**Question 2**

*R0 ->* Net reproductive rate

*R0 = (l1m1) + (l2m2) ………….* for 2 different time periods

*Mx ->* average number of offspring produced at each age

*lx ->* proportion of females surviving to each age

Assumption:

Provided *lx* is the proportion of females surviving to each age, from the data provided, it is assumed that the females that survived to a particular age produced offspring (i.e., N > 0, where “N” is number of offspring produced at that age) therefore, it is assumed that if N = 0, then the female did not survive to that age.

1. From the function, the estimated R0 for the provided data is **10.575**

Mean of bootstrap sample = **10.57611**

Standard Error = **0.8184001**

95% Confidence Interval is given as:

**interval**

2.5% 8.64

97.5% 11.97

1. It can be seen that the true estimate is similar in value to the mean estimate from the bootstrap sample of 100,000 estimated observations with a standard deviation of approximately 0.818. The interval shows that 95% of the tome the resampling method is used to generate the estimate, we are confident that the Net reproductive rate (R0) estimate will fall between the interval of 8.64 and 11.97.

**Question 3**

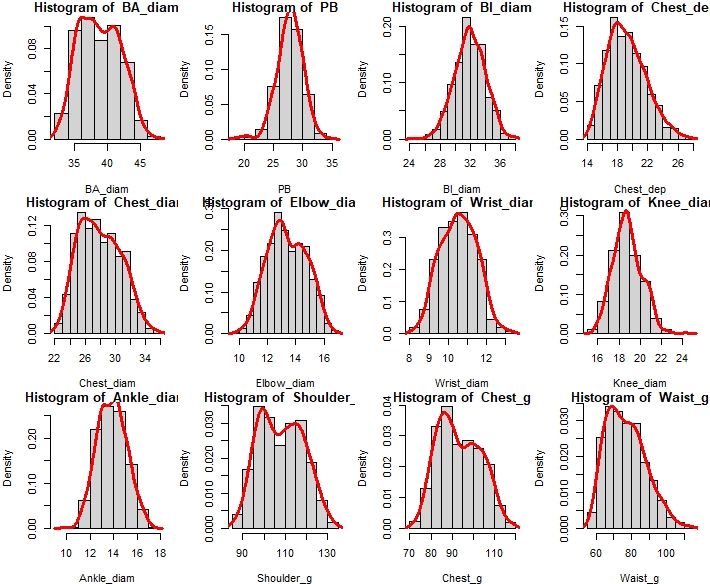
****

Figure : Histogram-Density Plot Distribution 1

