Assginment 5 Report, DATA 400

Intructor: Dr. Spectrum Han

Yohen Thounaojam, 56112204

Q1.

Definitions:

- A **leverage point** is an observation that has an unusual predictor value (very different from the bulk of the observations).
- An **influence point** is an observation whose removal from the data set would cause a large change in the estimated regression model coefficients.

Since the location of points in x space determines their leverage on the regression model, which is measured by the diagonal elements h_{ii} of the hat matrix $\mathbf{H} = \mathbf{X}(\mathbf{X'X})^{-1}\mathbf{X'}$ where $h_{ii} = \mathbf{x_i'}(\mathbf{X'X})^{-1}\mathbf{x_i}$ it is traditionally assumed that any observation x_i with

$$h_{ii} > 2\overline{h} = 2(k+1)/n$$

is remote enough to be considered a leverage point.

Influence Points can be calculated for any given model using Cook's Distance. The formula is:

$$D_k = \frac{1}{(q+1)\hat{\sigma}^2} \sum_{i=1}^{n} (\hat{y}_{i(k)} - y_i)^2$$

Q2.

Importing Data

```
##    temp  usage
## 1    21 185.79
## 2    24 214.47
## 3    32 288.03
## 4    47 424.84
## 5    50 454.68
## 6    59 539.03
## 7    68 621.55
## 8    74 675.06
## 9    62 562.03
## 10   50 452.93
## 11   41 369.95
## 12   30 273.98
```

Data Description

From the table above, we see that the p2.12 data frame has 12 observations on the number of pounds of steam used per month at a plant and the average monthly ambient temperature.

This data frame contains the following columns:

- **temp**: ambient temperature (in degrees F)
- **usage**: usage (in thousands of pounds)

Analysis and Results

Let us perform a Linear Regression to analyze the relationship between Temperature and Usage. The following are the results:

```
##
## Call:
## lm(formula = usage ~ temp, data = temp_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.5629 -1.2581 -0.2550  0.8681  4.0581
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -6.33209    1.67005  -3.792  0.00353 **
## temp         9.20847    0.03382 272.255  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.946 on 10 degrees of freedom
## Multiple R-squared:  0.9999, Adjusted R-squared:  0.9999
## F-statistic: 7.412e+04 on 1 and 10 DF,  p-value: < 2.2e-16
```

Table 1. Results of Linear Regression

Now, let us create a plot of the regression line.

