HW5_Liu

Xueying Liu

10/27/2020

Problem 3

```r
dim(EdStatsData)
```

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

There are 886930 data in total in the complete data set.

```r
# clean the data
# We will only consider the data till 2020
cleaned <- EdStatsData[,1:53]

# remove the row that is NA in all the year from 1970 to 2020
cleaned <- cleaned[apply(cleaned[,5:53],1,function(x)any(!is.na(x))),]
dim(cleaned)
```

```
## [1] 354575     53
```

After we removed the rows that the data is missing from 1970 to 2020, there are 354575 data remains.

```r
## choose 2 countries
country <- rbind(cleaned[cleaned$`Country Code`=="EAP",], cleaned[cleaned$`Country Code`=="ARB",])
country$`Indicator Code`<- factor(country$`Indicator Code`)

## Choose useful indicator to compare
enrolment <- apply(country[country$`Indicator Name`=="Adjusted net enrolment rate, primary, both sexes
population <- apply(country[country$`Indicator Name`=="Population, total",5:53],1,mean,na.rm=TRUE)
unemployment <- apply(country[country$`Indicator Name`=="Unemployment, total (% of total labor force)",5
literacy <- apply(country[country$`Indicator Name`=="Adult literacy rate, population 15+ years, both se
primary <- apply(country[country$`Indicator Name`=="Enrolment in primary education, both sexes (number)
secondary <- apply(country[country$`Indicator Name`=="Enrolment in secondary education, both sexes (numb
tertiary <- apply(country[country$`Indicator Name`=="Enrolment in tertiary education, all programmes, bo
GDP <- apply(country[country$`Indicator Name`=="GDP per capita (current US$)",5:53],1,mean,na.rm=TRUE)
GNI <- apply(country[country$`Indicator Name`=="GNI (current US$)",5:53],1,mean,na.rm=TRUE)
labor <- apply(country[country$`Indicator Name`=="Labor force, total",5:53],1,mean,na.rm=TRUE)
populationgrowth <- apply(country[country$`Indicator Name`=="Population growth (annual %)",5:53],1,mean
table <- format(rbind(population,populationgrowth,GDP,GNI,enrolment,unemployment,labor,literacy,
                      primary,secondary,tertiary),scientific = FALSE,digits = 2)
```

	East Asia & Pacific	Arab World
Population	1618223153.4	233553870.8
Population growth (annual %)	1.4	2.6
GDP	1260.6	2549.4
GNI	2342408277734.5	759146544901.7

	East Asia & Pacific	Arab World
Adjusted net enrolment rate	94.9	72.2
Unemployment, total (% of total labor force)	4.5	12.4
Labor force, total	1005704439.2	89649511.3
Adult literacy rate, population 15+ years, both sexes (%)	88.4	66.0
Enrolment in primary education, both sexes (number)	189592838.8	30077465.6
Enrolment in secondary education, both sexes (number)	97223200.0	17820070.9
Enrolment in tertiary education, all programmes, both sexes (number)	17068139.2	3939575.5

This is the summary table of the average of these indicators over 30 years from 1970 to 2020 in East Asia & Pacific area and Arab World area.

