Regression Modelling Assignment 1

Joshua Redolfi

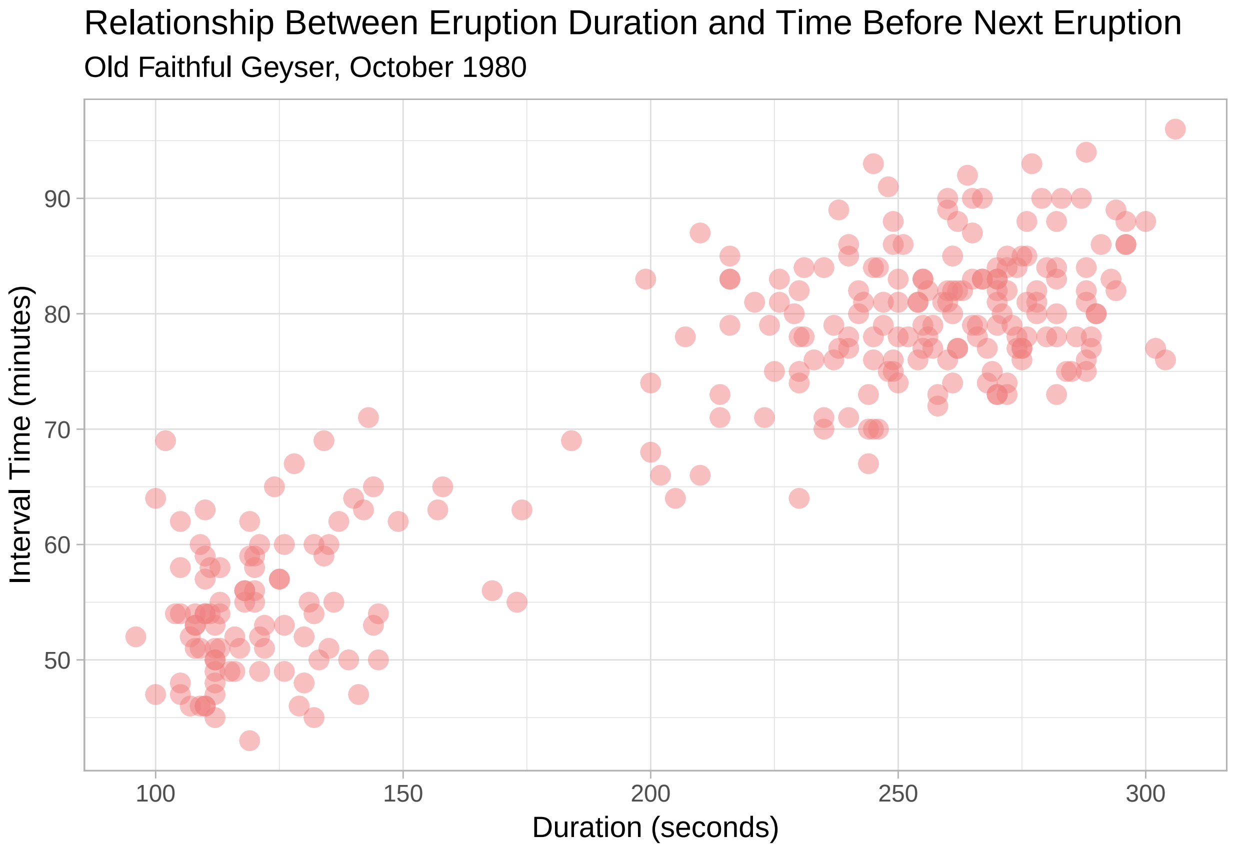
8 April 2019

Question 1

Data on eruptions of Old Faithful Geyser, in October 1980 was collected and stored in a *.csv* file *'oldfaithful'*. Variables are the duration in seconds of the current eruption, and the interval time in minutes to the next eruption. Data was not collected between approximately midnight and 6 AM. It is suspected that *Duration* is associated with the *Interval*

(a) Exploratory Data Analysis

We are interested in developing a linear regression model of the relationship between the duration of the current eruption and the interval of time preceding the next eruption of the Old Faithful Geyser using the data provided in a 1980 study. We begin our analysis by plotting the data below.



From the plot we see a strong positive correlation between Duration and Interval Time. We notice that the data is relatively clustered at the low end and the high end of the domain (Duration), however this characteristic shouldn't be a problem and overall a linear regression model would be a good fit. It is worth pointing out however that it is likely the observations are not independent, as the outcome of the previous case may impact the outcome of the next eruption, hence violating our assumption $\epsilon_i \overset{iid}{\sim} N(0,\sigma^2)$. For the purposes of this investigation however, we will ignore this concern.

We conduct a hypothesis test on the correlation between the predictor variable *Duration* and the response variable *Interval* as follows:

$$H_0 : \rho = 0, \; H_A : \rho \neq 0$$

The observed sample correlation $r = 0.8960697$ has a t-distribution with *268 degrees of freedom*:

$$t^* = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} = \frac{0.8960697\sqrt{270-2}}{\sqrt{1-0.8960697^2}} = 33.045$$

Computing the result in R, we have the following output:

```r
cor.test(x = Duration, y = Interval)
```

```
##
##  Pearson's product-moment correlation
##
## data:  Duration and Interval
## t = 33.045, df = 268, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.8697278 0.9173206
## sample estimates:
##       cor
## 0.8960697
```

The $p-value < 2.2 \times 10^{-16} << \alpha = 0.05$ so we reject the null hypothesis in favour of the alternative hypothesis. This concludes that the observed sample correlation is statistically significant and thereby suggests a strong positive correlation between the Duration and Interval Time.

