ECOM20001: Econometrics 1

Tutorial 6: Suggested Solutions

Earnings and Height

1. Regression Results

$$\widehat{Earnings}_i = \underset{(0.338)}{-0.051} + \underset{(0.002)}{0.028} Height_i, \quad R^2 = 0.011, SER = 2.678$$

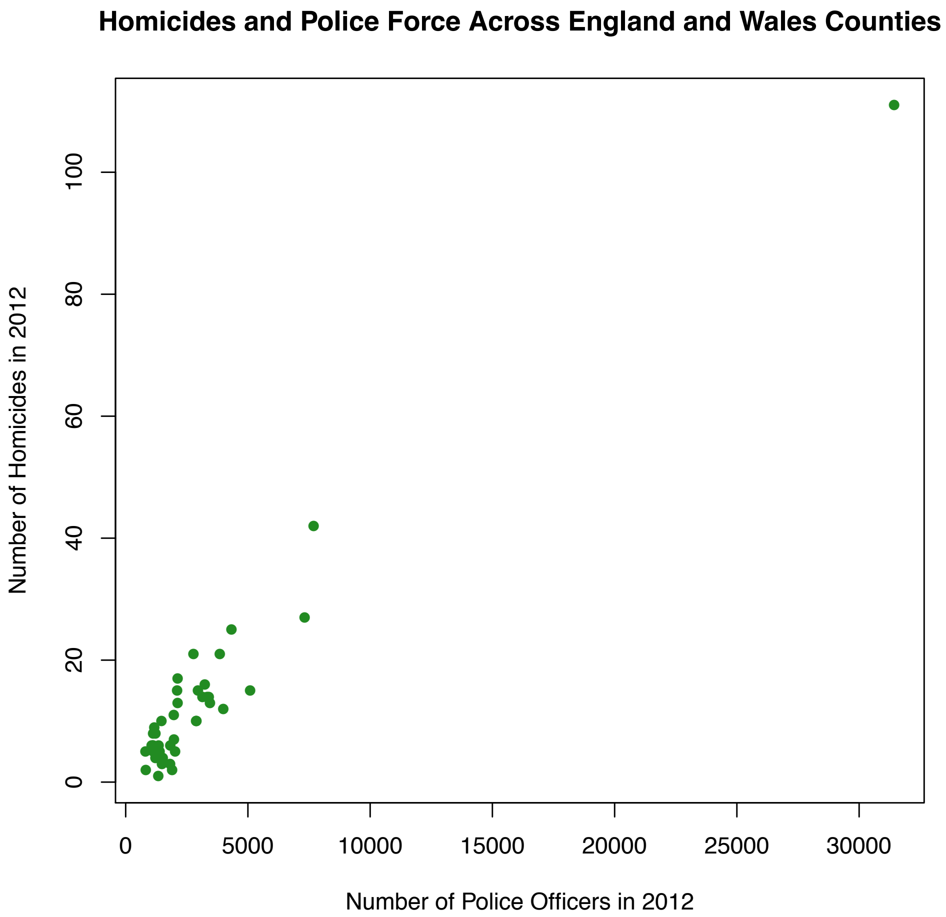
Increasing height by 1cm has a corresponding increase in annual earnings of \$1 x \$10000 x 0.028 = \$280. The 95% confidence interval (CI) is [0.024,0.032] so we fail to reject estimated relationships as small as \$240 annually and as large as \$320 annually at the 5% level of significance.

2. Increasing height by 100 cm yields a 0.028 x 100 cm x \$10000 = \$28000 increase in annual earnings. The 95% CI in terms of annual earnings is [\$24000, \$32000].

