Assignment 5. Non-Linear Multiple Regression

This assignment shows how to use a least squares program to fit a curve. The basic idea is to transform the data so your curved relationship becomes a linear relationship. Page 8 lists everything I'd like you to turn in.

A log-linear equation

We're going to re-do the AFDC-UP regression with this functional form:

$$AFDCUP\% \;=\; A \times UE\_RATE^{b_2} \times UI\_AVG^{b_3} \times INCOME^{b_4} \times HIGH\%^{b_5} \times NEED^{b_6} \times PAYMENT^{b_7} \times u$$

u is the error. Notice that the equation multiplies by *u*, rather than adding *u*. For this to work, *u* has to have an expected value of 1, and be always greater than 0. This is different from the linear equations we have been using so far, for which the error could be positive or negative and was assumed to have an expected value of 0. In the above equation, when u is bigger than 1, AFDCUP% is above its expected value. When u is 1, actual AFDCUP% is right on its expected value. When u is between 0 and 1, AFDCUP% is below its expected value.

AFDCUP% can't possibly be 0 or negative with this equation, so long as all the values of the right-side variables are bigger than 0.

For your write-up: Why does this assure that we get a more sensible-looking prediction that what we got last week?

Taking the ln (ln means $\log_e$) of both sides of the above equation gives you:

$$Ln(AFDCUP\%) = ln(A) + b_2 \times ln(UE\_RATE) + b_3 \times ln(UI\_AVG) + b_4 \times ln(INCOME) + b_5 \times ln(HIGH\%) + b_6 \times ln(NEED) + b_7 \times ln(PAYMENT) + ln(u)$$

This is now linear in the parameters. If we define:

LnAFDCUP% as equal to *Ln(AFDCUP%),*
a as equal to *ln(A)*,
LnUE_RATE equal to *ln(UE_RATE)*,
LnUI_AVG equal to *ln(UI_AVG)*,
LnINCOME equal to *ln(INCOME)*,
LnHIGH% equal to *ln(HIGH%)*,
LnNEED equal to *ln(NEED)*,
LnPayment equal to *ln(PAYMENT)*, and
v equal to *ln(u)*,

then we have this linear equation:

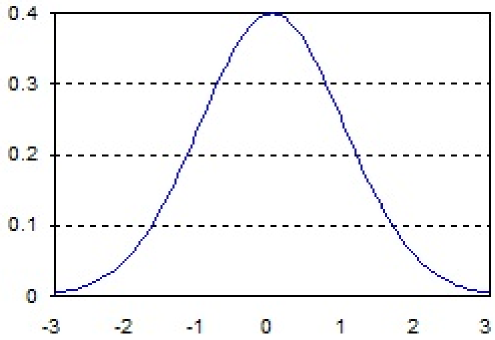
$$LnAFDCUP\% = a + b_2 \times LnUE\_RATE + b_3 \times LnUI\_AVG + b_4 \times ln(INCOME) + b_5 \times LnHIGH\% + b_6 \times LnNEED + b_7 \times LnPAYMENT + v$$

With this, we can use the linear least squares estimator (meaning, the least squares method we have been using in assignments 3 and 4).

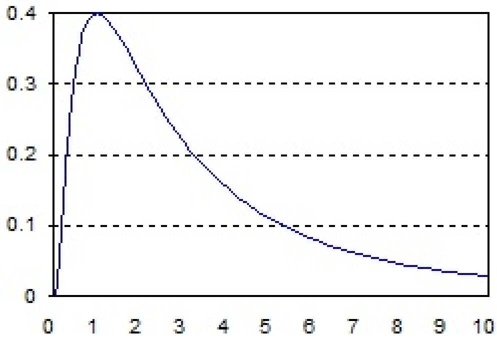
This is why the error *u* is multiplicative in the equation for AFDCUP%. In other words, this is why the near the top of page 1 ends with ×*u* rather than +*u*. It is set up that way so that, when we take logs of both sides, we get an error *v* that is added or subtracted. That is the form we want if we are using least squares.

Similarly, the first equation's error u has an expected value of 1, so that the error *v* in the transformed equation has an expected value of 0. We want that, too, if we are going to use least squares. (The logarithm of 1 is 0.)

To do conventional hypothesis testing, we further assume that *v* has the normal distribution. This implies that *u* has what's called a log-normal distribution.