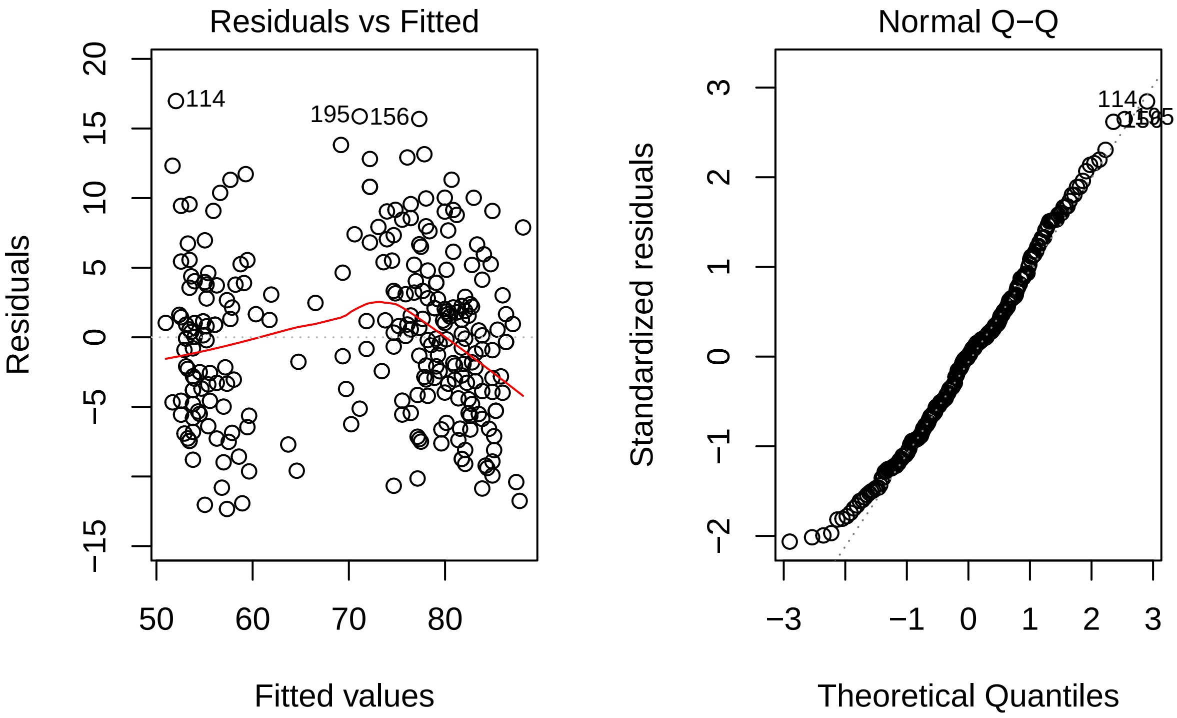
(b) Regression Diagnostics

We fit the following model:

$$Interval_i = \beta_0 + \beta_1(Duration_i) + \epsilon_i, \;\; \epsilon_i \overset{iid}{\sim} N(0,\sigma^2)$$

```r
lm.oldfaithful <- lm(Interval ~ Duration)
lm.oldfaithful
```

```
##
## Call:
## lm(formula = Interval ~ Duration)
##
## Coefficients:
## (Intercept)     Duration
##     33.9878       0.1769
```



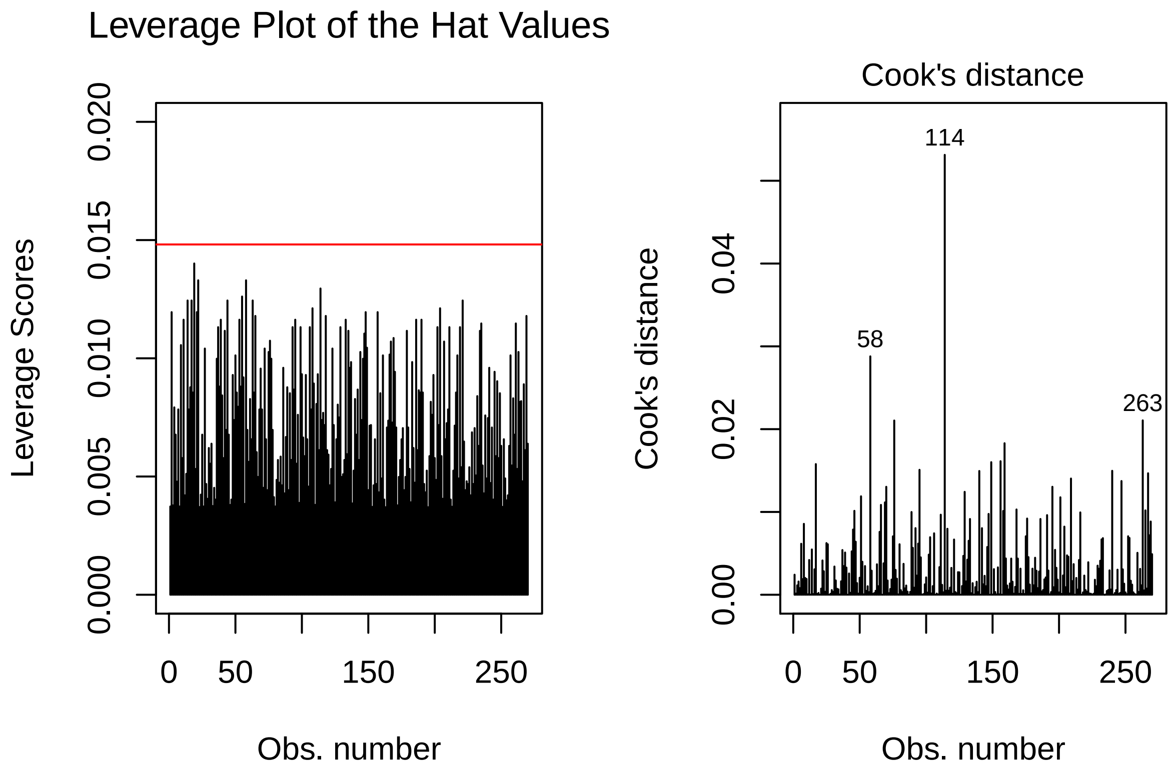
Upon careful observation of the residuals plots there are a few characteristics worth pointing out:

1. The residuals vs fitted values plot shows relatively constant variance. Despite the data being mainly clustered into two groups, we cannot assume that there is a distinctive pattern in the variability of the data due to limited observations in the low to mid range of the fitted values. Hence, it is reasonable to assume that our assumption $Var(\epsilon_i) = \sigma^2$ is consistent.

2. There appears to be a slight negative mean in the residuals corresponding to higher fitted values, however we can observe that in general the smoothing line tends to remain close to 0, indicating that $\mathbb{E}(\epsilon_i) = 0$.

3. The 114th, 195th and 156th observations do appear to induce a positive bias on the residual mean. However upon further examination of the normal quantile-quantile plot, these observations only deviate slightly from the normal line, and thus don't have much of an impact on the overall normality of the model.

4. A slight upward deviation from the normal line in the lower extreme of the theoretical quantiles scale indicate a slight fat left tail in the distribution of residuals, however this shouldn't cause any problems to the overall fit of the model.



We can observe from the leverage plot of the hat values that there are no major concerns. No extreme leverages can be observed, based on our $> 2\frac{p}{n} = \frac{4}{270}$ cut-off.

