HW5_hwenjun

Wenjun Han

10/29/2020

Problem 3

Get and clean the following data on education from the World Bank.

```r
# Read the csv data from the dowwloaded csv files
edstat_raw <- read.csv(file = 'EdStatsData.csv')

# Clean the data through tidyverse
# First we restack the data into a right format
edstat <-edstat_raw %>%
                      gather(key = "Time",value = "value", X1970:X2100)
# We found that the data has an X variable which not exist in orginal data
edstat = subset(edstat, select = -X)

#Change name of the variables
names(edstat)[1] <- "Country.Name"
head(edstat)
```

```
##   Country.Name Country.Code
## 1   Arab World          ARB
## 2   Arab World          ARB
## 3   Arab World          ARB
## 4   Arab World          ARB
## 5   Arab World          ARB
## 6   Arab World          ARB
##                                                          Indicator.Name
## 1            Adjusted net enrolment rate, lower secondary, both sexes (%)
## 2                Adjusted net enrolment rate, lower secondary, female (%)
## 3 Adjusted net enrolment rate, lower secondary, gender parity index (GPI)
## 4                  Adjusted net enrolment rate, lower secondary, male (%)
## 5                    Adjusted net enrolment rate, primary, both sexes (%)
## 6                        Adjusted net enrolment rate, primary, female (%)
##   Indicator.Code  Time    value
## 1     UIS.NERA.2 X1970       NA
## 2   UIS.NERA.2.F X1970       NA
## 3 UIS.NERA.2.GPI X1970       NA
## 4   UIS.NERA.2.M X1970       NA
## 5    SE.PRM.TENR X1970 54.82212
## 6 SE.PRM.TENR.FE X1970 43.35110
```

```
# Since we found that there are many missing data
# We remove the missing data from our table
edstat <- drop_na(edstat)
head(edstat)
```

```
##   Country.Name Country.Code
## 1   Arab World          ARB
## 2   Arab World          ARB
## 3   Arab World          ARB
## 4   Arab World          ARB
## 5   Arab World          ARB
## 6   Arab World          ARB
##                                                                  Indicator.Name
## 1                     Adjusted net enrolment rate, primary, both sexes (%)
## 2                         Adjusted net enrolment rate, primary, female (%)
## 3          Adjusted net enrolment rate, primary, gender parity index (GPI)
## 4                           Adjusted net enrolment rate, primary, male (%)
## 5 Adjusted net intake rate to Grade 1 of primary education, both sexes (%)
## 6     Adjusted net intake rate to Grade 1 of primary education, female (%)
##   Indicator.Code  Time    value
## 1    SE.PRM.TENR X1970 54.82212
## 2 SE.PRM.TENR.FE X1970 43.35110
## 3 UIS.NERA.1.GPI X1970  0.65857
## 4 SE.PRM.TENR.MA X1970 65.82623
## 5     UIS.NIRA.1 X1970 52.44892
## 6   UIS.NIRA.1.F X1970 44.34249
```

```
dim(edstat)
```

```
## [1] 5082201       6
```

