ECOM20001 Econometrics 1

Tutorial 6 Semester 1, 2022

Chin Quek

Department of Economics

Introduction

- Linear Regression - Hypothesis Testing

File downloaded from Canvas

- `tute6.R`
- `tute6_height.csv`
- `tute6_crime.csv`

Dataset 1: Earnings and Height

The first (micro) dataset, `tute6_height.csv`, has the following 5 variables:

- `id`: worker identifier
- `earnings`: annual labour earnings in $10,000's (in real terms, 2012=100)
- `height`: height without shoes in centimetres
- `weight`: weight without shoes in kilograms
- `male`: binary variable that equals 1 if worker is male and 0 otherwise
- `age`: age of the worker at time of survey

In total, the dataset contains this information for n = 17,870 U.S. workers.

```
# Loading the dataset
mydata1=read.csv("tute6_height.csv")
```

```
head(mydata1)
```

```
##   ï..id earnings height weight male age
## 1     1 8.405475    165     60    0  48
## 2     2 1.402139    165     70    0  41
## 3     3 8.405475    152     49    0  26
## 4     4 8.405475    170     68    0  37
## 5     5 2.856039    173     82    0  35
## 6     6 2.336287    160     46    0  25
```

Question 1

Estimate the following single linear regression model for worker$_i$:

$$\text{Earnings}_i = \beta_0 + \beta_1 \, \text{Height}_i + u_i$$

Present the regression results, including discussion of statistical significance of the OLS estimate against a null of no relationship between earnings and height, for the change in earnings associated with a one-unit change in height.

Also present the 95% confidence interval for the relationship between earnings and a one-unit change in height.

```r
earn_reg1=lm(earnings~height,data=mydata1)

summary(earn_reg1)
```

```
##
## Call:
## lm(formula = earnings ~ height, data = mydata1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.7972 -2.1909 -0.7923  3.4421  5.0579
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.051174   0.338050  -0.151     0.88
## height       0.027859   0.001984  14.042   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.678 on 17868 degrees of freedom
## Multiple R-squared:  0.01092,    Adjusted R-squared:  0.01086
## F-statistic: 197.2 on 1 and 17868 DF,  p-value: < 2.2e-16
```

```
stargazer(earn_reg1,
          type="text",
          title = "Q1. Earnings vs. Height")
```

```
##
## Q1. Earnings vs. Height
## ================================================
##                          Dependent variable:
##                      ---------------------------
##                                earnings
## ------------------------------------------------
## height                         0.028***
##                                 (0.002)
##
## Constant                        -0.051
##                                 (0.338)
##
## ------------------------------------------------
## Observations                    17,870
## R2                               0.011
## Adjusted R2                      0.011
## Residual Std. Error      2.678 (df = 17868)
## F Statistic          197.190*** (df = 1; 17868)
## ================================================
## Note:                *p<0.1; **p<0.05; ***p<0.01
```

```r
# Obtain regression coefficients from earn_reg1 regression
beta=coef(summary(earn_reg1))[, "Estimate"]

## Obtain standard errors from coefficients earn_reg1 regression
se=coef(summary(earn_reg1))[, "Std. Error"]

## Compute 95% CI of the regression slope coefficient by hand
CI95_low=beta[2]-1.96*se[2]     # lower bound of 95% CI
CI95_upp=beta[2]+1.96*se[2]     # upper bound of 95% CI
```

or by using the *confit* function in R

```r
confint(earn_reg1, 'height', level=0.95)
```

```
##                2.5 %     97.5 %
## height 0.02397003 0.03174726
```

Question 2

Discuss the statistical significance of the OLS regression estimate against a null of no relationship between earnings and height, for the change in earnings associated with a 100 cm change in height.

