Module 5 Exercise

```
# Initialize libraries
library(tidyverse)
```

```
## -- Attaching packages ----------------------------------------------------------------------- tidy
```

```
## v ggplot2 2.2.1     v purrr   0.2.4
## v tibble  1.4.2     v dplyr   0.7.4
## v tidyr   0.8.0     v stringr 1.2.0
## v readr   1.1.1     v forcats 0.2.0
```

```
## -- Conflicts -------------------------------------------------------------------------- tidyverse_
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(rio)
library(plm)
```

```
## Loading required package: Formula
```

```
##
## Attaching package: 'plm'
```

```
## The following objects are masked from 'package:dplyr':
##
##     between, lag, lead
```

1. Preliminaries

a. Create a Project and RMarkdown document

Either in a preexisting RStudio Project folder, or in a new Project, create a RMarkdown document for the exercise.

b. Initialize the Ecdat library

For the first part of the exercise, you will be working with the Ecdat package, which has many economics datasets.

- Install and initalize the package.

```
library(Ecdat)
```

```
## Loading required package: Ecfun
```

```
##
## Attaching package: 'Ecfun'
```

```
## The following object is masked from 'package:base':
##
##     sign
```

```
##
## Attaching package: 'Ecdat'
```

```
## The following object is masked from 'package:datasets':
##
##     Orange
```

2. Extramarital Affairs Exercise

b. Create an "affairs" tibble from the Fair data

For the first part of the exercise, you will be working the extramaritals data from Fair (Econometrica 1977).

- Create an "affairs" data frame from the Fair data in Ecdat as follows:

```
affairs <- Fair
```

- Turn the data set into a tibble
- Select the following variables: `nbaffairs, ym, child, rate, age`.
- Rename the variables as follows:
  - Change "nbaffairs" to "affairs"
  - Change "ym" to "yrsmarr"
  - Change "rate" to "mrating"

```
library(Ecdat)
affairs <- Fair %>% as.tibble() %>% select(nbaffairs, ym, child, rate, age) %>%
  rename(affairs = nbaffairs, yrsmarr= ym, mrating = rate)
```

c. Convert mrating to a factor

```
affairs$mrating <- ordered(affairs$mrating)
```

d. Perform regression

- Create a regression object that regresses affairs on age, child, yrsmarr and $yrsmarr^2$.

```
affairs_reg <- lm(affairs ~ age + child + yrsmarr + I(yrsmarr^2), data = affairs)
```

- Then view the summary output of the regression.

```
summary(affairs_reg)
```

```
##
## Call:
## lm(formula = affairs ~ age + child + yrsmarr + I(yrsmarr^2),
##     data = affairs)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.7476 -1.8476 -0.9871 -0.3325 11.6622
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)   1.207231   0.653281   1.848   0.0651 .
## age          -0.040625   0.023029  -1.764   0.0782 .
## childyes     -0.208227   0.380922  -0.547   0.5848
```

```
## yrsmarr       0.301967   0.126622   2.385   0.0174 *
## I(yrsmarr^2) -0.007684   0.007076  -1.086   0.2780
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.238 on 596 degrees of freedom
## Multiple R-squared:  0.04319,    Adjusted R-squared:  0.03677
## F-statistic: 6.726 on 4 and 596 DF,  p-value: 2.681e-05
```

f. Perform model diagnostics

- **Test for heteroskestacity.**
- **Test for missing polynomial terms.**

Write the statistical decision from each test in your RMarkdown report.

```
library(lmtest)
```

```
## Loading required package: zoo
```

```
##
## Attaching package: 'zoo'
```

```
## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric
```

```
bptest(affairs_reg)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  affairs_reg
## BP = 23.818, df = 4, p-value = 8.687e-05
```

```
resettest(affairs_reg)
```

