

Assignment 5. Non-Linear Multiple Regression

© 2006 Samuel L. Baker

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. As explained in the non-linear regression handout, we multiply by u , rather than adding u , as we did in preceding assignments. u has to have an expected value of 1, and will always be greater than 0. When u is bigger than 1, AFDCUP% is above its expected value. When u is 1, the actual AFDCUP% is right on its expected value. When u is between 0 and 1, AFDCUP% is below its expected value.

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, AFDCUP% can't possibly be 0 or negative, 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 will get a more sensible-looking prediction than what we got last week?

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

$$\begin{aligned} \ln(AFDCUP\%) = & \ln(A) + b_2 \times \ln(UE_RATE) + b_3 \times \ln(UI_AVG) + b_4 \times \ln(INCOME) + \\ & b_5 \times \ln(HIGH\%) + b_6 \times \ln(NEED) + b_7 \times \ln(PAYMENT) + \ln(u) \end{aligned}$$

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

$\ln AFDCUP\%$ as equal to $\ln(AFDCUP\%)$,
 a as equal to $\ln(A)$,
 $\ln UE_RATE$ equal to $\ln(UE_RATE)$,
 $\ln UI_AVG$ equal to $\ln(UI_AVG)$,
 $\ln INCOME$ equal to $\ln(INCOME)$,
 $\ln HIGH\%$ equal to $\ln(HIGH\%)$,
 $\ln NEED$ equal to $\ln(NEED)$,
 $\ln Payment$ equal to $\ln(PAYMENT)$, and
 v equal to $\ln(u)$,

then we have this linear equation (next page):

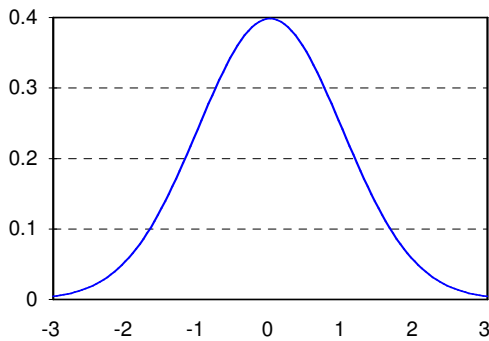
$$\text{LnAFDCUP}\% = a + b_2 \times \text{LnUE_RATE} + b_3 \times \text{LnUI_AVG} + b_4 \times \ln(\text{INCOME}) + b_5 \times \text{LnHIGH}\% + b_6 \times \text{LnNEED} + b_7 \times \text{LnPAYMENT} + v$$

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

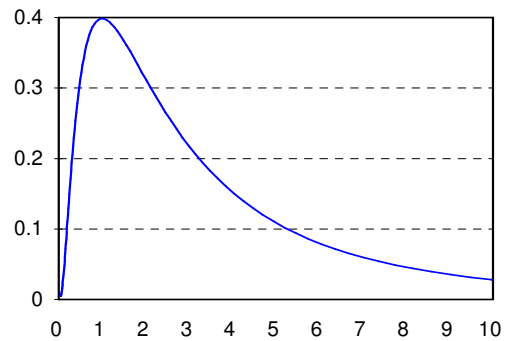
This is why the error u is multiplicative in the equation for AFDCUP%. In other words, this is why the near the top of page 1 ends with $\times u$ rather than $+u$. It is set up that way so that, when we take logs of both sides, we get an error v that is added or subtracted. That is the form we want 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 $\text{LnAFDCUP}\%$, LnUE_RATE , LnUI_AVG , LnINCOME , $\text{LnHIGH}\%$, 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 (next page):

	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

We 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 data 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, where they will be next to the observation names in column A.

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

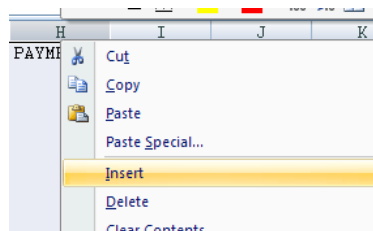
	A	B	C	D	E	F	G	H
1		AFDCUP%	UE_RATE	UI_AVG	INCOME	HIGH%	NEED	PAYMENT
2	CA 1979	0.307764	6.2	77.41	9825	73.5	444	423
3	CO 1979	0.0561941	4.8	100.53	9839	78.6	290	290
4	CT 1979	0.0462188	5.1	92.49	10368	70.3	384	384
5	DE 1979	0.106087	8	94.6	9159	68.6	287	287
6	HI 1979	0.2175923	6.3	94.76	9129	73.8	533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

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 shaded.

	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

Right-click on any of the shaded column-heading letters, B through H. (I clicked on H.)

Click on Insert in the pop-up menu.



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.

