HW5_tianjiang

Tian Jiang

10/30/2020

P3

How many data points were there in the complete dataset?

```
df<-read.csv("/Users/mac/Documents/5014/Edstats_csv/EdStatsData.csv", header = TRUE, sep = ",")
China<-df[df$Country.Name=='China',]
UnitedStates<-df[df$Country.Name=='United States',]
China1<-China[rowSums(is.na(China)) != 66, ]
UnitedStates1<-UnitedStates[rowSums(is.na(UnitedStates)) != 66, ]
China2<-China[seq(152,158,2),]
UnitedStates2<-UnitedStates[seq(152,158,2),]
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
China3<-China2 %>% select_if(~all(!is.na(.)))
China5<-China3 %>% rowwise() %>% mutate(mean = mean(c_across(X1970:X2010)))
China6<-China5 %>% rowwise() %>% mutate(max = max(c_across(X1970:X2010)))
China7<-China6 %>% rowwise() %>% mutate(min = min(c_across(X1970:X2010)))
China8<-China7 %>% rowwise() %>% mutate(median = median(c_across(X1970:X2010)))
summaryChina<-China8[,c(3,14:17)]
summaryChina
```

```
## # A tibble: 4 x 5
## # Rowwise:
##   Indicator.Name                                          mean   max   min median
##   <chr>                                                  <dbl> <dbl> <dbl>  <dbl>
## 1 Barro-Lee: Average years of primary schooling, age 1~  5.11  5.5   4.77   4.98
## 2 Barro-Lee: Average years of primary schooling, age 2~  4.97  5.5   4.41   4.93
## 3 Barro-Lee: Average years of primary schooling, age 2~  3.41  4.81  1.6    3.6
## 4 Barro-Lee: Average years of primary schooling, age 2~  4.85  5.5   3.89   4.89
```

```
UnitedStates3<-UnitedStates2 %>% select_if(~all(!is.na(.)))
UnitedStates5<-UnitedStates3 %>% rowwise() %>% mutate(mean = mean(c_across(X1970:X2010)))
UnitedStates6<-UnitedStates5 %>% rowwise() %>% mutate(max = max(c_across(X1970:X2010)))
UnitedStates7<-UnitedStates6 %>% rowwise() %>% mutate(min = min(c_across(X1970:X2010)))
UnitedStates8<-UnitedStates7 %>% rowwise() %>% mutate(median = median(c_across(X1970:X2010)))
summaryUnitedStates<-UnitedStates8[,c(3,14:17)]
summaryUnitedStates
```

```
## # A tibble: 4 x 5
## # Rowwise:
##   Indicator.Name                                          mean   max   min median
##   <chr>                                                  <dbl> <dbl> <dbl>  <dbl>
## 1 Barro-Lee: Average years of primary schooling, age 1~  5.96  6     5.92   5.97
## 2 Barro-Lee: Average years of primary schooling, age 2~  5.96  5.99  5.91   5.97
## 3 Barro-Lee: Average years of primary schooling, age 2~  5.88  5.95  5.68   5.91
## 4 Barro-Lee: Average years of primary schooling, age 2~  5.95  5.98  5.86   5.96
```

P4

```
China9<-China3[1:2,4:13]
China10<-t(China9)
colnames(China10)<-c('1519','2024')
China11<-China10[2:10,]
China12<-as.data.frame(China11)
China12$`1519`<-as.numeric(China12$`1519`)
China12$`2024`<-as.numeric(China12$`2024`)
lmfit<-lm(China12$`2024`~China12$`1519`)
summary(lmfit)
```

```
##
## Call:
## lm(formula = China12$`2024` ~ China12$`1519`)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.48294 -0.12672  0.00223  0.15758  0.70539
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)      2.3259     2.4606   0.945    0.376
## China12$`1519`   0.5175     0.4810   1.076    0.318
##
## Residual standard error: 0.3749 on 7 degrees of freedom
## Multiple R-squared:  0.1419, Adjusted R-squared:  0.01933
## F-statistic: 1.158 on 1 and 7 DF,  p-value: 0.3176
```

```
par(mfrow=c(2,3))

plot(fitted(lmfit),residuals(lmfit),xlab="Fitted",ylab="Residuals")
```

```r
abline(h=0)

plot(x = fitted.values(lmfit), y = rstudent(lmfit))

x <- model.matrix(lmfit)
lev <- hat(x)
sum(lev)
```

