HW5_Shi

Liang Shi

11/3/2020

Problem 3

```
df <- read.csv('Edstats_csv/EdStatsData.csv')
df_clean <- df[apply(df, 1, function(x) !all(is.na(x[5:70]))),]
df_clean[is.na(df_clean)] <- 0
print('Row data:')
```

```
## [1] "Row data:"
```

```
dim(df)
```

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

```
print('Cleaned data:')
```

```
## [1] "Cleaned data:"
```

```
dim(df_clean)
```

```
## [1] 357405     70
```

```
c_1 <- df_clean[df_clean$Country.Code=="CPV",][5:70]
s_1 <- summary(c_1)
kable(s_1, caption = "summary for China")
```

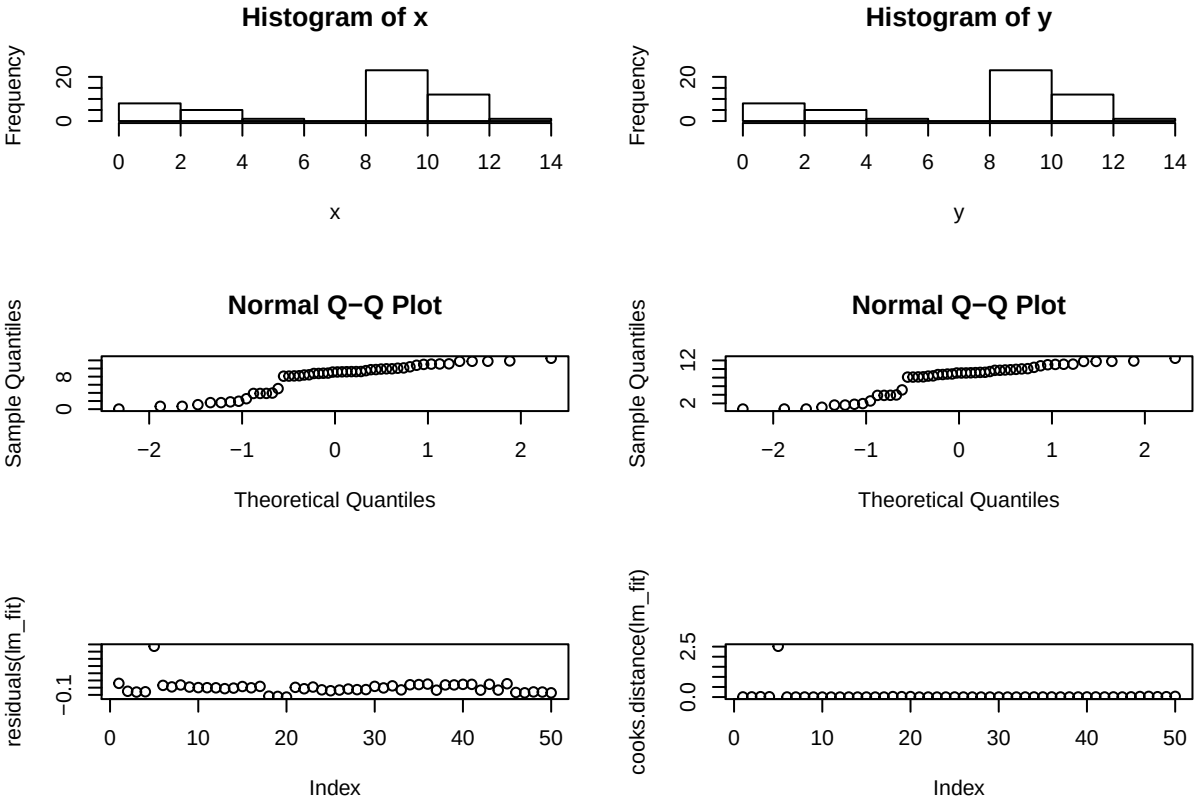
X1970	X1971	X1972	X1973	X1974	X1975	X1976
Min. : 0	Min. : 0	Min. : 0	Min. : -0.24	Min. : -0.26	Min. : 0	Min. : 0
1st Qu.: 0	1st Qu.: 0	1st Qu.: 0	1st Qu.: 0.00	1st Qu.: 0.00	1st Qu.: 0	1st Qu.: 0
Median : 0	Median : 0	Median : 0	Median : 0.00	Median : 0.00	Median : 0	Median : 0
Mean : 1010	Mean : 1050	Mean : 1086	Mean : 1080.94	Mean : 1161.86	Mean : 1174	Mean : 1167
3rd Qu.: 0	3rd Qu.: 0	3rd Qu.: 0	3rd Qu.: 0.00	3rd Qu.: 0.00	3rd Qu.: 0	3rd Qu.: 0
Max. :270198	Max. :272992	Max. :273651	Max. :273005.00	Max. :272292.00	Max. :272423	Max. :273652

```
c_2 <- df_clean[df_clean$Country.Code=="KEN",][5:70]
s_2<- summary(c_2)
kable(s_2, caption = "summary for Kenya")
```