The Cook's Distance plot shows that observations 58, 114 and 263 are discernibly high in comparison to the other values. However as explained in our analysis, observation 114 shouldn't be of any concern. Observation 58 and 263 don't show up as a potential issues in the other diagnostic plots, and are less influential than observation 114 so overall we can conclude that the model shows no obvious problems.

(c) Analysis of Variance and F-Test

We conduct the following hypothesis test to assess the fit of the model:

$$H_0 : \frac{\sigma^2_{Model}}{\sigma^2_{Error}} = 1, \ H_A : \frac{\sigma^2_{Model}}{\sigma^2_{Error}} > 1$$

The observed test statistic F^* has an f-distribution with *1* and *268 degrees of freedom*:

$$F^* = \frac{\sum_{i=1}^{n}(\hat{y}_i - \overline{y_i})^2/1}{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2/(n-2)} = \frac{SSR/1}{SSE/n-2} = \frac{MSR}{MSE}$$

```
anova(lm.oldfaithful)
```

```
## Analysis of Variance Table
##
## Response: Interval
##             Df Sum Sq Mean Sq F value    Pr(>F)
## Duration    1  39358   39358    1092 < 2.2e-16 ***
## Residuals 268   9659      36
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

From the above ANOVA table, we have $MSE = 39358$ and $MSE = 36$, thereby computing the ratio derives the $F^* = 1092$. Our result yields a $p - value < 2.2 \times 10^{-16} << \alpha = 0.05$, hence there is significant evidence to reject the null hypothesis in favour of the alternative hypothesis, and conclude that the variance explained by the model is larger than the residual variance. Thus, it can be deduced from the F-Test that the Duration is explaining a significant proportion of the variability in Interval Time.

The Coefficient of Determination R^2 is a measure of how closely fit the data are about the regression line. It is evaluated from the ratio

$$R^2 = \frac{SSR}{SST} = \frac{\Sigma(\hat{y}_i - \overline{y})^2}{\Sigma(y_i - \overline{y})^2}$$

which represents the proportion of variability that can be explained by the model. In the model we've fitted, $R^2 = \frac{SSR}{SSR+SSE} = \frac{39358}{39358+9659} = 0.8029$, so based on this result, the model *Duration* is explaining 80.29% of the variability in *Interval*.

(d) T-Test for Regression Coefficients

```
summary(lm.oldfaithful)
```

```
##
## Call:
## lm(formula = Interval ~ Duration)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -12.3337  -4.5250   0.0612   3.7683  16.9722
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 33.987808   1.181217   28.77   <2e-16 ***
## Duration     0.176863   0.005352   33.05   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.004 on 268 degrees of freedom
## Multiple R-squared:  0.8029, Adjusted R-squared:  0.8022
## F-statistic:  1092 on 1 and 268 DF,  p-value: < 2.2e-16
```

From the above output, we see that the slope estimate is $\hat{\beta_1} = 0.176863$, and its associated standard error is $SE(\hat{\beta_1}) = 0.005352$. Likewise, the intercept estimate is $\hat{\beta_0} = 33.987808$ and $SE(\hat{\beta_0}) = 1.181217$. Using these results we conduct the following two hypothesis tests for SLR coefficients as follows:

$$H_0 : \beta_1 = 0,\ H_A : \beta_1 \neq 0 \ and \ H_0 : \beta_0 = 0,\ H_A : \beta_0 \neq 0$$

From the above R output we can evaluate the corresponding t^* for the two hypothesis tests, where both coefficient estimates have a *t-distribution* with *268 degrees of freedom*:

$$t^*_{\beta_1} = \frac{\hat{\beta}_1 - 0}{SE(\hat{\beta}_1)} = \frac{0.176863}{0.005352} = 33.05,\ t^*_{\beta_0} = \frac{\hat{\beta}_0 - 0}{SE(\hat{\beta}_0)} = \frac{33.987808}{1.181217} = 28.77$$

In the test for β_1, $p-value < 2.2 \times 10^{-16} << \alpha = 0.05$, therefore we reject H_0 in favour of H_A and thus conclude that there is a significant linear relationship between Duration and Interval Time. The estimated slope coefficient suggests that there is a ~0.177 minute increase in interval time before the next eruption for every 1 second increase in duration of the current eruption.

In the test for β_0, $p-value < 2.2 \times 10^{-16} << \alpha = 0.05$, therefore we reject H_0 in favour of H_A and thus conclude that β_0 is significantly different from 0. However, in this study, the intercept parameter isn't a useful measure, as the interval time preceding the next eruption cannot be determined if the duration of the current eruption is 0.

(e) Interval Estimation

If there is an eruption which lasted for 120 seconds, then the interval of time before the next eruption can be computing by taking the prediction interval:

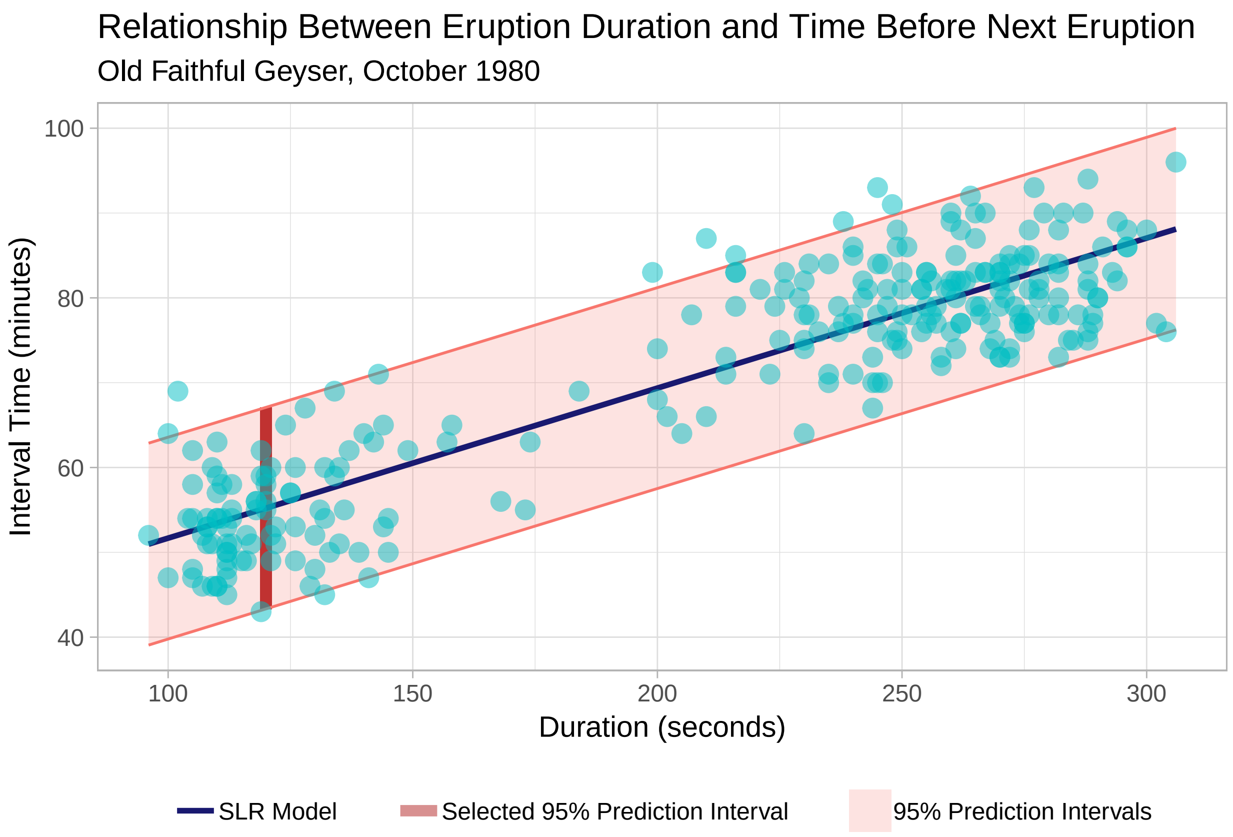
$$\hat{y_h} \pm t_{(\alpha/2,\ n-2)} \sqrt{MSE\Big(1 + \frac{1}{n} + \frac{(x_h - \overline{x})^2}{\sum_{i=1}^{n}(x_i - \overline{x})^2}\Big)}$$

$$\left(\hat{y_h} - t_{(0.025,\ 268)} \sqrt{39358\Big(1 + \frac{1}{270} + \frac{(120 - 209.8778)^2}{(S_{xx})^2}\Big)}\ ,\ \hat{y_h} + t_{(0.975,268)} \sqrt{39358\Big(1 + \frac{1}{270} + \frac{(120 - 209.8778)^2}{(S_{xx})^2}\Big)}\right)$$

