Lecture 5 EEP 118

Spring 2025

Sofia Villas-Boas

Lecture Plan- Lecture 5

Finish notes Lecture 4

4. Statistical Properties of Estimator

TODAY: Together here in lecture, we will go over R code in Jupyter Notebook using Crime Rate and Police Per Capita Data

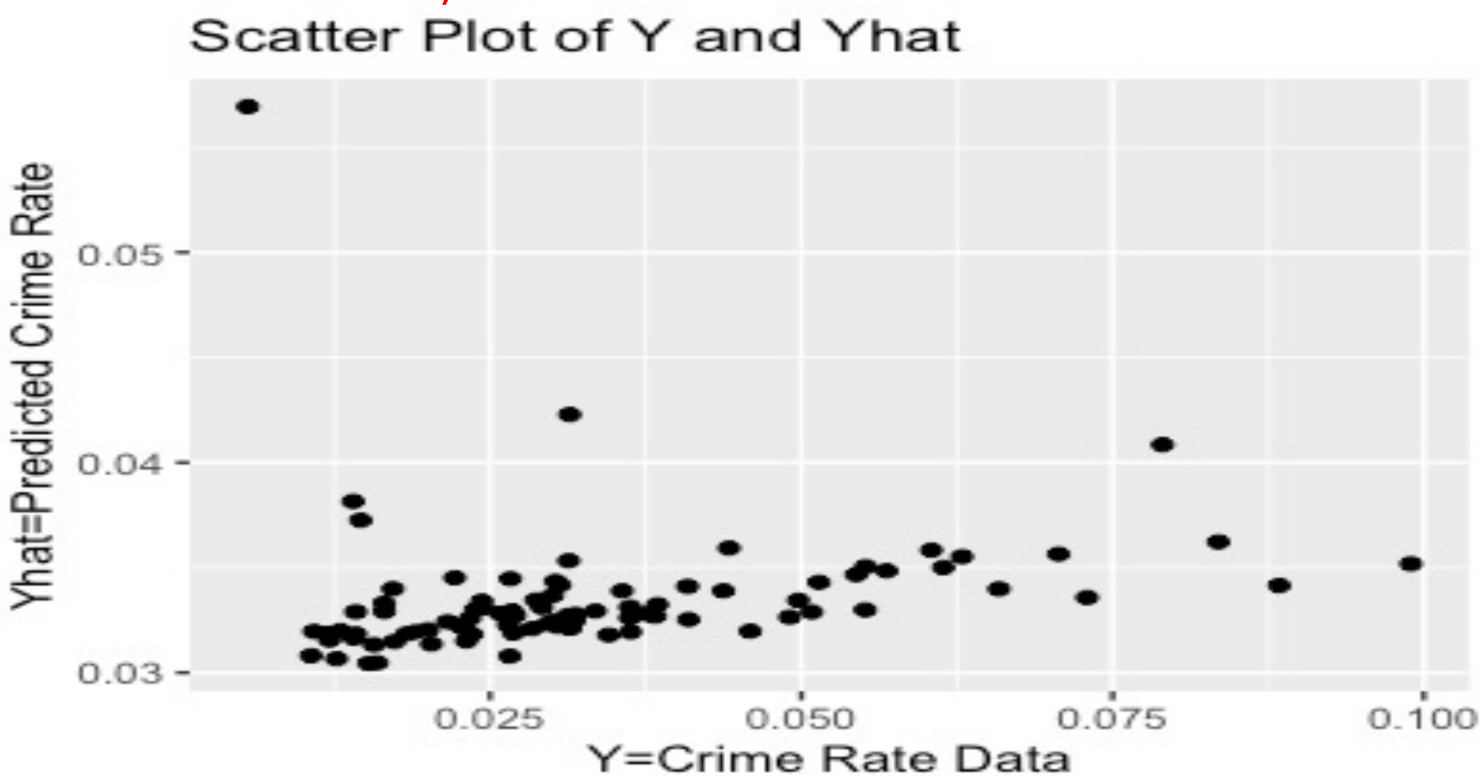
Study chapter 2

P set 1 posted and due date posted– follow P set write up instructions

Daily Assignment Lecture 5 posted - ungraded

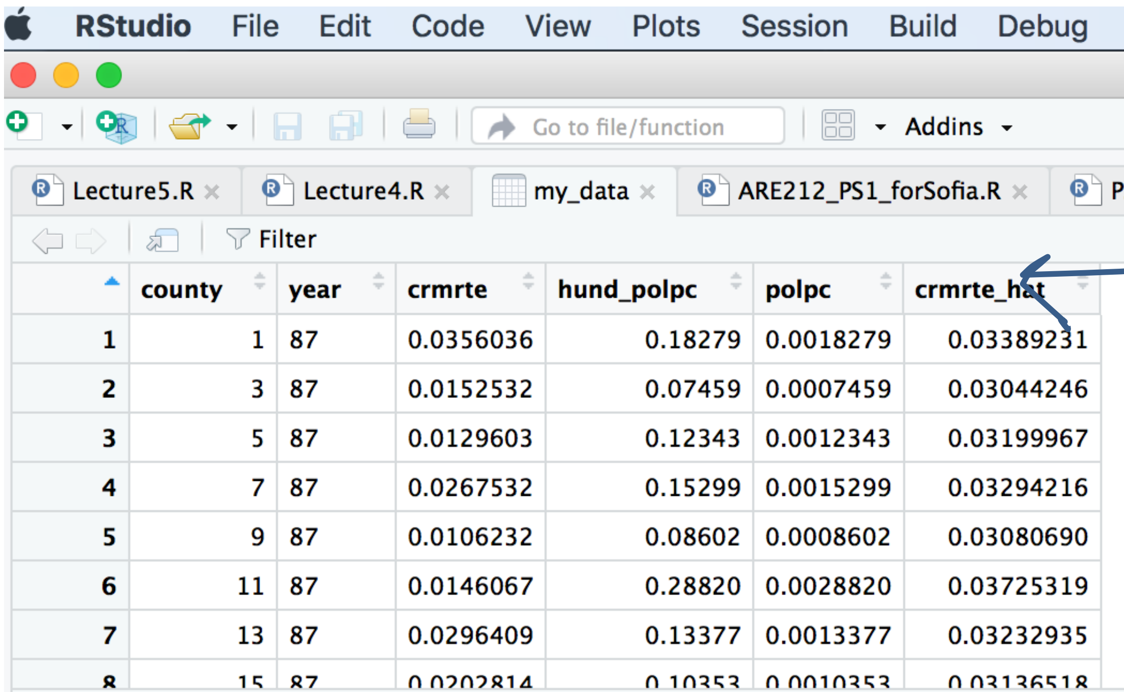
From Lecture 4

```
#------------------------------------------
scatter_Lect5 <- ggplot(my_data, aes(x=crmrte, y=crmrte_hat)) + # initiate plot
  geom_point() + # add points data
  labs(x = "Y = Crime Rate Data", # add labels
       y = "Yhat = Predicted Crime Rate",
       title = "Scatter Plot of Y and Yhat")
scatter_Lect5
```



3. Properties in relation to the sample

#create a new column of predicted crime rate, call it crmrte_hat
my_data$crmrte_hat <- regLecture4$fitted.values



	county	year	crmrte	hund_polpc	polpc	crmrte_hat
1	1	87	0.0356036	0.18279	0.0018279	0.03389231
2	3	87	0.0152532	0.07459	0.0007459	0.03044246
3	5	87	0.0129603	0.12343	0.0012343	0.03199967
4	7	87	0.0267532	0.15299	0.0015299	0.03294216
5	9	87	0.0106232	0.08602	0.0008602	0.03080690
6	11	87	0.0146067	0.28820	0.0028820	0.03725319
7	13	87	0.0296409	0.13377	0.0013377	0.03232935
8	15	87	0.0202814	0.10353	0.0010353	0.03136518

Where is a good prediction, which row of data and which county?

