

## Multiple Regression Theory

© 2006 Samuel L. Baker

Multiple regression is regression with two or more independent variables on the right-hand side of the equation. Use multiple regression if more than one cause is associated with the effect you wish to understand. That bears repeating:

### Why do multiple regression?

For prediction: Multiple regression lets you use more than one factor to make a prediction. Simple regression only allows you one causal factor.

For explanation: Multiple regression lets you separate causal factors, analyzing each one's influence on what you are trying to explain.

### The equation and the true plane

For the case of two independent variables, you can write the equation for a multiple regression model this way:

$$Y = \alpha + \beta X + \gamma Z + \text{error}$$

Imagine that the X- and Z-axes are on a table in front of you, with the X-axis pointing to the right and the Z-axis pointing directly away from you. The Y-axis is standing vertically, straight up from the table.

$Y = \alpha + \beta X + \gamma Z$  is the formula for a flat plane that is floating in the three-dimensional space above the table.

$\alpha$  is the height of the plane above the point on the table where  $X=0$  and  $Z=0$ .

$\beta$  is the slope of the plane in the X direction, how fast the plane rises as you go to the right. If  $\beta$  is bigger than 0, the plane is tilted so that the part to your right is higher than the part to your left.

$\gamma$  is the slope of the plane in the Z direction, how fast the plane rises as it goes away from you. If  $\gamma$  is bigger than 0, the plain is tilted toward you. If  $\gamma$  is negative, the plane is tilted away from you.

The error, in  $Y = \alpha + \beta X + \gamma Z + \text{error}$ , means that the data points do not lie right on this plane. Instead, the data points form a cluster or a cloud above and below the plane described by  $Y = \alpha + \beta X + \gamma Z$ .

When we collect data, we do not get to see the true plane. All we have is the cloud of points, floating in space. Multiple regression with two independent variables tries to find the plane that best fits that cloud of points, in the hope that the regression plane will be close to the true plane.

If you have more than two independent variables, it is conventional to go to a subscript notation for the variables and the slope parameters, like this:

$$Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_p X_p + \text{error}$$

This is why some statistics programs refer to the coefficients of the independent variables as “betas.”

With more than two independent variables, you are fitting hyper planes to data points floating in hyper space. That is not easy for most people to visualize.

**What  $\beta$  means:** The interpretation of the coefficient  $\beta$  in  $Y = \alpha + \beta X + \gamma Z + \text{error}$  is:

A change in  $X$  of 1 with no change in  $Z$  changes  $Y$  by  $\beta$ .

A change in  $X$  of  $\Delta X$  with no change in  $Z$  changes  $Y$  by  $\beta \Delta X$ .

$\beta$  is said to be the effect of a change in  $X$  on  $Y$  "holding  $Z$  constant." The phrase "holding ... constant" is used a lot, but it is a simplification. For example, suppose you one of your independent variables is time. Time might be the year number. It goes up by 1 every year. A time variable like that lets you measure a time trend. Does it make sense to say "holding time constant"? Not really. For that reason, here is a more precise way of saying it:

**What  $\beta$  means, longer version:**

When  $X$  is a unit bigger than what you would expect  $X$  to be, based on what  $Z$  is, then  $Y$  will tend to be  $\beta$  units bigger than what you would have predicted  $Y$  to have been, based on what  $Z$  is.

Whew! "The effect of  $Y$  on  $X$ , holding  $Z$  constant" is easier to say, but this longer way of saying it is more precise.

An example: Suppose  $Y$  is some measure of the ability to exercise.  $X$  is the person's age.  $Z$  is the person's weight. Our sample is people, all the same gender, aged 30 to 50. In that age range, there is a tendency for people to get heavier as they get older. That's certainly not true of every individual, but it tends to be true on average. If you know  $Z$ , a person's weight, you can make a guess about  $X$ , the person's age. This won't be a very good guess, but it will be a little bit better than guessing a person's age without knowing the person's weight.

When we use multiple regression to predict exercise ability based on age and weight, our equation is  $Y$  (Exercise ability) =  $\alpha + \beta \text{Age} + \gamma \text{Weight}$ . The meaning of  $\beta$ , the coefficient of Age, is this: When a person is one year older than you would expect, based on the person's weight, then exercise ability tends to be  $\beta$  units bigger than you would expect, based on the person's weight.

That is what we mean when we say that  $\beta$  measures the effect of Age, "holding" weight "constant." ( $\beta$  is probably negative, by the way.)

**Doing multiple regression by a series of simple regressions**

To further illustrate this idea of what a multiple regression coefficient means, you can do a multiple regression by means of a series of simple regressions.

Here is how you could estimate  $\beta$  in  $Y = \alpha + \beta X + \gamma Z$ , in five steps.

