Yahya Alhinai

EE5373

October 15th, 2019

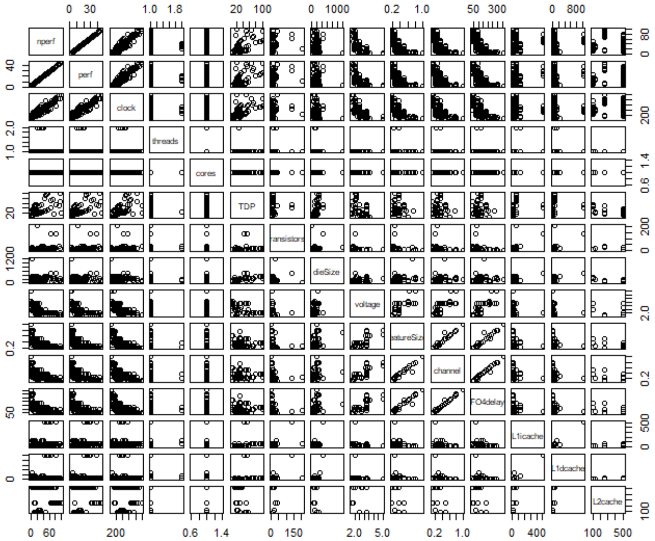
Multi-factor Regression Models

OBJECTIVE FOR THIS LAB:

- Generate a multi-factor regression model for the following benchmark programs in CPU DB: **int95**, **int06**, **fp95**, **fp06**.

INT1995:

1. **Pair-wise comparisons plot:**
   This plot does not include L3 column because the face it was not implemented in the processor.



2. **List of potential predictors with explanation of why those predictors are included:**

   List of predictors that are insufficient to have an impact on the model or the data frame are not specified:

   L3caches: the data frame values were not specified due to the face that L3 was not a implemented back then.

   Threads & core: those two sets of data has a single value for each. Meaning, it will not have an impact on the model.

   TDP: This value is provided by the chip manufacturer to indicate the expected chip's thermal. Thus, the value does not directly affect the performance.

   The rest of predictors have direct correlation to the performance which impact the model:

Clock, Transistors, DieSize, Voltage, FeatureSize, Channel, FO4delay, and L1/L2caches

3. **List of predictors eliminated during backward elimination process, with explanations:**
We started having all the predictors selected be evaluated for the regression model. We obtained the following:

$$R^2: 0.9985 \;||\; Adjusted - R^2: 0.9978$$

FO4delay: seems to have low number of values variation where it has only 23 different values:

$$R^2: 0.9985 \;||\; Adjusted - R^2: 0.9978$$

There are a lot of missing value in **dieSize** (35% of values are NAs). The following R-value after removing both:

$$R^2: 0.9985 \;||\; Adjusted - R^2: 0.9978$$

There are a lot of missing value in **voltage** (39% of voltage values are NAs). The following R-value after removing both:

$$R^2: 0.9985 \;||\; Adjusted - R^2: 0.9978$$

L1icache and **sqrt(L1icache)** have several values missing (L1i has 10% NA values) as well as the p-value is larger the desired value which is 0.05 and 0.2 for L1icache and sqrt(L1icache) respectively:

$$R^2: 0.9985 \;||\; Adjusted - R^2: 0.9978$$

L1dcache and **sqrt(L1dcache)** have several values missing (L1i has 12% NA values) as well as the p-value is larger the desired value which is 0.055 and 0.21 for L1dcache and sqrt(L1dcache) respectively:

$$R^2: 0.9889 \;||\; Adjusted - R^2: 0.9853$$

sqrt(L2cache) and L2cache are fundamentally the same. **sqrt(L2cache)** is to be removed and the following results are obtained:

$$R^2: 0.9869 || \; Adjusted - R^2: 0.9836$$

The **featureSize** predictor has a p-value that is 0.42 which is out of the desirable accuracy:

$$R^2: 0.9864 \;||\; Adjusted - R^2: 0.9846$$

Channel is missing 22% of its value; therefore, it’s to be removed:

$$R^2: 0.9864 \; || \; Adjusted - R^2: 0.9846$$

4. **Final model with each predictor and coefficient values:**

```
Residuals:
    Min      1Q  Median      3Q     Max 
-5.4675 -1.1531  0.2401  0.8489  7.3726 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept) -20.705076   1.399265 -14.797 3.04e-13 ***
clock         0.072889   0.003748  19.445 8.92e-16 ***
transistors   0.998533   0.101900   9.799 1.12e-09 ***
L2cache       0.028204   0.003116   9.052 4.83e-09 ***
```