X1970	X1971	X1972	X1973	X1974	X1975
Min. :0.000e+00	Min. :0.000e+00	Min. :0.000e+00	Min. :0.000e+00	Min. :0.000e+00	Min. :0.000
1st Qu.:0.000e+00	1st Qu.:0.000e+00	1st Qu.:0.000e+00	1st Qu.:0.000e+00	1st Qu.:0.000e+00	1st Qu.:0.00
Median :0.000e+00	Median :0.000e+00	Median :0.000e+00	Median :0.000e+00	Median :0.000e+00	Median :0.0
Mean :4.776e+06	Mean :5.658e+06	Mean :6.642e+06	Mean :7.260e+06	Mean :7.910e+06	Mean :8.23
3rd Qu.:0.000e+00	3rd Qu.:0.000e+00	3rd Qu.:0.000e+00	3rd Qu.:0.000e+00	3rd Qu.:0.000e+00	3rd Qu.:0.0
Max. :6.764e+09	Max. :8.263e+09	Max. :9.675e+09	Max. :1.025e+10	Max. :1.066e+10	Max. :1.07

Problem 4

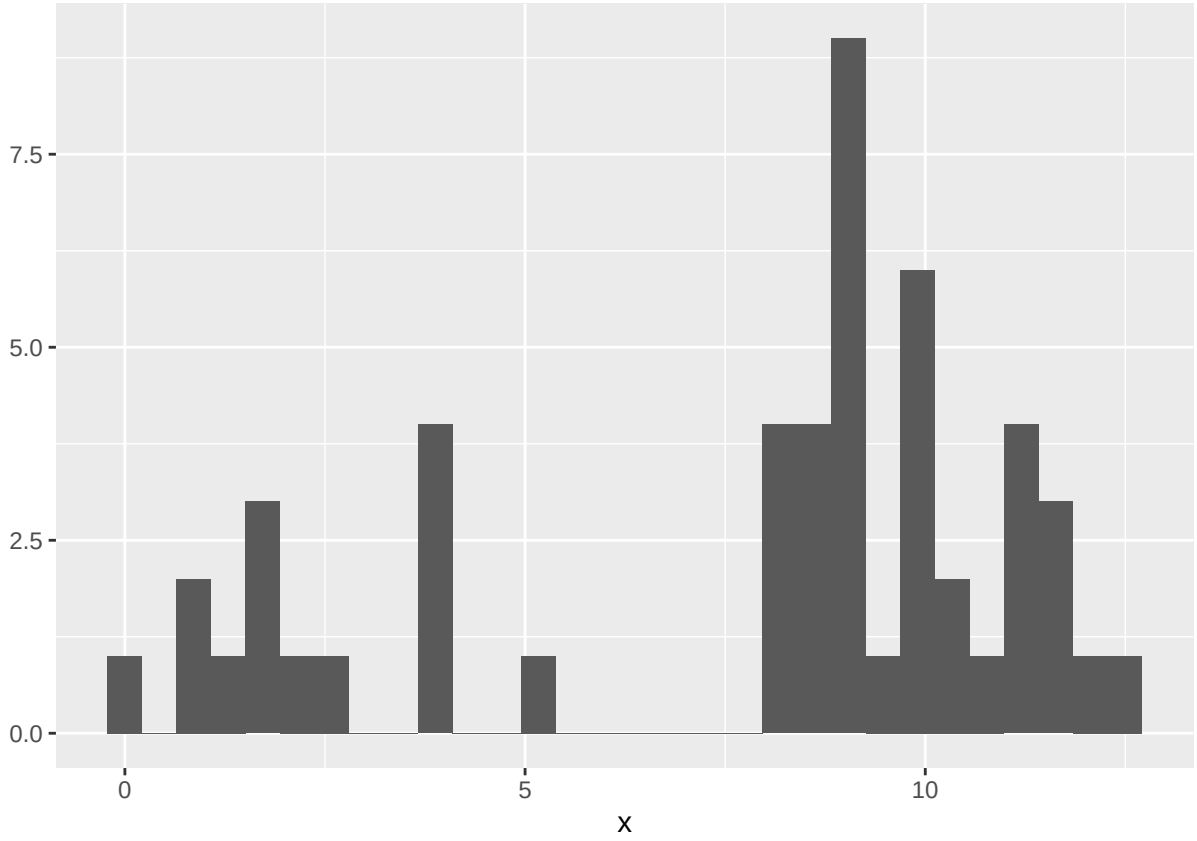
```
df_first_1 <- c_1[1:2]
df_first_1 = df_first_1[which(df_first_1 > 0),  ,drop=T]
df_first_1 <- df_first_1[apply(df_first_1, 1, function(x) !all(is.na(x))),]
y = log(df_first_1$X1970)
x = log(df_first_1$X1971)
lm_fit <- lm(y ~ x)
par(mfrow=c(3,2))
hist(x)
hist(y)
qqnorm(x)
qqnorm(y)
plot(residuals(lm_fit))
plot(cooks.distance(lm_fit))
```



