Lecture12notebook

February 23, 2025

1 Lecture 12 - EEP 118 Spring 2025

Birth weight and then also college analysis that I ask you to do as daily assignment - see the code provided on bcourses and type your commands in the bottom of this notebook yourself

This is the notebook for Lecture 12. This concludes the lecture material for the midterm.

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 CONTROL ENTER at same time

[2]:
```
#Lecture12_birthWeight.R
#LECTURE 12, EEP 118

install.packages("pacman")
# 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␣
 ↪installed on your machine, p_load() install the package for you, rather than␣
 ↪throwing an error. For instance, let's install and load one final package␣
 ↪named ggplot2.
p_load(ggplot2)

pacman::p_load(lfe, lmtest, haven, sandwich, tidyverse,psych,car)
# 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


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

```
Installing package into ‘/srv/r’
```

```
(as 'lib' is unspecified)
```

[3]:
```
#-----------------------------------------
#1. Read in data and see the top rows to see column names etc
#-----------------------------------------

#read in a Stata dataset
my_data <- read_dta("Lecture12BWGHT.dta")
head(my_data)
```

A tibble: 6 × 15

faminc <dbl>	cigtax <dbl>	cigprice <dbl>	bwght <dbl>	fatheduc <dbl>	motheduc <dbl>	parity <dbl>	male <dbl>	white <dbl>	cigs <dbl>
22.5	21.0	136.7	73	14	14	1	1	1	0
17.5	31.0	150.6	125	2	5	4	1	1	0
65.0	2.5	109.4	116	10	12	1	1	0	0
65.0	33.0	149.1	98	14	12	1	0	0	0
42.5	30.0	138.3	127	12	12	1	0	1	0
47.5	18.0	120.5	101	12	14	3	0	0	0

[4]:
```
#summary stats of birth weight and parity cigs faminc fatheduc motheduc
#one way describes all data:
describe(my_data)
```

A psych: 15 × 13

	vars <int>	n <dbl>	mean <dbl>	sd <dbl>	median <dbl>	trimmed <dbl>	mad <dbl
faminc	1	767	31.88461538	18.0280112	27.5000000	30.85121951	14.82
cigtax	2	767	19.64276402	7.8620297	20.0000000	19.76747967	8.895
cigprice	3	767	130.59648027	10.2946554	132.6999969	130.60178873	9.340
bwght	4	767	119.37548892	19.5227610	120.0000000	119.73333333	17.79
fatheduc	5	767	13.20860495	2.7394762	12.0000000	13.26341463	2.965
