STAT 423 Homework 3

1. Two runners: (Marcel and Dani) are put under a cardiac stress test (Conconi test) which involves running on a treadmill. The test is conducted as follows:

- The athlete warms up for 10 minutes.
- The assistant sets the treadmill speed to the runners desired start speed.
- The assistant records the heart rate of the runner every 200 meters (.125 miles).
- The assistant increases the treadmill speed every 200 meters by 0.5km/hr (0.31 mph).
- The assistant stops the stopwatch when the athlete is unable to continue.

(a) We first need to preprocess the data. Create a data frame that contains all (non NA) observations of the variables `pulse`, `speed`, and `runner`. The `pulse` is the response and `speed` and `runner` are predictors, where `runner` should be a categorical predictor with the levels "Dani" and "Marcel" (0 and 1). Hint: There should be 39 samples. **Print your processed data frame.**

In this sub-part, we will preprocess the data from the file `runners.RDS`. In particular, we will create a data frame with all of the non NA observations of the `pulse`, `speed`, and `runner` variables. Where `runner` should be a categorical predictor with the levels "Dani" and "Marcel" (0 and 1). In total there should be 39 samples/rows. This data frame will be displayed in its entirety below.

```
## # A tibble: 39 x 3
##    speed runner pulse
##    <dbl>  <dbl> <int>
##  1   9        1   145
##  2   9.5      1   148
##  3  10        1   152
##  4  10.5      1   156
##  5  11        1   156
##  6  11.5      1   163
##  7  12        0   130
##  8  12        1   159
##  9  12.5      0   136
## 10  12.5      1   166
## # i 29 more rows
```

As can be seen by the above output, the data frame contains the expected columns, as well as the expected number of samples/rows.

(b) Print the scatter plot of `pulse` vs. `speed` with different colored points indicating each of the runners. Which model do you think is reasonable in this case?

In this sub-part, we will print the scatter plot of `pulse` vs. `speed` with different colored points indicating each of the runners. In addition to this, we will also comment on which model is reasonable in this case. The scatter plot of `pulse` vs. `speed` is shown below.

As can be seen from the above scatter plot of `pulse` vs. `speed` with different colored points indicating each of the runners, it seems as if a linear regression model is appropriate. However, including the `runner` variable as a covaraiate will be necessary to change the intercept dependent on who the runner is.

(c) Now fit an OLS regression model: `pulse ~ speed + runner`. What does this model assume with respect to the average starting pulse of each runner? What does it assume about the average increase in pulse for a 1 km/hr increase in speed for each of the two runners.

In this sub-part, we will fit an OLS regression model: `pulse ~ speed + runner` and interpret what some of the estimated coefficients mean with respect to the problem. The output from the `summary()` function is displayed below.

```
##
## Call:
## lm(formula = pulse ~ speed + as.factor(runner), data = runners_clean)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -6.364 -3.340  0.217  2.992  7.411
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)          66.3510     3.7310   17.78   <2e-16 ***
## speed                 5.1611     0.2169   23.80   <2e-16 ***
## as.factor(runner)1   37.0789     1.4096   26.30   <2e-16 ***
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.811 on 36 degrees of freedom
## Multiple R-squared:  0.959,  Adjusted R-squared:  0.9568
## F-statistic: 421.5 on 2 and 36 DF,  p-value: < 2.2e-16
```

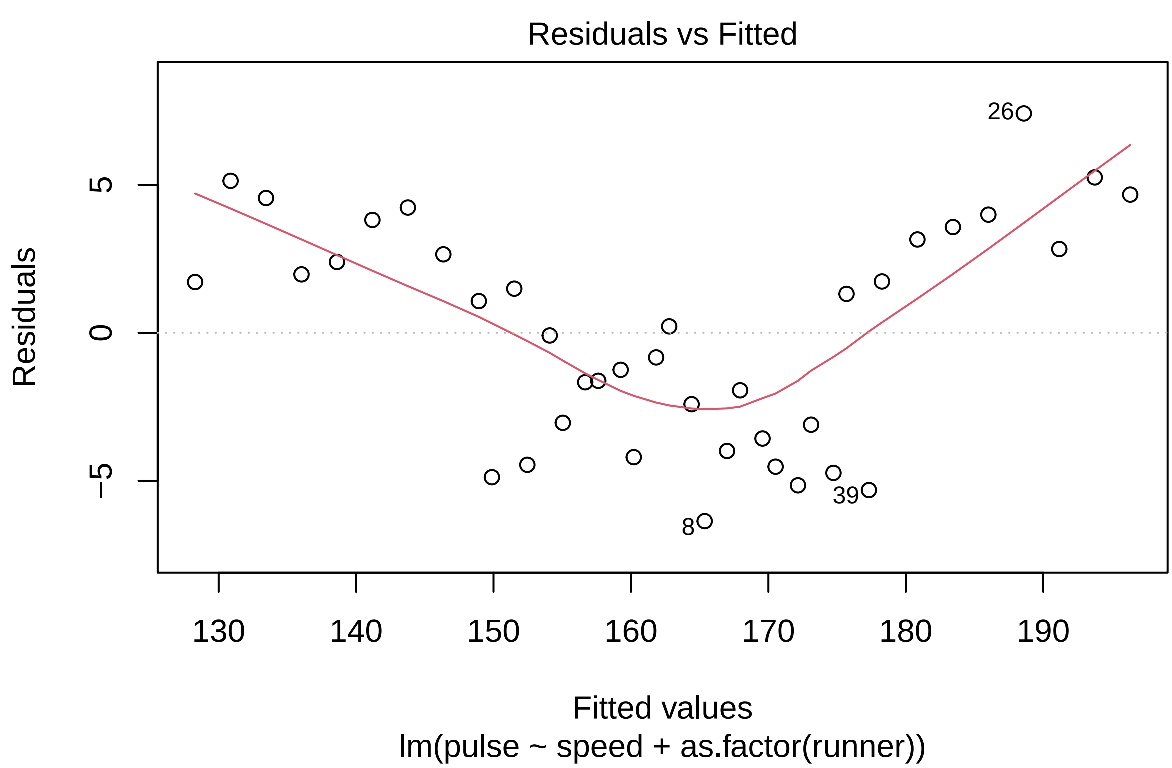
As seen by the above output, with 0/Dani being the reference group for the `runner` variable, the coefficient estimates are $\hat{\beta}_0 = 66.3510$, $\hat{\beta}_1 = 5.1611$, and $\hat{\beta}_2 = 37.0789$, which are all significant at any reasonable α level (even without any FWER or FDR corrections).

In terms of the average starting pulse of each runner, the model assumes that, for Dani, his starting pulse is $\hat{\beta}_0 = 66.3510$, while for Marcel, his starting pulse is $\hat{\beta}_0 + \hat{\beta}_2 = 66.3510 + 37.0789 = 103.4299$.

In terms of the average increase in pulse for a 1 km/hr increase in speed, the model assumes that, for both runners, the average increase in pulse is $\hat{\beta}_1 = 5.1611$. This is the same for both runners as there are no interaction terms, and hence, each category only contributes to a different intercept, not slope.

(d) Perform a residual analysis by plotting the "residuals vs. fitted" plot and the Normal QQ plot. Which model violations can we detect? **State all the assumptions you can check with these plots and whether you think they are satisfied.**

In this sub-part, we will perform a residual analysis of the above model, by plotting the "residuals vs. fitted" plot and the Normal QQ plot. For each plot, we will state all of the assumptions that we can check with these plots and whether we think they are satisfied or not. We will start with the "residuals vs. fitted" plot below.