Problem 5

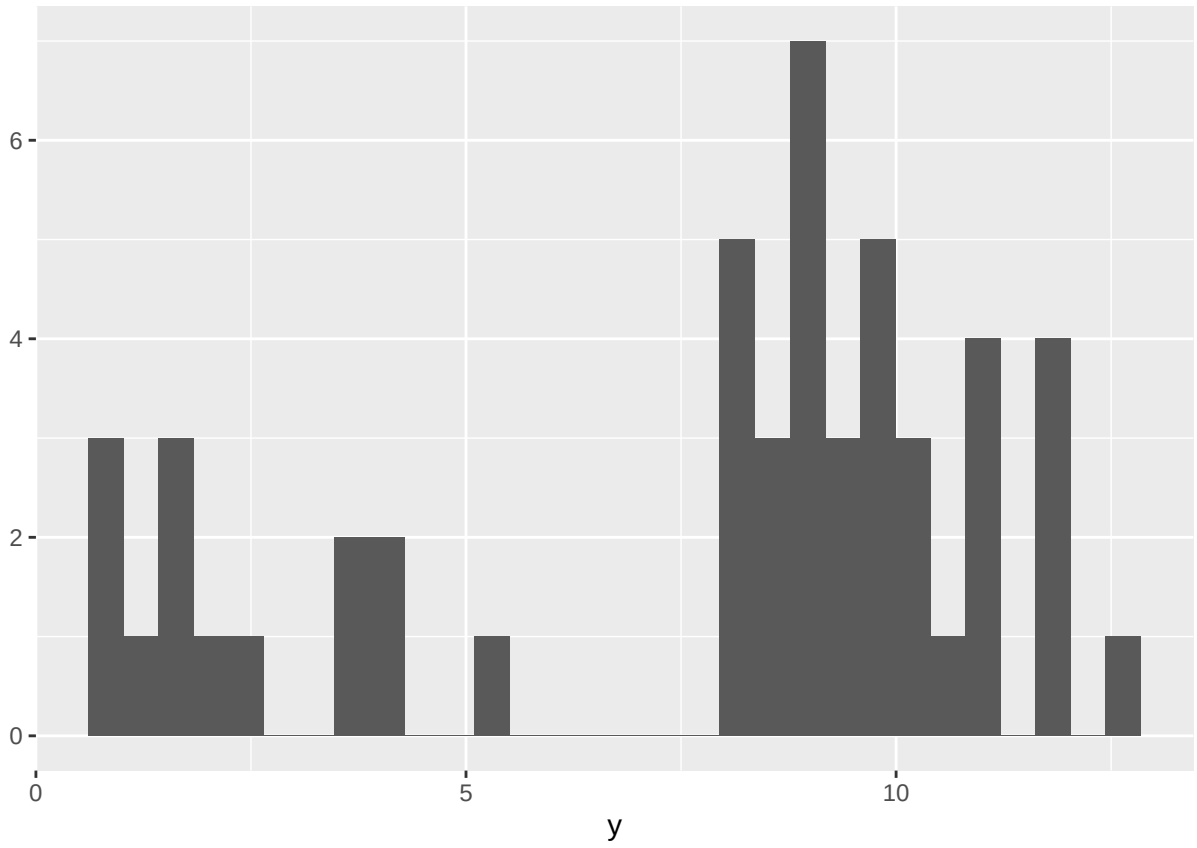
```
qplot(x, geom="histogram")