Where is a pretty bad prediction, which county?

3. Properties in relation to the sample

lets show these properties

$(\bar{x}, \bar{y})$ on regression line.

$\sum \hat{u} = 0.$

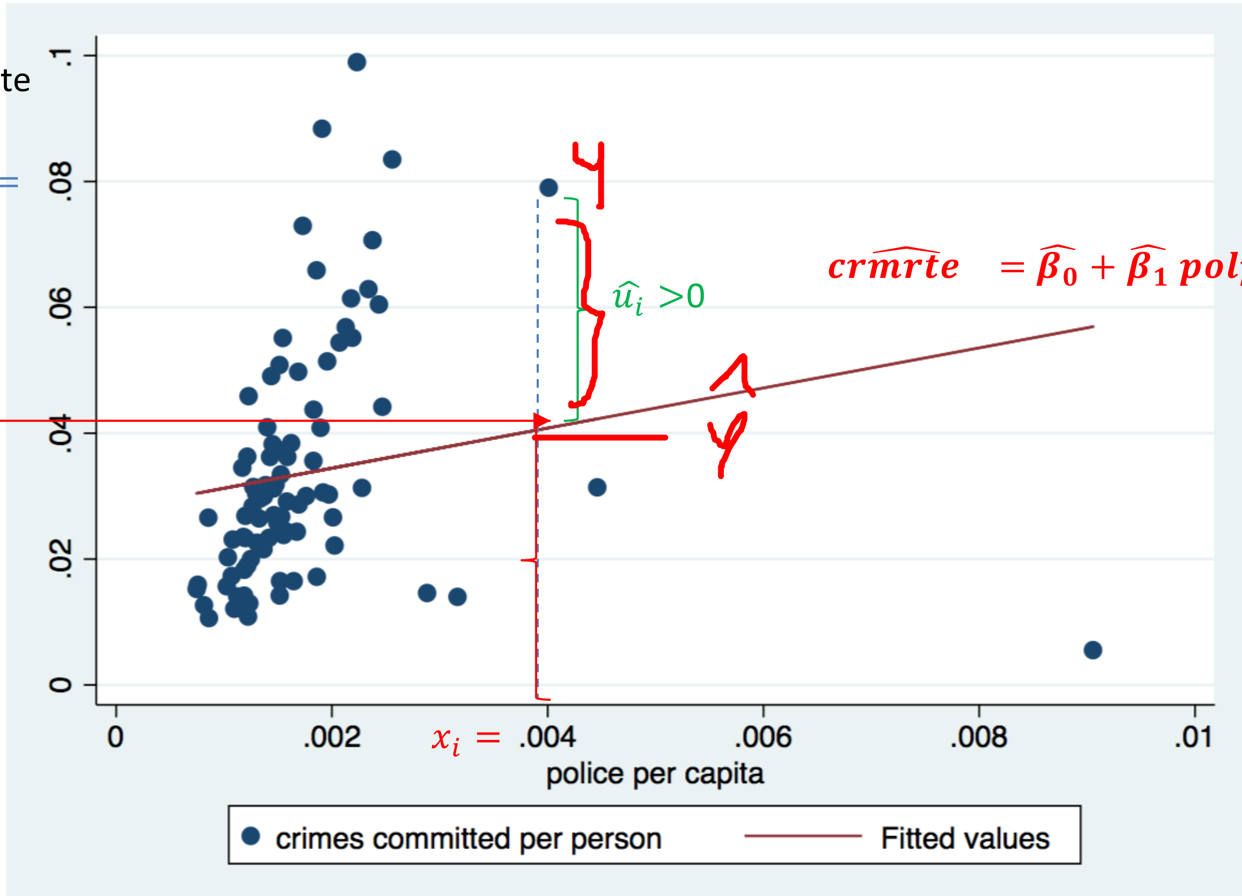
$Y - \hat{y} = \hat{u}$

$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$

Crime Rate

$y_i =$

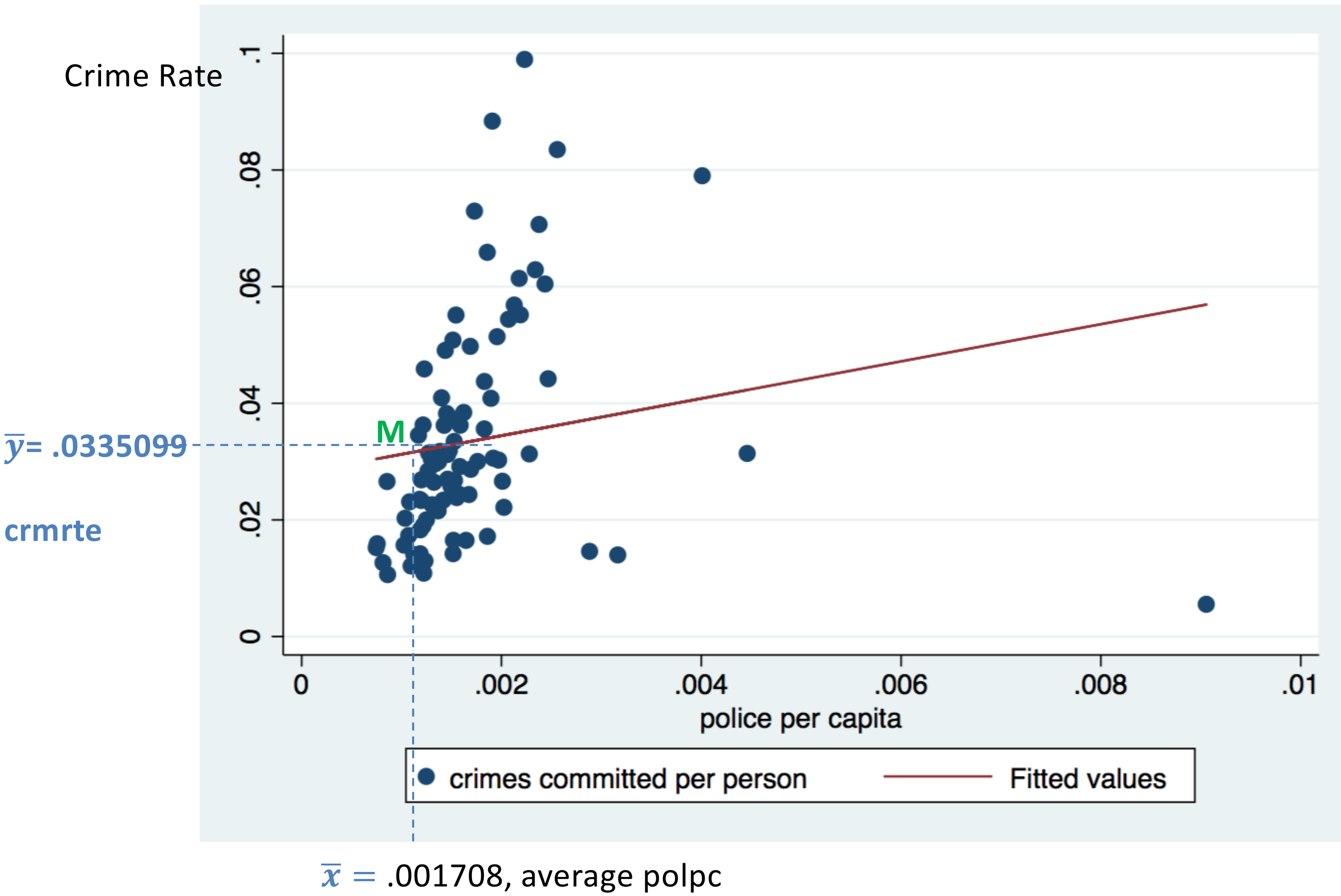
$\hat{y}_i$



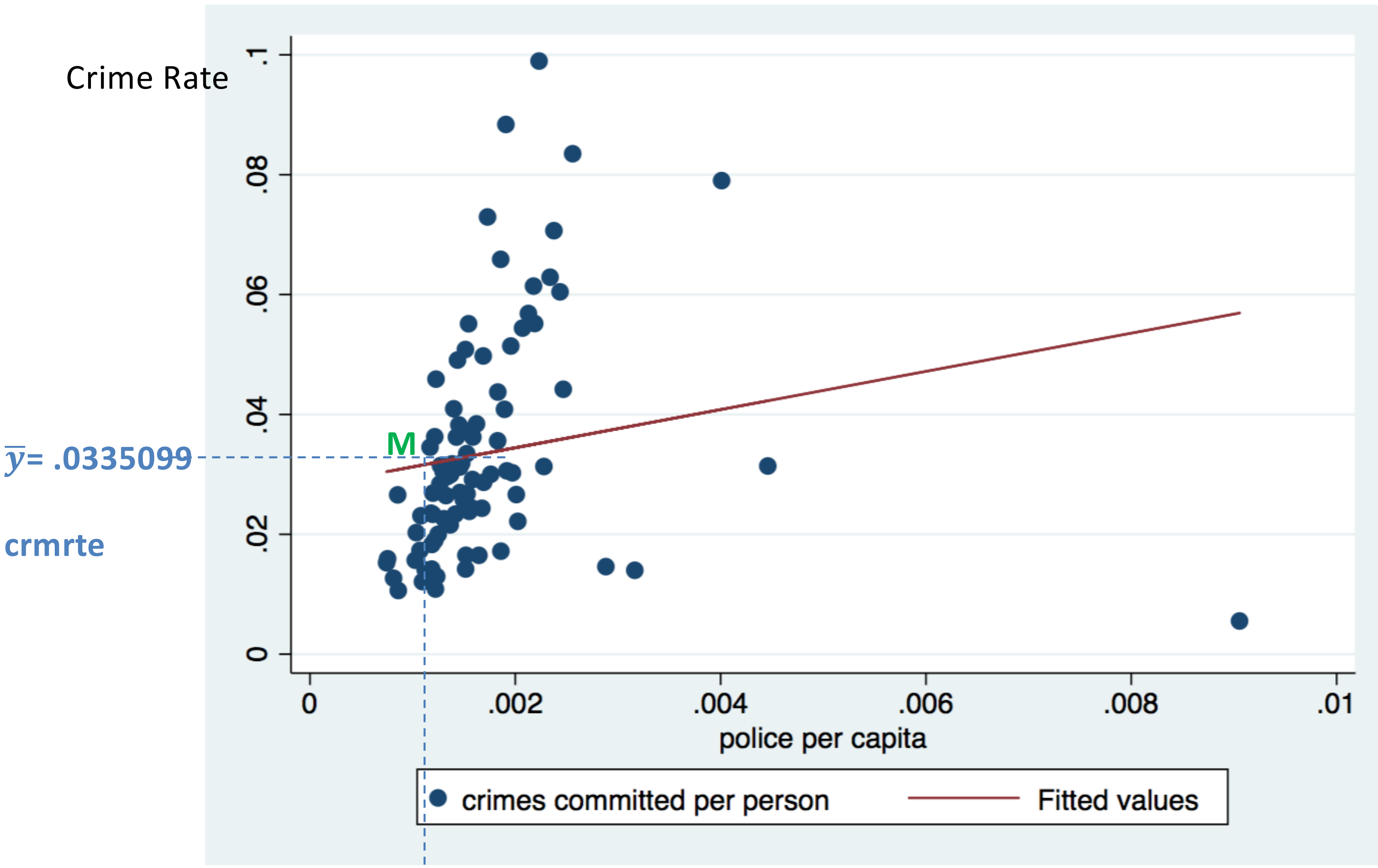
$\widehat{crmrte} = \widehat{\beta_0} + \widehat{\beta_1}\, polpc$ = 0.028 + 3.19 polpc

$$\bar{y} = \widehat{\beta_0} + \widehat{\beta_1}\ \bar{x}. \quad \text{Point M}$$

PROPERTY : $\sum \hat{u} = 0.$ → PROPERTY : ($\bar{x} = .001708$, $\bar{y}$= .0335099) on regression line, point M.



PROPERTY : $\sum \hat{u} = 0.$ → **PROPERTY : ($\bar{x} = .001708$, $\bar{y}$= .0335099) on regression line, point M.**



