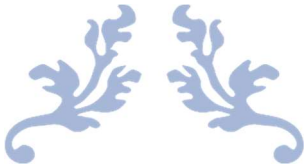


MATH 664 HOMEWORK 2

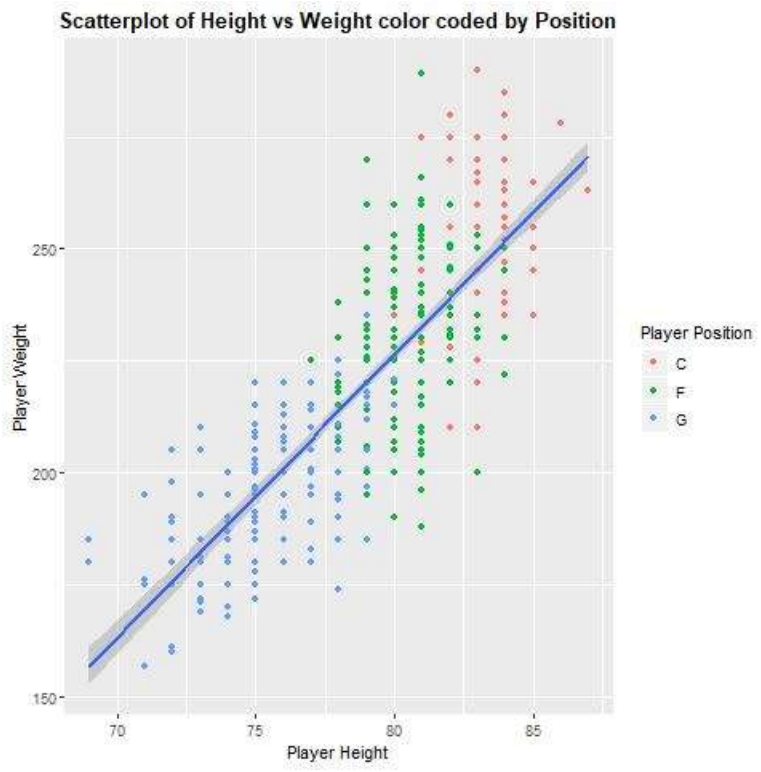
METHODS OF STATISTICAL CONSULTING

PRITHWISH GANGULY

Question 1

This dataset contains 505 observations of Players. The Player data comprises of Player Name, Player Pos, Height, Weight, Age, BMI.

Relationship between Height, Weight & Position



The plot above shows a clear positive linear relationship between player Height and Weight, as seen by the regression line. The strength of the relationship can be verified by the correlation matrix below.

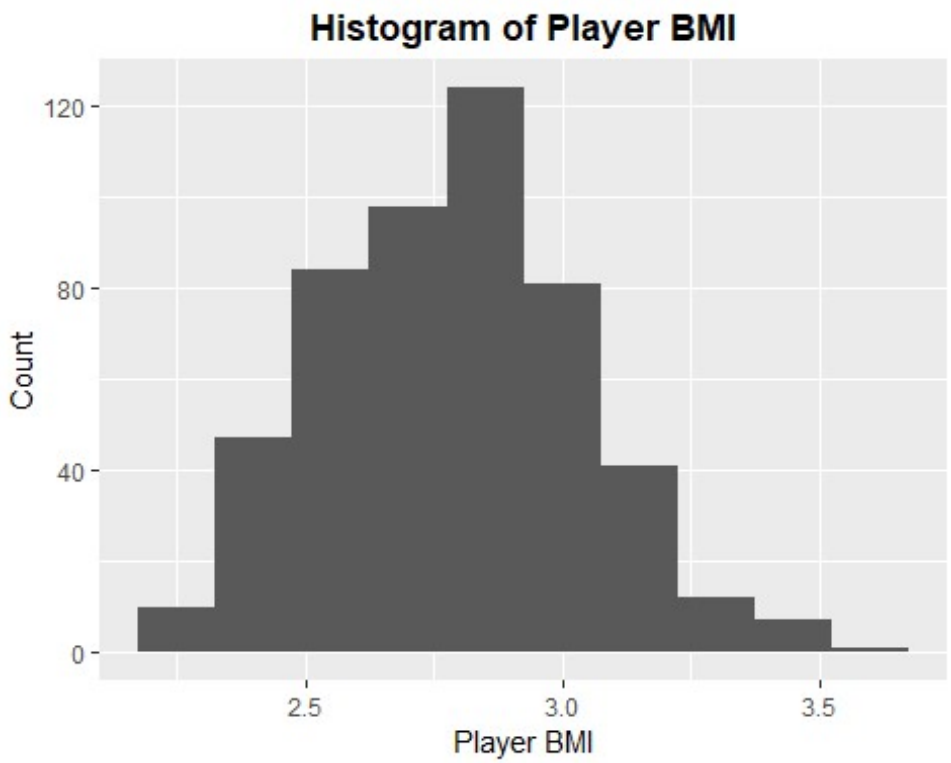
```
##             Height    Weight         Age
## Height 1.000000000 0.8207192 0.005010247
## Weight 0.820719235 1.0000000 0.109814367
## Age    0.005010247 0.1098144 1.000000000
```

