MEM5220 - Microeconometrics Self-Evaluation 1

Taltech - DEF

YOUR NAME HERE

04 April, 2021

Preface

This first R econometrics self-evaluation assignment is focused on data cleaning, data manipulation, plotting and estimating and interpreting simple linear regression models.
Some packages I have used to solve the exercises:

```
library(tidyverse)
library(modelsummary)
library(broom)
library(lmtest)
library(sandwich)
library(car)
```

You can use **any** additional packages for answering the questions.

Note:

- This assignment has to be solved in this R Markdown document and you should be able to "knit" the document without errors.
- Fill out your name in "yaml" - block on top of this document
- Use the R markdown syntax:
  - Write your code in code chunks
  - Write your explanations including the equations in markdown syntax
- If you have an error in your code use `#` to comment the line out where the error occurs but do not delete the code itself. I want to see your coding errors so I can give feedback!

For more information on using R Markdown for class exercises see https://ntaback.github.io/UofT_STA130/Rmarkdownforclassreports.html

---

You will be working with the dataset `Caschool` from the **Ecdat** package, the dataset `Wage1` from the **wooldrige** package. In the last two exercises, you will be working with a simulated data.

Try to answer by questions by including a code chunk and then a written answer. The answer to question 1.1 serves as a template. Please proceed with the rest of the questions in a similar way.

You can use any plotting package but the figures should have a research project style quality (eg. axis labels, figure legend, figure title and if necessary figures notes).

Caschool exercises ## Question: Load the dataset `Caschool` from the **Ecdat** package.

The Caschool-dataset contains the average test scores of 420 elementary schools in California along with some additional information. `r # install.packages("Ecdat") library("Ecdat") data("Caschool", package = "Ecdat")` ## Question:

What are the dimensions of the `Caschool` dataset?

A:

```r
dim(Caschool)
[1] 420 17
```

The dataset has 420 rows and 17 columns.

Question: Does the `Caschool` dataset contain missing observations?

A:

```r
sum(is.na(Caschool))
[1] 0
```

There are no missing observations in the dataset. ## Question: Display the structure of the `Caschool` dataset. Which variable are encoded as factors? **A**: `r str(Caschool)`

```
'data.frame':
420
obs.
of
17
variables:
$
distcod
:
int
75119
61499
61549
61457
61523
62042
68536
63834
62331
67306
...
$
county
:
Factor
w/
45
levels
"Alameda","Butte",..:
1 2
2 2
2 6
29
11 6
25
...
$
district:
Factor
w/
409
levels
"Ackerman
Elementary",..:
362
214
367
132
270
```

County, district and grspan are encoded as factors. ## Question: Provide a summary statistic of the data. **A**: `r summary(Caschool)`

"'
dist-
cod
county
dis-
trict
grspan
Min.
:61382
Sonoma
: 29
Lake-
side
Union
Ele-
men-
tary:
3
KK-
06:
61
1st
Qu.:64308
Kern
: 27
Moun-
tain
View
Ele-
men-
tary :
3 KK-
08:359
Me-
dian
:67760
Los
Ange-
les:
27
Jef-
fer-
son
Ele-
men-
tary :
2
Mean
:67473

computer
testscr
comp-
stu
expn-
stu
Min.
: 0.0
Min.
:605.5
Min.
:0.00000
Min.
:3926
1st
Qu.:
46.0
1st
Qu.:640.0
1st
Qu.:0.09377
1st
Qu.:4906
Me-
dian :
117.5
Me-
dian
:654.5
Me-
dian
:0.12546
Me-
dian
:5215
Mean
:
303.4
Mean
:654.2
Mean
:0.13593
Mean
:5312
3rd
Qu.:
375.2
3rd
Qu.:666.7

str
avginc
elpct
read-
scr
Min.
:14.00
Min.
:
5.335
Min.
:
0.000
Min.
:604.5
1st
Qu.:18.58
1st
Qu.:10.639
1st
Qu.:
1.941
1st
Qu.:640.4
Me-
dian
:19.72
Me-
dian
:13.728
Me-
dian :
8.778
Me-
dian
:655.8
Mean
:19.64
Mean
:15.317
Mean
:15.768
Mean
:655.0
3rd
Qu.:20.87
3rd
Qu.:17.629
3rd

mathscr
Min.
:605.4
1st
Qu.:639.4
Me-
dian
:652.5
Mean
:653.3
3rd
Qu.:665.9
Max.
:709.5
“‘
##
Ques-
tion:
What
are
the
names
of
the
vari-
ables
in
the
dataset?
A:
`r`
`names(Caschool)`

[1] "distcod" "county" "district" "grspan" "enrltot" "teachers" [7] "calwpct" "mealpct" "computer" "testscr" "compstu" "expnstu" [13] "str" "avginc" "elpct" "readscr" "mathscr"

Question: How many unique observations are available in the variable "county"

A: r unique(Caschool$county)

[1]
Alameda
Butte
Fresno
San
Joaquin
[5]
Kern
Sacramento
Merced
Tulare
[9]
Los
Angeles
Imperial
Monterey
San
Diego
[13]
San
Bernardino
San
Mateo
Ventura
Riverside
[17]
Santa
Clara
Madera
Santa
Barbara
Orange
[21]
Kings
Sonoma
Contra
Costa
Humboldt
[25]
Siskiyou
Lake
Sutter
Mendocino
[29]
San
Benito
Shasta
Tehama
Stanislaus

Question: Summarize the mean number of students grouped by county. **A**:

```r
mean_countCaschool <- Caschool %>% group_by(county) %>% summarise(mean_count = mean(enrltot)) %>% arrange(desc(mean_count))
mean_countCaschool
```

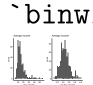
```
# A
tibble:
45 x
2
county
mean_count
<fct>
<dbl>
1
Orange
8224.
2
San
Bernardino
6470.
3
San
Diego
6170.
4
Santa
Clara
5932.
5
Los
Angeles
5831.
6
Ventura
4628.
7
Monterey
3549.
8
Sacramento
3511.
9
San
Mateo
3289.
10
Kern
3108.
#
...
with
35
more
rows
```

Question:

Calculate the log of average income from of the Caschool dataset. Call the variable **logavginc** and add this variable to the dataset. Then, plot a histogram of the average income vs. a histogram of log average income. What do you observe?