*1. Regress Y on Z.*

That means: Do a regression of the form  $Y = \alpha + \beta Z$ .

This is if you were trying to predict Y based on Z alone. In other words, as if you were trying to predict your dependent variable (Y) using the other independent variable, the one that is not X.

If you were doing this using your spreadsheet template, the B column would have the data for Z, and the C column would have the data for Y.

*2. Keep the residuals from this regression.*

Call those the Y-Z residuals.

Positive residuals from this regression indicate Y values that are higher than what you would predict, based on Z alone. Negative residuals indicate Y values that are lower than what you would predict, based on Z.

*3. Regress X on Z.*

That means: Do a regression of the form  $X = \alpha + \beta Z$ .

Now you are trying to predict X based on Z. In other words, you are trying to predict your X independent variable using the other independent variable.

If you were doing this using your spreadsheet template, the B column would now have the data for Z, and the C column would have the data for X.

*4. Keep the residuals from this regression.*

Call those the X-Z residuals.

Positive residuals indicate X values that are higher than what you would predict, based on the Z value. Negative residuals indicate X values that are lower than you would predict, based on the Z value.

*5. Regress the Y-Z residuals on the X-Z residuals.*

If you were doing this using your template, the B column would have the X-Z residuals and the C column would have the Y-Z residuals.

The slope coefficient in this regression is equal to the estimate of  $\beta$  in the multiple regression equation  $Y = \alpha + \beta X + \gamma Z$ .

**An Example of Simple Regressions to make a Multiple Regression**

Here is an example of how you could implement the above steps to do a multiple regression. The data are similar to what you will be using in the upcoming assignment.

These data are the results of a slightly botched agricultural experiment. We wanted to test the effect of different amounts of fertilizer on the amount of grain harvested. We set up seven plots of land and put from 100 to 700 pounds of fertilizer per acre on the plots. For each plot, we measured how many bushels of grain we got at harvest time. The botched part is that nature did not cooperate. Different plots got different amounts of rain during the growing season. Fortunately, we measured how much rain each plot got during the season. (Actually, we didn't do any of those things. These data are lifted from Wonnacott and Wonnacott, *Econometrics*, 1979.)

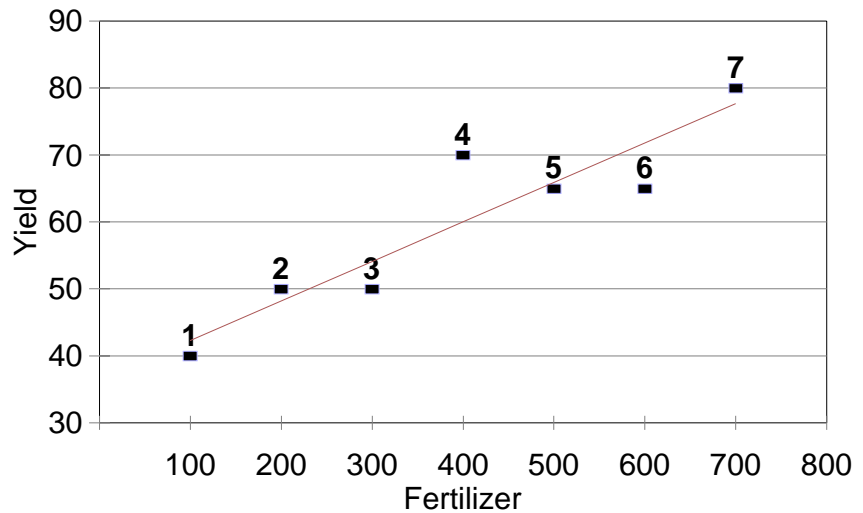
	A	B	C	D
1		Yield	Fertilizer	Rain
2	Plot 1	40	100	10
3	Plot 2	50	200	20
4	Plot 3	50	300	10
5	Plot 4	70	400	30
6	Plot 5	65	500	20
7	Plot 6	65	600	20
8	Plot 7	80	700	30

This graph shows the data on fertilizer and yield. Each data point is labeled with the plot number.

There is a clear positive relationship between fertilizer and yield. The simple least squares regression line, shown in this diagram, has a slope of 0.059. This indicates that, on average, adding one pound of fertilizer adds 0.059 bushels to the harvest per acre.

This may not be an valid assessment of the effect of fertilizer, because the plots got different amounts of rain.