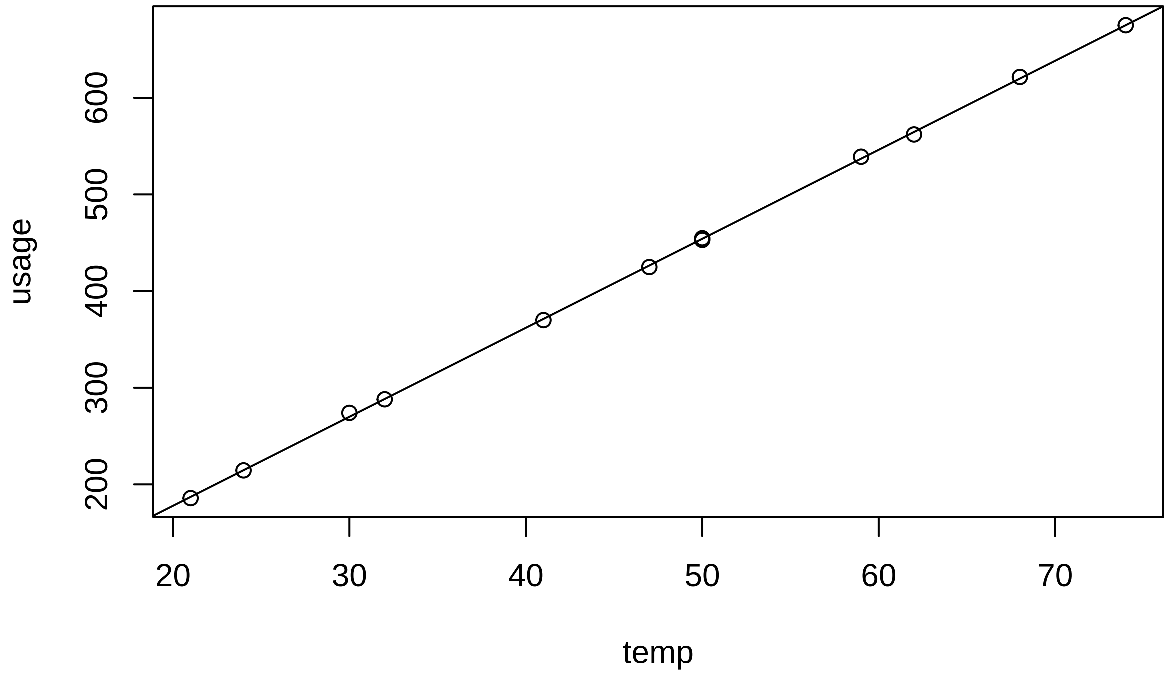

Figure 1: Linear Regression of Usage and Temp

In *Figure 1*, we notice a clear positive linear correlation between *temp* and *usage*. Now looking at the output from our Linear Regression model:

- Intercept: -6.33209
- Slope: 9.20847

Using the values above, we can form the following equation to represent the relationship between *temp* and *usage* from the following equation of a line:

$$y = m * x + b$$

Now, m is the Slope and b is the Intercept, so,

$$y = Slope * x + Intercept$$

$$y = 9.20847 * x - 6.33209$$

$$usage = 9.20847 * temp - 6.33209$$

Hypothesis Testing

Let, α=0.05.

Our null and alternate hypothesis are:

$$H_0 : m = 0;\ there\ is\ no\ relationship\ between\ x\ and\ y,\ therefore\ slope\ m = 0.$$

$$H_1 : m \neq 0;\ there\ is\ a\ relationship\ between\ x\ and\ y,\ therefore\ slope\ m \neq 0.$$

Now from *Table 1*, we see that the \$p-value < 0.0001 \$; hence, $p - value < \alpha$. Therefore, we reject the null hypothesis H_0.

Regression Diagnostics

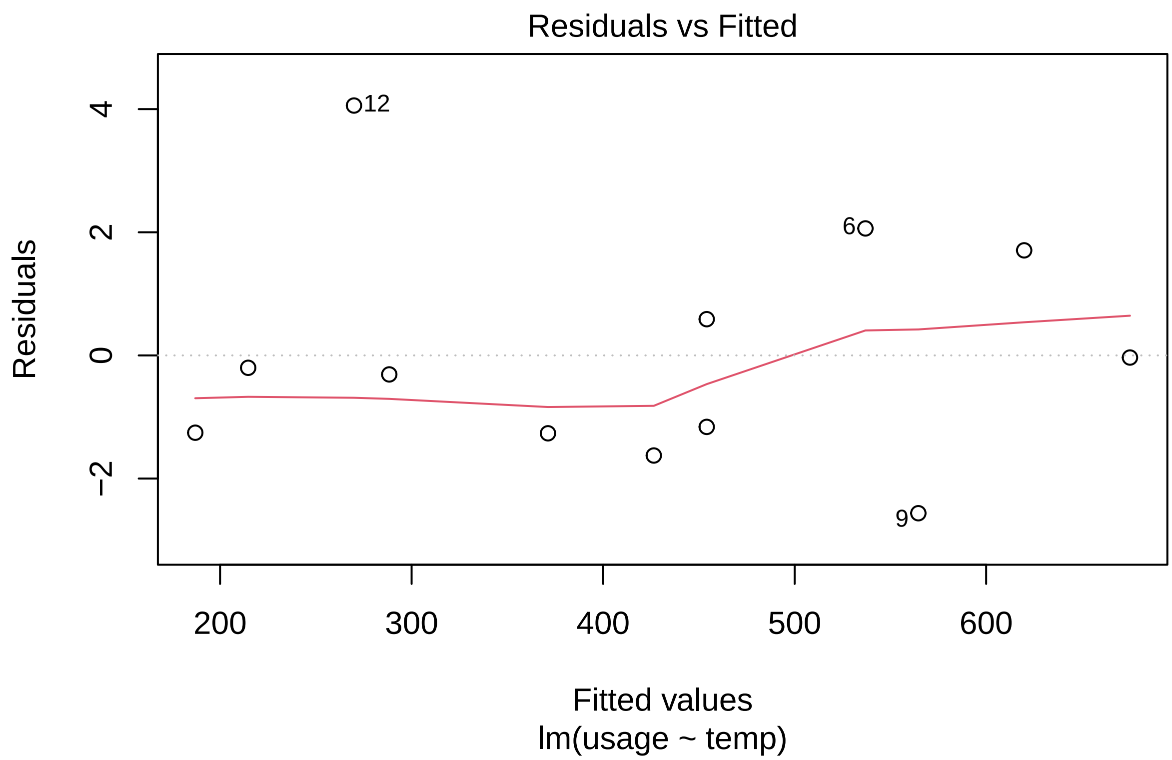

Figure 2: Residual Plot of Linear Regression Model

In *Figure 2*, where the Fitted Values are plotted against the Residuals, we observe the following:

1. The residuals are randomly and evenly distributed.
2. They are clustered around the horizontal band.
3. We notice a possible outlier, data-point 12.

After considering all of the above, we can conclude that our Linear Regression model used above is accurate and well-behaved. Furthermore, the *Adjusted* R^2 value of 0.999, as this signifies that the *temp* explains 99% of the variation of *usage* values.

Q3.

The *table.b3* data frame has observations on gasoline mileage performance for 32 different auto- mobiles. This data frame contains the following columns:

- y: Miles/gallon
- $x1$: Displacement (cubic in)
- $x2$: Horsepower (ft-lb)
- $x3$: Torque (ft-lb)
- $x4$: Compression ratio
- $x5$: Rear axle ratio
- $x6$: Carburetor (barrels)
- $x7$: No. of transmission speeds
- $x8$: Overall length (in)
- $x9$: Width (in)
- $x10$: Weight (lb)
- $x11$: Type of transmission (1=automatic, 0=manual)

Analysis 1

```
##
## Call:
## lm(formula = y ~ x1 + x2 + x3 + x4 + x5 + x6 + x7 + x8 + x9 +
##     x10 + x11, data = gas_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.3441 -1.6711 -0.4486  1.4906  5.2508
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 17.339838  30.355375   0.571   0.5749
## x1          -0.075588   0.056347  -1.341   0.1964
## x2          -0.069163   0.087791  -0.788   0.4411
## x3           0.115117   0.088113   1.306   0.2078
## x4           1.494737   3.101464   0.482   0.6357
## x5           5.843495   3.148438   1.856   0.0799 .
## x6           0.317583   1.288967   0.246   0.8082
## x7          -3.205390   3.109185  -1.031   0.3162
## x8           0.180811   0.130301   1.388   0.1822
## x9          -0.397945   0.323456  -1.230   0.2344
## x10         -0.005115   0.005896  -0.868   0.3971
## x11          0.638483   3.021680   0.211   0.8350
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.227 on 18 degrees of freedom
##   (2 observations deleted due to missingness)
## Multiple R-squared:  0.8355, Adjusted R-squared:  0.7349
## F-statistic:  8.31 on 11 and 18 DF,  p-value: 5.231e-05
```

Table 2. $gasModel_1$**: Full Multiple Linear Regression on Gas Data**

Now, we will plot the residual and QQ-Norm plots.