Correlation of 0.82 exhibits a very strong linear relationship between the 2 features.

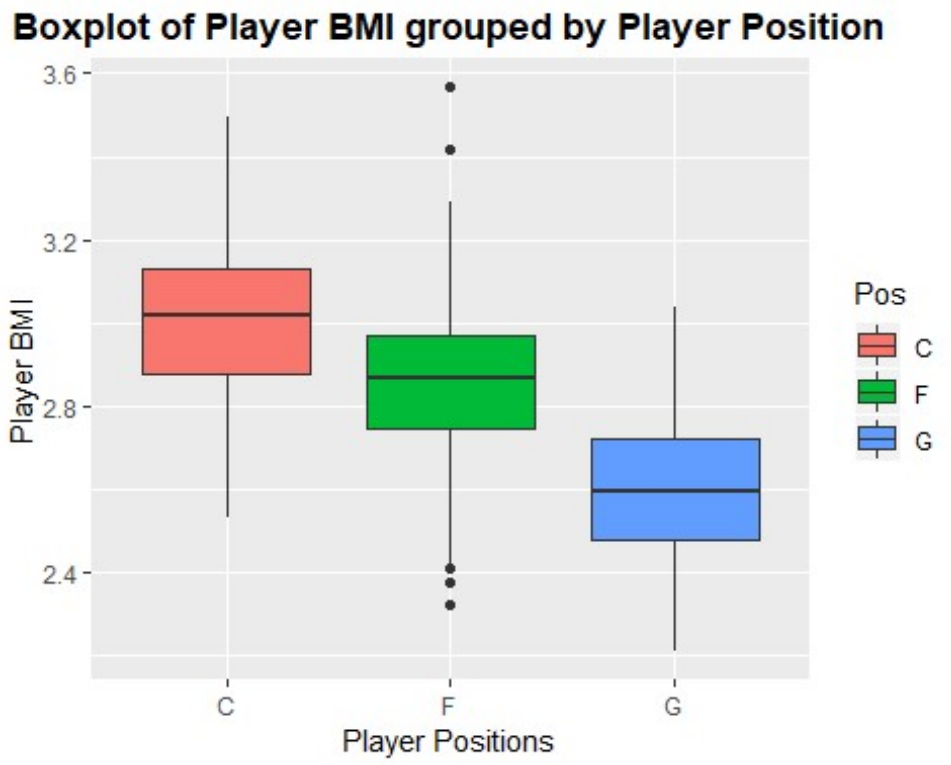
We also see a clear distribution of Player Position across scatterplot. Position G seems to be populated mainly by people in the relatively lower Height and Weight range , whereas Position F is populated by the medium Height and Weight range and Position C is mainly populated by the higher Height and Weight range.

Maybe BMI (Weight/Height) would be a better metric for our analysis?