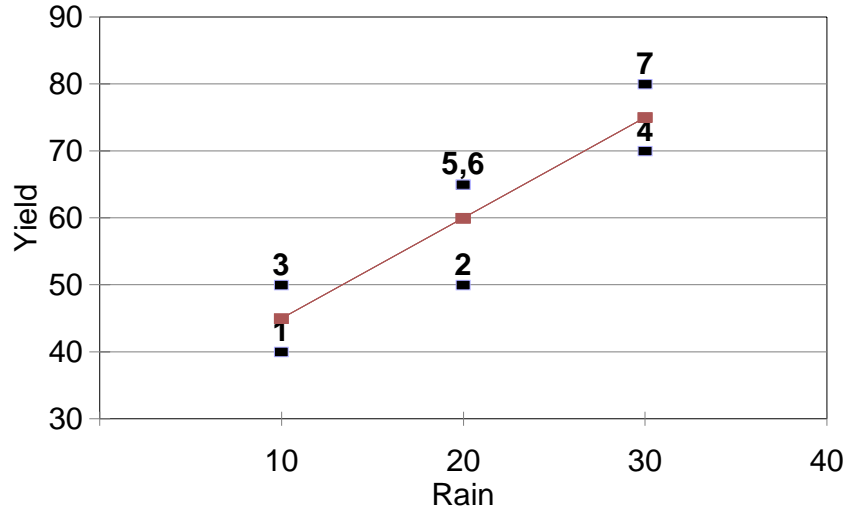
*The equation for this simple regression is  $Y = 36.43 + 0.059F$*



This diagram shows how much rain each field got, and how much yield. Fields 5 and 6 had the same rain and the same yield, so there are two points there.

There is a strong linear relationship between rain and yield in our data. The slope of the least squares regression line in this graph is 1.5, indicating that an extra inch of rain adds 1.5 bushels to the harvest.

*This regression line's equation is  $Y = 30 + 1.5R$*

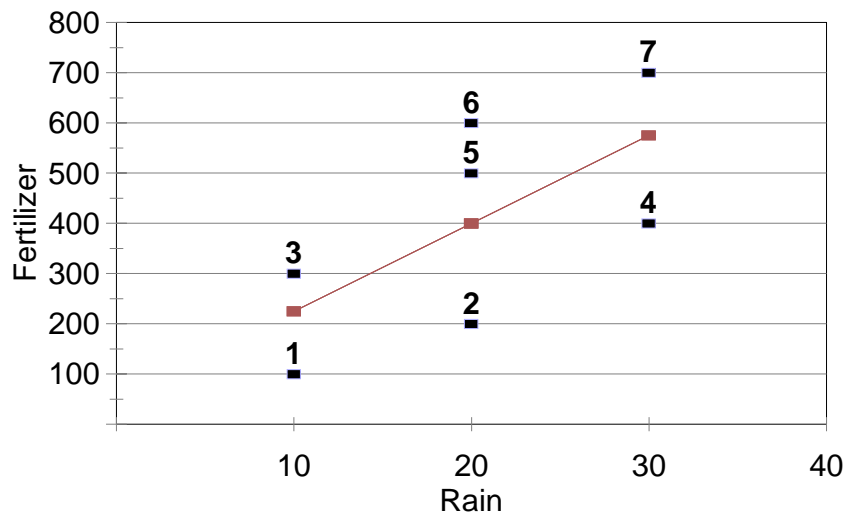


So, we have two variables, fertilizer and rain, that both seem to affect yield positively.

Here is what makes it complicated: The fields that got more rain tended to be the ones that got more fertilizer.

The diagram here shows this. The diagram has Rain on one axis and Fertilizer on the other axis. Each little rectangular dot represents a field and shows how much rain and fertilizer that field got.

*This regression line's equation is  $F = 50 + 17.5R$*



The least squares regression line through these data has a slope of 17.5. This tells us that, on average, fields that got one inch more rain got 17.5 pounds more fertilizer.

This does *not* mean that adding rain *causes* more fertilizer. That would be silly! Correlation does not prove causation. It does mean that fertilizer and rain are correlated *in our data*, and we have to deal with that.

We see that rain seemed to affect yield, and we see that the fields that got more fertilizer also got more rain. Wait! How do we know that the fertilizer had any effect of its own? Maybe the whole effect we saw in the diagram on page 4 was due to rain, not fertilizer. Maybe field 7 had more yield than field 1 because it got more rain, not because it got more fertilizer!

We could ask this the other way around, too. Maybe the effect of rain shown in the upper diagram on page 5 was really due to fertilizer, not rain.

To sort this out, we can look at the residuals, the vertical distances between the numbered points and the least squares lines in each graph.