	A	B	C	D	E	F	G	H	I
78	WI 1981								0.37496
79	CA 1982								0.64601
80	CO 1982								0.12202
81	CT 1982								0.08048
82	DE 1982								0.14724
83	HI 1982								0.27913
84	IL 1982								0.26606
85	KS 1982								0.16463
86	MD 1982								0.08536
87	MA 1982								0.12722
88	MI 1982								0.98711
89	MN 1982								0.25
90	NE 1982								0.10058
91	NJ 1982								0.16996
92	NV 1982								0.17978
93	OH 1982								0.5662
94	PA 1982								0.26052
95	RI 1982								0.09103
96	VT 1982								0.24986
97	WV 1982								0.68351
98	WI 1982								0.53181
99	CA 1983								0.69410
100	CO 1983								0.1637
101	CT 1983								0.10184
102	DE 1983								0.10896
103	HI 1983								0.29244
104	IL 1983								0.3063
105	IA 1983								0.26006
106	KS 1983								0.2381
107	MD 1983								0.08812
108	MA 1983								0.10145
109	MI 1983								1.10633
110	MN 1983								0.3392
111	NE 1983								0.16704
112	NJ 1983								0.17655
113	NY 1983								0.21713
114	OH 1983								0.75924
115	PA 1983								0.28540
116	RI 1983								0.10715
117	VT 1983								0.33024
118	WV 1983								0.89612
119	WI 1983								0.71716

Next, type the names for the new variables that we'll be creating in cells B1 through H1. I suggest making the new names Ln plus the old name. The new unemployment rate variable name, for example, will be LnUE_RATE.

You can type all the variable names, or try this: Type ="Ln"&I1 in cell B1. The quotation marks designate Ln as a string of characters. The & works like a + sign for text. It concatenates text strings. The cell reference to I1 brings in the text in I1. Press **Enter** and you should see LnAFDCUP%.

Copy that cell and paste to B1:H1. All the variable names will be created. We can now fill in the numbers below them.

Move the cell selector to B2, under LnAFDCUP%. Type =ln(I2) and press **Enter**. You should see -1.178422 in B2. This is the base e logarithm of 0.307764, the number in I2. $e^{-1.178422} = 0.307764$.

Move the cell selector to B2. Copy that cell to the clipboard, using **Ctrl**+C or the Copy icon.

For the destination, select (turn shaded and outlined) 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, by 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 shaded background.

Use **Save As** from the Office Button 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 storage device.

You will be copying and pasting columns A through H into LS. First, though, stay with Excel for a few more minutes, to do a calculation that you will need later. You will be making a prediction for S.C. To do this with this logarithmic model, you will need the logarithms of S.C.'s numbers for the independent

variables.

Here are the S.C. numbers, the same as you used last week to make the prediction:

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

Switch to a new worksheet by clicking on Sheet2 at the bottom of your spreadsheet.

On the blank page, 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 logarithms for later use.

Now switch back to Sheet1, that has the AFDCUP data.

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

$$\text{LnAFDCUP}\% = a + b_2 \times \text{LnUE_RATE} + b_3 \times \text{LnUI_AVG} + b_4 \times \ln(\text{INCOME}) + b_5 \times \text{LnHIGH}\% + b_6 \times \text{LnNEED} + b_7 \times \text{LnPAYMENT} + v$$

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.

Use LS

Start up LS by going to the syllabus and clicking the LS link. This opens <http://hspm.sph.sc.edu/COURSES/J716/ls-b/ls.html>

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 available 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 in the lower box.

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. $\hat{\beta}$ 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.

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

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 on Sheet2 of your spreadsheet.

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

Transforming the prediction 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. Sheet2 would be good. Type the three prediction numbers from the LS results (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 transforms your prediction and confidence interval for $\text{LnAFDCUP}\%$ into a prediction and confidence interval for $\text{AFDCUP}\%$.

For example, if your three prediction numbers are in cells D1, D2, and D3, move the cell selector to E1 and type $=\exp(D1)$. Then copy that cell to E2 and E3.

For your write-up: Report your prediction for $\text{AFDCUP}\%$ 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 $\text{AFDCUP}\%$ is in terms of percent of the labor force. You multiply by 1.7 million because that was the S.C. labor force.)

(The formula for dividing what's in E1 by 100 and multiplying by 1.7 million is: $=E1/100*1.7E6$.)

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?

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 $\text{LnAFDCUP}\%$, the \log_e of $\text{AFDCUP}\%$, 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 $\text{LnAFDCUP}\%$ prediction for South Carolina and the 90% confidence interval for that

prediction.

- Report the antilogs ($=\exp$) of the prediction for $\text{LnAFDCUP}\%$, the upper end of the 90% confidence interval, and the lower end of the confidence interval.
- Translate those figures, which are for $\text{AFDCUP}\%$ -- 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.