A normal density function. In linear regression with hypothesis testing, we assume the error is like this. We assume *v* is like this. (This example has a standard deviation of 1.)



The corresponding log-normal density function. We assume that *u* is like this.

To implement the new equation, we have to create the variables *LnAFDCUP%* , *LnUE_RATE*, *LnUI_AVG*, *LnINCOME*, *LnHIGH%*, *LnNEED*, and *LnPayment* from our data. Here's how:

Transforming the AFDC data to logarithms

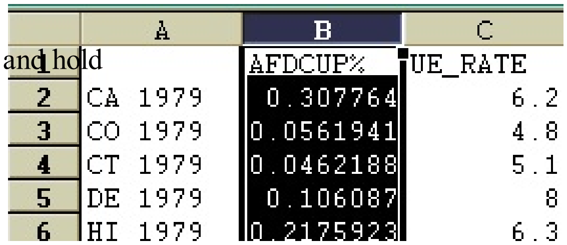
Open AFDC.XLS , the Assignment 4 data file, in your spreadsheet program. If you did not save it last week, get it from the Data for Assignments 4 & 5 link on the syllabus.

The first few rows of the AFDC data file look like this:

WWe must calculate the logarithms of all seven of the variables in the data set. We'll create seven more variables, each in its own column.

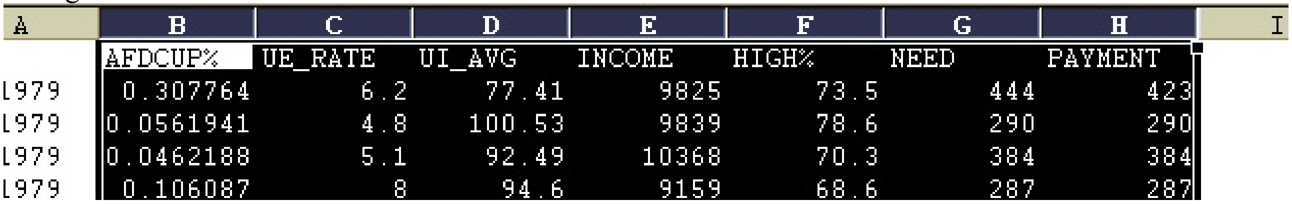
Let's move the whole data set to the right by seven columns, opening up columns B through H. This is so we can put the logarithms in columns B through H.

To insert the columns, move the mouse pointer to the B box at the top of the B column. Press and hold down the left mouse button. The B column should turn black. Don't let up on the mouse button yet.



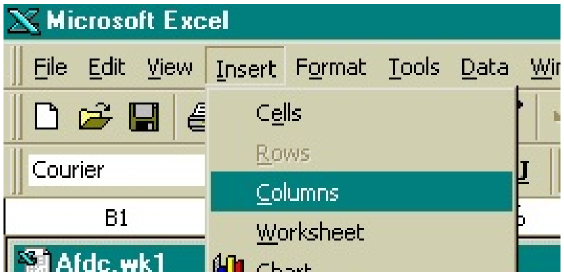
	A	B	C
1		AFDCUP%	UE_RATE
2	CA 1979	0.307764	6.2
3	CO 1979	0.0561941	4.8
4	CT 1979	0.0462188	5.1
5	DE 1979	0.106087	8
6	HI 1979	0.2175923	6.3

With the mouse button down, drag the pointer over to the top of the H column. All of the columns from B through H should turn black.



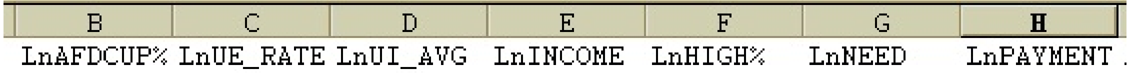
A	B	C	D	E	F	G	H	I
	AFDCUP%	UE_RATE	UI_AVG	INCOME	HIGH%	NEED	PAYMENT	
1979	0.307764	6.2	77.41	9825	73.5	444	423	
1979	0.0561941	4.8	100.53	9839	78.6	290	290	
1979	0.0462188	5.1	92.49	10368	70.3	384	384	
1979	0.106087	8	94.6	9159	68.6	287	287	

Go up to the top menu and click on Insert, then Columns.



