Homework 5

Youhui Ye

10/23/2020

Problem 3

```
library(tidyverse)
## import data
dat <- read_csv("Edstats_csv/EdstatsData.csv")
dim(dat)[1]
```

```
## [1] 886930
```

```
## tidy data using functions in tidyverse
tidy_dat <- dat %>%
  select(-70) %>% # delete empty column
  gather(key = "Year", value = "Value", 5:69, na.rm = TRUE)
dim(tidy_dat)[1]
```

```
## [1] 4205910
```

There are 886930 data points in the complete data set. After I reshaped the data, there are 4205910 data points.

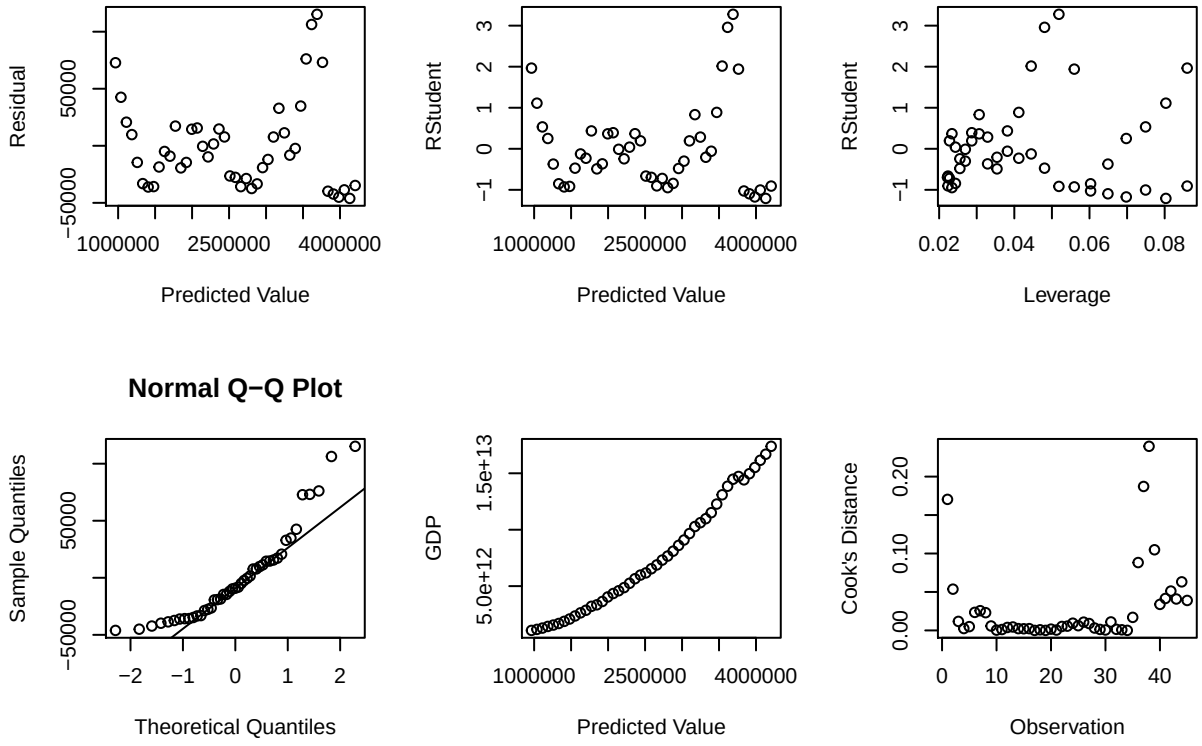
```
table1 <- tidy_dat %>%
  filter(get('Country Code') %in% c("CHN", "USA") ) %>%
  group_by(`Indicator Code`, `Country Code`) %>%
  summarise(Mean = mean(Value), "Standard Deviation" = sd(Value), .groups = "drop")
## First 10 row are displayed
knitr::kable(table1[1:10,])
```

Indicator Code	Country Code	Mean	Standard Deviation
BAR.NOED.1519.FE.ZS	CHN	3.5055556	2.7358413
BAR.NOED.1519.FE.ZS	USA	0.4144444	0.3586820
BAR.NOED.1519.ZS	CHN	5.9822222	7.5631637
BAR.NOED.1519.ZS	USA	0.5511111	0.5387357
BAR.NOED.15UP.FE.ZS	CHN	22.6011111	11.5245677
BAR.NOED.15UP.FE.ZS	USA	0.7266667	0.3357082
BAR.NOED.15UP.ZS	CHN	20.2288889	11.4435063
BAR.NOED.15UP.ZS	USA	0.8300000	0.4043204
BAR.NOED.2024.FE.ZS	CHN	3.9355556	3.1682097
BAR.NOED.2024.FE.ZS	USA	0.3488889	0.2163588