```r
predict(lm.oldfaithful, newdata = data.frame(Duration = 120), interval = 'prediction')
```

```
##        fit      lwr      upr
## 1 55.21136 43.33156 67.09116
```

From the output in R we see $\hat{y_h} = 55.21136$ and the 95% Prediction Interval is $(43.33156,\ 67.09116)$. This suggests that although there is a statistically significant probability of linear association between duration and interval time as confirmed by the T-Tests for $\rho\ and\ \beta_1$ and the F-Test, our model cannot accurately predict the next outcome as the 95% prediction interval bounds are relatively far apart. In the case of $x_h = 120$ seconds for instance, the range of the prediction interval is 23.7596, spanning 44.83% of the range of interval times across the whole sample as seen in the graph below.

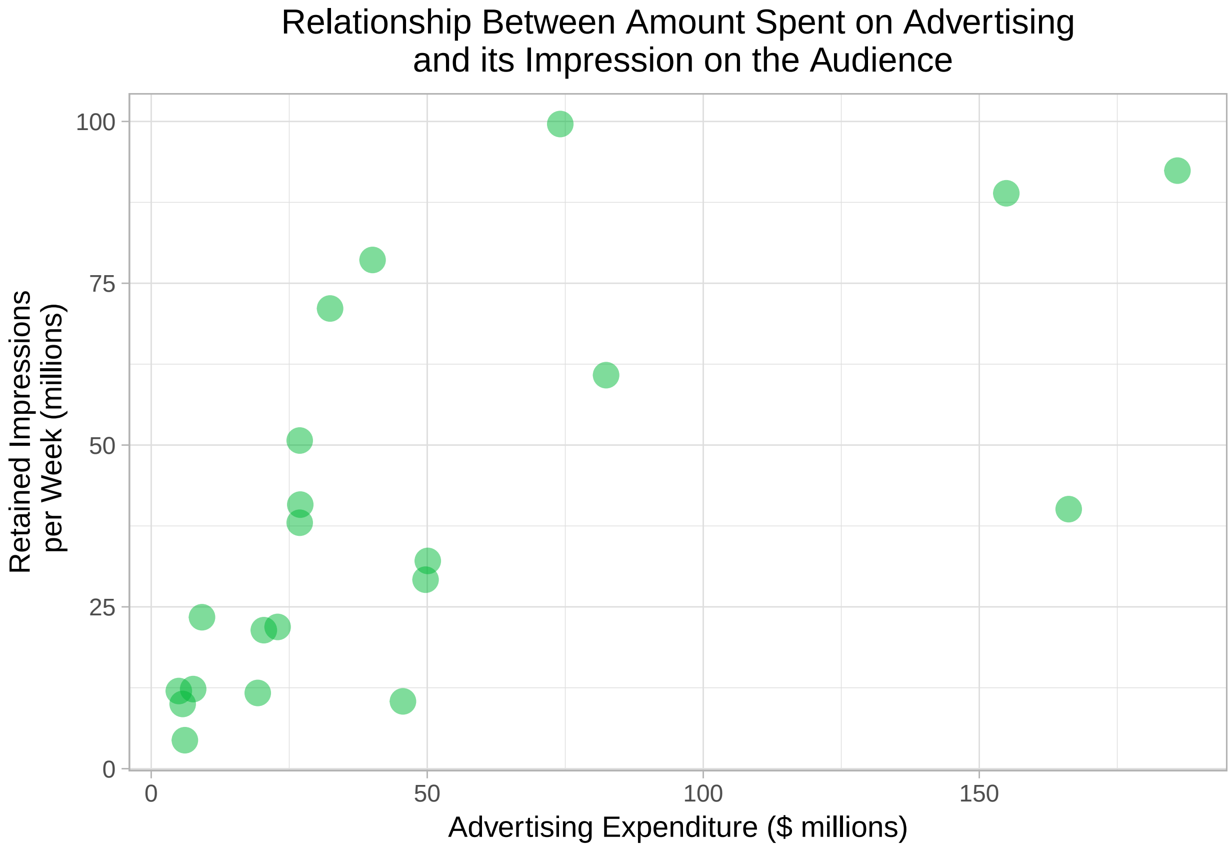


Question 2

On March 1, 1984, the Wall Street Journal published a survey of television advertisements conducted by Video Board Test, Inc., a New York ad-testing company that interviewed 4000 adults. These respondents were regular product users who were asked to cite a commercial they had seen for that product category in the past week. In this case, the response is the number of millions of retained impressions per week (*return*). The predictor, (*spend*), is the amount of money (in $ millions) spent by the firm on advertising. The data is available on wattle in .csv file called *advertising*.

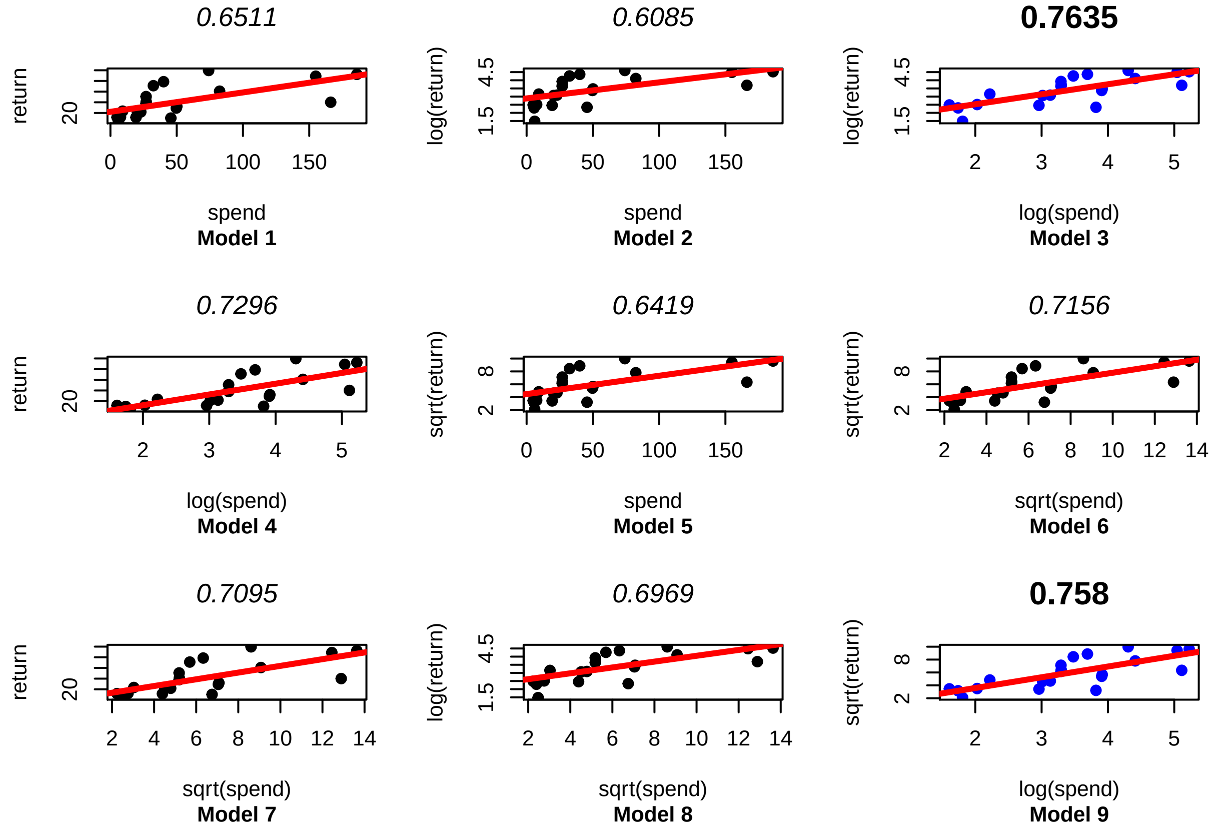
(a) Examining Possible Transformations

We are interested in developing a linear regression model of the relationship between firms' spending on television advertising, and the number of retained impressions per week from regular product users, using the data provided in a 1984 study. We begin our analysis by plotting the relationship below and examining some summary statistics concerning the data.