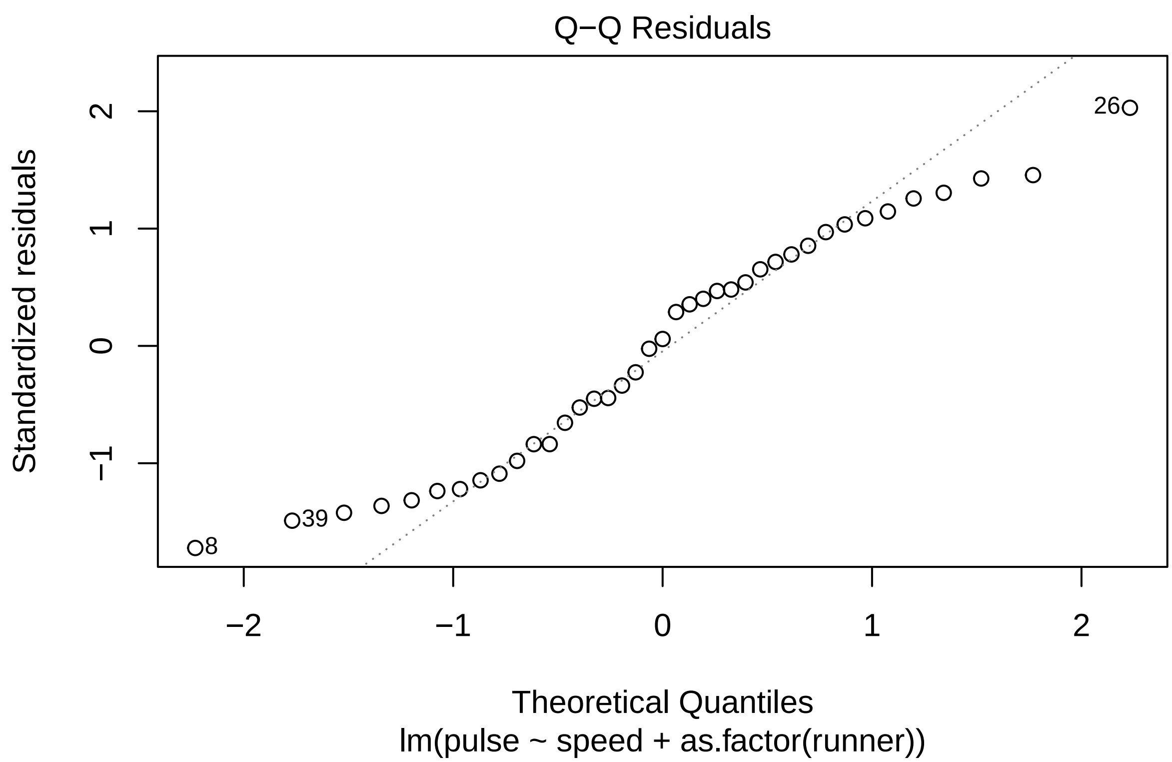


In a Tukey-Anscombe/"residuals vs. fitted" plot, one can check two assumptions. These two assumptions are: if $E[\epsilon_i] = 0$ is satisfied, and if $Var[\epsilon_i] = \sigma^2$ is satisfied.

Based on the above plot, since the plot does not show a flat scatter around 0, which is apparent due to the curvature of the loess curve, we have evidence that $E[\epsilon_i] \neq 0$. This also indicates the presence of non-linearity/the omission of an important predictor.

Furthermore, based on the above plot, the width of the points is greater in the middle range of the fitted values than it is for the lower and upper range of the fitted values. Thus, we have evidence that there is non-constant variance, that is, $Var[\epsilon_i] \neq \sigma^2$. It is important to note that the width across the ranges isn't drastically different, so this violation of the assumption would need to be tested further.

We will now plot the QQ plot of the residuals below.



Although it is possible to check that $E[\epsilon_i] = 0$ and $Var[\epsilon_i] = \sigma^2$ assumption with a QQ plot, the main assumption that is checked with a QQ plot is the $\epsilon_i \sim N(0, \sigma^2)$ assumption. If $\epsilon_i \sim N(0, \sigma^2)$, then the ordered standardized residuals $(\hat{\epsilon}_{(1)}, \ldots, \hat{\epsilon}_{(n)})$, should correspond linearly with the quantiles of a standard normal distribution.

Based on the above plot, it is apparent that the ordered standardized residuals, $(\hat{\epsilon}_{(1)}, \ldots, \hat{\epsilon}_{(n)})$, do not correspond linearly with the quantiles of a standard normal distribution. Hence we can see that the normality of the residuals assumption is violated in this model.

(e) Now, fit a model with an interaction term between `speed` and `runner`. What does this model assume with respect to the average starting pulse of each runner? What does it assume about the average increase in pulse for a 1 km/hr increase in speed for each of the two runners?

In this sub-part, we will fit an OLS regression model: `pulse ~ speed + runner + speed:runner` and interpret what some of the estimated coefficients mean with respect to the problem. The output from the `summary()` function is displayed below.

```
##
## Call:
## lm(formula = pulse ~ speed + as.factor(runner) + speed:as.factor(runner),
##     data = runners_clean)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.4947 -0.9034  0.2667  1.0588  3.6737
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)               84.2383     2.3574  35.734  < 2e-16 ***
## speed                      4.0932     0.1387  29.512  < 2e-16 ***
## as.factor(runner)1         2.3722     3.1330   0.757    0.454
## speed:as.factor(runner)1   2.3138     0.2042  11.333 2.91e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.788 on 35 degrees of freedom
## Multiple R-squared:  0.9912, Adjusted R-squared:  0.9905
## F-statistic:  1319 on 3 and 35 DF,  p-value: < 2.2e-16
```

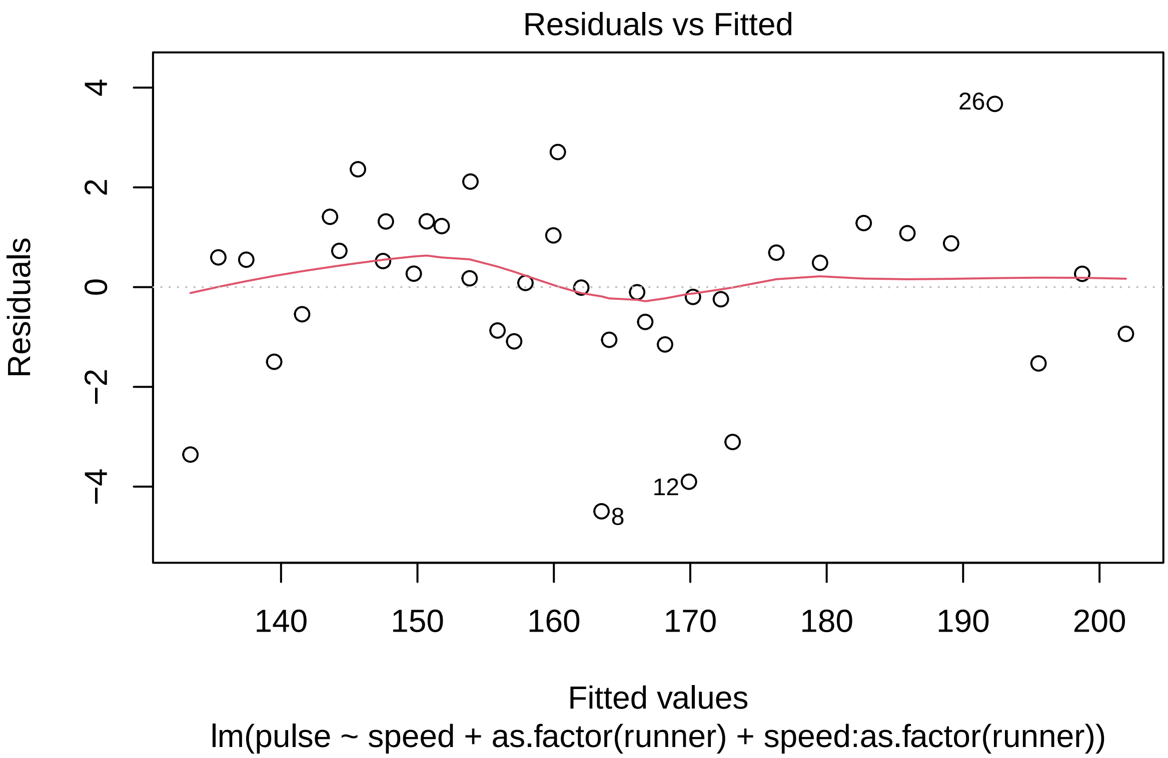
As seen by the above output, with 0/Dani being the reference group for the `runner` variable, the coefficient estimates are $\hat{\beta}_0 = 84.2383$, $\hat{\beta}_1 = 4.0932$, $\hat{\beta}_2 = 2.3722$, and $\hat{\beta}_3 = 2.3138$, the coefficients for the intercept, `speed`, and `speed:runner` are significant at any reasonable α level (even without ant FWER or FDR corrections). While the coefficient for `runner` is not significant at any reasonable α level.

In terms of the average starting pulse of each runner, the model assumes that, for Dani, his starting pulse is $\hat{\beta}_0 = 84.2383$, while for Marcel, his starting pulse is $\hat{\beta}_0 + \hat{\beta}_2 = 84.2383 + 2.3722 = 86.6105$.

In terms of the average increase in pulse for a 1 km/hr increase in speed, the model assumes that, for Dani, the average increase in pulse is $\hat{\beta}_1 = 4.0932$ for a 1 km/hr increase in speed, while for Marcel, the average increase in pulse is $\hat{\beta}_1 + \hat{\beta}_3 = 4.0932 + 2.3138 = 6.407$ for a 1 km/hr increase in speed.

(f) Perform a residual analysis (TA plot and Normal QQ plot) and discuss the model assumptions. **State all the assumptions you can check with these plots and whether you think they are satisfied.**

In this sub-part, we will perform a residual analysis of the model with an interaction, by plotting the "residuals vs. fitted" plot and the Normal QQ plot. For each plot, we will state all of the assumptions that we can check with these plots and whether we think they are satisfied or not. We will start with the "residuals vs. fitted" plot below.