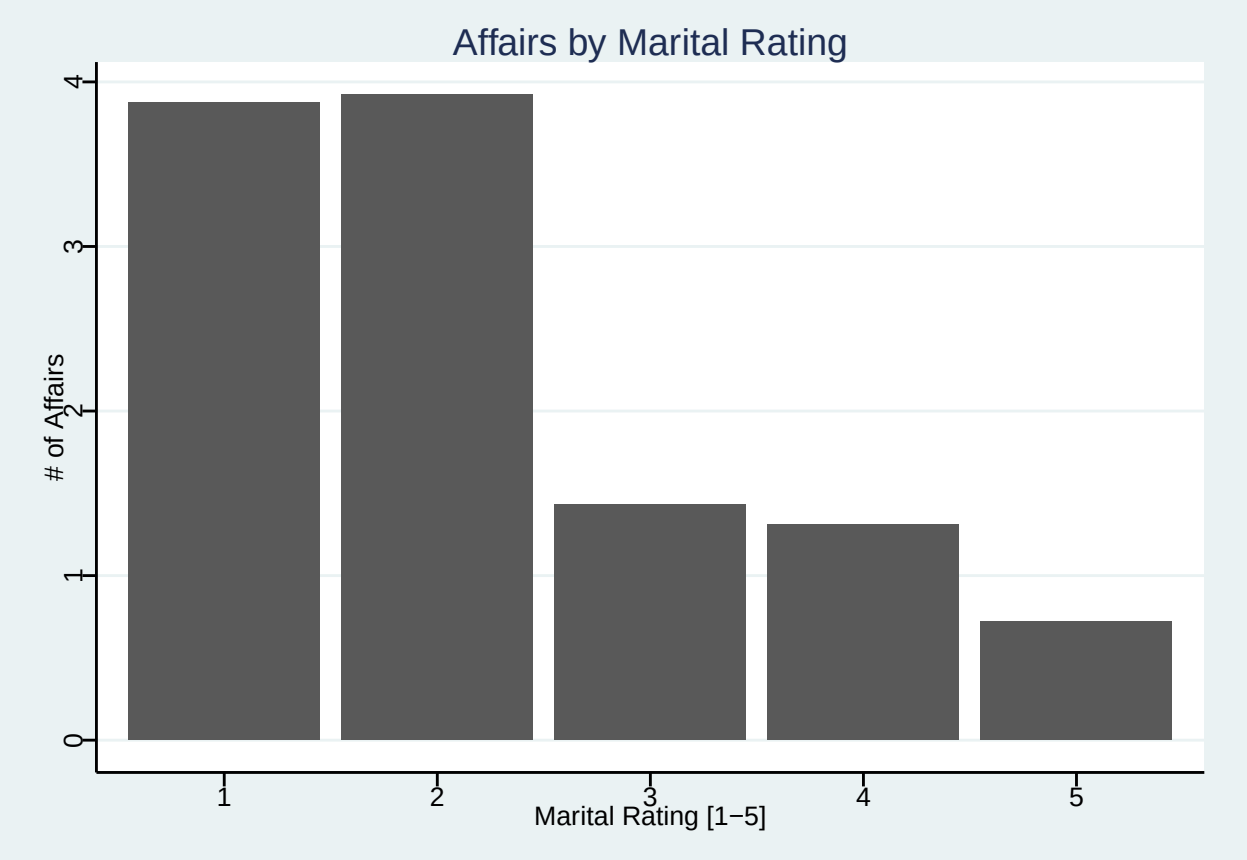
```
##
##  RESET test
##
## data:  affairs_reg
## RESET = 0.68819, df1 = 2, df2 = 594, p-value = 0.5029
```

g. Create a bar plot of average number of affairs by marital rating

- Create a summary data frame of number of affairs by marital rating
- Create a bar plot of the results, adding a graph title, axis labels, and style it like a Stata graph using **theme_stata()** from the ggthemes package.

```
library(ggthemes)

ggplot(affairs,
aes(x=mrating, y=affairs)) +
geom_bar( stat = "summary", fun.y = "mean") +
ggtitle("Affairs by Marital Rating") +
xlab("Marital Rating [1-5]") + ylab("# of Affairs") +
theme_stata()
```



3. Inequality in the World Development Indicators Example

For the next part of the exercise, let's revisit the inequality indicators from the World Development Indicators dataset.

- Import "wdi_data", which is just the data from Exercise 2, summarized by country and year from 2000.