There are 57650450 data points in complete dataset, and 5082201 data points we have in cleaned dataset. We chose country Australia and Hungary.

```
# Choose country Australia and Hungary
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
aus <- edstat[edstat$Country.Name == "Australia",]
head(aus)
```

```
##      Country.Name Country.Code
## 6404    Australia          AUS
## 6405    Australia          AUS
## 6406    Australia          AUS
## 6407    Australia          AUS
## 6408    Australia          AUS
## 6409    Australia          AUS
##                                                         Indicator.Name
## 6404   Barro-Lee: Average years of primary schooling, age 15+, female
## 6405    Barro-Lee: Average years of primary schooling, age 15+, total
## 6406 Barro-Lee: Average years of primary schooling, age 15-19, female
## 6407  Barro-Lee: Average years of primary schooling, age 15-19, total
## 6408 Barro-Lee: Average years of primary schooling, age 20-24, female
## 6409  Barro-Lee: Average years of primary schooling, age 20-24, total
##            Indicator.Code  Time value
## 6404 BAR.PRM.SCHL.15UP.FE X1970  5.59
## 6405    BAR.PRM.SCHL.15UP X1970  5.62
## 6406 BAR.PRM.SCHL.1519.FE X1970  5.93
## 6407    BAR.PRM.SCHL.1519 X1970  5.94
## 6408 BAR.PRM.SCHL.2024.FE X1970  5.88
## 6409    BAR.PRM.SCHL.2024 X1970  5.87
```

```r
hun <- edstat[edstat$Country.Name == "Hungary",]
head(hun)
```

```
##       Country.Name Country.Code
## 29365      Hungary          HUN
## 29366      Hungary          HUN
## 29367      Hungary          HUN
## 29368      Hungary          HUN
## 29369      Hungary          HUN
## 29370      Hungary          HUN
##                                                          Indicator.Name
## 29365   Barro-Lee: Average years of primary schooling, age 15+, female
## 29366    Barro-Lee: Average years of primary schooling, age 15+, total
## 29367 Barro-Lee: Average years of primary schooling, age 15-19, female
## 29368  Barro-Lee: Average years of primary schooling, age 15-19, total
## 29369 Barro-Lee: Average years of primary schooling, age 20-24, female
## 29370  Barro-Lee: Average years of primary schooling, age 20-24, total
##             Indicator.Code  Time value
## 29365 BAR.PRM.SCHL.15UP.FE X1970  7.26
## 29366    BAR.PRM.SCHL.15UP X1970  7.31
## 29367 BAR.PRM.SCHL.1519.FE X1970  7.84
## 29368    BAR.PRM.SCHL.1519 X1970  7.82
## 29369 BAR.PRM.SCHL.2024.FE X1970  7.82
## 29370    BAR.PRM.SCHL.2024 X1970  7.80
```

```r
# Create a summary table of indicators for comparison
summary(aus)
```

```
##  Country.Name       Country.Code       Indicator.Name     Indicator.Code
##  Length:28083       Length:28083       Length:28083       Length:28083
##  Class :character   Class :character   Class :character   Class :character
```

```
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##      Time               value
##  Length:28083       Min.   :0.000e+00
##  Class :character   1st Qu.:3.000e+00
##  Mode  :character   Median :5.000e+01
##                     Mean   :4.754e+09
##                     3rd Qu.:1.345e+05
##                     Max.   :1.567e+12
```

```r
summary(hun)
```

```
##  Country.Name       Country.Code       Indicator.Name     Indicator.Code
##  Length:33046       Length:33046       Length:33046       Length:33046
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##      Time               value
##  Length:33046       Min.   :-2.100e+01
##  Class :character   1st Qu.: 2.000e+00
##  Mode  :character   Median : 3.700e+01
##                     Mean   : 6.324e+08
##                     3rd Qu.: 3.778e+04
##                     Max.   : 2.651e+11
```

```r
by_aus <- aus %>% group_by(Indicator.Code) %>% summarise(mean = mean(value), n = n())
```

```
## `summarise()` ungrouping output (override with `.groups` argument)
```

```r
head(by_aus)
```

```
## # A tibble: 6 x 3
##   Indicator.Code       mean     n
##   <chr>               <dbl> <int>
## 1 BAR.NOED.1519.FE.ZS 0.88      9
## 2 BAR.NOED.1519.ZS    0.724     9
## 3 BAR.NOED.15UP.FE.ZS 1.12      9
## 4 BAR.NOED.15UP.ZS    0.991     9
## 5 BAR.NOED.2024.FE.ZS 0.339     9
## 6 BAR.NOED.2024.ZS    0.338     9
```

```r
by_hun <- hun %>% group_by(Indicator.Code) %>% summarise(mean = mean(value), n = n())
```

```
## `summarise()` ungrouping output (override with `.groups` argument)
```

```r
head(by_hun)
```

```
## # A tibble: 6 x 3
##   Indicator.Code       mean     n
##   <chr>               <dbl> <int>
## 1 BAR.NOED.1519.FE.ZS 0.96      9
## 2 BAR.NOED.1519.ZS    0.886     9
## 3 BAR.NOED.15UP.FE.ZS 1.29      9
## 4 BAR.NOED.15UP.ZS    1.07      9
## 5 BAR.NOED.2024.FE.ZS 0.511     9
## 6 BAR.NOED.2024.ZS    0.53      9
```

The summary of indicators of Australia data and Hungary data is shown as above tables.