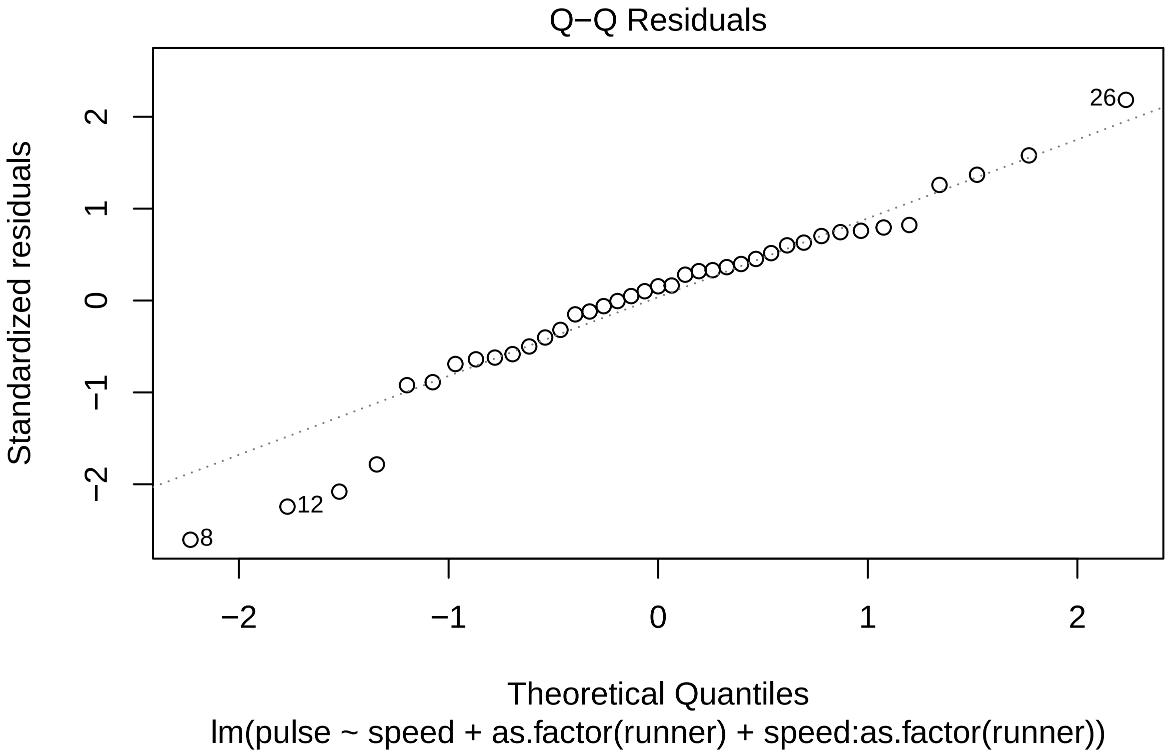


In a Tukey-Anscombe/"residuals vs. fitted" plot, one can check two assumptions. These two assumptions are: if $E[\epsilon_i] = 0$ is satisfied, and if $Var[\epsilon_i] = \sigma^2$ is satisfied.

Based on the above plot, since the plot shows a relatively flat scatter around 0, which is apparent due to the lack of curvature of the loess curve, we have evidence that the $E[\epsilon_i] = 0$ assumption is not violated. It is important to note that, due to the relatively small sample size, there are some parts of this plot that don't seem to follow the random scatter, for the most part this assumption looks good, but more testing would need to be done to confirm this.

Furthermore, based on the above plot, the width of the points seems to be relatively constant across all fitted values. Thus, we have evidence that there is constant variance, that is, $Var[\epsilon_i] = \sigma^2$. Again, it is important to note that, due to the relatively small sample size, there are some parts of this plot that seem to have a narrower width than the rest of the plot, for the most part this assumption looks good, but more testing would need to be done to confirm this.

We will now plot the QQ plot of the residuals below.



Although it is possible to check that $E[\epsilon_i] = 0$ and $Var[\epsilon_i] = \sigma^2$ assumption with a QQ plot, the main assumption that is checked with a QQ plot is the $\epsilon_i \sim N(0, \sigma^2)$ assumption. If $\epsilon_i \sim N(0, \sigma^2)$, then the ordered standardized residuals $(\hat{\epsilon}_{(1)}, \ldots, \hat{\epsilon}_{(n)})$, should correspond linearly with the quantiles of a standard normal distribution.

Based on the above plot, it is apparent that the ordered standardized residuals, $(\hat{\epsilon}_{(1)}, \ldots, \hat{\epsilon}_{(n)})$, seem to correspond linearly with the quantiles of a standard normal distribution. However, at the lower tail, there seems to be 4 points that noticeably differentiate themselves from the normal line. Some of these points are deemed as outliers, so more research on these points would need to be done in order to understand why they differ so much. Overall, we can see that, for the most part, the normality of the residuals assumption is not violated in this model.

(g) Using the full model (with interaction), compute the estimates of the average initial pulse (i.e. when `speed=0`) for each runner, as well as the estimates of the average pulse increase with every additional 1 km/hr in speed (for each runner).

In this sub-part, although we have already discussed this in part (e), we will compute the estimates of the average initial pulse (i.e. when `speed=0`) for each runner, as well as the estimates of the average pulse increase with every additional 1 km/hr in speed (for each runner).

For Dani, the estimate of his average initial pulse (i.e. when `speed=0`) is $\hat{\beta}_0 = 84.2383$. Furthermore, the estimate of his average pulse increase with every additional 1 km/hr in speed is $\hat{\beta}_1 = 4.0932$.

For Marcel, the estimate of his average initial pulse (i.e. when `speed=0`) is $\hat{\beta}_0 + \hat{\beta}_2 = 84.2383 + 2.3722 = 86.6105$. Furthermore, the estimate of his average pulse increase with every additional 1 km/hr in speed is $\hat{\beta}_1 + \hat{\beta}_3 = 4.0932 + 2.3138 = 6.407$.

2. The Australian Bureau of Agricultural and Resource Economics conducts an annual survey of the agroindustry. In 1991, 451 farms in New South Wales took part. The raw data is contained in the file `farm.RDS` available on Canvas. The variables have the following meanings.

`revenue`: target variable, total revenue of the farm.
`costs`: predictor, total costs of the farm.
`region`: predictor, code for different regions within New South Wales.
`industry`: predictor, code for the cultivation (1=(wheat), 2=(wheat, sheep, cattle), 3=(sheep), 4=(cattle), 5=(sheep, cattle)).

The aim is to fit a suitable regression model that explains the revenue of a farm. You will need to perform the following steps:

(a) Preprocess the data as needed, i.e define the necessary factor variables, assess whether transformations are necessary, etc. Check whether there are sufficiently many observations for all levels of the factor variables. The recommendation is that there are at least five observations for each level.

In this sub-part, we will preprocess the data as needed, that is, we will define the necessary factor variables, assess whether transformations are necessary, and check whether there are sufficiently many observations for all levels of the factor variables. In particular, the recommendation is that there are at least five observations for each level. We will start by pre-processing the data below.

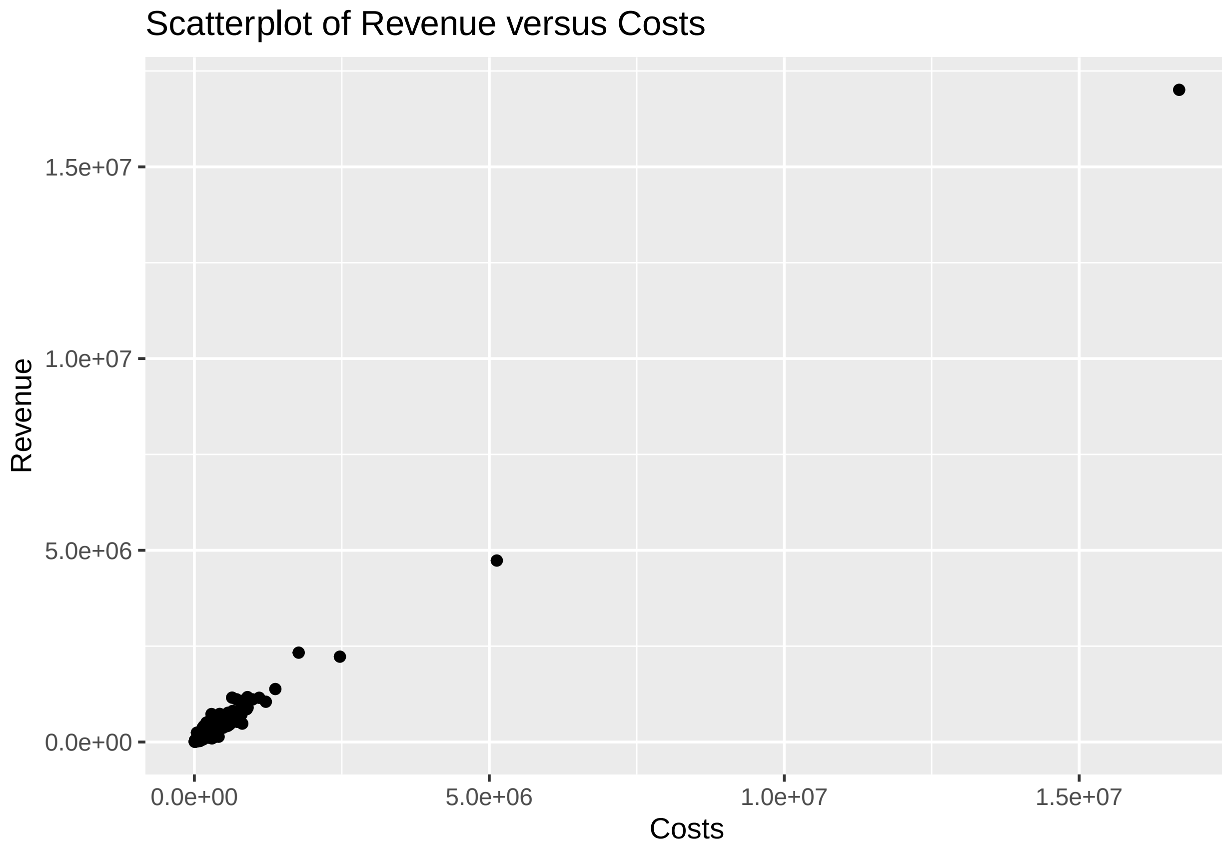
```
## [1] "region"   "industry" "costs"    "revenue"
```