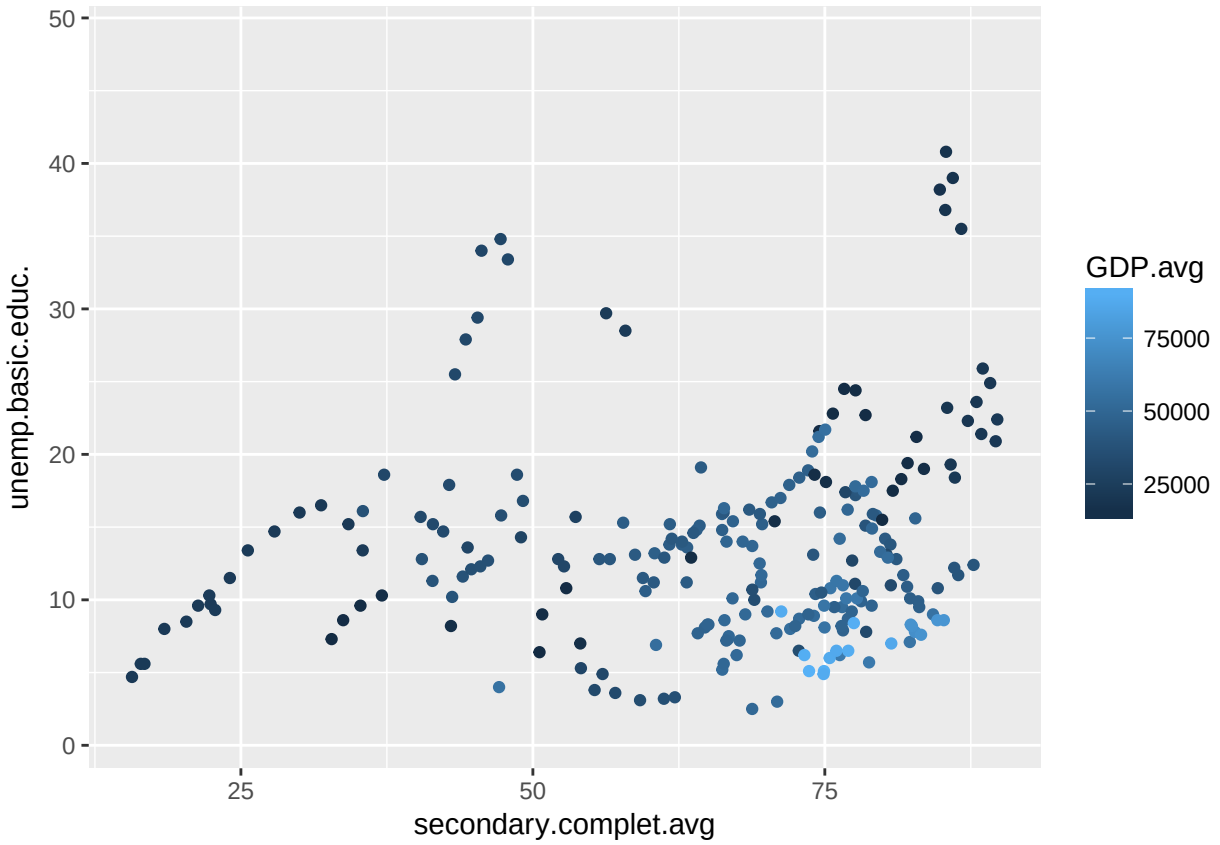
```
wdi_data <- import("C:/Users/AN.4271/OneDrive - Handelshögskolan i Stockholm/Teaching/R Course/Modules/
```

a. Create scatter plots

- Create twos scatterplots: each with secondary school completion as the x-axis.
  - For the first scatterplot, choose female-to-male employment ratio as the y-variable.
  - For the second scatterplot, make unemployment among those with basic education the y-variables:
- Color the points based on the value of GDP.
- Add titles and axis labels.