$\boldsymbol{clock}$	$\mathbf{0.0728}$
$\boldsymbol{transistors}$	0.9985
$\boldsymbol{L2cache}$	0.0282

$$nperf = -20.70 + 0.0728 * clock + 0.9985 * transistors + 0.0282 * L2cache$$

5. **Explanation of quality analysis:**

a. **P-values:**

i. The predictors that were chosen at the end are relevant. The predetermined threshold is less than 1% which indicated a high relevance between the actual value and the model.

ii. The p-value of the coefficient is tiny (almost zero). Which gives a percentage of almost 100% chance that this specific is relevant to the model.

b. **Residual standard error:**

i. Ideally, residual standard error * 1.5 = 2.09 is the distance of Q1 and Q3 from the mean

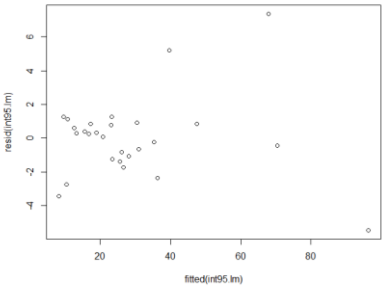
ii. Residual standard error does not follow Gaussian distribution since Q1 and Q3 are not 2.09 apart from the mean

c. $\boldsymbol{R^2}$ **values:**

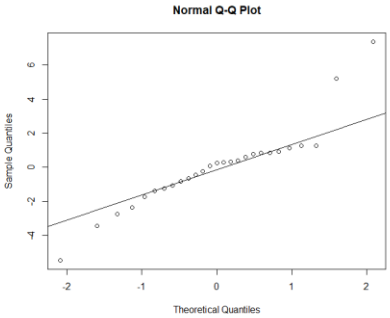
i. Values of R^2 that are closer to one indicate a better-fitting model. Our model has a value of 0.9864 which means that the model explains 98.64% of the data’s variation. R^2 for this model is pretty good even though it is described by only 3 predictors.

d. **residual analysis**

i. The following figure shows that residuals appear to be scattered around 0. This is a good indication that the model is highly representative of the actual values.

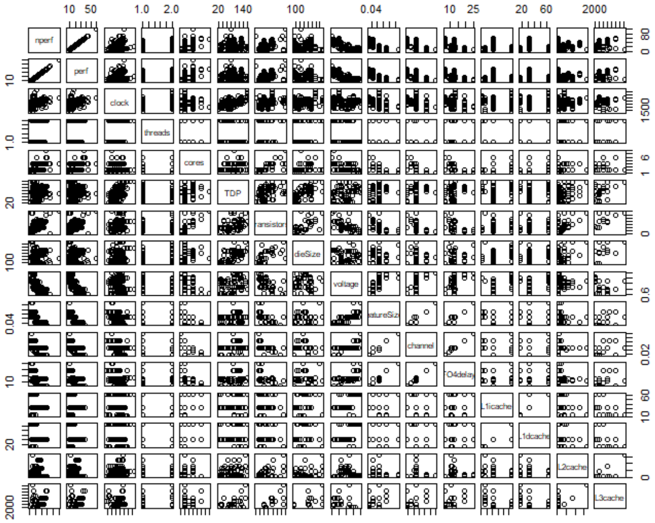


ii. As the figure below shows the theoretical vs actual samples, we see the that residuals roughly follow the indicated line. There is nonlinearity in the plot which leads to conclude that residuals are not perfectly normally distributed. Overall, although the model is not perfect, it describes data well.



INT2006:

1. **Pair-wise comparisons plot:**



2. **List of potential predictors with explanation of why those predictors are included:**

List of predictors that are insufficient to have an impact on the model or the data frame are not specified:

Threads & core: those two sets of data have max of only 4 values to represents. Meaning, it will not have an impact on the model.

L3caches: this predictor has a great deal of unspecified values (NAs); therefore, it's going to have an insufficient remake to our model.