As can be seen, the appropriate data attributes are present in the data frame, however, we must convert the `region` and `industry` variables to factors in order to create accurate models when using the `lm()` function. This code is shown below.

```
# Turn the region variable into a factor
farm_data$region <- as.factor(farm_data$region)

# Turn the industry variable into a factor
farm_data$industry <- as.factor(farm_data$industry)
```

We will now assess if any transformations need to be applied. Since `revenue` and `costs` are the only variables where transformations make sense, we will observe a scatter plot of `revenue` versus `costs` in order to see if a transformation is necessary.



Due to these extreme values of `costs` and `revenue`, it appears as if a transformation is necessary in order to account for the large outliers in the data set. Had we made a histogram, we would've came to the conclusion that our data is highly right skewed for both `costs` and `revenue`. This further emphasizes the need to make a transformation.

We will now check whether there are sufficiently many observations for all levels of the factor variables. In particular, the recommendation is that there are at least five observations for each level. This is done below.

```
## # A tibble: 6 x 2
##   region count
##   <fct>  <int>
## 1 111       30
## 2 121       95
## 3 122      103
## 4 123      108
## 5 131       81
## 6 132       34
```

As can be seen by the above table, each region has over 30 observations, hence there are sufficiently many observations for all levels of the `region` variable.

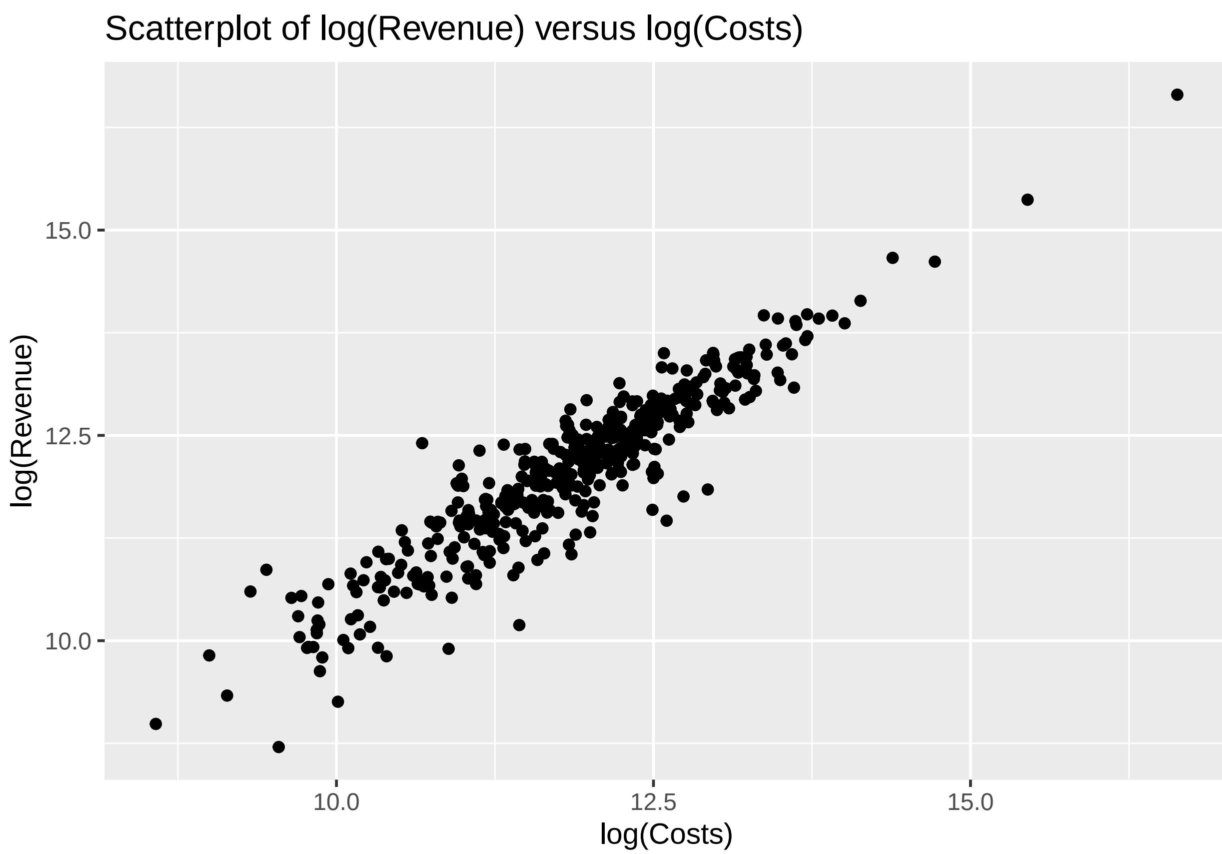
```
## # A tibble: 5 x 2
##   industry count
##   <fct>    <int>
## 1 1           37
## 2 2          137
```

```
## 3 3           100
## 4 4            86
## 5 5            91
```

As can be seen by the above table, each region has over 37 observations, hence there are sufficiently many observations for all levels of the `industry` variable.

(b) Explore the relationship between `revenue` and `costs` and choose a suitable transformation for `revenue` and `costs`. Fit a model that uses the transformed variables, along with `region` and `industry`, and perform a residual analysis (using the TA and QQ plots). Comment on the possible assumption violations. **State all the assumptions you can check with these plots and whether you think they are satisfied**

In this sub-part, we will explore the relationship between `revenue` and `costs` and choose a suitable transformation for `revenue` and `costs`. As was noticed in part (a), there seems to be a heavy right skew in the distribution of `revenue` and `costs`. Thus, even though there is a strong linear relationship between the two variables, the outliers make these relationships hard to detect. In order to scale down these large values, a *log* transformation might be appropriate. We will make a scatter plot of these transformed variables to assess the appropriateness.



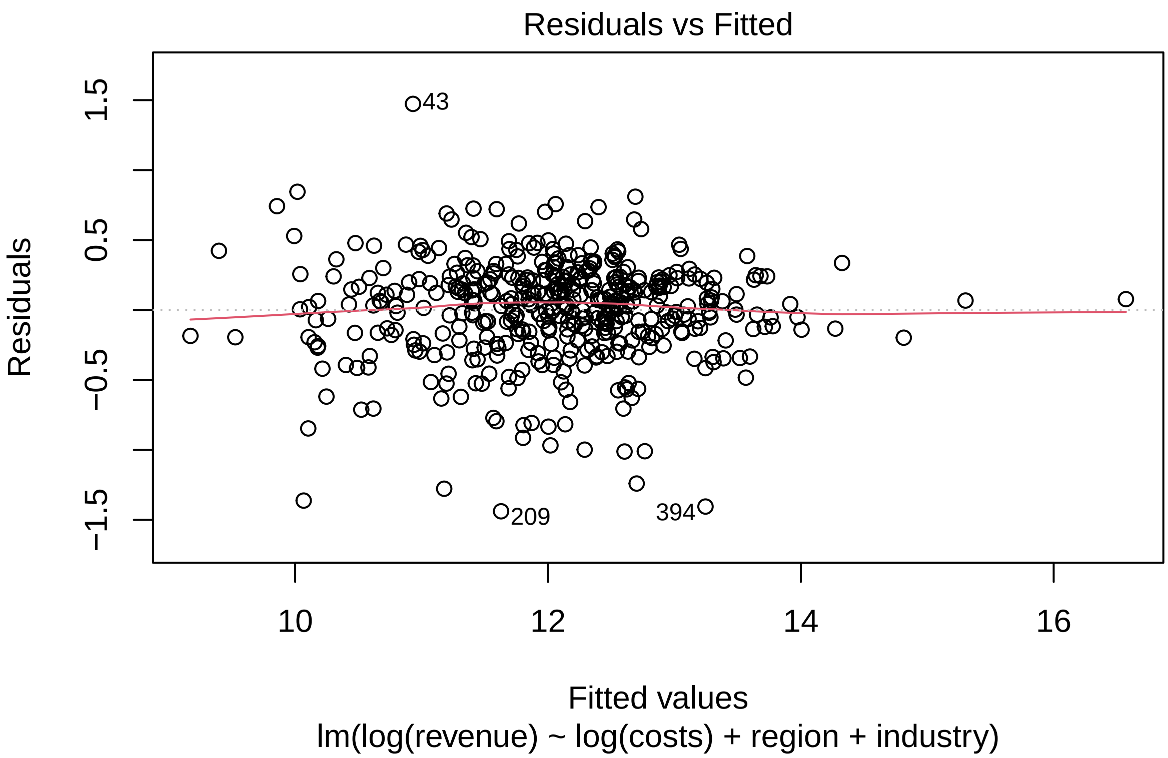
As can be seen from the above scatter plot of `log(revenue)` versus `log(costs)`, there still appears to be a strong positive linear relationship between the two variables. However, as opposed to the scatter plot of the original variables, this relationship is more discernible with no noticeable outliers. Hence it seems like the *log* transformation is appropriate.