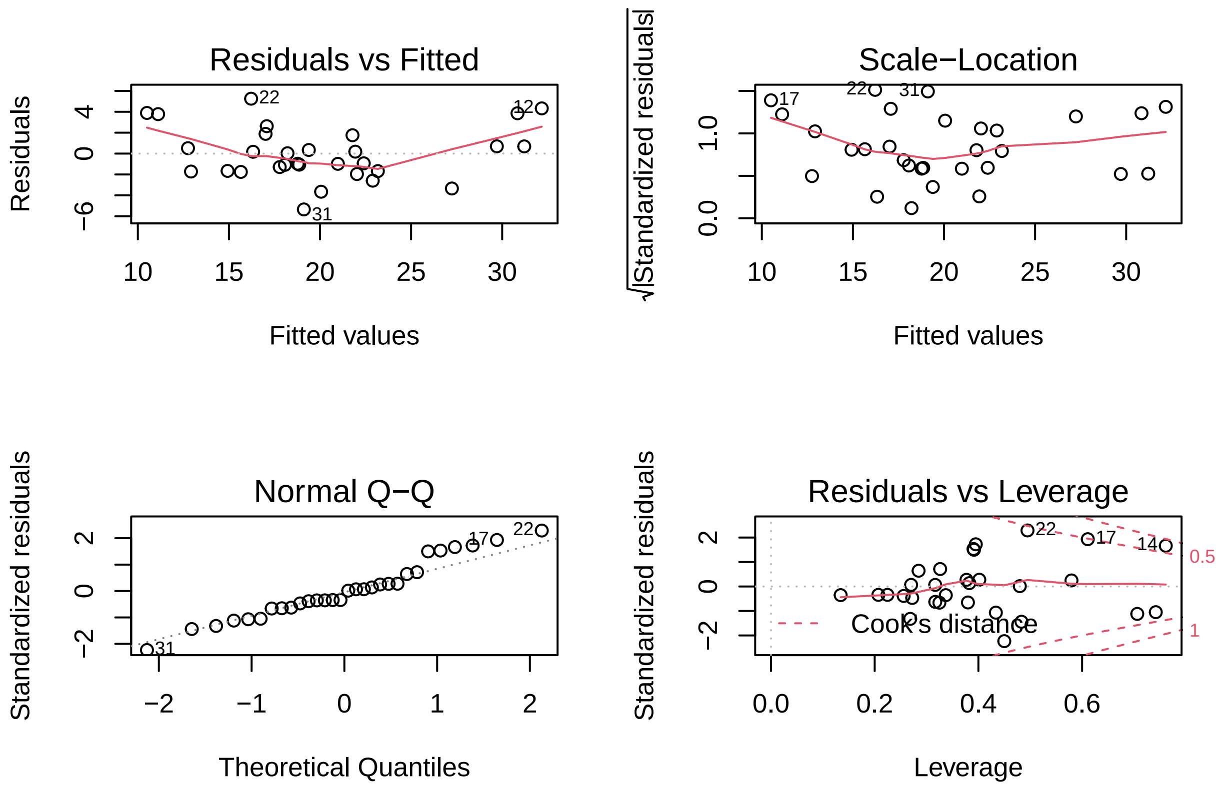

Figure 3: Diagnostic Plots of Multiple Linear Regression of Gas Data

Results 1

The goal of this question to fit a multiple linear regression model. To start off, the $p-values$ of all the variables are very insignifcant for the model to be considered. This could be solved by keeping only a few variables with significant $p-values$.

After diagnostic checking using residual and QQ-Norm plots, we will fit another model using only the $x5, x8, x10$ variables. Then we will compare the two models.

From the *Residuals vs Fitted* plot in *Figure 3*, we see that the data has multiple outliers and also has a cluster below the 25 value on the x-axis. Also, from the *QQ-Norm plot* in the same figure, we notice that the data does follow a normal distribution.

Analysis 2

As instructed by the quesiton, we will now use only the $x5, x8, x10$ variables in the regression model. Then, we ill compare the results with those from **Results 1**

```
##
## Call:
## lm(formula = y ~ x5 + x8 + x10, data = gas_data)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -4.512 -1.945 -0.631  1.931  6.003
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  5.010946  11.275042   0.444   0.6601
## x5           2.625031   1.202720   2.183   0.0376 *
## x8           0.211874   0.078850   2.687   0.0120 *
## x10         -0.009334   0.001702  -5.485 7.37e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.859 on 28 degrees of freedom
## Multiple R-squared:  0.8151, Adjusted R-squared:  0.7953
## F-statistic: 41.14 on 3 and 28 DF,  p-value: 2.156e-10
```

Table 3. $gasModel_2$**: Multiple Linear Regression on Gas Data with** $x5, x8, x10$ **variables.**

Now, we will plot the residual and QQ-Norm plots.

