Lab 3: Regression with multiple explanatory variables

July 6, 2016

Today's lab will have less instruction, so it is on you, as a budding statistician to provide a bit of creativity and apply what we have learned so far. Previously, we have only considered regression with one X variable and one Y variable. That means were are assuming models of the form

$$Y_i = a + bX_i + \epsilon_i.$$

However, we can also easily add in other explanatory variables so that

$$Y_i = a + b_1 X_{1i} + b_2 X_{2i} + \ldots + b_p X_{pi} + \epsilon_i.$$

The formula for fitting a best fit line (in terms of minimizing the squared errors) is a bit more complicated, but we use the same idea picking the coefficients (the b's) to minimize-

$$\sum_i (y_i - \hat{y}_i)^2.$$

Today, we will be looking at recent data from the UK "Brexit" vote. If you, like many Britons[1], aren't familiar with the European Union is, you can read more about the whole story here

http://www.vox.com/2016/6/17/11963668/brexit-uk-eu-explained.

In particular, the response variable we will be using is the percentage of individuals who voted to remain in the European Union in each local authority. We will be looking at several explanatory variables including

- Percentage of individuals born in the UK
- Percentage of individuals with no formal education beyond compulsory education
- Percentage of individuals working in manufacturing
- Percentage of individuals working in finance
- Percentage of individuals over the age of 60
- Percentage of individuals between the ages of 20 and 35

Each row in the data represents a local authority/distict in either England or Wales. The "Brexit" vote took place in 2016, and the explanatory variables were collected in the 2011 census. Local Authorities with missing data have been removed.

```r
# To get off website
brexit.data <- read.csv("http://www.stat.washington.edu/~ysamwang/notes/brexit_data.csv")

# If you have downloaded it locally
brexit.data <- read.csv("brexit_data.csv")
head(brexit.data)
```

[1] Google searches for "What is the EU" spiked in the UK. Unfortunately, the spike occured after the vote had occured. http://www.npr.org/sections/alltechconsidered/2016/06/24/480949383/britains-google-searches-for-what-is-the-eu-spike-after-brexit-vote

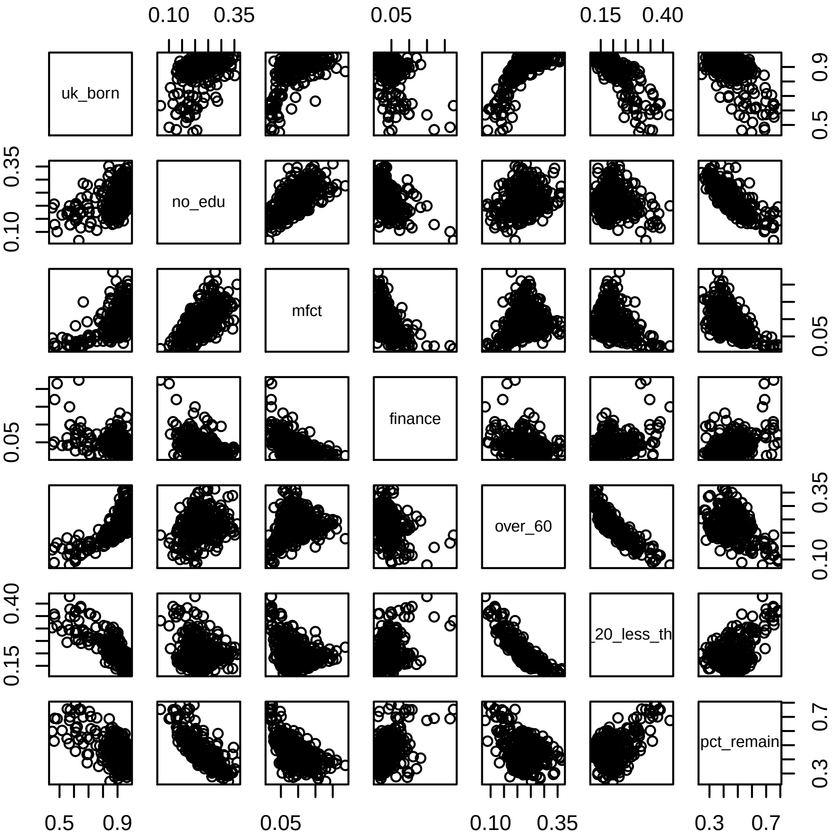
```
##              geography   uk_born    no_edu       mfct    finance   over_60
## 1           Darlington 0.9475295 0.2481226 0.09997144 0.03835639 0.2263366
## 2        County Durham 0.9675338 0.2750048 0.13156555 0.02221647 0.2360407
## 3           Hartlepool 0.9721932 0.3065959 0.11676861 0.02089125 0.2211066
## 4        Middlesbrough 0.9178539 0.2989068 0.08121437 0.02489596 0.1934081
## 5       Northumberland 0.9717588 0.2387226 0.09236833 0.02368942 0.2635178
## 6 Redcar and Cleveland 0.9776663 0.2842061 0.10318700 0.01957270 0.2520029
##   over_20_less_than35 pct_remain
## 1           0.1926604     0.4382
## 2           0.1937371     0.4245
## 3           0.1911049     0.3043
## 4           0.2263821     0.3452
## 5           0.1636817     0.4589
## 6           0.1780406     0.3381
```