From the plot we see a general positive relationship between Advertising Expenditure and Retained Impressions per Week (the predictor variable *spend* has been renamed as *Advertising Expenditure* and the response variable *return* as *Retained Impressions per Week*). However the data in our plot appear to exhibit a curvilinear trend. This observation would indicate a violation of our assumption that $\epsilon_i \overset{iid}{\sim} N(0, \sigma^2)$, and so it would be unreasonable to fit a linear model on the raw data.

Therefore it is necessary to transform the data to produce a more appropriate model. We do so by selecting the model with the best fit out of nine possible combinations of transformations using the `log()` and `sqrt()` function on the predictor and response variables. Our selection criteria is based on the *Coefficient of Determination* R^2.



For each model the correlation coefficient r is plotted above its respective graph, and from this we determined the model with the greatest R^2 to be **Model 3**. It is however worth further investigating **Model 9**, as this model has a similarly high R^2. Lets analyse this potential association further by conducting a test for correlation in Model 3.

We conduct a hypothesis test on the correlation between the predictor variable log(spend) and the response variable log(return) as follows:

$$H_0 : \rho = 0, \;\; H_A : \rho \neq 0$$

The observed sample correlation $r = 0.7634849$ has a t-distribution with *19 degrees of freedom*:

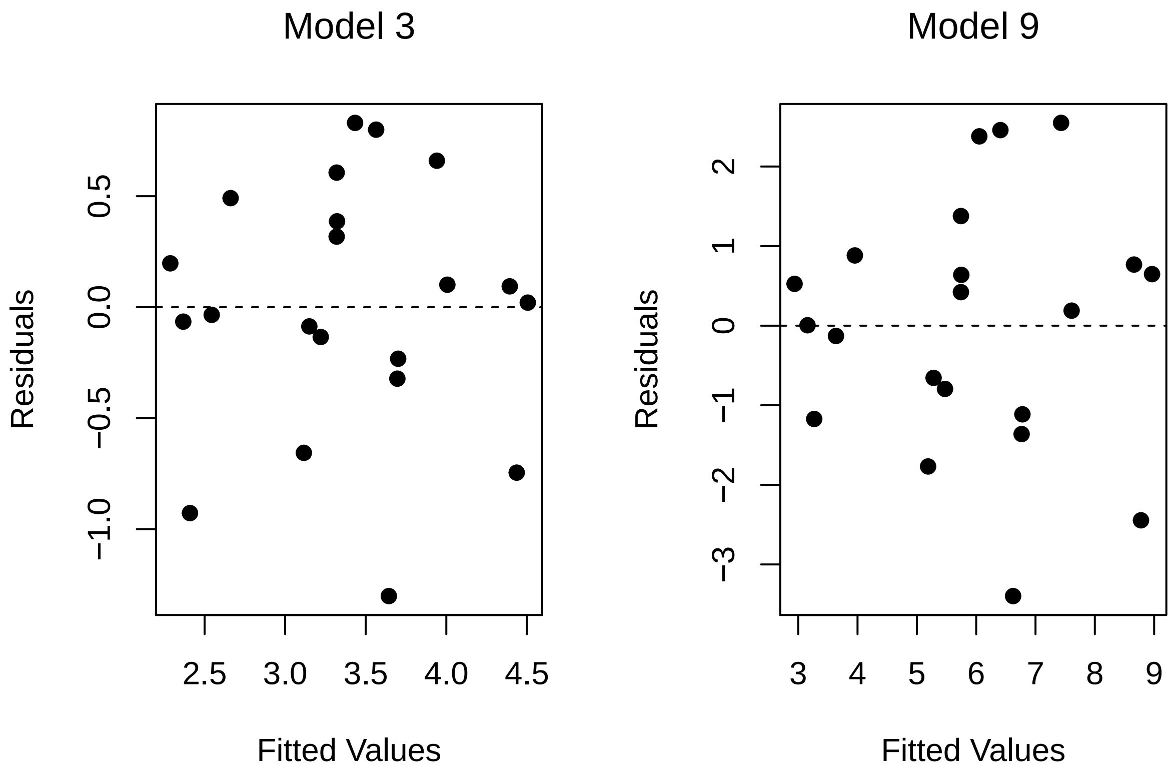
$$t^* = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} = \frac{0.7634849\sqrt{21-2}}{\sqrt{1-0.7634849^2}} = 5.153$$

Computing the result in R, we have the following output:

```
cor.test(x = log(spend), y = log(return))
```

```
##
##  Pearson's product-moment correlation
##
## data:  log(spend) and log(return)
## t = 5.153, df = 19, p-value = 5.655e-05
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.4949155 0.8989049
## sample estimates:
##       cor
## 0.7634849
```

The $p - value = 5.655 \times 10^{-5} << \alpha = 0.05$ so we reject the null hypothesis in favour of the alternative hypothesis. This concludes that the observed sample correlation is statistically significant and thereby suggests a strong positive correlation between log(spend) and log(return). It is also worth noting that a model, cannot simply be selected by it's R^2. This statistic is merely an indication of linear association, but fails to fully explain how the data is distributed and how potential influential points have affected the model. These properties will be analysed further in subsequent sections. Furthermore, there may also be concerns about the small sample size and of the independence of the data. We assume that of the 4000 respondents to the investigated survey, the responses of one individual have no influence on any other individual. For now however, lets look that the residuals vs fitted plots for **Model 3** and **Model 9**.



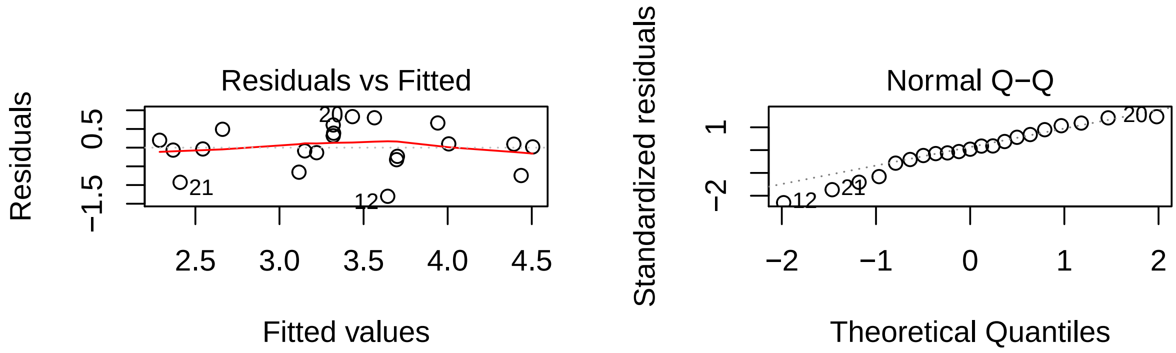
Examining the fitted values vs residual plots of the two models, we find that the data in Model 9 are heteroscedastic, whereby the variability in the data increases for higher fitted values. Model 3 however doesn't exhibit this behaviour, and is characterised by a far more constant variance across fitted values. Based on this criteria, we select Model 3:

$$\log(Return_i) = \beta_0 + \beta_1 \log(Spend_i) + \epsilon_i, \epsilon_i \overset{iid}{\sim} N(0, \sigma^2)$$

(b) Regression Diagnostics