motheduc	6	767	13.11342894	2.4662557	12.0000000	13.14634146	1.482
parity	7	767	1.58409387	0.8445231	1.0000000	1.42764228	0.000
male	8	767	0.52281617	0.4998051	1.0000000	0.52845528	0.000
white	9	767	0.84615385	0.3610366	1.0000000	0.93170732	0.000
cigs	10	767	1.88657106	5.4470069	0.0000000	0.32682927	0.000
lbwght	11	767	4.76714934	0.1823095	4.7874918	4.78012814	0.152
bwghtlbs	12	767	7.46096806	1.2201726	7.5000000	7.48333333	1.111
packs	13	767	0.09432855	0.2723503	0.0000000	0.01634146	0.000
lfaminc	14	767	3.25909216	0.7360736	3.3141861	3.33299788	0.645
indx	15	767	0.31652593	0.1781760	0.3198221	0.31640191	0.227

[5]:
```
#to describe only a subset of the variables in the data:
data2<-cbind(my_data$bwght,my_data$parity,my_data$cigs,my_data$faminc,my_data$fatheduc,my_data
##Renaming first four columns columns
colnames(data2) <- c("bwght", "parity", "cigs", "faminc", "fatheduc",␣
 ↪"motheduc")
```

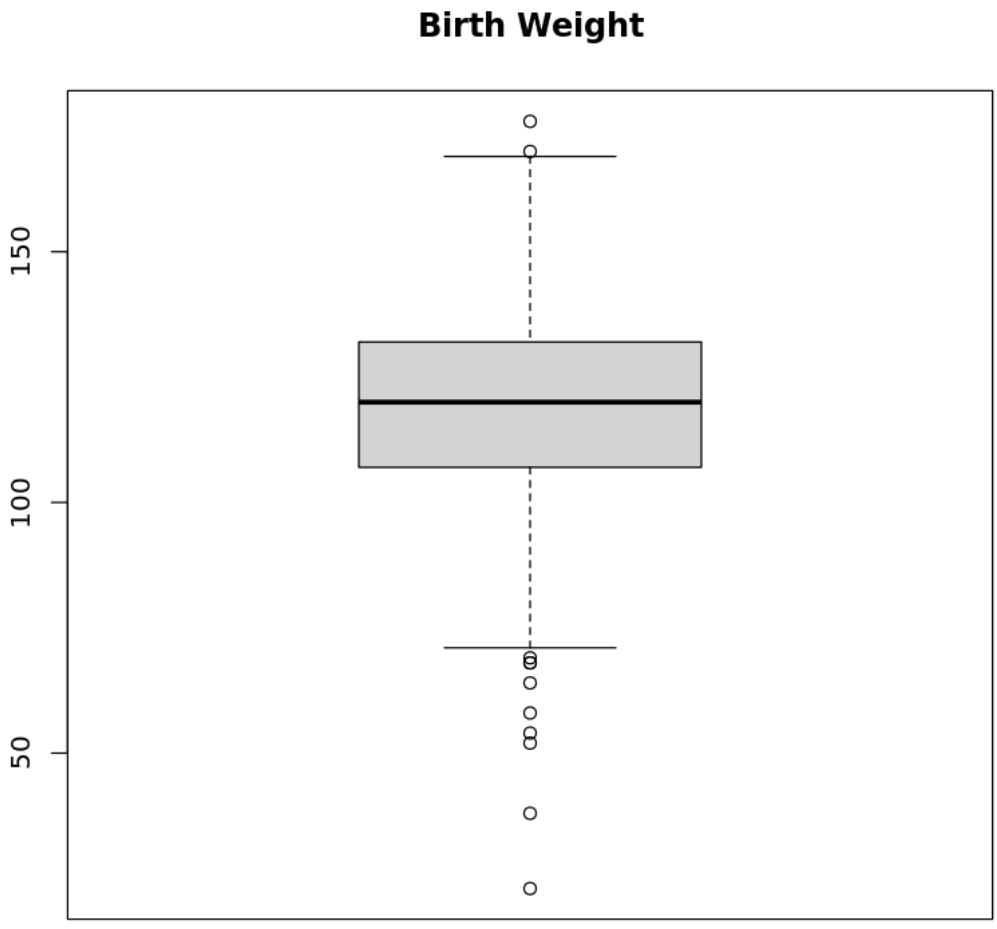
```
describe(data2)
```

A psych: 6 × 13

	vars <int>	n <dbl>	mean <dbl>	sd <dbl>	median <dbl>	trimmed <dbl>	mad <dbl>	min <db
bwght	1	767	119.375489	19.5227610	120.0	119.7333333	17.7912	23.0
parity	2	767	1.584094	0.8445231	1.0	1.4276423	0.0000	1.0
cigs	3	767	1.886571	5.4470069	0.0	0.3268293	0.0000	0.0
faminc	4	767	31.884615	18.0280112	27.5	30.8512195	14.8260	0.5
fatheduc	5	767	13.208605	2.7394762	12.0	13.2634146	2.9652	2.0
motheduc	6	767	13.113429	2.4662557	12.0	13.1463415	1.4826	4.0

box plot of birthweight of babies

[6]:
```
#box plot of birth Weight
boxplot(my_data$bwght, main="Birth Weight" )
# box plot for 'bweight above'
```



TESTING FOR q=2 restrictions on parameters of a linear regression model

[7]:
```
#use F test
#get SSR of the unrestricted model, several things are saved in reg12u
# a list of 12 things actually, see the Golbal envinoment window on the right␣
 ↪near reg12u


#regression unrestricted model
reg12u <- lm(bwght~cigs + faminc + motheduc + fatheduc + parity, my_data)
#show output
summary(reg12u)
```

```
Call:
lm(formula = bwght ~ cigs + faminc + motheduc + fatheduc + parity,
    data = my_data)

Residuals:
    Min      1Q  Median      3Q     Max
-95.701 -11.902   0.471  11.530  60.392

