HW5_wliu3

Wei Liu

11/3/2020

Problem 1

Done

Problem 2

Done

Problem 3

```
library(data.table)
```

```
##
## Attaching package: 'data.table'

## The following objects are masked from 'package:dplyr':
##
##     between, first, last

## The following object is masked from 'package:purrr':
##
##     transpose
```

```
bank_data<-fread("EdStatsData.csv",header = T)

#create function to extract numbers from a character string
numextract <- function(string){
  str_extract(string, "\\-*\\d+\\.*\\d*")
}

#change name of columns for dataset
name_bank<-names(bank_data)
name_bank[-(1:4)]<-numextract(name_bank[-(1:4)])
name_bank[1]<-"Country.Name"
names(bank_data)<-name_bank
dim(bank_data)
```

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

There are 886930 observations of 70 variables in the complete dataset.

```
data1<-bank_data%>%
  gather("year", "values", 5:69)

data2<-data1[,-5]
data3<-data2 %>% drop_na()
```

There are 5082201 observations of 6 variables in the cleaned dataset.

```
China_data <- data3[data3$Country.Name == "China" & data3$`Indicator Code`== "BAR.PRM.SCHL.15UP" , c(1,4

Japan_data <- data3[data3$Country.Name == "Japan" & data3$`Indicator Code` == "BAR.PRM.SCHL.15UP",c(1,4

China_data1 <- data3[data3$Country.Name == "China" & data3$`Indicator Code`== "BAR.PRM.SCHL.2024" , c(1

Japan_data1 <- data3[data3$Country.Name == "Japan" & data3$`Indicator Code` == "BAR.PRM.SCHL.2024",c(1,4

data4<-rbind(China_data, China_data1, Japan_data ,Japan_data1)
data5<- data4 %>% spread(`Indicator Code`, values)

data7<-summary(data5[1:9,])
data8<-summary(data5[10:18,])
knitr::kable(data7,caption="Summary table of two indicators from China")
```

Table 1: Summary table of two indicators from China

Country.Name	year	BAR.PRM.SCHL.15UP	BAR.PRM.SCHL.2024
Length:9	Length:9	Min. :2.990	Min. :4.760
Class :character	Class :character	1st Qu.:3.930	1st Qu.:4.850
Mode :character	Mode :character	Median :4.230	Median :4.910
NA	NA	Mean :4.231	Mean :5.047
NA	NA	3rd Qu.:4.810	3rd Qu.:5.240
NA	NA	Max. :5.070	Max. :5.590

```
knitr::kable(data7,caption="Summary table of two indicators from Japan")
```

Table 2: Summary table of two indicators from Japan

Country.Name	year	BAR.PRM.SCHL.15UP	BAR.PRM.SCHL.2024
Length:9	Length:9	Min. :2.990	Min. :4.760
Class :character	Class :character	1st Qu.:3.930	1st Qu.:4.850
Mode :character	Mode :character	Median :4.230	Median :4.910
NA	NA	Mean :4.231	Mean :5.047
NA	NA	3rd Qu.:4.810	3rd Qu.:5.240
NA	NA	Max. :5.070	Max. :5.590

