Class	1
Full Name	TAN CHUAN XIN
Matriculation Number	U1821755B

Declaration of Academic Integrity

By submitting this assignment for assessment, I declare that this submission is my own work, unless otherwise quoted, cited, referenced or credited. I have read and understood the Instructions to CBA.PDF provided and the Academic Integrity Policy.

I am aware that failure to act in accordance with the University's Academic Integrity Policy may lead to the imposition of penalties which may include the requirement to revise and resubmit an assignment, receiving a lower grade, or receiving an F grade for the assignment; suspension from the University or termination of my candidature.

I consent to the University copying and distributing any or all of my work in any form and using third parties to verify whether my work contains plagiarised material, and for quality assurance purposes.

Please insert an "X" within the square brackets below to indicate your selection.

[X] I have read and accept the above.

Table of Contents

For each question, please start your answer in a new page.

Answer to Q1:

Create the BMI variable based on CDC definition . Show your code.

From the website, `BMI is a person's weight in kilograms divided by the square of height in meters`. We shall implement this formula to create the BMI variable.

We reference Appendix A of `CBA Question Paper.pdf` and see that the `Height` variable is given in cm, while the `Weight` variable is given in kg. Therefore, we will have to perform some conversion for the height variable before using it.

The formula is therefore `BMI = Weight / (Height/100)^2`. We will keep BMI as a variable type of Double because the decimal points are important. However, we will round off to just one decimal point.

```
# create the BMI variable
premium2.dt$BMI <- round(premium2.dt$Weight / ((premium2.dt$Height/100)^2), 1)
head(premium2.dt$BMI)
```

```
## [1] 23.7 22.5 23.6 27.8 31.9 27.0
```

```
summary(premium2.dt$BMI)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   15.20   23.38   27.15   27.45   30.80   50.00
```

Answer to Q2:

There are many categorical variables with integer coded values (e.g. Diabetes, HighBloodPressure, Transplant…etc.) Is it necessary to convert them to factor datatype in R?

Yes, it is necessary to convert them into factor datatype in R. If we wrongly interpret the categorical variables as continuous variables, this allows for numeric concepts on continuous numbers such as fractions to become applicable.

Taking the example of the `Gender` variable, a value of 0.5 would mean that the person is halfway male and female, which is not possible. Or for the `Allergy` variable, the interpretation would be that the person has half an allergy which makes no sense as well.

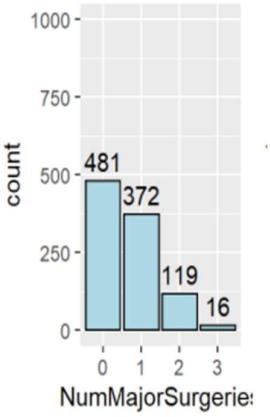
Hence, it is necessary to convert categorical variables with integer coded values into factor datatype in R.

Answer to Q3:

Explore the data and report on your key findings.

A lot of exploratory data analysis was conducted. The most important ones will be mentioned here.

NumMajorSurgeries



Findings

NumMajorSurgeries is continuous variable of integer variable type.

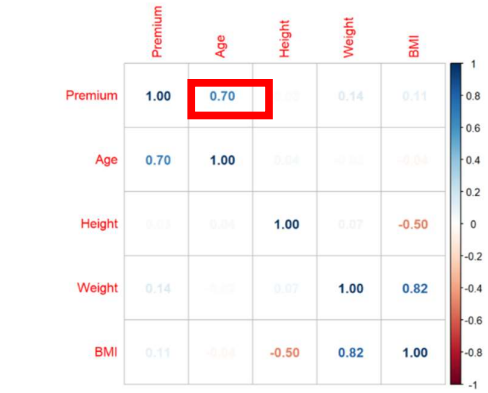
While NumMajorSurgeries could theoretically take on any continuous integer value, it practically is limited to very few values because a human can only go through that many major surgeries. Therefore, an argument can be made for this to be a factor variable.

Particularly in this case, where NumMajorSurgeries only takes on integers 0,1,2,3, we should treat it as a factor variable. It is closer to a factor variable than a continuous variable in the context of this dataset.