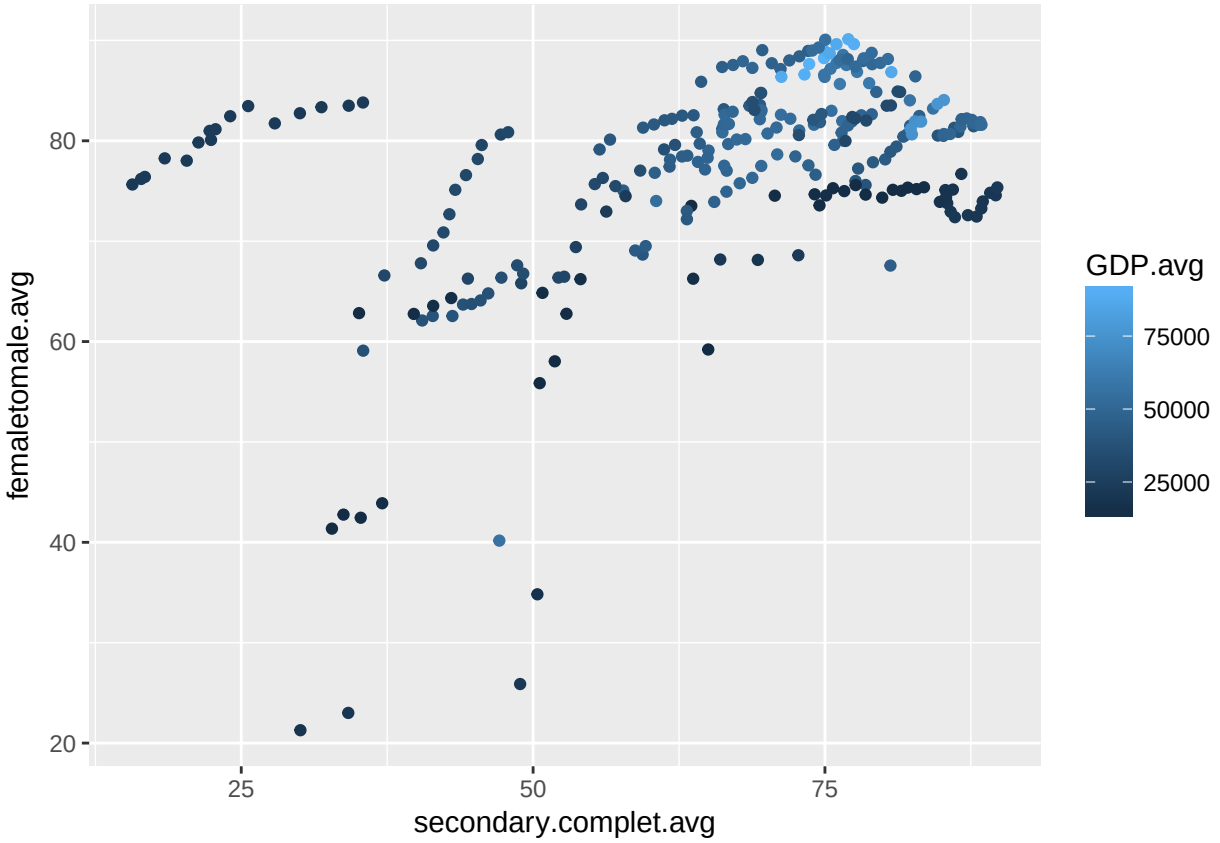
```
ggplot(wdi_data, aes(x=secondary.complet.avg,y =unemp.basic.educ.)) +
  geom_point(aes(color=GDP.avg))
```

```
## Warning: Removed 325 rows containing missing values (geom_point).
```




```
ggplot(wdi_data, aes(x=secondary.complet.avg,y =femaletomale.avg)) +
  geom_point(aes(color=GDP.avg))
```

```
## Warning: Removed 290 rows containing missing values (geom_point).
```



b. Perform regression

Perform a pooled OLS regression of GDP average on secondary school completion and view summary information from the regression.

```
WDI_regs <- lm(GDP.avg ~ secondary.complet.avg, data= wdi_data)
summary(WDI_regs)
```

```
##
## Call:
## lm(formula = GDP.avg ~ secondary.complet.avg, data = wdi_data)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -32665  -7370   1937   8322  49249
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)           13888.36    3906.06   3.556 0.000446 ***
## secondary.complet.avg   386.71      57.15   6.767 8.36e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 16320 on 266 degrees of freedom
##   (290 observations deleted due to missingness)
## Multiple R-squared:  0.1469, Adjusted R-squared:  0.1437
```

```
## F-statistic: 45.79 on 1 and 266 DF,  p-value: 8.363e-11
```