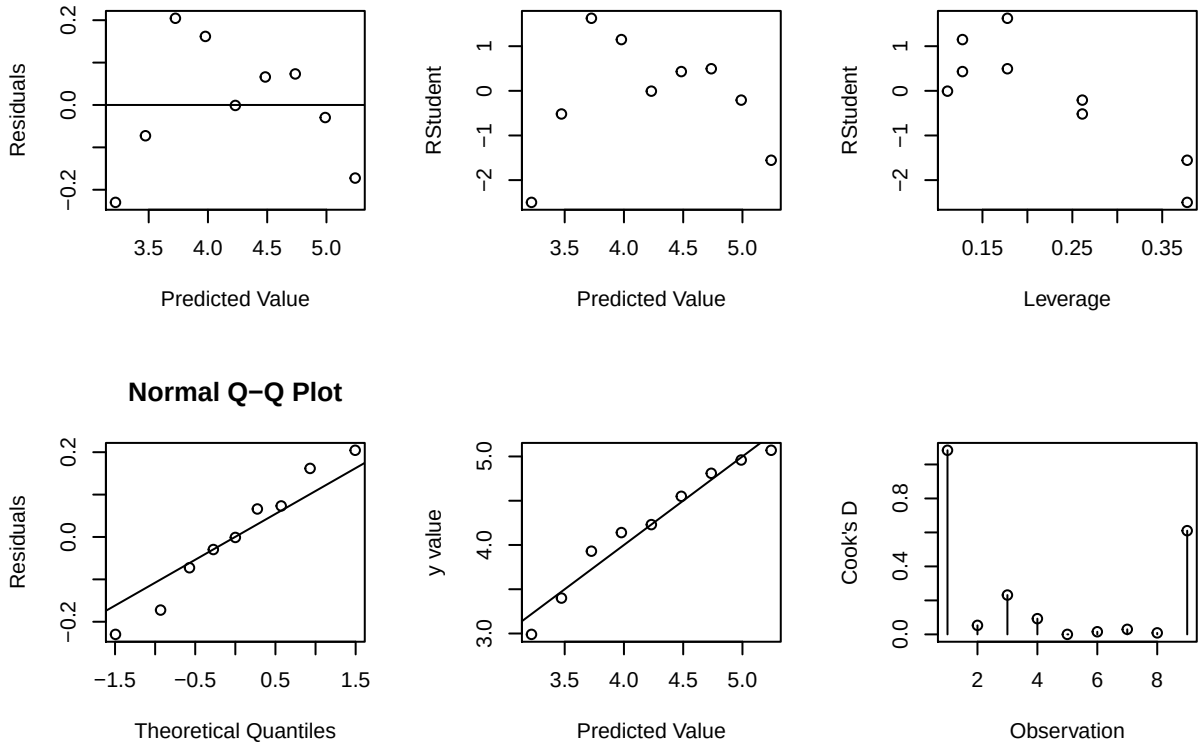
Problem 4

```
#linear regression model of x and y
x<-c(1:9)
y<-data5$BAR.PRM.SCHL.15UP[1:9]
lmfit<-lm(y~x)


library(MASS)
```

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

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

```
library(stats)
library(faraway)
# combine plots together
par(mfcol=c(2,3))

plot(fitted(lmfit),residuals(lmfit),xlab="Predicted Value",ylab="Residuals")
abline(h=0)
qqnorm(residuals(lmfit),ylab="Residuals")
qqline(residuals(lmfit))

#Studentized residuals

plot(fitted(lmfit), studres(lmfit), xlab="Predicted Value",ylab="RStudent")
plot(fitted(lmfit), y, xlab="Predicted Value",ylab="y value")
abline(a=0,b=1)
#leverage values

plot(hatvalues(lmfit), studres(lmfit), xlab="Leverage",ylab="RStudent")

#cook distance
cook<-cooks.distance(lmfit)
plot(x, cook, xlab="Observation",ylab="Cook's D")
lines(x, cook,type="h")
```



Problem 5

```
library(ggplot2)
# 6 figures
s1<-ggplot(lmfit, aes(x = fitted(lmfit), y = residuals(lmfit))) +
  geom_point() +
  labs(x="Predicted Value", y="Residuals")+
geom_hline(yintercept=0,linetype="dashed",color = "blue")+theme_classic()


s4<-qplot(sample = residuals(lmfit), data = lmfit)+
labs(x="Theoretical quantiles", y = "Residuals")+
  theme_classic()+stat_qq_line()

s2<-ggplot(lmfit, aes(x = fitted(lmfit), y = studres(lmfit))) +
  geom_point() +
  labs(x="Predicted Value", y="RStudent")+
  theme_classic()

s5<-ggplot(lmfit, aes(x = fitted(lmfit), y = y)) +
  geom_point() +
  labs(x="Predicted Value", y="y values")+
  geom_abline(slope=1)+
```

```
  theme_classic()

s3<-ggplot(lmfit, aes(x = hatvalues(lmfit), y = studres(lmfit))) +
  geom_point() +
  labs(x="Leverage", y="RStudent")+
  theme_classic()

s6<-ggplot(lmfit, aes(x = x, xend=x, y=0, yend=cook)) +
  geom_segment() +
  labs(x="Observation", y="Cook's D")+
  theme_classic()

#combine 6 figures together
library(ggpubr)
```

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

```
figure_final<-ggarrange(s1,s2,s3,s4,s5,s6,ncol=3,nrow=2)
figure_final
```