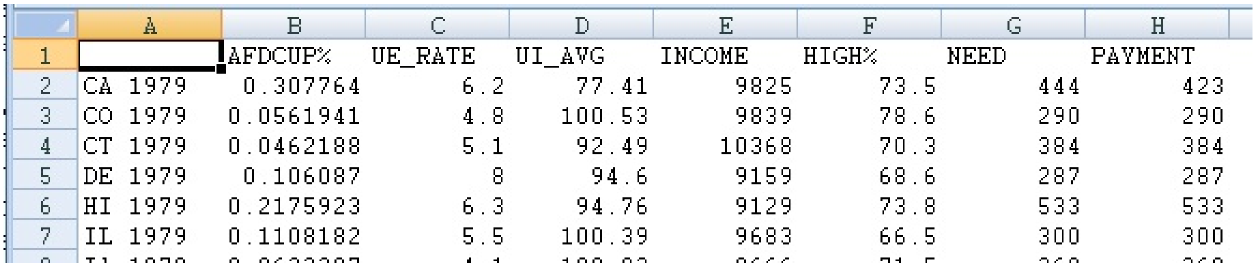
The contents of columns B through H will move to the right, to columns I through N. Columns B through H will now be blank.

Next, click on cell B1 to turn the inserted area white. Then put the names for the new variables we'll be creating in cells B1 through H1. I suggest making the new names "Ln" plus the old name, like this.



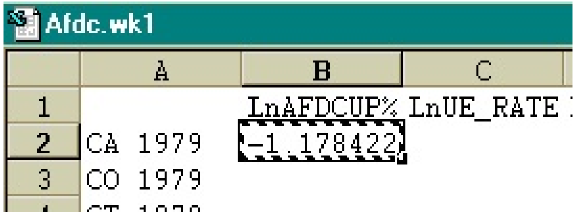
B	C	D	E	F	G	H
LnAFDCUP%	LnUE_RATE	LnUI_AVG	LnINCOME	LnHIGH%	LnNEED	LnPAYMENT

Move the cell selector to B2, under LnAFDCUP%. Type `=ln(I2)` and press Enter ↵. You should see



	A	B	C	D	E	F	G	H
1		AFDCUP%	UE_RATE	UI_AVG	INCOME	HIGH%	NEED	PAYMENT
2	CA 1979	0.307764	6.2	77.41	9825	73.5	444	423
3	CO 1979	0.0561941	4.8	100.53	9839	78.6	290	290
4	CT 1979	0.0462188	5.1	92.49	10368	70.3	384	384
5	DE 1979	0.106087	8	94.6	9159	68.6	287	287
6	HI 1979	0.2175923	6.3	94.76	9129	73.8	533	533
7	IL 1979	0.1108182	5.5	100.39	9683	66.5	300	300

`-1.178422` in B2. This is the base e logarithm of 0.307764, the number in I2. $e^{-1.178422}$= 0.307764.



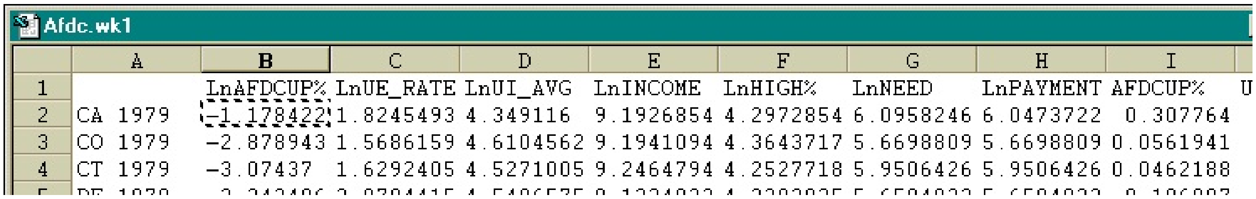
	A	B	C
1		LnAFDCUP%	LnUE_RATE
2	CA 1979	-1.178422	
3	CO 1979		

Move the cell selector to B2. Copy that cell to the clipboard, using the menu (`Edit Copy`), or the keyboard (Ctrl+C), or the copy icon 📋 .

This copies B2 to the Windows clipboard.

For the destination, select (turn black) the entire block from B2 down and over to H119. Do this by clicking on B2 and dragging to H119. Alternatively, click on B2, then press and hold ⇧Shift while you use the arrow keys to extend the block over and down to H119.

