Assignment 4. Multiple Regression and Collinearity

We're going to do some multiple regression analysis on a file of data on state AFDC-UP programs (Aid for Families with Dependent Children -- Unemployed Parents). These are public assistance programs for two-parent families. South Carolina started this program in 1986. Some states had had it since 1961. (In 1996, Temporary Assistance for Needy Families [TANF] replaced AFDC. Eligibility criteria are similar, but a work requirement was added.)

Imagine that it's 1985. You have been hired to predict how many families will be on AFDC-UP so that the legislature can decide how much to budget for the program. You decide to make the projection based on the experience of other states that already have AFDC-UP.

Get the data

Go to the course syllabus, at http://sambaker.com/courses/716.php . Click the link for "Data for assignments 4&5." The link is to an Excel file, AFDC.XLS. Your browser should give you the choice of saving it or opening it directly in Excel. Either choice is OK.

If you need the data in a different file format, such as a format for Lotus 1-2-3, please e-mail me.

If you downloaded the file, double-click its icon on your screen. That should start your spreadsheet and open the file.

Here are the first few lines of AFDC.XLS. Yours should be identical.

	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	0.056194	4.8	100.53	9839	78.6	290	290
4	CT 1979	0.046219	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	0.217592	6.3	94.76	9129	73.8	533	533
7	IL 1979	0.110818	5.5	100.39	9683	66.5	300	300
8	IA 1979	0.06234	4.1	108.03	8666	71.5	369	369
9	KS 1979	0.030259	3.4	88.79	9223	73.3	306	306

...

Row 1 has the names of the variables, starting in cell B1. Each row below that gives the values of those variables for one state in one year. Column A contains the state name abbreviation and the year, to identify the data in that row.

We have two goals with this little piece of research:

(1) We want to understand what affects the AFDC-UP caseload in a state.
(2) We want to predict what the caseload will be in South Carolina, when AFDC-UP is running.

Here are details about those variables:

AFDCUP% is AFDC-UP families relative to the size of the state labor force, expressed as a percentage. In other words, it is the number of families on AFDC-UP divided by the number of people in the state labor force and multiplied by 100%.

AFDCUP% will be our dependent ("Y") variable. The variables below will be independent ("X") variables,

UE_RATE is the percentage of the state labor force unemployed, averaged over the year. AFDC-UP is specifically designed to aid the unemployed. We expect a positive relationship -- the higher the unemployment rate, the higher the percentage of people on AFDC-UP.

UI_AVG is the average weekly unemployment insurance payment. Unemployment insurance gives the unemployed money for a period of time. It competes with AFDC-UP because recipients have to choose which assistance program they want. The better unemployment insurance pays, the fewer unemployed will choose AFDC-UP.

INCOME is the average annual income per capita in the state. States with richer people should have fewer people on public assistance, other things equal.

HIGH% is the percentage of people over 25 years old who had at least one year of high school. Does a more educated labor force mean fewer people on this form of welfare?

NEED is the AFDC eligibility level of monthly income for family of four. Only families whose income is less than this amount can get AFDC-UP assistance. The higher the NEED level, the more families qualify for AFDC-UP. (The name "NEED" is a euphemism. Think of it as the eligibility income ceiling for assistance.)

PAYMENT is the monthly AFDC payment for a family of four if that family has no income. Some economists argue that welfare attracts people away from work with its benefits. If this is true, then the higher the welfare PAYMENT is, the more families will choose welfare rather than work.

Get the data into LS and do the regression

In Excel, select all of the cells, from A1 at the upper left down to H119 as the lower right. Copy this block, by pressing Ctrl+C or Edit and Copy from the menu.

Start your web browser and click the LS link on the syllabus. This loads http://sambaker.com/ls

After LS loads, paste into the upper box. To paste, right-click in the upper box, bringing up a context menu, and click Paste. Alternatively, click in the upper box and press Ctrl+V .

Click button number 2 to read the data.

This might be a good time to start your word processor, such as Microsoft Word, so that you can paste in your regression results from LS's lower box and add comments.

Alternatively, you can do all of the LS work first, then copy the entire contents of LS's lower box for pasting into Word. Then come back here and read through the instructions, adding to your document the comments that I ask for.

Look at the Correlations table, in LS's lower box.

The correlations table gives you the simple correlation (r) between any two variables. For example, to find the correlation between NEED and PAYMENT, look in the row for NEED and the column for PAYMENT (or the column for NEED and the row for PAYMENT).

For your write-up:

- What is the correlation is between NEED and PAYMENT? Look in the row for NEED and the column for PAYMENT, or the row for PAYMENT and the column for NEED.

On the LS screen, below the lower box, click to choose AFDCUP% for the dependent variable. This is what we want to explain and predict.

The independent variables we want are all of the others, so you can leave them all checked.

Click the Do Regression button. Your regression results should appear in the lower box. Select them, copy them, and paste them into your word processor. To make the columns of numbers in the results line up properly, select them and change the font to Courier New or some other monospace font. In that same selected text, if the lines of text are too long and wrap to the next lines, reduce the size of the font.

Interpret the results

Being able to write up your results coherently is as important as being able to run the computer program.

For your write-up:

- Report each of the coefficients and say whether or not each is significantly different from 0. Use the p-values for this. Coefficients with p-values less than 0.05 are significant at the 0.05 level. Coefficients with p-values greater than 0.05 are not significant at the 0.05 level.