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

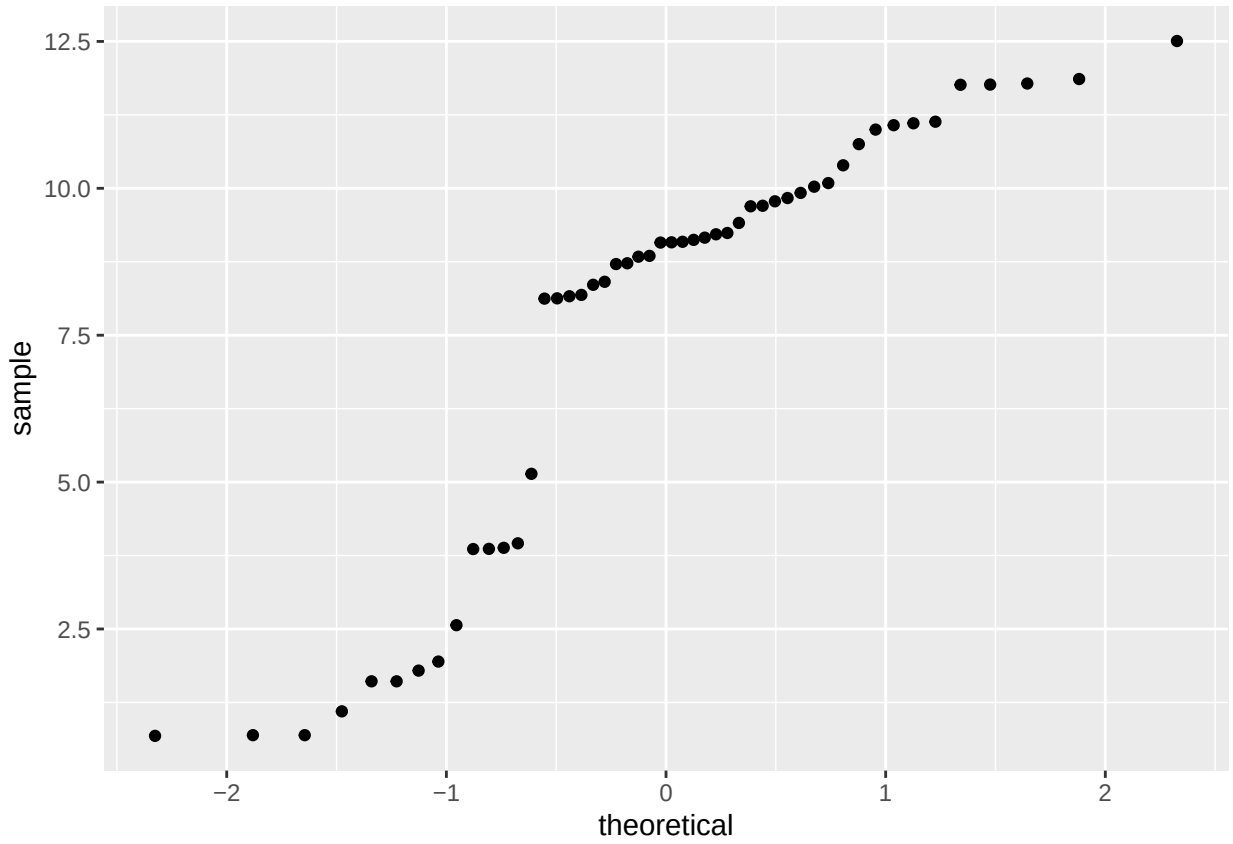



