text{tr}}\mathbf{X})^{-1}$ is the inverse of the matrix formed by multiplying \mathbf{X}^{tr} by \mathbf{X} .

Fortunately, in practice you do not ever need to work with the formula, because a computer will do the calculations for you. The computer will report estimated coefficient values, as well as an r^2 value that tells you whether there is a good fit for the data. As with simple regression, an r^2 value of 1 means a perfect fit; an r^2 value of 0 means that 0% of the variation in the dependent variable is accounted for by variations in the independent variables.

The computer will also report a standard error for each coefficient; a larger value for the standard error means that there is more uncertainty about the true value of that coefficient. Dividing the estimated coefficient by the corresponding standard error gives a quantity known as the t statistic, which is used for hypothesis tests about whether or not a particular variable belongs in the regression equation. If the true value of that coefficient is zero, then the t statistic will come from a t distribution with $n - m$ degrees of freedom (n is the number of observations; m is the number of coefficients that are estimated, including the constant term). If the absolute value of the reported t statistic is greater than the absolute value of the critical value from the t distribution table, then reject the null hypothesis—the true coefficient is not zero, and the variable belongs.

The computer will also report an F statistic, which is used to test the hypothesis that the coefficients of all of the independent variables are zero. If the null hypothesis is true, and the

coefficients are all zero (meaning the regression calculation is worthless for predicting the value of Y), then the F statistic will come from an F distribution with $m - 1$ numerator degrees of freedom and $n - m$ denominator degrees of freedom. If the reported value is greater than the critical value for those degrees of freedom, then reject the null hypothesis.

In practice the difficult matter with regression analysis is determining which variables to include, and the exact form to use for the equation. You can test to see whether a variable that is included really belongs; however, you might have left out variables that should be included. The restriction that the equation be linear is not a big problem; if the true relationship involves a quadratic curve, such as $Y = aX_1 + bX_1^2$, then simply include both X_1 and X_1^2 as independent variables in your regression analysis. If the true relation is of the form $Y = X_1^{B_1} X_2^{B_2}$, see **logarithm** for information on transformations that convert it to a linear form.

Some other problems that can arise with multiple regression include **multicollinearity** (when two or more of the independent variables are highly correlated); **heteroscedasticity** (when the variances of the error terms are different for different observations); and **serial correlation** (when the errors for successive observations of time series data are correlated with each other). These problems make it more difficult for the regression calculation to estimate the coefficients accurately.

Here are some sample data:

| X_1 | X_2 | X_3 | X_4 | Y |
|-------|-------|-------|-------|-----|
| 2 | 1 | 3 | 0 | 12 |
| 3 | 3 | 8 | 0 | 23 |
| 2 | 5 | 3 | 1 | 29 |
| 2 | 5 | 7 | -1 | 27 |
| 5 | 1 | 3 | -1 | 20 |
| 5 | 1 | 8 | 0 | 21 |
| 4 | 6 | 4 | 1 | 39 |
| 5 | 5 | 8 | 0 | 37 |

The value of Y is given by the equation:

$$Y = 2 + 3X_1 + 4X_2 + X_4$$

However, remember that in reality we will not be able to see the true equation as we can in this artificial example. If we knew that X_1 , X_2 , and X_4 all should be included, then we would run the regression with those independent variables and we would find an r^2 value of 1, with each of the true coefficients found exactly. However, in reality we don't know for sure which variables should be included. Suppose we perform a multiple regression calculation with X_1 and X_2 as the independent variables. The results are:

| Variable | X_1 | X_2 | Constant term |
|----------------|---------|---------|---------------|
| Coefficient | 2.9780 | 4.1538 | 1.5581 |
| Standard error | 0.2191 | 0.1452 | 1.0339 |
| t statistic | 13.5890 | 28.6001 | 1.5070 |

The r^2 value is 0.9943. We do not have a perfect fit, because we left out the variable X_4 . However, X_4 has only a very small influence on Y , so leaving it out has not hurt our regression equation noticeably. The estimated coefficients (2.978 and 4.1538) are close to the true

values (3 and 4, respectively). The t statistics need to be compared against a t distribution with $8 - 3 = 5$ degrees of freedom, which gives a critical value of 2.571 using the 5% significance level (page 209). The two t statistics (13.589 and 28.6) are way above the critical value, so we can clearly reject the hypothesis that the true coefficients are zero.

The F statistic for this regression is reported to be 438.4; this needs to be compared against an F distribution with $3 - 1 = 2$ numerator degrees of freedom and $8 - 3 = 5$ denominator degrees of freedom. The critical value is 5.79 (page 211). The observed value is way above this limit, so the null hypothesis that both coefficients are truly zero can clearly be rejected.

Now suppose we perform a regression calculation with X_2 and X_3 as the independent variables. We know from the true equation that X_3 doesn't belong in the equation, but X_1 and X_4 do. Unfortunately, the researcher in the field does not see the true equation, and will not always know if important variables have been left out. In this case the resulting regression equation is:

| Variable | X_2 | X_3 | Constant term |
|----------------|--------|--------|---------------|
| Coefficient | 3.6834 | 0.4550 | 11.0663 |
| Standard error | 0.8464 | 0.7373 | 5.0216 |
| t statistic | 4.3520 | 0.617 | 2.2038 |

The r^2 value falls to .8001. The estimated coefficient for X_2 is still close to its true value of 4; its t statistic (4.352) is still above the critical value so we reject the null hypothesis that the true coefficient of X_2 is zero. The t statistic for X_3 falls inside the interval -2.571 to 2.571 , so we accept the null hypothesis that the true coefficient of X_3 is zero. This happens to be correct, because we know that X_3 is not included in the true equation. However, the regression results provide no way of testing for the fact that there are variables that should be included (X_1 and X_4) but are missing. In this case the missing variables do not hurt us too badly, but in other cases missing variables can wreak havoc on our ability to estimate the coefficients of the variables that are included.