A:

```
r
Caschool$logavginc
<-
log(Caschool$avginc)
r
library(patchwork)
library(ggpubr)
p1
<-
ggplot(Caschool,
aes(avginc))
+
geom_histogram(show.legend
=
FALSE)
+
labs(title="Average
income",
x
="Avg.
income")+
theme_pubr()
p2
<-
ggplot(Caschool,
aes(logavginc))
+
geom_histogram(show.legend
=
FALSE)
+
labs(title="Average
income",
x
="Log
Avg.
income")+
theme_pubr()
```

r
patchwork
<-
(p1
+
p2)
patchwork
`stat_bin()`
using
`bins
=
30`.
Pick
better
value
with
`binwidth`.
`stat_bin()`
using
`bins
=
30`.
Pick
better
value
with
`binwidth`.


Average income is clearly leftward-skewed. The log of averge income looks more like a normal distribution. ## Question:

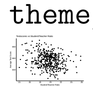
We want to create now a subset of counties that have the ten highest district average income and that have the ten lowest district average income. Call this subset *Caschool_lowhighincome*.

Hint: One way is the create two subsets (eg. Cascholl_highincome and Caschool_lowincome and the use the `rbind()` function to bind them together.). **A**: “‘r Caschool_highincome <- Caschool %>% arrange(desc(avginc)) %>% head(10) Caschool_lowincome <- Caschool %>% arrange((avginc)) %>% head(10)

Caschool_lowhighincome
<-
rbind(Caschool_highincome,Caschool_highincome)
“‘
##
Ques-
tion:

Let us test wether a high student/teacher ratio will be associated with higher-than-average test scores for the school? Create a scatter plot for the full dataset (*Caschool*) for the variables **testscr** and **str**.
A:

```
r
ggplot(mapping
=
aes(x
=
str,
y =
testscr),
data
=
Caschool)
+ #
base
plot
geom_point()
+ #
add
points
scale_y_continuous(name
=
"Average
Test
Score")
+
scale_x_continuous(name
=
"Student/Teacher
Ratio")
+
labs(title="Testscores
vs
Student/Teacher
Ratio")+
theme_pubr()
```



Question:

Suppose a policymaker is interested in the following linear model:

$$testscr = \beta_0 + \beta_1 str + u \quad (1)$$

Where *testscr* is the average test score for a given school and *str* is the Student/Teacher Ratio (i.e. the average number of students per teacher).

Estimate the specified linear model. Is the estimated relationship between a schools Student/Teacher Ratio and its average test results positive or negative?
A:

```r
fit_single <- lm(formula = testscr ~ str, data = Caschool)
summary(fit_single)
```

Call:
lm(formula = testscr ~ str, data = Caschool)
Residuals:
Min 1Q Median 3Q Max -47.727 -14.251 0.483 12.822 48.540

Coefficients:	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	698.9330	9.4675	73.825	< 2e-16
str	***-2.2798***	***0.4798***	***-4.751***	***2.78e-06***