We will now fit a model that uses the transformed variables, along with `region` and `industry`. This model is fit and shown below.

```
##
## Call:
## lm(formula = log(revenue) ~ log(costs) + region + industry, data = farm_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.43881 -0.17143  0.03773  0.22168  1.47317
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.379636   0.248432   5.553 4.86e-08 ***
## log(costs)   0.917954   0.018617  49.306  < 2e-16 ***
## region121   -0.076883   0.077353  -0.994  0.32081
## region122   -0.082997   0.076912  -1.079  0.28113
## region123   -0.036680   0.076151  -0.482  0.63027
## region131   -0.003855   0.079775  -0.048  0.96148
## region132   -0.243938   0.100536  -2.426  0.01565 *
## industry2   -0.155614   0.068023  -2.288  0.02263 *
## industry3   -0.222879   0.071421  -3.121  0.00192 **
## industry4    0.002649   0.075844   0.035  0.97215
## industry5   -0.171106   0.072947  -2.346  0.01944 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3612 on 440 degrees of freedom
## Multiple R-squared:  0.8712, Adjusted R-squared:  0.8683
## F-statistic: 297.7 on 10 and 440 DF,  p-value: < 2.2e-16
```

As can be seen above, based on the F-statistic value of 297.7 and corresponding p-value of less than 2.2×10^{-16}, this model is very significant when compared to the empty model. However, now we will perform a residual analysis, as done below.

We will now perform a residual analysis of the above model, by plotting the "residuals vs. fitted" plot and the Normal QQ plot. For each plot, we will state all of the assumptions that we can check with these plots and whether we think they are satisfied or not. We will start with the "residuals vs. fitted" plot below.

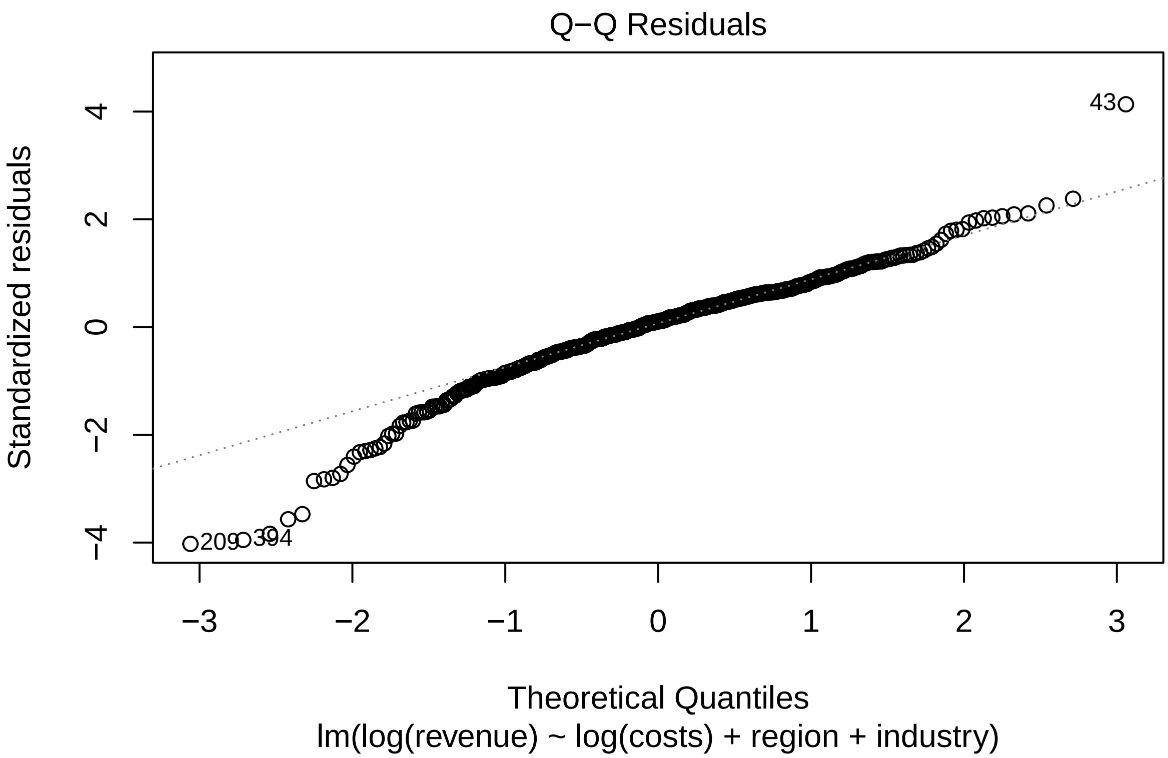


In a Tukey-Anscombe/"residuals vs. fitted" plot, one can check two assumptions. These two assumptions are: if $E[\epsilon_i] = 0$ is satisfied, and if $Var[\epsilon_i] = \sigma^2$ is satisfied.

Based on the above plot, since the plot does shows a flat scatter around 0, which is apparent due to the lack of curvature in the loess curve, we have evidence that $E[\epsilon_i] = 0$.

Furthermore, based on the above plot, the width of the points is greater in the lower and middle range of the fitted values, than it is for the upper range of the fitted values. However, the difference in width is not great, thus we have evidence that there is constant variance, that is, $Var[\epsilon_i] = \sigma^2$. It is important to note that, although the width across the ranges isn't drastically different, there still exists a difference, so this assumption would need to be tested further.

We will now plot the QQ plot of the residuals below.



Although it is possible to check that $E[\epsilon_i] = 0$ and $Var[\epsilon_i] = \sigma^2$ assumption with a QQ plot, the main assumption that is checked with a QQ plot is the $\epsilon_i \sim N(0, \sigma^2)$ assumption. If $\epsilon_i \sim N(0, \sigma^2)$, then the ordered standardized residuals $(\hat{\epsilon}_{(1)}, \ldots, \hat{\epsilon}_{(n)})$, should correspond linearly with the quantiles of a standard normal distribution.

Based on the above plot, it is apparent that the ordered standardized residuals, $(\hat{\epsilon}_{(1)}, \ldots, \hat{\epsilon}_{(n)})$, do not correspond linearly with the quantiles of a standard normal distribution, due to the noticeable deviation in the lower tail. Hence we can see that the normality of the residuals assumption is violated in this model. It is important to note that, since only the lower tail deviates from the normal line, it is possible that the normality assumption is met, therefore more testing would need to be done in order to confirm this conclusion.

(c) What is the expected revenue of a cattle farm. in region 111 with costs of 100,000?

In this sub-part, we will calculate the expected revenue of a cattle farm in region 111 with costs 100,000. We can do this by using the `predict()` function in R. This is done below.

As calculated in R above, the expected `log(revenue)` of a cattle farm in region 111 with costs of 100,000 is 11.95063, however, after exponentiating this `log(revenue)` value, we obtain the expected `revenue` value of 154914.2

(d) Test whether `region` has an influence on `revenue` when the other predictors are given at the 1% level.

In this sub-part, we will test whether `region` has an influence on `revenue` when the other predictors are given at the 1% level. We can do this by comparing a model with `region` to the model without region using the `anova()` function in R. This is done below.

```
## Analysis of Variance Table
##
## Model 1: log(revenue) ~ log(costs) + industry
## Model 2: log(revenue) ~ log(costs) + region + industry
##   Res.Df    RSS Df Sum of Sq      F  Pr(>F)
## 1    445 58.775
## 2    440 57.411  5    1.3639 2.0906 0.06551 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

As can be seen above, due to the p-value of 0.06551, we fail to reject the null hypothesis that the model with `revenue` is better than the model without `revenue` at the 1% level of significance. Hence we have no significant evidence that `region` has an influence on `revenue` when the other predictors are given.

(e) Add an interaction term between `region` and `industry`: `fit.farm <- lm(log(revenue) ~ log(costs) + region + industry + region:industry, data=farm)`.

In this sub-part, we will add an in interaction term between `region` and `industry`. We will refit this new model and show its output using the `summary()` function in R. This is done below.