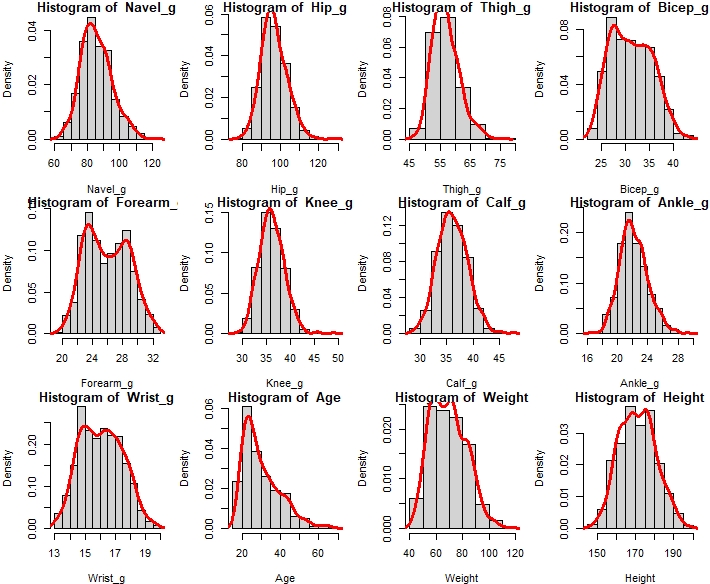
****

Figure : Histogram-Density Plot Distribution 2

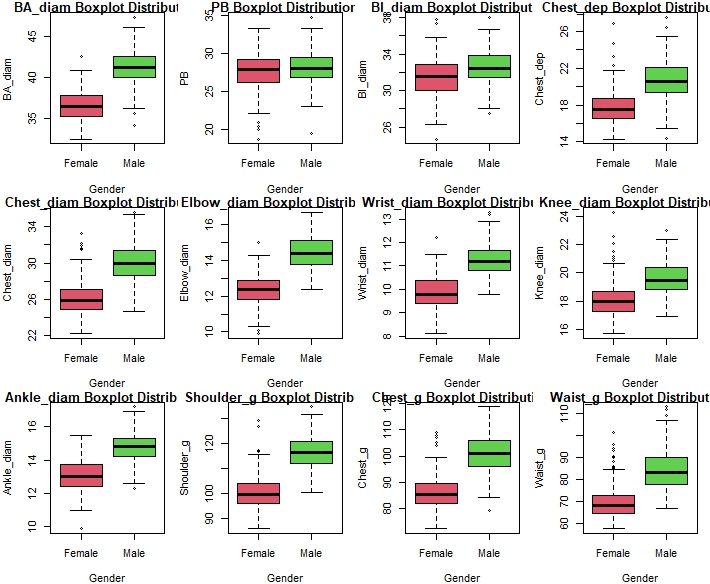
****

Figure : Boxplot Distribution of variables with respect to Gender 1

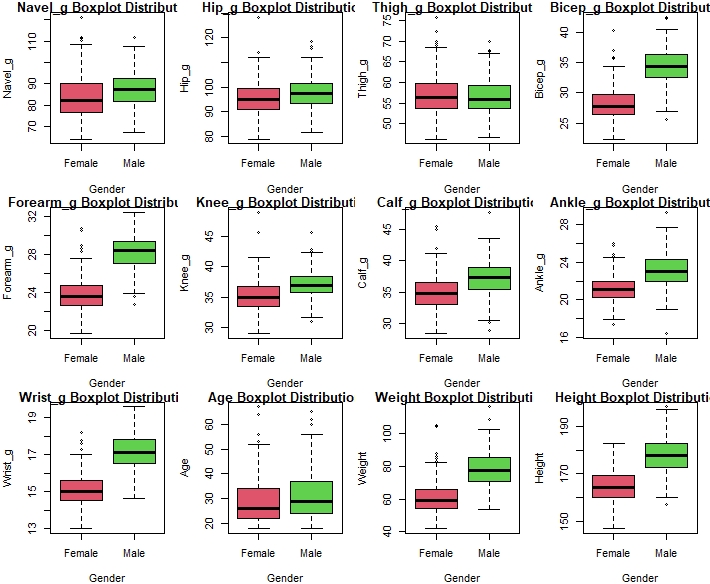
****

Figure : Boxplot Distribution of variables with respect to Gender 2

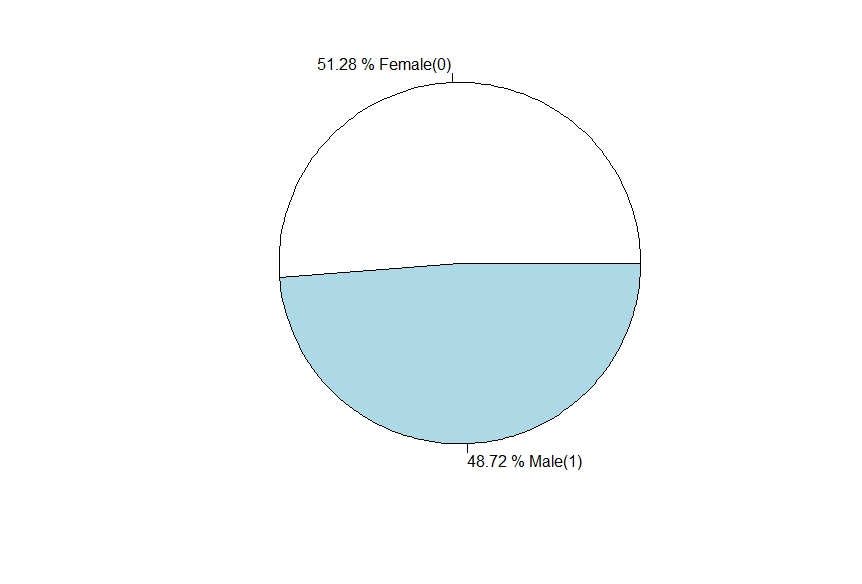
****

Figure : Pie Chart Showing Gender Proportions

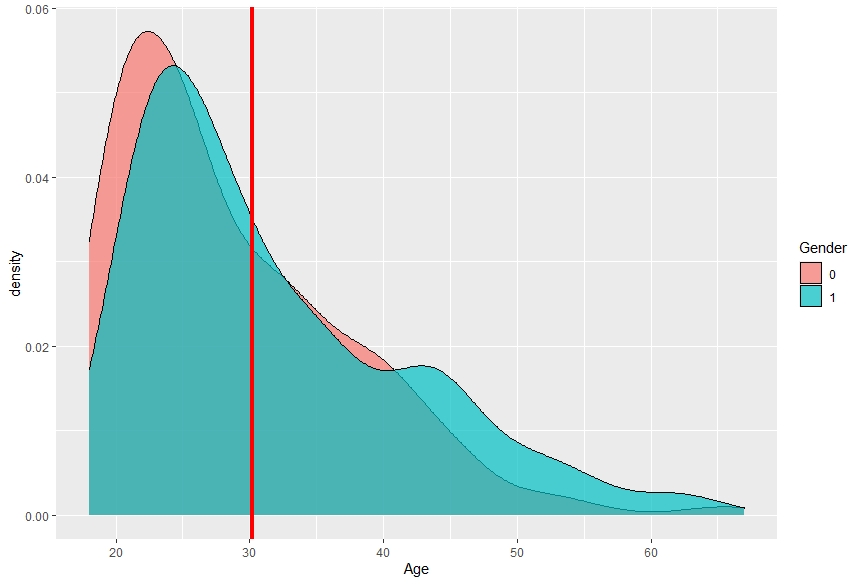
****

Figure : Age Distribution by Gender

* Figure 1 and 2 shows the histogram distribution of the numeric variables contained in the dataset, and it can be observed that majority observations for the variables in the data collected are normally distributed as indicated by the density plot (red-line).
* Figure 3 and 4 shows a boxplot distribution for all variables in the data with respect to the stated gender. It can easily be observed that mean estimate for male observation values for all variables are more than the mean estimates for females. Also, some outlier observation can be noticed for both gender for all variables.
* The pie chart in figure 5 shows the proportion of gender in the population when compared to each other, and it can be seen that of the total observed population, the female gender consists of about 51.28% while the male gender is less with 48.72%.