Problem 4

Using base plotting functions to create a single figure as similar to the example. Use simple linear regression method for plotting data.

```r
library(MASS)
```

```
##
## Attaching package: 'MASS'

## The following object is masked from 'package:dplyr':
##
##     select
```

```r
library(faraway)
library(olsrr)
```

```
## Registered S3 methods overwritten by 'car':
##   method                          from
##   influence.merMod                lme4
##   cooks.distance.influence.merMod lme4
##   dfbeta.influence.merMod         lme4
##   dfbetas.influence.merMod        lme4

##
## Attaching package: 'olsrr'

## The following object is masked from 'package:faraway':
##
##     hsb

## The following object is masked from 'package:MASS':
##
##     cement
```

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

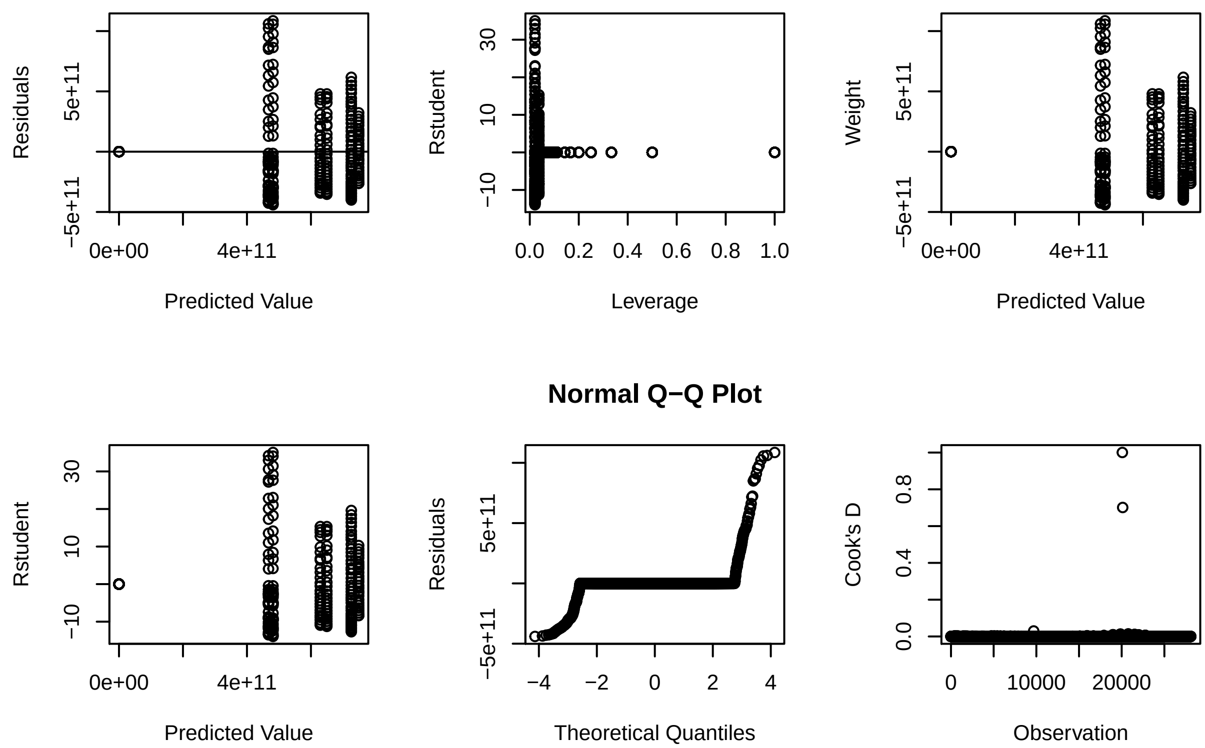
```r
library(sur)

# Plot country Australia
# Fit data with linear regression model
lmfit_aus <- lm(aus$value~aus$Indicator.Code)

# Plot 1
par(mfcol=c(2,3))
plot(fitted(lmfit_aus),residuals(lmfit_aus),xlab = "Predicted Value",ylab = "Residuals")
abline(h=0)

# Plot 2
plot(fitted(lmfit_aus),studres(lmfit_aus),xlab = "Predicted Value",ylab = "Rstudent")
```

```
## Warning in sqrt((n - p - sr^2)/(n - p - 1)): NaNs produced
```

```r
# Plot 3
plot(leverage(lmfit_aus),studres(lmfit_aus),xlab = "Leverage",ylab = "Rstudent")
```

```
## Warning in sqrt((n - p - sr^2)/(n - p - 1)): NaNs produced
```

```r
# Plot 4
qqnorm(residuals(lmfit_aus),ylab = "Residuals")

# Plot 5
plot(fitted(lmfit_aus),weighted.residuals(lmfit_aus, drop0 = TRUE),xlab = "Predicted Value",ylab = "Wei

# Plot 6
plot(cooks.distance(lmfit_aus),xlab = "Observation",ylab = "Cook's D")
```