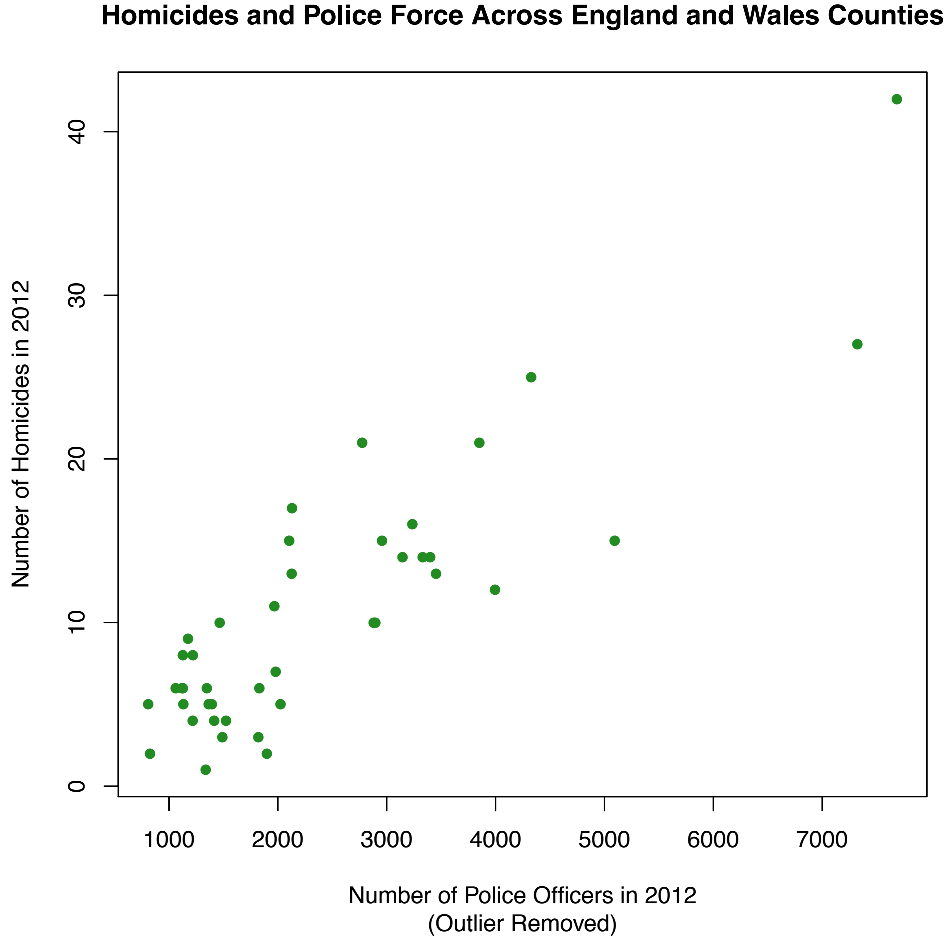
3. t-statistic for testing a null equal to \$3000 increase in annual earnings for a 10cm increase in height the same as a 1cm increase in height having a \$300 increase in earnings. Hence the t-statistic for the test is t=(0.028-0.03) / 0.002 = -1.08 which has an associated p-value of 2 x Phi(-|-1.08|) = 0.280, where Phi( ) is the CDF of the N(0,1) distribution. So fail to reject the null of a \$3000 increase in annual earnings from a 10 cm increase in height at the 5% level of significance.

<u>Homicides and Police</u>

1. Raw scatter plot below, positive relationship. Somewhat surprising (to me at least) because I would have initially thought more police means less crime/ homicides.



2. 'Metropolitan Police' is the data point in the top right corner of the graph with more than 100 homicides a year, which is a potential outlier. It's a potential outlier because it corresponds to London, which is substantially larger than all other markets in the UK. Using the scatter plot, we examine whether there are any potential outliers. Second scatter plot with this potential outlier removed shows a similar positive relationship, but the scale of the axes are no longer stretched out by the potential outlier.



3. Estimation results

$$\widehat{Homicides}_i = \underset{(0.7718)}{1.9963} + \underset{(0.0001)}{0.0036} Police_i, \quad R^2 = 0.942, SER = 4.221$$

with p-value on the police coefficient less than 0.000001, so reject the null that there is no relationship between homicides and police. 95% CI for the coefficient on police is [0.0033, 0.0038]. Interpreting the results, we find that adding one more police officer (e.g. a one unit change) is associated with a 0.0036 increase in homicides in a county.

It is a bit difficult to interpret this result since the scale of the police variable is in the 1000s so adding just one police officer seems like a minuscule change. The other issue is we have had to keep four digits past the decimal in our regression results to maintain some degree of accuracy in the results. This is not preferred because there's too many digits relative to what's needed for interpreting the magnitudes of the effects of interest. It would be preferred to have two or three digits past the decimal and no more. We need to rescale the variable to fix this issue.

4. Results after rescaling the police variable:

$$\widehat{Homicides}_i = \underset{(0.772)}{1.996} + \underset{(0.139)}{3.566}\,Police_i, \quad R^2 = 0.942, SER = 4.221$$

with p-value on the police coefficient unchanged at less than 0.000001, so reject the null that there is no relationship between homicides and police. Interpretation now is an increase of police by 1000 has an associated increase of 3.566 homicides. 95% CI is [3.286,3.847]. We no longer have to stretch the significant digits for accuracy; in fact the re-scaled results are more accurate than the unscaled ones. Much cleaner and more relevant/clearer interpretation of results given the magnitude of the police variable in the raw data.

In passing, here are the results keeping only two digits after the decimal. This is even cleaner in terms of presentation and interpretation without giving up anything in terms of relevant level of accuracy:

$$\widehat{Homicides}_i = \underset{(0.77)}{2.00} + \underset{(0.14)}{3.57}\,Police_i, \quad R^2 = 0.94, SER = 4.22$$