* The density plot in figure 6 shows the age distribution for each gender. The plot helps visualise the claim that majority of the population for each gender is either in their 20’s or 30’s. The red line indicates the mean age of the entire population and it is seen that majority of the observations for each gender is seen before the line. Therefore, the claim on the age distribution is justified.

**FORWARD STEPWISE REGRESSION**

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) -119.62205 3.45704 -34.602 < 2e-16 \*\*\*

Waist\_g 0.41204 0.03644 11.307 < 2e-16 \*\*\*

Knee\_diam 0.58496 0.17190 3.403 0.000782 \*\*\*

Forearm\_g 0.57658 0.13283 4.341 2.10e-05 \*\*\*

Hip\_g 0.18290 0.05486 3.334 0.000992 \*\*\*

Height 0.29188 0.02275 12.831 < 2e-16 \*\*\*

Calf\_g 0.46003 0.08065 5.704 3.45e-08 \*\*\*

Chest\_g 0.19206 0.04509 4.259 2.95e-05 \*\*\*

Thigh\_g 0.15700 0.06615 2.373 0.018413 \*

Gender1 -2.15675 0.65923 -3.272 0.001227 \*\*

Knee\_g 0.23881 0.09798 2.437 0.015529 \*

Shoulder\_g 0.08829 0.03631 2.432 0.015758 \*

Age -0.04069 0.01675 -2.429 0.015866 \*

Chest\_dep 0.22246 0.10065 2.210 0.028029 \*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.04 on 239 degrees of freedom