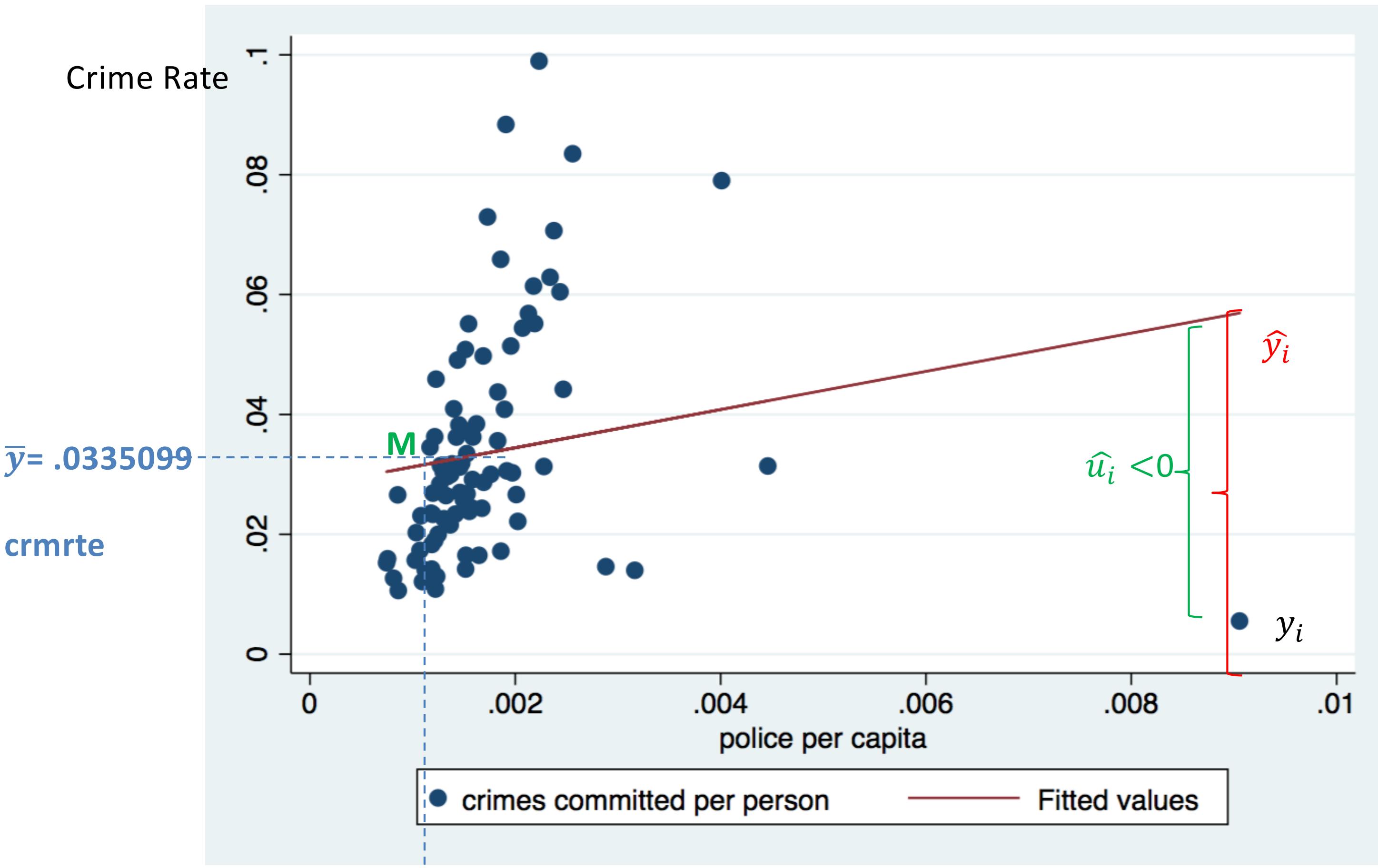
$\bar{x}$ = .001708, average polpc

Can you find a $\widehat{u_i} < 0$?

Recall that $\hat{u} = y - \hat{y}$

$\hat{u}$ = y- $\widehat{\beta_0} - \widehat{\beta_1} x$

PROPERTY : $\sum \hat{u} = 0.$ → **PROPERTY : ($\bar{x} = .001708$, $\bar{y}$= .0335099) on regression line, point M.**



$\bar{x}$ = .001708, average polpc

Can you find a $\widehat{u_i} < 0$?

Recall that $\hat{u} = y - \hat{y}$

$\hat{u}$ = y- $\widehat{\beta_0} - \widehat{\beta_1}x$

Showed that $\frac{1}{\mathrm{n}}\sum \hat{u}_i \quad = 0 \rightarrow \bar{y} = \widehat{\beta_0} + \widehat{\beta_1}\bar{x}$

We also know SST=$\underbrace{\sum_i (y_i - \bar{y})^2}_{\text{Sum Squares Total}} = \underbrace{\sum_i (\widehat{y_i} - \bar{y})^2}_{\text{Sum Squares Explained}} + \underbrace{\sum_i \hat{u}_i^2}_{\text{Sum Squares Residual}}$

Then Goodness of Fit, fraction of variation of Y explained by the model is the R squared defined by

$R^2 = \frac{SSE}{SST} = \frac{\sum_i (\widehat{y_i} - \bar{y})^2}{\sum_i (y_i - \bar{y})^2}$ = goodness of fit or we can write also R^2 =1- $\frac{SSR}{SST}$

R Regression output

$$\widehat{crmrte} = \widehat{\beta_0} + \widehat{\beta_1}\, polpc$$ = 0.028 + 3.19 polpc