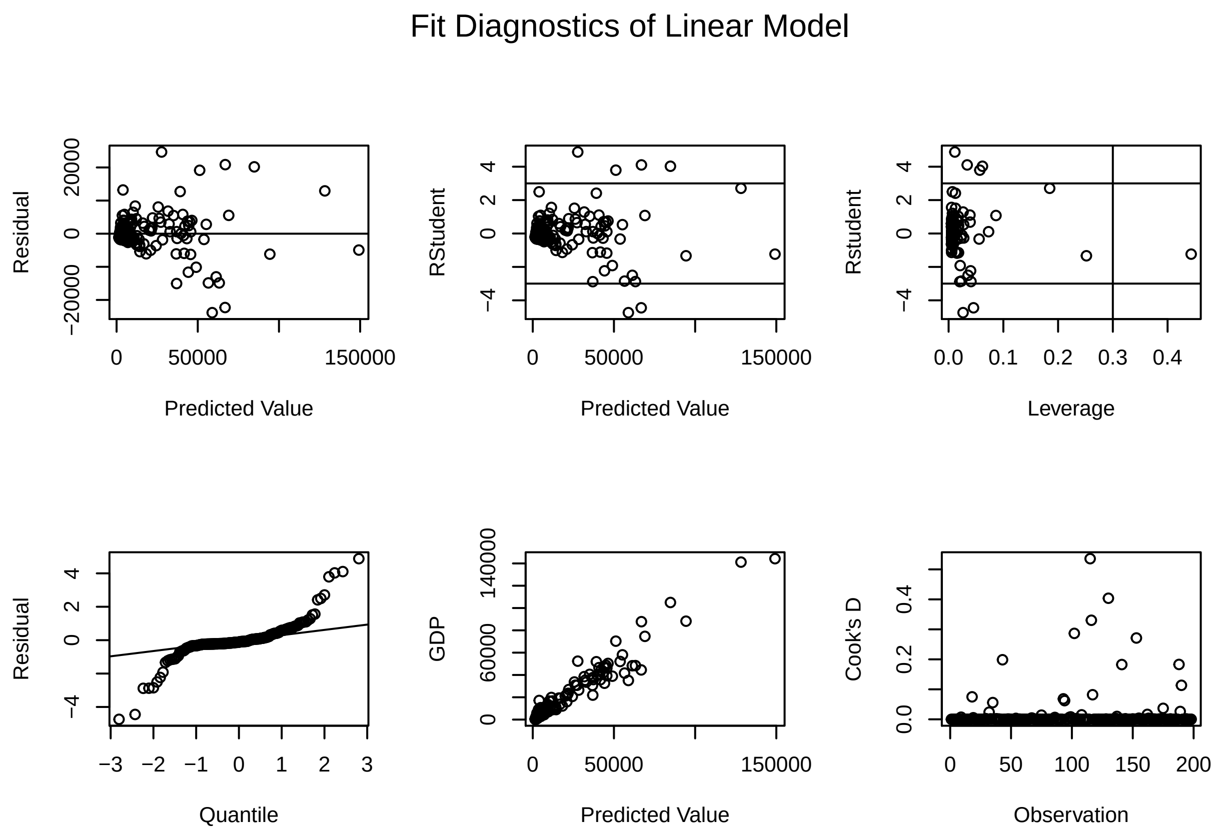
Problem 4

In this problem, we are going to model the linear relationship between the GDP per capita in 2010 and the GDP per capita in 2000 and 1990.

```
## Only explore GDP indecator in each country
data.4 <- cleaned[cleaned$`Indicator Name`=="GDP per capita (current US$)",]

## select independent variable 1990 and 2000 and dependent variable 2010
data.4 <- data.4[complete.cases(data.4[ , c('1990','2000','2010')]),]

## Do the linear regression
fit <- lm(data.4$`2010`~data.4$`1990`+data.4$`2000`)
```

Create plots for this linear regression model:

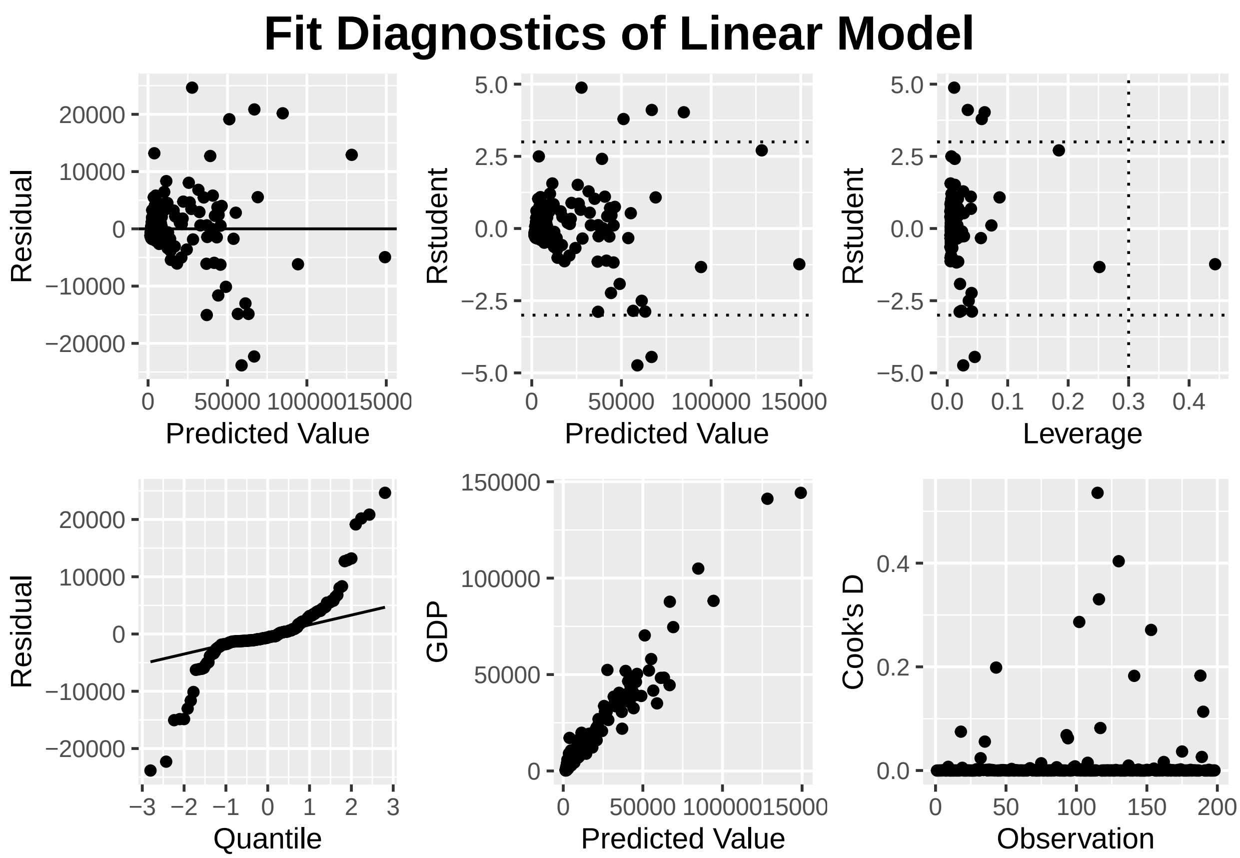
```
par(mfrow=c(2,3),oma = c(0, 0, 2, 0))
plot(predict(fit),residuals(fit),xlab = "Predicted Value",ylab = "Residual")
abline(h=0)



plot(predict(fit),rstudent(fit),xlab = "Predicted Value",ylab = "RStudent")
abline(h=-3)
abline(h=3)

plot(leverage(fit),rstudent(fit),xlab = "Leverage",ylab = "Rstudent")
abline(h=-3)
abline(h=3)
abline(v=0.3)

qqnorm(rstudent(fit),xlab = "Quantile",ylab = "Residual", main = "")
qqline(rstudent(fit))

plot(predict(fit),data.4$`2010`,xlab = "Predicted Value",ylab = "GDP")

cooks <- cooks.distance(fit)
plot(cooks,xlab = "Observation",ylab = "Cook's D")
mtext("Fit Diagnostics of Linear Model",outer = TRUE)
```



Problem 5

We can recreate the plot in problem 3 using ggplot2 functions.

```
fig1 <- qplot(x=predict(fit),y=residuals(fit))+xlab("Predicted Value")+ylab("Residual")+
  geom_hline(yintercept =0)

fig2 <- qplot(x=predict(fit),y=rstudent(fit))+xlab("Predicted Value")+ylab("Rstudent")+
  geom_hline(yintercept =3,linetype="dotted")+geom_hline(yintercept =-3,linetype="dotted")

fig3 <- qplot(x=leverage(fit),y=rstudent(fit))+xlab("Leverage")+ylab("Rstudent")+
  geom_hline(yintercept =3,linetype="dotted")+geom_hline(yintercept=-3,linetype="dotted")+
  geom_vline(xintercept =0.3,linetype="dotted")

fig4 <- ggplot()+geom_qq(aes(sample=residuals(fit)))+geom_qq_line(aes(sample=residuals(fit)))+
  xlab("Quantile")+ylab("Residual")

fig5 <- qplot(x=predict(fit),y=data.4$`2010`)+xlab("Predicted Value")+ylab("GDP")

fig6 <- qplot(y=cooks)+xlab("Observation")+ylab("Cook's D")

figlist <- list(fig1,fig2,fig3,fig4,fig5,fig6)
fitdiag <- ggarrange(plotlist = figlist, nrow = 2, ncol = ceiling(length(figlist)/2))+ggtitle("Fit Diag
annotate_figure(fitdiag,
                top = text_grob("Fit Diagnostics of Linear Model", face = "bold", size = 18))
```


Fit Diagnostics of Linear Model
Residual
Predicted Value
Rstudent
Predicted Value
Rstudent
Leverage
Residual
Quantile
GDP
Predicted Value
Cook's D
Observation