The rest of predictors have direct correlation to the performance which impact the model:

Clock, Transistors, TDP, DieSize, Voltage, FeatureSize, Channel, FO4delay, and L1/L2caches

3. **List of predictors eliminated during backward elimination process, with explanations:**
We started having all the predictors selected be evaluated for the regression model. We obtained the following:

$$R^2\text{: } 0.9612 \text{ || } Adjusted - R^2\text{: } 0.9577$$

dieSize has the largest value of p-value which is 0.96. This indicate its low relevance to the model.

$$R^2\text{: } 0.9612 \text{ || } Adjusted - R^2\text{: } 0.958$$

voltage has a p-value of 0.29. This indicate its low relevance to the model.

$$R^2\text{: } 0.9604 \text{ || } Adjusted - R^2\text{: } 0.9575$$

FO4delay has the largest value of p-value which is 0.04. This indicate its low relevance to the model.

$$R^2\text{: } 0.9593 \text{ || } Adjusted - R^2\text{: } 0.9566$$

TDP has the largest value of p-value which is 0.016. This indicate its low relevance to the model.

$$R^2\text{: } 0.9494 \text{ || } Adjusted - R^2\text{: } 0.9464$$

L2cache and sqrt(L2cache) are fundamentally the same. **L2cache** is to be removed and the following results are obtained:

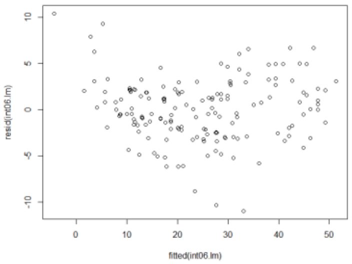
$$R^2\text{: } 0.932 \text{ || } Adjusted - R^2\text{: } 0.9285$$

4. **Final model with each predictor and coefficient values:**

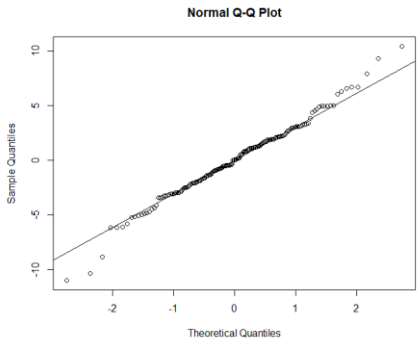
```
Residuals:
     Min       1Q   Median       3Q      Max
-10.9811  -2.1105   0.0113   2.0294  10.3902

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)    -2.380e+02  1.807e+01 -13.171  < 2e-16 ***
clock           1.078e-02  6.588e-04  16.368  < 2e-16 ***
transistors     1.514e-02  1.527e-03   9.913  < 2e-16 ***
```

```
featureSize    -3.359e+02  4.174e+01  -8.046 2.09e-13 ***
channel         5.228e+02  6.514e+01   8.026 2.34e-13 ***
L1icache       -1.280e+01  1.274e+00 -10.043  < 2e-16 ***
sqrt(L1icache)  7.739e+01  5.230e+00  14.797  < 2e-16 ***
L1dcache        6.608e+00  9.113e-01   7.251 1.84e-11 ***
sqrt(L2cache)  -1.743e-01  1.056e-02 -16.500  < 2e-16 ***
Residual standard error: 3.472 on 155 degrees of freedom
```

5. **Explanation of quality analysis:**
   - **a. P-values:**
     - i. This value indicates the probability that the intercept is not relevant. The predictors that were chosen at the end are relevant. The predetermined threshold is less than 1% which indicated a high relevance between the actual value and the model.
     - ii. The p-value of the coefficient is tiny for all predictors. Which gives a percentage of almost 100% chance that this specific is relevant to the model.
   - **b. Residual standard error:**
     - i. Ideally, residual standard error * 1.5 = 5.2 is the difference distance of Q1 and Q3
     - ii. Residual standard error seems to follow Gaussian distribution since Q1 and Q3 apart from each other. As well as it almost symmetrical with mean close to zero. The distribution is pretty good.
   - **c. R^2 values:**
     - i. Our model has a value of 0.932 which means that the model explains 93.20% of the data's variation. R^2 for this model is good since any other removal of predictors will make this value drops significantly.
   - **d. residual analysis**
     - i. The following figure shows that residuals appear to be scattered around 0. This is a good indication that the model is highly representative of the actual values.

