Lecture 5 - EEP 118 Spring 2025¶

This is the notebook for Lecture 5 where we learn about the statistical properties of the estimators from our linear models.

We will see how the estimated coefficients, the standard errors of the estimated coefficients, and the R Squared change when the sample size in the data increases when we run a linear least squares regression estimation procedure. We do this in Lecture 5 using R.

To run, hit the `i>|Run` button on top middle bar and keep hitting and it will run line by line,

OR

To run a line that starts with In [ ]: highlight the content and hit SHIFT ENTER at same time

Let the unknown population model be

$crmrte=\beta_0+ \beta_1 polpc +u$

We will use a sample to estimate $\hat{\beta_0}$ and $\hat{\beta_1}$

As shown in lecture, $\hat{\beta_0}$ and $\hat{\beta_1}$ are random variables.

Given the estimated parameters, then the predicted crime rate, crmrte HAT, is equal to: $\hat{crmrte_i}= \hat{\beta_0}+\hat{\beta_1} polpc_i $

Where $\hat{\beta_1}$ is the marginal effect of police per capita on predicted crime rate, namely $\hat{crmrte}$.

β_0 and β_1 are true unknown values from the population regression

$\hat{\beta_0}$ and $\hat{\beta_1}$ are estimators, (formulas) to compute an estimate (a value) with a sample

If we use a different sample we get different values of $\hat{\beta_0}$ and $\hat{\beta_1}$

If we repeat for many samples we get a distribution of $\hat{\beta_0}$ and $\hat{\beta_1}$

If certain assumptions hold the distribution of $\hat{\beta_0}$ and $\hat{\beta_1}$ will be related to β_0 and β_1

In [ ]:
```
#load needed packages
library(tidyverse)
```

Data Sample N=630

Open the data set that we obtained by drawing a sample of size N=630 from the population of US counties, and estimate the linear model $crmrte=\beta_0+ \beta_1 polpc +u$ by minimizing the sum of squared residuals to get $\hat{\beta_0}$ and $\hat{\beta_1}$

In [9]:
```
#------------------------------------------
#1. Read in data and see the top rows to see column names etc
#------------------------------------------
my_data <- read.csv("Lecture5.csv")
head(my_data)
```

A data.frame: 6 × 4

	county	year	crmrte	polpc
	<int>	<int>	<dbl>	<dbl>
1	1	81	0.0398849	0.0017868
2	1	82	0.0383449	0.0017666
3	1	83	0.0303048	0.0018358
4	1	84	0.0347259	0.0018859
5	1	85	0.0365730	0.0019244
6	1	86	0.0347524	0.0018952

In [10]:
```
#regression
regLectureN630 <- lm(crmrte ~ polpc, my_data)
#show output
summary(regLectureN630)
```

```
Call:
lm(formula = crmrte ~ polpc, data = my_data)

Residuals:
      Min        1Q    Median        3Q       Max 
-0.045791 -0.012476 -0.002669  0.007157  0.098927 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 0.0292403  0.0008673  33.713  < 2e-16 ***
polpc       1.2246077  0.2598397   4.713 3.01e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.01782 on 628 degrees of freedom
Multiple R-squared:  0.03416,   Adjusted R-squared:  0.03262 
F-statistic: 22.21 on 1 and 628 DF,  p-value: 3.009e-06
```

you get $\hat{\beta_0}=0.0292$ and $\hat{\beta_1}=1.2246$.

Given the above estimation, the predicted model is

$\hat{crmrte_i}=0.0292+1224 polpc_i$

I drew two additional samples of the same N=630 and got the following for the second sample

$\hat{crmrte_i}=0.03+1.776 polpc_i$

and for the third sample of N=630

$\hat{crmrte_i}=0.03 +2.02 polpc_i$

The estimated parameters change across samples (like you see in problem set 1).

If we were to average all three intercept estimated coefficients, for example, you would find on average an estimate that has as expected value the TRUE population parameter beta for the intercept.