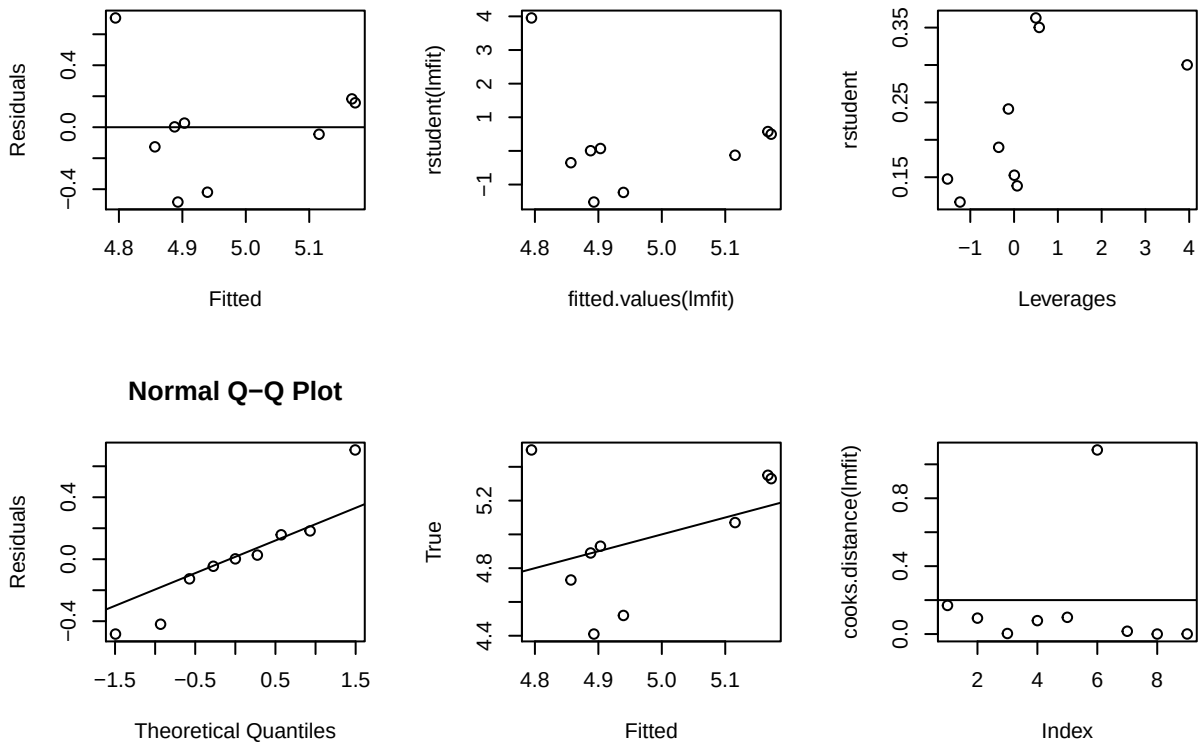
```
## [1] 2
```

```r
plot(rstudent(lmfit),lev,xlab="Leverages",ylab="rstudent")
abline(h=2*sum(lev)/9)

qqnorm(residuals(lmfit), ylab="Residuals")
qqline(residuals(lmfit))#????????????????????????????????????????different from ggplot     normal

plot(fitted(lmfit),China12$`2024`,xlab="Fitted",ylab="True")
abline(coef = c(0,1))

plot(cooks.distance(lmfit))
abline(h=0.2)
```




```r
#layout matrix
```

P5

```
#install.packages("broom")
library(broom)
library(ggplot2)
require(gridExtra)
```

```
## Loading required package: gridExtra
```

```
##
## Attaching package: 'gridExtra'
```