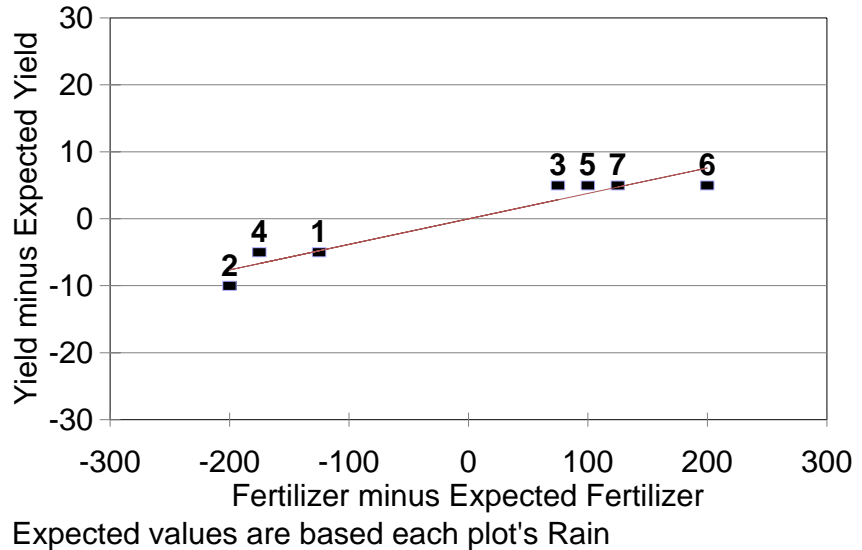
If we compare each point's residual in the top diagram on page 5 with that point's residual in the second diagram on page 5, we get this:

Plot number	Residual in Yield-Rain diagram	Residual in Fertilizer-Rain diagram
1	Negative residual. This point is below the least squares line. This plot had less yield than expected, given how much rain it got.	Negative residual. The point is below the least squares line. This plot had less fertilizer than expected, given how much rain it got.
2	Negative. The point is below the line.	Negative. The point is below the line.
3	Positive. The point is above the line. This plot had more yield than expected, given how much rain it got.	Positive. The point is above the line. This plot had more fertilizer than expected, given how much rain it got.
4	Negative. The point is below the line.	Negative. The point is below the line.
5	Positive. The point is above the line.	Positive. The point is above the line.
6	Positive. The point is above the line.	Positive. The point is above the line.
7	Positive. The point is above the line.	Positive. The point is above the line.

The residuals match pretty well between the two diagrams. Plots with positive residuals in one diagram have positive residuals in the other. Plots with negative residuals in one diagram have negative residuals in the other.

This means that plots that had more fertilizer than you would have predicted (from the amount of rain they got) did get more yield than you would have predicted (from the amount of rain they got). Plots that had less fertilizer than you would have predicted (from the amount of rain they got) did get less yield than you would have predicted (from the amount of rain they got). This pattern lets us conclude that fertilizer does have its own effect, separate from whatever effect rain has.

To quantify fertilizer’s effect, we take the residuals from the two diagrams and plot them. Our horizontal axis variable is the residuals in the fertilizer-rain regression. Those residuals show how much each plot’s fertilizer amount differed from what would have been predicted from the rain amount. For example, Plot 2 got 200 pounds less fertilizer than the 400 pounds you would have predicted from the 20 inches of rain it got.



Our vertical axis variable is the residuals in the yield-rain regression. These residuals show how much each plot’s yield differed from what would have been predicted from the rain amount. For example, Plot 2 got 10 bushels less yield than the 60 that you would have predicted from the 20 inches of rain it got.

The least squares regression line for this residuals vs. residuals graph is shown. The slope is 0.038. This is, in fact, the very number you would get as the estimated coefficient  $\beta$  for Fertilizer in the least squares multiple regression  $Yield = \alpha + \beta Fertilizer + \gamma Rain$ . When Fertilizer is a unit bigger than what you would expect Fertilizer to be, based on what Rain is, then Yield will tend to be 0.038 bushels bigger than what you would have predicted for Yield based on what Rain is.

This is what it really means to say that adding a pound of fertilizer increases Yield by 0.038 bushels, “holding rain constant.” We did not succeed in holding rain constant in the experiment. Using multiple regression, we can infer what would have happened if we had held rain constant.

In general, in the multiple regression equation  $Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_p X_p$ , when  $X_1$  is one unit bigger than you would have predicted, based on the values of  $X_2$  through  $X_p$ , then  $Y$  will tend to be  $\beta_1$  units bigger than you would have predicted, based on the values of  $X_2$  through  $X_p$ .

This procedure – doing a multiple regression by means of series of simple regressions – is complicated! Can we do it more simply with a formula?

**The formula for a least squares coefficient in a multiple regression**

Here is the formula for a least squares coefficient in an equation with two independent variables.