Also present the 95% confidence interval for the relationship between earnings and a 100 cm change in height.

```r
CI95_low_100=100*(beta[2]-1.96*se[2])    # lower bound of 95% CI
CI95_upp_100=100*(beta[2]+1.96*se[2])    # upper bound of 95% CI
paste("95% CI lower bound for 100cm increase in earnings is:",
      CI95_low_100)
```

```
## [1] "95% CI lower bound for 100cm increase in earnings is: 2.39702224050
```

```r
paste("95% CI upper bound for 100cm increase in earnings is:",
      CI95_upp_100)
```

```
## [1] "95% CI upper bound for 100cm increase in earnings is: 3.17470701036
```

Additional workings

Question 3

Present the regression results, including discussion of statistical significance of the OLS regression estimate against a null that a 10 cm increase in height has an associated \$3,000 increase in annual earnings.

There is no R code provided for this question.

Please write our own code to answer this question (hint: look at the code used for Question 2).

. . . discussion of statistical significance of the OLS regression estimate against a null that a 10 cm increase in height has an associated $3,000 increase in annual earnings.

```
##                Estimate Std. Error    t value     Pr(>|t|)
## (Intercept) -0.05117363 0.33805005 -0.1513789 8.796786e-01
## height       0.02785865 0.00198389 14.0424369 1.479037e-44
```

```
beta
```

```
## (Intercept)      height
## -0.05117363  0.02785865
```

```
se
```

```
## (Intercept)      height
##  0.33805005  0.00198389
```

```r
## t-statistic and p-value for null that slope=0.03
tstat2=(beta[2]-0.03)/se[2]
pval2=2*pnorm(-abs(tstat2))
paste("pvalue for 2-sided test of null that slope=0.03 is:", pval2)
```

```
## [1] "pvalue for 2-sided test of null that slope=0.03 is: 0.2804222126497
```

Additional workings

$$10\text{cm} \uparrow \text{in height} \implies \$3000 \uparrow \text{annual earnings}$$

$$\text{earnings} = \beta_0 + \beta_1 \, \text{height} + u_i$$

Dataset 2: Police and Homicide

The second (county-level) dataset from England and Wales, `tute6_crime.csv`, has the following 3 variables:

- `county`: county name
- `police`: number of police officers in 2012
- `homicide`: number of homicides in 2012

In total, the dataset contains this information for n $= 43$ counties.

```
# Loading the dataset
mydata2=read.csv(file="tute6_crime.csv")

head(mydata2)
```

```
##              county police homicides
## 1 Avon and Somerset   2957        15
## 2      Bedfordshire   1128         8
## 3    Cambridgeshire   1348         6
## 4          Cheshire   2025         5
## 5         Cleveland   1490         3
## 6           Cumbria   1128         6
```

Question 4

How would you describe a typical county?

Present a scatter plot with police on the horizontal axis and homicides on the vertical axis. What relationship do you see? Do you think this relationship is surprising?

Using the scatter plot above, examine whether there are any potential outliers.

- If there are, discuss which data point(s) appear to be outliers and why they may be outliers.
- Identify which counties, if any, are outlier(s).

If you consider a county to be an outlier, then remove that observation and then run a regression of Homicides = f(Police) with and without the outlier(s) you have identified.

