**ISQA 8720 Assignment 1 Due: By Tuesday, Sep. 22 2020, 5:30 PM (see Canvas for potential changes of the due date)**

**Name: \_\_Martin Nwadiugwu\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

**NOTES:**

1. Use R for the calculations and implementation.
2. You need to submit this answer sheet and your R code.
3. Submit all documents in a zip file and upload it to Canvas. Name your Zip Folder with your name, A1, and the course # (Example: LastName-A1-ISQA 8720).

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

1. **Regression Analysis (60 points)**

This question should be answered using the Carseats data set that you have available in Canvas. You can find a detailed description of the data variables by looking up the help function about the data set in the package ISLR: (library(ISLR) ?Carseats)

1. Fit a multiple regression model to predict Sales, using Price, Urban, and US as predictor variables. Show the R output for the lm model here.

Call:

lm(formula = Sales ~ Price + Urban + US, data = Carseats)

Residuals:

Min 1Q Median 3Q Max

-6.9206 -1.6220 -0.0564 1.5786 7.0581

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 13.043469 0.651012 20.036 < 2e-16 \*\*\*

Price -0.054459 0.005242 -10.389 < 2e-16 \*\*\*

UrbanYes -0.021916 0.271650 -0.081 0.936

USYes 1.200573 0.259042 4.635 4.86e-06 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.472 on 396 degrees of freedom

Multiple R-squared: 0.2393, Adjusted R-squared: 0.2335

F-statistic: 41.52 on 3 and 396 DF, p-value: < 2.2e-16

1. Write out the model in equation form (note that you have qualitative predictors in the model).

**The model equation is:**

E(Y) = B0 + B1x1 + B2X2 + B3X3

X1 = Price, X2 = Urban, X3 = US

Sales = 13.043469 + (-0.054459) \* Price + (-0.021916) \* UrbanYes + (1.200573) \* USYes + e

UrbanYes = 1; UrbanNo = 0; USYes = 1; USNo = 0

**Average Sales for Urban and US residents if Price is zero**

If x2= Yes and X3 = Yes

E(Y|x1) = B0 + B2 + B3 + B1x1

Sales = 13.043469 + (-0.021916) + 1.200573

**Average Sales for Non-Urban and Non-US residents if Price is zero**

If x2= 0 and X3 = 0

E(Y|x1) = B0 + B1x1

Sales = 13.043469

**Average Sales for Urban and Non-US residents if Price is zero**

If x2= 1 and X3 = 0

E(Y|x1) = B0 + B2 + B1x1

Sales = 13.043469 + (-0.021916)

**Average Sales for Non-Urban and US residents if Price is zero**

If x2= 0 and X3 = 1

E(Y|x1) = B0 + B3 + B1x1

Sales= 13.043469 + 1.200573

1. Provide an interpretation of each coefficient in the model. What does it tell you about the relationship between the target and each of the predictor variables?

**Interpretation of the coefficients**:

1. Price: There will be an average -0.054459 or 54.459-unit decrease in sales for a unit increase in Price, all other variables kept constant.
2. Urban: There will be an average -0.021916 or 21.916-unit decrease in sales for residents in Urban dwellings, all other variables kept constant
3. US: There will be an average 1.200573 or 1200.573 units increase in sales for US residents, all other variables kept constant
4. For which of the predictors can you reject the null hypothesis H0: βj = 0?

**H0: βj = 0**: Price, Urban and US customers has no difference in Sales

**H1: βj != 0**: There is a difference in sales for various Price, Urban and US customers.

With p-value < 0.001 for both Price and US predictor variables, we reject the null hypothesis (**for Price and US variables**) and conclude that we are 95% confident that there is a difference in sales for various Price, and US customers.

1. Based on your response to the previous question, fit a smaller model that only uses the predictors for which there is evidence of association with the outcome (significance).