Player BMI Analysis

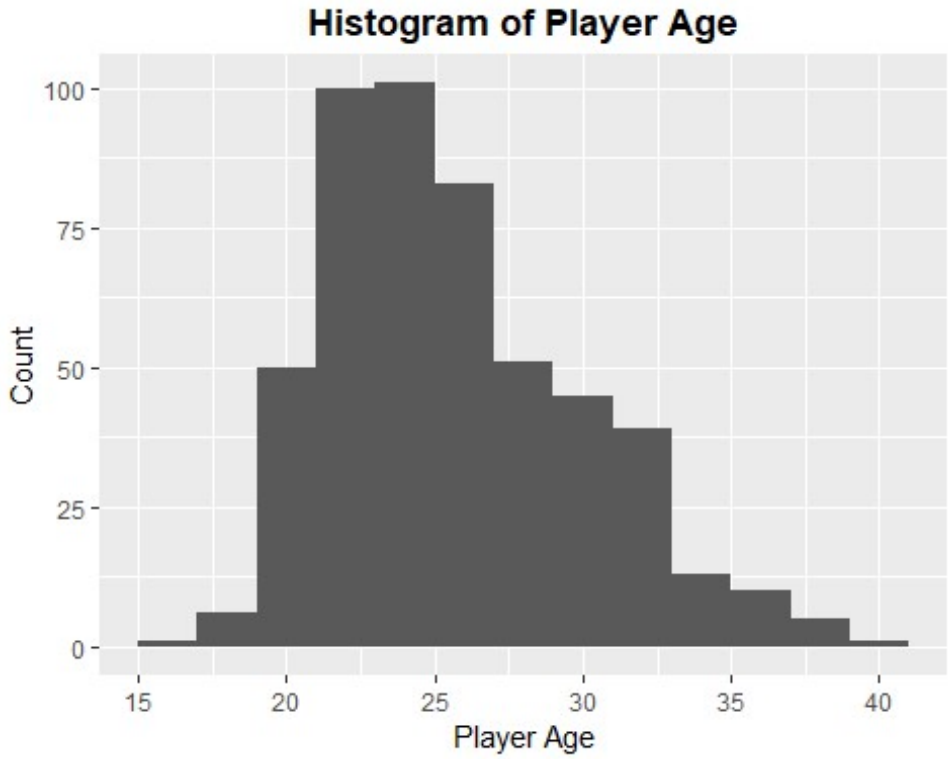


BMI has an Uniform distribution with a mean of 2.784 and a variance of 0.06107



As in our earlier plot, players in position C have the highest BMI, followed by position F and finally position G has the lowest BMI.

Player Age Analysis



The histogram of Player Age exhibits a slight right-skewed plot. Despite there being a lot of players of age 24, the mean gets pulled up due to there being relatively a few older players.

Age Summary Data:

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   15.00   23.00   25.00   26.19   29.00   40.00

## [1] "Mode =  24"
```

The mean, median and mode of Age exhibit the same behaviour since, Mode < Median < Mean, which is natural for right-skewed data.

An anomaly we can observe from the Age Summary Data is the minimum Age of a Player. The youngest player is a teenager of 15 years, 10 years younger than the average player and 25 years younger than the oldest player.

Features Summary

```
##  Pos         Height          Weight           Age              BMI
##  C: 92   Min.   :69.00   Min.   :157.0   Min.   :15.00   Min.   :2.211
##  F:211   1st Qu.:76.00   1st Qu.:200.0   1st Qu.:23.00   1st Qu.:2.597
```

```
##  G:202   Median :80.00   Median :220.0   Median :25.00   Median :2.785
##          Mean   :79.07   Mean   :220.7   Mean   :26.19   Mean   :2.784
##          3rd Qu.:82.00   3rd Qu.:240.0   3rd Qu.:29.00   3rd Qu.:2.938
##          Max.   :87.00   Max.   :290.0   Max.   :40.00   Max.   :3.568

##             Player Pos Height Weight Age
## 326 Jarvis Varnado   F     81    230  15
```

The 15 year old does seem like a true anomaly, despite being the youngest his weight is higher than the median Weight and his height is higher than the median height.

Question 2

This dataset contains 18 observations. Mortgage Yield is the quantitative response and the rest are predictors. Below is a preview of the dataset we will now try to model.

```
##                   SMSA Mortgage.Yield Loan Distance Savings.unit
## 1 Los Angeles-Long Bea           6.17 78.1     3042         91.3
## 2               Denver           6.06 77.0     1997         84.1
## 3 San Francisco-Oaklan           6.04 75.7     3162        129.3
## 4    Dallas-Fort Worth           6.04 77.4     1821         41.2
## 5                Miami           6.02 77.4     1542        119.1
## 6              Atlanta           6.02 73.6     1074         32.3
##   Savings.capita Pop.inc Banks.Mortgage
## 1         1738.1    45.5           33.1
## 2         1110.4    51.8           21.9
## 3         1738.1    24.0           46.0
## 4          778.4    45.7           51.3
## 5         1136.7    88.9           18.7
## 6          582.9    39.9           26.6
```

To model and analyze this dataset we must essentially ask ourselves 4 questions:

- Is atleast one of the predictors useful in predicting the response?
- Do all the predictors help to explain the response, or is only a subset of predictors useful?
- How well does the model fit the data?
- Given a set of predictor values, what response value should we predict, and how accurate is our predictors?

Is atleast one of the predictors useful in predicting the response?