Problem 4

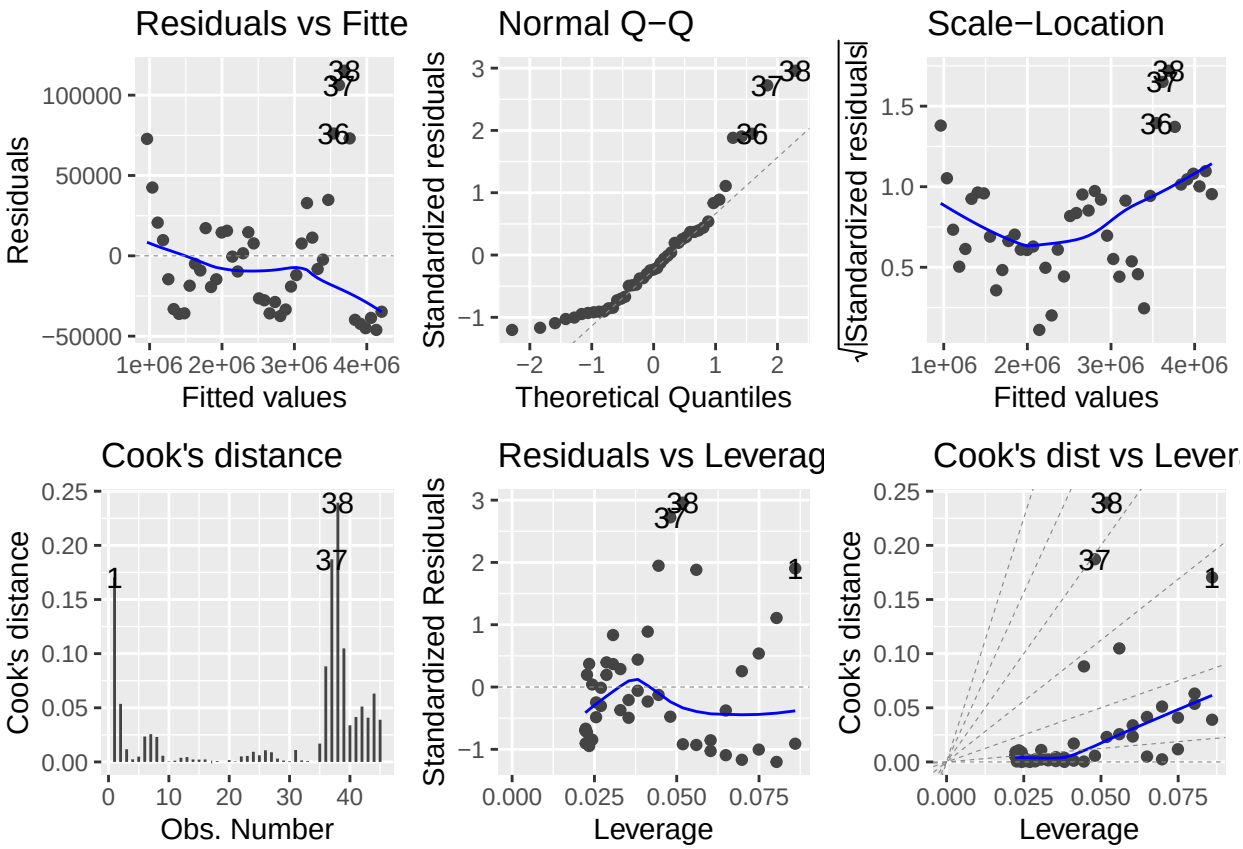
```
library(tidyverse)
USA_GDP <- tidy_dat %>%
  filter(get('Country Code') == "USA" & get('Indicator Name') == "GDP at market prices (current US$)" )
  select(Year, Value)

library(MASS) # studendized residuals
lmfit <- lm(sqrt(Value) ~ as.numeric(Year), data = USA_GDP)
par(mfrow = c(2, 3))
plot(lmfit$fitted.values, lmfit$residuals, xlab = "Predicted Value", ylab = "Residual")
plot(lmfit$fitted.values, studres(lmfit), xlab = "Predicted Value", ylab = "RStudent")
plot(hatvalues(lmfit), studres(lmfit), xlab = "Leverage", ylab = "RStudent")
qqnorm(lmfit$residuals)
qqline(lmfit$residuals)
plot(lmfit$fitted.values, USA_GDP$Value, xlab = "Predicted Value", ylab = "GDP")
plot(cooks.distance(lmfit), ylab = "Cook's Distance", xlab = "Observation")
```



Problem 5

```
library(ggfortify)
autoplot(lmfit, which = 1:6, nrow = 2)
```


Residuals vs Fitte
Residuals
Fitted values
Normal Q−Q
Standardized residuals
Theoretical Quantiles
Scale−Location
√|Standardized residuals|
Fitted values
Cook's distance
Cook's distance
Obs. Number
Residuals vs Leverag
Standardized Residuals
Leverage
Cook's dist vs Lever
Cook's distance
Leverage