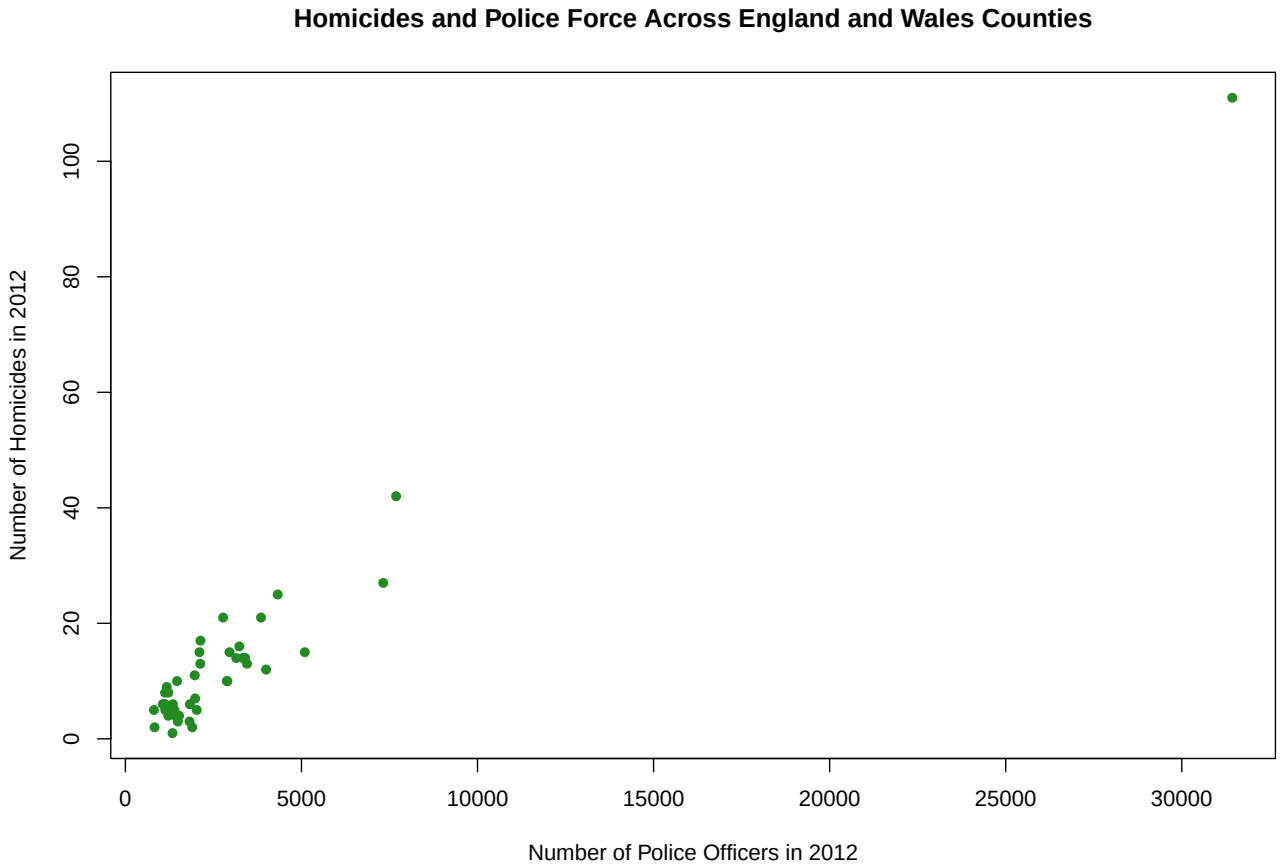
Present the regression results in one table using `stargazer`. Then interpret the results of these regression(s).