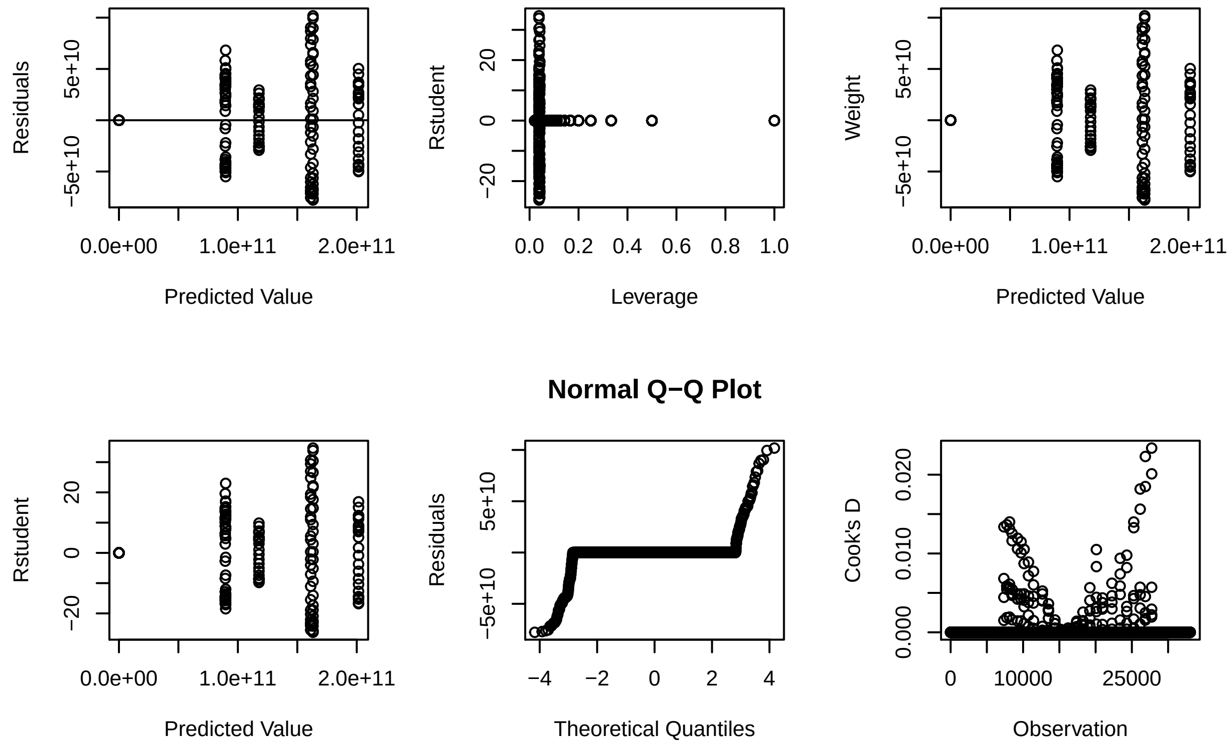

```
# Plot country Hungary
lmfit_hun <- lm(hun$value~hun$Indicator.Code)

par(mfcol=c(2,3))
plot(fitted(lmfit_hun),residuals(lmfit_hun),xlab = "Predicted Value",ylab = "Residuals")
abline(h=0)

# Plot 2
plot(fitted(lmfit_hun),studres(lmfit_hun),xlab = "Predicted Value",ylab = "Rstudent")
```

```
## Warning in sqrt((n - p - sr^2)/(n - p - 1)): NaNs produced
```

```
# Plot 3
plot(leverage(lmfit_hun),studres(lmfit_hun),xlab = "Leverage",ylab = "Rstudent")
```

```
## Warning in sqrt((n - p - sr^2)/(n - p - 1)): NaNs produced
```

```
# Plot 4
qqnorm(residuals(lmfit_hun),ylab = "Residuals")

# Plot 5
plot(fitted(lmfit_hun),weighted.residuals(lmfit_hun, drop0 = TRUE),xlab = "Predicted Value",ylab = "Wei

# Plot 6
plot(cooks.distance(lmfit_hun),xlab = "Observation",ylab = "Cook's D")
```



