Lecture 15 EEP118

In [1]:

```
#install packaged and load the data
#Lecture15.R
#LECTURE 15

# Load the 'pacman' package
library(pacman)
#packages to use load them now using the pacman "manager"
p_load(dplyr, haven, readr)
#Another great feature of p_load(): if you try to load a package that is not
p_load(ggplot2)

pacman::p_load(lfe, lmtest, haven, sandwich, tidyverse)
# lfe for running fixed effects regression
# lmtest for displaying robust SE in output table
# haven for loading in dta files
# sandwich for producing robust Var-Cov matrix
# tidyverse for manipulating data and producing plots


#if in R studio comment the following line to change into Lecture 15 directo
#setwd("/Users/sofiavillas-boas/Dropbox/EEP118_Spring2024/Lectures/Lecture15

#set scientific display off, thank you Roy
options(scipen=999)

#read in a Stata dataset
my_data <- read_dta("Lecture14hprice1.dta")
head(my_data)

#    Variable |       Obs        Mean    Std. Dev.       Min        Max
#-------------+---------------------------------------------------------
#  price |      88     293.546    102.7134        111        725         pri
#bdrms |       88    3.568182    .8413926          2          7         bdr
#lotsize |     88    9019.864    10174.15       1000      92681         lot
#sqrft |       88    2013.693    577.1916       1171       3880         sqr
#colonial |     88    .6931818    .4638161          0          1
```

```
Installing package into ‘/srv/r’
(as ‘lib’ is unspecified)



lfe installed
```

A tibble: 6 × 10

price	assess	bdrms	lotsize	sqrft	colonial	lprice	lassess	llotsize
<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
300.000	349.1	4	6126	2438	1	5.703783	5.855359	8.720297
370.000	351.5	3	9903	2076	1	5.913503	5.862210	9.200593
191.000	217.7	3	5200	1374	0	5.252274	5.383118	8.556414
195.000	231.8	3	4600	1448	1	5.273000	5.445875	8.433811
373.000	319.1	4	6095	2514	1	5.921578	5.765504	8.715224
466.275	414.5	5	8566	2754	1	6.144775	6.027073	9.055556

How to obtain an estimate of the Prediction for the average value of all houses with bdrms=3, sqrft=2000, and lotsize=9000, that is get a prediction for the value of a house on average with certain characteristics?

The trick is to transform the data and run a regression with the transformed data so that then the estimate of the constant is the average prediction you want and you also gets its standard error

In [2]:

```
#trick
#Prediction for the average value of all houses with bdrms=3, sqrft=2000, ar
#generate transfored variables such that then the estimated constant gives u
#gen bdrms0=bdrms-3
#gen sqrft0=sqrft-2000
#gen lotsize0=lotsize-9000
#reg price bdrms0 sqrft0 lotsize0

my_data$bdrms0<-my_data$bdrms-3
my_data$sqrft0<-my_data$sqrft-2000
my_data$lotsize0<-my_data$lotsize-9000
reg14rev <- lm(price~bdrms0+lotsize0+sqrft0, my_data)
summary(reg14rev)
```

```
Call:
lm(formula = price ~ bdrms0 + lotsize0 + sqrft0, data = my_data)

Residuals:
     Min       1Q   Median       3Q      Max 
-120.026  -38.530   -6.555   32.323  209.376 

Coefficients:
               Estimate  Std. Error t value             Pr(>|t|)    
(Intercept) 283.9529869   8.1210704  34.965 < 0.0000000000000002 ***
bdrms0       13.8525217   9.0101454   1.537              0.12795    
lotsize0      0.0020677   0.0006421   3.220              0.00182 ** 
sqrft0        0.1227782   0.0132374   9.275   0.0000000000000166 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 59.83 on 84 degrees of freedom
Multiple R-squared:  0.6724,	Adjusted R-squared:  0.6607 
F-statistic: 57.46 on 3 and 84 DF,  p-value: < 0.00000000000000022
```

From the output we can get a confidence interval

t table got 5% critical from the t 88-4=84 degrees of freedom,, tc14

95% conf interval for predicted average prices of a 3 bd rooom house 2000 sqrt, 9000lot

283.95 - tc14 * 8.121 ; 283.95 + tc14 * 8.121

The critical value for the 95% confidence interval is 2 (between 1.987 and 2.000 to be exact). The CI for average price E is thus

283.95 ±2 (8.12) ≈ 267.7 : 300.2 .

We predict the mean price to be between 267.7 thousand dollars and 300.2 thousand dollars. This was the CI for the average house E (of the above characteristics). The CI for the average house E is not the same as the CI for p, the price of a particular house of the above characteristics!!!

In [3]:

```r
#regression  in logs and in levels
# and how to choose:

#in logs
reg15log <- lm(lprice~bdrms+lotsize+sqrft, my_data)
summary(reg15log)
lprice_hat<-reg15log$fitted.values
price_hat<-exp(lprice_hat)
aa<-exp(0.1899*0.1899*0.5)
my_data$price_hat<-price_hat*aa
#correlate then
cor(my_data$price,my_data$price_hat)
```

```
Call:
lm(formula = lprice ~ bdrms + lotsize + sqrft, data = my_data)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.73389 -0.10792 -0.01595  0.11181  0.63914 

Coefficients:
               Estimate  Std. Error t value             Pr(>|t|)    
(Intercept) 4.759375453 0.093536105  50.883 < 0.0000000000000002 ***
bdrms       0.025238784 0.028592798   0.883              0.37992    
lotsize     0.000005602 0.000002038   2.749              0.00732 ** 
sqrft       0.000364117 0.000042008   8.668    0.000000000000277 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1899 on 84 degrees of freedom
Multiple R-squared:  0.6223,	Adjusted R-squared:  0.6088 
F-statistic: 46.13 on 3 and 84 DF,  p-value: < 0.00000000000000022
```

0.8372370977254

In [4]:

```
#in levels now:
reg15lev <- lm(price~bdrms+lotsize+sqrft, my_data)
summary(reg15lev)
my_data$price_hat2<-reg15lev$fitted.values
#correlate them
cor(my_data$price,my_data$price_hat2)
```

```
Call:
lm(formula = price ~ bdrms + lotsize + sqrft, data = my_data)

Residuals:
     Min       1Q   Median       3Q      Max 
-120.026  -38.530   -6.555   32.323  209.376 

Coefficients:
               Estimate  Std. Error t value           Pr(>|t|)    
(Intercept) -21.7703081  29.4750419  -0.739            0.46221    
bdrms        13.8525217   9.0101454   1.537            0.12795    
lotsize       0.0020677   0.0006421   3.220            0.00182 ** 
sqrft         0.1227782   0.0132374   9.275 0.0000000000000166 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 59.83 on 84 degrees of freedom
Multiple R-squared:  0.6724,	Adjusted R-squared:  0.6607 
F-statistic: 57.46 on 3 and 84 DF,  p-value: < 0.00000000000000022
```

0.819976968080505

We pick the one with the biggest correlation between price and predicted price.

In this case:

correlate them

cor(my_data$price, my_data$price_hat2) [1] 0.819977 with levels

cor(my_data$price, my_data$price_hat) [1] 0.8372371 with logs

If reg in levels, the correlation between price and predicted price is 0.8199

If in logs , correlation of price and resulting predicted price is 0.8372, so…

in logs is chosen