```
qplot(y, geom="histogram")
```

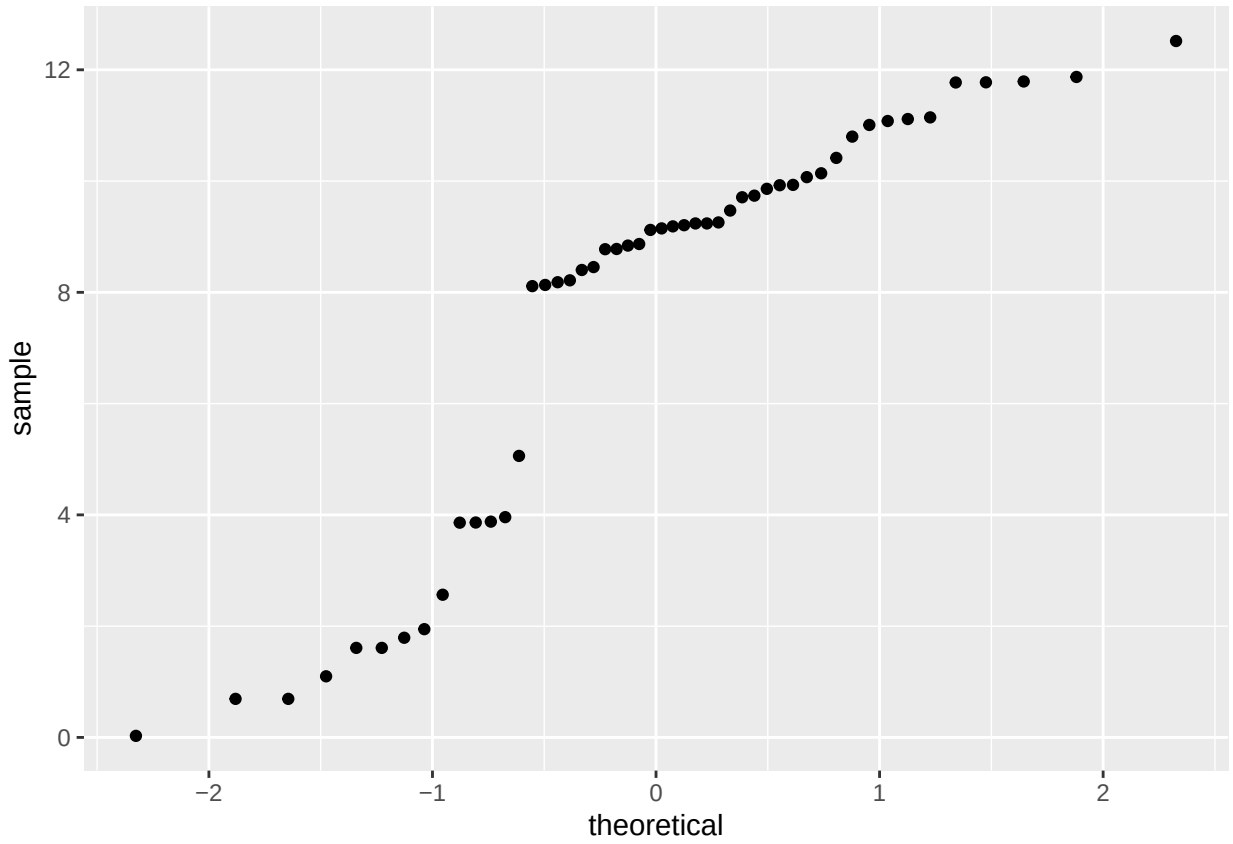