Signif. codes: 0 '' ***0.001*** '' *0.01* '' 0.05 '.' 0.1 ' ' 1

Residual standard error: 18.58 on 418 degrees of freedom Multiple R-squared: 0.05124, Adjusted R-squared: 0.04897 F-statistic: 22.58 on 1 and 418 DF, p-value: 2.783e-06

~~~~
## Question:

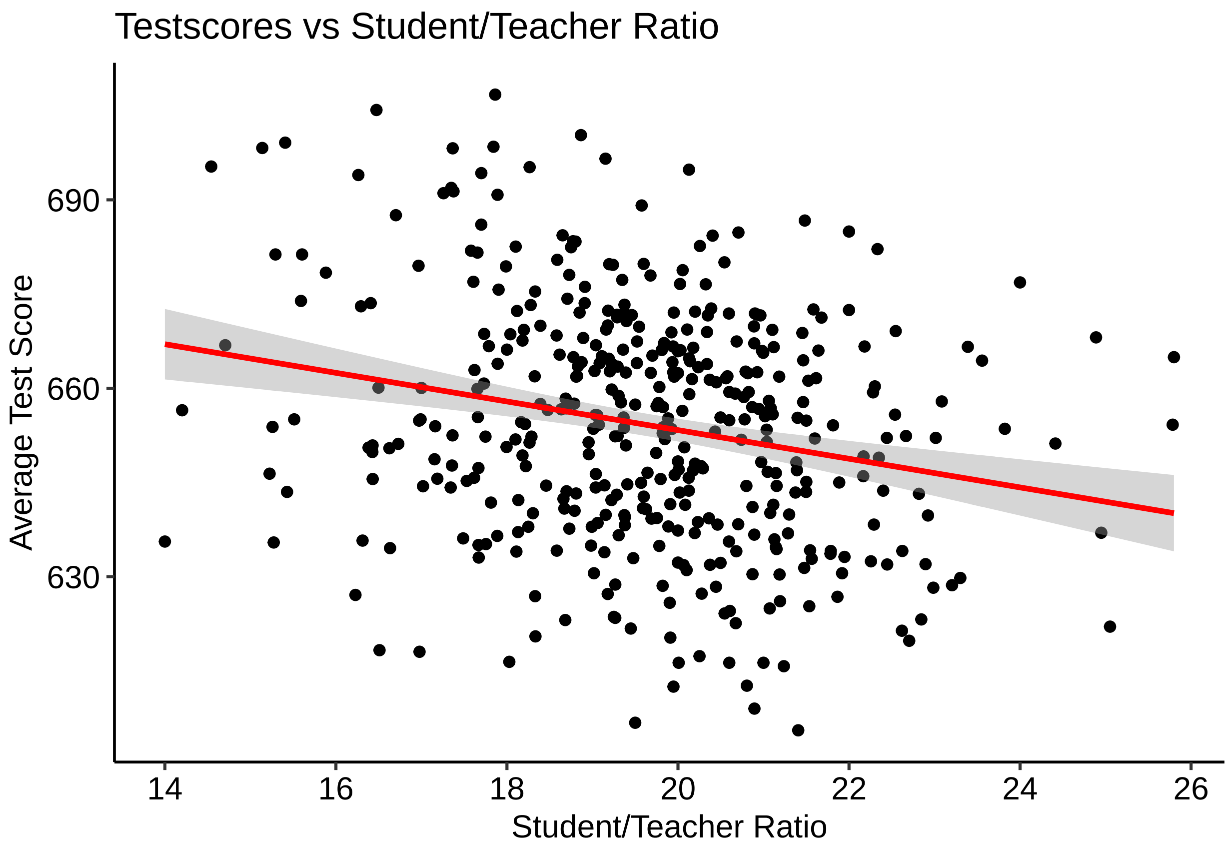
Now, plot the regression line for the model we have just estimated.

**A**:


```r
ggplot(mapping = aes(x = str, y = testscr), data = Caschool) + # base plot
  geom_point() + # add points
  geom_smooth(method = "lm", size=1, color="red") + # add regression line
~~~~

```
  scale_y_continuous(name = "Average Test Score") +
  scale_x_continuous(name = "Student/Teacher Ratio") +
  labs(title="Testscores vs Student/Teacher Ratio")+
  theme_pubr()
```

```
`geom_smooth()` using formula 'y ~ x'
```



0.1 Question:

Let us extend our example of student test scores by adding families' average income to our previous model:

$$testscr = \beta_0 + \beta_1 str + \beta_2 avginc + u \tag{2}$$

A:

```r
fit_multivariate <- lm(formula = "testscr ~ str + avginc", data = Caschool)
summary(fit_multivariate)
```

```
Call:
lm(formula = "testscr ~ str + avginc", data = Caschool)

Residuals:
```

```
    Min      1Q  Median      3Q     Max 
-39.608  -9.052   0.707   9.259  31.898 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 638.72915    7.44908  85.746   <2e-16 ***
str          -0.64874    0.35440  -1.831   0.0679 .  
avginc        1.83911    0.09279  19.821   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 13.35 on 417 degrees of freedom
Multiple R-squared:  0.5115,    Adjusted R-squared:  0.5091 
F-statistic: 218.3 on 2 and 417 DF,  p-value: < 2.2e-16
```

Adding the explanatory variable "avginc" to the model, the estimated coefficient of the student/ teacher ratio becomes first smaller compared to the previous model and second insignificant at conventional levels.

0.2 Question:

Assume know that "str" depends also on the value of yet another regressor, "avginc". Estimate the following model. Compare the sign of the estimate of β_2 and β_3. Interpret the results.

$$testscr = \beta_0 + \beta_1 str + \beta_2 avginc + \beta_3 (str \times avginc) + u \tag{3}$$

A:

```
fit_inter = lm(formula = testscr ~ str + avginc + str*avginc, data = Caschool)
summary(fit_inter)
```

```
Call:
lm(formula = testscr ~ str + avginc + str * avginc, data = Caschool)

Residuals:
    Min      1Q  Median      3Q     Max 
-41.346  -9.260   0.209   8.736  33.368 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 689.47473   14.40894  47.850  < 2e-16 ***
str          -3.40957    0.75980  -4.487 9.34e-06 ***
avginc       -1.62388    0.85214  -1.906   0.0574 .  
str:avginc    0.18988    0.04646   4.087 5.24e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

	Model 1	Model 2	Model 3
(Intercept)	698.933***	638.729***	689.475***
	(9.467)	(7.449)	(14.409)
str	-2.280***	-0.649*	-3.410***
	(0.480)	(0.354)	(0.760)
avginc		1.839***	-1.624*
		(0.093)	(0.852)
str × avginc			0.190***
			(0.046)
Num.Obs.	420	420	420
R2	0.051	0.511	0.530
R2 Adj.	0.049	0.509	0.527
AIC	3650.5	3373.7	3359.2
BIC	3662.6	3389.9	3379.4
Log.Lik.	-1822.250	-1682.856	-1674.587
F	22.575	218.302	156.585

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

```
Residual standard error: 13.1 on 416 degrees of freedom
Multiple R-squared:  0.5303,    Adjusted R-squared:  0.527
F-statistic: 156.6 on 3 and 416 DF,  p-value: < 2.2e-16
```

We observe also that the estimate of β_2 changes signs and becomes negative, while the interaction effect β_3 is positive.

