

hw5_tdolkar

Tsering Dolkar

10/26/2020

Problem 1.

Worked through the Swirl “Exploratory_Data_Analysis” lesson parts 1-10.

Problem 2.

Created the Rmd file

Problem 3.

```
## getting http://databank.worldbank.org/data/download/Edstats_csv.zip

#read in data, looks like csv dump, blah
data_raw <- read.csv('EdStatsData.csv', sep = ",")
```

```
destroyX = function(es) {
  f = es
  for (i in c(1:length(f))){ #for each value in Year
    if (startsWith(f[i], "X") == TRUE) { #if starts with 'X' ..
      f[i] <- substr(f[i], 2, 100) #get rid of it
    }
  }
  return(f)
}
```

```
tidy_data <- data_raw %>%
  gather(key = "Year", value = "Data", 5:70)
tidy_data$Year <- destroyX(tidy_data$Year)
tidy_data <- na.omit(tidy_data)
completeDataset_dataPoints <- dim(data_raw)
completeDataset_dataPoints
```

```
## [1] 886930    70
```

```
cleanedDataset_dataPoints <- dim(tidy_data)
cleanedDataset_dataPoints
```

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

```
chosenCountries <- tidy_data %>%  
  filter(Country.Code == "EMU" | Country.Code == "ECS")  
summary_stat <- by(chosenCountries, chosenCountries$Indicator.Name, summary)  
length(summary_stat)
```

```
## [1] 367
```

```
head(summary_stat)
```

```
## $'Adjusted net enrolment rate, primary, both sexes (%)'  
## Country.Name Country.Code Indicator.Name Indicator.Code  
## Length:88 Length:88 Length:88 Length:88  
## Class :character Class :character Class :character Class :character  
## Mode :character Mode :character Mode :character Mode :character  
##  
##  
##  
## Year Data  
## Length:88 Min. :93.05  
## Class :character 1st Qu.:95.01  
## Mode :character Median :96.72  
## Mean :96.41  
## 3rd Qu.:97.51  
## Max. :99.35  
##  
## $'Adjusted net enrolment rate, primary, female (%)'  
## Country.Name Country.Code Indicator.Name Indicator.Code  
## Length:87 Length:87 Length:87 Length:87  
## Class :character Class :character Class :character Class :character  
## Mode :character Mode :character Mode :character Mode :character  
##  
##  
##  
## Year Data  
## Length:87 Min. :92.44  
## Class :character 1st Qu.:94.60  
## Mode :character Median :96.73  
## Mean :96.31  
## 3rd Qu.:97.36  
## Max. :99.43  
##  
## $'Adjusted net enrolment rate, primary, gender parity index (GPI)'  
## Country.Name Country.Code Indicator.Name Indicator.Code  
## Length:87 Length:87 Length:87 Length:87  
## Class :character Class :character Class :character Class :character  
## Mode :character Mode :character Mode :character Mode :character  
##  
##  
##  
## Year Data  
## Length:87 Min. :0.9810
```

```

## Class :character 1st Qu.:0.9907
## Mode :character Median :0.9985
## Mean :0.9971
## 3rd Qu.:1.0040
## Max. :1.0112
##
## $'Adjusted net enrolment rate, primary, male (%)'
## Country.Name Country.Code Indicator.Name Indicator.Code
## Length:87 Length:87 Length:87 Length:87
## Class :character Class :character Class :character Class :character
## Mode :character Mode :character Mode :character Mode :character
##
##
##
## Year Data
## Length:87 Min. :93.64
## Class :character 1st Qu.:95.30
## Mode :character Median :96.72
## Mean :96.58
## 3rd Qu.:97.78
## Max. :99.31
##
## $'Adjusted net intake rate to Grade 1 of primary education, both sexes (%)'
## Country.Name Country.Code Indicator.Name Indicator.Code
## Length:87 Length:87 Length:87 Length:87
## Class :character Class :character Class :character Class :character
## Mode :character Mode :character Mode :character Mode :character
##
##
##
## Year Data
## Length:87 Min. :89.44
## Class :character 1st Qu.:91.62
## Mode :character Median :92.91
## Mean :93.45
## 3rd Qu.:95.91
## Max. :97.30
##
## $'Adjusted net intake rate to Grade 1 of primary education, female (%)'
## Country.Name Country.Code Indicator.Name Indicator.Code
## Length:53 Length:53 Length:53 Length:53
## Class :character Class :character Class :character Class :character
## Mode :character Mode :character Mode :character Mode :character
##
##
##
## Year Data
## Length:53 Min. :89.56
## Class :character 1st Qu.:93.80
## Mode :character Median :95.85
## Mean :95.42
## 3rd Qu.:96.99
## Max. :97.80

