HW5_linal20

Lina Lee

10/27/2020

Problem 3

Using tidy concepts, get and clean the following data on education from the World Bank. http://databank.worldbank.org/data/download/Edstats_csv.zip How many data points were there in the complete dataset? In your cleaned dataset? Choosing 2 countries, create a summary table of indicators for comparison.

```
setwd("C:\\Users\\linal\\Documents\\STAT_5014_2020_linal20")
edDat<-read.csv("EdStatsData.csv",header = TRUE)

# the number of rows of the complete dataset
nrow(edDat)
```

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

there are 886930 rows in the complete dataset.

```
# filter data for two countries, Korea and USA
KORdat<-edDat%>%filter(Country.Code=="KOR")
USAdat<-edDat%>%filter(Country.Code=="USA")

# delete columns including only NA
ind <- apply(KORdat[,5:50], 1, function(x) all(is.na(x)))
KORdat_sub <- KORdat[ !ind, ]

ind <- apply(USAdat[,5:50], 1, function(x) all(is.na(x)))
USAdat_sub <- USAdat[ !ind, ]

# indicators that we consider are:
# Barro-Lee: Percentage of female population age 40-44 with no education
# Barro-Lee: Percentage of female population age 45-49 with no education
# Barro-Lee: Percentage of female population age 50-54 with no education
# Barro-Lee: Percentage of female population age 55-59 with no education
# Barro-Lee: Percentage of female population age 60-64 with no education

# filter indicators including Percentage of female population with no education

feNoEd_kor<-KORdat_sub%>%filter(Indicator.Name=="Barro-Lee: Percentage of female population age 40-44 wi


feNoEd_us<-USAdat_sub%>%filter(Indicator.Name=="Barro-Lee: Percentage of female population age 40-44 wit
feNoEd_kor$age<-c("40-44","45-49","50-54","55-59","60-64")
feNoEd_us$age<-c("40-44","45-49","50-54","55-59","60-64")
```

```
# remove rows with only NA
feNoEd_kor <- feNoEd_kor %>% select_if(~all(!is.na(.)))

feNoEd_us <- feNoEd_us %>% select_if(~all(!is.na(.)))
```

```
# combind datasets for two countries
feNoEd<-rbind(feNoEd_kor,feNoEd_us)

# the number of rows
nrow(feNoEd)
```

```
## [1] 10
```

the number of row for the cleaned dataset is 10.

```
# transform wide data form into long data form.
feNoEd_long<-gather(feNoEd, key = "year", value = "value",5:13)
feNoEd_long<-feNoEd_long%>%mutate(year=substr(year,2,5))
feNoEd_long$year<-as.numeric(feNoEd_long$year)
feNoEd_long$age<-as.factor(feNoEd_long$age)



# filter for the women age of 40-44 (we will focus on age 40-44)
feNoEd_kor_long_age1<-feNoEd_long%>%filter(Country.Code=="KOR")%>%filter(age=="40-44")
feNoEd_usa_long_age1<-feNoEd_long%>%filter(Country.Code=="USA")%>%filter(age=="40-44")

# select the columns including year and percentage of female population with no education
feNoEd_kor_summary<-feNoEd_kor_long_age1[,6:7]
feNoEd_usa_summary<-feNoEd_usa_long_age1[,6:7]
names(feNoEd_kor_summary)<-c("year","percentage of female with no education")
names(feNoEd_usa_summary)<-c("year","percentage of female with no education")
```

```
knitr::kable(summary(feNoEd_kor_summary),align=rep('c', 2),caption = "The summary of    percentage of fe
```

Table 1: The summary of percentage of female with age of 40-44 with no education in Korea

year	percentage of female with no education
Min. :1970	Min. : 0.300
1st Qu.:1980	1st Qu.: 0.910
Median :1990	Median : 3.140
Mean :1990	Mean : 8.479
3rd Qu.:2000	3rd Qu.:10.300
Max. :2010	Max. :35.200

```
knitr::kable(summary(feNoEd_usa_summary),align=rep(’c’, 2),caption = "The summary of    percentage of fe
```

Table 2: The summary of percentage of female with age of 40-44 with no education in USA

year	percentage of female with no education
Min. :1970	Min. :0.1500
1st Qu.:1980	1st Qu.:0.3100
Median :1990	Median :0.4900
Mean :1990	Mean :0.4978
3rd Qu.:2000	3rd Qu.:0.6000
Max. :2010	Max. :0.8100

Problem 4

Using base plotting functions, create a single figure that is composed of the first two rows of plots from SAS's simple linear regression diagnostics as shown here: https://support.sas.com/rnd/app/ODSGraphics/examples/reg.html. Demonstrate the plot using suitable data from problem 3.