Conversion to factor is performed!

```
premium2.dt$NumMajorSurgeries <- factor(premium2.dt$NumMajorSurgeries, levels = c(0, 1, 2, 3), labels = c(0, 1, 2, 3))
```

Correlation for continuous variables

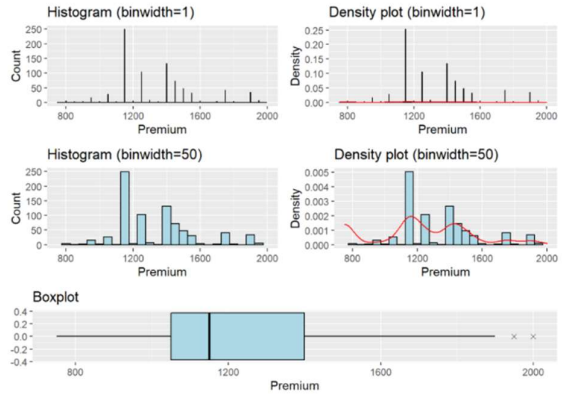


Findings

High correlation between Age and Premium suggests that Age could be one of the most important predictors for Premium within the continuous variables.

BMI has high correlation with Weight and moderately high correlation with Height, which could result in BMI being removed in the models subsequently. This makes sense - BMI was derived from Height and Weight

Premium




```
uniqueN(premium2.dt$Premium)
```

```
## [1] 24
```

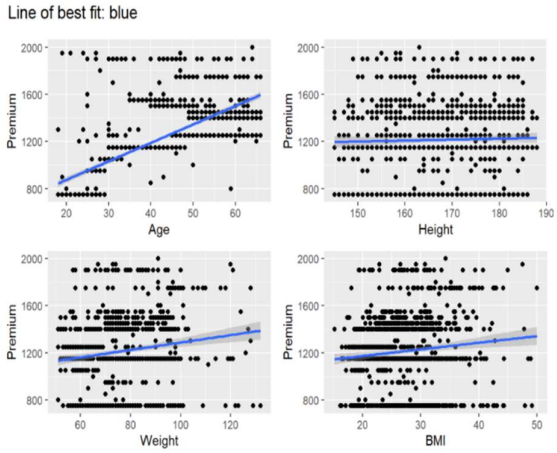
Findings

Most of the population have a premium between $1100 to $1400, but the majority has a premium of around $1150.

However, we can also note that the Premium variable takes only several distinct values (24 unique ones) between the range of 750-2000.

Having very few unique values of Premium might affect our model's error later.

Continuous variables



Findings

Age

- The older you are, the higher the premium tends to be. Likely an important variable
- Appears to be the most important continuous variable due to the high slope

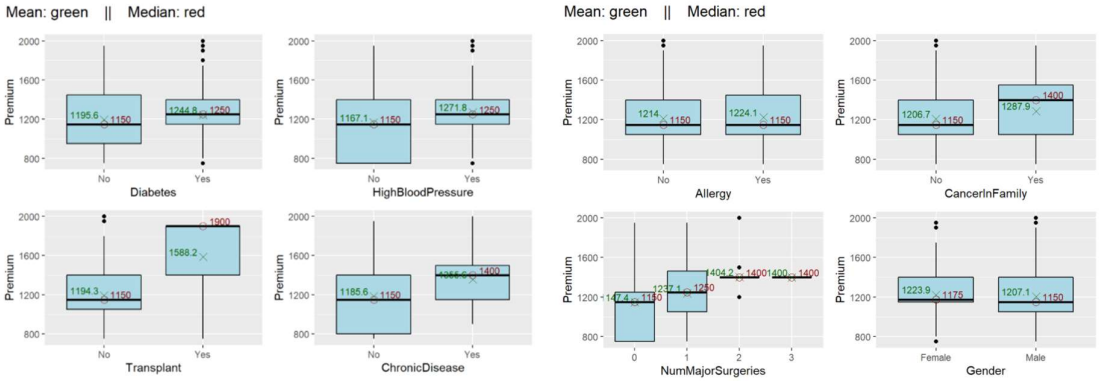
Height

- Appears to have almost no relationship with Premium

Weight

- Appears to be important as well, but less important than age

Categorical variables



Findings

- Categorical variables that are important
  - Diabetes, HighBloodPressure, CancerInFamily, Transplant, ChronicDisease, NumMajorSurgeries
  - These six variables show a clear separation between the mean and median values of Premium for their different categories
- Transplant looks to be the most important
  - The interquartile range of the two different categories of Transplant has almost zero overlap, hence it can be used to predict the Premium range with high levels of confidence
- Allergy and Gender are likely unimportant
  - Their distributions for both categories are similar

Answer to Q4:

Using 1 SE optimal CART and one other technique learnt in this course:

Given that the dependent variable is Premium, which is a continuous variable, the other technique chosen is linear regression.

a. What is the 10-fold cross validation RMSE and number of splits in the 1SE Optimal CART?