This means that an increase in str reduces average student scores (more students per teacher make it harder to teach effectively); that an increase in average district income in isolation actually reduces scores; and that the interaction of both increases scores (more students per teacher are actually a good thing for student performance in richer areas).

0.3 Question:

You have fitted 3 specifications for the Caschool example. Report the regression results of equation 1, 2 and 3, in a formatted table regression output table. Discuss the model fit and model selection.

A:

```
library(modelsummary)
modelsummary(list(fit_single, fit_multivariate, fit_inter),stars = TRUE )
```

The adjusted R^2 is highest for the model 3, the model that includes an interaction term. AIC and BIC, two widely used information criteria, would also select model 3, relative to each of the other models (The relatively quality of the model is maximized when the information criterion is minimized).

1 Wage1 excercises

Wage data: These are data from the 1976 Current Population Survey. Source of the data is Wooldrige. Familiarize yourself with the dataset if necessary.

```
# install.packages("wooldridge")
library("wooldridge")
data("wage1", package = "wooldridge")
```

1.1 Question

First, estimate the following model and test again for heteroscedasticity.

$$wage = \beta_0 + \beta_1 female + \beta_3 educ + \beta_4 exper + u \tag{4}$$

A:

```
lm3_wage1 <- lm(wage~female+educ+exper, data=wage1)
summary(lm3_wage1)
```

```
Call:
lm(formula = wage ~ female + educ + exper, data = wage1)

Residuals:
    Min      1Q  Median      3Q     Max 
-6.3856 -1.9652 -0.4931  1.1199 14.8217 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) -1.73448    0.75362  -2.302   0.0218 *  
female      -2.15552    0.27031  -7.974 9.74e-15 ***
educ         0.60258    0.05112  11.788  < 2e-16 ***
exper        0.06424    0.01040   6.177 1.32e-09 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.078 on 522 degrees of freedom
Multiple R-squared:  0.3093,    Adjusted R-squared:  0.3053 
F-statistic: 77.92 on 3 and 522 DF,  p-value: < 2.2e-16
```

```
lm3_bptest <- bptest(lm3_wage1)
lm3_bptest
```

```
	studentized Breusch-Pagan test

data:  lm3_wage1
BP = 36.904, df = 3, p-value = 4.821e-08
```

The test statistic of the BP-test is 36.9043336 and the corresponding p-value is smaller than 4.8208966×10^{-8}, so we can reject homoscedasticity for all reasonable significance levels.

```
# coeftest(lm3_wage1, vcov=hccm)
cov3 <- hccm(lm3_wage1, type="hc3") # hc3 is the standard method
lm3_robust <- coeftest(lm3_wage1, vcovHC)
lm3_robust
```

```
t test of coefficients:

             Estimate Std. Error t value  Pr(>|t|)    
(Intercept) -1.734481   0.868647 -1.9968   0.04637 *  
female      -2.155517   0.260249 -8.2825 1.020e-15 ***
educ         0.602580   0.065005  9.2697 < 2.2e-16 ***
exper        0.064242   0.010113  6.3521 4.626e-10 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

1.2 Question:

Now, estimate the following model:

$$log(wage) = \beta_0 + \beta_1(married \times female) + \beta_3 educ + \beta_4 exper + beta_5 exper^2 + \beta_6 tenure + \beta_7 tenure^2 + u \quad (5)$$

1. What is the reference group in this model?
2. Ceteris paribus, how much more wage do single males make relative to the reference group?
3. Ceteris paribus, how much more wage do single females make relative to the reference group?
4. Ceteris paribus, how much less do married females make than single females?
5. Do the results make sense economically. What socio-economic factors could explain the results?

A:

```
lm2_wage1 <- lm(log(wage)~married*female+educ+exper+I(exper^2)+tenure+I(tenure^2), data=
summary(lm2_wage1)
```

```
Call:
lm(formula = log(wage) ~ married * female + educ + exper + I(exper^2) + 
    tenure + I(tenure^2), data = wage1)
```

```
Residuals:
     Min       1Q   Median       3Q      Max 
-1.89697 -0.24060 -0.02689  0.23144  1.09197 

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)    
(Intercept)     0.3213781  0.1000090   3.213 0.001393 ** 
married         0.2126757  0.0553572   3.842 0.000137 ***
female         -0.1103502  0.0557421  -1.980 0.048272 *  
educ            0.0789103  0.0066945  11.787  < 2e-16 ***
exper           0.0268006  0.0052428   5.112 4.50e-07 ***
I(exper^2)     -0.0005352  0.0001104  -4.847 1.66e-06 ***
tenure          0.0290875  0.0067620   4.302 2.03e-05 ***
I(tenure^2)    -0.0005331  0.0002312  -2.306 0.021531 *  
married:female -0.3005931  0.0717669  -4.188 3.30e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3933 on 517 degrees of freedom
Multiple R-squared:  0.4609,    Adjusted R-squared:  0.4525 
F-statistic: 55.25 on 8 and 517 DF,  p-value: < 2.2e-16
```

```r
library(scales) #  percent
```

```
Attaching package: 'scales'

The following object is masked from 'package:purrr':

    discard

The following object is masked from 'package:readr':

    col_factor
```

```r
df_lm2_wage1 <- tidy(lm2_wage1)
# Singe male
marriedmale <- df_lm2_wage1 %>%
  filter(term == "married") %>%
  dplyr::select(estimate) %>%
  pull() # pull out the single coefficient value of the dataframe
# Single female
singlefemale <- df_lm2_wage1 %>%
  filter(term == "female") %>%
  dplyr::select(estimate) %>%
  pull() # pull out the single coefficient value of the dataframe
marriedfemale <- df_lm2_wage1 %>%
```