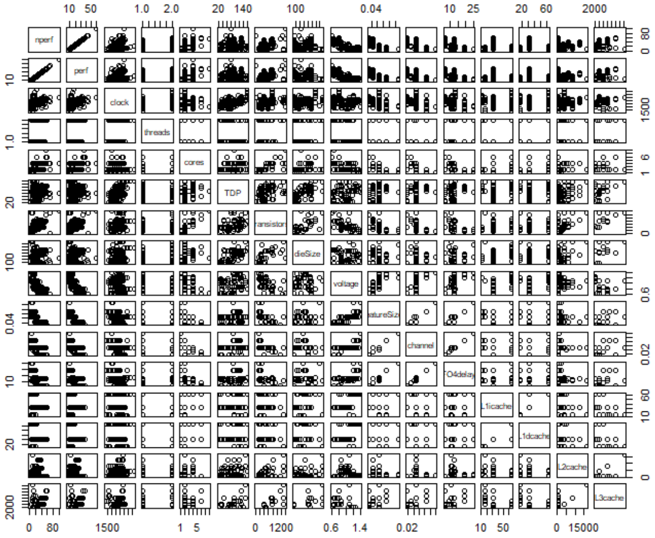


     - ii. As the figure below shows the theoretical vs actual samples, we see the that residuals roughly follow the indicated line. There is nonlinearity in the plot, especially at the beginning and in the end, which leads to conclude that residuals are not perfectly normally distributed, but it's a really good representation of the actual data.



FP1995:

1. **Pair-wise comparisons plot:**



2. **List of potential predictors with explanation of why those predictors are included:**

List of predictors that are insufficient to have an impact on the model or the data frame are not specified:

<u>L3caches</u>: the data frame values were not specified due to the face that L3 was not a implemented back then.

<u>Threads & core:</u> those two sets of data has a single value for each. Meaning, it will not have an impact on the model.

<u>TDP:</u> This value is provided by the chip manufacturer to indicate the expected chip's thermal. Thus, the value does not directly affect the performance.

<u>Channel:</u> there are not enough variations of value to impact the model represented (only 4 different variations value) and therefore it will not impact the performance that much.

The rest of predictors have direct correlation to the performance which impact the model:

Clock, Transistors, DieSize, Voltage, FeatureSize, FO4delay, and L1/L2caches

3. **List of predictors eliminated during backward elimination process, with explanations:** We started having all the predictors selected be evaluated for the regression model. We obtained the following:

$$R^2: 0.9712 \,||\, Adjusted - R^2: 0.9567$$

sqrt(L2cache) and L2cache are fundamentally the same. **sqrt(L2cache)** is to be removed and the following results are obtained:

$$R^2: 0.9712 \,||\, Adjusted - R^2: 0.9567$$

voltage has the largest value of p-value which is 0.20. This indicate its low relevance to the model.

$$R^2: 0.9533 \,||\, Adjusted - R^2: 0.9326$$

FO4delay has a value of p-value 0.24. This indicate its low relevance to the model.

$$R^2: 0.9593 \,||\, Adjusted - R^2: 0.9309$$

The **featureSize** predictor has a p-value that is 0.012 which is out of the desirable accuracy:

$$R^2: 0.9287 \,||\, Adjusted - R^2: 0.9074$$

dieSize is the next to remove because of their irrelevant correlation described by p-value which has exceeded 0.15:

$$R^2: 0.921 || \, Adjusted - R^2: 0.9021$$

sqrt(L2cache) and L2cache are fundamentally the same. **sqrt(L2cache)** is to be removed and the following results are obtained:

$$R^2: 0.921 || \, Adjusted - R^2: 0.9021$$

sqrt(L1icache) and L1icache are fundamentally the same as well as the large p-value which is 0.32. **sqrt(L1icache)** is to be removed and the following results are obtained:

$$R^2: 0.921 || \, Adjusted - R^2: 0.9021$$

sqrt(L1dcache) p-value is given NAs which mean it has no correlation to our model. **sqrt(L1icache)** is to be removed and the following results are obtained:

$$R^2: 0.9152|| \ Adjusted - R^2: 0.8997$$

Transistor seems to have the largest p-value at 0.78; therefore, it's to be removed and the following results are obtained:

$$R^2: 0.9408 \ || \ Adjusted - R^2: 0.9375$$

L1icache a very large p-value at 0.77. This predictor is to be removed.

$$R^2: 0.9408 \ || \ Adjusted - R^2: 0.9375$$

L1dcache a very large p-value at 0.77. This predictor is to be removed.