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

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
```

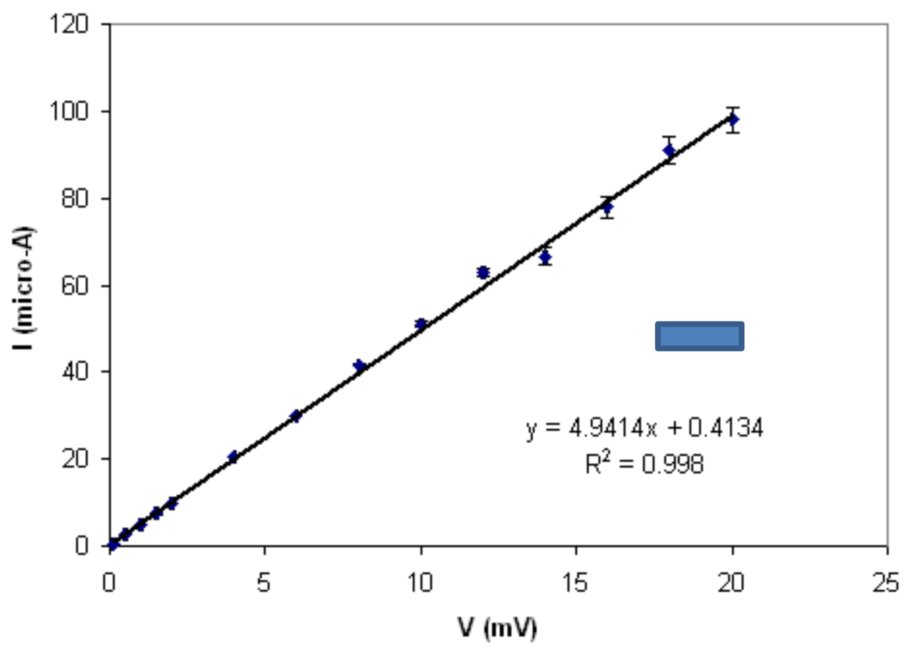
Beta0_hat

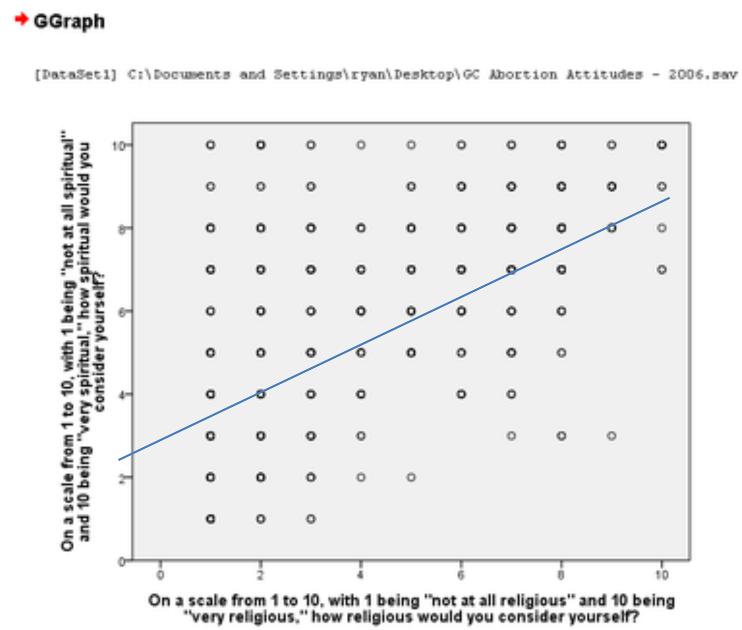
Beta1_hat

$$R^2 = \frac{SSE}{SST} = \frac{\sum_i (\hat{y}_i - \bar{y})^2}{\sum_i (y_i - \bar{y})^2}$$

R^2 ==0.02798

Which scatter plot has a linear model with R-squared closer to 1 and R-squared closer to 0?





Take away Lecture 4

1. Population model

$y \quad = \beta_0 + \beta_1 x \quad + u$, Model Linear in parameters;

Assumption 1: E[u]=0 Assumption 1: E[u|x]=0

2. Estimation based on a Sample

$y \quad = \widehat{\beta_0} + \widehat{\beta_1} x \quad + \hat{u}, \widehat{\beta_1} = \frac{cov(x,y)}{var(x)}, \widehat{\beta_0} = \bar{y} - \widehat{\beta_1}\bar{x},$

you need to be able to interpret.

3. Properties in relation to the sample – Goodness of Fit

($\bar{x}$,$\bar{y}$) on regression line.

$\sum \hat{u} = 0.$

$R^2 = \frac{SSE}{SST} = \frac{\sum_i(\widehat{y_i}-\bar{y})^2}{\sum_i(y_i-\bar{y})^2}$, goodness of fit

where SST=$\sum_i(y_i - \bar{y})^2 = \sum_i(\widehat{y_i} - \bar{y})^2 + \sum_i \hat{u}_i^2$

Lecture 5

4. Statistical Properties of Estimator

TODAY: Together here in lecture we will go over R code and Jupyter notebook using Crime Rate and Police Per Capita Data

Study chapter 2

Pset 1 posted and due date posted– follow Pset write up instructions

Daily Assignment Lecture 5 posted - ungraded

Now and your Daily Assignment for lecture 5 go to Jupyter

Go to Bcourses and get the link to datahub,

And click on it there and access the notebook for Lecture 5 where you can do all we did today there

To go to R studio in data hub go to Bcourses and click the link to R studio in datahub – we did this in Lecture 4

$\widehat{\beta_0}$ *and* $\widehat{\beta_1}$ are random variables

β_o *and* β_1 are <u>true unknown values</u> from the population regression

$\widehat{\beta_0}$ *and* $\widehat{\beta_1}$ are <u>estimators</u>, (formulas) to compute an <u>estimate</u> (a value) with a sample

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

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

If certain assumptions hold the distribution of $\widehat{\beta_0}$ *and* $\widehat{\beta_1}$ will be related to β_o *and* β_1

Model

- $crmrte = \beta_0 + \beta_1\, polpc + u$ not known

- We used a sample and estimated $\widehat{\beta_0}$ *and* $\widehat{\beta_1}$
- $\widehat{\beta_0}$ *and* $\widehat{\beta_1}$ are <u>random variables</u>
- Beta hat : estimated parameters

  $$\widehat{crmrte}_i = \widehat{\beta_0} + \widehat{\beta_1}\ polpc_i$$

$\widehat{\beta_1}$ is the marginal effect of policy per capita on predicted crime rate, namely on $\widehat{crmrte}$