```
##
## Call:
## lm(formula = Mortgage.Yield ~ ., data = df2[-1])
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.145030 -0.017814  0.001474  0.034316  0.134565
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
```

```
## (Intercept)     4.285e+00  6.682e-01   6.413 4.99e-05 ***
## Loan            2.033e-02  9.308e-03   2.184   0.0515 .
## Distance        1.359e-05  4.692e-05   0.290   0.7775
## Savings.unit   -1.584e-03  7.532e-04  -2.103   0.0593 .
## Savings.capita  2.017e-04  1.124e-04   1.794   0.1002
## Pop.inc         1.283e-03  1.765e-03   0.727   0.4826
## Banks.Mortgage  2.357e-04  2.302e-03   0.102   0.9203
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.09991 on 11 degrees of freedom
## Multiple R-squared:  0.8706, Adjusted R-squared:    0.8
## F-statistic: 12.33 on 6 and 11 DF,  p-value: 0.0002523
```

Since our F-statistic > 1, we can be sure atleast 1 of the predictor is useful in predicting the response.

Do all the predictors help to explain the response, or is only a subset required?

Here, I have decided to go with Mixed selection method to conduct Variable selection.

We start off by trying out all the variables individually and pick the one with the best model fit.

```
##
## Call:
## lm(formula = Mortgage.Yield ~ Loan, data = df2[-1])
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.29757 -0.05335  0.01026  0.09776  0.21019
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.37736    0.63115   3.767  0.00169 **
## Loan         0.04720    0.00859   5.495 4.89e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1355 on 16 degrees of freedom
## Multiple R-squared:  0.6536, Adjusted R-squared:  0.632
## F-statistic:  30.2 on 1 and 16 DF,  p-value: 4.892e-05
```

Loan gives the best fit when testing with individual variables. It had the highest adjusted R^2 and the lowest RSE.

Now, we need to add variables and monitor the p-value. We will drop any variable with a p-value > 0.4.

```
##
## Call:
## lm(formula = Mortgage.Yield ~ Loan + Distance, data = df2[-1])
```

```
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.233685 -0.057813  0.004149  0.031559  0.202924
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 3.216e+00  5.742e-01   5.601 5.05e-05 ***
## Loan        3.389e-02  8.116e-03   4.175 0.000812 ***
## Distance    9.947e-05  3.186e-05   3.122 0.006991 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.109 on 15 degrees of freedom
## Multiple R-squared:  0.7901, Adjusted R-squared:  0.7621
## F-statistic: 28.23 on 2 and 15 DF,  p-value: 8.23e-06
```

The addition of Distance gives us a very good improvement in R^2 and RSE.

```
##
## Call:
## lm(formula = Mortgage.Yield ~ Loan + Distance + Savings.unit,
##     data = df2[-1])
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.16081 -0.04076 -0.01115  0.03487  0.17548
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)   3.620e+00  5.941e-01   6.093 2.78e-05 ***
## Loan          3.003e-02  8.011e-03   3.749  0.00216 **
## Distance      7.028e-05  3.482e-05   2.019  0.06311 .
## Savings.unit -5.028e-04  3.010e-04  -1.671  0.11700
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.103 on 14 degrees of freedom
## Multiple R-squared:  0.825,  Adjusted R-squared:  0.7875
## F-statistic:    22 on 3 and 14 DF,  p-value: 1.455e-05
```

Savings.unit doesn't vary our fitness metrics but it is within the p-value bounds, so we can keep it.

```
##
## Call:
## lm(formula = Mortgage.Yield ~ Loan + Distance + Savings.unit +
##     Savings.capita, data = df2[-1])
##
## Residuals:
```

```
##       Min        1Q    Median        3Q       Max
## -0.162840 -0.021711  0.000036  0.038421  0.140622
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)     4.184e+00  6.158e-01   6.794 1.27e-05 ***
## Loan            2.260e-02  8.262e-03   2.736   0.0170 *
## Distance        1.273e-05  4.349e-05   0.293   0.7744
## Savings.unit   -1.710e-03  6.800e-04  -2.514   0.0259 *
## Savings.capita  2.039e-04  1.051e-04   1.940   0.0743 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.09412 on 13 degrees of freedom
## Multiple R-squared:  0.8643, Adjusted R-squared:  0.8225
## F-statistic:  20.7 on 4 and 13 DF,  p-value: 1.524e-05
```

Distance has now crossed the p-value bounds, so we drop it.

