Homework 5

Due Wednesday Nov 4, 2020

Yixin Chen

2020-11-02

Problem 3

```
#library(downloader)
#download("http://databank.worldbank.org/data/download/Edstats_csv.zip",dest="Edstats_csv.zip")
#unzip("world_bank.zip", exdir="./")
Edstats<-read.csv("/Users/volderay/STAT5014_yixinc/Edstats_csv/EdStatsData.csv")
dim(Edstats)
```

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

```
#complete re-organized data
Edstats.complete <- Edstats %>%
     gather(key = "Year", value = "value", 5:70)
dim(Edstats.complete)[1]
```

```
## [1] 58527380
```

```
#Clean data
#Delete rows with "NA" values
Edstats.clean <- Edstats[,-70]
Edstats.clean <- Edstats.clean %>%
     gather(key = "Year", value = "Value", 5:69, na.rm = TRUE)
dim(Edstats.clean)[1]
dat<-Edstats.clean
```

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

In the complete dataset(including `NA` values), there are $886930 \times (70-4) = 58537380$ data points(observations). After deleting the observations with `NA` values, there are 5082201 data points left in the clean dataset.

I chose China and USA.

```
CHN.ind<-as.factor(dat[dat$Country.Code == "CHN",3])
USA.ind<-as.factor(dat[dat$Country.Code == "USA",3])
sum.ind<-data.frame( t(c(length(levels(CHN.ind)),length(levels(USA.ind)))) )
colnames(sum.ind)<-c("CHN","USA")
rownames(sum.ind)<-"Number of Indicators"
kable(sum.ind,"latex", booktabs = T) %>% kable_styling(position = "center")
```

	CHN	USA
Number of Indicators	1703	1870

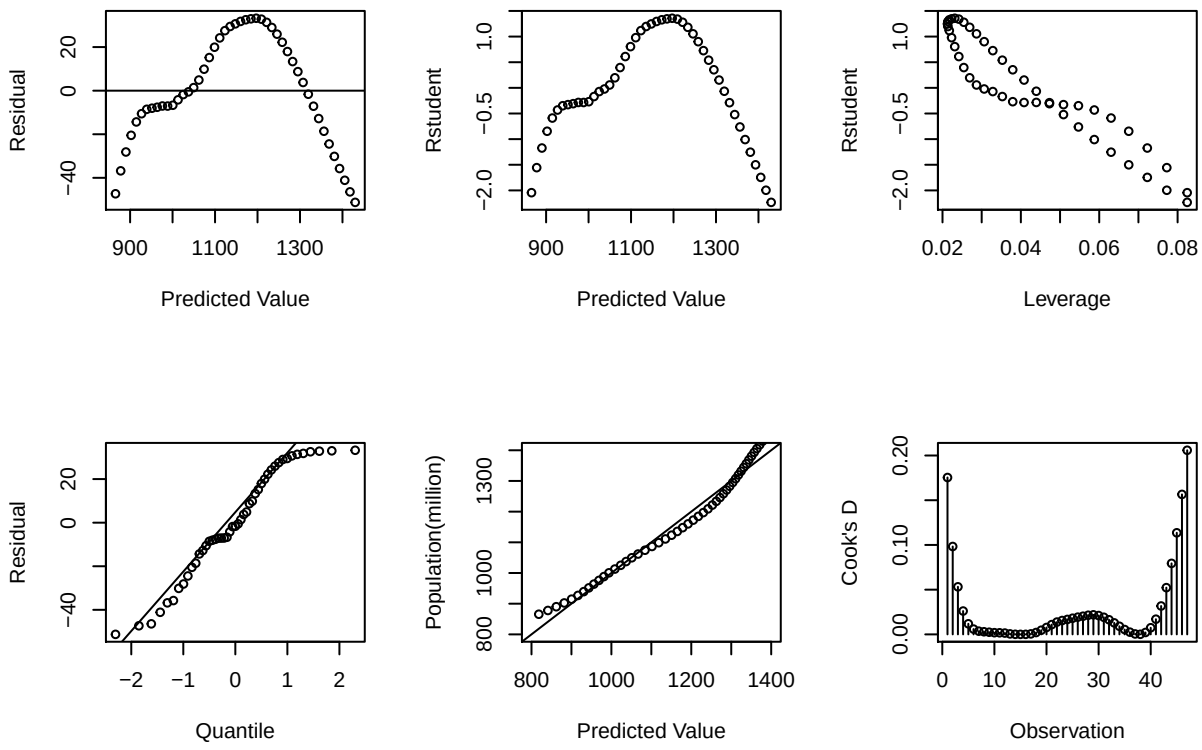
Problem 4

```
#China population(Y) and year(X)
CHN<-dat[dat$Country.Code == "CHN",]
CHN.pop<-CHN[CHN$Indicator.Code == "SP.POP.TOTL",5:6]
#Turn Year into numeric values
year<-sapply(CHN.pop$Year, str_split,pattern = "X")
CHN.pop$Year<-as.numeric(sapply(year, function(x){x[2]}))
#rescale
CHN.pop$Value<-CHN.pop$Value/1e+6
fit<-lm(Value~Year,data = CHN.pop)
library(MASS)
```