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



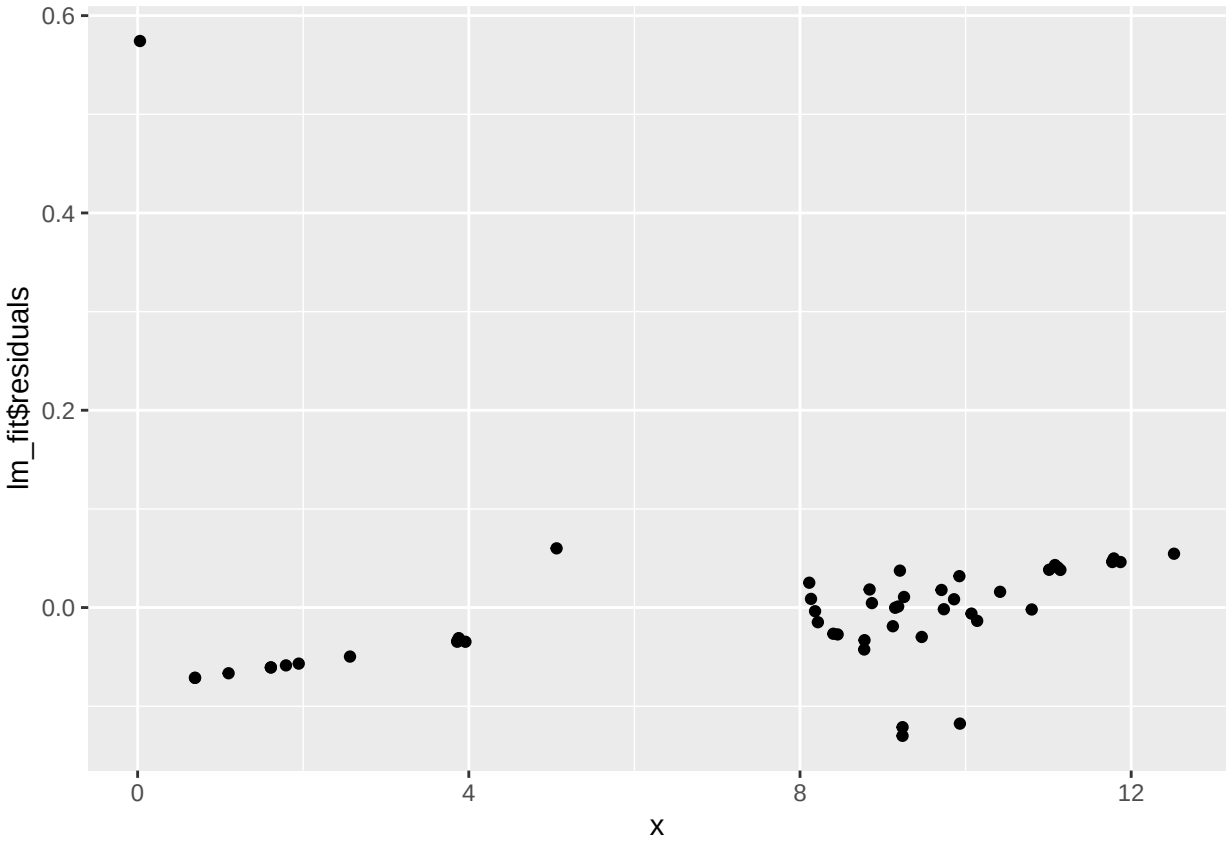