```
##
## Call:
## lm(formula = Mortgage.Yield ~ Loan + Savings.unit + Savings.capita +
##     Pop.inc, data = df2[-1])
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.15689 -0.01822  0.00736  0.03907  0.14017
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)     4.326e+00  6.052e-01   7.148  7.5e-06 ***
## Loan            2.010e-02  8.532e-03   2.356  0.03486 *
## Savings.unit   -1.775e-03  4.391e-04  -4.043  0.00139 **
## Savings.capita  2.268e-04  7.547e-05   3.004  0.01015 *
## Pop.inc         1.254e-03  1.630e-03   0.769  0.45549
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.09235 on 13 degrees of freedom
## Multiple R-squared:  0.8693, Adjusted R-squared:  0.8291
## F-statistic: 21.62 on 4 and 13 DF,  p-value: 1.196e-05
```

Pop.inc crosses the p-value threshold so we drop it.

```
##
## Call:
## lm(formula = Mortgage.Yield ~ Loan + Distance + Savings.unit +
##     Savings.capita + Banks.Mortgage, data = df2[-1])
##
## Residuals:
##       Min        1Q    Median        3Q       Max
```

```
## -0.160935 -0.022812  0.000104  0.038353  0.141154
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)     4.184e+00  6.407e-01   6.531 2.81e-05 ***
## Loan            2.254e-02  8.622e-03   2.614   0.0226 *
## Distance        1.195e-05  4.594e-05   0.260   0.7991
## Savings.unit   -1.695e-03  7.228e-04  -2.345   0.0370 *
## Savings.capita  2.027e-04  1.101e-04   1.840   0.0906 .
## Banks.Mortgage  2.216e-04  2.256e-03   0.098   0.9234
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.09793 on 12 degrees of freedom
## Multiple R-squared:  0.8644, Adjusted R-squared:  0.8079
## F-statistic:  15.3 on 5 and 12 DF,  p-value: 7.577e-05
```

Banks.Mortgage crosses the p-value bounds too, so we drop it.

Applying logic, I decided to apply an interaction term between Loan and Pop.inc

```
##
## Call:
## lm(formula = Mortgage.Yield ~ Loan + Savings.unit + Savings.capita +
##     I(Loan/Pop.inc), data = df2[-1])
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.149894 -0.027263  0.004744  0.040772  0.128029
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      4.042e+00  5.128e-01   7.881 2.63e-06 ***
## Loan             2.515e-02  7.015e-03   3.585  0.00333 **
## Savings.unit    -1.008e-03  5.176e-04  -1.946  0.07354 .
## Savings.capita   2.109e-04  6.509e-05   3.240  0.00645 **
## I(Loan/Pop.inc) -4.939e-02  2.122e-02  -2.327  0.03677 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.07934 on 13 degrees of freedom
## Multiple R-squared:  0.9036, Adjusted R-squared:  0.8739
## F-statistic: 30.45 on 4 and 13 DF,  p-value: 1.717e-06
```

We get a far better model fit with high R^2 and low RSE and our p-values are all statistically significant.

Since our dataset is small, our coefficient estimates maybe have high variance. We could use boostrap to get more robust coefficients.

Bootstrap done with 500 bootstrap replicates.

Bootstrap Coefficients