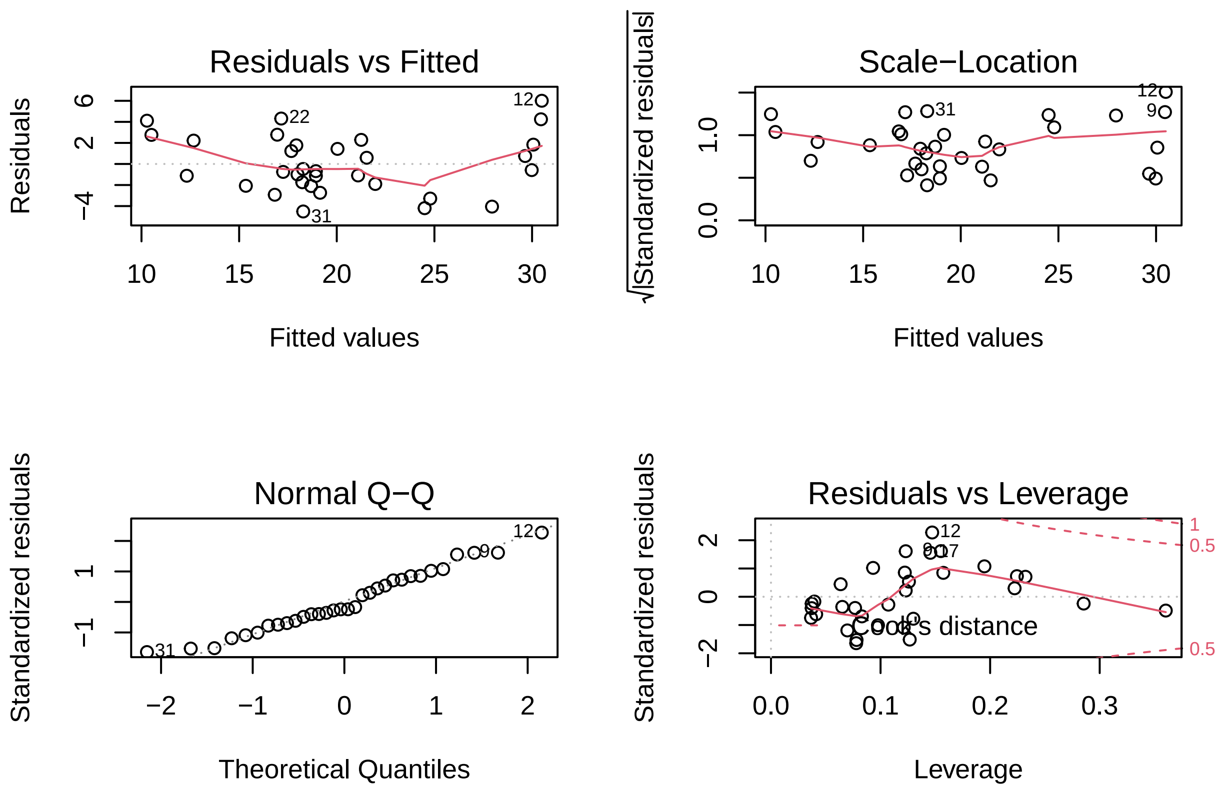

Figure 4: Diagnostic Plots of Multiple Linear Regression of Gas Data

Results 2

From *Table 3*, we find that all three variables have statistically significant $p - values$ for $\alpha = 0.05$.

Furthermore, in *Figure 4 - Residuals vs Fitted* plot, we see that all residual values are randomly and equally distributed. On top of that, the QQ-Norm plot in the same figure shows a hihgly normal distribution of the data. ALl of these combined show a reliable linear regression model.

Comparison of Results

Numeric: In *Analysis 1*, p-values of the variables used in the linear regression model ($gasModel_1$) did not show significance (except for $x5$). However, in *Analysis 2*, we see that all the p-values used in the model ($gasModel_2$) are statistically significant for $\alpha = 0.05$.

Graphical: Here, our main focus is on two graphs in particular: *Residuals vs Fitted* & *QQ-Norm*. In $gasModel_1$ from *Figure 3*, clusters are formed in the *Residual vs Fitted* plot. However, in $gasModel_2$ from *Figure 4 - Residuals vs Fitted* plot, we see that all residual values are randomly and equally distributed. This means that the variables used in $gasModel_2$ form a better model.

When looking at the Nornaml Q-Q plots of the two models. We see a better fitting plot for $gasModel_2$ in *Figure 4 - Normal Q-Q* to the normal distribution.

Preference: From our comparison of the two models, $gasModel_1$ and $gasModel_2$, it is clear that $gasModel_2$ is better preferred because:

1. The model shows higher significance of $p - value$.
2. Residual Plot shows better random distribution of the residual values without any clusters.
3. Shows a better normal distribution in the Normal Q-Q plot.

Q4.

The goal of the question is to generate the $ANOVA$ table of the data in Table 5.1 in the TextBook using the linear regression model. So, let us first import the dataset.

```
data("weightgain", package = "HSAUR3")
```

Next, we will perform Linear Regression and then use the *anova*() function to generate the $ANOVA$ table.

```
weightgain_lm <- lm(weightgain~source+ type, data=weightgain)
anova(weightgain_lm)
```

```
## Analysis of Variance Table
##
## Response: weightgain
##           Df Sum Sq Mean Sq F value  Pr(>F)
## source     1  220.9  220.90  0.9150 0.34501
## type       1 1299.6 1299.60  5.3829 0.02596 *
## Residuals 37 8933.0  241.43
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Explanation: When we perform an Analysis of Variance, categorical data such as the one in *weightgain* data set are coded with 1's and -1's so that each catrgory's mean is compared to the grand mean. However, in regression, the categorical variables are dummy coded, which means that each variable's intercept is compared to the reference group's intercept. To check this, we will use the *aov*() function wihtout linear regression to confirm the correctness of our result above.