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

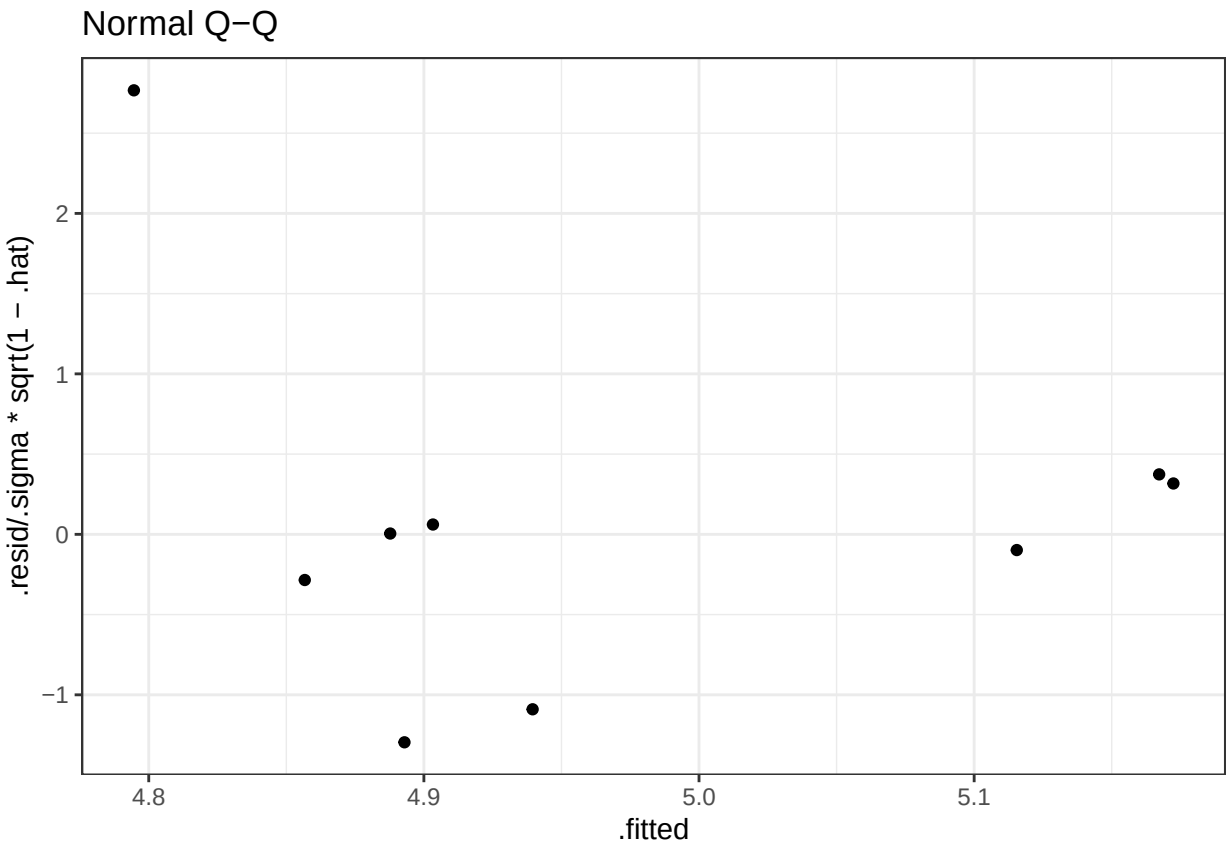
```
df <- augment(lmfit)

p1<-ggplot(df, aes(x = .fitted, y = .resid)) + geom_hline(yintercept = 0, linetype=2, color="darkgrey")

p2<-ggplot(df, aes(x = .fitted, y = .resid / .sigma * sqrt(1 - .hat))) +  geom_point()

p3<-ggplot(df, aes(x =.hat, y = .resid / .sigma * sqrt(1 - .hat))) + geom_hline(yintercept = 2*sum(lev),

#?????????????
p4<-ggplot(df,aes(sample=.resid))+geom_qq()
p4<-p2+geom_abline()+xlab("Theoretical Quantiles")+ylab("Residuals")
p4<-p2+ggtitle("Normal Q-Q")+theme_bw()
p4
```