```
##
## Call:
## lm(formula = log(revenue) ~ log(costs) + region + industry +
##     region:industry, data = farm_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.37258 -0.18678  0.03876  0.21251  1.47728
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)          1.54428    0.43571   3.544 0.000438 ***
## log(costs)           0.91217    0.01887  48.342  < 2e-16 ***
## region121           -0.13496    0.37154  -0.363 0.716598
## region122           -0.23159    0.38986  -0.594 0.552813
## region123           -0.16758    0.37553  -0.446 0.655657
## region131           -0.13986    0.50982  -0.274 0.783962
## region132           -0.19606    0.51073  -0.384 0.701264
## industry2           -0.29756    0.38005  -0.783 0.434095
## industry3           -0.31046    0.37178  -0.835 0.404156
## industry4            0.29796    0.51009   0.584 0.559442
## industry5           -0.31699    0.41646  -0.761 0.446998
## region121:industry2  0.17383    0.39686   0.438 0.661595
## region122:industry2  0.21391    0.41083   0.521 0.602864
## region123:industry2  0.14175    0.39769   0.356 0.721684
## region131:industry2  0.19572    0.54109   0.362 0.717748
## region132:industry2 -0.72288    0.63562  -1.137 0.256069
## region121:industry3  0.01654    0.39026   0.042 0.966206
## region122:industry3  0.13837    0.40869   0.339 0.735109
## region123:industry3  0.13416    0.39466   0.340 0.734077
## region131:industry3  0.19389    0.52297   0.371 0.711021
## region132:industry3 -0.62464    0.57689  -1.083 0.279527
## region121:industry4 -0.40906    0.52414  -0.780 0.435574
```

```
## region122:industry4 -0.19486    0.54065  -0.360 0.718718
## region123:industry4 -0.08965    0.53142  -0.169 0.866113
## region131:industry4 -0.41839    0.63026  -0.664 0.507153
## region132:industry4 -0.39939    0.62825  -0.636 0.525311
## region121:industry5  0.09692    0.43733   0.222 0.824713
## region122:industry5  0.12975    0.44978   0.288 0.773124
## region123:industry5  0.17872    0.43678   0.409 0.682613
## region131:industry5  0.21976    0.55412   0.397 0.691869
## region132:industry5  0.11321    0.57940   0.195 0.845174
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3604 on 420 degrees of freedom
## Multiple R-squared:  0.8777, Adjusted R-squared:  0.8689
## F-statistic: 100.5 on 30 and 420 DF,  p-value: < 2.2e-16
```

i. How many parameters are estimated in total?

In this sub-section, we will count the total number of parameters that are estimated. As can be seen above, there are 30 parameters estimated (31 if you include the intercept estimation, and 32 if you include the variance estimation).

ii. Is the interaction term significant at the 1% level?

In this sub-section, we will test whether the interaction term is significant/has an influence on `revenue` when the other predictors are given at the 1% level. We can do this by comparing a model with the interaction term to the model without the interaction term using the `anova()` function in R. This is done below.

```
## Analysis of Variance Table
##
## Model 1: log(revenue) ~ log(costs) + region + industry
## Model 2: log(revenue) ~ log(costs) + region + industry + region:industry
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1    440 57.411
## 2    420 54.540 20    2.8706 1.1053 0.3404
```

As can be seen above, due to the p-value of 0.3404, we fail to reject the null hypothesis that the model with the interaction term is better than the model without the interaction term at the 1% level of significance. Hence, we have no significant evidence that the interaction term has an influence on `revenue` when the other predictors are given.

iii. Based on this whole exercise, which model would you choose to predict the revenue of a farm?

In this sub-section, based on this whole exercise, I will decide which model I would choose to predict the revenue of a farm. Due to the fact that we failed to reject the null hypothesis that the interaction term has an influence on `revenue` when the other predictors are given, when compared to the model with no interaction term, I would choose the smaller model (the model from part (b)) to predict the `revenue` of a farm. By choosing this model we save a lot of degrees of freedom that can be used to make our estimates more precise.

3. Run the following code to create the vectors `x1`, `x2`, and `y` `> set.seed(1)`
```
> n <- 100
> x1 <- runif(n)
> x2 <- runif(n,10,20)
> y <- 2+2*x1+0.3*x2+rnorm(n)
```

(a) The last line of the code above corresponds to creating a linear model in which `y` is a function of `x1` and `x2`. Write out the form the linear model. What are the values of the regression coefficients β_0, β_1, and β_2? What is the value of σ^2?

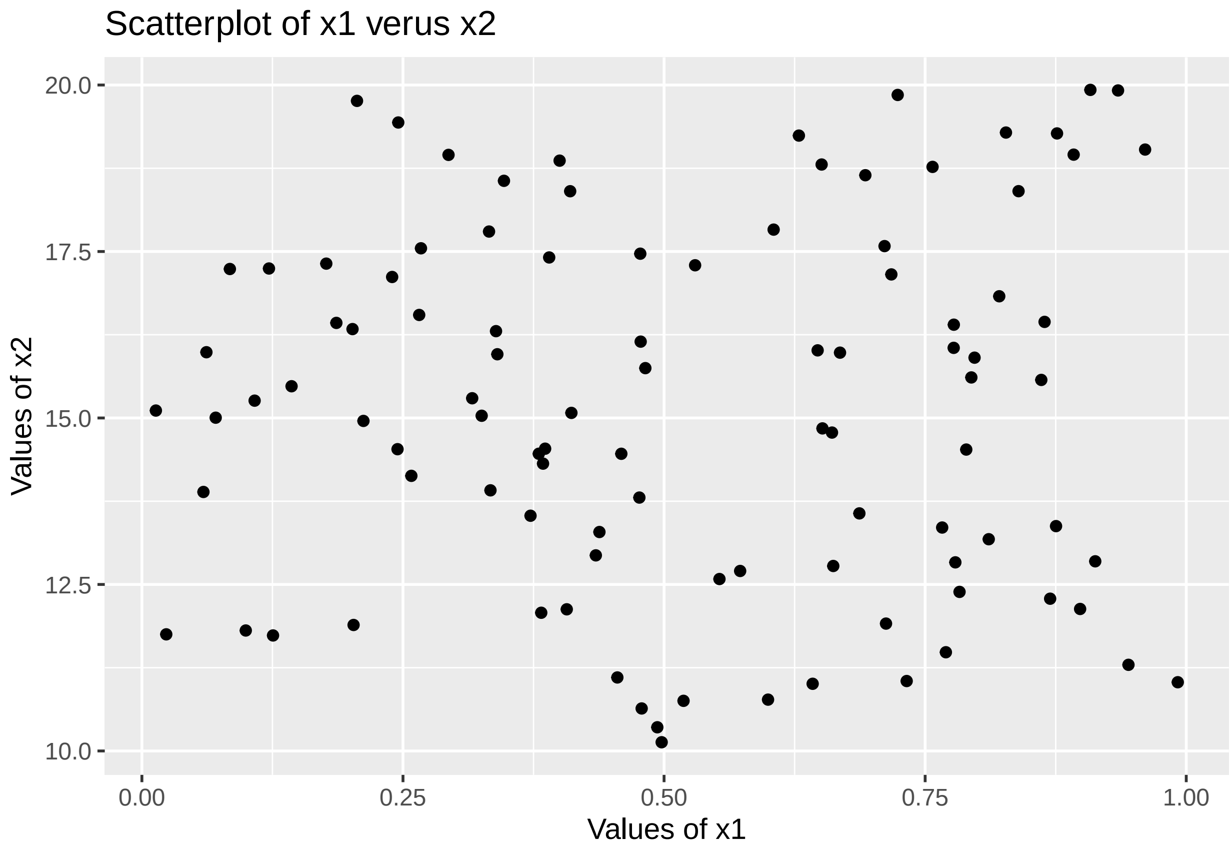
In this sub-part, we will write out the form the linear model and discern the values of the regression coefficients β_0, $beta_1$, and β_2, as well as the value of σ^2. This is done below.

The linear model that is represented by the line of code `y <- 2+2*x1+0.3*x2+rnorm(n)`, is $y_i = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon_i$. Therefore, the values of the regression coefficients β_0, β_1, and β_2, are 2, 2, and 0.3, respectively. Furthermore, the term `rnorm(n)` tells us the that errors are randomly drawn from a `N(0, 1)` distribution (standard normal). Therefore, we can see that the value of σ^2 is 1.

(b) Use the function `cor()` to calculate the correlation coefficient between `x1` and `x2`. Create a scatter plot using `ggplot2` displaying the relationship between the variables `x1` and `x2`. What can you say about the direction and strength of their relationship.

In this sub-part, we will use the `cor()` function to calculate the correlation coefficient between `x1` and `x2`, as well as make a scatter plot of these two variables to further analyze the relationship between them, this is done below.

As computed in R, the correlation coefficient between `x1` and `x2` is 0.01703215, which represents a very weak positive linear relationship, thus we expect to see little to no linear relationship between the two variables when making a scatter plot of them. We will use `ggplot2` to make this scatter plot below.



As seen by the above scatter plot of `x2` versus `x1`, we have confirmed that there appears to be no linear relationship between the two variables (furthermore, there appears to be no relationship at all). This provides us evidence that there is no multicollinearity between the two predictor variables, which is a good thing.

(c) Fit a least squares regression to predict `y` using `x1` and `x2`. Describe the obtained results. What are the values of $\hat{\beta}_0$, $\hat{\beta}_1$ and $\hat{\beta}_2$? How do these relate to the true values of β_0, β_1 and β_2? What is the value of s and how does it relate to the true value of σ^2? Can you reject the null hypothesis $H_0 : \beta_1 = 0$? How about the null hypothesis $H_0 : \beta_2 = 0$?

In this sub-part, we will fit a least squares regression to predict `y` using `x1` and `x2`, and describe the results appearing in the output of the `summary()` function. This is done below.