```
aov(weightgain~ source+ type, data=weightgain)
```

```
## Call:
##    aov(formula = weightgain ~ source + type, data = weightgain)
##
## Terms:
##                  source   type Residuals
## Sum of Squares    220.9 1299.6    8933.0
## Deg. of Freedom       1      1        37
##
## Residual standard error: 15.5381
## Estimated effects may be unbalanced
```

Q5.

Removing the galaxies having leverage higher than 0.8. To do so, we will simply set the weights of the data points with leverage higher than 0.8 to 0 and update the mode.

First, let us generate a zero intercept linear model of the galaxy data:

```
galaxy_model <- lm(y ~ x-1, data=hubble)
print(coef(galaxy_model))
```

```
##        x
## 76.58117
```

Now, we will find data points with *leverage > 0.8* using *cooks.distance()*. Then, we will remove those data points from the data set and re-generate out zero intercept linear model.

```
high_leverage <- hatvalues(galaxy_model) > 0.08
hubble2 <- hubble[!high_leverage,]
galaxy_model <- lm(y ~ x-1, data=hubble2)
print(coef(galaxy_model))
```

```
##        x
## 79.78791
```

Lastly, we will perform the unit conversion below (referred from pg. 109 of HSAR) and calculate the new age of the universe.

```
Mpc <- 3.09*10^19
ysec <- 60^2 * 24 * 365.25
Mpcyear <- Mpc/ ysec
1/(coef(galaxy_model)/Mpcyear)
```