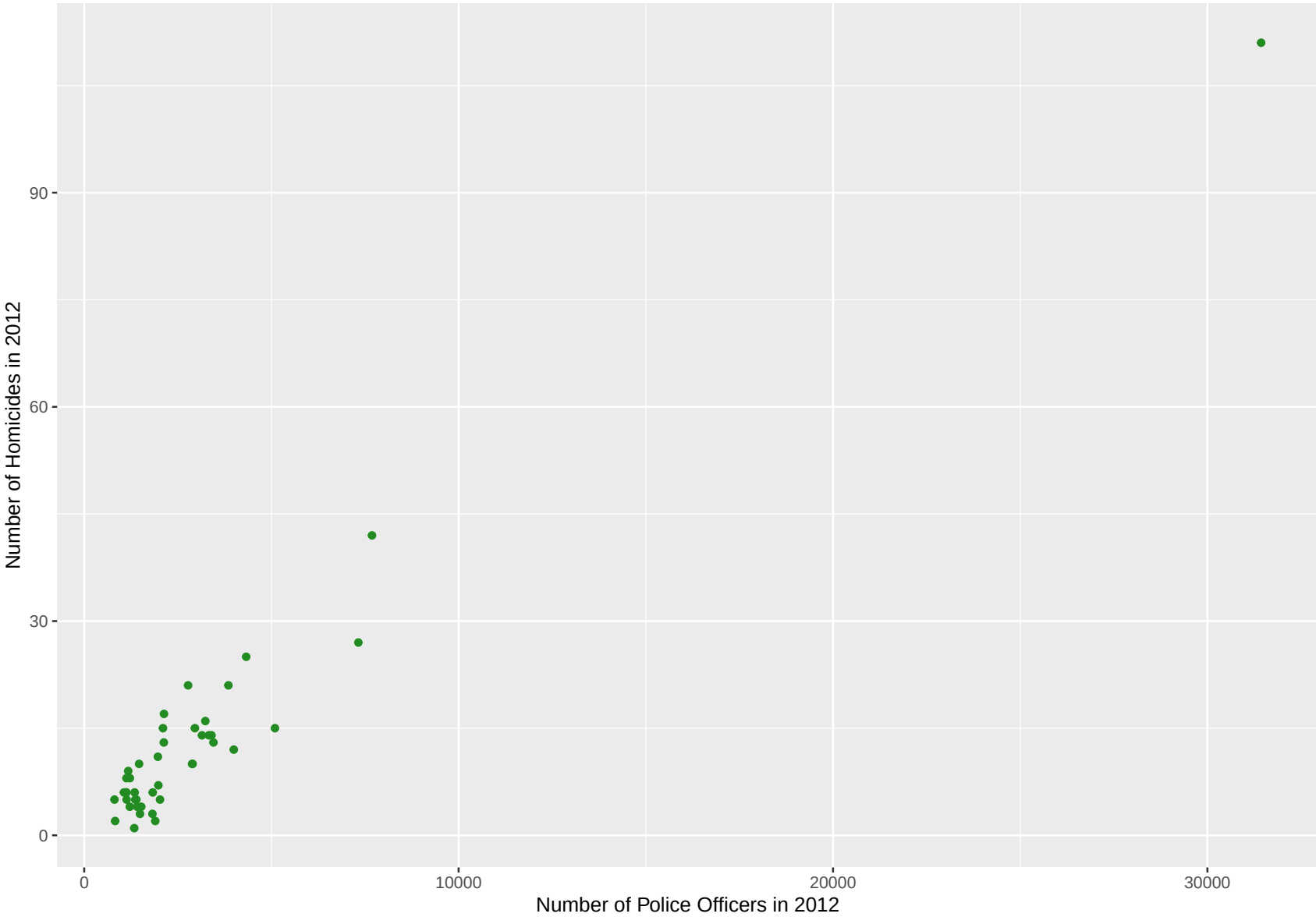
Produce a second scatter plot with the outliers removed.

```
summary(mydata2)
```

```
##     county              police        homicides
##  Length:43          Min.   :  809   Min.   :  1.00
##  Class :character   1st Qu.: 1342   1st Qu.:  5.00
##  Mode  :character   Median : 1968   Median :  9.00
##                     Mean   : 3066   Mean   : 12.93
##                     3rd Qu.: 3191   3rd Qu.: 14.50
##                     Max.   :31435   Max.   :111.00
```

```
plot(mydata2$police, mydata2$homicides,
     main="Homicides and Police Force Across England and Wales Counties",
     xlab="Number of Police Officers in 2012",
     ylab="Number of Homicides in 2012",
     col="forestgreen", pch=16)
```




Number of Homicides in 2012
90
60
30
0
0
10000
20000
30000
Number of Police Officers in 2012

Regression using all observations

```
crime_reg1=lm(homicides~police, data=mydata2)
summary(crime_reg1)
```

```
##
## Call:
## lm(formula = homicides ~ police, data = mydata2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.7686 -2.7359 -0.8036  2.2218 12.5901
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 1.9963839  0.7718408   2.587   0.0133 *
## police      0.0035662  0.0001389  25.678   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.221 on 41 degrees of freedom
## Multiple R-squared:  0.9415, Adjusted R-squared:   0.94
## F-statistic: 659.4 on 1 and 41 DF,  p-value: < 2.2e-16
```