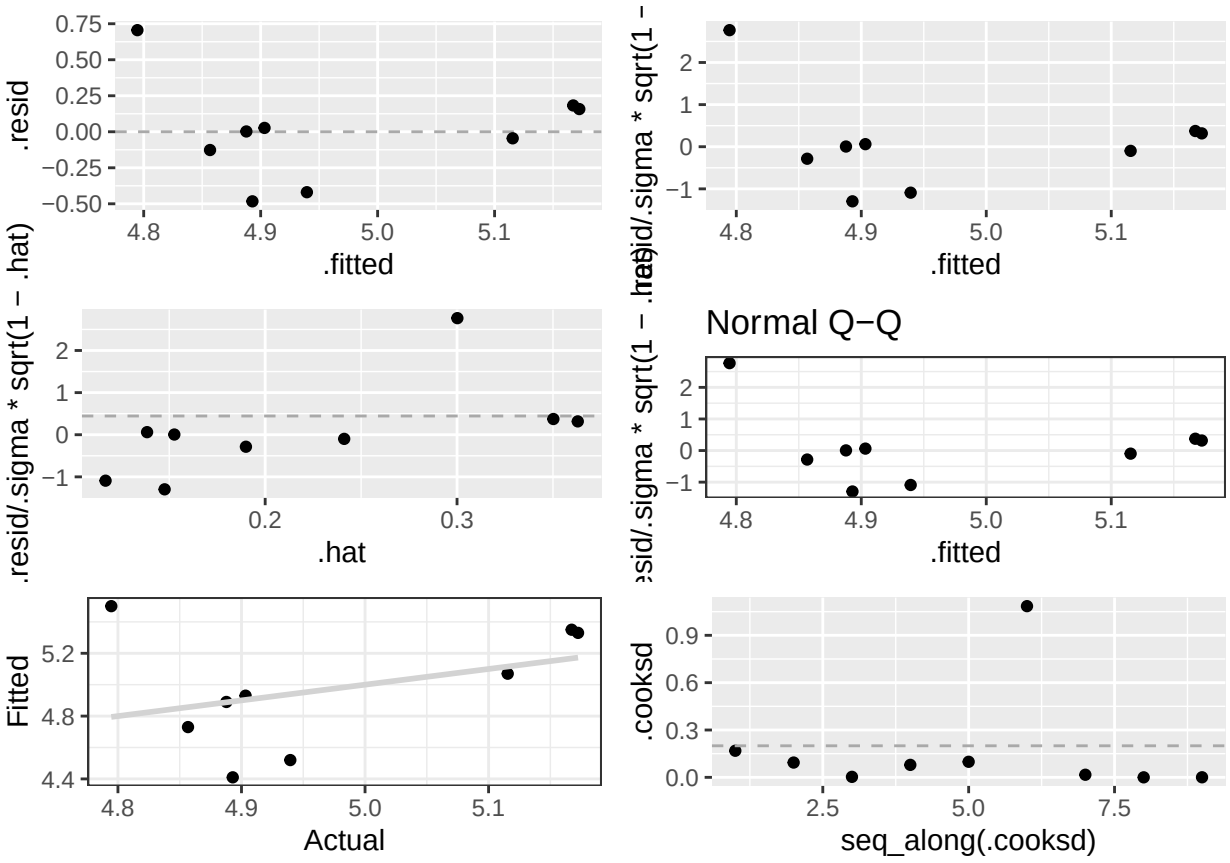



```
p5<-lmfit %>% augment() %>%
  ggplot()  +
  geom_point(aes(.fitted, China12$`2024`)) +
  geom_smooth(aes(.fitted, China12$`2024`), method = "lm", se = FALSE, color = "lightgrey") + labs(x =
  theme_bw()

p6<-ggplot(df, aes(seq_along(.cooksd), .cooksd)) +
  geom_point()+geom_hline(yintercept = 0.2, linetype=2, color="darkgrey")

grid.arrange(p1, p2, p3,p4,p5,p6)
```

```
## `geom_smooth()` using formula 'y ~ x'
```


.resid
.fitted
.resid/.sigma * sqrt(1 − .hat)
.hat
Normal Q−Q
Fitted
Actual
.cooksd
seq_along(.cooksd)