

Assignment 3. Multiple Regression

© 2011 Samuel L. Baker

In this assignment, you'll first do a simple regression on some data that I'll give you. From that regression, you'll gauge the effect of one thing on another thing and make a prediction. Then you'll use multiple regression to gauge the effect again and redo the prediction.

To do your regressions, you'll use LS, a regression analysis program available over the internet. Page 10 has a check list of what to turn in. Pages 8-9 have instructions for how to get results out of LS and into a word-processor document for editing and printing.

Get the data

Open your web browser and go to <http://sambaker.com/courses/J716/Data3.html>, or go to the syllabus and click the link for "Getting data for assignment 3." Follow the directions on that web page to get your data.

For your write-up: Include a copy of your data. (This will happen automatically if you follow the instructions to come.)

These data show the results of an agricultural experiment, the purpose of which was to determine the effect of fertilizer on the yield of corn. Seven plots of land in widely scattered locations were planted with the same crop. Different amounts of fertilizer were used. Also measured was the amount of rainfall during the season on each plot. Yield is in bushels. Fertilizer is in pounds. Rain is in inches.

For the benefit of non-U.S. students, "corn" is maize. A bushel is 35.24 liters. A pound is 0.454 kg. An inch is 2.54cm.

Use Excel or other spreadsheet to get the data ready for the LS program

(These directions show how to put data into LS by way of Excel from scratch. The Data3 web page suggests faster methods for getting the data into LS.)

How your data should look:

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

Start up your spreadsheet program, with a new blank notebook.

The picture shows the arrangement that the LS regression program requires. The columns are for the variables. The rows are for the observations.

In the first row, leave cell A1 blank.

In the rest of the first row, put the names of your variables, going across to the right. LS counts how many names there are in the first row to determine how many data columns to read, so be sure to name every variable.

Starting with row 2, the A column is for the names of the observations. If you don't name your observations, leave the A column blank, as it is in the picture.

The numerical data for each variable go under the variable's name, starting in row 2, column B.

Your data should form a rectangular block. If there are blanks or words where numbers should be, LS will give you an error message.

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

Start the process of copying your data and pasting into LS. Move the cell selector to one corner of your data. Cell A1 will do fine. Click and drag with your mouse, or press ⇧ and the arrow keys to move the cell selector to the opposite corner. If you started in A1, go over and down to D8. The selected cells should have a heavy outline and be shaded.

Be sure to include the A column, even though it's empty here.

When your spreadsheet looks like this, press Ctrl+C

(or use the menu's Copy command).

Starting LS

After your selection is copied, minimize your Excel window. To do this in Windows, click on the □ nearest the upper right corner of your Excel window.

In your web browser, go to the syllabus and click one of the links for LS. The URL is: <http://sambaker.com/ls>

After a pause, while Java is loading, you should see LS's main screen. It has two large boxes, called "text areas." The upper text area is for your data. The lower text area is for results.

Click in the upper text area. Paste here what you copied from the spreadsheet. On a PC, press Ctrl+V or right-click and select paste from the context menu.

Macs require an extra step. Open Word and use Paste Special to paste as Unformatted Text. Then select what you pasted, copy that, and paste it into LS's upper box.

Click on the "Click here to read and check your data" button. LS will read your data. If everything is OK, the lower window will show the means and simple correlations of your variables, and some radio buttons will appear below the lower window. If no radio buttons appear, look in the lower window for a message indicating what problem LS found with your data.

If your data have errors, go back to your spreadsheet and see if you can fix what is wrong. Perhaps you chose the wrong area to copy, or perhaps one of your cell entries was mistyped. Once the data are fixed, select the block again and copy it to the clipboard. Return to LS. Erase all the

contents of the upper box in LS (in most browsers you can click in the box, press **Ctrl**+A , and then press **Delete**). Then paste the fixed data in.

A copy of your data will also be in the lower text area. Scroll up from the Means table if you want to see them. Later, if you copy the entire contents of this text area to your word processor, the data will be the first thing in the file.

Look at the means to be sure they seem reasonable. This is a way to check for mistakes in data entry. For example, in this homework, if your mean for Yield is not between 50 and 70, something is wrong.

