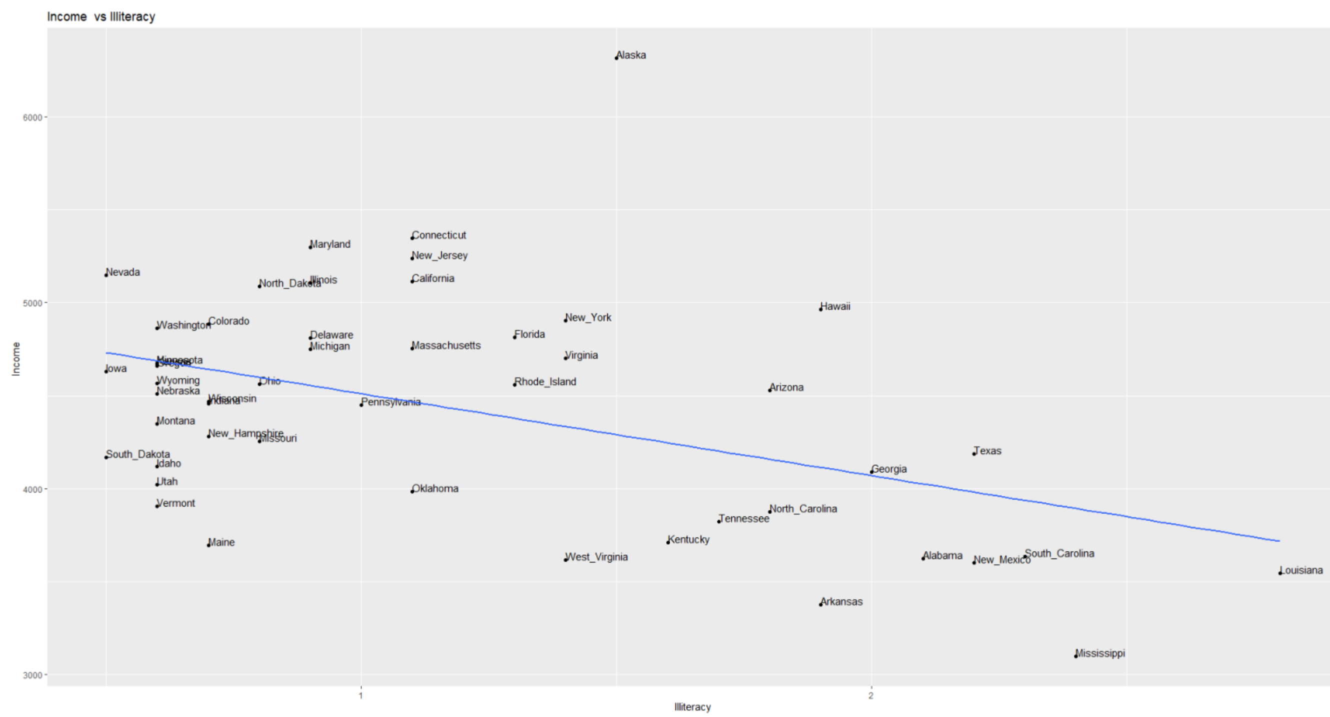
MSIS 5503 – Statistics for Data Science – Fall 2021 - Assignment 12 Solution

Note: Your solution might vary depending on your conclusions on which observations are outliers.

I am showing one of the correct solutions by a student

1) Plot Income vs Illiteracy, with name of the State as text for the data points. Based on a *visual* inspection of this plot, comment on:

```
#
```




```
Clear the Environment
rm(list=ls())
library(MASS)
# Read csv file as a DataFrame
#
setwd(r"{C:\Users\pramodh\Documents\Coursework\MSIS-5503\week13\Assignment}")
df <- read.table('States.csv',
                 header = TRUE, sep = ',')
#Assign variable names to DataFrame Column objects
State <- df$State
Population<-df$Population
Income <- df$Income
Illiteracy<-df$Illiteracy
Life_Exp <- df$Life_Exp
Murder<- df$Murder
HS_Grad <-df$HS_Grad
Frost <- df$Frost
Area <- df$Area
#
library(ggplot2)
#
ggplot(df, aes(x=Illiteracy , y=Income, label=State)) +
  geom_point() +
  geom_text(aes(label=State),hjust=0, vjust=0) +
  geom_smooth(method = "lm", se = FALSE) +
  ggtitle("Income  vs Illiteracy") +
  xlab("Illiteracy") +
  ylab("Income ")
```

a. Linearity, Heteroscedasticity, Potential Outliers
Answer : As shown by the blue line there seems to be a -ve linear relationship between Income and Illiteracy.
The error or the distance of the data points from the linear fit (blue line) seems to be constant on an average for all values of Illiteracy . Thus , the there does not seem to be heteroscedasticity
Alaska , whose Income is very far away from the blue line is one potential outlier

2) **Model 1:**

a. Develop a Regression model that predicts Income using Illiteracy. ***Write your conclusions*** about the model based on the summary output.
Answer :
```
#Model 1 Income based on Illiteracy
mod1 <- lm(Income ~ Illiteracy)
summary(mod1)
```
```
> #Model 1 Income based on Illiteracy
> mod1 <- lm(Income ~ Illiteracy)
> summary(mod1)

Call:
lm(formula = Income ~ Illiteracy)

Residuals:
    Min      1Q  Median      3Q     Max 
-948.89 -376.20  -49.77  347.00 2024.60 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)   4951.3      172.3  28.739  < 2e-16 ***
Illiteracy    -440.6      130.9  -3.367  0.00151 ** 
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 558.4 on 48 degrees of freedom
Multiple R-squared:  0.191,	Adjusted R-squared:  0.1742 
F-statistic: 11.34 on 1 and 48 DF,  p-value: 0.001505
```