Call:

lm(formula = Sales ~ Price + US, data = Carseats)

Residuals:

Min 1Q Median 3Q Max

-6.9269 -1.6286 -0.0574 1.5766 7.0515

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 13.03079 0.63098 20.652 < 2e-16 \*\*\*

Price -0.05448 0.00523 -10.416 < 2e-16 \*\*\*

USYes 1.19964 0.25846 4.641 4.71e-06 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.469 on 397 degrees of freedom

Multiple R-squared: 0.2393, Adjusted R-squared: 0.2354

F-statistic: 62.43 on 2 and 397 DF, p-value: < 2.2e-16

**Model:**

E(Y) = B0 + B1x1 + B2X2

X1 = Price, X2 = US

Sales = 13.03079 + -0.05448 \* Price + 1.19964 \* USYes

1. How well do the models in (a) and (e) fit the data? Use both R2 and MSE/RMSE for this.

Although the R2 confirms both (a) and (e) are only responsible for 23.93% of the variance, the adjusted R2 for (e) is marginally higher (0.2354) than that of (a) (0.2335). The changes in the standard error was also marginal, with (e) having a smaller residual standard error of 2.469 compared to (a) with 2.472.

ANOVA shows a p-value>0.05 confirming we do not get a statistical significance by adding the Urban predictor to the model

Analysis of Variance Table

Model 1: Sales ~ Price + US

Model 2: Sales ~ Price + Urban + US

Res.Df RSS Df Sum of Sq F Pr(>F)

1 397 2420.9

2 396 2420.8 1 0.03979 0.0065 0.9357

More also, this implies that other variables within the dataset are likely responsible for the remaining 75% variance as can be seen below when all predictors are used.

Call:

lm(formula = Sales ~ ., data = Carseats)

Residuals:

Min 1Q Median 3Q Max

-2.8409 -0.6817 0.0127 0.6468 3.4684

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 5.7285968 0.6110801 9.375 < 2e-16 \*\*\*

X -0.0003284 0.0004538 -0.724 0.470

CompPrice 0.0930031 0.0041583 22.366 < 2e-16 \*\*\*

Income 0.0156505 0.0018582 8.422 7.3e-16 \*\*\*

Advertising 0.1238581 0.0111803 11.078 < 2e-16 \*\*\*

Population 0.0002157 0.0003708 0.582 0.561

Price -0.0953564 0.0026727 -35.678 < 2e-16 \*\*\*

ShelveLocGood 4.8520250 0.1532252 31.666 < 2e-16 \*\*\*

ShelveLocMedium 1.9579029 0.1261938 15.515 < 2e-16 \*\*\*

Age -0.0461835 0.0031894 -14.480 < 2e-16 \*\*\*

Education -0.0224532 0.0198208 -1.133 0.258

UrbanYes 0.1278481 0.1132532 1.129 0.260

USYes -0.1853717 0.1499447 -1.236 0.217

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.02 on 387 degrees of freedom

Multiple R-squared: 0.8736, Adjusted R-squared: 0.8697

F-statistic: 222.9 on 12 and 387 DF, p-value: < 2.2e-16

Analysis of Variance Table

Model 1: Sales ~ Price + US

Model 2: Sales ~ X + CompPrice + Income + Advertising + Population + Price +

ShelveLoc + Age + Education + Urban + US

Res.Df RSS Df Sum of Sq F Pr(>F)