```
ggplot(df_first_1, aes(sample=log(X1970)))+stat_qq()
```



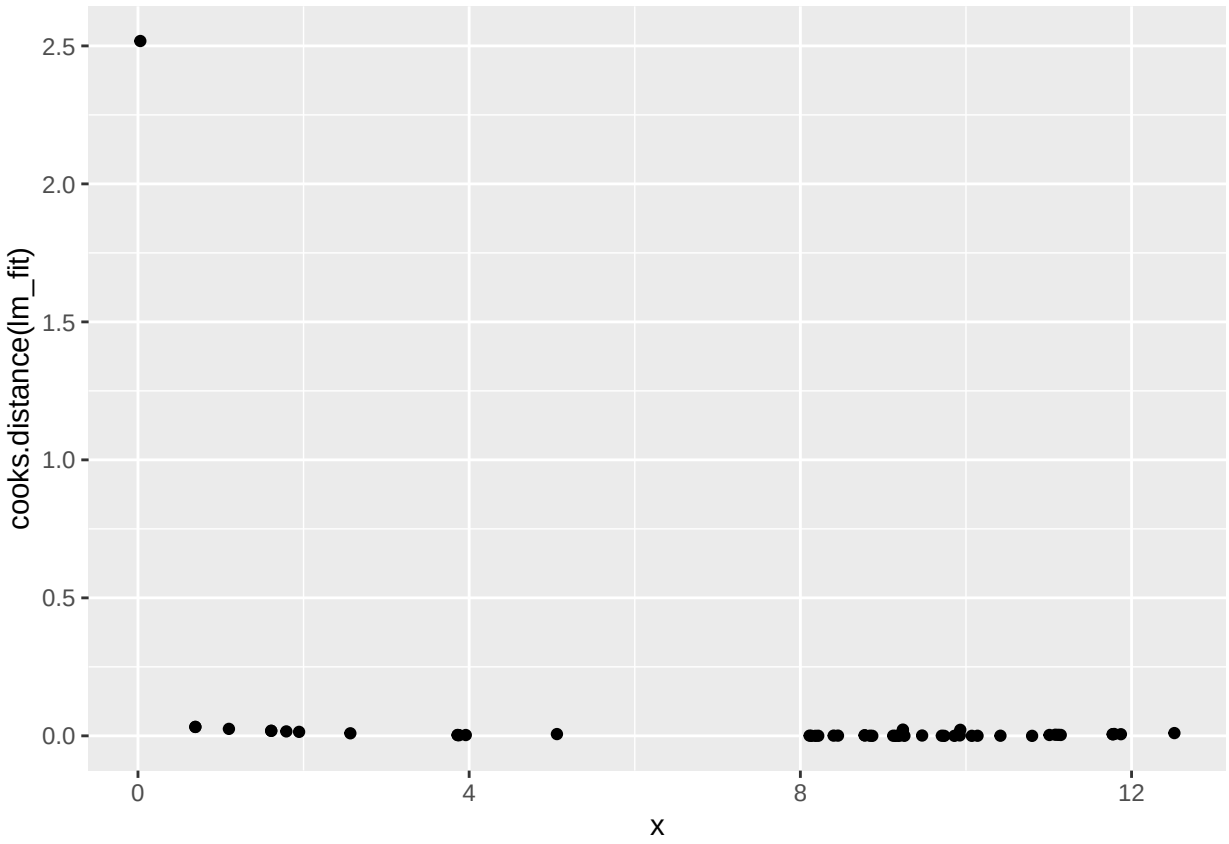

```
ggplot(df_first_1, aes(sample=log(X1971)))+stat_qq()
```




```
ggplot(lm_fit, aes(x=x, y=lm_fit$residuals)) + geom_point()
```




```
ggplot(lm_fit, aes(x=x, y=cooks.distance(lm_fit))) + geom_point()
```


cooks.distance(lm_fit)
2.5
2.0
1.5
1.0
0.5
0.0
0
4
8
12
x