```
##
## Call:
## lm(formula = y ~ x1 + x2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.8579 -0.6167 -0.1432  0.5352  2.3318
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.97628    0.57973   3.409 0.000951 ***
## x1           1.93074    0.36345   5.312 6.89e-07 ***
## x2           0.30144    0.03578   8.425 3.33e-13 ***
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9675 on 97 degrees of freedom
## Multiple R-squared:  0.5095, Adjusted R-squared:  0.4994
## F-statistic: 50.38 on 2 and 97 DF,  p-value: 9.917e-16
```

As can be seen by the above R about from fitting `y~x1+x2`, we can see that we obtained an F-statistic of 50.38 with a p-value of 9.917×10^{-16}, which means that the model is significant when compared to the empty model.

Furthermore, our estimates of $\hat{\beta}_0$, $\hat{\beta}_1$ and $\hat{\beta}_2$, were 1.97628, 1.93074 and 0.30144, respectively. These estimates are close to the true values of β_0, β_1 and β_2, which were 2, 2 and 0.3, respectively.

Also, as computed in R, the value of s is 0.9675158. This means that s^2 is 0.9360869. These estimates are close to the true values of σ and σ^2, which are both 1.

Lastly, the p-values associated with $\hat{\beta}_1$ and $\hat{\beta}_2$, were 6.89×10^{-7} and 3.33×10^{-13}, respectively. Thus, at any reasonable α level, we reject the null hypotheses $H_0 : \beta_1 = 0$ and $H_0 : \beta_2 = 0$.

(d) Now fit a least squares regression to predict `y` using only `x1`. Comment on your results. What are the values of $\hat{\beta}_0$ and $\hat{\beta}_1$? How do these relate to the true values of β_0 and β_1? What is the value of s and how does it relate to the true value of σ^2? Cab you reject the null hypothesis $H_0 : \beta_1 = 0$?

In this sub-part, we will fit a least squares regression to predict `y` using `x1`, and describe the results appearing in the output of the `summary()` function. This is done below.

```
##
## Call:
## lm(formula = y ~ x1)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.48228 -0.97125 -0.03059  0.93666  2.78169
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   6.5235     0.2770  23.548  < 2e-16 ***
## x1            1.9829     0.4758   4.168 6.64e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.267 on 98 degrees of freedom
## Multiple R-squared:  0.1506, Adjusted R-squared:  0.1419
## F-statistic: 17.37 on 1 and 98 DF,  p-value: 6.638e-05
```

As can be seen by the above R about from fitting `y~x1`, we can see that we obtained an F-statistic of 17.37 with a p-value of 6.638×10^{-5}, which means that the model is significant when compared to the empty model. However, both the F-statistic and the p-value of this model are less significant than the previous model that included `x2`.

Furthermore, our estimates of $\hat{\beta}_0$ and $\hat{\beta}_1$, were 6.5235 and 1.9829, respectively. The estimate for $\hat{\beta}_0$ was not close to the true value of β_0 which was 2. However, the estimate for $\hat{\beta}_1$ was close to the true value of β_1 which was 2. It seems as if the intercept is compensating for the missing predictor that should be there.

Also, as computed in R, the value of s is 1.266699. This means that s^2 is 1.604525. These estimates are somewhat close to the true values of σ and σ^2, which are both 1. However, the estimates for these values are farther from the true value than those for the full model.

Lastly, the p-value associated with $\hat{\beta}_1$ was 6.64×10^{-5}. Thus, at any reasonable α level, we reject the null hypothesis $H_0 : \beta_1 = 0$. Due to the fact that the estimate of β_1 was so close to its true value, we still say that the predictor is significant, even though the model is missing an important predictor.

(e) Now fit a least squares regression to predict `y` using only `x2`. Comment on your results. What are the values of $\hat{\beta}_0$ and $\hat{\beta}_2$? How do these relate to the true values of β_0 and β_2? What is the value of s and how does it relate to the true value of σ^2? Cab you reject the null hypothesis $H_0 : \beta_2 = 0$?

In this sub-part, we will fit a least squares regression to predict `y` using `x2`, and describe the results appearing in the output of the `summary()` function. This is done below.

```
##
## Call:
## lm(formula = y ~ x2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.3432 -0.8776 -0.1927  0.7798  2.5804
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.92698    0.62330   4.696 8.64e-06 ***
## x2           0.30467    0.04044   7.534 2.46e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.094 on 98 degrees of freedom
## Multiple R-squared:  0.3668, Adjusted R-squared:  0.3603
## F-statistic: 56.77 on 1 and 98 DF,  p-value: 2.465e-11
```

As can be seen by the above R about from fitting `y~x2`, we can see that we obtained an F-statistic of 56.77 with a p-value of 2.465×10^{-11}, which means that the model is significant when compared to the empty model. However, both the F-statistic and the p-value of this model are less significant than the model including both of the predictors. However, the F-statistic and the p-value of this model are more significant than the previous model including only `x1`.

Furthermore, our estimates of $\hat{\beta}_0$ and $\hat{\beta}_2$, were 2.92698 and 0.30467, respectively. The estimate for $\hat{\beta}_0$ was somewhat close to the true value of β_0 which was 2. However, the estimate for $\hat{\beta}_2$ was close to the true value of β_2 which was 0.3. It seems as if the intercept is compensating for the missing predictor that should be there, but not as much as it was in the previous sub-part.

Also, as computed in R, the value of s is 1.09366. This means that s^2 is 1.196093. These estimates are close to the true values of σ and σ^2, which are both 1. However, the estimates for these values are farther from the true value than those for the model including both predictors, but are closer to the true value than the model including only `x1`.

Lastly, the p-value associated with $\hat{\beta}_2$ was 2.46×10^{-11}. Thus, at any reasonable α level, we reject the null hypothesis $H_1 : \beta_2 = 0$. Due to the fact that the estimate of β_2 was so close to its true value, we still say that the predictor is significant, even though the model is missing an important predictor. Furthermore, due to the observed differences in the models containing only `x1`, and only `x2`, it appears that `x2` is a more significant and influential predictor as opposed to `x1`.

(f) Run the following code to create the vectors x1, x2, and y.