Below the means are the correlations. The correlations are in the form of a table, with a row for each variable and a column for each variable. Each number in this table shows the simple correlation the row variable and the column variable. The table has 1's down the diagonal, because any variable is perfectly positively correlated with itself. For other table entries, a correlation near 1 indicates a strong linear relationship between the two variables. A correlation near -1 indicates a strong negative relationship between the variables. A correlation near 0 indicates no relationship. This table can be useful for exploring data and deciding which variables to include in the analysis.

Make a note of the correlation between Fertilizer and Rain. You will use this number later in your write-up. In the Correlations table, you want the number in the column for Rain and the row for Fertilizer. The same number is in the row for Rain and the column for Fertilizer. The correlation table is symmetric around the diagonal from upper left to lower right.

Now you are ready to do a regression. Below the lower box, there should be the names of all the variables in your data, with a radio button for each one.

Click with your mouse on the radio button for Yield. This will select Yield as your dependent variable, your Y variable, the variable on the left-hand-side of your equation.

Selecting your dependent variable makes some new checkboxes appear. There is one checkbox for each possible independent (X or right-hand-side) variable. By default, all of these are checked. Leave them all checked.

Click on the Do regression button. The regression results will appear in the lower big box.

Here are my regression results. Yours will be like this, but with different numbers.

Yield is the dependent variable, with a mean of 60.0

| Variable | Coefficient | Std Error | T-statistic | P-Value |
|------------|-------------|-----------|-------------|-----------|
| Intercept | 28.095238 | 2.4914821 | 11.276516 | 3.5238E-4 |
| Fertilizer | 0.0380952 | 0.0058321 | 6.5319726 | 0.0028378 |
| Rain | 0.8333333 | 0.1543033 | 5.4006172 | 0.0056899 |

These columns test if each coefficient is 0.

Here are the same results, with comments added.

Yield is the dependent variable, with a mean of 60.0
 This line shows the dependent variable you chose and its mean.

| Variable | Coefficient | Std Error | T-statistic | P-Value |
|------------|-------------|-----------|-------------|-----------|
| Intercept | 28.095238 | 2.4914821 | 11.276516 | 3.5238E-4 |
| Fertilizer | 0.0380952 | 0.0058321 | 6.5319726 | 0.0028378 |
| Rain | 0.8333333 | 0.1543033 | 5.4006172 | 0.0056899 |

These columns test if
each coefficient is 0.

The columns below have the variable names, the estimated coefficient for each variable, the coefficient's standard error, the "t-statistic," and the P-value that corresponds to that t-statistic.

At the top of the list of variables is the intercept. In this case, the Intercept coefficient is our estimate of what Yield would be if there were no fertilizer or rain at all.

The Fertilizer coefficient is about 0.038. This means that adding 1 pound of fertilizer raises yield by an average of 0.038 bushels, if the rain amount does not change.

The Rain coefficient is about 0.833. This means that adding 1 inch of rain increases yield by an average of 0.833 bushels, if the fertilizer amount does not change.

In the T-statistic column, each T-statistic is the estimated coefficient divided by the standard error. To test the hypothesis that a particular coefficient is zero, you can compare the t-statistic on that variable's row with the critical value from the t table. Alternatively, you can use the P-value as a short cut. If the P-Value is less than 0.05, the t-statistic will be higher than the critical value in the t table at the 0.05 significance level and the proper number of degrees of freedom. You can then say that the coefficient of Fertilizer is significantly different from zero at the 5% level. You can, in other words, reject the hypothesis that the coefficient is 0, and know that the probability that you are making a mistake (a "type I error") is less than .05.

The same goes for saying whether the coefficient of rain is significant.

There is also a lower part to the regression results. Here you see the calculation of the degrees of freedom number used in the t tests (7 observations minus 3 parameters = 4 degrees of freedom). Also shown is the Standard Error of the Regression (“s”), the R-Squared, the Sum of Squared Residuals, and an F-test and associated P-value for assessing whether this equation is helpful for predicting Yield. We will use the F-test next week.

For your write up:

Write out your estimated equation.

It is of the form $\text{Yield} = \text{intercept} + \text{coefficient} \times \text{Fertilizer} + \text{coefficient} \times \text{Rain}$.

My estimated equation is:

$$\text{Predicted Yield} = 28.095 + 0.0381 \times \text{Fertilizer} + 0.83333 \times \text{Rain}$$

This is an important concept! Be sure you see how I got this equation from the Coefficient column in the regression results.

For each coefficient (but not the intercept), test the hypothesis that the coefficient is 0. Explain what you're doing in the hypothesis tests.