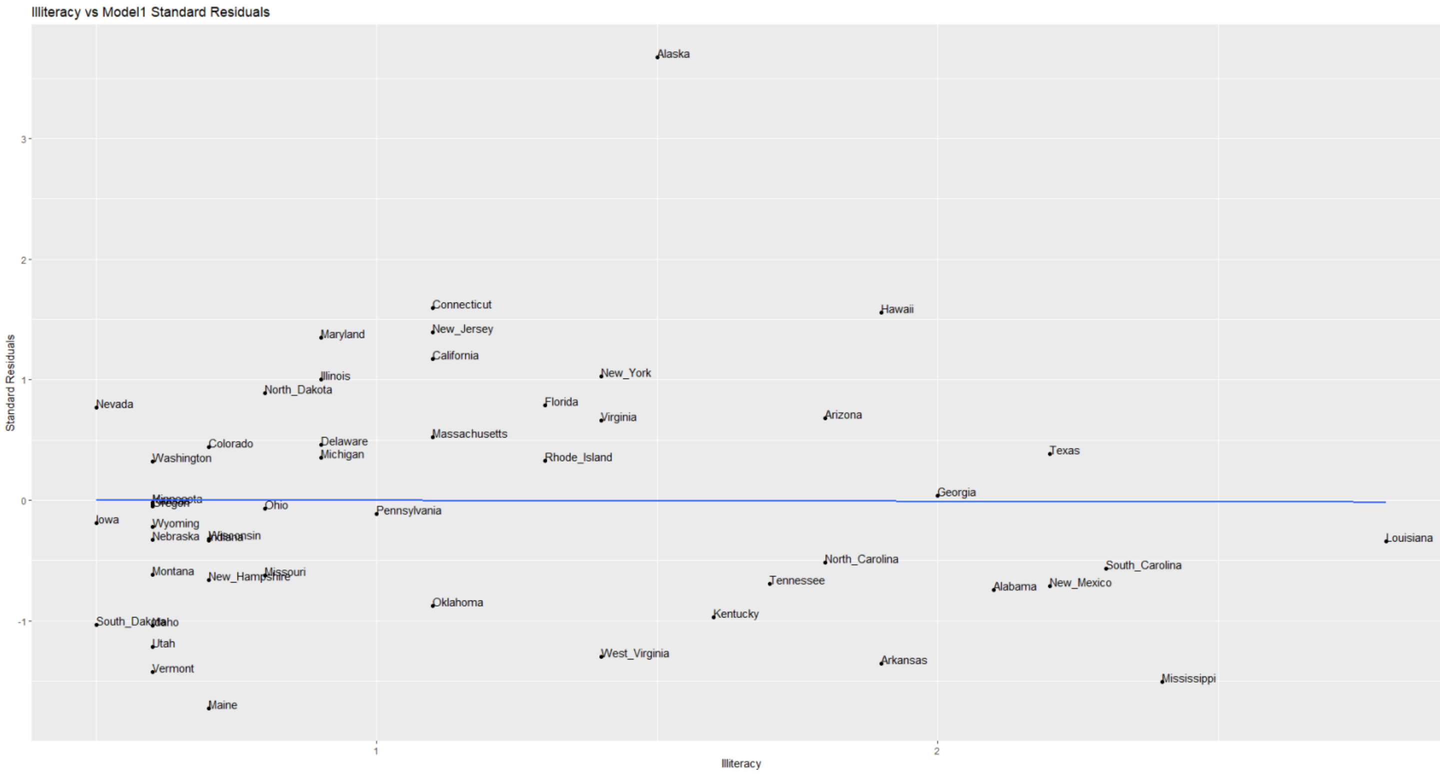
Null Hypothesis H_0 : Income is not linearly dependent on Illiteracy
Alternate Hypothesis H_A: Income is linearly dependent on Illiteracy

We can see that the p-value 0.00151 < 0.05 and thus we can reject the null hypothesis at significance level 5%. Thus, we can conclude that Income is linearly dependent on Illiteracy and with on 1% increase in Illiteracy the per capita income of a state decreases by $440.

The model explains only 19% of variability which is not too low , the remaining variability can be explained by variables not included here.

b. Obtain the standardized residuals for this model and plot them against Illiteracy, with name of the State as text for the data points. Based on a *visual* inspection of this plot, ***comment on*** Linearity, Heteroscedasticity, Potential Outliers
Answer :
```
#Standardized Residuals
library(moments)
mod1_rstand <-rstandard(mod1)

#b
mod1_rstand_ill<-
data.frame(as.numeric(mod1_rstand),as.numeric(Illiteracy),State)
ggplot(mod1_rstand_ill, aes(x=Illiteracy, y=mod1_rstand,label=State)) +
  geom_point()+
  geom_text(aes(label=State),hjust=0, vjust=0) +
  geom_smooth(method = "lm", se = FALSE) +
  ggtitle("Illiteracy vs Model1 Standard Residuals") +
  xlab("Illiteracy") +
  ylab("Standard Residuals")
```

