Labs for Unit B1: Correlation and Regression

Brendan Halpin

June 1, 2022

1 Lab 3

1.1 Predicting house prices

Data on house price and characteristics are available in this do file:

```
do http://teaching.sociology.ul.ie/so5032/labs/agrestiHousePrice.do
```

Using t-tests on individual parameters, and overall F-tests/adjusted R^2, search for a good model to predict house price. Use dummy variables where appropriate (see below).

1.2 Note: Dummy variables

If you have a categorical explanatory variable, you can enter it as a set of n-1 "dummy" variables, where n is the number of values. A dummy variable is a variable taking the values 0 and 1, indicating that the original variable takes the appropriate value:

Original	d1	d2	d3
1	1	0	0
2	0	1	0
3	0	0	1
4	0	0	0

In this example, the original value takes the values 1 to 4. There are three dummy variables, d1 to d3, taking the values 0 and 1, each corresponding to one value of the original variable. For value 4 of the original variable, all three dummy variables have the value 0. Once the dummy variables are entered in a regression analysis, the interpretation of their parameter estimates is the effect on the dependent variable of being in this category compared with category 4.

You can create dummy variables easily in Stata:

`tab bathroom, gen(d)`

However, you don't need to. You can simply use "factor notation": `reg price size i.bathrooms`. This gives the same result as putting in `d2` and `d3`. Try both ways to satisfy yourself.

1.3 Murder, mayhem and more model search

This data set:

```
use http://teaching.sociology.ul.ie/ssrm/unitb1/agrestistates.dta
```

contains information on US states, detailing various crime and other statistics. Explore the data set, and search for a good model to predict the murder rate, using t-tests, Adj-R^2, F as appropriate. Start by looking at bivariate regressions, then move on to fuller models.

Does it make logical sense to use the violent crime rate as a predictor?

Do you observe **multicollinearity** here? That it, correlated explanatory variables that are strong individually but both weak when entered together?

1.3.1 Residuals

When you have a model with which you are satisfied, generate residuals:

```
predict e, resid
```

Examine them: look at their distribution, their scatter plot with other variables, and examine cases with large residuals (positive or negative).

```
scatter e h
histogram e, normal
qnorm e
```

1.3.2 Cook's Distance

A measure of how much influence a case has on the model is "Cook's Distance". This is related to the residual, but is a better measure of how much the case affects the estimation. It is available from the same command as the residuals:

```
predict cook, cooksd
```

Generate it, and examine it in comparison with the other variables, and with the residuals. Ideally most cases will have similar values.

What case has the highest value?

Remove the case with the highest Cook's Distance from the data and fit the regression again. Do the results differ? If so, how would you explain that? In research terms, would it make sense to remove this case, would it change the substantive inferences you could make?

1.4 Non-linearity

1.5 Modelling a non linear effect

We can use logs, squared terms and grouped variables to model non-linear relationships with linear regression.

Load this file, which contains country-level statistics on GNP and birth rate, *inter alia*:

```
do http://teaching.sociology.ul.ie/so5032/birth.do
```

Make a scatterplot relating birth rate to GNP. What is the form of the relationship? Then fit models predicting birth rate using GNP as (i) a linear effect, (ii) a quadratic effect (GNP plus squared GNP) (iii) logged GNP and (iv) a grouped effect. Consider the fit of the four models.

Plot the four predicted values as lines/curves on the same graph:

```
sort gnp
scatter bir
```

How do they compare? Plot the residuals as well.

1.6 Non-linear age: quadratic, cubic and more?

This data file contains information on income and age.

```
use http://teaching.sociology.ul.ie/ssrm/unitb1/agenonlin
```

Examine the relationship between age and income, and fit regressions to capture this, using polynomials (start with squared age but increase as necessary). At each stage examine the fit by plotting the fitted values:

```
predict p
sort age
scatter income age if income<6000 || line p age
```

Two alternatives: group age into bands, and non-parametric curve drawing (lowess).

Grouping:

```
gen ageg = int(age/5)*5
reg income i.ageg
predict pg
scatter income pg age
```

Lowess draws a smooth curve. It's non-parametric (no simple statistical model, just summarises the data, more or less smoothly according to the `bw()` option).

```
lowess income age
lowess income age, bw(0.1)
```