```
lm(log(return) ~ log(spend))
```

```
##
## Call:
## lm(formula = log(return) ~ log(spend))
##
## Coefficients:
## (Intercept)   log(spend)
##      1.2999       0.6135
```

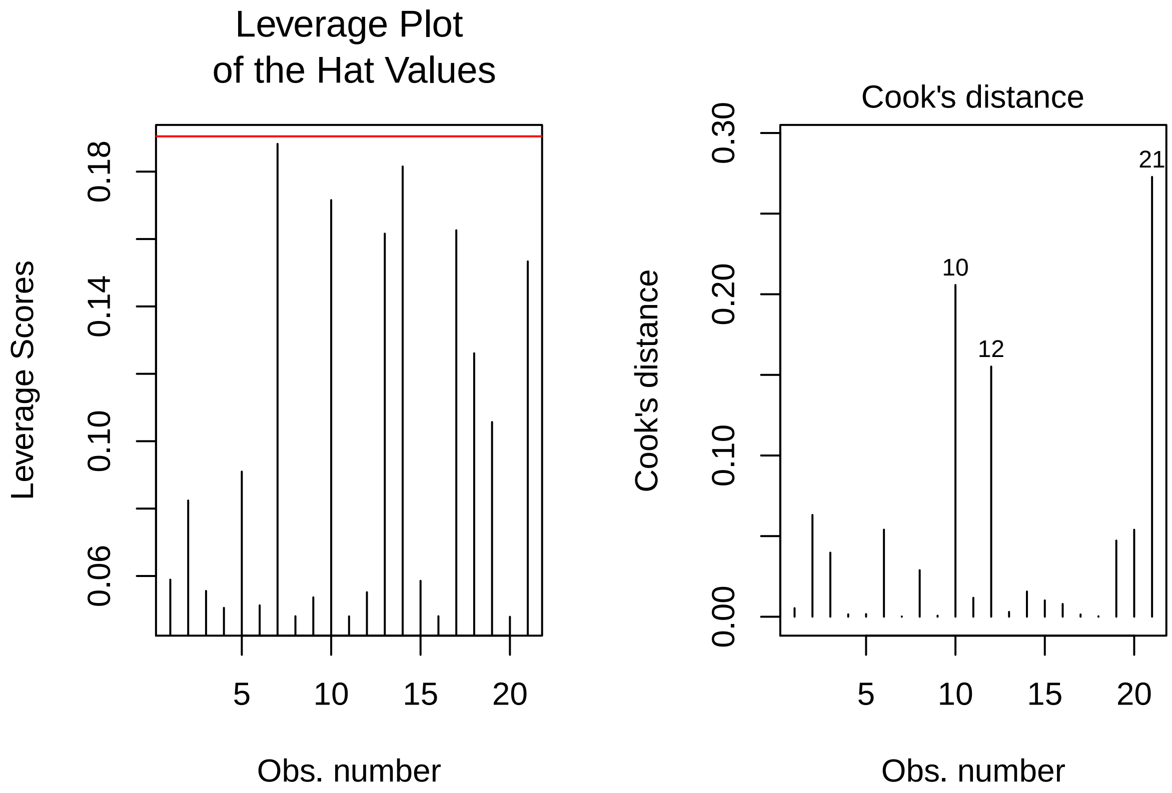


By observing our residual plots there are a few characteristics worth pointing out:

1. Firstly, as mentioned in our discussion in part (a), the residuals vs fitted values plot shows discernible patterns in the variance of the data, and so we can deduce that our e_i values are likely independent of each other, i.e. $Cov(e_i, e_j) = 0$, and that $Var(\epsilon_i) = \sigma^2$. Although the data appear to form a general elliptical shape, this is commonly observed in practice and so for the purposes of this investigation can be ignored.

2. We also note that the residuals are generally centred about 0, so that in our model $\mathbb{E}(\epsilon_i) = 0$. Although the 12th and possibly the 21st observation does appear to have a major influence on this condition, whether their influences on the model's fit is of concern or not, will be assessed further in relation to the other diagnostic plots. For now however, we assume normality amongst the residual term.

3. It can be identified from the normal quantile-quantile plot that the 12th and 21st observations do indeed deviate quite noticeably from the normal line, however considering that our sample is quite small this is to be expected. Overall, the model's fit does appear to exhibit a general normality.



Upon examination of the leverage plot of the hat values, we see that there are no major concerns. No extreme leverages can be observed, based on our $> 2\frac{p}{n} = \frac{4}{21}$ cut-off. In relation to the residual plots, although the 12th observation is quite deviant in the quantile-quantile plot, its leverage is very low. The 21st observation does have relatively high leverage, however its level of influence shouldn't have much of an impact on the overall fit of the model.

The Cook's Distance plot shows that observations 10, 12 and 21 are distinguishably high in comparison to the other values. However as explained in our analysis, observations 12 and 21 shouldn't be of any concern. Observation 10 doesn't show up as a potential issue in the other diagnostic plots, so overall we can conclude that the model shows no obvious problems.

(c) Analysis of Variance and F-Test

We conduct the following hypothesis test to assess the fit of the model:

$$H_0 : \frac{\sigma^2_{Model}}{\sigma^2_{Error}} = 1,\ \ H_A : \frac{\sigma^2_{Model}}{\sigma^2_{Error}} > 1$$

The observed test statistic F^* has an f-distribution with *1* and *19 degrees of freedom*:

```
anova(lm(log(return)~log(spend)))
```

```
## Analysis of Variance Table
##
## Response: log(return)
##            Df Sum Sq Mean Sq F value    Pr(>F)
## log(spend)  1 8.9623  8.9623  26.554 5.655e-05 ***
## Residuals  19 6.4128  0.3375
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

From the above ANOVA table, we have $MSE = 8.9623$ and $SSE = 0.3375$, thereby computing the ratio; $F^* = \frac{MSR}{MSE} = 26.554$. Our result yields a $p - value = 5.655 \times 10^{-5} << \alpha = 0.05$, hence there is significant evidence to reject the null hypothesis in favour of the alternative hypothesis, and conclude that the variance explained by the model is larger than the residual variance. Thus, it can be deduced from the F-Test that the predictor variable log(spend) is explaining a significant proportion of the variability in log(return).

As aforementioned in our analysis, the Coefficient of Determination R^2 is a measure of how closely fit the data are about the regression line. In the model we've fitted, $R^2 = \frac{SSR}{SSR+SSE} = \frac{8.9623}{8.9623+6.4128} = 0.5829$, so based on this result, the model log(spend) is explaining 58.29% of the variability in log(return).