Least squares estimated coefficient  
of X in  $Y = \alpha + \beta X + \gamma Z$

$$\hat{\beta} = \frac{\sum_{i=1}^N (X_i - \bar{X})(Y_i - \bar{Y}) \sum_{i=1}^N (Z_i - \bar{Z})^2 - \sum_{i=1}^N (Z_i - \bar{Z})(Y_i - \bar{Y}) \sum_{i=1}^N (X_i - \bar{X})(Z_i - \bar{Z})}{\sum_{i=1}^N (X_i - \bar{X})^2 \sum_{i=1}^N (Z_i - \bar{Z})^2 - \left( \sum_{i=1}^N (X_i - \bar{X})(Z_i - \bar{Z}) \right)^2}$$

That is pretty complicated, too! If there are more than two independent variables, the formula gets even more complicated. These formulas can be written more elegantly if matrix algebra is used. That does not, however, make them easier to calculate in a spreadsheet.

Because of this complexity, we will not attempt to implement formulas like this in spreadsheet templates. Instead, we will use a computer program specifically designed for the purpose of calculating multiple regression coefficients from data. That will be in assignment 3.

Suppose we have done all that. We have used the computer program and we have a multiple regression equation with estimated coefficients. What can we do with that equation?

**Prediction**

Prediction with multiple regression is done by plugging numbers in for the variables. Suppose our estimated equation is  $Y = 28.10 + 0.038F + 0.83R$ . F is fertilizer and R is rain. (You don't *have* to use "X" and "Z".) To predict yield if we have 800 pounds of fertilizer and 20 inches of rain, we calculate this way:

$$Y = 28.10 + 0.038 \times 800 + 0.83 \times 20$$

$$Y = 75.24 \text{ bushels}$$

The prediction you get from a multiple regression is generally different from what you get from a simple regression. Let's continue this example to see why:

Our simple least squares regression equation, when we just looked at yield and fertilizer, was  $Y = 36.43 + 0.059F$ , where Y is yield and F is fertilizer.

What will yield be if we use 800 pounds of fertilizer? From this simple regression equation, our prediction for the yield is

$$Y = 36.43 + 0.059 \times 800$$

$$Y = 83.57 \text{ bushels}$$

As shown above, with 800 pounds of fertilizer and 20 inches of rain, we get a prediction of 75.24 bushels.

What happened? Why is this prediction smaller? It is because of the amount of rain I chose to use. I

chose to base my prediction on 20 inches of rain. That is the average amount of rain that the fields got.

However, 800 pounds of fertilizer is not an average amount of fertilizer. 800 is higher than the highest amount of fertilizer in the data set.

The multiple regression equation handles that just fine because it separates the effects of fertilizer and rain. The simple regression of fertilizer and yield cannot make that separation. It does not see that the fields that got more fertilizer also got more rain. If you use the simple regression to predict, you are assuming implicitly that if a plot gets a lot of fertilizer, it also will get a lot of rain.

To demonstrate this, let us try 30 inches of rain with 800 pounds of fertilizer in our multiple regression equation. Now we have an above-average amount of rain to go with our above-average amount of fertilizer.

$$Y = 28.10 + 0.038 \times 800 + 0.83 \times 30$$

$$Y = 83.57 \text{ bushels}$$

This gives the same prediction that we got from the simple regression for 800 pounds of fertilizer. It worked out that way because 30 inches of rain happens to be the amount of rain that you would expect a field with 800 pounds of fertilizer to get, based on the data we have.

So, when is the simple regression good enough, and when do you need multiple regression?

The simple regression model is OK if the past relationship of rain and fertilizer will continue in the future.

The multiple regression model is better if we are not sure that the past relationship of rain and fertilizer that will continue in the future.

In my judgement, the multiple regression model is a better choice in this case. It is not a good bet that the fertilizer-rain relationship of the past will continue. That was just one year's funny luck. Putting more fertilizer on a plot is not going to make it rain more. We are therefore better off predicting with the multiple regression equation.

**Specification Bias**

Specification is the jargon term for choosing what variables to use in a regression. Bias is the term for when your estimator tends to be off target, when the expected value of the estimate is not equal to the true value.

Specification bias can happen when you leave an important variable out of your regression model, and if that left-out variable is correlated with one or more of the variables that are in the model. That is happening in the yield-fertilizer-rain example. The simple regression leaves out an important variable, Rain. Rain happens to be correlated with Fertilizer. This causes the Fertilizer coefficient in the simple regression to be biased upward. Fertilizer's simple regression coefficient, 0.059, is too high. It overstates the actual effect of fertilizer. The multiple regression, which includes Rain, gives a lower coefficient for Fertilizer, 0.038. Presumably, that estimate of fertilizer's effect is not biased (unless there is another left-out variable that we do not know about).