Paste from the Windows clipboard, by selecting Edit Paste from the menu, pressing Ctrl+V, or clicking the paste icon 📋 . In an eye-blink, Excel will calculate and display 826 logarithms for you. Click on any spreadsheet cell to get rid of the black background.



	A	B	C	D	E	F	G	H	I
1		LnAFDCUP%	LnUE_RATE	LnUI_AVG	LnINCOME	LnHIGH%	LnNEED	LnPAYMENT	AFDCUP%
2	CA 1979	-1.178422	1.8245493	4.349116	9.1926854	4.2972854	6.0958246	6.0473722	0.307764
3	CO 1979	-2.878943	1.5686159	4.6104562	9.1941094	4.3643717	5.6698809	5.6698809	0.0561941
4	CT 1979	-3.07437	1.6292405	4.5271005	9.2464794	4.2527718	5.9506426	5.9506426	0.0462188

Use `File Save As` from the menu to save the spreadsheet. Save under a different name than AFDC.XLS, just in case you messed up and need to start over. If you are in the lab, save to your device in E: or your disk in A:.

Before you cut and paste your data into LS, take a moment to use Excel do a calculation. To make a prediction for S.C. with this model, you will need the logarithms of S.C.'s numbers. Those logarithms are what you will plug in to the prediction form in LS.

Here are the S.C. numbers from last week's assignment:

UE_Rate = 6.6
UI_AVG = 95
INCOME = 10729
HIGH% = 53.7
NEED = 425
PAYMENT = 238

Bring up a new blank spreadsheet by selecting File New... from Excel's menu. Then select a plain Workbook.

A new blank spreadsheet appears. (The AFDCUP data spreadsheet you just worked on is still available, so don't worry.)

On the blank spreadsheet, type the S.C. numbers into a column – cells A1 through A6 will do fine. In the next column over, calculate the natural log of each of them with the =ln function. For example, if the S.C. numbers are in A1 through A6, cells B1 through B6 should have the formulas `=ln(A1)` through `=ln(A6)`. Print the spreadsheet or write down the values of the logs for later use.

Now switch back to the spreadsheet that has the AFDCUP data. In Excel, you can use the Window item on the menu to switch among active spreadsheets.

We are ready to copy the data from this spreadsheet and paste them into LS.

Select the block of cells from A1 in the upper left down to H119. Notice that we are not selecting all of the columns. We are just selecting the column of observation names and the columns for the Ln variables. The reason for this is that we only plan to use the Ln variables in our regression equation. After all, our formula is

$$\begin{aligned} LnAFDCUP\% = \quad & a + b_2 \times LnUE\_RATE + b_3 \times LnUI\_AVG + b_4 \times ln(INCOME) + b_5 \times LnHIGH\% + \\ & b_6 \times LnNEED + b_7 \times LnPAYMENT + v \end{aligned}$$

All the variables in that formula are log variables.

When you have A1:H119 selected, copy that block of cells to the clipboard by pressing Ctrl+C or whatever other way you prefer.

Run the regressions

Start up LS by going to the syllabus and clicking the LS link. This opens http://sambaker.com/ls

Click in the upper box on that page and give the paste command (Ctrl+v should work). The box should fill in with your data.

Click button 2. If everything is OK, the dependent variable checkboxes will appear.

Select LnAFDCUP% as your dependent variable.

The independent variables should be the Intercept and the logs of all the other independent variables. If you selected the A1:H119 block for copying, as I suggested above, you can leave all the check marks where they are. If your pasted data included any other variables, uncheck them.

Click on Go. The regression results should appear.

For your writeup: Report your coefficients. Which variables have significant coefficients and which do not? In particular, what happened to Need and Payment compared with last week? Which of these two variables now appears to really affect AFDC-UP caseload? (Note: If an X variable has a significant coefficient, then it affects [or, at least, is related to] the Y variable. If an X variable has an insignificant coefficient, it does not affect the Y variable.)

The coefficients in a logarithmic equation like this are estimated elasticities. For those of you who didn't just take HSPM 712, the elasticity of Y with respect to X is the percentage change in Y that you get if X changes by 1%.