10-fold CV RMSE in 1SE Optimal CART = 161.6693
Number of splits in 1SE Optimal CART = 7

Read Appendix I for very detailed explanations. Else skip to question b

b. Identify the key predictors of premium.

For CART, the key predictors are (in order of importance)

Age > Transplant > NumMajorSurgeries > Weight > ChronicDisease > CancerInFamily

For linear regression, the key predictors are (not in order of importance)

Age, Transplant, NumMajorSurgeries, Weight, ChronicDisease, CancerInFamily

Both models have the same key predictors of Premium.

Read Appendix II for very detailed explanations. Else skip to question c

c. Is BMI or Gender important in determining premium?

	CART	Linear regression
BMI	Important (as surrogate in node 13, 6, 27)	Not important
Gender	Not important	Not important

Read Appendix III for very detailed explanations. Else skip to question d

d. <u>Evaluate and compare the predictive accuracy of the two techniques on a 70-30 train-test split. Present testset RMSE results in a table.</u>

	CART (train-test)	Linear regression (train-test)
Testset RMSE	160.9869	182.4656

```
# predict on the testset
cart_predict <- predict(cart4, newdata = testset)

# calculating rmse
cart_error <- testset$Premium - cart_predict
cart_square_error <- cart_error^2
cart_mean_square_error <- mean(cart_square_error)
cart_root_mean_square_error <- sqrt(cart_mean_square_error)

summary(abs(cart_error))
```

```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
##    4.636   45.364   64.344  101.973   78.857 1154.636
```

```
cart_root_mean_square_error
```

```
## [1] 160.9869
```

```
# predict on the testset
linreg_predict <- predict(linreg5, newdata = testset)

# calculating rmse
linreg_error <- testset$Premium - linreg_predict
linreg_square_error <- linreg_error^2
linreg_mean_square_error <- mean(linreg_square_error)
linreg_root_mean_square_error <- sqrt(linreg_mean_square_error)

summary(abs(linreg_error))
```

```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
##    0.208   56.302  114.143  135.139  179.595 1145.290
```

```
linreg_root_mean_square_error
```

```
## [1] 182.4656
```

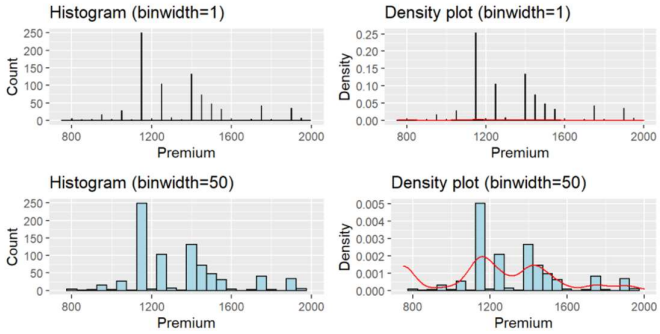
Compare

The testset RMSE was calculated as per the screenshots, using models trained on trainset data. From the table above, we can see that CART has a lower testset RMSE than linear regression. Therefore, we conclude that the predictive accuracy of the CART model is better than the linear regression model.

It is strange that the CART model has a lower RMSE than the linear regression. The dependent variable Premium is a continuous variable, and CART model we have only has eight leaf nodes, hence can only predict at most eight values. One should expect linear regression to perform better.

Evaluate

However, this can be explained by the distribution of Premium. Looking at the histograms and density plots for Premium, we can see that Premium does not take on many unique values, as seen from the large spikes. Indeed, there are actually only 24 unique values for Premium, out of 988 records.




```
uniqueN(premium2.dt$Premium)
```

```
## [1] 24
```

```
uniqueN(testset$Premium)
```

```
## [1] 20
```

```
uniqueN(cart_predict)
```

```
## [1] 7
```

```
uniqueN(linreg_predict)
```

```
## [1] 295
```

Therefore for the testset, the fewer values (7) that the CART model predicts actually match the 20 unique values of Premium better, versus the many values (295) that the linear regression model predicts.