Questions

- What direction do you think the association is between each of these variables?
- What strength do you think the association is between each of these variables?

As shown in lab 2, we can use the `pairs` command to plot the many pairs of variables at once. Note that we've excluded the first column here, since that's just the name of local authority

```
pairs(brexit.data[, -1])
```



Questions

- Does this look like what you might expect?
- What sticks out?
- Do the relationships look roughly linear?

1 Multivariate Regression

When there are multiple variables, we can still use the regular `lm` command, but we need to specify more variables in our formula. Notice now on the right hand side of the ∼, we have multiple variables which are seperated by the + sign. We can add additional variables simply by using the + sign.

```
output <- lm(pct_remain ~ uk_born + no_edu, data = brexit.data)
summary(output)
```

```
##
## Call:
## lm(formula = pct_remain ~ uk_born + no_edu, data = brexit.data)
##
## Residuals:
```

```
##       Min        1Q    Median        3Q       Max 
## -0.132640 -0.035044 -0.005769  0.030399  0.206090 
## 
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept)  1.01641    0.02606   39.00   <2e-16 ***
## uk_born     -0.32934    0.03220  -10.23   <2e-16 ***
## no_edu      -1.19710    0.06604  -18.13   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 0.05556 on 341 degrees of freedom
##   (1 observation deleted due to missingness)
## Multiple R-squared:  0.6818,Adjusted R-squared:  0.6799 
## F-statistic: 365.3 on 2 and 341 DF,  p-value: < 2.2e-16
```

We can see from the summary of our model that the estimated model is

$$Y_i = a + b_{uk\_born} X_{uk\_born,i} + b_{no\_edu} X_{no\_edu,i} + \epsilon_i,$$

where $b_{uk\_born} = -.33$ and $b_{no\_edu} = -1.20$.

