Labs for Unit B1: Correlation and Regression

Brendan Halpin

2013-01-13 Sun

1 Lab 2

1.1 Crime: spurious association

The text book (Agresti & Finlay) discusses county-level data on crime. You can load it into Stata as follows:

```
use http://teaching.sociology.ul.ie/so5032/labs/agresticounties.dta
```

Graph crime rate against education, and against income. Note the relationship.
For example:

```
scatter c hs
```

Repeat the exercise distinguishing between counties according to level of urbanisation. E.g.

```
scatter c hs if u<=60 || scatter c hs if u>60
```

(Even better, make a variable with several levels of urbanisation, to get more detail in the graph.) Does this change your interpretation?

Approach the same issue using regression:

```
reg c hs
```

What is the effect of education on crime? Is this plausible? Now add the counties' level of urbanisation to the regression. What changes and what does it mean?

```
reg c hs u
```

1.2 Direct and indirect effects

Intergenerational transmission can often work through chains of causality. This data set contains information on the respondent's job prestige score, his/her father's job prestige score when he/she was growing up, and his/her level of education:

```
clear
use http://teaching.sociology.ul.ie/ws/halpin/indirect
```

Look at the relationships between all pairs of variables, and then focus on the father's score/own score relationship.

1.3 Interaction effects

Interaction is where the effect of one variable depends on the level of another. For instance, in predicting income, both age and education may have direct effects, but the effect of education may be less for older people (due to the time elapsed since leaving education). That is, an additional year of education may boost income more for young people than for older people. We can accommodate this by adding an extra variable to the model, consisting of age multiplied by education. If this is significant, it tells us how the effect of education changes for each year of age.

1.3.1 Interaction between binary and continuous variables

A small data set drawn from the British Household Panel Survey allows us to look at an interaction between gender and hours in predicting income. Load the data and execute the following syntax:

```
clear
use http://teaching.sociology.ul.ie/so5032/labs/ojbhrs.dta
gen female = osex==2
gen inter = female*ojbhrs
reg ofimn female
reg ofimn female ojbhrs
reg ofimn female ojbhrs inter
```

Draw the regression lines for the second and third models. Fit bi-variate models for men and women separately (i.e. only ojbhrs as the explanatory variable) and compare the findings from those models.

1.3.2 Interaction between two continuous variables

Use the NLSW88 data file, and create the log of wage as before. Fit a regression model that predicts the log of wage using `ttl_exp` and `grade`. We can create an interaction between `ttl_exp` and `grade` by multiplying them together: add this variable to the model.

Note that Stata has syntax that can fit interactions more conveniently:

```
clear
sysuse nlsw88
gen lw = log(wage)
gen intvar = ttl_exp*grade
reg lw ttl_exp grade intvar

reg lw c.ttl_exp##c.grade
```

How do you interpret the model? Try by drawing the line relating `ttl_exp` to `lw` for grade at specific values (say 6, 10, 14). Do the same thing for `grade`'s effect on `lw` setting `ttl_exp` at 6, 10, 14.

1.4 Inference when adding variables

1.4.1 F-tests

Stata's regression output presents the result of an F-test against the null model (top-right of output) but doesn't do incremental F-tests. A handy add-on for this can be installed using `ssc install ftest`. Using it means you need to fit a model, store its details, fit another and compare the two:

```
ssc install ftest
clear
use http://teaching.sociology.ul.ie/so5032/labs/agresticounties
reg c u
estimates store urban
reg c u i hs
ftest urban
```

Interpret that result, and compare it with the result of testing `reg c i hs` and `reg c i hs u`.

1.4.2 Adjusted R^2

F-tests can be used to globally test a model, and also do compare two models, one with extra variables. An approximate but quicker way to do this is to look at Adjusted R^2, which is R^2 scaled to take account of the number of cases and number of parameters, in a calculation similar to that for the F-statistic. Adjusted R^2 *can* fall as variables are added to the model, unlike R^2, if their contribution is insignificant.