If the left-out variable is negatively correlated with a variable in the equation, the bias goes the other way. The effect of the variable that is in the equation is biased downward, tending to be smaller than the true effect.

**Standard errors of the coefficients**

In a multiple regression, we typically want to do hypothesis tests to tell us which of the independent variables really affect the dependent variable. For that, we need the standard error of each coefficient.

The standard error of the least squares estimate of  $\beta$  in the multiple regression,  $y = \alpha + \beta X + \gamma Z + \text{error}$ :

$$s_{\beta} = \frac{s}{\sqrt{\sum_{i=1}^N (X_i - \bar{X})^2 - \frac{\left(\sum_{i=1}^N (X_i - \bar{X})(Z_i - \bar{Z})\right)^2}{\sum_{i=1}^N (Z_i - \bar{Z})^2}}}$$

$s_{\beta}$ , the standard error of the coefficient  $\beta$

where  $s = \sqrt{(\text{Sum of squared residuals} / (\text{Observations} - 3))}$

A formula like this defies intuitive understanding, but notice that:

1. The observed Z values appear in the formula for the standard error of X's coefficient (in the form of Z values' deviations from the mean of the Z's.) This means: Both the coefficient and the standard error of any variable depend on what other variables are in the regression.
2. The denominator of  $s_{\beta}$  gets closer to zero the more closely X and Z are correlated. (Not obvious, so here's a brief explanation: If X and Z are perfectly correlated, then every  $X_i - \bar{X}$  is some multiple of  $Z_i - \bar{Z}$ . If you substitute  $m(X_i - \bar{X})$  in the above equation everywhere you see  $Z_i - \bar{Z}$ , then the entire denominator becomes zero. I encourage the algebraically inclined to try this. If the denominator is 0 then the whole fraction, the standard error of the coefficient, is infinite. The

practical implication is that the standard error of each coefficient goes up when there's a lot of collinearity among the independent variables. More on collinearity in a moment.

3. If  $X$  and  $Z$  are not correlated – meaning that the two do not have a linear relationship – then the part of the denominator after the minus sign becomes 0 and disappears. The standard error of the least squares estimate of  $\beta$  becomes the same as if you had ignored  $Z$  and done a simple regression. Actually, if  $X$  and  $Z$  are not correlated, then the humongous formula on page 8 reduces to the simple regression formula that you implemented in your spreadsheet.

If you have more than two right-hand side variables, the standard-error formula gets even more horrendous, unless matrix notation is used.

### **t-tests on the estimated coefficients**

t-tests on the coefficients for multiple regression are done similarly to t-tests for simple regression. You divide the estimated coefficient from the equation, the  $\hat{\beta}$ , by its estimated standard error, from the formula above. Fortunately, your regression analysis program will do this calculation for you.

The  $\hat{\beta}/s_{\hat{\beta}}$  ratio is your t-statistic, which you can compare with the appropriate critical value from a t-table. To find the right t-value, use  $N - P$  for your degrees of freedom.  $N$  is the number of observations.  $P$  is the number of parameters, including the intercept, on the right-hand side of the equation.

**Collinearity and multicollinearity** (From here on, the material is mainly for Assignment 4.)

Collinearity means that, in your data, two of your independent variables' values have a close linear relationship.

Multicollinearity means that, in your data, one of your independent variable's values is close to being a linear function of some of the other independent variables' values.

Perfect collinearity is usually the result of a mistake in setting up your model, such as:

1. Putting the same variable twice into a regression equation.
2. Putting in two variables that measure the same thing two different ways. For example, one variable is the person's weight in pounds and another variable is the person's weight in kg.
3. Using too many dummy variables. We will discuss this in Assignment 6.

Less-than-perfect collinearity is fairly common, particularly in studies in the social sciences. Assignments 3 and 4 will have some examples.

Visualizing collinearity:

You have a three-dimensional graph. The  $X$  and  $Z$  axes lie on a table. The  $Y$  axis points straight up. Your data points form a cloud floating over the table. The cloud is shaped like a thin cigar, because the points are all nearly above a line on the table, in the  $X$ - $Z$  plane. You are trying to fit a plane to the cigar shape. Because the cigar is thin, that plane can tilt easily from side to side, without much affecting how close the plane is to the points. You can imagine that your plane is a rectangular flat piece of cardboard. The diagonal of the cardboard runs through the long dimension of the cigar. You can rotate the cardboard, with the diagonal still running through the cigar. As you do this, the slope along the diagonal stays the same. However, the slopes along the

edges of the cardboard change greatly. The coefficients in a multiple regression are determined by the slopes of the planes in the directions of the X and Z axes. So those coefficients are indeterminate.