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

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

```
par(mfrow=c(2,3))
#par(mar = c(3,3,3,3)) # Set the margin on all sides to 2
#par(cex.axis = 0.6)
#fitted vs. Residual
plot(fitted(fit),residuals(fit),xlab = "Predicted Value",ylab = "Residual",cex=0.8)
abline(h=0)
#fitted vs. Studentized Residual
plot(fitted(fit),studres(fit),xlab = "Predicted Value",ylab = "Rstudent",cex=0.8)
#Leverage vs. Studentized Residual
plot(hatvalues(fit),studres(fit),xlab = "Leverage",ylab = "Rstudent",cex=0.8)
#QQ-plot
qqnorm(residuals(fit),xlab = "Quantile",ylab = "Residual",main = "",cex=0.8)
qqline(residuals(fit))
#fitted vs. actual
plot(CHN.pop$Value,fitted(fit),xlab = "Predicted Value",ylab = "Population(million)",xlim = c(800,1400)
segments(700,700,1500,1500)
#Cook's D
plot(c(1:47),cooks.distance(fit),xlab="Observation",ylab="Cook's D",cex=0.8)
segments(1:47,0,1:47,cooks.distance(fit))
```



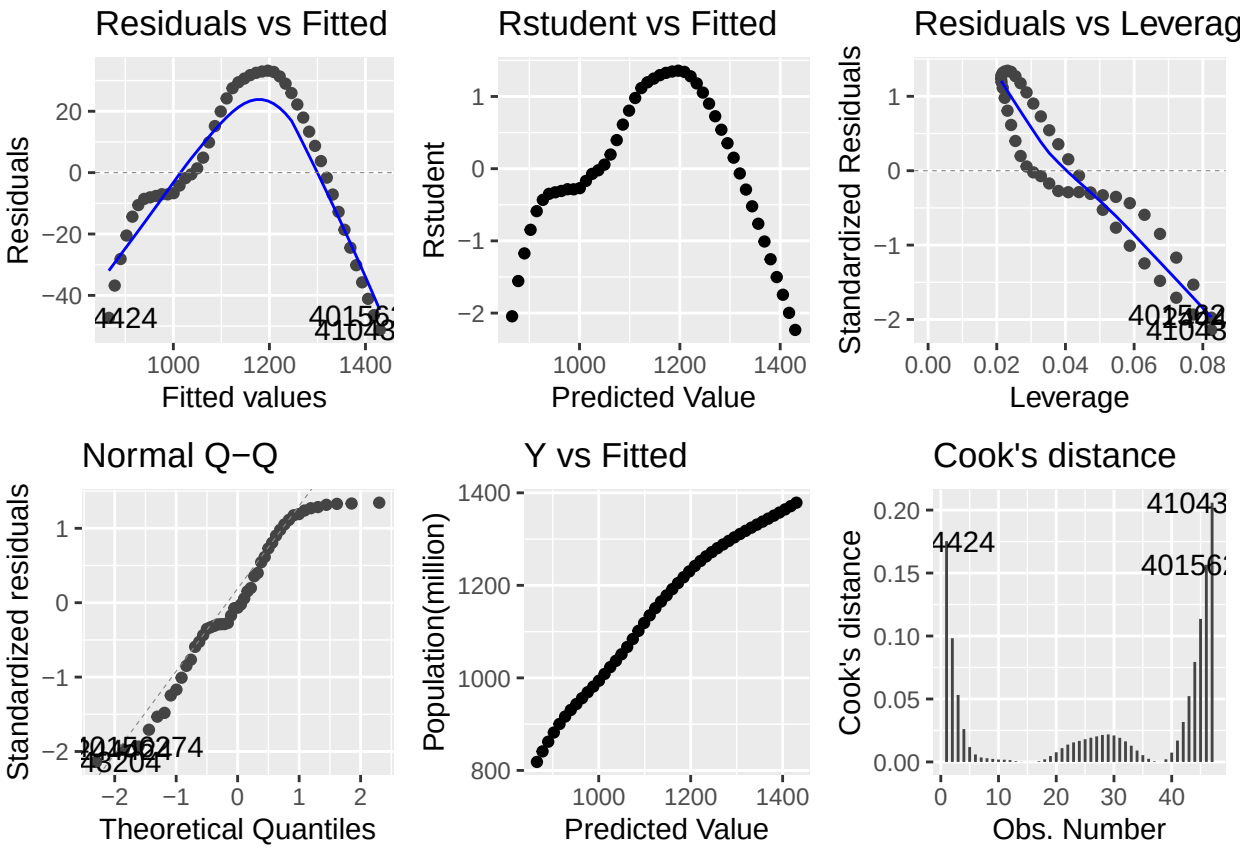
Using *base* plotting functions, create a single figure that is composed of the first two rows of plots from SAS's simple linear regression diagnostics as shown here: https://support.sas.com/rnd/app/ODSGraphics/examples/reg.html. Demonstrate the plot using suitable data from problem 3.

Problem 5

```
library(ggfortify)
library(ggpubr)
plots<-autoplot(fit,which = 1:6)
```

```
## Warning: `arrange_()` is deprecated as of dplyr 0.7.0.
## Please use `arrange()` instead.
## See vignette('programming') for more help
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_warnings()` to see where this warning was generated.
```

```
fitstu<-data.frame(cbind(fitted(fit),studres(fit)))
fitvalue<-data.frame(cbind(fitted(fit), CHN.pop$Value))
fit_stures<-ggplot(fitstu,aes(x=X1,y=X2))+geom_point()+xlab("Predicted Value")+ylab("Rstudent")+ggtitle
fit_value<-ggplot(fitvalue,aes(x=X1,y=X2))+geom_point()+xlab("Predicted Value")+ylab("Population(millio
ggarrange(plots[[1]], fit_stures, plots[[5]], plots[[2]], fit_value, plots[[4]],nrow = 2,ncol = 3)
```



Problem 6

Finish this homework by pushing your changes to your repo.

Only submit the .Rmd and .pdf solution files. Names should be formatted HW5_lastname_firstname.Rmd and HW5_lastname_firstname.pdf