c. Test for autocorrelation

```
bgtest(WDI_regs)
```

```
##
##  Breusch-Godfrey test for serial correlation of order up to 1
##
## data:  WDI_regs
## LM test = 185.67, df = 1, p-value < 2.2e-16
```

d. Fixed effects regression

Now instead perform the regression of GDP on secondary school completion as a fixed effects regression (including fixed effects for country and year).

```
wdi_fixedeffects <- plm(GDP.avg ~ secondary.complet.avg, data=wdi_data,
                index=c("country","year"), model="within", effect="twoways")

summary(wdi_fixedeffects)
```

```
## Twoways effects Within Model
##
## Call:
## plm(formula = GDP.avg ~ secondary.complet.avg, data = wdi_data,
##     effect = "twoways", model = "within", index = c("country",
##         "year"))
##
## Unbalanced Panel: n = 36, T = 1-16, N = 268
##
## Residuals:
##       Min.    1st Qu.     Median    3rd Qu.       Max.
## -5422.2788  -578.3148     1.5143   667.1695  5071.2981
##
## Coefficients:
##                       Estimate Std. Error t-value Pr(>|t|)
## secondary.complet.avg  -33.861     38.989 -0.8685   0.3861
##
## Total Sum of Squares:    498670000
## Residual Sum of Squares: 496940000
## R-Squared:      0.0034796
## Adj. R-Squared: -0.23181
## F-statistic: 0.75422 on 1 and 216 DF, p-value: 0.38611
```

```
wdi_fixedeffects2 <- plm(GDP.avg ~ secondary.complet.avg + year, data=wdi_data,
                index=c("country","year"), model="within")
summary(wdi_fixedeffects2)
```

```
## Oneway (individual) effect Within Model
##
## Call:
## plm(formula = GDP.avg ~ secondary.complet.avg + year, data = wdi_data,
##     model = "within", index = c("country", "year"))
```

```
##
## Unbalanced Panel: n = 36, T = 1-16, N = 268
##
## Residuals:
##       Min.    1st Qu.     Median    3rd Qu.       Max.
## -5422.2788  -578.3148     1.5143   667.1695  5071.2981
##
## Coefficients:
##                       Estimate Std. Error t-value  Pr(>|t|)
## secondary.complet.avg  -33.861     38.989 -0.8685 0.3861073
## year2001              1941.493   1364.354  1.4230 0.1561752
## year2002              2402.227   1820.831  1.3193 0.1884648
## year2003               849.561   1365.019  0.6224 0.5343481
## year2004              1107.999   1105.294  1.0024 0.3172493
## year2005              1760.988   1076.690  1.6356 0.1033885
## year2006              2835.815   1097.091  2.5848 0.0104003 *
## year2007              4170.795   1099.328  3.7939 0.0001925 ***
## year2008              3985.862   1104.016  3.6103 0.0003802 ***
## year2009              2222.496   1098.738  2.0228 0.0443293 *
## year2010              3095.764   1104.752  2.8022 0.0055364 **
## year2011              3652.585   1118.910  3.2644 0.0012753 **
## year2012              3612.263   1140.049  3.1685 0.0017541 **
## year2013              4063.443   1129.358  3.5980 0.0003976 ***
## year2014              4152.850   1138.440  3.6478 0.0003315 ***
## year2015              4815.988   1163.905  4.1378 5.025e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Total Sum of Squares:    730780000
## Residual Sum of Squares: 496940000
## R-Squared:      0.31999
## Adj. R-Squared: 0.15943
## F-statistic: 6.35256 on 16 and 216 DF, p-value: 1.4177e-11
```

e. Autocorrelation robust inference

Test the coefficients of the model using the Stata-style HC_1 estimation of Newey-West heteroskedasticity and autocorrelated (HAC) robust standard errors.

- Use the function vcovNW() for specifying the variance method inside of coeftest().

```
coeftest(wdi_fixedeffects,vcov = vcovNW(wdi_fixedeffects,type="HC1"))
```

```
##
## t test of coefficients:
##
##                       Estimate Std. Error t value Pr(>|t|)
## secondary.complet.avg  -33.861     40.427 -0.8376   0.4032
```