Coefficients:
             Estimate Std. Error t value             Pr(>|t|)
(Intercept) 114.11003    4.51768  25.259 < 0.0000000000000002 ***
cigs         -0.64010    0.13182  -4.856          0.00000146 ***
faminc        0.02761    0.04448   0.621              0.53488
motheduc     -0.59267    0.38908  -1.523              0.12811
fatheduc      0.72225    0.34586   2.088              0.03710 *
parity        2.41438    0.82393   2.930              0.00349 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 19.09 on 761 degrees of freedom
Multiple R-squared:  0.04965,        Adjusted R-squared:  0.0434
F-statistic: 7.951 on 5 and 761 DF,  p-value: 0.0000002572
```

[8]:
```
# display the SSRU
sum(reg12u$residuals^2)
```

277457.11631579

do the restricted regresison now and get SSR resticted

[9]:
```
#regression restricted model
reg12r<-lm( bwght ~ cigs + faminc +parity,my_data)
#show output
summary(reg12r)
```

```
Call:
lm(formula = bwght ~ cigs + faminc + parity, data = my_data)

Residuals:
    Min      1Q  Median      3Q     Max
-95.694 -12.098   0.478  11.900  57.306

Coefficients:
             Estimate Std. Error t value             Pr(>|t|)
(Intercept) 115.21941    2.00477  57.473 < 0.0000000000000002 ***
cigs         -0.62879    0.12923  -4.866          0.00000139 ***
faminc        0.04288    0.03905   1.098              0.27246
parity        2.50932    0.82069   3.058              0.00231 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 19.13 on 763 degrees of freedom
Multiple R-squared:  0.04397,        Adjusted R-squared:  0.04021
F-statistic:  11.7 on 3 and 763 DF,  p-value: 0.0000001686
```

[10]:
```
#display the restricted SSR   = SSRR
sum(reg12r$residuals^2)
```

279114.544350498

[11]:
```
#the parts we need for the F test are:

SSRr<-sum(reg12r$residuals^2)
SSRu<-sum(reg12u$residuals^2)
dfu<-reg12u$df.residual
```

[12]:
```
#compute the F statistic, call it F1
q<-2
F1<-(SSRr-SSRu)/q
F1<-F1/(SSRu/dfu)
F1

#to construct F stat get SSR u and SSR r and use formula
#given that se*se=SSR/(N-K-1)
#then To get SSR= se*se*(N-K-1)
#where q=# restrictions;
```

```
#  N-k-1 = Degrees of freedom unrestricted model
#N = # observations
#  K = # explanatory variables

# F stat=
# =( Rr-  u)/    divided by   (  w( - -1))
```

2.27296879453079

We get an F1=2.273, from above statistic constructed using the SSR formula.

Alternatively, use R squared to compute the F stat value for your test, call this one F2. F2=F1, see that below and compare to F1 above you computed before.

[13]:
```
#get R squared unrestricted
summary(reg12u)
# get R square from the output
r2u<-0.04965
```

```
Call:
lm(formula = bwght ~ cigs + faminc + motheduc + fatheduc + parity,
    data = my_data)

Residuals:
    Min      1Q  Median      3Q     Max
-95.701 -11.902   0.471  11.530  60.392

Coefficients:
             Estimate Std. Error t value             Pr(>|t|)
(Intercept) 114.11003    4.51768  25.259 < 0.0000000000000002 ***
cigs         -0.64010    0.13182  -4.856              0.00000146 ***
faminc        0.02761    0.04448   0.621                 0.53488
motheduc     -0.59267    0.38908  -1.523                 0.12811
fatheduc      0.72225    0.34586   2.088                 0.03710 *
parity        2.41438    0.82393   2.930                 0.00349 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 19.09 on 761 degrees of freedom
Multiple R-squared:  0.04965,        Adjusted R-squared:  0.0434
F-statistic: 7.951 on 5 and 761 DF,  p-value: 0.0000002572
```

[14]:
```
#get R squared restricted
summary(reg12r)
r2r<-0.04397
```

```
Call:
lm(formula = bwght ~ cigs + faminc + parity, data = my_data)

Residuals:
    Min      1Q  Median      3Q     Max
-95.694 -12.098   0.478  11.900  57.306

Coefficients:
             Estimate Std. Error t value             Pr(>|t|)
(Intercept) 115.21941    2.00477  57.473 < 0.0000000000000002 ***
cigs         -0.62879    0.12923  -4.866          0.00000139 ***
faminc        0.04288    0.03905   1.098              0.27246
parity        2.50932    0.82069   3.058              0.00231 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 19.13 on 763 degrees of freedom
Multiple R-squared:  0.04397,        Adjusted R-squared:  0.04021
F-statistic:  11.7 on 3 and 763 DF,  p-value: 0.0000001686
```

[15]:
```
# compute the F using the Rsquared version formula
topF2<-(r2u-r2r)/2
bottomF2<-(1-r2u)/dfu
F2<-topF2/bottomF2
F2
```

2.27415162834745

[16]:
```
#you see that F2=2.274
#just like F1 was, these are two alternative ways to et the F stat value given␣
 ↪your estimates