And the same for the slope,

because E($\hat{\beta}$)=β if we make four assumptions.

Statistical Properties of Estimator $\hat{\beta}$

Let the model, general y and x notation, be given by

$y=\beta_0+ \beta_1 x +u$

1. $\hat{\beta}=(\hat{\beta_0},\hat{\beta_1})$ are random variables
2. $\hat{\beta}$ are unbiased, both the intercept and the slope:

E($\hat{\beta_0})=\beta_0$,

E($\hat{\beta_1})=\beta_1$

if SLR1+SLR2+SLR3+SLR4 assumptions hold.

Where SLR is short for simple Linear Regression (SLR)

What does each SLR assumption mean?

SLR1, Y linear in parameters , that is $ y=\beta_0+ \beta_1 x +*u*$

SLR2, { (x_i,y_i), i=1,...n} random sample in the population, then we can write for each observation i the following $ y_i=\beta_0+ \beta_1 x_i +*u*_i$

SLR3 There is variation in x in the sample (the sample variance of x cannot be zero), that is x needs to be varrying in the sample.

SLR4 $E(u|x)=0$, that is there is zero conditional mean of the disturbance u and *for the random sample* $E(u_i|x_i)=0$, i=1,2, ..., n

4. Repeating the same random sampling of N=630 observations gives different estimates, but if you were to average them up, you would find an unbiased estimator for the population parameter, as we will show next under four assumptions. We will show that the expected value of the estimator is the true parameter, E($\hat{\beta})=\beta$. We do not have a bias.

We will show that our estimator is unbiased for the true parameter of the population

No description has been provided for this image

Proof of Unbiasedness

One can show that E($\hat{\beta_1})=\beta_1$ if all four assumptions hold.

Proof in book chapter 2, page 54 in 3rd edition Below is an illustration , if you take a more theoretical class we would go over it in great detail...

No description has been provided for this image
No description has been provided for this image
No description has been provided for this image

Now lets compare what happens when we change the sample size N

Let's reduce the sample we use to estimate the model. Let us keep only year 87 , save as my_data2 dataframe.

In [11]:
```
my_data2 <- filter(my_data, year == 87)
head(my_data2)
```

A data.frame: 6 × 4

	county	year	crmrte	polpc
	<int>	<int>	<dbl>	<dbl>
1	1	87	0.0356036	0.0018279
2	3	87	0.0152532	0.0007459
3	5	87	0.0129603	0.0012343
4	7	87	0.0267532	0.0015299
5	9	87	0.0106232	0.0008602
6	11	87	0.0146067	0.0028820

Regression of Crime Rate on Police Per Capita for Year 1987 only N=90

In [12]:
```
#regression
regLectureN90 <- lm(crmrte ~ polpc, my_data2)
#show output
summary(regLectureN90)
```

```
Call:
lm(formula = crmrte ~ polpc, data = my_data2)

Residuals:
      Min        1Q    Median        3Q       Max 
-0.051400 -0.011799 -0.003837  0.006455  0.063787 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  0.02806    0.00395   7.105 2.99e-10 ***
polpc        3.18839    2.00318   1.592    0.115    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.01873 on 88 degrees of freedom
Multiple R-squared:  0.02798,	Adjusted R-squared:  0.01694 
F-statistic: 2.533 on 1 and 88 DF,  p-value: 0.115
```

Lets compare with the regression using N=630

In [13]:
```
summary(regLectureN630)
```

```
Call:
lm(formula = crmrte ~ polpc, data = my_data)

Residuals:
      Min        1Q    Median        3Q       Max 
-0.045791 -0.012476 -0.002669  0.007157  0.098927 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 0.0292403  0.0008673  33.713  < 2e-16 ***
polpc       1.2246077  0.2598397   4.713 3.01e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.01782 on 628 degrees of freedom
Multiple R-squared:  0.03416,	Adjusted R-squared:  0.03262 
F-statistic: 22.21 on 1 and 628 DF,  p-value: 3.009e-06
```

The estimated coefficients change, the sample changed, so no surprise there.