Lets use now a dataset for more years and counties such that N=630

I drew ten independent samples of N=630:

Repeating the same random sampling of 630 observations gives different estimates of $\widehat{\beta_0}$ *and* $\widehat{\beta_1}$

ENV ECON 118 / IAS 118 - Introductory Applied Econometrics
Lecture 5

Table 1: 10 Bootstrap Sample BetaHat Estimates

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
police per capita	2.017***	0.971***	0.975***	0.930***	1.157***	1.403***	1.776***	1.828***	1.815***	1.973***
	(0.223)	(0.309)	(0.318)	(0.241)	(0.187)	(0.222)	(0.210)	(0.191)	(0.188)	(0.159)
Constant	0.029***	0.031***	0.031***	0.032***	0.031***	0.030***	0.030***	0.030***	0.028***	0.028***
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
Num of Obs.	630	630	630	630	630	630	630	630	630	630
R squared	0.12	0.02	0.01	0.02	0.06	0.06	0.10	0.13	0.13	0.20
Mean Dep Var										

Dependent Variable is Number of Crimes Per Capita (Crime Rate).
$*p < 0.10$, $**p < 0.05$, $***p < 0.01$

$$\widehat{crmrte}_i = 0.029 + 2.02\ \ polpc$$

$$\widehat{crmrte}_i = 0.03 + 1.776\, polpc$$

but if you were to average them up, you would find on average an estimate that has as expected value the population parameter beta, because $E\left(\hat{\beta}\right) = \beta$.

Average of(2.017, 0.971, 0.975, 0.93, 1.157, 1.403, 1.776, 1.828, 1.815, 1.973)

You can add them up and divide by ten = 14.845/ 10 = **1.4845**

Simple Linear Regression (SLR) Assumptions

Show that if SLR1-SLR4 then $\hat{\beta}$ unbiased, that is E[$\hat{\beta}$]= β

E[beta0hat]=beta0
E[beta1hat]=beta1 if

$\left\{ \right.$ SLR1, Y linear in parameters: y=beta0+beta1 x+u
SLR2, { (xi,yi), i=1,…n} random sample in the population
SLR3 variation in x in sample: var(x) not zero
SLR4 Zero conditional Mean E(u|x)=0

SLR1 The population Model is linear in parameters

$$y \quad = \beta_0 + \beta_1 x + u \qquad (1)$$

SLR2 Random sampling, then we can write (1) in terms of the random sample as

$$y_i = \beta_0 + \beta_1 x_i + u_i \quad , i=1,2, …, n$$

SLR3 sample variance of x cannot be zero, that is, the x's cannot be all equal.

SLR4 zero conditional mean of the disturbance u is that $E[u|x] = 0$

$for\ the\ random\ sample\ E[u_i|x_i]$=0, i=1,2, …, n

Show that $\widehat{\beta_0}$ $\widehat{\beta_1}$ unbiased, that is E[$\hat{\beta}$]= β if SLR1+ SLR2+SLR3+SLR4

$$\widehat{\beta_1} = \frac{cov(x,y)}{var(x)} = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})/\cancel{(n-1)}}{\sum_i (x_i - \bar{x})^2/\cancel{(n-1)}}$$

$$\widehat{\beta_1} = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sum_i (x_i - \bar{x})^2 = SSTx} = \frac{1}{SSTx} \sum_i (x_i - \bar{x})(y_i - \bar{y})$$

SLR1+SLR2

Substituting

$y_i = \beta_0 + \beta_1 x_i + u_i$

And $\bar{u}$ =0