- Write a sentence for each variable explaining what you found. For example, you might write, "The positive and significant coefficient for UE_RATE means that the higher the unemployment rate, the more people are on AFDC-UP." If a variable is not significant, write that it does not have a significant relationship to AFDC-UP. (You can check your results so far with the answer checker, at http://sambaker.com/courses/J716/a04/AFDC.html .)

- Calculate (instructions are just below) and report the 95% confidence interval for the coefficient of UE_RATE.

The 95% confidence interval for a coefficient is:

Estimated coefficient ± (Coefficient's standard error)×(critical T-value from t-table)

The t-table (in 716-tables for hypothesis tests.pdf) doesn't have a row for 111 degrees of freedom. Look in the 0.05 column. You can use the 120 row to get the critical value. If you want to be safer, use the 60 row. If you want to be fancier, you can interpolate a number between what you get from the 120 row and

what you get from the 60 row.

While you are looking at the results, write down the sum of squared residuals. We'll use it later in an F test.

Residuals

Click the red button for the residuals plot.

Click the choice for sorting the residuals by the predicted value of the dependent variable. This is the right choice when the data are not in time order, or in order by one of the other X variables.

The residual plot can reveal patterns that can help you tell whether the assumptions behind using least squares are appropriate for your data. A curved relationship between the dependent variable and any of the independent variables will show up as a curve in the residual plot and a low number for the Durbin-Watson statistic. The residual plot, when ordered by the predicted value of the dependent variable, can also reveal if the residuals tend to get bigger or smaller as you go from small predicted values of the dependent variable to large values. If the residuals spread out, this suggests **heteroskedasticity**, the term for error variances being different for different observations. Such patterns in the residuals imply that there may be a better model than the linear one for your data.

Copy and paste the residuals plot into your word processor. Change the font to a font for which all the letters and spaces have the same width. Courier New is a good choice. If the lines of text in the residuals plot wrap from one line to the next, make the font smaller. You want one observation per line in the residuals plot.

For this data set, the residuals plot may take up two or three pages. Look up and down the whole plot, and then...

For your write-up:

- Describe the pattern of the residuals in a sentence. (If you do not see anything special in the residuals, that is OK. We will discuss this in class.) Report your Durbin-Watson statistic. What does it imply about whether there is serial correlation? (There is a Durbin-Watson table in the handout with the t- and F-tables. The Durbin-Watson table only goes up to 5 variables, so use the 5 column to get your critical value.)

Prediction

Calculate a predicted value of AFDCUP% for South Carolina, based on:

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

To do this, click the blue button in LS. Type the above numbers in the spaces on the form that appears.

For your write-up:

- Report your prediction as it is, even if it seems illogical. Report also the 90% confidence interval for the prediction.

Translate your prediction for AFDCUP% into a number of families. Here is how: The S.C. labor force was about 1.7 million people. Multiply that by the predicted AFDCUP% you just got. Then divide the result by 100, because AFDCUP% is a *percentage* of the labor force. (I put the "%" sign in the variable name to remind me of that.) For example, if your predicted value of AFDCUP% for South Carolina is 0.02, make this calculation: 1,700,000 × 0.02 × 0.01 = 340. Again, report your result, even if it seems illogical. The answer checker can verify your prediction.
http://sambaker.com/courses/J716/a04/AFDC.html

(Spreadsheet note: The formula for dividing what's in B1 by 100 and multiplying by 1.7 million is: `=B1/100*1.7E6`. Excel may change the 1.7E6 to 1700000 after you complete entering the formula.)

For your write-up:

- How many AFDC-UP families will we have statewide? How many families will there be if AFDCUP% is at the high end of the 90% confidence interval?

- Does the prediction you got (and the pattern of the residuals, if you saw something remarkable) make you think this model appropriate or inappropriate? Explain.

A test of the hypothesis that both NEED and PAYMENT have coefficients of 0.

Do another regression on the same data. In the main LS window, click to take the check marks out of the boxes for NEED and PAYMENT in the list of independent variables. That takes these two variables out of the equation, creating a "reduced model." Click the Go button. Write down the sum of squared residuals from this reduced model regression.

Use this formula to calculate an F statistic:

$$F = \frac{[\ SSR(RM) - SSR(FM)\ ] / [\ P_{FM} - P_{RM}\ ]}{SSR(FM) / [\ Observations - P_{FM}\ ]}$$

SSR(RM) means the sum of squared residuals from the reduced model, the model with some variables taken out.
SSR(FM) means the sum of squared residuals from the full model, the model with all the variables in it.
P_{FM} is the number of parameters (coefficients) in the full model, counting the intercept.
P_{RM} is the number of parameters (coefficients) in the reduced model, counting the intercept.
Observations is how many data points you have.

When you look up the critical value in the F table, [P_{FM} - P_{RM}] is the numerator degrees of freedom. [Observations - P_{FM}] is the denominator degrees of freedom. The F-table in the handout has no row for 111. Use the 60 row if you want to be safe, or interpolate between the 60 and 120 numbers if you want to be fancy.

The answer checker (http://sambaker.com/courses/J716/a04/AFDC.html) has help with the F-test.

For your write-up:

- From the F-test result, can you reject the hypothesis that both coefficients are 0?

- When you use t-tests (the p-values are t-tests) to judge whether NEED and PAYMENT are significant, you come to a conclusion. When you use the F test to judge whether NEED and PAYMENT are significant, do you come to a different conclusion? If so, explain why this happened. Are NEED and PAYMENT important determinants of AFDC-UP or not?