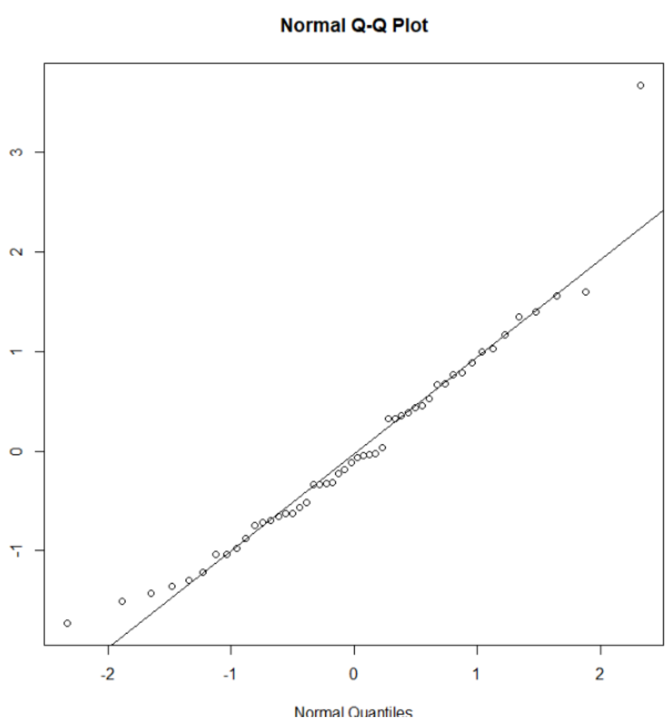


- We can see that the standardized residuals do not have any pattern , the plot is a null plot and except for Alaska they are between -3 and 3, Thus we can conclude that there is linearity
- We can also see that the standardized residuals do not change with Illiteracy in any systematic or functional manner. Thus, there is no Heteroscedasticity.
- We can see also see Only Alaska has a high residual value and thus is a potential outlier

c. **<u>Comment on the normality</u>** of the standardized residuals, using a Q-Q plot

Answer :

```
#c. Comment on the normality of the standardized residuals, using a Q-Q plot
par(mfrow=c(1,1))
qqnorm(mod1_rstand, ylab="Standardized Residuals of Illiteracy", xlab="Normal Quantiles")
qqline(mod1_rstand)
```



The Q-Q plot is somewhat bow-shaped (indicating some skewness) and also S-shaped (crossing the line in both directions) indicating Kurtosis different from a normal distribution.

d. Calculate the skewness and Kurtosis of the standardized residuals ***and interpret them***

Answer :

```
> qqline(mod1_rstand)
> print(skewness(mod1_rstand))
[1] 0.9175981
> print(kurtosis(mod1_rstand))
[1] 4.852361
```

The skewness is positive which means the standardized residuals a skewed to the right and 0.9 indicates the skewness is acceptable

The kurtosis is positive indicating that the standardized residuals are Leptokurtotic

e. Develop a table showing Illiteracy, Income, State and Cook's D. ***Identify*** State or States that may be an outlier, based on Cook's D. ***Comment*** on the Illiteracy and Income of the outlier state(s).

Answer :

```
cook_dist <- cooks.distance(mod1)
#
df_mod1 <-data.frame(Illiteracy, Income, State, cook_dist)
library(dplyr)
arrange(df_mod1,desc(cook_dist))
```

Illiteracy	Income	State	cook_dist
1.5	6315	Alaska	0.180011
2.4	3098	Mississippi	0.130175
1.9	4963	Hawaii	0.062985
0.7	3694	Maine	0.049528
1.9	3378	Arkansas	0.04737
0.6	3907	Vermont	0.03988
0.6	4022	Utah	0.028987
1.1	5348	Connecticut	0.026303
0.5	4167	South_Dakota	0.02496
0.9	5299	Maryland	0.022382
2.2	3601	New_Mexico	0.021443
0.6	4119	Idaho	0.021147
2.1	3624	Alabama	0.020122
1.1	5237	New_Jersey	0.020095
1.4	3617	West_Virginia	0.019804
2.3	3635	South_Carolina	0.01602
1.6	3712	Kentucky	0.014679
1.1	5114	California	0.01419
0.5	5149	Nevada	0.013709
1.4	4903	New_York	0.012436
0.9	5107	Illinois	0.012323
2.8	3545	Louisiana	0.011395
0.8	5087	North_Dakota	0.01112
1.8	4530	Arizona	0.010092
1.7	3821	Tennessee	0.008877
1.1	3983	Oklahoma	0.007921
0.6	4347	Montana	0.007576
0.7	4281	New_Hampshire	0.007204
1.3	4815	Florida	0.00667
2.2	4188	Texas	0.006272

1.8	3875	North_Carolina	0.005856
0.8	4254	Missouri	0.005548
1.4	4701	Virginia	0.005169
0.7	4884	Colorado	0.003198
1.1	4755	Massachusetts	0.002816
0.9	4809	Delaware	0.002612
0.6	4508	Nebraska	0.002099
0.6	4864	Washington	0.002055
0.7	4458	Indiana	0.00188
0.7	4468	Wisconsin	0.001682
0.9	4751	Michigan	0.001556
1.3	4558	Rhode_Island	0.001128
0.6	4566	Wyoming	0.000959
0.5	4628	Iowa	0.000833
1	4449	Pennsylvania	0.000138
0.8	4561	Ohio	6.68E-05
0.6	4660	Oregon	4.76E-05
2	4091	Georgia	4.57E-05
0.6	4669	Kansas	2.11E-05
0.6	4675	Minnesota	9.36E-06