1 397 2420.87

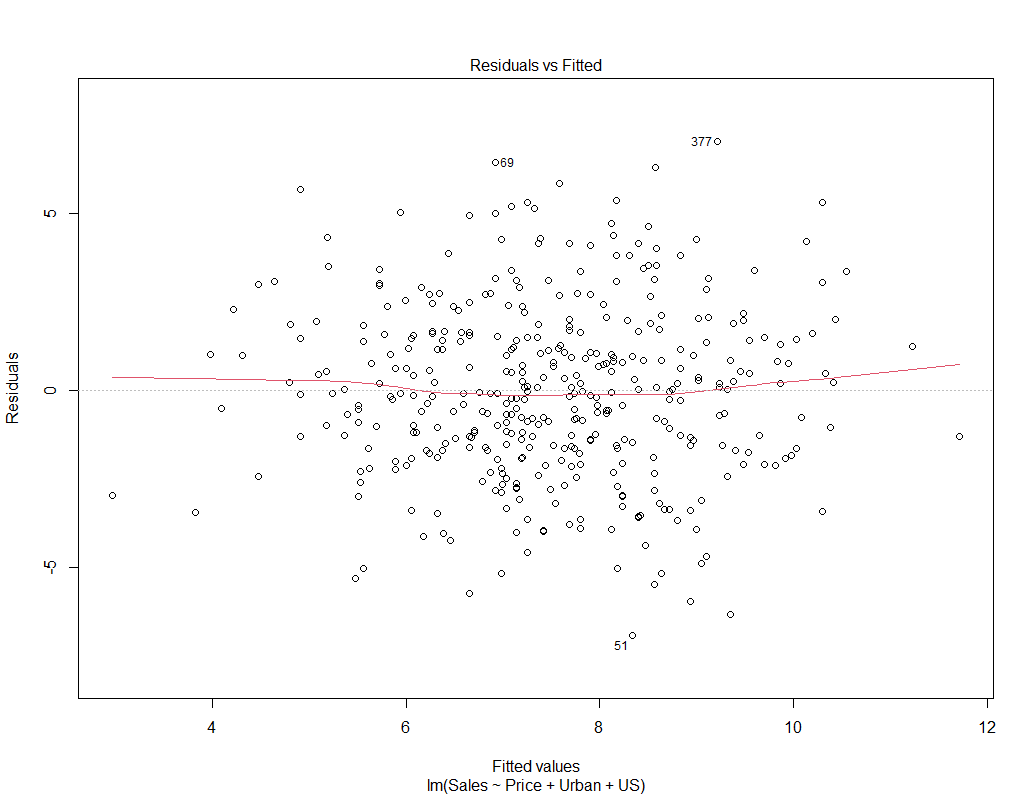
2 387 402.29 10 2018.6 194.19 < 2.2e-16 \*\*\*

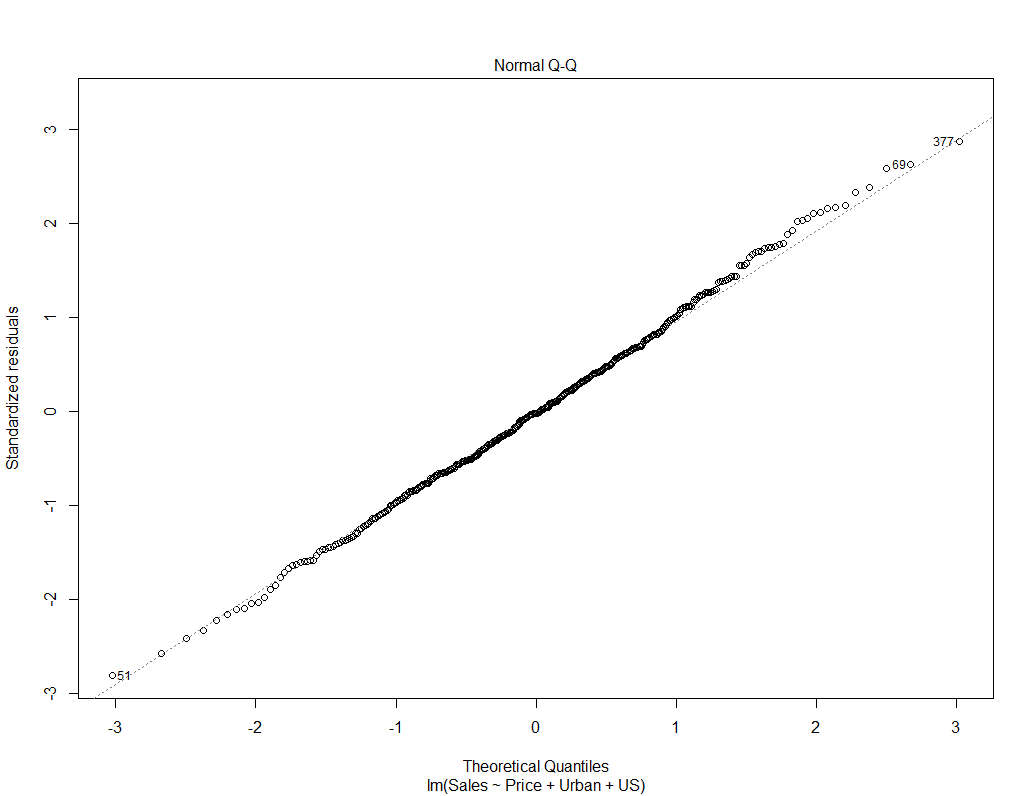
---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