And $\bar{y} = \beta_0 + \beta_1 \bar{x} + \bar{u}$

$$\widehat{\beta_1} = \frac{1}{SSTx} \sum_i (x_i - \bar{x})\{\beta_1 (x_i - \bar{x}) + u_i\}$$

Proof in book chapter 2, page 54 in 3rd edition

Cont $\widehat{\beta_1}$ unbiased proof

$$\widehat{\beta_1} = \frac{1}{SSTx} \sum_i (x_i - \bar{x})(\beta_1(x_i - \bar{x}) + u_i)$$

$$= \frac{1}{SSTx} \{\beta_1 \sum_i (x_i - \bar{x})(x_i - \bar{x}) + \sum_i (x_i - \bar{x})u_i)\}$$

$$= \frac{1}{SSTx} \{\beta_1 SSTx + \sum_i (x_i - \bar{x})u_i \}$$

$$= \beta_1 \frac{\cancel{SSTx}}{\cancel{SSTx}} + \frac{\sum_i (x_i - \bar{x})u_i}{SSTx}$$

$$E[\widehat{\beta_1} | x] = \beta_1 + E[\frac{\sum_i (x_i - \bar{x})\, u_i}{SSTx} | x] = \beta_1 + \frac{1}{SSTx} \sum_i (x_i - \bar{x}) \underbrace{E(u_i | x]}_{=0}$$

SSTx not 0

So $E[\widehat{\beta_1}] = \beta_1$

SLR3

SLR4

What about $\widehat{\beta_0}$?

$$\widehat{\beta_0} = \bar{y} - \widehat{\beta_1}\bar{x}$$

since $\bar{y} = \beta_0 + \beta_1 \bar{x} + \underbrace{\bar{u}}_{=0}$

Then $\widehat{\beta_0} = \beta_0 + \beta_1 \bar{x} - \widehat{\beta_1}\bar{x} = \beta_0 + (\beta_1 - \widehat{\beta_1})\, \bar{x}$

$E[\widehat{\beta_0}\,|\,x] = \beta_0 + \underbrace{(\beta_1 - E[\widehat{\beta_1}|x])}_{=0}\, \bar{x} = \beta_0$

SLR5

IF SLR5 "HOMOSKEDASTICITY", i.e., var[u|x] =var(u)= σ_u^2 , i.e., does not depend on x, then

$$\mathrm{var}\left(\hat{\beta}_1\right) = \frac{\sigma_u^2}{\mathrm{SST}_x} = \frac{\sigma_u^2}{(N-1)S_x^2} \text{ and } \mathrm{var}\left(\hat{\beta}_0\right) = \frac{\sigma_u^2}{SST_x}\frac{\sum x_i^2}{N}$$

given that σ_u^2 is unknown it can be estimated by $\dfrac{\sum \hat{u}_i^2}{N-2}$

(later in class, how to deal with this)

Illustration of Homoskedasticity

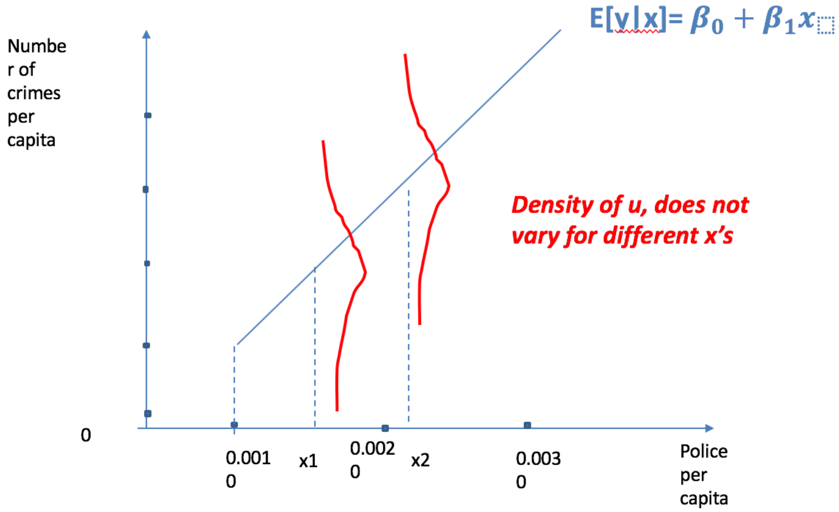
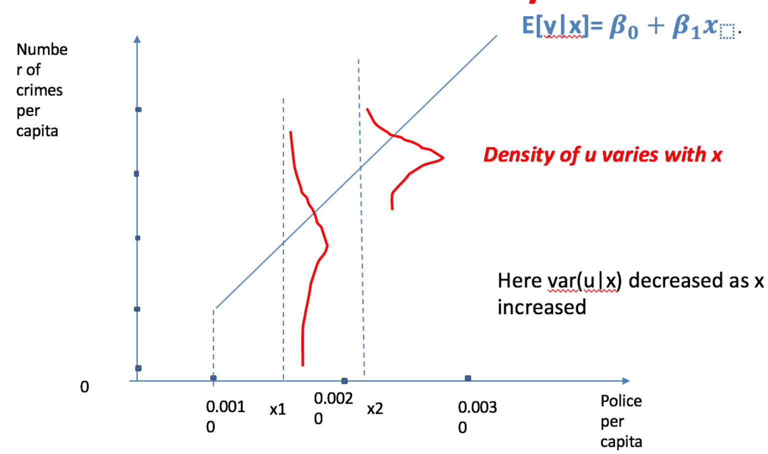


Illustration of Heteroskedasticity



Here var(u|x) decreased as x increased

Illustration of Homoskedasticity of population disturbance u

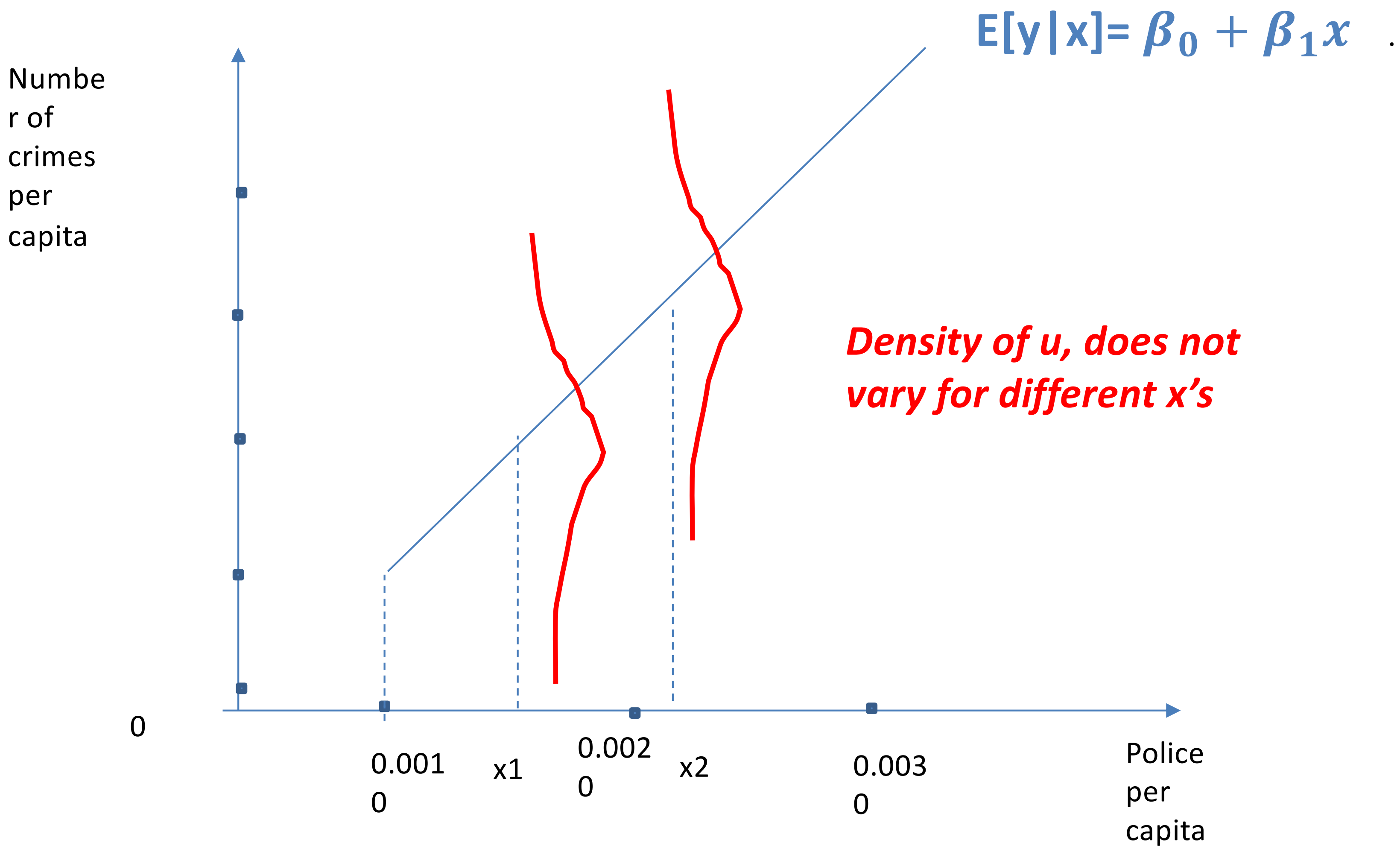
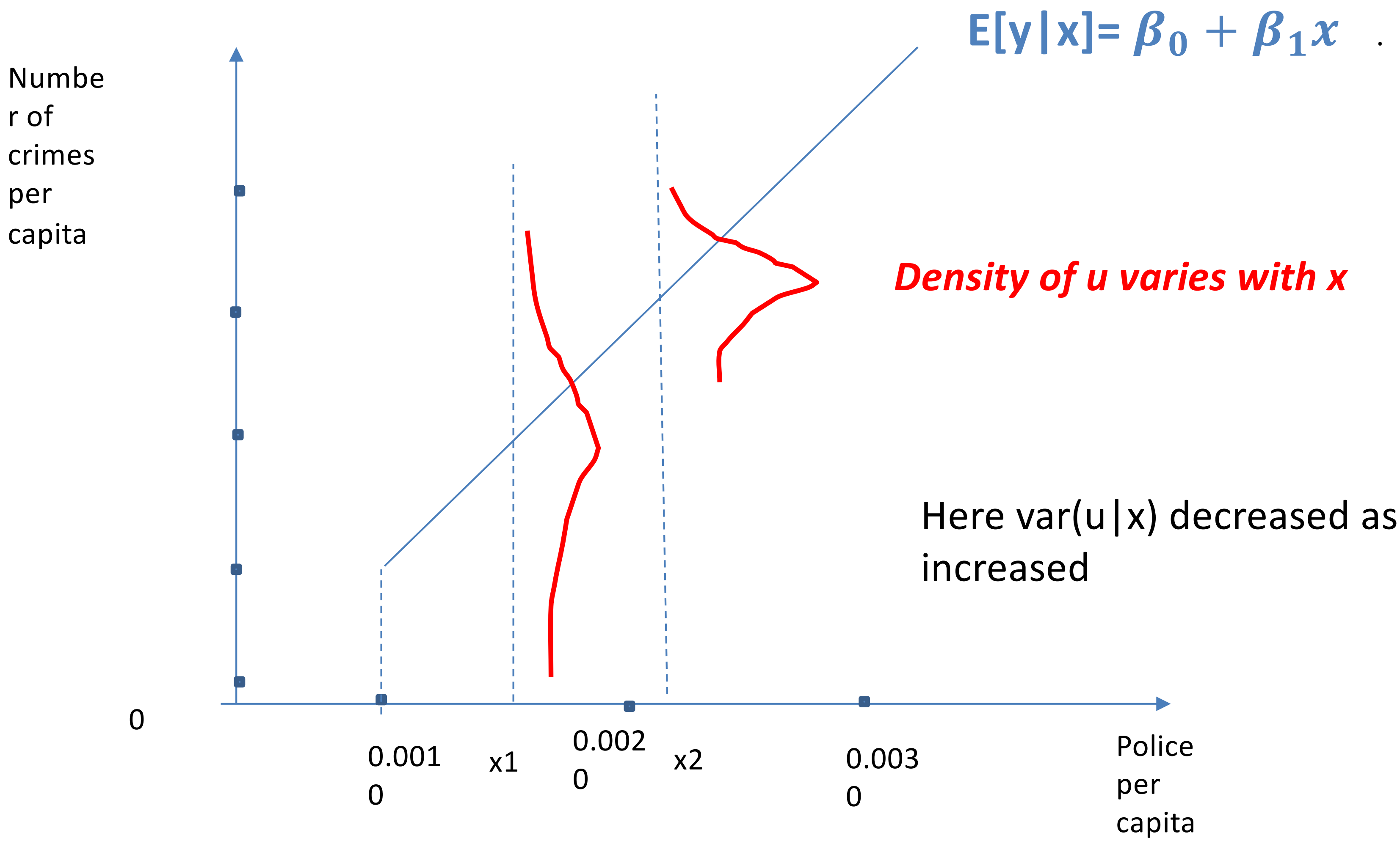


Illustration of Heteroskedasticity of population disturbance u



Here var(u|x) decreased as x increased

Take away Lecture 5 :
Statistical Properties of Estimator betahat

1. beta_hat are random variables
2. beta_hat are unbiased (E(Beta0hat)=Beta0, E(beta1hat)=beta1 if

Simple Linear Regression (SLR) Assumptions

SLR1, Y linear in parameters
SLR2, { (xi,yi), i=1,…n} random sample in the population
SLR3 variation in x in 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

$$E\left(\hat{\beta}\right) = \beta$$

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

$$\operatorname{var}\left(\hat{\beta}_1\right) = \frac{\operatorname{var}(u)}{\mathrm{SST}_x}$$

Comparing se of beta_hat of polpc

N=90