This results in a lower RMSE value for the CART model.

Answer to Q5:

Explain the limitations of your analysis. [Max 1 page.]

1. Small sample size not giving the correct picture
    a. It is possible that the sample size of 988 records for premium2.csv is a very small dataset
    b. Indeed, when we look at the size of the insurance industry, 988 records is very small and therefore might not accurately represent the industry
    c. This will cause problems when we try and extend the current CART/linear regression models to new data
    d. We also need to know if the distributions present in the small sample size are representative of the wider population
2. Granularity of the data
    a. Within a category, there could be different levels of severity that have not been captured by this dataset
    b. For example, the factor variable of HighBloodPressure has only the levels “Yes” and “No”. This could be further broken down into four levels like “Severe”, “Moderate”, “Mild”, “No”
    c. Increasing the granularity of the data captured could lead to more accurate models
3. Lack of domain knowledge
    a. The Premium variable has only 24 unique values, in 988 records. However, this analysis interprets Premium as a continuous variable, a typical representation of monetary variables
    b. It is possible that Premium might only take on very few values because of the structure of the insurance industry. Without domain knowledge, we would not know if the insurance industry has some typical values for Premium
    c. If this were the case, the analysis could be wrong
4. Assumed linear relationship
    a. With linear regression, there is the fundamental assumption that the independent variables will have a linear relationship with the dependent variable. However it is possible that the relationship is non-linear, such as exponential
    b. Therefore, non-linear relationships cannot be captured by linear regression and that would cause the linear regression model to fail

Answer to Q6:

Is CART successful in this application? Explain. [Max 1 page.]

Yes

1. The optimal tree is simple to use
    a. A very simple set of decision rules has been generated for CART in the optimal tree
    b. This will allow for quick decision making, and will be good to use as a ballpark figure when the insurance company wants to quote their clients for insurance premium payables

Maybe

1. The error generated might be worth the cost savings
    a. The RMSE of the optimal tree on the testset is 160.9869, while the minimum value and maximum value of Premium is 750 and 2000 respectively. This represents an error range of 8% to 21.5% in the prediction for Premium
    b. The cost incurred because of this error range in prediction could be significantly lower than the savings that we gain because the CART model could lead to huge benefits through time savings, cost savings and manpower savings
2. The error generated could be lowered, with more training data
    a. With only 988 records, this CART model can be a proof of concept. By feeding this model more data, we can generate better predictions in the future

No

1. Non-parametric nature of CART
    a. The nature of CART is non-parametric – it is sensitive to the data that it has been trained on, and does not draw any inference on the relationships between the independent and dependent variables
    b. Hence, CART is non-generalizable and might not apply well to new data that is unseen, or has a significantly different distribution from the original data that the model was trained on
2. Model outputs for Premium are rather big buckets
    a. Our CART model has only depth 5, and outputs the same Premium value for many different records because there are only 8 leaf nodes
    b. Therefore, the model will quote the same premium value for a lot of people, and this might cause the insurance company to not be able to maximize their profits as they might undercharge some people, or overcharge others who might be unable to pay
    c. There is no granularity in the prediction for CART, which might be a problem when trying to maximize profit

References:

plotgrid references
https://wilkelab.org/cowplot/articles/plot_grid.html

ggplot references
https://ggplot2.tidyverse.org/reference/ggplot.html

Histogram code
http://www.sthda.com/english/wiki/ggplot2-histogram-plot-quick-start-guide-r-software-and-data-visualization

Boxplot code
https://stackoverflow.com/questions/19876505/boxplot-show-the-value-of-mean

Line of best fit code
https://stackoverflow.com/questions/15633714/adding-a-regression-line-on-a-ggplot

Decision trees are immune to multicollinearity
https://towardsdatascience.com/why-feature-correlation-matters-a-lot-847e8ba439c4#:~:text=Multicollinearity%20happens%20when%20one%20predictor,immune%20to%20multicollinearity%20by%20nature%20.