1. Test the assumptions of the linear regression model, i.e., use the plot() function (or autoplot()) to test the constant variance and normality assumptions. Do they seem to be reasonable?

The models appear to be marginally linear but largely meets normality assumptions as seen in Q-Q plot although there are a few outliers in the residual plots.





1. Using the model from (e), obtain 95% confidence intervals for the coefficient(s).

confint(Sales\_predict2)

2.5 % 97.5 %

(Intercept) 11.79032020 14.27126531

Price -0.06475984 -0.04419543

USYes 0.69151957 1.70776632

1. Using the model from (e), predict the Sales for following observation:

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| CompPrice | Income | Advertising | Population | Price | ShelveLoc | Age | Education | Urban | US |
| 138 | 73 | 11 | 276 | 120 | Bad | 42 | 17 | Yes | Yes |

**Using only Price and US variables as predictors selected in (e)?**

E(Y) = B0 + B1x1 + B2X2

X1 = Price, X2 = US

Sales = 13.03079 + (-0.05448 \* 120) + (1.19964 \* 1)

Sales = 7.69283

1. **Regression Analysis – Collinearity (40 points)**

This problem focuses on the *collinearity* problem and its potential effects on linear regression models.

1. Perform the following commands in R (use your NU-ID as random number seed):

set .seed (‘replace with your NU-ID’)

x1 <- runif (100)

x2 <- 0.5\* x1 + rnorm (100) /10

y <- 2 + 2\* x1 +0.3\* x2+rnorm (100)

The last line corresponds to creating a linear model in which y is

a function of x1 and x2. **Write out the parametrized form of the linear model.**

Y = 2 + 2X1 + 0.3X2 + e

**What are the regression (beta) coefficients?**

B0 = 2 + rnorm(100)

B1 = 2

B2 = 0.3

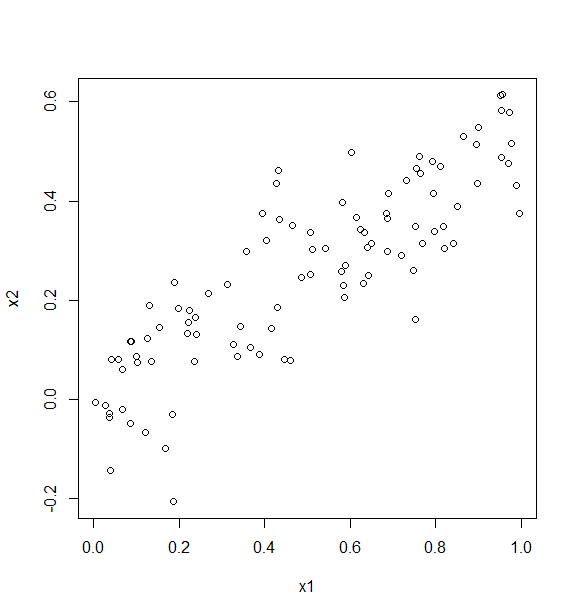
1. What is the correlation coefficient between x1 and x2? Create a scatterplot displaying the relationship between the variables.