Regression excluding Metropolitan

```
exGLA <- subset(mydata2, county !="Metropolitan Police")

crime_reg2=lm(homicides~police, data = exGLA)
summary(crime_reg2)
```

```
##
## Call:
## lm(formula = homicides ~ police, data = exGLA)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -7.677 -2.629 -0.385  2.548  8.687
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.083628   1.138030  -0.073    0.942
## police       0.004467   0.000400  11.168 7.27e-14 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.999 on 40 degrees of freedom
## Multiple R-squared:  0.7572, Adjusted R-squared:  0.7511
## F-statistic: 124.7 on 1 and 40 DF,  p-value: 7.272e-14
```

Confidence intervals

```
confint(crime_reg1, 'police', level=0.95)
```

```
##                2.5 %      97.5 %
## police 0.003285738 0.003846692
```

```
confint(crime_reg2, 'police', level=0.95)
```

```
##                2.5 %      97.5 %
## police 0.003658761 0.005275659
```

Using stargazer

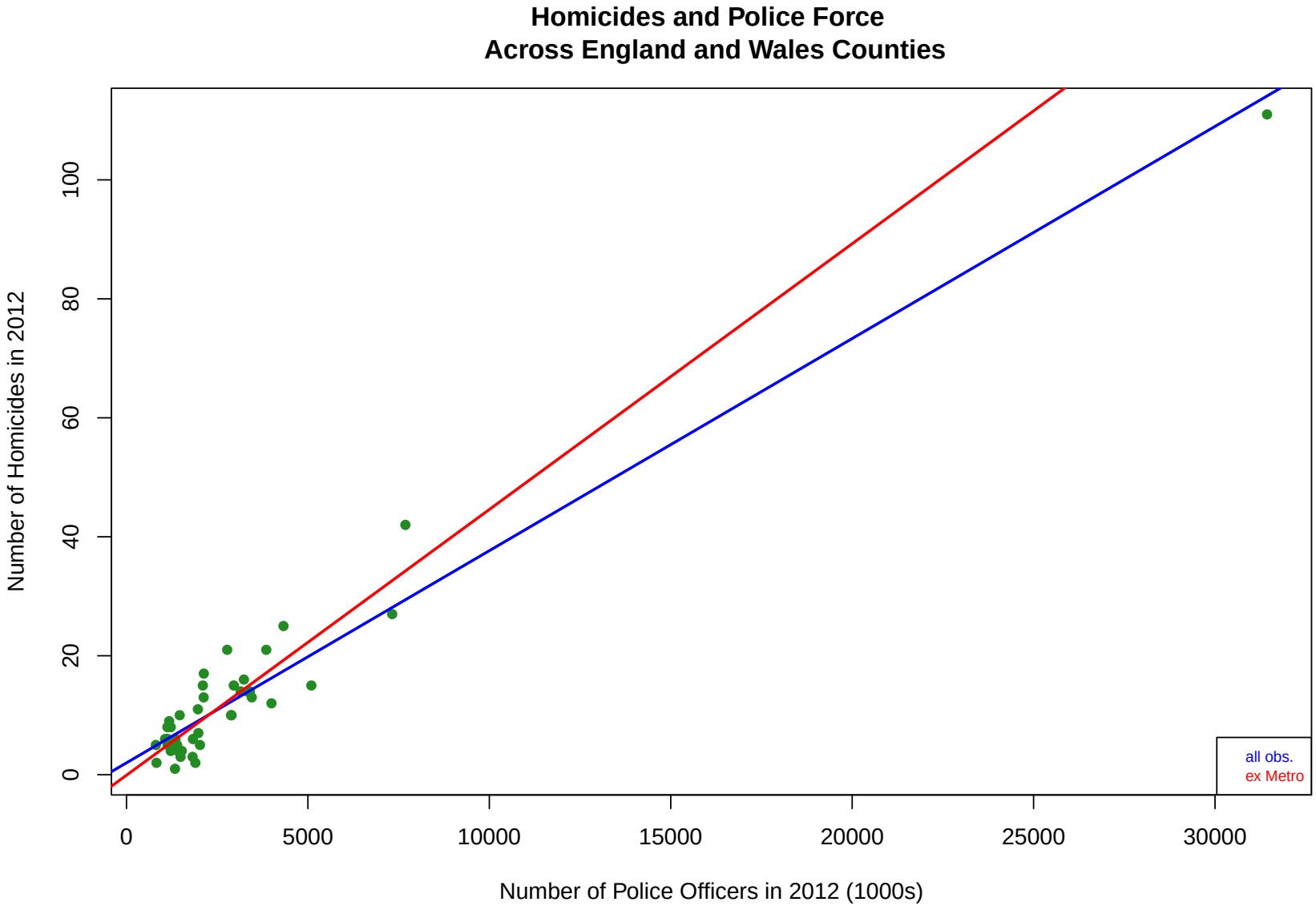
```
stargazer(crime_reg1, crime_reg2,
          title= "Homicides and Police: England and Wales Counties",
          type = "text",
          column.labels=c("all obs.","excl. Metropolitan"),
          digits=6,
          intercept.bottom = FALSE)
```