For your write-up: Test the hypothesis that the true coefficient of LnNEED is 1.0, reporting your method and result, including what the result implies . An elasticity of 1 would mean that a 1% increase in the eligibility level would bring in 1% more AFDC-UP cases. To do the hypothesis test, calculate the expression to the right. β is the estimated coefficient, of LnNEED in this case. β_0 is the hypothetical value you want to test, 1 in this case. This expression has the t distribution, with degrees of freedom equal to the number of observations minus the number of coefficients (including the intercept) in your equation.

$$\frac{(\hat{\beta}-\beta_0)}{\textit{Standard Error of } \hat{\beta}}$$

For a hypothesis test, compare this with the value from the t table.

Also test the hypothesis that the true coefficient of LnUE_RATE is 1.0. If you reject this hypothesis, you've found that there's a more-than-proportional response of welfare caseload to changes in unemployment. Write a sentence saying that after you show your work. An elasticity greater than 1 for unemployment would make sense because to qualify for welfare you must not only have very little income, you must also have almost no savings or other assets. When unemployment rates are high, more people have been out of work longer, so more of them have spent down their savings and sold their assets, thus making themselves poor enough for welfare.

View the residuals plot. Sort them by the predicted value of the dependent variable, LnAFDCUP%.

For your write-up: Comment on the pattern of the residuals. How does it compare with last time?

Next, predict LnAFDCUP% for South Carolina, using the independent variables' logarithm values that you calculated according to the instructions on page 4.

For your write-up: Report your prediction and its 90% confidence interval.

Transforming the predictions back from logarithms

We're not done yet! You must convert your prediction and confidence interval for LnAFDCUP% into a prediction and confidence interval for AFDCUP%. This requires raising e to the power of the numbers that you just reported.

Get back to Excel. Use a new spreadsheet or a blank portion of your current spreadsheet. Type the three prediction numbers (the prediction itself, the upper end of the 90% confidence interval, and the lower end of the confidence interval) in a column. In the next column over, calculate =exp of each of those numbers. This will give you your prediction and confidence interval for LnAFDCUP%.

For example, if your three prediction numbers are in cells A1, A2, and A3 of an otherwise blank spreadsheet, move the cell selector to B1 and type =exp(A1) . Then copy that cell to B2 and B3. The B column will have your prediction and confidence interval for AFDCUP%.

For your write-up: Report your prediction for AFDCUP% and your 90% confidence interval. Notice that this confidence interval is not symmetrical around the predicted value.

In another column over, divide each of those predicted numbers by 100 and multiply by 1.7 million (which is 1.7E6). That gives you the prediction and confidence interval in terms of numbers of people on AFDC-UP. (You divide by 100 because AFDCUP% is in terms of percent of the labor force. You multiply by 1.7 million because that was the S.C. labor force.)

(Spreadsheet note: The formula for dividing what's in B1 by 100 and multiplying by 1.7 million is: `=B1/100*1.7E6` .)

For your write-up: Report your prediction for the number of AFDC-UP families, showing how you calculated it. Does this prediction make more sense than the one you got last week? Why?

When they planned the program, state authorities projected a caseload of 3500. Actual caseload in 1987 and 1988 averaged about 450.

For your write-up: Whose prediction turned out to be more accurate, yours or theirs?

(See next page for hand-in check-list.)

Assignment 5 Checklist

What to Hand In

- Explain why this week's equation is likely to be better for predictions than the linear form used last week.

- Regress LnAFDCUP%, the $\log_e$ of AFDCUP%, on all of the other Ln variables (and the intercept). Report your coefficients. Which variables have significant coefficients and which do not? In particular, what happened to Need and Payment compared with last week? Which variable now appears to have the real impact on AFDC-UP caseload?

- Test the hypothesis that the true coefficient of LnNEED is 1.0, showing your method. Write a sentence saying what your finding implies.

- Test the hypothesis that the true coefficient of LnUE_RATE is 1.0, showing your method. Write a sentence saying what your finding implies.

- Comment on the pattern of the residuals. How does it compare with last time's linear model?

- Report the LnAFDCUP% prediction for South Carolina and the 90% confidence interval for that prediction.

- Report the antilogs (=exp) of the prediction for LnAFDCUP%, the upper end of the 90% confidence interval, and the lower end of the confidence interval.

- Translate those figures, which are for AFDCUP% -- a percentage of the labor force -- into numbers of people (families). Show your work. Does this prediction make more sense than the one you got last week? Why?

- Comment on whose prediction turned out to be more accurate, yours or the state's.