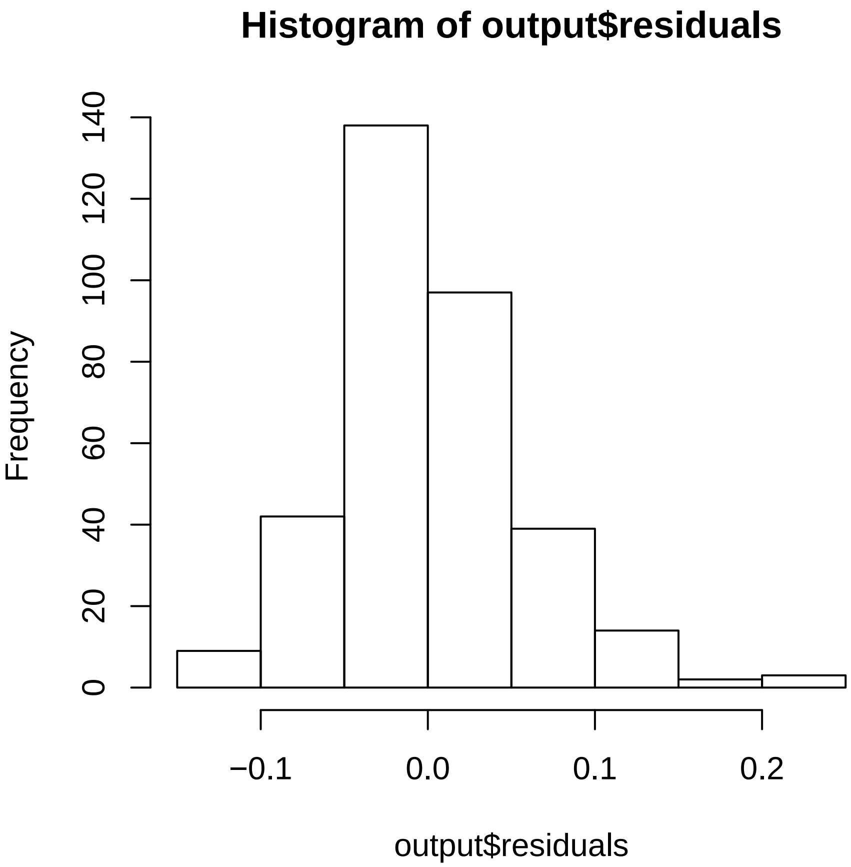
Before, we interpreted the model as “a change in X is associated with a b unit change in Y.” Now that we have multiple explanatory variables, we can make a slightly more sophisticated statement- ”When fixing all other explanatory variables, a change in X_1 is associated with a b_1 unit change in Y. So the way to think about this is to imagine two local authorities which are the same in all other explanatory variables, except one local authority has an X_1 value which is larger than the other local authority by 1 unit. Then we would expect the Y value to be larger by b_1 units. This allows us to start decomposing the association between the response variable and the explanatory variables into different components and measure the relative impact (via the size of the coefficient) of each variable.

Questions

- How would you interpret each of the estimated coefficients above?
- Does the magnitude (size) of the coefficients agree with what you would’ve guessed?

We can get the residuals from the fitted model using the $residuals command. When plotting a histogram of the residuals, we can see that there might be an outlier on the top end of the range.

```
# plot a histogram of the residuals
hist(output$residuals)
```



```
# find which residual has the largest absolute value
which.max(abs(output$residuals))
```

```
## 48
## 48
```

```
# look at which county has a large residuals
brexit.data[48,]
```

```
##    geography   uk_born    no_edu       mfct    finance   over_60
## 48 Liverpool 0.9011074 0.2872204 0.06270661 0.03650511 0.1822454
##    over_20_less_than35 pct_remain
## 48           0.2766485     0.5819
```

We can see that for a local authority with its percentage of UK born residents and percentage of individuals with no non-compulsory education, Liverpool had a surprisingly large percentage of people who voted to remain in the EU. While there might be various reasons for this, it could be because Liverpool has a large tourism industry, or is home to many international shipping lines.

2 Fit your own regressions

Now is your chance to explore the data yourself. Using the form above, fit a regression and include variables which you think might be associated with the percentage of people voting to remain in the EU. As you fit your models, check to make sure that the associations are roughly linear, and take a log transformation if necessary. Look back to lab 2 to remember how to take transformations.

If you find an outlier, check out what county that corresponds to and dig around on google to find out why that local authority might be an outlier.

Try fitting multiple models (at least 3 or 4) and think about what makes sense to investigate and what variables might need transformations.

Questions

- Look at the r^2 value for each model. As you include more variables, what happens to the r^2 value?
- When you include more variables, how do the regression coefficients change for the existing variables?
- Why do you think this happens?

After you are done, discuss your findings with your neighbor and pat yourself on the back. Congratulations, you're on your way to being a statistician!

Questions

Questions to discuss with your neighbor.

- How did you decide which variables to include and which variables not to include?
- What is the proper interpretation of your regression coefficients?
- What are the signs of each of the coefficients?
- What are the relative sizes of the coefficients?
- Does this make sense with what we know about the world?
- What would we need to be careful about in interpreting these models?
- What other variables (that weren't available) would also be good to include?