```
  filter(term == "married:female") %>%
  dplyr::select(estimate) %>%
  pull() # pull out the single coefficient value of the dataframe
married<- df_lm2_wage1 %>%
  filter(term == "married") %>% #
  dplyr::select(estimate) %>%
  pull() # pull out the single coefficient value of the dataframe
```

```
lm2_robust <- coeftest(lm2_wage1, vcovHC)
lm2_robust
```

```
t test of coefficients:

                  Estimate  Std. Error t value  Pr(>|t|)
(Intercept)     0.32137810  0.11151141  2.8820 0.0041157 **
married         0.21267568  0.05816944  3.6561 0.0002822 ***
female         -0.11035021  0.05785288 -1.9074 0.0570190 .
educ            0.07891028  0.00757178 10.4216 < 2.2e-16 ***
exper           0.02680057  0.00520490  5.1491 3.727e-07 ***
I(exper^2)     -0.00053525  0.00010817 -4.9480 1.016e-06 ***
tenure          0.02908752  0.00737872  3.9421 9.190e-05 ***
I(tenure^2)    -0.00053314  0.00027213 -1.9591 0.0506349 .
married:female -0.30059307  0.07328034 -4.1020 4.758e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

1. Reference group: *single* and *male*
2. Cp. married males make 21% (`percent(marriedmale)`) more than single males.
3. Cp. a single female makes -11% (`percent(singlefemale)`) less than the reference group.
4. Married females make 9% (`percent(abs(marriedfemale) - abs(married))`) less than single females.
5. There seems to be a marriage premium[1] for men but for women the marriage premium is negative.

1.3 Question:

Test for heteroscedasticity test in the estimated regression of the wage1 dataset. Do we reject homoscedasticity for all reasonable significance levels? Adjust for heteroscedasticity by using refined White heteroscedasticity-robust SE.

A:

[1]There is clearly a correlation between men having children and men getting higher salaries, and the reverse for women. However, this may reflect the fact that women are more likely to withdraw from work to take care of children (regardless of whether they'd prefer to), and men may double down on work.

```
bptest(lm2_wage1)
```

```
	studentized Breusch-Pagan test

data:  lm2_wage1
BP = 13.189, df = 8, p-value = 0.1055
```

We do not reject the null hypothesis at conventional significance levels.

1.4 Question

Create a regression table showing the results from equation 4 and 5. Show a specification where the SE have not been adjusted for heteroscedasticity and another specification where the SE have been adjusted for heteroscedasticity.

A:

```
modelsummary(list(lm3_wage1, coeftest(lm3_wage1, vcovHC), lm2_wage1, coeftest(lm2_wage1,
```

```
Original model not retained as part of coeftest object. For additional model summary inf
This message is displayed once per session.
```

2 Collinearity exercises

This exercise focuses on the **collineartiy** problem.

2.1 Question:

Run the following commands in R:

```
set.seed(1)
x1 <- runif(100)
x2 <- 0.5 * x1 + rnorm(100)/10
y <- 2 +2*x1 + 0.3 *x2 +rnorm(100)
```

The last line corresponds to creating a linear model in which y is a function of x_1 and x_2. Write out the form of the linear model. What are the regression coefficients?

A:

$y = 2 + 2x_1 + 0.3x_2 + \epsilon$

$\beta_0 = 2$, $\beta_1 = 2$, $\beta_3 = 0.3$

2.2 Question:

What is the correlation between x_1 and x_2? Create a scatterplot displaying the relationship between the variables.

	Model 1	Model 2	Model 3	Model 4
(Intercept)	-1.734**	-1.734**	0.321***	0.321***
	(0.754)	(0.869)	(0.100)	(0.112)
female	-2.156***	-2.156***	-0.110**	-0.110*
	(0.270)	(0.260)	(0.056)	(0.058)
educ	0.603***	0.603***	0.079***	0.079***
	(0.051)	(0.065)	(0.007)	(0.008)
exper	0.064***	0.064***	0.027***	0.027***
	(0.010)	(0.010)	(0.005)	(0.005)
married			0.213***	0.213***
			(0.055)	(0.058)
I(exper^2)			-0.001***	-0.001***
			(0.000)	(0.000)
tenure			0.029***	0.029***
			(0.007)	(0.007)
I(tenure^2)			-0.001**	-0.001*
			(0.000)	(0.000)
married × female			-0.301***	-0.301***
			(0.072)	(0.073)
Num.Obs.	526	526	526	526
R2	0.309		0.461	
R2 Adj.	0.305		0.453	
AIC	2681.5	2681.5	521.9	521.9
BIC	2702.8	2702.8	564.6	564.6
Log.Lik.	-1335.736	-1335.736	-250.955	-250.955
F	77.920		55.246	