```
##           x
## 12272058551
```

So, the new age is \$13037418512 \$ years. Or, 13.04 *billion* years old.

Q6.

Data Description

Data on distances and velocities of 24 galaxies containing Cepheid stars, from the Hubble space telescope key project to measure the Hubble constant.

A data frame with 3 columns and 24 rows. The columns are:

- Galaxy: A (factor) label identifying the galaxy.
- y: The galaxy's relative velocity in kilometers per second.
- x: The galaxy's distance in Mega parsecs. 1 parsec is 3.09e13 km.

Plots (Linear and Quadratic)

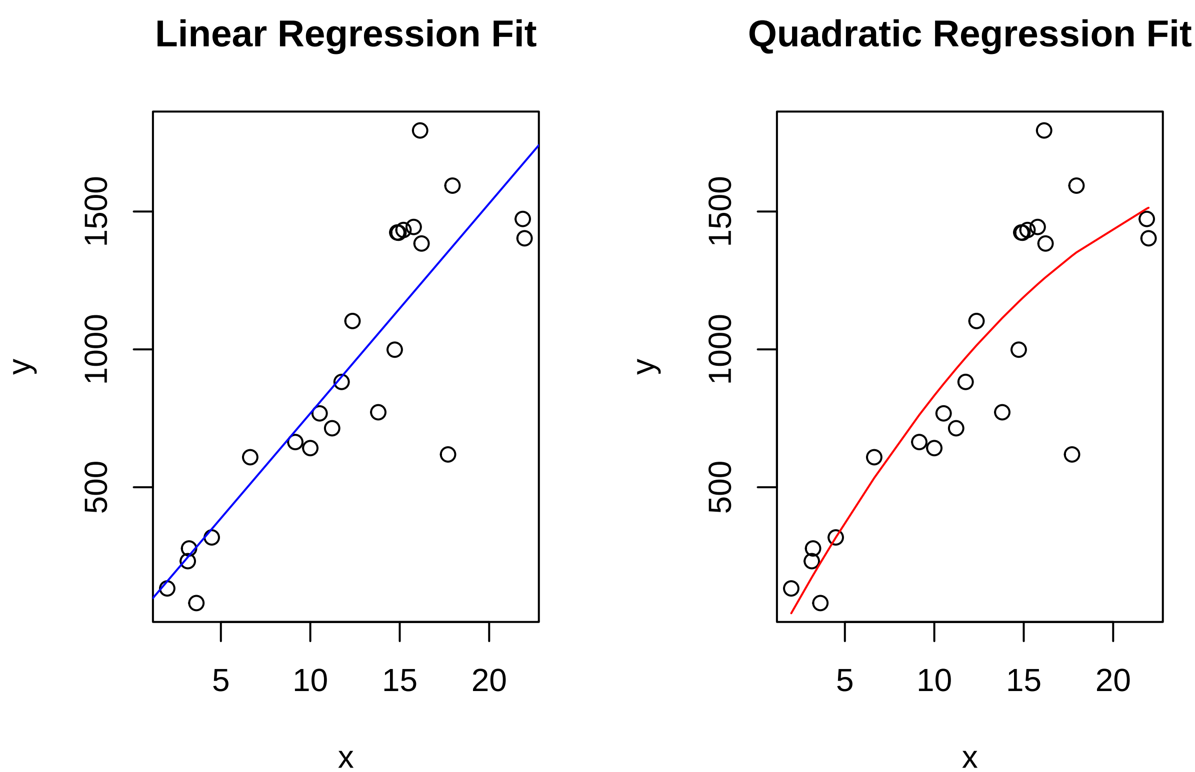

Figure 5: Regression Model fits on Scatter Plot for Linear and Quadratic

Which is better?

Just from looking at *Figure 5* and the scatter plot of the data points, we can see that the Quadratic Regression has a better fit for the data as it accounts for the data's non-linear behavior. This can also be confirmed in *Figure 6.a* as we see a cluster/patch of many positive residuals in the middle.

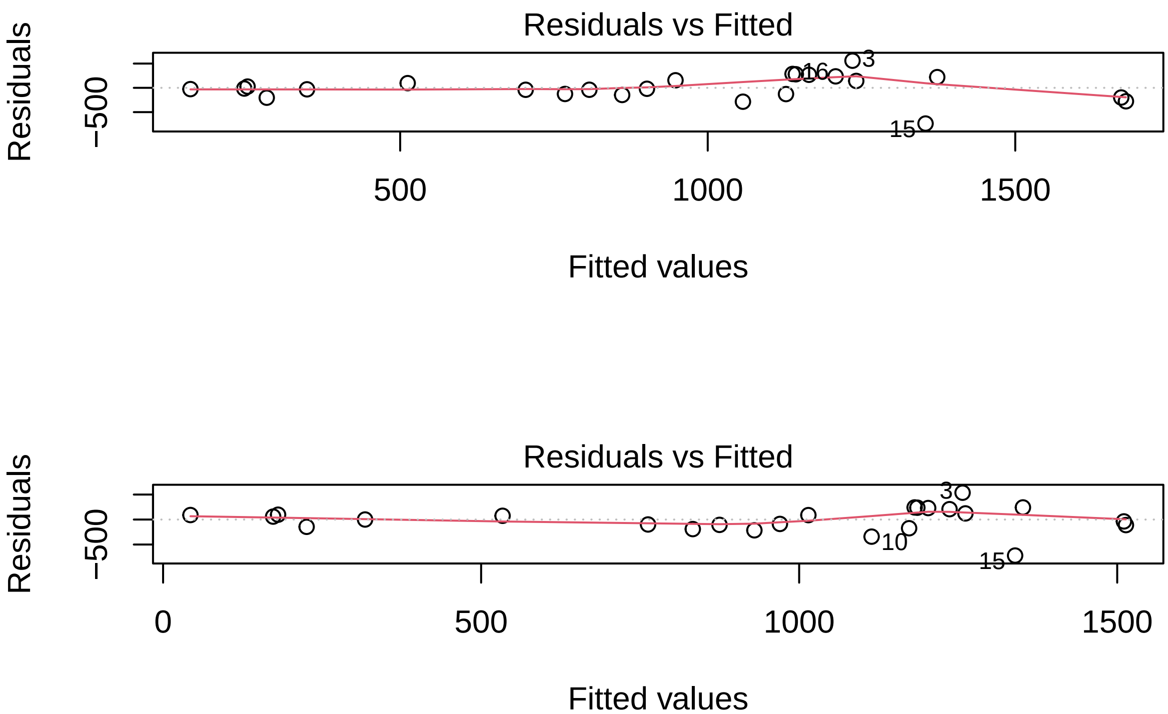

Figure 6: Top: Linear Model, Bottom: Quadratic Model