Labs for Unit B1: Correlation and Regression

Brendan Halpin

January 2018

1 Lab 1

1.1 Correlation

Give Stata the following command to load a file, which contains six pairs of variables, `x1` and `y1`, `x2` and `y2` etc:

```
use http://teaching.sociology.ul.ie/so5041/labs/correl
```

First, graph all six pairs in scatterplots: `scatter x1 y1` etc. What sort of association do you see in each case (positive, negative, none, strong, weak)? Make a guess what the value of the correlation coefficient might be (write it down).

Then for each graph, get the correlation coefficient: e.g., `corr x1 y1`. How do the reported correlation coefficients correspond with those you guessed?

1.2 Linear Regression

The following code will run a regression with county-level crime rate as the outcome, and county-level median income as the explanatory variable (the data are from Florida).

```
clear
use http://teaching.sociology.ul.ie/ws/data/floridacrime

scatter crime income

reg crime income
```

Run the code and examine the output.

- Write out the $Y = a + bX$ equation
- Report R^2
- Test the hypothesis that income is associated with crime
- Predict crime for income = 20 and income = 30
- Draw the regression line (on paper)
- Calculate the predicted value and residual (error term) for Columbia county and relate them to the observed value

You can verify your predicted values and the line, by getting Stata to do the work:

```
predict ypred
list income crime ypred
line ypred income || scatter crime income
```

1.3 NLSW:

Execute the following commandsto load the *National Longitudinal Study of Women* data set that comes with Stata.

```
clear
sysuse nlsw88
```

Generate a new variable containing the natural log of wage:

```
gen lw = ln(wage)
```

For technical reasons, the log of variables like wage often work better in regression models. Considering the following list of variables:

- age
- ttl_exp, total lifetime work experience
- grade, years of education
- union, whether a member of a union

Let's consider wage (logged) as the "dependent variable", to be explained by the others (ignoring union for the moment as it only has two values). Create scatterplots for log-wage (on the Y-axis) compared with each of the other variables. Consider the correlations too (e.g., corr ttl_exp lw). Can you see much of a relationship?

Now do regression analyses: reg lw *varname*, with each of the other variables **one at a time** as the independent. There are two things to look at: the R^2 figure and the parameter estimate (B for the independent variable, along with its significance). Which variables affect wage much? Do any not affect it at all?

Interpret the results: in each case ask the question, "what happens to the predicted value of income, if the value of X were to change by one unit?". For two different values of the independent variable (X) calculate the predicted value of income – see where these fall on the scatterplot, and see where the regression line would lie. Does it seem like a good summary of the relationship?

If R^2 is big, the independent variable "explains" the dependent variable "a lot". However, it is possible for R^2 to be small and yet for the independent variable to a systematic effect (i.e. very low p-value for significance): this independent variable may be only one thing among many that affect the dependent variable.

1.4 Union effects

Test the effect of union on logged wage. Use a t-test in the first instance, and then fit a regression. Compare the results.

Do the same relating grade to union. Note that unionised workers tend to earn more and be better educated. Could it be that the union effect is simply due to them being better educated? That is, for workers with similar education does union status matter?

Fit the wage/grade regression for unionised and non-unionised workers separately, and think about the results (make scatterplots too). Do:

```
reg lw grade if union==0
predict p0
reg lw grade if union==1
predict p1
```

Having saved the predicted values, we can plot the two regressions simultaneously:

```
line p0 p1 grade || scatter lw grade
```

The || syntax allows us to combine plots. The following example exploits this a little more to make a plot that distinguishes by union even better:

```
line p0 p1 grade || scatter lw grade if union==0 || scatter lw grade if union==1
```

1.5 Two explanatory variables

You can also fit a model with both union status and grade explaining wage. Fit a regression with both `grade` and `union` as explanatory variables. Interpret the parameter estimates.

Compare your results to the previous separate regressions, and the t-test.