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

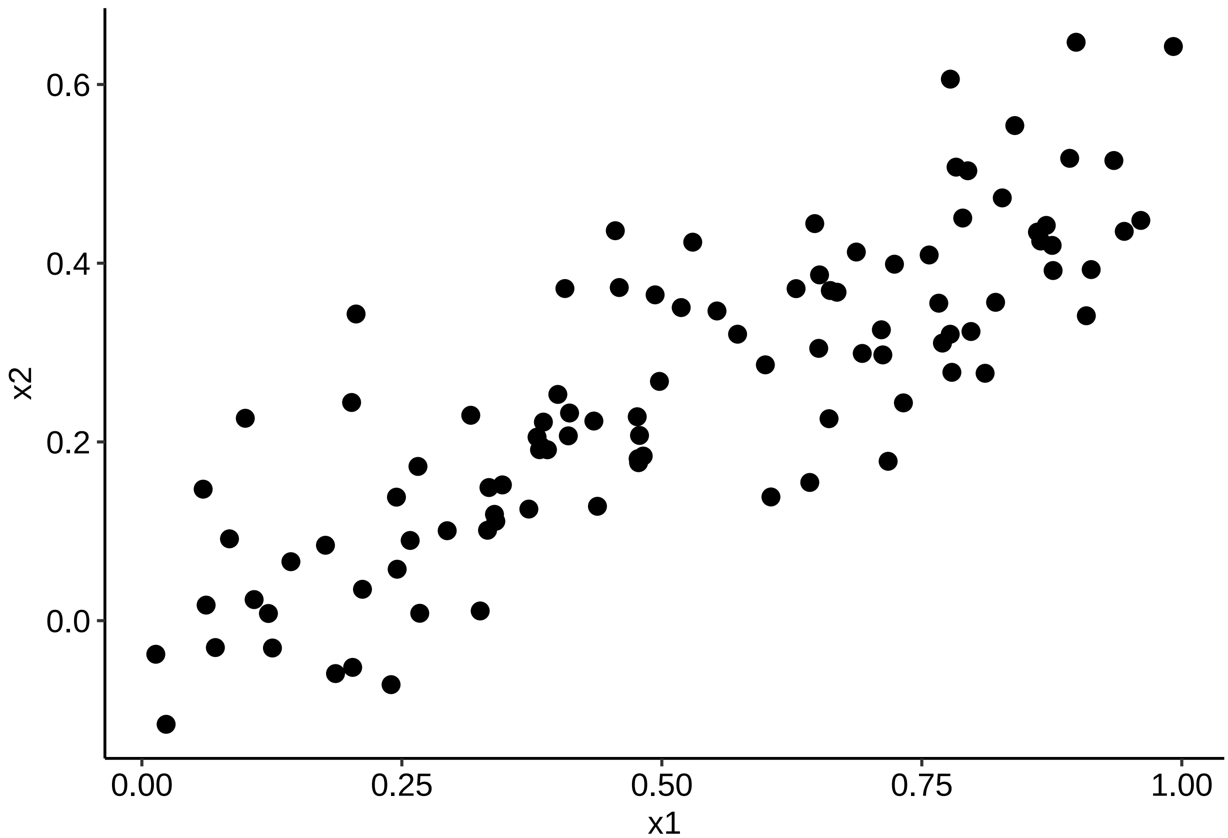
Model 2 and Mode 4 have been estimated with robust standard errors.

A:

```
cor(x1, x2)
```

```
[1] 0.8351212
```

```
d <-  data.frame(x1,x2)
ggplot(d, aes(x1, x2)) +
  geom_point(shape = 16, size = 3, show.legend = FALSE) +
  theme_pubr()
```



2.3 Question:

Using this data, fit a least squares regression to predict y using x_1 and x_2. Describe the results obtained. What are $\hat{\beta}_0$, $\hat{\beta}_1$ and $\hat{\beta}_2$? How do these relate to the true β_0, β_1 and β_2? Can you reject the null hypothesis $H_0 : \beta_1 = 0$? How about the null hypothesis $H_0 : \beta_2 = 0$?

A:

```
lm.fit = lm(y~x1+x2)
summary(lm.fit)
```

```
Call:
lm(formula = y ~ x1 + x2)
```

```
Residuals:
    Min      1Q  Median      3Q     Max 
-2.8311 -0.7273 -0.0537  0.6338  2.3359 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)   2.1305     0.2319   9.188 7.61e-15 ***
x1            1.4396     0.7212   1.996   0.0487 *  
x2            1.0097     1.1337   0.891   0.3754    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.056 on 97 degrees of freedom
Multiple R-squared:  0.2088,    Adjusted R-squared:  0.1925 
F-statistic:  12.8 on 2 and 97 DF,  p-value: 1.164e-05
```

The regression coefficients are close to the true coefficients, although with high standard error. We can reject the null hypothesis for β_1 because its p-value is below 5%. We cannot reject the null hypothesis for β_2 because its p-value is much above the 5% typical cutoff, over 60%.

2.4 Question:

Now fit least squares regression to predict y using only x_1. Comment on your results. Can you reject the null hypothesis $H_0 : \beta_1 = 0$?

A:

```
lm.fit = lm(y~x1)
summary(lm.fit)
```

```
Call:
lm(formula = y ~ x1)

Residuals:
     Min       1Q   Median       3Q      Max 
-2.89495 -0.66874 -0.07785  0.59221  2.45560 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)   2.1124     0.2307   9.155 8.27e-15 ***
x1            1.9759     0.3963   4.986 2.66e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.055 on 98 degrees of freedom
Multiple R-squared:  0.2024,    Adjusted R-squared:  0.1942 
```

```
F-statistic: 24.86 on 1 and 98 DF,  p-value: 2.661e-06
```

Yes, we can reject the null hypothesis for the regression coefficient given the p-value for its t-statistic is near zero.

2.5 Question:

Now fit least squares regression to predict y using only x_2. Comment on your results. Can you reject the null hypothesis $H_0 : \beta_2 = 0$?