**Correlation coefficient**

> cor(x1, x2)

[1] 0.8523826

**Scatter Plot**



1. Using this data, fit a least squares regression to predict y using x1 and x2. Describe the results obtained. What are , , and ? How do these relate to the true *, ,* and ? Can you reject the null hypothesis *H*0 : = 0? How about the null hypothesis *H*0: = 0?

P\_y = lm(y~x1+x2)

> summary(P\_y)

Call:

lm(formula = y ~ x1 + x2)

Residuals:

Min 1Q Median 3Q Max

-2.29756 -0.70943 0.09078 0.68527 2.67031

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 1.8154 0.2133 8.511 2.17e-13 \*\*\*

x1 1.3232 0.7044 1.879 0.0633 .

x2 1.6269 1.1555 1.408 0.1623

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.098 on 97 degrees of freedom

Multiple R-squared: 0.2742, Adjusted R-squared: 0.2592

F-statistic: 18.32 on 2 and 97 DF, p-value: 1.776e-07

**What are , , and ?**

= 1.8154

= 1.3232

=1.6269

**How do these relate to the true *, ,* and ?**

They are not very closely related

**Can you reject the null hypothesis *H*0 : = 0?**

Given p-value > 0.05 we cannot reject the null hypothesis.

**How about the null hypothesis *H*0: = 0?**

Given p-value > 0.05 we cannot reject the null hypothesis.

1. Now fit a least squares regression to predict y using only x1. Comment on your results. Can you reject the null hypothesis *H*0: = 0?

P\_y1 = lm(y~x1)

> summary(P\_y1)

Call:

lm(formula = y ~ x1)

Residuals:

Min 1Q Median 3Q Max

-2.4120 -0.7449 0.1274 0.6813 2.7500

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 1.8140 0.2144 8.462 2.59e-13 \*\*\*

x1 2.1686 0.3702 5.858 6.26e-08 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.104 on 98 degrees of freedom

Multiple R-squared: 0.2594, Adjusted R-squared: 0.2518

F-statistic: 34.32 on 1 and 98 DF, p-value: 6.265e-08

**Can you reject the null hypothesis *H*0: = 0?**

Given p-value < 0.01 we can reject the null hypothesis *H*0: = 0 and X1 can explain (R2) 25.94% of the variance of y.

1. Now fit a least squares regression to predict y using only x2. Comment on your results. Can you reject the null hypothesis *H*0: = 0?

P\_y2 = lm(y~x2)

> summary(P\_y2)

Call:

lm(formula = y ~ x2)

Residuals:

Min 1Q Median 3Q Max

-2.26626 -0.72027 0.08919 0.80829 2.72431

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 1.9967 0.1927 10.364 < 2e-16 \*\*\*

x2 3.4772 0.6120 5.682 1.37e-07 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.112 on 98 degrees of freedom

Multiple R-squared: 0.2478, Adjusted R-squared: 0.2401

F-statistic: 32.28 on 1 and 98 DF, p-value: 1.368e-07

**Can you reject the null hypothesis *H*0: = 0?**

Given p-value < 0.01 we can reject the null hypothesis *H*0: = 0 and X2 can explain (R2) 24.78% of the variance of y.

1. Do the results obtained in (c)–(e) contradict each other? If yes, explain why this happens in this example.

Yes, they do. In (c) both x1 and x2 were not statistically significant (p>0.05) to explain the variance around y, but they were significant (p < 0.001) in both (d) and (e) and responsible for 25.94% and 24.78% variance of y respectively.

1. Now suppose we obtain one additional observation, which was unfortunately incorrectly measured.

x1 <- c(x1 , 0.1)

x2 <- c(x2 , 0.8)

y <- c(y,6)

Re-fit the linear models from (c) to (e) using this new data. What effect does this new observation have on the each of the models?

In each model, is this observation an outlier? A high-leverage point? Both? Explain your answers.