But, more importantly look what happened to the standard errors for the estimated coeffiicents...

No description has been provided for this image

Homoskedasticity Assumption

Why is that?

Let us introduce a fifth Assumption for the linear model and derive the formula for the standard errors of our estimated coefficients and see how the standard errors change as N changes

What is the homoskedasticity Assumption?

No description has been provided for this image

No description has been provided for this image

How do we obtain the estimated variance (or standard errors) of the estimated parameters?

No description has been provided for this image

No description has been provided for this image

Let us get the sample Variance of X and then SSTx

$\hat{var(x)}=\frac{SSTx}{N-1}$

So SSTx=$\hat{var(x)}$ (N-1)

In this case x is polpc

In [16]:
```
#get SSTx

xbar<-mean(my_data$polpc)
my_data$xMxbar<-my_data$polpc-xbar
SSTx=sum(my_data$xMxbar*my_data$xMxbar)
SSTx
```

A data.frame: 6 × 4

	county	year	crmrte	polpc
	<int>	<int>	<dbl>	<dbl>
1	1	81	0.0398849	0.0017868
2	1	82	0.0383449	0.0017666
3	1	83	0.0303048	0.0018358
4	1	84	0.0347259	0.0018859
5	1	85	0.0365730	0.0019244
6	1	86	0.0347524	0.0018952

0.00470481740776927

In [17]:
```
# add predicted crime rate to my_data
my_data <- mutate(my_data, crmrte_hat = regLectureN630$fitted.values)

#generate uhats to get variance of uhats
my_data <- mutate(my_data, uhat = regLectureN630$residuals)

head(my_data)
```

A data.frame: 6 × 7

	county	year	crmrte	polpc	xMxbar	crmrte_hat	uhat
	<int>	<int>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
1	1	81	0.0398849	0.0017868	-1.299959e-04	0.03142840	0.008456501
2	1	82	0.0383449	0.0017666	-1.501959e-04	0.03140366	0.006941238
3	1	83	0.0303048	0.0018358	-8.099587e-05	0.03148840	-0.001183605
4	1	84	0.0347259	0.0018859	-3.089587e-05	0.03154976	0.003176143
5	1	85	0.0365730	0.0019244	7.604127e-06	0.03159690	0.004976095
6	1	86	0.0347524	0.0018952	-2.159587e-05	0.03156115	0.003191254

Lets get the Sum of squared residuals SSR, sum of squared uhats

Since uhat_bar is zero

Then the variance of uhats is sum uhat_i squared, which is SSR divided by N-2

We divide by N-2) because the model lost two degrees of freedom a constant and an x)

In [18]:
```
#get Sum of squared residuals SSR, sum of squared uhats
#Since uhat_bar is zero
SSR<-sum(my_data$uhat*my_data$uhat)
SSR
```

0.199486423339062

In [19]:
```
#Then the variance of uhats is sum uhat_i squared,
#which is SSR divided by N-2
#we divide by N-2) because the model lost two degrees of freedom a constant

(varuhat<-SSR/(630-2))
```

0.000317653540348825

No description has been provided for this image

In [20]:
```
#so varhat of betapolice per capita hat

vhat_beta_polpc_hat<-varuhat/SSTx

(sehat_beta_polpc_hat<-sqrt(vhat_beta_polpc_hat))
```

0.259839674416346

Generate Predicted Crime Rate using b0 and b1 estimates of the regression you estimated