We can see none of the states have cook D value above 0.5 , but we can consider Alaska and Mississippi with the highest Cook D values as outliers .
Alaska :Its Income is the highest among all states , justifying considering it as outlier and Illiteracy is moderately high compared to other states.
Mississippi : Its income is a bit low compared to other states but illiteracy is the highest justifying considering it as an outlier

3) **Model 2:**
   a. Create a new data frame that contains the name of the state, Illiteracy and Income but excludes the outlier state or states identified from Model 2. Use the subset() function for this. (See "Selecting Observations" in https://www.statmethods.net/management/subset.html
      Answer :
```
#Model 2 Income based on Illiteracy without outlier
df2 <- subset(df , !(State %in% c('Alaska','Mississippi')))
```

b. Repeat steps (a) through (e) from Model 1.

```
mod2 <- lm(df2$Income ~ df2$Illiteracy)
summary(mod2)
```

```
Call:
lm(formula = df2$Income ~ df2$Illiteracy)

Residuals:
    Min      1Q  Median      3Q     Max 
-916.12 -338.06  -20.28  332.56  907.57 

Coefficients:
               Estimate Std. Error t value Pr(>|t|)    
(Intercept)      4907.1      148.2  33.106  < 2e-16 ***
df2$Illiteracy   -424.2      115.8  -3.663 0.000642 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 470.9 on 46 degrees of freedom
Multiple R-squared:  0.2259,	Adjusted R-squared:  0.209 
F-statistic: 13.42 on 1 and 46 DF,  p-value: 0.0006415
```

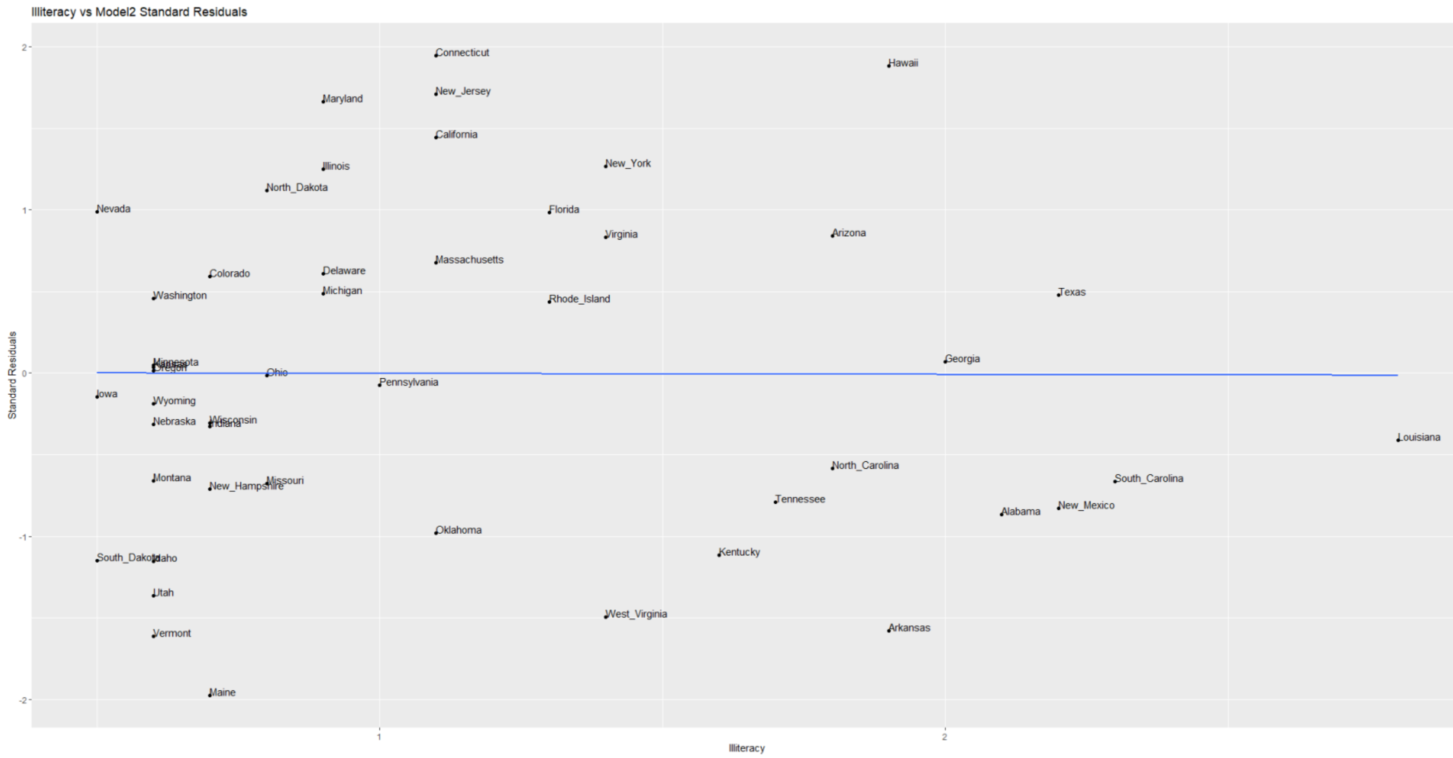
Null Hypothesis H_0 : Income is not linearly dependent on Illiteracy
Alternate Hypothesis H_A: Income is linearly dependent on Illiteracy

We can see that the p-value 0.0006415 < 0.05 and thus we can reject the null hypothesis at significance level 5%. Thus, we can conclude that Income is linearly dependent on Illiteracy and with on 1% increase in Illiteracy the per capita income of a state decreases by $424.2

The model explains only 22% of variability which is better than the previous model.