A:

```r
lm.fit = lm(y~x2)
summary(lm.fit)
```

```
Call:
lm(formula = y ~ x2)

Residuals:
     Min       1Q   Median       3Q      Max 
-2.62687 -0.75156 -0.03598  0.72383  2.44890 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)   2.3899     0.1949   12.26  < 2e-16 ***
x2            2.8996     0.6330    4.58 1.37e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.072 on 98 degrees of freedom
Multiple R-squared:  0.1763,    Adjusted R-squared:  0.1679 
F-statistic: 20.98 on 1 and 98 DF,  p-value: 1.366e-05
```

Yes, we can reject the null hypothesis for the regression coefficient given the p-value for its t-statistic is near zero.

2.6 Question:

Do the results from the previous questions contradict each other? Explain your answer.

A:

No, because x_1 and x_2 have collinearity, it is hard to distinguish their effects when regressed upon together. When they are regressed upon separately, the linear relationship between y and each predictor is indicated more clearly.

3 Simulation exercises

3.1 Question:

The probability that a baby is girl or boy is approximately 48.8% or 51.2%, respectively, and these do not much very much across the world. Suppose that 400 babies are born in a hospital in a given year. How many will be girls?

Set a seed (eg. `set.seed(123)`) to make the result reproducible!

A:

```
set.seed(123)
n_girls <- rbinom(1, 400, 0.488)
n_girls
```

```
[1] 189
```

3.2 Question:

Simulate the process 1000 times and plot the distribution. Indicate the mean in the distribution plot.

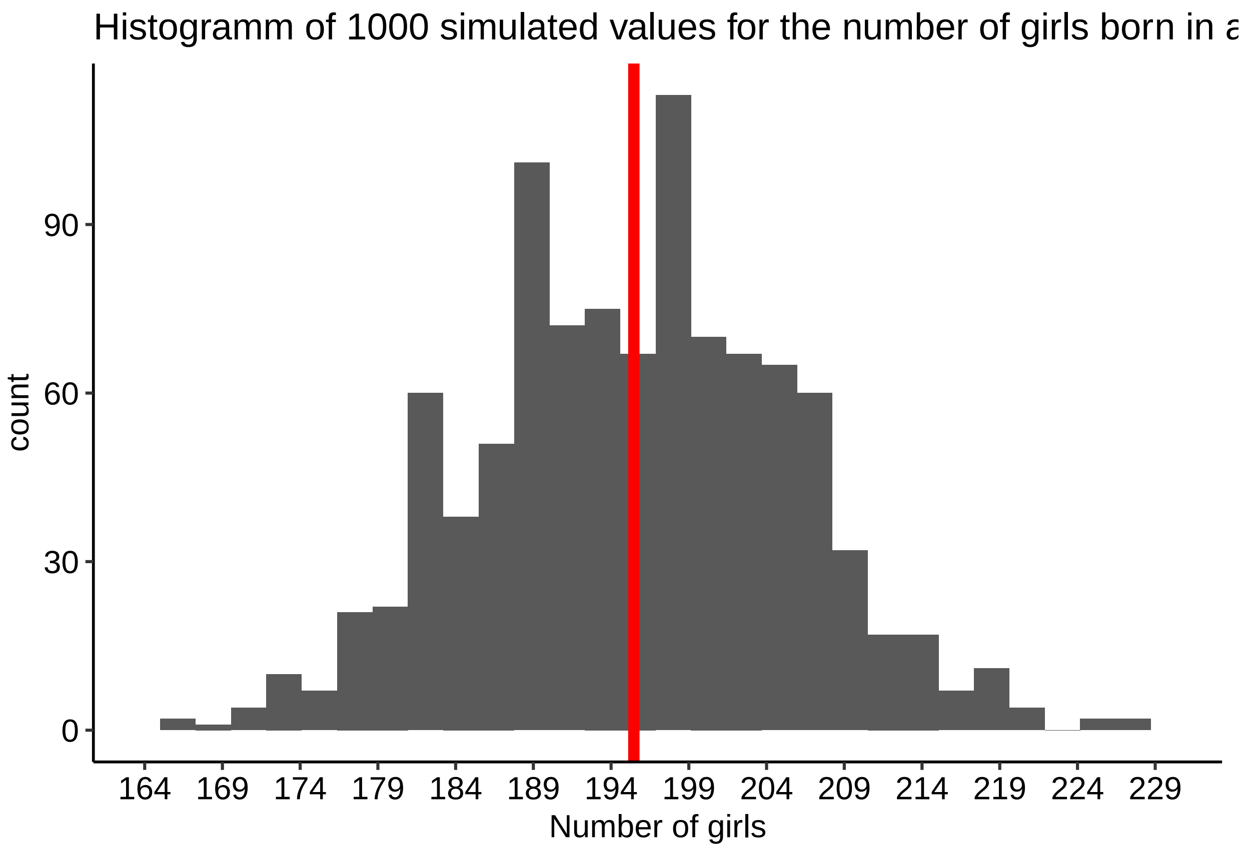
A:

```
n_sims <- 1000
n_girls <- rbinom(n_sims, 400, 0.488)
```

```
n_girls <- as.data.frame(n_girls)
ggplot(n_girls, aes(n_girls)) +
  geom_histogram(show.legend = FALSE) +
  scale_x_continuous(breaks = seq(min(n_girls), 300, 5), lim = c(min(n_girls), max(n_gir
  geom_vline(aes(xintercept = mean(n_girls)),col='red',size=2)+
  labs(title="Histogramm of 1000 simulated values for the number of girls born in a hosp
        x ="Number of girls")+
  theme_pubr()
```

```
`stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