**Model (c)**

x1 <- c(x1 , 0.1)

> x2 <- c(x2 , 0.8)

> y <- c(y,6)

> R\_y = lm(y~x1+x2)

> summary(R\_y)

Call:

lm(formula = y ~ x1 + x2)

Residuals:

Min 1Q Median 3Q Max

-2.32787 -0.78877 0.07435 0.81687 2.61557

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 1.8702 0.2144 8.722 7.15e-14 \*\*\*

x1 0.5301 0.5833 0.909 0.36574

x2 3.0103 0.9253 3.253 0.00156 \*\*

---

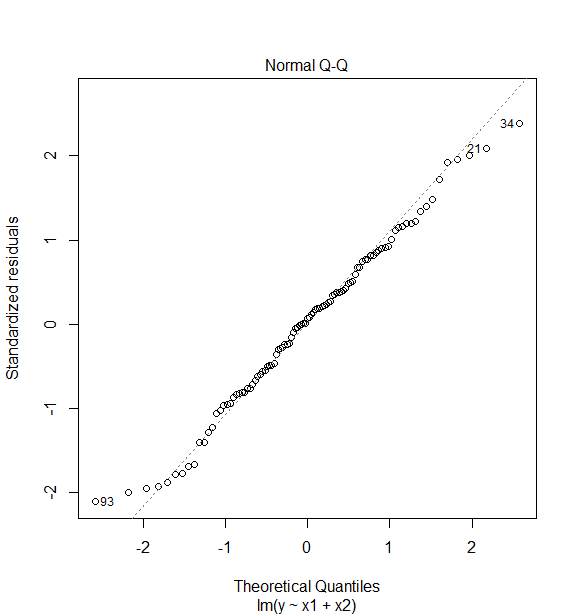
Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

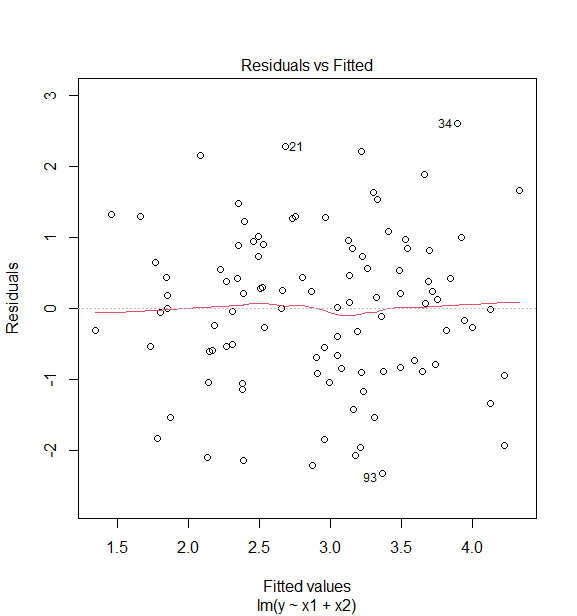
Residual standard error: 1.114 on 98 degrees of freedom

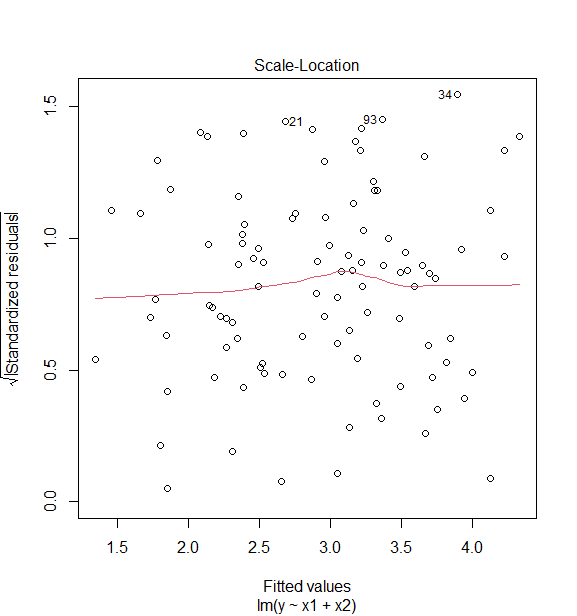
Multiple R-squared: 0.288, Adjusted R-squared: 0.2735