```
> set.seed(1)
> n <- 100
> x1 <- runif(n)
> x2 <- 0.5*x1+rnorm(n,0,0.01)
> y <- 2+2*x1+0.3*x2+rnorm(n)
```

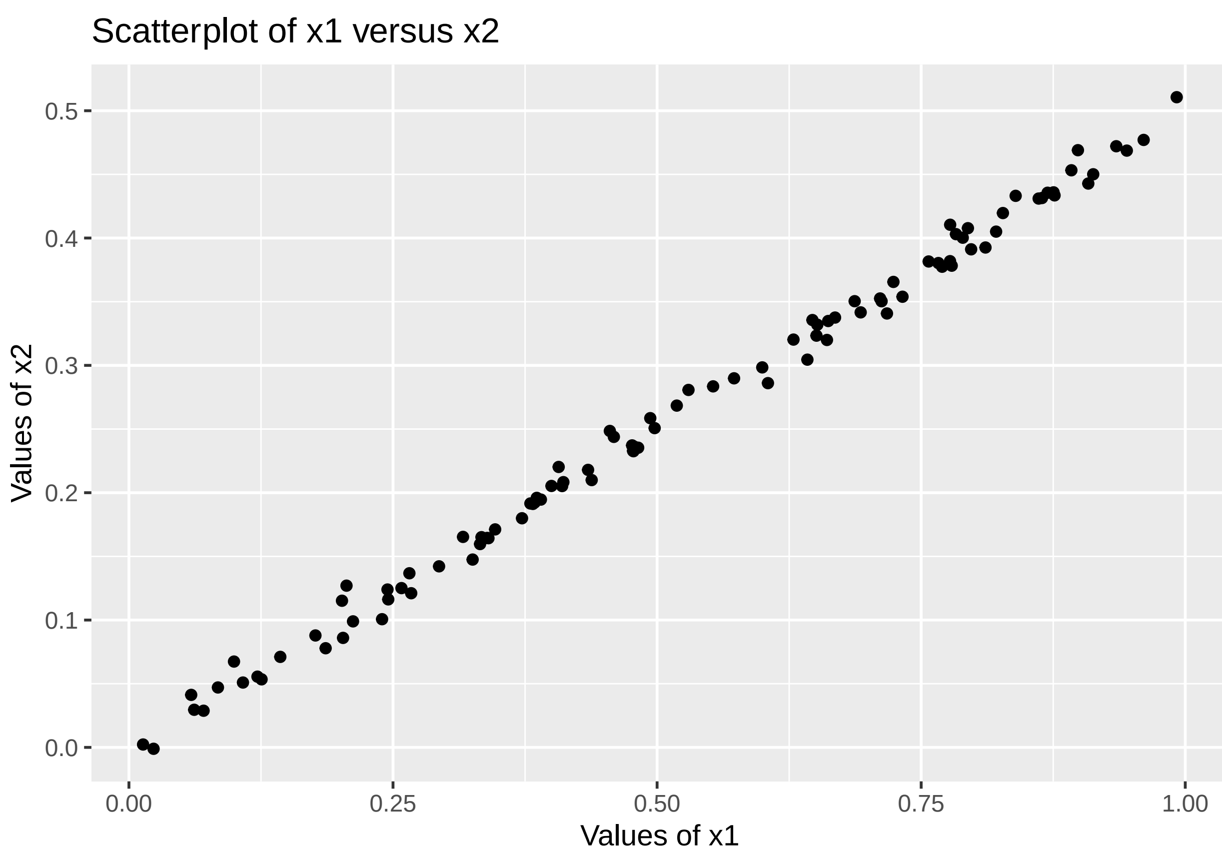
Repeat parts (b), (c), (d), and (e) using the new vectors x1, x2 and y. What differences do you see between? Explain why these differences occur.

In this sub-part, we will repeat parts (b), (c), (d), and (e) using the new vectors x1, x2 and y. After which we will compare the differences and explain why these differences occur.

Repeating part (b):

We will start off by repeating part (b). To do this, we will use the cor() function to calculate the correlation coefficient between x1 and x2, as well as make a scatter plot of these two variables to further analyze the relationship between them, this is done below.

As computed in R, the correlation coefficient between x1 and x2 is 0.9975904, which represents a very strong positive linear relationship, thus we expect to see a very strong linear relationship between the two variables when making a scatter plot of them. We will use ggplot2 to make this scatter plot below.



As can be seen from the above scatter plot of x2 versus x1, we have confirmed that there appears to be a strong positive linear relationship between the two variables. This provides us evidence that there is strong multicollinearity between the two predictor variables, which is not a good thing. This multicollinearity is caused by x2's dependence on x1.

Repeating part (c):

We will now repeat part (c) and fit a least squares regression to predict `y` using `x1` and `x2`, and describe the results appearing in the output of the `summary()` function. This is done below.

```
##
## Call:
## lm(formula = y ~ x1 + x2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.8311 -0.7273 -0.0537  0.6338  2.3359
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   2.1305     0.2319   9.188 7.61e-15 ***
## x1           -1.7540     5.7178  -0.307    0.760
## x2            7.3967    11.3372   0.652    0.516
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.056 on 97 degrees of freedom
## Multiple R-squared:  0.2051, Adjusted R-squared:  0.1887
## F-statistic: 12.51 on 2 and 97 DF,  p-value: 1.465e-05
```

As can be seen by the above R about from fitting `y~x1+x2`, we can see that we obtained an F-statistic of 12.51 with a p-value of 1.465×10^{-5}, which means the model is significant when compared to the empty model.

Furthermore, our estimates of $\hat{\beta}_0$, $\hat{\beta}_1$ and $\hat{\beta}_2$, were 2.1305, -1.7540 and 7.3967, respectively. Other than the intercept, these estimates are very different when compared to the true values of β_0, β_1 and β_2, which were 2, 2 and 0.3, respectively. In particular, $\hat{\beta}_1$ underestimates β_1 by a large amount, and $\hat{\beta}_2$ overestimates β_2 by an even larger amount.

Also, as computed in R, the value of s is 1.0561788. This means that s^2 is 1.115512. These estimates are close to the true values of σ and σ^2, which are both 1.

Lastly, the p-values associated with $\hat{\beta}_1$ and $\hat{\beta}_2$, were 0.760 and 0.516, respectively. Thus, at any reasonable α level, we fail to reject the null hypotheses $H_0 : \beta_1 = 0$ and $H_0 : \beta_2 = 0$.

Repeating part (d):
We will now repeat part (d) and fit a least squares regression to predict `y` using `x1`, and describe the results appearing in the output of the `summary()` function. This is done below.

```
##
## Call:
## lm(formula = y ~ x1)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.87789 -0.68357 -0.07517  0.61429  2.40388
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   2.1172     0.2303   9.193 6.83e-15 ***
```

```
## x1            1.9675     0.3955   4.974 2.79e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.053 on 98 degrees of freedom
## Multiple R-squared:  0.2016, Adjusted R-squared:  0.1934
## F-statistic: 24.74 on 1 and 98 DF,  p-value: 2.795e-06
```

As can be seen by the above R about from fitting `y~x1`, we can see that we obtained an F-statistic of 24.74 with a p-value of 2.795×10^{-6}, which means the model is significant when compared to the empty model. However, both the F-statistic and the p-value of this model are more significant than the previous model including both of the predictors.

Furthermore, our estimates of $\hat{\beta}_0$ and $\hat{\beta}_1$, were 2.1172 and 1.9675, respectively. These estimates were close to the true values of β_0 and β_1, which were 2 and 2, respectively. This model produces more accurate estimates due to the lack of multicollinearity now that `x2` is gone.

Also, as computed in R, the value of s is 1.053078. This means that s^2 is 1.108974. These estimates are close to the true values of σ and σ^2, which are both 1. These values are closer to the true value than the model with both `x1` and `x2`.

Lastly, the p-value associated with $\hat{\beta}_1$ was 2.79×10^{-6}. Thus, at any reasonable α level, we reject the null hypothesis $H_0 : \beta_1 = 0$.

Repeating part (e):
We will now repeat part (e) and fit a least squares regression to predict `y` using `x2`, and describe the results appearing in the output of the `summary()` function. This is done below.

```
##
## Call:
## lm(formula = y ~ x2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.85470 -0.68465 -0.06898  0.60983  2.34499
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   2.1199     0.2282   9.288 4.24e-15 ***
## x2            3.9273     0.7829   5.016 2.35e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.051 on 98 degrees of freedom
## Multiple R-squared:  0.2043, Adjusted R-squared:  0.1962
## F-statistic: 25.16 on 1 and 98 DF,  p-value: 2.35e-06
```

As can be seen by the above R about from fitting `y~x2`, we can see that we obtained an F-statistic of 25.16 with a p-value of 2.35×10^{-06}, which means the model is significant when compared to the empty model. However, both the F-statistic and the p-value of this model are less significant than the previous model including only `x1`. However, the F-statistic and the p-value of this model are more significant than the model including both predictors.

Furthermore, our estimates of $\hat{\beta}_0$ and $\hat{\beta}_2$, were 2.1199 and 3.9273, respectively. The estimate for $\hat{\beta}_0$ was close to the true value of β_0 which was 2. However, the estimate for $\hat{\beta}_2$ was not that close to the true value of β_2

which was 0.3. It is important to note that, although this estimate of β_2 isn't great, it is still better than the estimate of β_2 in the full model.

Also, as computed in R, the value of s is 1.051285. This means that s^2 is 1.051285. These estimates are close to the true values of σ and σ^2, which are both 1. These estimates for s and s^2 are very comparable to the model with only `x1` (but slightly closer to the true value), and are closer to the true value than the model with both `x1` and `x2`.

Lastly, the p-value associated with $\hat{\beta}_2$ was 2.25×10^{-6}. Thus, at any reasonable α level, we reject the null hypothesis $H_1 : \beta_2 = 0$.

Conclusions:
As we can see from the output of repeating parts (b)-(e) with the new sample, this problem shows us that multicollinearity can lead to very inaccurate estimates of regression parameters. Since `x1` and `x2` were highly correlated, the standard error of their parameter estimates were very high, which led to small t-statistics and large p-values. This problem was remedied when only `x1` or `x2` were included, though the parameter estimates for the `x1` only model were more accurate than the `x2` only model. In comparison, the original parts (b)-(e) saw estimates that more closely resembled the true parameter values (especially in the full model case), due to the fact that no multicollinearity was present.

(g) Use `x1`, `x2` and `y` from Part (f) and suppose that we obtain one additional observation, which was unfortunately mismeasured.

```
> x1 <- c(x1, 0.1)
> x2 <- c(x2, 0.8)
> y <- c(y, 6)
```

Re-fit the linear models from parts (c), (d) and (e) using this new data. What effect does this new observation have on each of the models? In each model, is this observation an outlier? A high-leverage point? Both? Explain your answers.

As mentioned in Canvas messages and during lecture, we are skipping this problem due to the fact that we do not have the proper definitions for leverage points and outliers yet.