Multiple R-squared: 0.9775, Adjusted R-squared: 0.9763

F-statistic: 797.9 on 13 and 239 DF, p-value: < 2.2e-16

The above output is the result of the forward selection regression. The data used to carry out the regression is the training dataset which is 50% of the enter data. The model generated from the regression has selected 13 out of 24 variables as the optimal subset of predictors for the response variable. This model explains 97.63% of the variability found in the response variable (weight). It can also be observed that all selected predictors in this model are considered significant from the output using the p-value of their respective t-value. Also, the p-value of the F-statistic signifies that the overall model is considered significant.

The ANOVA analysis in the output below provides a better confirmation of the relevance and significance of adding the selected predictors to the model. And it would be observed that the overall model still remains significant with the addition of all optimal predictors which signifies their relevance to the response variable (weight).

Analysis of Variance Table

Response: Weight

Df Sum Sq Mean Sq F value Pr(>F)

Waist\_g 1 36466 36466 8762.8216 < 2.2e-16 \*\*\*

Knee\_diam 1 3311 3311 795.5774 < 2.2e-16 \*\*\*

Forearm\_g 1 1056 1056 253.8196 < 2.2e-16 \*\*\*

Hip\_g 1 762 762 183.0396 < 2.2e-16 \*\*\*

Height 1 868 868 208.5545 < 2.2e-16 \*\*\*

Calf\_g 1 291 291 69.8402 5.303e-15 \*\*\*

Chest\_g 1 216 216 51.9510 7.423e-12 \*\*\*

Thigh\_g 1 69 69 16.6413 6.156e-05 \*\*\*

Gender 1 40 40 9.6853 0.002084 \*\*

Knee\_g 1 26 26 6.2304 0.013233 \*

Shoulder\_g 1 23 23 5.4667 0.020208 \*

Age 1 19 19 4.6079 0.032831 \*

Chest\_dep 1 20 20 4.8856 0.028029 \*

Residuals 239 995 4

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

**Model Performance Metrics**

**RMSE Rsquared MAE**

2.267 0.973 1.738

The output above gives the Root Mean Square Error, Rsquared value and Mean Absolute Error between the actual response of the model and the predictions.

**Finding Accuracy**

**actuals predicteds**

**actuals** 1.000 0.986

**predicteds** 0.986 1.000

A high correlation value of 98.6 % indicates that actuals values and predicted values have similar directional movement. And it is observed that the actual values and predicted ones seem very close to each other. A good metric to see how much they are close is the min-max accuracy, that considers the average between the minimum and the maximum prediction. The min-max accuracy is 97.5% which further indicates that the predicted and actual values are very similar.

Finally, the Mean Absolute percentage error/deviation of the prediction is about 2.5% which is very low. This means that prediction accuracy of the model is good.