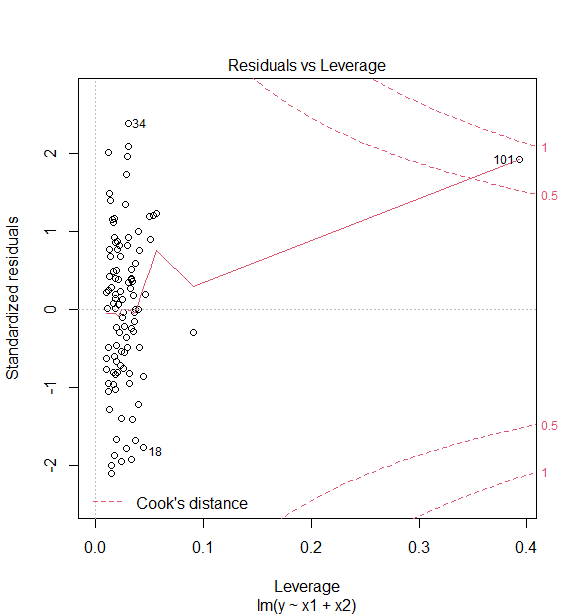
F-statistic: 19.82 on 2 and 98 DF, p-value: 5.9e-08

In this model, the R2 of 28.8 is slightly better than that in (c) 27.42. Also, x2 that was not significant is now significant with a p-value of 0.00156. The model shows some relative outlier as seen in the Normal Q-Q and residual plots.









**Model (d)**

#for (d)

> R\_y1 = lm(y~x1)

> summary(R\_y1)

Call:

lm(formula = y ~ x1)

Residuals:

Min 1Q Median 3Q Max

-2.4357 -0.7788 0.0634 0.6098 3.8623

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 1.9381 0.2235 8.671 8.58e-14 \*\*\*

x1 1.9963 0.3878 5.147 1.34e-06 \*\*\*

---

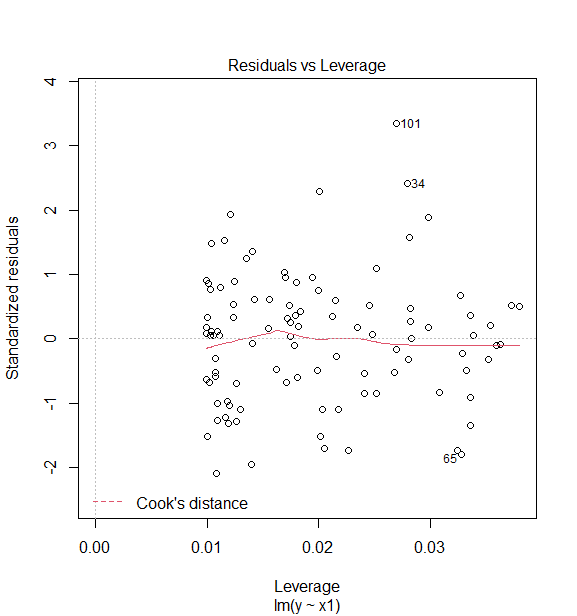
Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

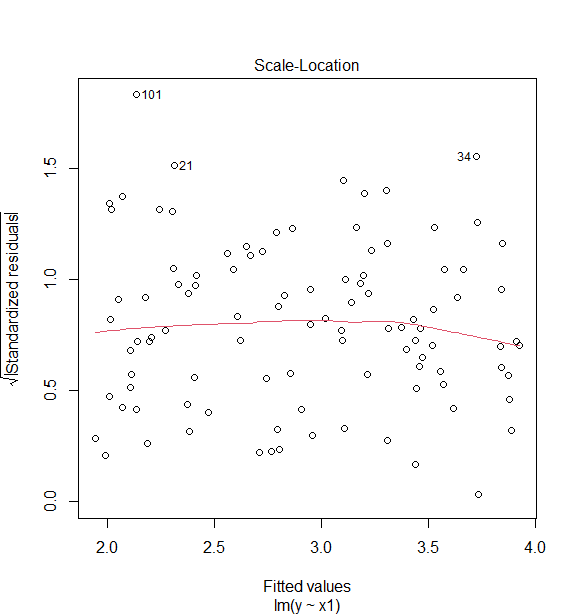
Residual standard error: 1.166 on 99 degrees of freedom