Problem 5

Recreate the plot in the last problem using ggplot2 functions.

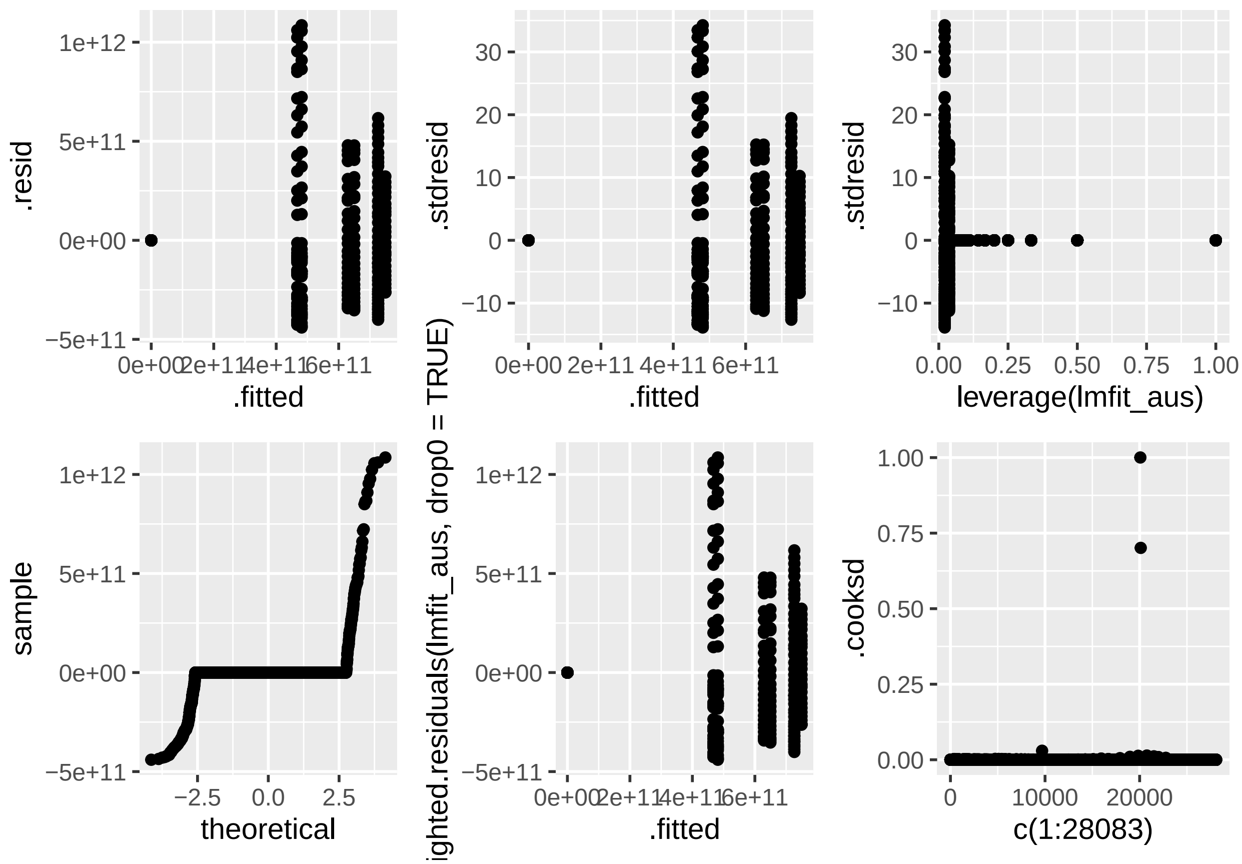
```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.0 --

## v ggplot2 3.3.2     v purrr   0.3.4
## v tibble  3.0.4     v stringr 1.4.0
## v readr   1.4.0     v forcats 0.5.0

## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## x MASS::select()  masks dplyr::select()
```

```
library(ggplot2)
library(ggpubr)

# Plot country Australia

# Plot 1
```

```
lmfit_ausf <- fortify(lmfit_aus)
a1 <- ggplot(lmfit_ausf, aes(x = .fitted, y = .resid)) + geom_point()

# Plot 2
b1 <- ggplot(lmfit_ausf, aes(x = .fitted, y = .stdresid)) + geom_point()

# Plot 3
c1 <- ggplot(lmfit_ausf, aes(x = leverage(lmfit_aus), y = .stdresid)) + geom_point()

# Plot 4
d1 <- ggplot(lmfit_ausf, aes(sample=.resid))+stat_qq()

# Plot 5
e1 <- ggplot(lmfit_ausf, aes(x = .fitted, y = weighted.residuals(lmfit_aus, drop0 = TRUE))) + geom_poin

# Plot 6
f1 <- ggplot(lmfit_ausf, aes(x= c(1:28083), y=.cooksd)) + geom_point()

ggarrange(a1, b1, c1, d1, e1, f1,
          ncol = 3, nrow = 2)