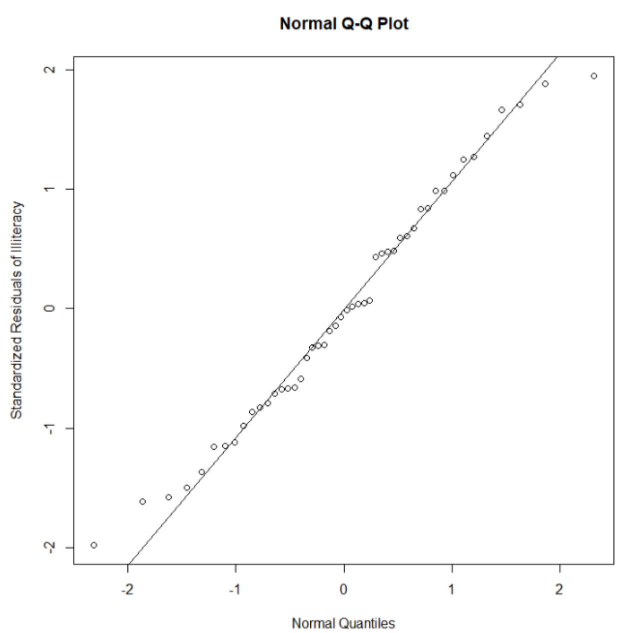
```
#Standardized Residuals
mod2_rstand <-rstandard(mod2)


mod2_rstand_ill<-
data.frame(as.numeric(mod2_rstand),as.numeric(df2$Illiteracy),df2$State)
ggplot(mod2_rstand_ill, aes(x=df2$Illiteracy, y=mod2_rstand,label=df2$State))
+
  geom_point()+
  geom_text(aes(label=df2$State),hjust=0, vjust=0) +
  geom_smooth(method = "lm", se = FALSE) +
  ggtitle("Illiteracy vs Model2 Standard Residuals") +
  xlab("Illiteracy") +
  ylab("Standard Residuals")
```



- We can see that the standardized residuals do not have any pattern , the plot is a null plot and they are all between -3 and 3, Thus we can conclude that there is linearity
- We can also see that the standardized residuals do not change with Illiteracy in any systematic or functional manner. Thus, there is no Heteroscedasticity.

```
#Comment on the normality of the standardized residuals, using a Q-Q plot
par(mfrow=c(1,1))
qqnorm(mod2_rstand, ylab="Standardized Residuals of Illiteracy",
xlab="Normal Quantiles")
qqline(mod2_rstand)
```



The Q-Q plot is not much bow-shaped (**indicating skewness has reduced**) but S-shaped (crossing the line in both directions) indicating Kurtosis different from a normal distribution..

```r
#Calculate the skewness and Kurtosis of the standardized residuals and interpret them
print(skewness(mod2_rstand))
print(kurtosis(mod2_rstand))
```

```
> #d.   Calculate the skewness a
> print(skewness(mod2_rstand))
[1] 0.1321421
> print(kurtosis(mod2_rstand))
[1] 2.173803
```

The skewness is positive but quite less which means that the standardized residuals are slightly right skewed

The Kurtosis is less than 3 indicating that the standard residuals are Platykurtotic

```r
# Cook distance
cook_dist <- cooks.distance(mod2)
#
df_mod2 <-data.frame(df2$Illiteracy, df2$Income, df2$State, cook_dist)
write.csv(arrange(df_mod2,desc(cook_dist)),"cook_d2.csv",row.names = FALSE)
```

Illiteracy	Income	State	cook_dist
1.9	4963	Hawaii	0.105297
1.9	3378	Arkansas	0.074092
0.7	3694	Maine	0.065525
0.6	3907	Vermont	0.051924
1.1	5348	Connecticut	0.040536
0.6	4022	Utah	0.037141
0.9	5299	Maryland	0.03438
2.2	3601	New_Mexico	0.03366
0.5	4167	South_Dakota	0.031332
2.1	3624	Alabama	0.031291
1.1	5237	New_Jersey	0.031227
1.4	3617	West_Virginia	0.028745
0.6	4119	Idaho	0.026593
2.3	3635	South_Carolina	0.025226
0.5	5149	Nevada	0.023171
1.1	5114	California	0.022328
1.6	3712	Kentucky	0.021748
1.4	4903	New_York	0.020636
2.8	3545	Louisiana	0.019523
0.9	5107	Illinois	0.019434
0.8	5087	North_Dakota	0.017836
1.8	4530	Arizona	0.017593
1.7	3821	Tennessee	0.013023
1.3	4815	Florida	0.011173
2.2	4188	Texas	0.011117
1.1	3983	Oklahoma	0.010298
1.4	4701	Virginia	0.008922
0.6	4347	Montana	0.008721
1.8	3875	North_Carolina	0.008487
0.7	4281	New_Hampshire	0.008457
0.8	4254	Missouri	0.006509
0.7	4884	Colorado	0.005856
1.1	4755	Massachusetts	0.00487
0.9	4809	Delaware	0.004623