Linear regression with k-fold cross validation using **caret** library
http://www.sthda.com/english/articles/38-regression-model-validation/157-cross-validation-essentials-in-r/

CART and linear regression code
RE6013 slides

Appendix

Appendix I

a. What is the 10-fold cross validation RMSE and number of splits in the 1SE Optimal CART?

CART

The optimal tree was generated based on the full dataset. Checking the cp table, the root node error was obtained, and the xerror at the final split was obtained. These will allow us to calculate the 10-fold cross validation RMSE for the 1SE Optimal CART. The number of splits in the tree can also be determined by reading the cp table.

```
##
## Regression tree:
## rpart(formula = Premium ~ ., data = premium2.dt, method = "anova",
##     control = rpart.control(minsplit = 2, cp = 0))
##
## Variables actually used in tree construction:
## [1] Age              CancerInFamily    ChronicDisease    NumMajorSurgeries
## [5] Transplant       Weight
##
## Root node error: 96355891/988 = 97526
##
## n= 988
##
##       CP nsplit rel error xerror   xstd
## 1 0.5110      0     1.000  1.003 0.0397
## 2 0.0907      1     0.489  0.491 0.0393
## 3 0.0746      2     0.398  0.401 0.0376
## 4 0.0253      3     0.324  0.327 0.0373
## 5 0.0185      4     0.298  0.311 0.0372
## 6 0.0180      5     0.280  0.299 0.0377
## 7 0.0135      6     0.262  0.297 0.0384
## 8 0.0135      7     0.248  0.268 0.0364
```

```
# calculate errors from inspecting the printcp table
cart2_root_node_error <- 97526 # copy from cp table
cart2_xerror <- 0.268 # take the very last value of xerror from cp table
cart2_cv_rmse <- sqrt(cart2_xerror * cart2_root_node_error)
cart2_cv_rmse
```

```
## [1] 161.6693
```

Linear regression (can skip this part)

For completeness sake, we can also obtain the 10-fold CV RMSE for the linear regression model. Cross-validation was performed for linear regression using the *caret* library, and the printout provides us with the RMSE.

10-fold CV RMSE in Optimal linear regression = 186.655

```
print(linreg4)
```

```
## Linear Regression
##
## 988 samples
##   6 predictor
##
## No pre-processing
## Resampling: Cross-Validated (10 fold)
## Summary of sample sizes: 890, 889, 890, 889, 889, 889, ...
## Resampling results:
##
##   RMSE     Rsquared   MAE
##   186.655  0.6427536  133.3114
```

Appendix II

b. Identify the key predictors of premium.