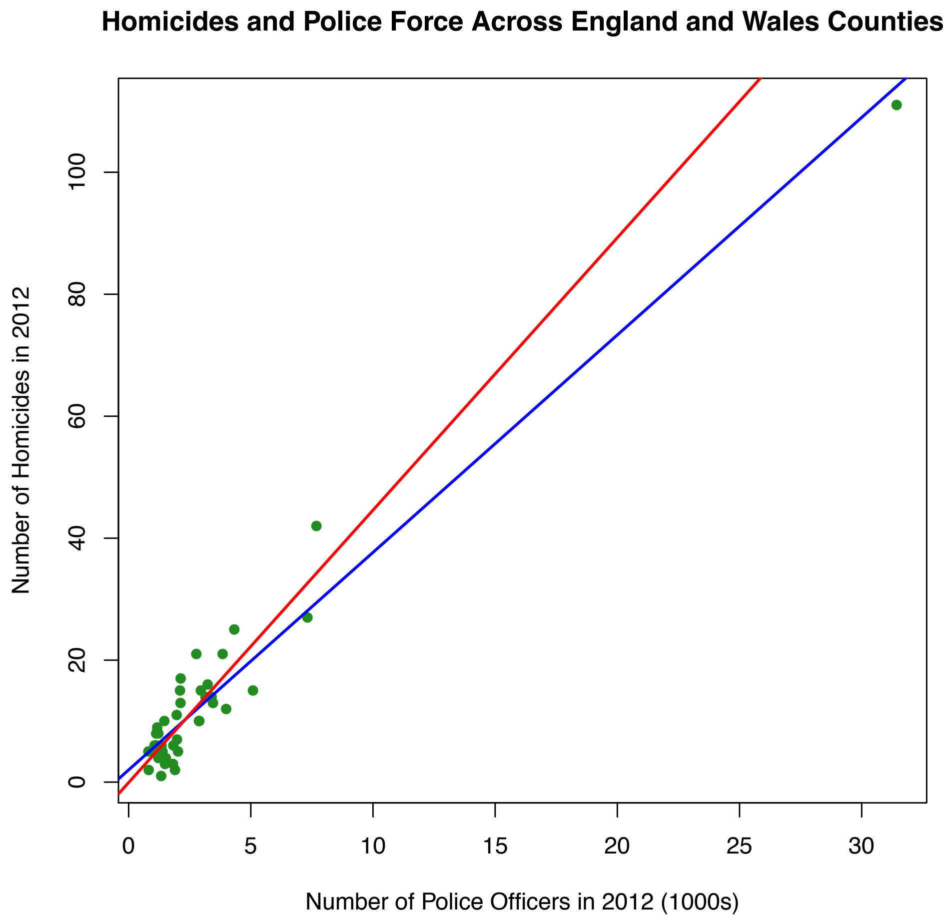
5. Results with rescaled police variable and with removing outlier:

$$\widehat{Homicides}_i = \underset{(1.14)}{-0.08} + \underset{(0.40)}{4.48}\,Police_i, \quad R^2 = 0.76, SER = 4.00$$

Here, there is a large increase in the coefficient on police. It increases from 3.57 to 4.48, which is a 100 x (4.48-3.57) / 3.57 = 25% increase. The coefficient has a p-value less than 0.000001 so is still statistically significant at conventional levels of significance such as the 1% level. Interpretation now is an increase of police by 1000 has an associated increase of 4.48 homicides. 95% CI is [3.66,5.28]. The large change in the slope coefficient estimate strongly suggests that the outlier should be removed.

Notice in passing how the standard error on the policy coefficient rises substantially, from 0.14 to 0.40, which is a near tripling of the standard error magnitude. Why does this occur? Because when we drop the outlier, we get a substantial decrease in the variance of our independent variable, which recall from the class notes is inversely related to the standard error of the OLS regression slope estimate.

Scatter plot which highlights the differences in the results with and without the outlier using the rescaled police variable. Clearly shows how the outlier “drags down” the OLS slope coefficient estimate.



6. Answers

- I found the scatter plot surprising originally because I focused on the impact of police presence on homicides, which I thought would be negative not positive.
- But we could find a *positive* relationship if the government actively puts more police officers in areas that tend to have higher crime rates. That is, all else equal, higher crime areas would attract more police officers if governments

actively targeted the police force to be in high violence/homicide areas to maximise the public benefit from having police officers around.

- Alternatively, it could be that putting more police officers in an area reduces the homicides since having more police around increases the chances of getting caught and prosecuted as a suspect, and hence increases the cost of engaging in crime, or in the extreme, homicide.
- The OLS estimates cannot be interpreted as causal since you cannot disentangle the "government targeting" and "homicide reducing" influences on the correlation between homicides and police officers. There may be even more explanations that we cannot disentangle out, but having just two is enough to undermine a causal interpretation of our OLS estimates in questions 4. and 5. That is, the OLS estimates reflect a mix of the "government targeting" and "homicide reducing" influences on the correlation between homicides and police officers which work in opposite directions. The fact that the correlation is positive suggests that the "government targeting" force dominates the "homicide reducing", but there is no way to figure out how large these two forces are empirically (yet!).
- An experiment would see you randomly put police officers in some counties and randomly withhold police officers in other counties, and then track the relative homicide rates over time. We would expect at least a negative relationship if having police officers around were a good deterrent for homicides. So we would expect the OLS estimate from a regression using the experimental data relating homicides and police offer counts to be negative.