0.6	4864	Washington	0.004177
0.9	4751	Michigan	0.002926
1.3	4558	Rhode_Island	0.002169
0.6	4508	Nebraska	0.001952
0.7	4458	Indiana	0.001807
0.7	4468	Wisconsin	0.001577
0.6	4566	Wyoming	0.0007
0.5	4628	Iowa	0.000504
2	4091	Georgia	0.000178
1	4449	Pennsylvania	5.94E-05
0.6	4675	Minnesota	4.71E-05
0.6	4669	Kansas	2.53E-05
0.6	4660	Oregon	5.19E-06
0.8	4561	Ohio	2.97E-06

c. ***Compare*** Model 1 and Model 2 and whether excluding the outliers may be justified in your opinion.
Answer :We saw that the Rsquare improved, all standard residuals fall between -3,3 and skewness also reduced. Thus we should prefer model 3 after removing the outliers Alaska and Mississippi

4) Add Murder to the data frame created for Model 2. Create a new column in this data frame that contains a Dummy variable taking a value of 1 if the Murder rate is greater than the mean of Murder rate and 0, otherwise. ***Print out the data frame***.
Answer :

```
df3<-subset(df2, select=c(State,Illiteracy,Income,Murder))
murder_mean<-mean(df3$Murder)
df3$d_murder= ifelse(df2$Murder>murder_mean,1,0)
write.csv(df3,"d_murder.csv",row.names = FALSE)
```

State	Illiteracy	Income	Murder	d_murder
Alabama	2.1	3624	15.1	1
Arizona	1.8	4530	7.8	1
Arkansas	1.9	3378	10.1	1
California	1.1	5114	10.3	1
Colorado	0.7	4884	6.8	0
Connecticut	1.1	5348	3.1	0
Delaware	0.9	4809	6.2	0
Florida	1.3	4815	10.7	1
Georgia	2	4091	13.9	1
Hawaii	1.9	4963	6.2	0
Idaho	0.6	4119	5.3	0
Illinois	0.9	5107	10.3	1
Indiana	0.7	4458	7.1	0
Iowa	0.5	4628	2.3	0
Kansas	0.6	4669	4.5	0
Kentucky	1.6	3712	10.6	1
Louisiana	2.8	3545	13.2	1
Maine	0.7	3694	2.7	0
Maryland	0.9	5299	8.5	1

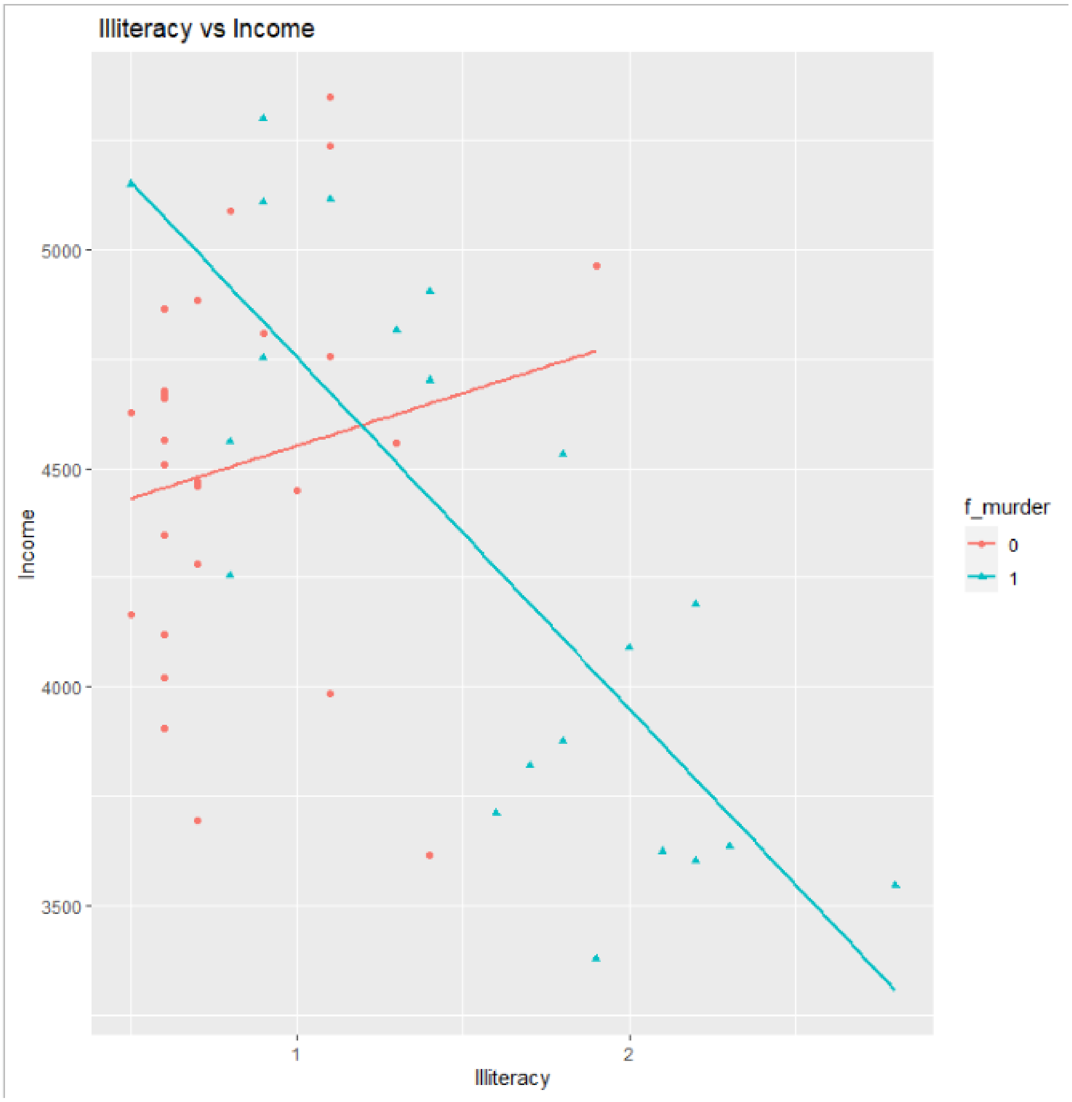
Massachusetts	1.1	4755	3.3	0
Michigan	0.9	4751	11.1	1
Minnesota	0.6	4675	2.3	0
Missouri	0.8	4254	9.3	1
Montana	0.6	4347	5	0
Nebraska	0.6	4508	2.9	0
Nevada	0.5	5149	11.5	1
New_Hampshire	0.7	4281	3.3	0
New_Jersey	1.1	5237	5.2	0
New_Mexico	2.2	3601	9.7	1
New_York	1.4	4903	10.9	1
North_Carolina	1.8	3875	11.1	1
North_Dakota	0.8	5087	1.4	0
Ohio	0.8	4561	7.4	1
Oklahoma	1.1	3983	6.4	0
Oregon	0.6	4660	4.2	0
Pennsylvania	1	4449	6.1	0
Rhode_Island	1.3	4558	2.4	0
South_Carolina	2.3	3635	11.6	1
South_Dakota	0.5	4167	1.7	0
Tennessee	1.7	3821	11	1
Texas	2.2	4188	12.2	1
Utah	0.6	4022	4.5	0
Vermont	0.6	3907	5.5	0
Virginia	1.4	4701	9.5	1
Washington	0.6	4864	4.3	0
West_Virginia	1.4	3617	6.7	0
Wisconsin	0.7	4468	3	0
Wyoming	0.6	4566	6.9	0