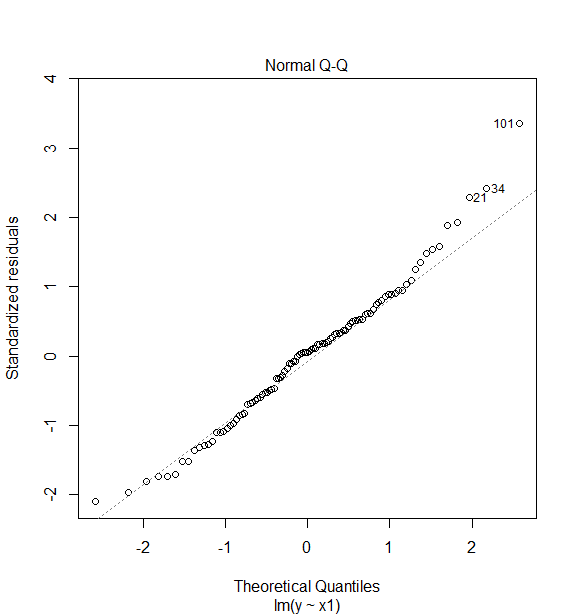
Multiple R-squared: 0.2111, Adjusted R-squared: 0.2032

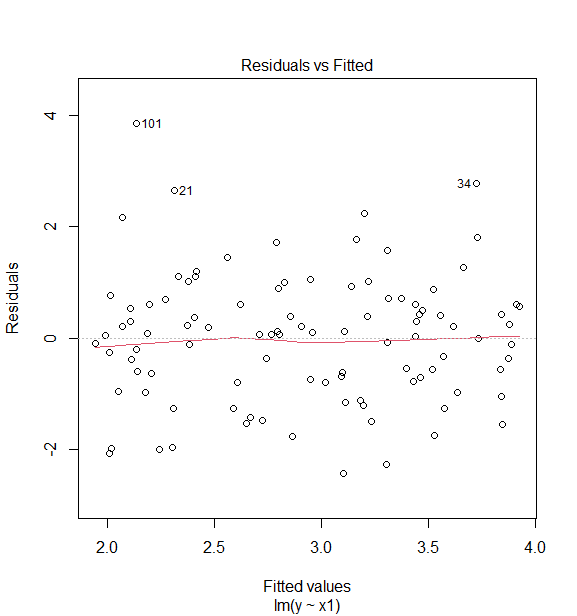
F-statistic: 26.49 on 1 and 99 DF, p-value: 1.342e-06

In this model, the R2 of 21.11 is slightly lower than that in (c) 25.94. Also, x1 is significant in both models with a p-value < 0.001. The model shows some relative outlier as seen in the Normal Q-Q and residual plots.









**Model (e)**

#for (e)

> R\_y2 = lm(y~x2)

> summary(R\_y2)

Call:

lm(formula = y ~ x2)

Residuals:

Min 1Q Median 3Q Max

-2.29894 -0.74483 0.04651 0.81311 2.66629

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 1.9608 0.1897 10.337 < 2e-16 \*\*\*

x2 3.6599 0.5869 6.236 1.11e-08 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.113 on 99 degrees of freedom

Multiple R-squared: 0.282, Adjusted R-squared: 0.2748

F-statistic: 38.89 on 1 and 99 DF, p-value: 1.109e-08

In this model, the R2 of 28.2 is slightly higher than that in (e) 24.78. Also, x2 is significant in both models with a p-value < 0.001. The model shows fewer relative outlier as seen in the Normal Q-Q and residual plots.