What difference does collinearity or multicollinearity make when you are doing regression? In terms of our two uses of regression, which are prediction and explanation:

Prediction can be good, if you are predicting for values of the independent variables that maintain the linear relationship. For example, if you do have the weight in pounds and the weight in kilograms as independent variables in the same model, you can predict for a new weight so long as the 1 pound = 0.454kg relationship is maintained. If the past relationship does not continue, you are out of luck! No good prediction is possible.

Explanation can be hard – too much collinearity makes separation of causes impossible. You can't tell the effect on Y of changing X and holding Z constant if in your data X and Z always change together. What you see in your regression results is large standard errors for your coefficients, and, consequently, low levels of significance as measured by t-tests.

#### **t-tests and F-tests with multicollinearity**

When you have near collinearity between two of your X variables, often both of those variables will flunk the t-test. For each variable, you will not be able to reject the hypotheses that its coefficient is zero. Yet, that variable may still be important to explaining what Y is. Its effect may be obscured by the presence of the other variable that is correlated with it.

Sometimes there are a bunch of variables that are all correlated with each other. All of the variables may show insignificant t-statistics, yet one or more of them may still be important to explaining Y. Here, too, the separate effect of any one variable is obscured by the effects of the other variables.

Some statisticians use a stepwise method to decide which of a group of highly collinear variables should be included and which should be excluded. I prefer this approach:

#### **Hypothesis tests on more than one coefficient at once**

If you have two or more variables with insignificant coefficients, but you suspect that this is because they are too collinear, here is how you can find out if one or more of them significantly relates to the dependent variable.

1. Do the regression with all the variables in the equation. This is the full model. You probably already did this.
2. Write down or circle on your printout the Sum of Squared Residuals.
3. Do another regression on the same data. This time leave out the independent variables that you think are too collinear. This model with some of the independent variables removed is called the "reduced model."
4. Print or write down the sum of squared residuals from this reduced model regression.
5. Plug the numbers into this formula:

$$F = \frac{(SSR_{RM} - SSR_{FM}) / (P_{FM} - P_{RM})}{SSR_{FM} / (Observations - P_{FM})}$$

- SSR<sub>FM</sub> = sum of squared residuals for the full model
- SSR<sub>RM</sub> = sum of squared residuals for the reduced model
- P<sub>FM</sub> = number of parameters in the full model
- P<sub>RM</sub> = number of parameters in the reduced model

6. The result of this formula has the F distribution with
  - P<sub>FM</sub> - P<sub>RM</sub> numerator degrees of freedom and
  - Observations - P<sub>FM</sub> denominator degrees of freedom

The numerator degrees of freedom is the number of coefficients that you are testing simultaneously.

7. Find the critical value in an F table (which you can find in 716-tables for hypothesis tests.pdf) corresponding to your numerator degrees of freedom, your denominator degrees of freedom, and the significance (“alpha”) level you choose. If the F value from the formula is greater than the F value from the table, you can reject the hypothesis that all the coefficients you are testing are 0.

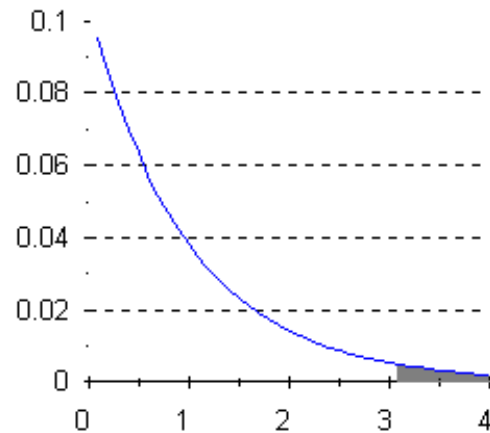
If you can reject the hypothesis that all of the coefficients are zero of the variables that you took out, this tells you that at least one of those variables does have a significant effect on the dependent variable. You cannot tell which of the variables is the one with the real effect, only that at least one of them does.

If a group of variables flunk this F test, then none of them matter. You can take all of them out of your equation. In other words, you can use the reduced model for prediction and explanation. That will save you a little time and work.

To reiterate: The two regressions are (1) a “full model” with all the X variables in, and (2) a “reduced model” with the suspect X variables taken out. The null hypothesis is that the full model and the reduced model are equally good for predicting Y. The key piece of information from each regression is the sum of the squares of its residuals. The F formula calculates a number from (1) the relative increase in the sum of squared residuals in going from the full model to the reduced model, (2) the difference in the number of parameters in the two models, and (3) the number of degrees of freedom in the full model.

The formula above starts with “F =” because the number it gives you is a standardized number that has the F distribution.