The easiest way is to use the p-values. In my results, both Fertilizer and Rain have P-values that are less than 0.05. Therefore, I can reject the hypothesis that Fertilizer has a zero coefficient. Likewise, I can reject the hypothesis that Rain has a zero coefficient. Your results may differ.

What is your coefficient for Fertilizer? Each additional pound of fertilizer increases yield by how many bushels, according to these figures?

What is your coefficient for Rain? How much extra yield do you get for each extra inch of rain, on the average?

Prediction:

Click the Prediction button. A form for predictions will appear below the buttons.

Type in 800 for Fertilizer and 20 for Rain. Press to get LS to calculate the prediction. The results will be in the box above.

Notice that I am asking you predict for an amount of fertilizer, 800, that is way bigger than the average amount of fertilizer that the fields got. The amount of rain, 20, is not above the average rain amount. This will be an important idea later.

In addition to the prediction, LS calculates some prediction confidence intervals. We do not use these this week. In case you use LS in the future, I should tell you that LS's prediction confidence intervals are accurate for models with up to 120 degrees of freedom. For big data sets with more than 120 degrees of freedom, the intervals may be too wide by up to 2%.

Checking for curvature or an outlier

Click the red Residuals Plot button. You will be given the choice of whether to sort the residuals. Check the second choice, to sort by the predicted value. (The rationale for this will be explained later.)

The residual plot will show in LS's lower window. Mine look like this:

| Observation | Predicted | Residual | -3s | -2s | -s | 0 | +s | +2s | +3s |
|-------------|-----------|------------|-----|-----|----|---|----|-----|-----|
| 1 | 40.238095 | -0.2380952 | | . | | * | | . | |
| 3 | 47.857143 | 2.1428571 | | . | | | * | . | |
| 2 | 52.380952 | -2.3809524 | | . | * | | | . | |
| 5 | 63.809524 | 1.1904762 | | . | | | * | . | |
| 6 | 67.619048 | -2.6190476 | | . | * | | | . | |
| 4 | 68.333333 | 1.6666667 | | . | | | * | . | |
| 7 | 79.761905 | 0.2380952 | | . | | * | | . | |

Durbin Watson = 3.4444444377777685

The residuals' plot has one line for each observation. On each residual's line, each * represents the residual. The residuals are standardized, meaning that each residual is divided by s, the standard error of the regression. For example, in the diagram above, the first residual is between -s and 0, at about -0.4s. This is because the first residual is -2.32, the s is 5.96, and -2.32/5.96 is about -0.4.

The purpose of the residual plot is to enable you to spot patterns that might indicate that least squares is not your best choice of method. A big curve or wave, for example, might indicate that a non-linear equation might be better, or that an important variable is left out. If you see individual residuals that stick way out, this might indicate that the errors do not have the same variance for all observations.

At the end of the residual plot, LS reports the Durbin-Watson statistic. This measures the serial correlation of the residuals, which means how much the residuals tend to follow each other. It detects snake or curved patterns. Here is a rough guide to interpreting the Durbin-Watson statistic:

| Durbin-Watson value | Indication |
|---------------------|--|
| less than 1 | Serial correlation evident. |
| near 2 | OK. No serial correlation. |
| more than 3 | Alternating positive-negative residuals. |

A more formal method of interpreting the Durbin-Watson statistic is with the Durbin-Watson table, in the downloadable file of tables.

If your Durbin-Watson statistic is near 2, that tells you that there is not much curvature in your data. There is no reason not to use the flat plane of linear multiple regression for prediction. (You see above that I got a Durbin-Watson value of 3.44. This high number suggests that the residuals are alternating, negative and positive. This is not atypical of data made up for a textbook exercise!)

Detecting outliers: As you see above, the plot goes from -3s to +3s. Residuals whose absolute values are less than 3 times the regression's standard error ("s") are shown with *'s. If there were a residual bigger than 3s or more negative than -3s, the value of the standardized residual (the residual divided by s) would be shown as a number, rather than with a *. If the errors truly are normally distributed and have equal

variance, only about three residuals in a thousand should be more than 3 times the size of s in absolute value. This implies that, if you are doing a study with your own data, and your data set is not very big, and you see a residual indicated with a number rather than a *, you probably have an outlier. In a data set as small as the one we are using here, even a * outside the 2's might indicate an outlier. Only about one residual in 20 should be outside $\pm 2s$.