5) Create a factor variable f_murder based on the dummy variable for murder from the previous question. Using ggplot(), plot Income vs Illiteracy with shape and color based on f_murder. Then, using geom_smooth() show the lines of interaction between Illiteracy and f_murder in determining Income. ***Interpret*** the interaction in your words i.e., what is the plot saying in terms of the effect of murder rate on the relationship between Illiteracy and Income.

Answer :

```
f_murder <- factor(df3$d_murder)

ggplot(df3, aes(x=Illiteracy, y=Income, shape=f_murder, color=f_murder)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE) +
  ggtitle(" Illiteracy vs Income") +
  xlab("Illiteracy") +
  ylab("Income")
```



- We can see that when f_muder =1 or murder > mean of murder income decreases with increase in Illiteracy whereas for f_murder = 0 income increases with increase in Illiteracy. This suggests an interaction between Muder and Illliteracy in determining Per capita Income .

6) **Model 3:**

   a. Develop a regression model that predicts Income using Illiteracy, Murder and their Interaction. ***Write your conclusions*** about the model based on the summary output.

   Answer :

```
mod3 <- lm(Income ~ Illiteracy+Murder+Illiteracy*Murder)
summary(mod3)
```

```
Call:
lm(formula = Income ~ Illiteracy + Murder + Illiteracy * Murder)

Residuals:
    Min      1Q  Median      3Q     Max 
-955.20 -325.99   10.66  299.96 1892.12 

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)    
(Intercept)        3822.61     405.33   9.431 2.54e-12 ***
Illiteracy          617.34     434.85   1.420  0.16245    
Murder              146.82      50.33   2.917  0.00544 ** 
Illiteracy:Murder  -117.10      40.13  -2.918  0.00544 ** 
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 520.1 on 46 degrees of freedom
Multiple R-squared:  0.3273,	Adjusted R-squared:  0.2834 
F-statistic: 7.461 on 3 and 46 DF,  p-value: 0.000359
```

- We can see that We can see Only Murder and interaction term of Illiteracy and Murder have significant p-values and significance level 0.05
- Rsquare indicates that the model explains 32.73% the variability in Income.
- We can see that the overall F-Tests p-value 0.000359 is less than 0.05 indicating that the model does explain a significant proportion of the variability in Income.

So, the model is

Income = 3822.61 + 146.82 * Murder -117.10 * Illiteracy*Murder

b. Using ols_vif_tol(), produce multicollinearity diagnostics ***and interpret them***.

Answer :

```
library(olsrr)
ols_vif_tol(mod3)
```

```
> ols_vif_tol(mod3)
          Variables  Tolerance       VIF
1        Illiteracy 0.07859297 12.723785
2            Murder 0.15997571  6.250949
3 Illiteracy:Murder 0.03913995 25.549345
>
```

The VIF values are very high indicating high multicollinearity between the predictors. Thus we look into centered versions of predictors to remedy this

7) **Model 4:**

a. Develop a regression model that predicts Income using the centered version of Illiteracy, the centered version of Murder and the Interaction between these centered versions. ***Write your conclusions*** about the model based on the summary output.