```
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  td. Error t  lue Pr(>|t|)    
(Intercept)  0.02806    0.00395    105 2.99e-10 ***
polpc        3.18839    2.00318    592    0.115    
---
Signif. codes:  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
```

N=630

```
lm(formula = crmrte ~ polpc, data = sample400)

Residuals:
      Min        1Q    Median        3Q       Max 
-0.049522 -0.012204 -0.002544  0.006653  0.091215 

Coefficients:
            Estimate  td. Error   value Pr(>|t|)    
(Intercept) 0.028714   0.001064   6.988  < 2e-16 ***
polpc       1.507435   0.285827   5.274  2.2e-07 ***
---
Signif. codes:  0 '*                  0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.01805 on 398 degrees of freedom
Multiple R-squared:  0.06532,	Adjusted R-squared:  0.06297 
F-statistic: 27.81 on 1 and 398 DF,  p-value: 2.196e-07
```

As N increases from N=90 to N=630 the standard error (se) of the beta hat for police per capita decreases !

A good way to present the regression results

Y_hat = 0.028 + 1.507 x

(0.001) (0.28)

R squared = 0.061

Where in parentheses under the estimates of beta_hats are the standard errors of the beta_hats

Increasing n/sample size increases the precision of the estimate, because $var(\widehat{\beta_1})$

$$\widehat{var(\widehat{\beta_1})} = \frac{\widehat{\sigma_u^2}}{SST_x} = \frac{SSR}{(n-2)\, SST_x}$$

WRITE ON THE BOARD

Practical take-away

$\widehat{var(\widehat{\beta_1})}$ is the measure of the variation we can expect across the different estimators $(\widehat{\beta}\ )$of β

Many samples, then many $\widehat{\beta}$, all distributed around the true (unknown) value of β with standard error $\widehat{se(\widehat{\beta}\ )}$

HOW TO REDUCE $\widehat{se(\widehat{\beta}\ )}$?

- Large n
- Large variation in x
- Small variation in u

Lets get $var\left(\widehat{\beta_1}\right)$

$\widehat{\sigma^2_{polpc}}$ =0.0027 * 0.0027

Variable	Obs	Mean	Std. Dev.	Min	Max
crmrte	630	.0315099 $= \bar{y}$	.0181229	.0018116	0.16
polpc	630	.001708 $= \bar{x}$	.0027	.000459	.0355

$$SST_x = SST_{polpc} = \widehat{\sigma^2_{polpc}}\ (n-1) = 0.0027^2(630-1)$$

Calculating $\mathrm{se}\left(\hat{\beta}_1\right)$ **next slide**

Back to sample N=630

Calculating $\mathrm{se}\left(\hat{\beta}_1\right)$

The unknown true values for the variance and standard deviation of the random variable $\widehat{\beta}$

$$\mathrm{var}\left(\hat{\beta}_1\right) = \frac{\sigma_u^2}{\mathrm{SST}_x} \qquad \mathrm{sd}\left(\hat{\beta}_1\right) = \frac{\sigma_u}{\sqrt{\mathrm{SST}_x}}$$

Estimation are obtained by replacing σ_u^2 by an estimation computed from the residuals

$$\widehat{\sigma_u^2} = \frac{\sum_i \widehat{u_i^2}}{n-2}$$

$$\mathrm{var}\left(\hat{\beta}_1\right)=\frac{\sigma_u^2}{\mathrm{SST}_x}=\frac{\sigma_u^2}{(N-1)S_x^2} \text{ and } \mathrm{var}\left(\hat{\beta}_0\right)=\frac{\sigma_u^2}{SST_x}\frac{\sum x_i^2}{N}$$ given that σ_u^2 is unknown it can be estimated by $\frac{\sum \hat{u}_i^2}{N-2}$

Given that $\widehat{\sigma^2_{polpc}}$ =0.0027 * 0.0027

$SST_x = SST_{polpc} = \widehat{\sigma^2_{polpc}}\ (n-1) = 0.0027^2(630-1)$ =0.0045

Compute $\widehat{\sigma_u^2} = \frac{\sum_i \widehat{u_i^2}}{n-2} = \frac{SSR}{n-2} = \frac{0.19948}{630-2} = 0.000317$ =0.01782*0.01782

```
lm(formula = crmrte ~ polpc, data = sample630)

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
```

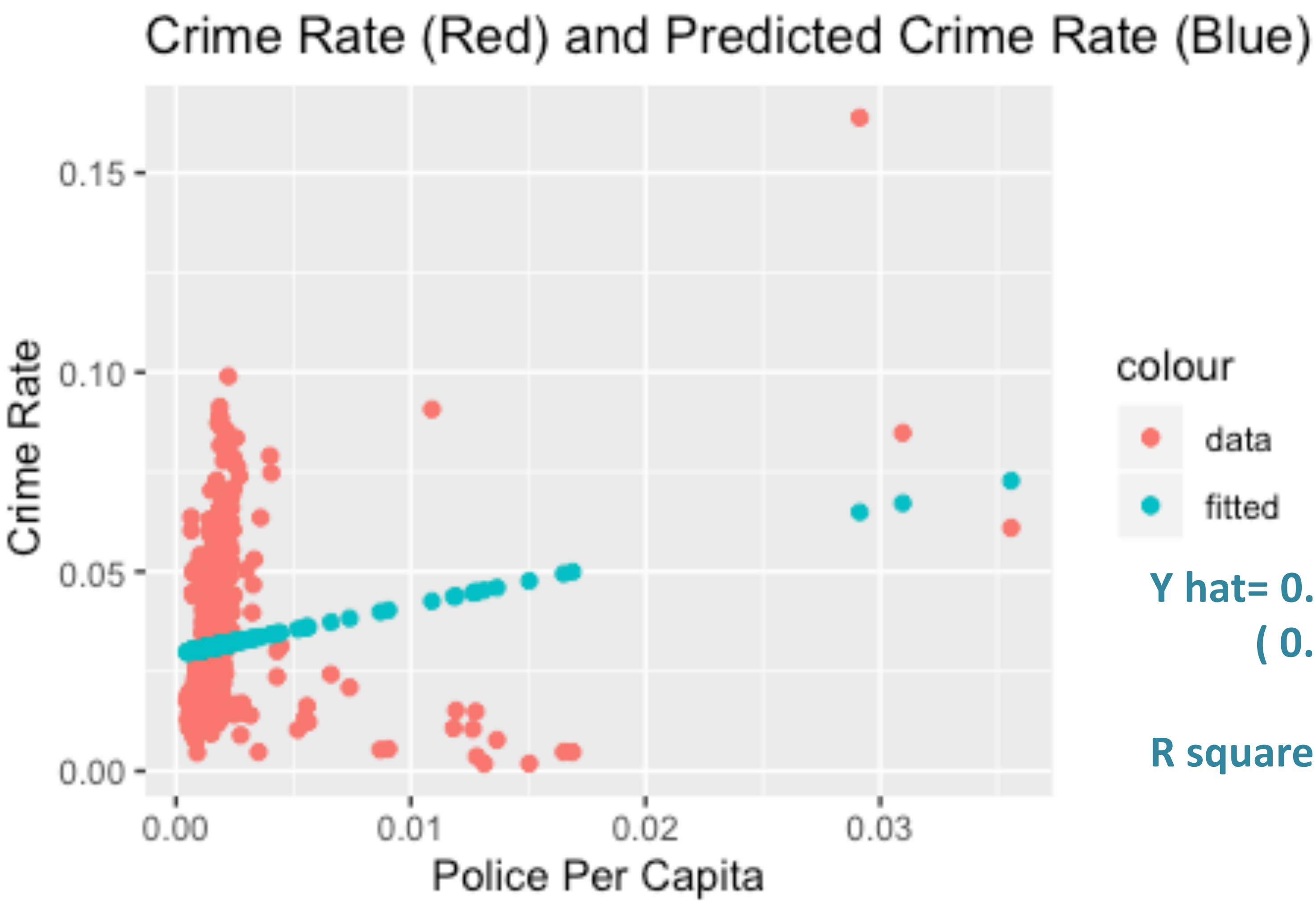
$$\widehat{var\left(\widehat{\beta_1}\right)} = \frac{\widehat{\sigma_u^2}}{SST_x}$$

$$= \frac{\widehat{\sigma_u^2}}{SST_{polpc}}$$

$$= \frac{0.000317}{0.0027^2(630-1)}$$

$$\widehat{se\left(\widehat{\beta_1}\right)} = \sqrt{0.00011}$$

$$= 0.2598397$$



Y hat= 0.029 + 1.225 x
(0.0008) (0.25)

R squared = 0.0036

R code for this figure in Lecture5.R