$$R^2: 0.9407 \ || \ Adjusted - R^2: 0.9385$$

4. **Final model with each predictor and coefficient values:**

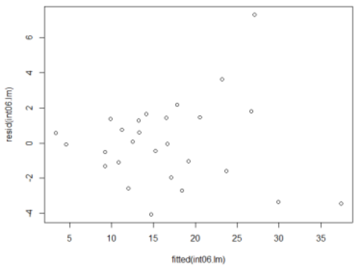
```
Residuals:
    Min      1Q  Median      3Q     Max 
-5.7196 -1.6937  0.0476  1.4941  6.0078 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) -1.553030   0.954646  -1.627  0.10949    
clock        0.042472   0.001450  29.290  < 2e-16 ***
L2cache     -0.004753   0.001772  -2.682  0.00964 ** 
Residual standard error: 2.51 on 55 degrees of freedom
```

5. **Explanation of quality analysis:**
   a. **P-values:**
      i. The predictors that were chosen at the end are relevant. The predetermined threshold is less than 1% which indicated a high relevance between the actual value and the model.
      ii. The p-value of the coefficient is tiny for all predictors. Which gives a percentage of more than 99% chance that this specific is relevant to the model.

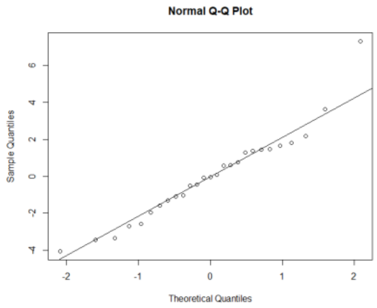
   b. **Residual standard error:**
      i. Ideally, residual standard error * 1.5 = 3.765 is the difference distance of Q1 and Q3
      ii. Residual standard error deviates slightly from following Gaussian distribution since Q1 and Q3 are less than 3.76 from each other. As well as it is not perfectly symmetrical, but the mean is close to zero. The distribution does not represents Gaussian distributed.

   c. **R^2 values:**
      i. Our model has a value of 0.9407 which means that the model explains 94.07% of the data's variation. R^2 for this model is good and it's represented by only two predictors.

   d. **residual analysis**
      i. The following figure shows that residuals appear to be scattered around 0. This is a good indication that the model is highly representative of the actual values.

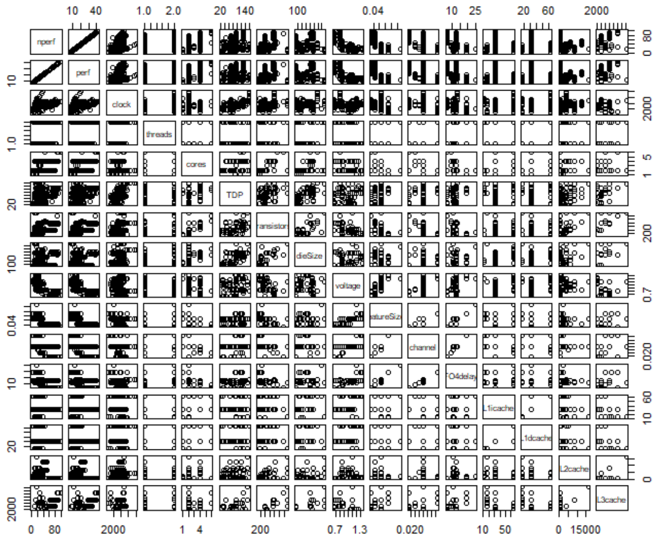


ii. As the figure below shows the theoretical vs actual samples, we see the that residuals roughly follow the indicated line. There is nonlinearity in the plot, especially at the beginning and in the end. This leads to conclude that residuals are not perfectly normally distributed, but it's a really good representation of the actual data.



FP2006:

1. **Pair-wise comparisons plot:**



2. **List of potential predictors with explanation of why those predictors are included:**

List of predictors that are insufficient to have an impact on the model or the data frame are not specified:

Threads & core: those two sets of data have max of only 4 values to represents. Meaning, it will not have an impact on the model.

L3caches: this predictor has a great deal of unspecified values (NAs); therefore, it's going to have an insufficient remake to our model.

TDP: This value is provided by the chip manufacturer to indicate the expected chip's thermal. Thus, the value does not directly affect the performance.

Channel: there are not enough variations, only three different values (only 4 different values), to impact the model represented and therefore it will not impact the performance that much.