Regression results

```
##
## Homicides and Police: England and Wales Counties
## ================================================================
##                                    Dependent variable:
##                          ------------------------------------------
##                                          homicides
##                                all obs.              excl. Metropolitan
##                                  (1)                       (2)
## ----------------------------------------------------------------
## Constant                       1.996384**                -0.083628
##                                (0.771841)                (1.138031)
##
## police                        0.003566***               0.004467***
##                                (0.000139)                (0.000400)
##
## ----------------------------------------------------------------
## Observations                      43                        42
## R2                              0.941459                  0.757162
## Adjusted R2                     0.940031                  0.751091
## Residual Std. Error       4.221425 (df = 41)        3.998949 (df = 40)
## F Statistic          659.364100*** (df = 1; 41) 124.719000*** (df = 1; 40)
## ================================================================
## Note:                                        *p<0.1; **p<0.05; ***p<0.01
```

Excluding Metropolitan

To see the effect of this outlier graphically, use a scatter plot:



```
plot(mydata2$police, mydata2$homicides,
     main="Homicides and Police Force \n Across England and Wales Counties"
     xlab="Number of Police Officers in 2012 (1000s)",
     ylab="Number of Homicides in 2012",
     col="forestgreen",
     pch=16)

abline(crime_reg1, col="blue", lwd=2)
abline(crime_reg2, col="red", lwd=2)

legend ("bottomright",
        c("all obs.","ex Metro"),
        text.col=c("blue","red"),
        cex = 0.7)
```

Question 5

$$\text{Homicides}_i = \beta_0 + \beta_1 \, \text{Police}_i + u_i$$

You will notice that the parameter estimate(s) for the police variable is quite small in magnitude.

We could **rescale** the `Police` variable from the raw data to being in terms of 1000s of police in a county.

Using this rescaled variable `police_1000`, estimate and then report a new set of regression results for the model you prefer from Question 4.

Interpret the results including discussion of statistical significance of the OLS regression estimate against a null of no relationship between homicides and police, for the change in homicides associated with a one-unit change in the rescaled police variable.

There is no R code provided for this question, so please write, then run your own code.