#get critical values for certain significance levels 5% or 10%

# decide reject null if F>c  or cannot reject null if F<c

# conclude

#see lecture notes for interpretation

#--------------------
# or get R^2 or R-squared for R and Unrestr UR and use formula also
# =(( _ ^2- _ ^2 )/)/((1- _ ^2 )/( - -1))


#get critical values for certain significance levels 5% or 10%
```

```
# decide reject null if F>c  or cannot reject null if F<c

# conclude
```

TESTING FOR LINEAR COMBIN OF PARAMETERS

[17]:
```
my_data$toteduc<-my_data$motheduc+my_data$fatheduc

reg12r2<-lm(bwght~ cigs+faminc+toteduc+parity,my_data)
summary(reg12r2)


#SSRU
sum(reg12u$residuals^2)

#SSRR
sum(reg12r2$residuals^2)


#construct F with SSR u and compare to critical value. see lecture notes
```

```
Call:
lm(formula = bwght ~ cigs + faminc + toteduc + parity, data = my_data)

Residuals:
    Min      1Q  Median      3Q     Max
-95.510 -11.965   0.396  11.794  57.147

Coefficients:
             Estimate Std. Error t value             Pr(>|t|)
(Intercept) 112.57340    4.46260  25.226 < 0.0000000000000002 ***
cigs         -0.61311    0.13141  -4.666          0.00000364 ***
faminc        0.02865    0.04456   0.643              0.52052
toteduc       0.11436    0.17230   0.664              0.50706
parity        2.54728    0.82298   3.095              0.00204 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 19.13 on 762 degrees of freedom
Multiple R-squared:  0.04452,        Adjusted R-squared:  0.03951
F-statistic: 8.877 on 4 and 762 DF,  p-value: 0.0000005227
```

277457.11631579

278953.271954953

and given the above SSRu=277457.11631579 and SSRr=278953.271954953 get the F for these

restrictions like we did above in the method F1. see slides for solutions.

Or, alternatively, use the canned package to test as below

[18]:
```
#linear restriction hypothesis testing
#make sure you have installed car package

linearHypothesis(reg12u, c("motheduc=0", "fatheduc= 0"))
#see lecture notes for interpretation

#end birthweight analysis
```

A anova: 2 × 6

	Res.Df <dbl>	RSS <dbl>	Df <dbl>	Sum of Sq <dbl>	F <dbl>	Pr(>F) <dbl>
1	763	279114.5	NA	NA	NA	NA
2	761	277457.1	2	1657.428	2.272969	0.1037048

you see that the F is 2.27 and pvalue is 0.103 so we cannot reject at 10 percent the hypothesis we specified. This is the end of the birthweight notebook portion.

2 Now you will switch to the new dataset on college attendance and wages.

Load the college data set and do the assignment yourself using the code provided on bcourses if needed. try first to code yourself

[ ]:
```
#load Lecture12twoyear.dta
```

Type the commands to regress the unrestricted model, log wage on jc, univ and experience and show the summary of the regression below

[ ]:
```

```

Type the command to test that the jc coeff equals the univ coeff

[ ]:
```

```

[19]:
```
#read in a Stata dataset
my_data <- read_dta("Lecture12twoyear.dta")
head(my_data)

#summary stats of all data
#one way describes all data:
describe(my_data)

#to describe only a subset
data2<-cbind(my_data$exper,my_data$jc,my_data$univ,my_data$lwage)

##Renaming first four columns columns
```

```r
colnames(data2) <- c("exper", "jc", "univ", "lwage")
describe(data2)


#TESTING FOR q=2 restrictions on parameters of a linear regression model

#regression unrestricted model
reg12college1<-lm( lwage~ jc + univ + exper, my_data)
#show
summary(reg12college1)