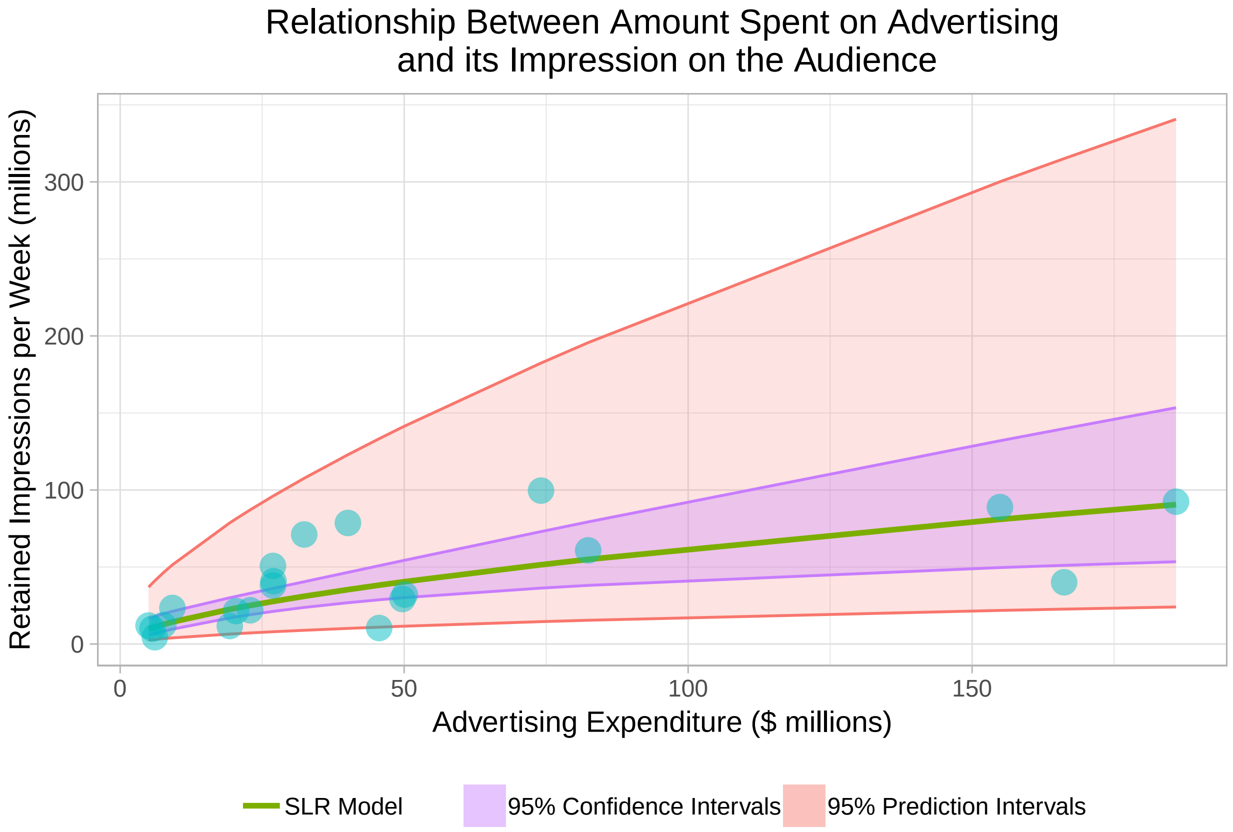
(d) Interpreting the Model

The true model $\log(Return_i) = \beta_0 + \beta_1 \log(Spend_i) + \epsilon_i, \epsilon_i \overset{iid}{\sim} N(0, \sigma^2)$ can be fitted using the regression estimates $\hat{\beta_1}$ and $\hat{\beta_0}$ as our best approximations for the true parameters β_1 and β_0. This relationship between the transformed variables log(spend) and log(return) is modelled by the equation

$$\log(\widehat{Return_i}) = \hat{\beta_0} + \hat{\beta_1} \log(Spend_i)$$

Expressing the predictor variable *spend* in terms of the response variable *return* we have

$$\widehat{Return_i} = e^{\hat{\beta_0}} (Spend_i)^{\hat{\beta_1}}$$



The fitted model can be visualised by mapping the transformed model onto the original scale as shown in the graph above. From this we can see that a $1 million increase in advertising expenditure leads to a non-linear increase in retained impressions per week. We can also observe from the graph that much of the data are scattered beyond the 95% confidence intervals, thereby indicating a lot of variability. Furthermore, the corresponding 95% prediction intervals expand in size significantly as advertising expenditure increases, and so advertising expenditure becomes less effective of an indicator of retained impressions per week. We can further investigate the effect of the coefficients on the response variable by equating the change in return as follows:

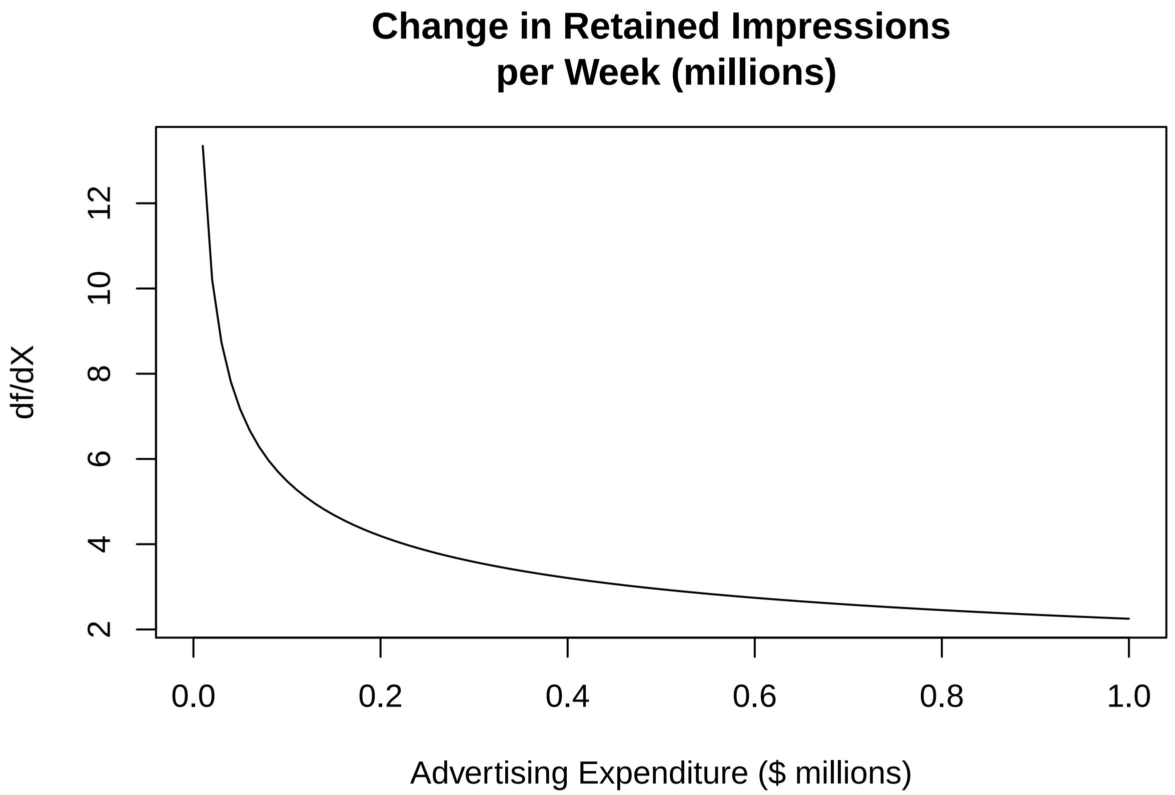
$$\therefore \Delta\ \widehat{Return}_i = \widehat{Return}_i - \widehat{Return}_{i-1} = e^{\hat{\beta}_0}(Spend_{i-1}+1)^{\hat{\beta}_1} - e^{\hat{\beta}_0}(Spend_{i-1})^{\hat{\beta}_1}$$

$$\Delta\ \widehat{Return}_i \approx 3.6691\left(\left(Spend_{i-1}+1\right)^{0.6135} - \left(Spend_{i-1}\right)^{0.6135}\right)$$

We can visualise this relationship by graphing the change in the slope of the SLR model. By computing the derivative of the model $f = e^{\beta_0}X_i^{\beta_1}$ w.r.t X_i, where $f = \widehat{Return}_i$ and $X_i = Spend_i$,

$$\frac{df}{dX_i} = \beta_1 e^{\beta_0}(X_i)^{\beta_1 - 1}$$

The graph of this equation as shown below shows us that there is an inverse relationship between the level of advertising expenditure and the change in retained impressions per week.


Change in Retained Impressions
per Week (millions)
df/dX
2
4
6
8
10
12
0.0
0.2
0.4
0.6
0.8
1.0
Advertising Expenditure ($ millions)