Scaling the police variable in mydata2

```
mydata2$police_1000=mydata2$police/1000
summary(mydata2)
```

```
##     county              police       homicides       police_1000
##  Length:43          Min.   :  809   Min.   :  1.00   Min.   : 0.809
##  Class :character   1st Qu.: 1342   1st Qu.:  5.00   1st Qu.: 1.343
##  Mode  :character   Median : 1968   Median :  9.00   Median : 1.968
##                     Mean   : 3066   Mean   : 12.93   Mean   : 3.066
##                     3rd Qu.: 3191   3rd Qu.: 14.50   3rd Qu.: 3.191
##                     Max.   :31435   Max.   :111.00   Max.   :31.435
```

```
crime_reg3=lm(homicides~police_1000,
              data=mydata2[mydata2$county!="Metropolitan Police",])

stargazer(crime_reg2, crime_reg3, intercept.bottom = FALSE,
          title= "Homicides and Police: England and Wales Counties",
          type = "text", column.labels=c("Unscaled","Rescaled"))
```

```
##
## Homicides and Police: England and Wales Counties
## ================================================================
##                                      Dependent variable:
##                                  ---------------------------
##                                            homicides
##                                    Unscaled      Rescaled
##                                      (1)            (2)
## ----------------------------------------------------------------
## Constant                            -0.084        -0.084
##                                    (1.138)        (1.138)
##
## police                             0.004***
##                                    (0.0004)
##
## police_1000                                       4.467***
##                                                    (0.400)
##
## ----------------------------------------------------------------
## Observations                          42             42
## R2                                  0.757          0.757
## Adjusted R2                         0.751          0.751
## Residual Std. Error (df = 40)       3.999          3.999
## F Statistic (df = 1; 40)          124.719***     124.719***
## ================================================================
## Note:                          *p<0.1; **p<0.05; ***p<0.01
```

Confidence interval of slope parameter (excluding Metropolitan)

```
confint(crime_reg2, 'police', level=0.95)
```

```
##                2.5 %      97.5 %
## police 0.003658761 0.005275659
```

```
confint(crime_reg3, 'police_1000', level=0.95)
```

```
##                  2.5 %   97.5 %
## police_1000 3.658761 5.275659
```

Question 6

Go back to the scatter plot in Question 4.

Provide an economic explanation for why you might find a positive relationship between the number of homicides and the number of police.

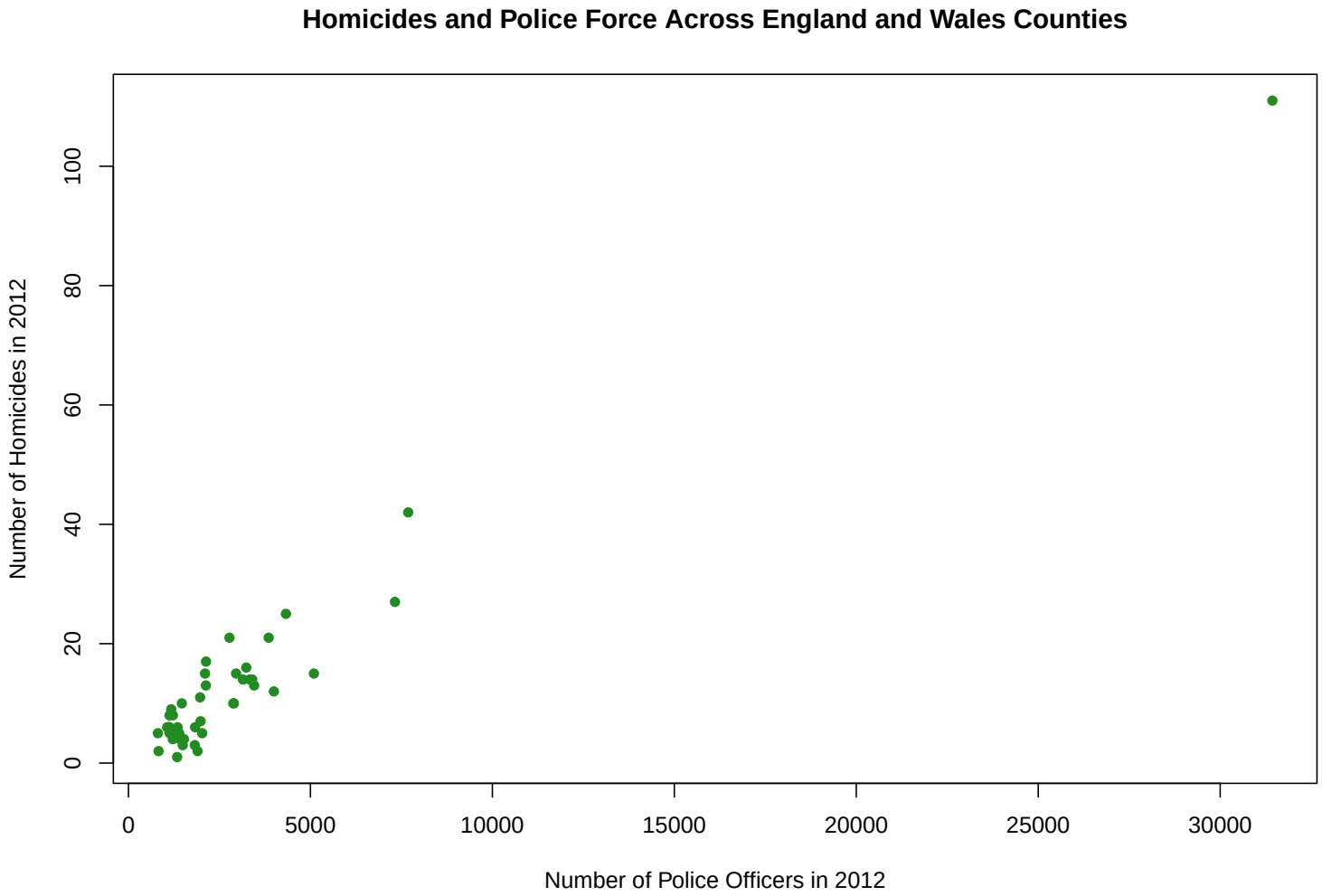
Provide a separate economic explanation for why you might find a negative relationship between the number of homicides and the number of police.