Now I want to explore **specification bias**. (“Specification bias” is a jargon term that makes you sound smart, so learn to say it.) Specification bias is the distortion of results that you get if you leave out a variable or put in a wrong variable. Avoiding specification bias is a big part of why you use multiple regression.

On the LS web page, look under the results box for “5. Choose your independent variables:” There should be checkboxes checked for the Intercept, Fertilizer, and Rain. Click the checkbox next to Rain to clear that checkbox.

Unchecking Rain means that now you just have Fertilizer as your one independent variable. You are ignoring Rain. We are back to a simple regression.

(Leave the Intercept’s box checked. The intercept is the α in $Y = \alpha + \beta X$. If you leave it off, your equation is just $Y = \beta X$. For all of this course's assignments, unless specifically told otherwise, leave the Intercept checked.)

Click on the Do regression button. The regression results will appear in the lower big box.

Here is how my results look. Yours will differ, because your data are different.

Yield is the dependent variable, with a mean of 60.0

| Variable | Coefficient | Std Error | T-statistic | P-Value |
|------------|-------------|-----------|-------------|-----------|
| Intercept | 36.428571 | 5.03812 | 7.2305883 | 7.8942E-4 |
| Fertilizer | 0.0589286 | 0.0112656 | 5.2308517 | 0.0033793 |

These columns test if
each coefficient is 0.

7 observations minus 2 parameters = 5 degrees of freedom.
 Standard Error of the Regression (s) = 5.961184
 R-Squared = 0.8454969 Sum of Squared Residuals = 177.67857
 Testing whole equation: F = 27.361809 P-value = 0.0033793

For your write up:

Tell me your estimated coefficient for Fertilizer from the simple regression. Each additional pound of fertilizer increases yield by how many bushels, according to your figures?

Write out your regression equation. It is of the form $Yield = intercept + coefficient \times Fertilizer$

Test the hypothesis that the true coefficient of Fertilizer is 0. In other words, tell me if you think the apparent relationship between Fertilizer and Yield was just a lucky coincidence. Is it likely that putting

fertilizer on a field truly adds nothing to yield at all? See the hypothesis testing discussions the earlier sections of this booklet on regression theory if you need more help with the t-test and its interpretation.

Why is your coefficient for fertilizer, showing the estimated the effect of fertilizer on yield, different in the simple regression from what it was in the multiple regression?

Hints:

Fertilizer and rain each affect the yield. (We can tell that from the Correlations table that LS calculates for you. The correlation between fertilizer and yield is pretty high, as is the correlation between rain and yield.) We can also tell from the correlations table (I asked you to notice this) that the plots that got more fertilizer also tended to get more rain. The plots that got more yield thanks to getting more fertilizer got a boost in yield from the extra rain they also got.

The simple regression gives fertilizer the credit for the extra yield that was actually due to the additional rain that the plots with more fertilizer also got. The fertilizer coefficient in the simple regression is therefore bigger than the actual contribution of fertilizer – by itself – to yield.

The multiple regression ($\text{Yield} = \alpha + \beta \cdot \text{Fertilizer} + \gamma \cdot \text{Rain} + \text{error}$) separates the effects of fertilizer and rain. The fertilizer coefficient in the multiple regression tells you the effect of fertilizer by itself, not confounded by the effect of more water.

Here is another way to think about it:

Imagine that the different amounts of water on the different fields came, not from the sky, but from a practical joker who deliberately put more water on the fields on which you put more fertilizer. He did this, knowing that water raises yield, to fool you into thinking that fertilizer's effect is bigger than it really is. (Maybe he wants to sell you fertilizer!) If you use the simple regression to assess the effect of fertilizer, you will fall for his scheme.

Using simple regression here causes specification bias. Leaving out a variable that actually matters is called mis-specification. The effect of mis-specification is to change – bias – the coefficient of the variable that is in the equation. Here, leaving out Rain biases the Fertilizer coefficient.

Specification bias also can affect the prediction. Let's see how that happens.

Click the blue Prediction button. Below the buttons, a form will appear with spaces into which you can type numbers for the X variables' values. The prediction button applies to whichever regression you did most recently, so the form will show Fertilizer and not Rain.

Type 800 in the space for Fertilizer. Press [Enter ↵] to proceed with the calculation. LS will tell you the predicted Y value, based on your X values.