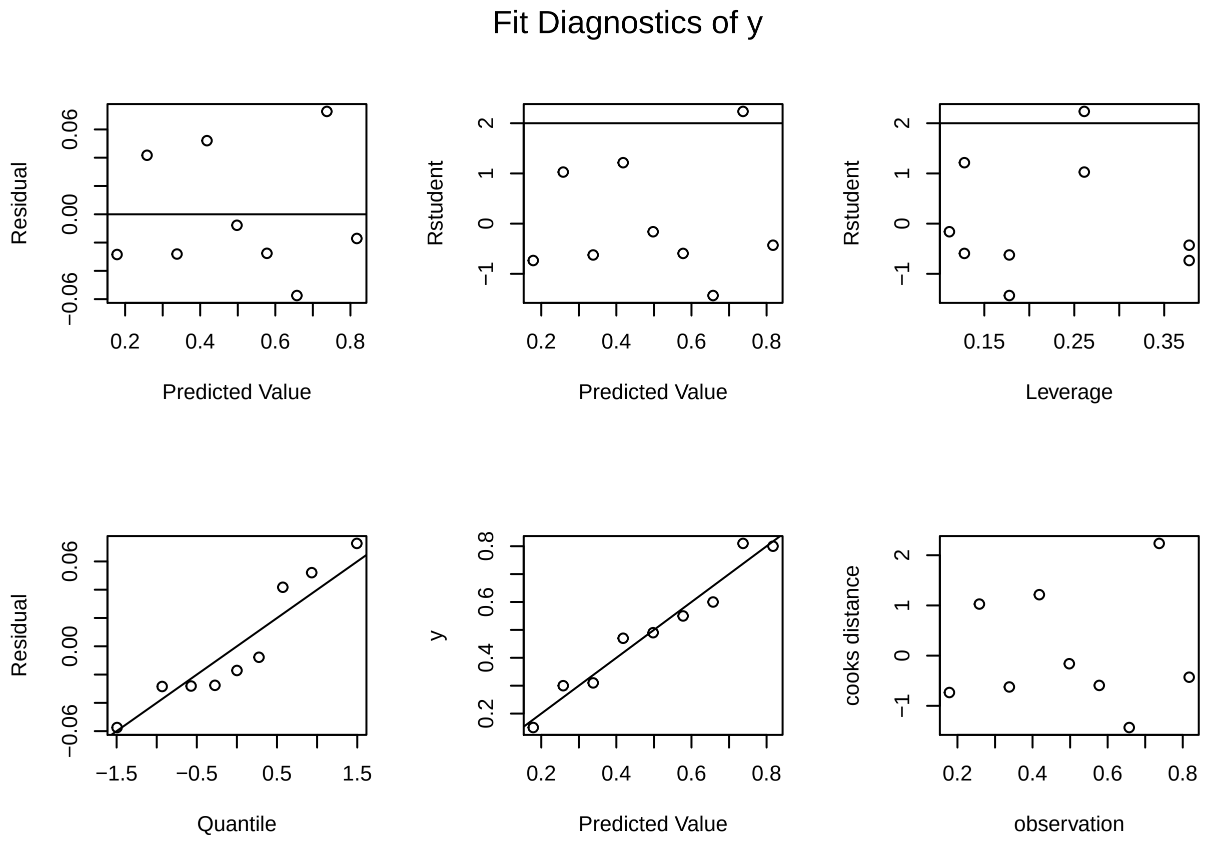
```
# fit regression
fit=lm(value~year,feNoEd_kor_long_age1)
```

```
library(sur)
library(stats)

# create values for fit diagnostics
fit=lm(value~year,feNoEd_usa_long_age1)
Edres<-resid(fit)
yhat<-fitted(fit)
lev<-leverage(fit)
rst<-rstudent(fit)
cooks<-cooks.distance(fit)
y<-feNoEd_usa_long_age1$value
q<-qnorm(ppoints(length(Edres)))
x<-1:9
```

```
#plot diagnostics with base R plot
par(mfrow=c(2,3))
plot(x=yhat,y=Edres,xlab="Predicted Value", ylab="Residual")
abline(h=0)
plot(x=yhat,y=rst,xlab="Predicted Value", ylab="Rstudent")
abline(h=2)
abline(h=-2)
plot(x=lev,y=rst,xlab="Leverage", ylab="Rstudent")
abline(h=2)
abline(h=-2)
qqplot(q,Edres,xlab="Quantile", ylab="Residual")
abline(0,0.12/3)
plot(x=yhat,y=y,xlab="Predicted Value", ylab="y")
abline(0,1)
```

```r
plot(x=yhat,y=rst,xlab="observation", ylab="cooks distance")
mtext("Fit Diagnostics of y", side = 3, line = -1.5, outer = TRUE)
```



Problem 5

Recreate the plot in problem 3 using ggplot2 functions. Note: there are many extension libraries for ggplot, you will probably find an extension to the ggplot2 functionality will do exactly what you want.

```r
#plot diagnostics with ggplot
p1<-ggplot()+
  geom_point(aes(x=yhat,y=Edres))+
  geom_abline(intercept = 0, slope = 0)+
  labs(x="Predicted Value",y="Residual")+
  theme_bw()


p2<-ggplot()+
  geom_point(aes(x=yhat,y=rst))+
  geom_abline(intercept = 2, slope = 0)+
  geom_abline(intercept = -2, slope = 0)+
  labs(x="Predicted Value",y="Rstudent")+
  theme_bw()
```