To give you an idea what the F distribution looks like, the diagram to the right shows a density function for an F



The F distribution with 2 and 100 degrees of freedom. 5% of the area under the curve is in the shaded tail. If the F statistic is in the tail, reject the null hypothesis.

distribution with 2 degrees of freedom for the numerator and 100 degrees of freedom for the denominator. The horizontal axis shows a range of possible values for F. The height of the curve, above any particular F value on the horizontal axis, is proportional to how likely it is that a formula like the one above could produce something close to that F value by pure luck, if there is really no difference between the two models. An F value near 0 is more likely to happen in this circumstance than an F value near 4, because the curve is higher over 0 than over 4.

In the diagram, the F distribution's density function is not symmetrical, like the bell-shaped t-distribution is. For one thing, F values start at 0 and go up. Negative F values are impossible. The shape of the F distribution depends on the numerator degrees of freedom. For two numerator degrees of freedom, as shown here, the curve starts high and falls as you go to the right. For three or more numerator degrees of freedom, the F distribution rises and then falls as you go to the right. Though the F curve does fall as you go to the right, never gets all the way down to 0 no matter how far to the right you go. This means that big values of F can happen by chance, but very rarely.

As you go to the right along the horizontal axis, you will come to an F value such that only 5% of the total area under the F curve is in the "tail" to the right of that F value. That F value is called the critical F value at the .05 level, and it is the number you get find when you look an the F table. If the null hypothesis (no difference between the models) is true, F statistics in that tail occur by chance only 5% of the time.

As with the t distribution, the F distribution is actually a family of distributions. There is one F distribution for every combination of numerator degrees of freedom (the difference in number of parameters between the full and reduced model) and denominator degrees of freedom (number of observations minus number of parameters in the full model). Finding the critical value in an F table involves picking the right column and row for your numerator and denominator degrees of freedom.

If the F statistic you calculate is higher than the critical value in the F table at the .05 level, reject the null hypothesis and conclude that there is a significant difference in predictive power between the two regressions. Put back in the variables that you took out to make the reduced model, because they do help predict Y.

Regression analysis programs, like the LS program that we will use in this course, include with the regression output an F value for the whole equation. This tells you if there is anything worthwhile in your equation, by testing all of the coefficients being 0 at once. It is the number you would get if you compared your equation with a reduced model that had no independent variables except the intercept.

If any one of the coefficients in an equation has a significant t statistic, then the F for Equation will be significant, too, except in freakish circumstances. Not as rare is to have none of your X variables have a significant t statistic, but have the F for Equation be significant. Collinearity among your X variables can cause this. Your equation is useful for prediction, but not for telling which X variables are most important.

It may have occurred to you that we have two ways now to test whether an independent variable's true coefficient is 0. One would be the t-test we discussed before. The other would be an F test, for which we could do a reduced model regression with that one variable taken out, and the plug the sums of squared residuals from the full and reduced models into the F test formula. The two methods are mathematically equivalent. To see this, look at the t and F tables (the link to download them is on the syllabus). Pick a

significance level, such as .05, and compare the numbers in the t table with the numbers in the F table's column for one numerator degree of freedom. You will see that the F table values are the squares of the t table values. The F statistic you would calculate using the method above in this paragraph would equal the square of the t statistic for that variable in the full model. Consequently, the t and F methods give the same answer when testing the significance of any one variable.

### **Advanced stuff: Confidence regions**

We have seen already how to use an estimated coefficient, its standard error, and a critical value from a t-table, to calculate a confidence interval. We say, for example, that the probability is 95% that the true value of the coefficient is between two numbers that we calculate.

When you have several variables, you can use a critical value from an F-table to calculate a joint confidence region. The confidence region is an ellipse in a plane whose coordinate axes are for  $\beta_1$  and  $\beta_2$ . A rectangle constructed with sides equal to the 95% confidence intervals for two coefficients individually would be a 90.25% confidence region ( $.95^2=.9025$ ).

If two independent variables,  $x_1$  and  $x_2$ , are correlated, their coefficients will have negative covariance.

This means: If the estimate  $\hat{\beta}_1$  happens to be higher than the true  $\beta_1$ , then  $\hat{\beta}_2$  will tend to be lower than the true  $\beta_2$ .

If collinearity is high, we won't have much idea what  $\beta_1$  and  $\beta_2$  are, but we will know what  $\beta_1/\hat{\beta}_1 + \beta_2/\hat{\beta}_2$  is.

In such a situation, the joint confidence region will be a tilted ellipse, with  $\beta_1/\hat{\beta}_1 + \beta_2/\hat{\beta}_2 = 0$  as its principal axis. The closer than  $x_1$  and  $x_2$  are correlated, the narrower the ellipse will be.