```
Warning: Removed 2 rows containing missing values (geom_bar).
```



3.3 Question:

In the previous exercise we simulated a discrete probability model. Now, we will simulate a mixed discrete/ continuous model.

In the United States 52% of the adults are women and 48% are men. The heights of the men are approximately normally distributed wit mean 69.1 inches and standard deviation 2.9. Women have a mean height of 63.7 inches and a standard deviation of 2.7.

Generate the height of one randomly chosen adult (random adult means that this can either be a man or a women). Don't forget to set a seed. How tall is that person? What gender does that random person probably have?

A:

```
set.seed(123)
N <- 10
male <- rbinom(1,1,0.48)
height <- ifelse(male==1, rnorm(N, 69.1, 2.9), rnorm(1, 63.7, 2.7))
avg_height <- mean(height)
```

3.4 Question:

Now, simulate the distribution of the average height by generating 1000 draws. Plot the distribution of the average height of those 10 adults.

A:

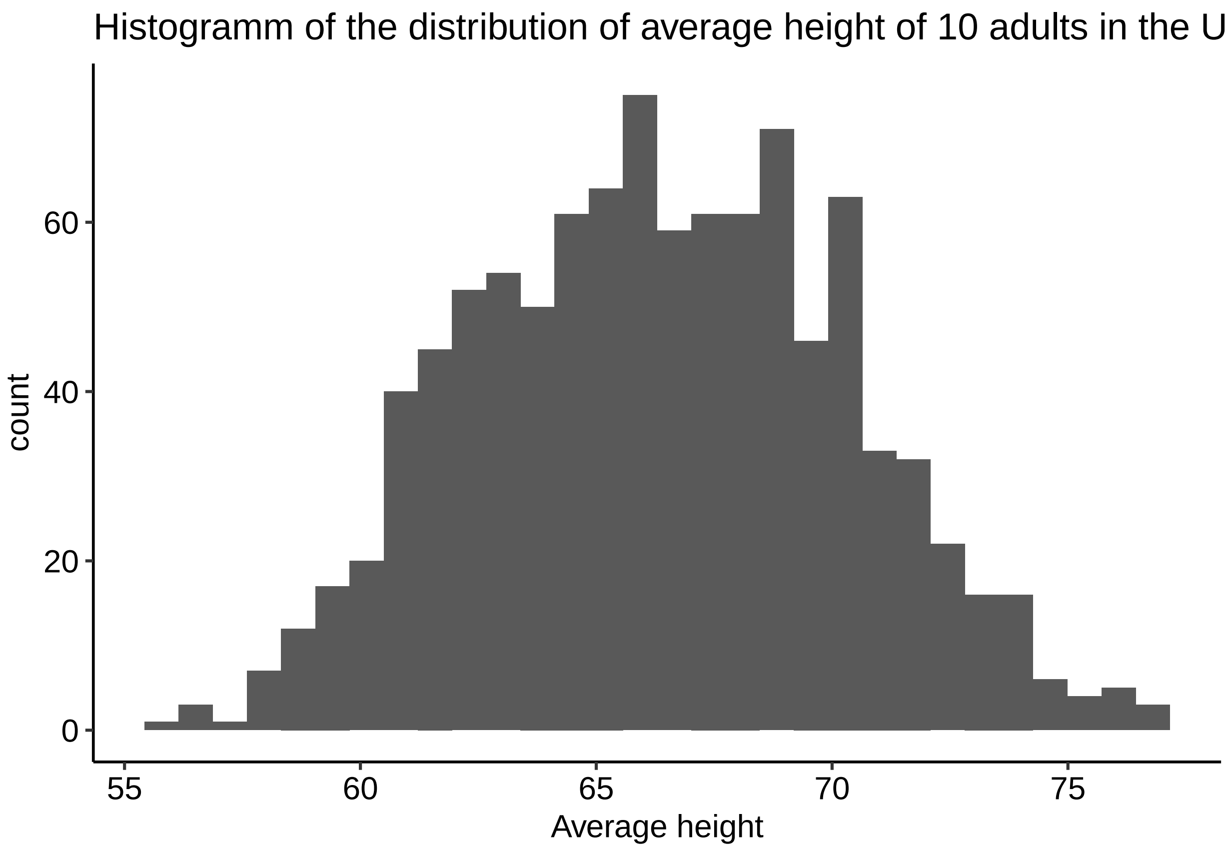
```
n_sims < 1000
```

```
[1] FALSE
```

```
avg_height <- rep(NA, n_sims)
for (s in 1:n_sims) {
  N <- 10
male <- rbinom(1,1,0.48)
height <- ifelse(male==1, rnorm(N, 69.1, 2.9), rnorm(1, 63.7, 2.7))
avg_height[s] <- mean(height)
}
```

```
avg_height <- as.data.frame(avg_height)
ggplot(avg_height, aes(avg_height)) +
  geom_histogram(show.legend = FALSE) +
  labs(title="Histogramm of the distribution of average height of 10 adults in the Unite
        x ="Average height")+
  theme_pubr()
```

```
`stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



3.5 Question:

Finally, instead of estimating the average height of 10 people, simulate the same model and extract the maximum height of 10 people. Plot the distribution.

A:

```
max_height <- rep(NA, n_sims)
for (s in 1:n_sims) {
  N <- 10
male <- rbinom(1,1,0.48)
height <- ifelse(male==1, rnorm(N, 69.1, 2.9), rnorm(1, 63.7, 2.7))
max_height[s] <- max(height)
avg_height[s] <- mean(height)
}
```

```
max_height <- as.data.frame(max_height)
ggplot(max_height, aes(max_height)) +
  geom_histogram(show.legend = FALSE) +
  labs(title="Histogramm of the distribution of maximum height of 10 adults in the Unite
        x ="Maximum height")+
  theme_pubr()
```

```
`stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```