(When you have done more than one regression, the Residuals and Prediction buttons give you the results for the last regression that you did. If you need residuals or a prediction based on a regression that you did earlier, do that regression again by selecting the variables you want with the check boxes and then clicking the Do regression button.)

For your write-up: How is this simple regression prediction different from the multiple regression prediction? Why is there this difference?

Hint: Build on the fact that the fields that got more fertilizer also tended to get more rain. The simple regression doesn't see rain or its effect. Using the simple regression is like assuming that if you add more fertilizer, you'll get more rain. Here, I asked you to predict for a large amount of fertilizer, but only an average amount of rain.

For another way to think about this, let's get bring in the practical joker again. This is the joker who deliberately put more water on the fields on which you put more fertilizer, to fool you into thinking that the fertilizer's effect is bigger than it really is. If you apply an amount of fertilizer that is way above the average of the other fields, and if the joker doesn't sneak in and add a way-above-average amount of water, the yield will disappoint you. You won't get as much as the simple regression predicted.

For more on this, see pages 8-9 of Multiple Regression Theory.

Working with a word-processor and LS

At this point, LS has put a lot of text into the results text area. Plus you have a lot of notes, either on paper or in a word processor file. Now let's combine the two.

Start a new blank document in a word processor, such as Microsoft Word. Leave your web browser with LS open.

In Windows, press **[Alt]+[Tab]**, or use the task bar at the bottom of the screen, to switch back to LS.

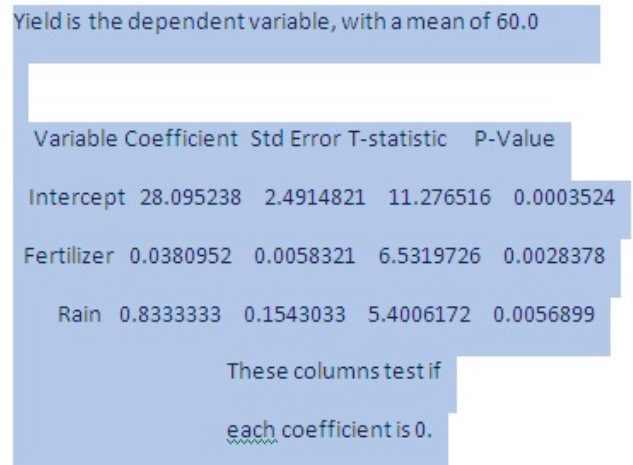
Click in the LS results window and press **[Ctrl]+A** on a PC, or **command+A** on a Mac. The whole background of the results window should change color to show the selection. (If that does not work, try right-clicking in the text area and then clicking Select All on the context menu that appears. If *that* does not work, click and drag with your mouse to select the whole contents.)

[Ctrl]+C should then copy all of this to the clipboard. Do not use Edit Copy from the menu at the top -- that may not work.

Press **[Alt]+[Tab]** or use the task bar at the bottom of the screen to switch back to your word processor. **[Ctrl]+V** pastes.

What you see in your word processor may look all out of alignment. (See the picture to the right.) To get the text to line up properly, select the text and change the font to Courier New or some other monospace font. If that causes some lines to be too long, so that they wrap to the next line, select the text again and change the font size to a smaller number.