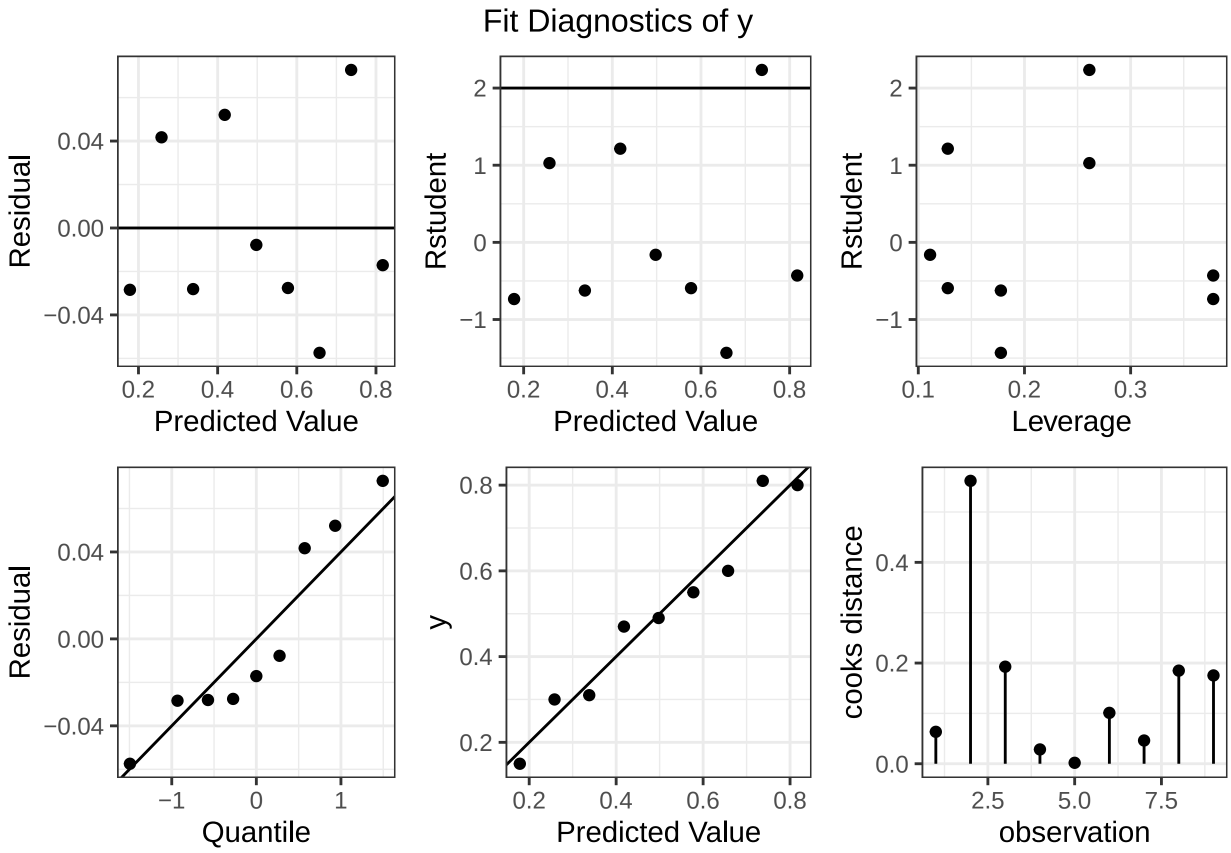
```
p3<-ggplot()+
  geom_point(aes(x=lev,y=rst))+
  labs(x="Leverage",y="Rstudent")+
  theme_bw()

p4<-ggplot()+
  stat_qq(aes(sample=Edres))+
  geom_abline(intercept = 0, slope = 0.12/3)+
  labs(x="Quantile",y="Residual")+
  theme_bw()

p5<-ggplot()+
  geom_point(aes(x=yhat,y=y))+
  geom_abline(intercept = 0, slope = 1)+
  labs(x="Predicted Value",y="y")+
  theme_bw()


p6<-ggplot() +
  geom_point(aes(x=x, y=cooks)) +
  geom_segment( aes(x=x, xend=x, y=0, yend=cooks))+
  labs(x="observation",y="cooks distance")+
  theme_bw()

gridExtra::grid.arrange(p1,p2,p3,p4,p5,p6, ncol=3, nrow=2, top = "Fit Diagnostics of y")
```


Fit Diagnostics of y
Residual
Predicted Value
Rstudent
Predicted Value
Rstudent
Leverage
Residual
Quantile
y
Predicted Value
cooks distance
observation