```

Problem 4.

```
options(scipen = 0)
options(digits = 2)

par(mfrow=c(2,3))

lmfit <- lm(chosenCountries$Data ~ chosenCountries$Year)

plot(fitted(lmfit),residuals(lmfit),pch=16,xlab = "Predicted Value", ylab = "Residual")
abline(h = 0)

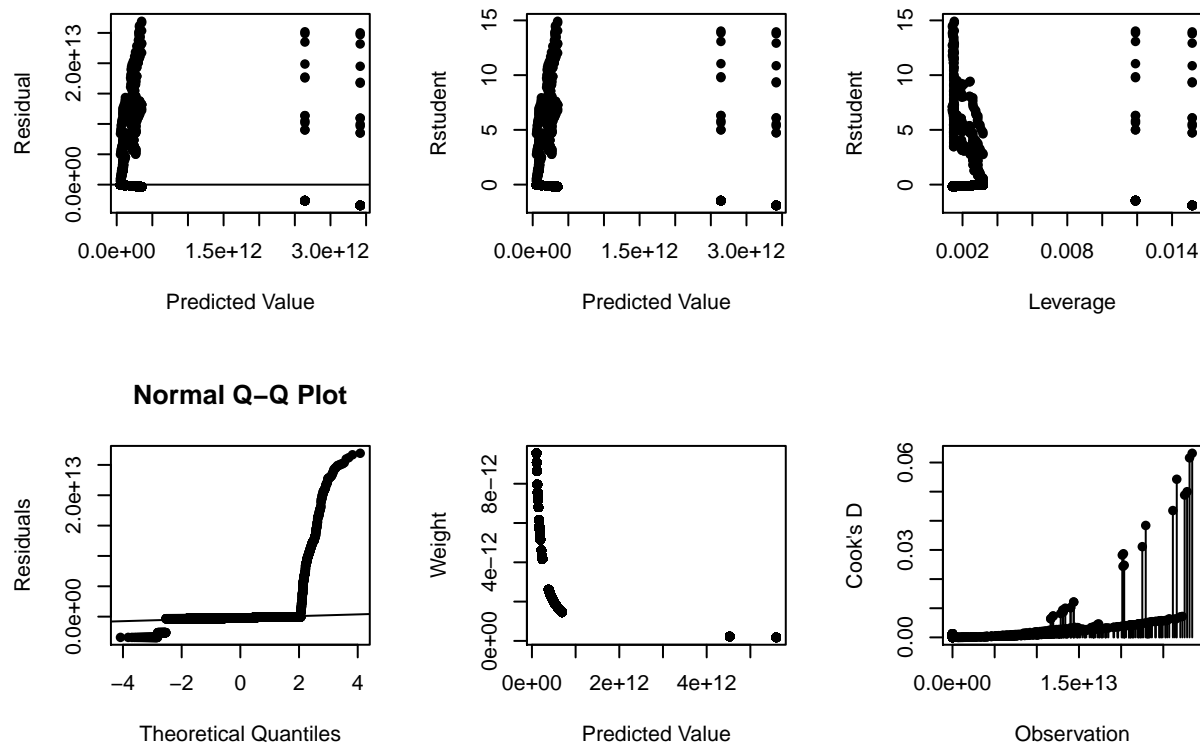
plot(fitted(lmfit),rstudent(lmfit),pch=16,xlab = "Predicted Value", ylab = "Rstudent")

plot(hatvalues(lmfit), rstudent(lmfit), pch = 16, xlab="Leverage", ylab = "Rstudent")

qqnorm(residuals(lmfit), ylab = "Residuals", pch = 16)
qqline(residuals(lmfit))

lmfitw <- lm(abs(residuals(lmfit)) ~ chosenCountries$Year)
w <- 1/abs(fitted(lmfitw))
yw <- w^0.5*chosenCountries$Data
plot(fitted(lmfitw), w, pch=16, xlab = "Predicted Value", ylab = "Weight")

plot(chosenCountries$Data, cooks.distance(lmfit), xlab = "Observation", ylab = "Cook's D", type = "h")
points(chosenCountries$Data, cooks.distance(lmfit), pch = 16)
```



Problem 5.

```
p1 <- ggplot(chosenCountries, aes(x=fitted(lmfit), y=residuals(lmfit), color="red")) +
  geom_point() +
  geom_hline(yintercept = 0) +
  labs(x = "Predicted Value", y = "Residual") +
  theme(axis.text.x = element_text(angle = 90, size = 5,
    vjust = 0.6), legend.position = "none")

p2 <- ggplot(chosenCountries, aes(x=fitted(lmfit), y=rstudent(lmfit), color="red")) +
  geom_point() +
  labs(x = "Predicted Value", y = "Rstudent") +
  theme(axis.text.x = element_text(angle = 90, size = 5,
    vjust = 0.6), legend.position = "none")

p3 <- ggplot(chosenCountries, aes(x=hatvalues(lmfit), y=rstudent(lmfit), color="red")) +
  geom_point() +
  labs(x = "Leverage", y = "Rstudent") +
  theme(axis.text.x = element_text(angle = 90, size = 5,
    vjust = 0.6), legend.position = "none")

p4 <- ggplot(chosenCountries, aes(sample=residuals(lmfit), color = "red")) +
  stat_qq() +
  stat_qq_line(line.p = c(0.25, 0.75)) +
```

```

    labs(y = "Residual") +
    theme(legend.position = "none")

p5 <-ggplot(chosenCountries, aes(x=fitted(lmfit), y= w, color="red")) +
  geom_point() +
  geom_hline(yintercept = 0) +
  labs(x = "Predicted Value", y = "Weight") +
  theme(axis.text.x = element_text(angle = 90, size = 5,
                                     vjust = 0.6),legend.position = "none")

p6 <-ggplot(chosenCountries, aes(x=Data, y= cooks.distance(lmfit), color="red")) +
  geom_point() +
  geom_segment(aes(x=Data, y = 0, xend = Data, yend= cooks.distance(lmfit))) +
  geom_hline(yintercept = 0) +
  labs(x = "Observation", y = "Cook's D") +
  theme(axis.text.x = element_text(angle = 90, size = 5,
                                     vjust = 0.6),legend.position = "none")

figure1 <- multi_panel_figure(columns = 3, rows = 2, panel_label_type = "none")
figure1 %<>%
  fill_panel(p1, column = 1, row = 1) %<>%
  fill_panel(p2, column = 2, row = 1) %<>%
  fill_panel(p3, column = 3, row = 1) %<>%
  fill_panel(p4, column = 1, row = 2) %<>%
  fill_panel(p5, column = 2, row = 2) %<>%
  fill_panel(p6, column = 3, row = 2)
figure1

```