```

```
## Warning: Removed 69 rows containing missing values (geom_point).

## Warning: Removed 69 rows containing missing values (geom_point).

## Warning: Removed 69 rows containing missing values (geom_point).
```




```r
# Plot country Hungary
lmfit_hunf <- fortify(lmfit_hun)
a2 <- ggplot(lmfit_hunf, aes(x = .fitted, y = .resid)) + geom_point()

# Plot 2
b2 <- ggplot(lmfit_hunf, aes(x = .fitted, y = .stdresid)) + geom_point()

# Plot 3
c2 <- ggplot(lmfit_hunf, aes(x = leverage(lmfit_hun), y = .stdresid)) + geom_point()

# Plot 4
d2 <- ggplot(lmfit_hunf, aes(sample=.resid))+stat_qq()

# Plot 5
e2 <- ggplot(lmfit_hunf, aes(x = .fitted, y = weighted.residuals(lmfit_hun, drop0 = TRUE))) + geom_poin

# Plot 6
f2 <- ggplot(lmfit_hunf, aes(x= c(1:33046), y=.cooksd)) + geom_point()

ggarrange(a2, b2, c2, d2, e2, f2,
          ncol = 3, nrow = 2)
```

```
## Warning: Removed 20 rows containing missing values (geom_point).

## Warning: Removed 20 rows containing missing values (geom_point).
```

```
## Warning: Removed 20 rows containing missing values (geom_point).
```