**RIDGE REGRESSION MODEL**

Fitting model with training response and predictor variable

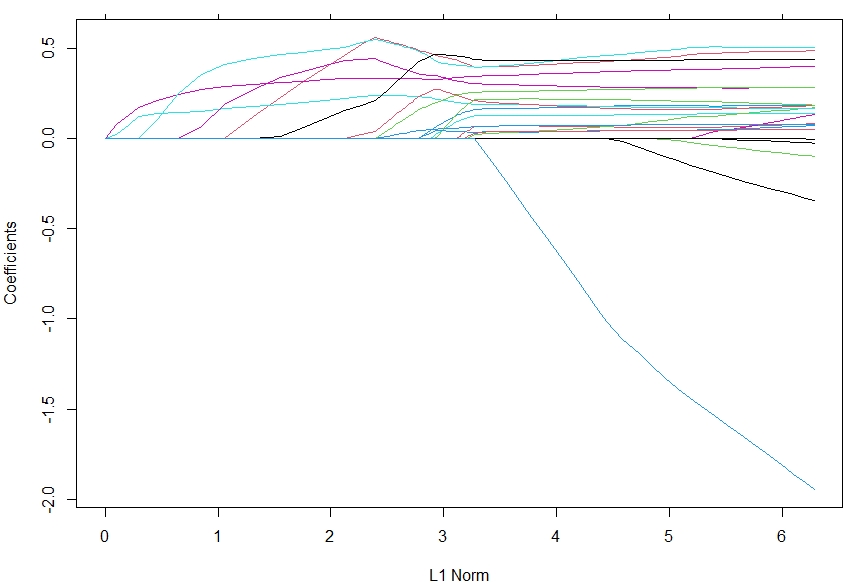


Figure 7: Fit Plot

Each curve corresponds to a variable. It shows the path of its coefficient against the ℓ1-norm of the whole coefficient vector at as λ varies.

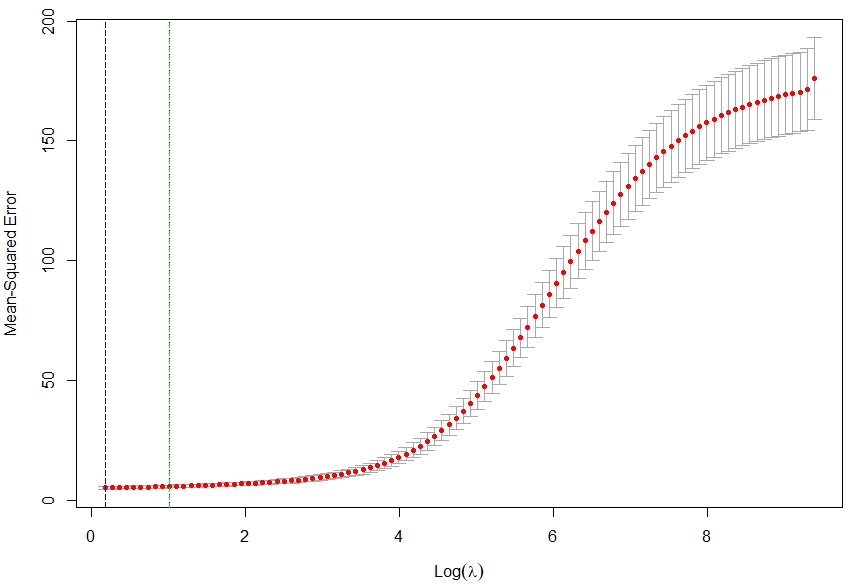


Figure 8: Log(λ) vs MSE

The above includes the cross-validation curve (red dotted line), and upper and lower standard deviation curves along the λ sequence (error bars). Two selected λ’s is indicated by the vertical dotted lines. The “**lambda.min**” (indicated on the blue dotted line) is the value of λ that gives minimum mean cross-validated error. The other λ saved is “**lambda.1se**” (indicated on the green dotted line), which gives the most regularized model such that error is within one standard error of the minimum.

The lowest point in the curve indicates the optimal lambda: the log value of lambda that best minimised the error in cross-validation. The best lambda value is 1.2

**Model Performance Metrics**

**RMSE Rsquared MAE**

2.369 0.970 1.824

The output above gives the Root Mean Square Error, Rsquared value and Mean Absolute Error between the actual response of the model and the predictions.

**Finding Accuracy**

**actuals2 predicteds2**

**actuals2** 1.000 0.984

**predicteds2** 0.984 1.000

A high correlation value of 98.4% indicates that actuals values and predicted values have similar directional movement. And it is observed that the actual values and predicted ones seem very close to each other. A good metric to see how much they are close is the min-max accuracy, that considers the average between the minimum and the maximum prediction. The min-max accuracy is 97.4% which further indicates that the predicted and actual values are very similar.

Finally, the Mean Absolute percentage error/deviation of the prediction is about 2.6% which is very low. This means that prediction accuracy of the model is good.