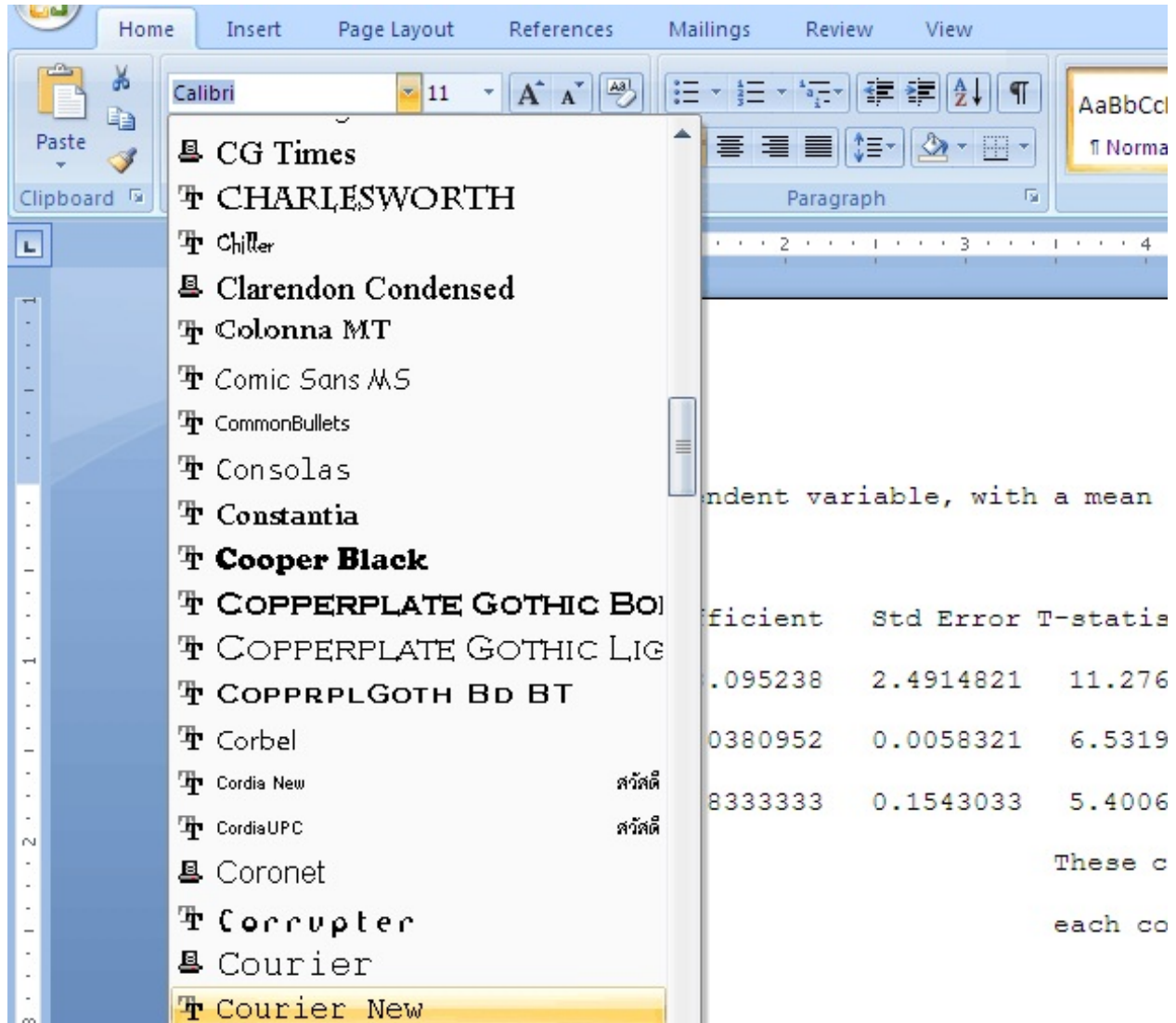
The picture to the right shows some text in Word after it is selected.



Under the Home tab (older Word versions are different), click on the font selector and pick Courier New, or any font that has all the letters the same width.

If a text table is too wide and the lines wrap around, select the text again and change the font size to a smaller number. (The residuals plot may need a font size of 10 or less.)

The font selector in Word



When you have all your results from LS in place, lined up, and readable, go through and type in or copy in at the appropriate places what you wrote when I asked you in the instructions above to do this or that “for your write-up.” That will give you a nice report! Save the file and submit it on Blackboard.

Assignment 3 hand-in check list

1. Your data.
2. The correlation between Fertilizer and Rain.

Multiple Regression

3. The multiple regression results.
4. Your multiple regression equation with the estimated coefficients in it.
5. How much extra yield you get from adding a pound of fertilizer, and from adding an inch of rain.
6. The residuals' plot
7. The tests of the hypotheses that the coefficients of Fertilizer and Rain are 0.
8. The predicted yield when using 800 pounds of fertilizer and 20 inches of rain.

Simple Regression

9. The simple regression results.
10. Your simple regression equation.
11. How much extra yield you get from adding a pound fertilizer, according to this regression.
12. The test of the hypothesis that the coefficient of Fertilizer is 0.
13. Your comment on why there is a difference between what the multiple regression says the effect of fertilizer on yield is and what the simple regression says fertilizer's effect is.
14. The predicted yield when using 800 pounds of fertilizer.
15. Your comment on why this prediction is larger than the prediction from the multiple regression.