Answer :

```
c_Illiteracy<-Illiteracy-mean(Illiteracy)
c_Murder<-Murder-mean(Murder)
cIll_cMur <- c_Illiteracy*c_Murder

mod4 <- lm(Income ~ c_Illiteracy+c_Murder+cIll_cMur, data = df)
summary(mod4)
```

```
Call:
lm(formula = Income ~ c_Illiteracy + c_Murder + cIll_cMur, data = df)

Residuals:
    Min      1Q  Median      3Q     Max 
-955.20 -325.99   10.66  299.96 1892.12 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept)  4617.315     96.338  47.928  < 2e-16 ***
c_Illiteracy -246.592    200.260  -1.231  0.22445    
c_Murder        9.815     28.802   0.341  0.73481    
cIll_cMur    -117.096     40.131  -2.918  0.00544 ** 
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 520.1 on 46 degrees of freedom
Multiple R-squared:  0.3273,	Adjusted R-squared:  0.2834 
F-statistic: 7.461 on 3 and 46 DF,  p-value: 0.000359
```

We can see that p-values for centered versions of Illiteracy and Murder are higher than 0.05 and thus there is no significant dependency of Income on them. But the Interaction term has significant p-value

Rsquare Indicates that the model explains 32.73 % of the variability

b. Using ols_vif_tol(), produce multicollinearity diagnostics ***<u>and interpret them relative to the multicollinearity diagnostics of Model 3</u>***.
Answer :

```
> ols_vif_tol(mod4)
     Variables Tolerance      VIF
1 c_Illiteracy 0.3705757 2.698504
2     c_Murder 0.4884383 2.047342
3    cIll_cMur 0.6779227 1.475095
> |
```

We can clearly see the VIF values have reduced compared to Model 3 indicating this lesser multicollinearity . This model is better for interpreting the slope coefficients.

8) **<u>Model 5:</u>**

a. Develop a regression model that predicts Income using only the interaction between the centered versions of Illiteracy and Murder. ***<u>Write your conclusions</u>*** about the model based on the summary output.
Answer :

```
mod5 <- lm(Income ~ cIll_cMur, data = df)
summary(mod5)

Call:
lm(formula = Income ~ cIll_cMur, data = df)

Residuals:
     Min       1Q   Median       3Q      Max 
-1073.31  -358.60    68.72   262.85  1841.03 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  4667.04      89.64  52.066  < 2e-16 ***
cIll_cMur    -149.18      33.04  -4.515 4.12e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 520.1 on 48 degrees of freedom
Multiple R-squared:  0.2981,	Adjusted R-squared:  0.2834 
F-statistic: 20.38 on 1 and 48 DF,  p-value: 4.116e-05

> |
```

- We can see that the P-values for both Overall f-test and for the slope coefficient of interaction terms are significant and α=0.05.

- The Rsquare indicates that it explains 29.81% variability. This is quite good and additionally even though it is a bit less compared to Model 4 the adjusted Rsquare is same Indicating there is no issue

b. Compare Models 4 and 5 and ***discuss which model you would choose and why***.

Answer :

Since in Model 4 only the interaction term is significant the Model equation is

Income = 4617.315 – 117.096*`cIll_cMur`

Fo Model 5

Income = 4667.04 – 149.18*`cIll_cMur`

We can see that there is considerable change slope coefficient of the interaction term from -117 to -149 form Model 4 to Model 5. Thus we need to go with Model 5 as the coefficient was distorted a lot in Model 5 by the non significant variables

9) **Final Model**

a. Write out the equation for the final model that you have chosen (between Model 4 and Model 5). The equation should be simplified so that Illiteracy, Murder, and the Interaction terms have their own numerical coefficients.

Answer : We are going with model 5 and thus the equation is

Income = 4667.04 – 149.18*(Illiteracy-1.17)*(Murder-7.378)

Or

Income = 3379.279 + 1100.65*Illiteracy + 174.54* Murder -149.18* Illiteracy* Murder

b. Predict Income for some value of Illiteracy and Murder of your choice.

Answer :

For Kansas State Illiteracy rate and

From Model 5 we get

Income = 3379.279 + 1100.65*0.6 + 174.54* 4.5 -149.18* 0.6* 4.5 = 4422.313

The actual value is 4669

10) **Final Summary**: In your own words explain what the final model is saying about the effect of Illiteracy by itself on Income, and when Illiteracy is combined with Murder rate.

Answer :

We can rewrite the equation as

Income = 3379.279 +(1100.65-149.18*Murder)*Illiteracy + 174.54*Murder

We can clearly see from this equation that Effect of changes in Illiteracy is dependent on Murder . This is what interaction means . Thus, we cannot interpret the effect of Illiteracy alone in a straightforward way.

But we can always interpret it based on the value of murder for the State in consideration and assuming it be constant. For example, In Kansas we have Murder 4.5 substituting this in the equation we get

Income = 3379.279 + 429.34* Illiteracy + 785.43.

So we can say if Murder is held constant at 4.5 then with 0.1 increase in Illiteracy rate Income would increase by $42.9