Plot Crime Rate and Predicted Crime Rate to see how well we are doing

Get regression line estimates and police per capita graph

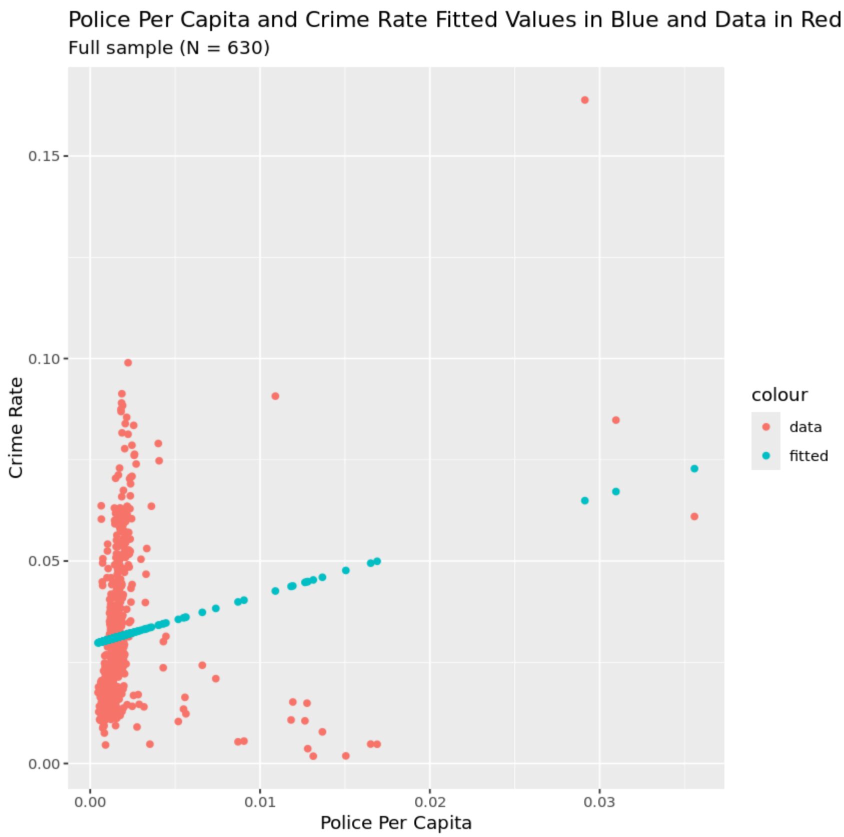
Use the full sample N=630

Combine the fitted values crime rate with the crime rate data on a scatterplot with police per capita on the horizontal x axis, for N=630

In [11]:
```
#We already generated fitted values

#you can do that with one the the two versions of commands below

#my_data$crmrte_hat<-regLectureN630$fitted.values
#my_data <- mutate(my_data, crmrte_hat = regLectureN630$fitted.values)
```

In [22]:
```
#make combined scatter plot of crime rate data and fitted values of crime ra
scatter_data_fittedVals <- ggplot(data = my_data) +
```

```
                                    geom_point(aes(x=polpc, y=crmrte, color = "data"
                                    geom_point(aes(x=polpc, y=crmrte_hat, color = "f
                                    labs(x = "Police Per Capita",
                                         y = "Crime Rate",
                                         title = "Police Per Capita and Crime Rate F
                                         subtitle = "Full sample (N = 630)")

scatter_data_fittedVals
```



In [ ]:

Take away from Lecture 5

Statistical Properties of Estimator $\hat{\beta}$

1. $\hat{\beta}$ are random variables

2. $\hat{\beta}$ are unbiased, both the intercept and the slope:

   E($\hat{\beta_0})=\beta_0$,

   E($\hat{\beta_1})=\beta_1$

   if SLR1+SLR2+SLR3+SLR4 assumptions hold.

   ```
   Where SLR is short for simple Linear Regression (SLR)

    SLR1, Y linear in parameters
   ```

```
SLR2, { $(x_i,y_i)$, i=1,…n}  random sample in the population

SLR3 There is variation in x in the sample

SLR4 E(u|x)=0
```

3. Repeating the same random sampling of N=630 observations gives different estimates, but if you were to average them up, you would find an unbiased estimator for the population parameter, because E($\hat{\beta})=\beta$.

4. Increasing sample size increases the precision of the estimate, or in other words, decreases the standard errors of the estimated coefficients, because given SLR5 (Homoskedasticity) Var($\hat{\beta})=\frac{var{u}}{SST_x}$.

In [ ]:

```
```