#slide
#how to test that param jc equal param univ?
linearHypothesis(reg12college1, "jc = univ")
```

A tibble: 6 × 23

female <dbl>	phsrank <dbl>	BA <dbl>	AA <dbl>	black <dbl>	hispanic <dbl>	id <dbl>	exper <dbl>	jc <dbl>	univ <dbl>
1	65	0	0	0	0	19	161	0.0000000	0.00000
1	97	0	0	0	0	93	119	0.0000000	7.03333
1	44	0	0	0	0	96	81	0.0000000	0.00000
1	34	0	0	0	1	119	39	0.2666667	0.00000
1	80	0	0	0	0	132	141	0.0000000	0.00000
0	59	0	0	0	0	156	165	0.0000000	0.00000

A psych: 23 × 13

	vars	n	mean	sd	median	trimmed
	<int>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
female	1	6763	0.51959190	0.4996530	1.0000000	0.52448716
phsrank	2	6763	56.15703090	24.2729635	50.0000000	56.79892811
BA	3	6763	0.30652077	0.4610827	0.0000000	0.25817779
AA	4	6763	0.04406329	0.2052509	0.0000000	0.00000000
black	5	6763	0.09507615	0.2933418	0.0000000	0.00000000
hispanic	6	6763	0.04687269	0.2113818	0.0000000	0.00000000
id	7	6763	40615.72319976	24980.6323852	39301.0000000	39768.2973572
exper	8	6763	122.38163537	33.4279875	129.0000000	126.09037147
jc	9	6763	0.33889456	0.7721268	0.0000000	0.12121914
univ	10	6763	1.92627423	2.2970005	0.1999997	1.63807036
lwage	11	6763	2.24809573	0.4876918	2.2763002	2.25698553
stotal	12	6763	0.04748291	0.8535441	0.0000000	0.08965340
smcity	13	6763	0.28537631	0.4516269	0.0000000	0.23175014
medcity	14	6763	0.11740352	0.3219243	0.0000000	0.02180743
submed	15	6763	0.06860861	0.2528061	0.0000000	0.00000000
lgcity	16	6763	0.09448470	0.2925235	0.0000000	0.00000000
sublg	17	6763	0.08709153	0.2819900	0.0000000	0.00000000
vlgcity	18	6763	0.05855390	0.2348052	0.0000000	0.00000000
subvlg	19	6763	0.06358125	0.2440235	0.0000000	0.00000000
ne	20	6763	0.21070531	0.4078396	0.0000000	0.13842173
nc	21	6763	0.29883188	0.4577798	0.0000000	0.24856773
south	22	6763	0.32707378	0.4691791	0.0000000	0.28386620
totcoll	23	6763	2.26516879	2.3302019	1.5066650	2.03939386

A psych: 4 × 13

	vars	n	mean	sd	median	trimmed	mad
	<int>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
exper	1	6763	122.3816354	33.4279875	129.0000000	126.0903715	32.6172000
jc	2	6763	0.3388946	0.7721268	0.0000000	0.1212191	0.0000000
univ	3	6763	1.9262742	2.2970005	0.1999997	1.6380704	0.2965196
lwage	4	6763	2.2480957	0.4876918	2.2763002	2.2569855	0.4959362

```
Call:
lm(formula = lwage ~ jc + univ + exper, data = my_data)

Residuals:
     Min       1Q   Median       3Q      Max
-2.10362 -0.28132  0.00551  0.28518  1.78167

Coefficients:
             Estimate Std. Error t value            Pr(>|t|)
(Intercept) 1.4723256  0.0210602  69.910 <0.0000000000000002 ***
jc          0.0666967  0.0068288   9.767 <0.0000000000000002 ***
univ        0.0768762  0.0023087  33.298 <0.0000000000000002 ***
exper       0.0049442  0.0001575  31.397 <0.0000000000000002 ***
---
```

```
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4301 on 6759 degrees of freedom
Multiple R-squared:  0.2224,        Adjusted R-squared:  0.2221
F-statistic: 644.5 on 3 and 6759 DF,  p-value: < 0.00000000000000022
```

A anova: 2 × 6

	Res.Df <dbl>	RSS <dbl>	Df <dbl>	Sum of Sq <dbl>	F <dbl>	Pr(>F) <dbl>
1	6760	1250.942	NA	NA	NA	NA
2	6759	1250.544	1	0.3985339	2.154016	0.142244

#ALTERNATIVELY

#slides of Lecture 12 notes

#regression restricted model such that a parameter is already the tested object,

#test coeff of beta_univ=0 is the null of whether univ and jc have similar returns on wages

create totcollege as the sum of jc and univ and add to the dataframe

reg12_college2<- lm(lwage ~ totcollege+ univ+ exper, my_data)

summary(reg12_college2)

what do you interpret when you type these commands below given the output, see slides for a check.