```
##
## ORDINARY NONPARAMETRIC BOOTSTRAP
##
##
## Call:
## boot(data = df2, statistic = boot.fn, R = 500)
##
##
## Bootstrap Statistics :
##          original        bias     std. error
## t1*  4.0419008721  2.523323e-02 6.954401e-01
## t2*  0.0251465944 -2.444235e-04 9.157711e-03
## t3* -0.0010075064 -3.897045e-05 5.606637e-04
## t4*  0.0002108905 -5.382411e-06 6.200778e-05
## t5* -0.0493897609  6.668275e-04 2.787224e-02
```

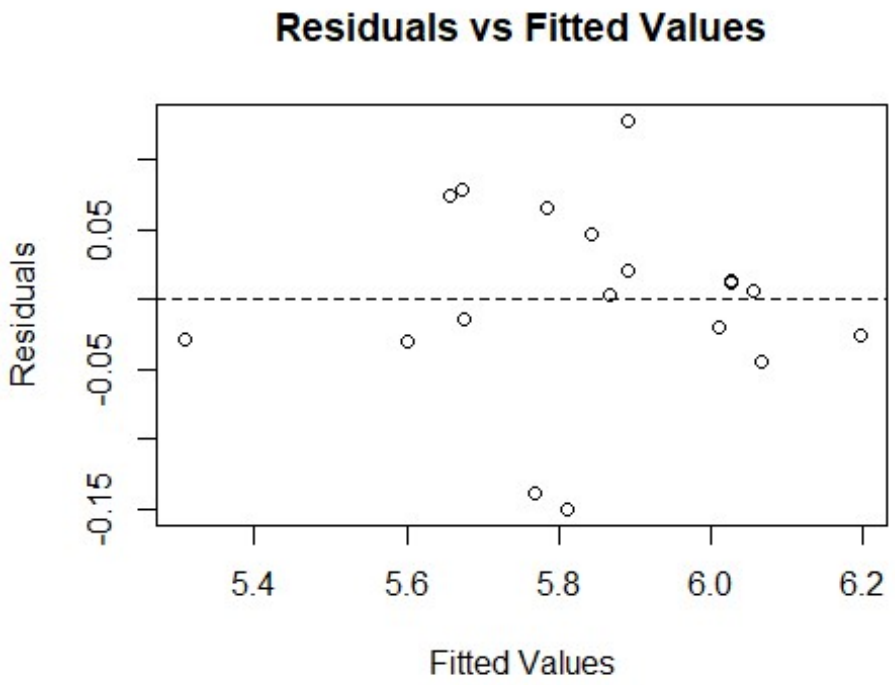
Bootstrap legend

t1 = Intercept, t2 = Loan, t3 = Savings.unit, t4 = Savings.capita, t5 = I(Loan/Pop.inc)

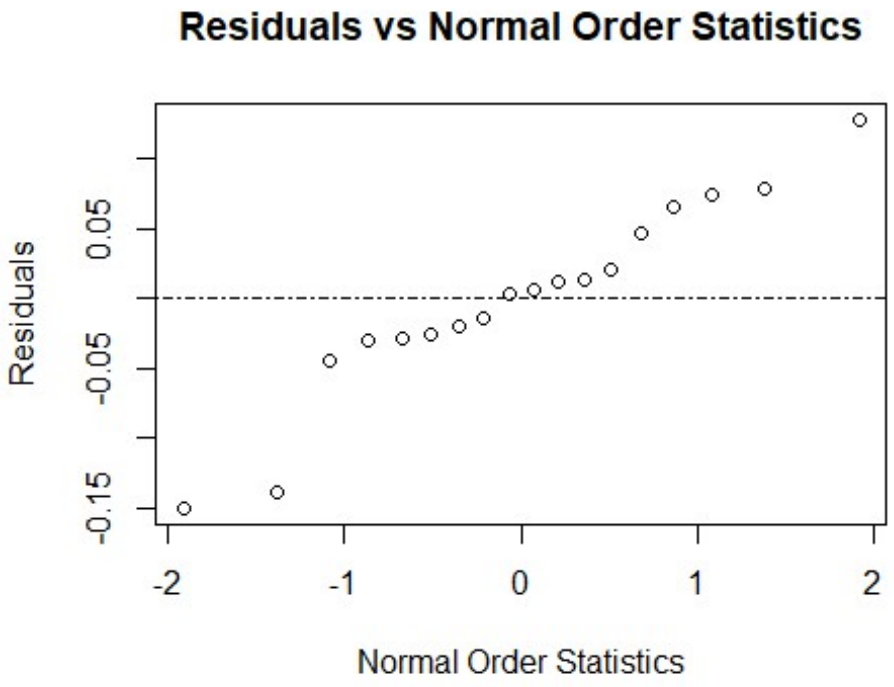
These coefficients would lead us to a more robust model.

Model Inference

Loan/Pop.inc and Savings.unit both have negative gradients and are inversely proportional to Mortgage.Yield while the other variables have positive gradients and are directly proportional.



The residuals vs Fitted values plot confirms our assumption that the relationship is linear



The above plot shows us that the residuals are mostly normally distributed.

Overall we can say that Average Loan/Mortgage Ratio, Savings per unit built, Savings per capita and ((Average Loan/Mortgage)/Pop inc) ratio together explain a 87.39% variance in Mortgage Yield.