The rest of predictors have direct correlation to the performance which impact the model:

Clock, Transistors, DieSize, Voltage, FeatureSize, FO4delay, and L2caches

3. **List of predictors eliminated during backward elimination process, with explanations:**
   We started having all the predictors selected be evaluated for the regression model. We obtained the following:

$$R^2: 0.9577 \;||\; Adjusted - R^2: 0.9542$$

voltage has the largest value of p-value which is 0.1. This indicate its low relevance to the model.

$$R^2: 0.955 \;||\; Adjusted - R^2: 0.9517$$

dieSize has the largest value of p-value which is 0.96. This indicate its low relevance to the model.

$$R^2: 0.9401 \;||\; Adjusted - R^2: 0.9362$$

sqrt(L2cache) and L2cache are fundamentally the same. **sqrt(L2cache)** is to be removed and the following results are obtained:

$$R^2: 0.9151|| \; Adjusted - R^2: 0.9101$$

FO4delay is next to remove because of its irrelevant correlation and because R^2 and $Adjusted - R^2$ increases: `

$$R^2: 0.9217 \;||\; Adjusted - R^2: 0.9171$$

featureSize is the next to remove because of its low value variations which has only 4 different values. This means that **featureSize** does not heavily effecting our model:

R^2: 0.9199 || $Adjusted - R^2$: 0.9158 **featureSize**

4. **Final model with each predictor and coefficient values:**

```
Residuals:
    Min      1Q  Median      3Q     Max
-64.226  -7.338   1.412   9.547  21.930

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)     84.625707  10.125587   8.358 3.31e-14 ***
clock            0.014011   0.002246   6.240 3.95e-09 ***
transistors      0.008095   0.005217   1.552 0.122771
sqrt(L1icache)  13.510344   5.193625   2.601 0.010180 *
sqrt(L1dcache) -20.903846   5.499547  -3.801 0.000206 ***
L2cache          0.010461   0.001451   7.210 2.26e-11 ***
sqrt(L2cache)   -1.602962   0.165333  -9.695  < 2e-16 ***

Residual standard error: 13.62 on 156 degrees of freedom
```

5. **Explanation of quality analysis:**

   a. **P-values:**

      i. The predictors that were chosen at the end are relevant beside transistors which has a high irrelevant to the model. The predetermined threshold is less than 1% for most of predictors beside transistors section which has indicated a high irrelevant at 12%.

      ii. Beside transistors, The p-value of the coefficient is tiny. Which indicates the relevance of the model to the data plotted.

   b. **Residual standard error:**

      i. Ideally, residual standard error * 1.5 = 20.44 is the difference distance of Q1 and Q3

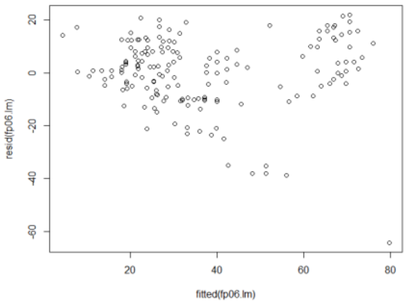
      ii. Residual standard error deviates from following Gaussian distribution since Q1 and Q3 are less than 20.44 from each other. As well as it is not perfectly symmetrical, and the mean is not centered at zero. The distribution does not represent Gaussian distributed.

   c. **R^2 values:**

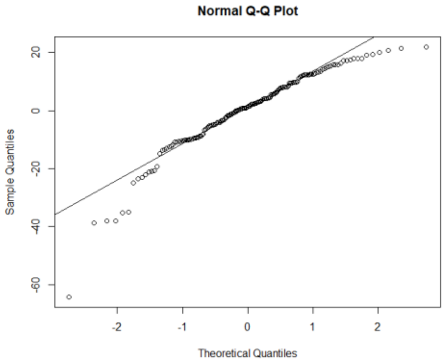
      i. Our model has a value of 0.9082 which means that the model explains 90.82% of the data's variation. R^2 for this model is good representation since there are no predictors left to drop without significantly effecting the value of R^2.

   d. **residual analysis**

      i. The following figure shows that residuals appear to be scattered around 0. This is a good indication that the model is highly representative of the actual values.



ii. As the figure below shows the theoretical vs actual samples, we see the that residuals roughly follow the indicated line. There is nonlinearity in the plot, especially at the beginning and in the end. This leads to conclude that residuals are not perfectly normally distributed. However, The model is a really good representation of the actual data.


Normal Q-Q Plot
Sample Quantiles
Theoretical Quantiles