Given your economic explanations for a positive and negative relationship, can you plausibly interpret your OLS estimates of the relationship between homicides and the number of police in questions 4. and 5. above as being "causal"? That is, can you interpret the estimated relationship as the causal impact of increasing the number of police on the number of homicides?

If you think the OLS estimate of the relationship is causal, explain why.

If you do not think the OLS estimate of the relationship is causal, briefly describe an experiment that could be used to estimate the causal impact of increasing the number of police on the crime rate in a county.

What would you expect the sign of the OLS estimate of the empirical relationship between the number of homicides and number of police with such experimental data to be?


Homicides and Police Force Across England and Wales Counties
Number of Homicides in 2012
0
20
40
60
80
100
0
5000
10000
15000
20000
25000
30000
Number of Police Officers in 2012

```
plot(mydata2$police,mydata2$homicides,
     main="Homicides and Police Force Across England and Wales Counties",
     xlab="Number of Police Officers in 2012",
     ylab="Number of Homicides in 2012",
     col="forestgreen", pch=16)
```

Positive and negative relationships between number of homicides and number of police

Positive and negative relationships between number of homicides and number of police

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.

Positive and negative relationships between number of homicides and number of police

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.

Negative relationship: 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.

Causal relationship?

Can the OLS estimates be interpreted as causal?

Causal relationship?

Can the OLS estimates be interpreted as causal?

No. We cannot disentangle the "government targeting" and "homicide reducing" influences on the correlation between homicides and police officers, which work in **opposite directions**.

- There could be more explanations that we cannot disentangle out
- 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!).

Conducting an experiment

How would you conduct an experiment to estimate the *causal impact* of increasing number of police on the crime rate?

- Randomly put police officers in some counties and randomly withhold police officers in other counties.
- Then track the relative homicide rates over time.

What do you expect the sign of the OLS estimate of the empirical relationship between number of homicides and number of police of such experimental data to be?

Conducting an experiment

How would you conduct an experiment to estimate the *causal impact* of increasing number of police on the crime rate?

- Randomly put police officers in some counties and randomly withhold police officers in other counties.
- Then track the relative homicide rates over time.

What do you expect the sign of the OLS estimate of the empirical relationship between number of homicides and number of police of such experimental data to be?

- 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.