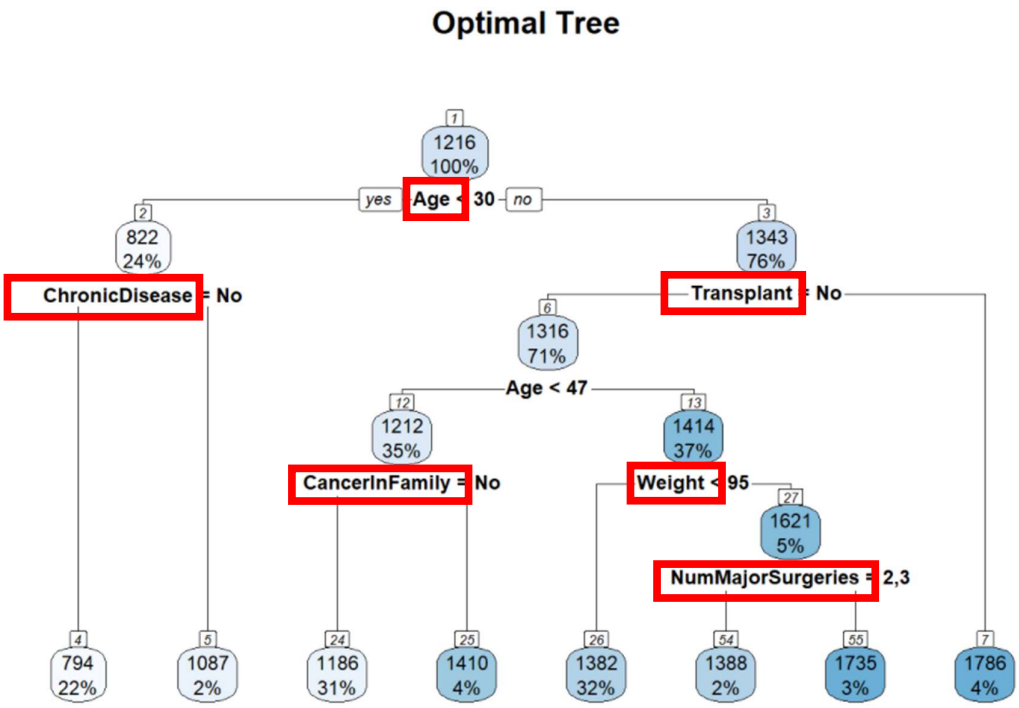
CART

Using the optimal tree that was trained on the full dataset, we can check the most important variables by simply calling up the variable importance table from the 1SE optimal cart model.

```
# check the variable importance
cart2$variable.importance
```

```
##              Age        Transplant NumMajorSurgeries            Weight
##         56591531           8735453           3514205           3373486
##    ChronicDisease    CancerInFamily               BMI HighBloodPressure
##          1784659           1733030           1349207           1096521
##         Diabetes
##          1012173
```

However, the 1SE optimal tree does not use all these variables. Therefore, we can refer to the optimal tree to pick just the key predictors, which are variables actually used in the optimal tree.



From this, the order of variable importance for the key predictors of Premium are:

Age > Transplant > NumMajorSurgeries > Weight > ChronicDisease > CancerInFamily

Six variables have been selected

Linear regression

Using the optimal linear regression model that was trained on the full dataset, we can check the most important variables by checking what variables were preserved in the final linear regression model.

```
# now that we know the optimal linear regression model, we put it through 10-fold cross validation, same as CART
# specify the 10-fold cross validation
train.control <- trainControl(method = "cv", number = 10)
linreg4 <- train(Premium ~ Age + Transplant + ChronicDisease + Weight + CancerInFamily + NumMajorSurgeries,
                 data = premium2.dt, method = "lm", trControl = train.control)
summary(linreg4)
```

```
##
## Call:
## lm(formula = .outcome ~ ., data = dat)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -657.10 -109.02  -13.45   96.87 1185.54
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)         195.2795    38.4173   5.083 4.45e-07 ***
## Age                  16.7306     0.4855  34.461  < 2e-16 ***
## TransplantYes       393.3761    25.8685  15.207  < 2e-16 ***
## ChronicDiseaseYes   134.3662    15.4937   8.672  < 2e-16 ***
## Weight                3.5603     0.4160   8.559  < 2e-16 ***
## CancerInFamilyYes    92.2126    20.1551   4.575 5.37e-06 ***
## NumMajorSurgeries1   12.7562    14.1536   0.901 0.367665
## NumMajorSurgeries2  -83.6129    21.4483  -3.898 0.000103 ***
## NumMajorSurgeries3 -163.0772    49.0348  -3.326 0.000915 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 186.2 on 979 degrees of freedom
## Multiple R-squared:  0.6479, Adjusted R-squared:  0.645
## F-statistic: 225.2 on 8 and 979 DF,  p-value: < 2.2e-16
```

From this, the key predictors of Premium are:

> Age, Transplant, NumMajorSurgeries, Weight, ChronicDisease, CancerInFamily

However, it should be noted that we can only tell what the key predictors are, but not the order of importance between them.

CART vs Linear regression

Given that the key predictors of Premium for both CART and linear regression are the exact same, we can confirm that the key predictors of Premium are definitely the following six variables

> Age, Transplant, NumMajorSurgeries, Weight, ChronicDisease, CancerInFamily

Appendix III

CART

From part b, we established that the key predictors of Premium for both models are:

Age, Transplant, NumMajorSurgeries, Weight, ChronicDisease, CancerInFamily

We might be tempted to say that BMI or Gender are unimportant because they do not appear in the six variables above. However, they are still important because they can be used as **surrogates** in CART. This is in the event where the key predictor is missing for a record, then we will need the next best alternative variable to make the split decision at a node that requires the missing key predictor.

Below shows the details for node number 13 of the 1SE optimal CART.

```
## Node number 13: 365 observations,    complexity param=0.0253133
##   mean=1413.836, MSE=25472.96
##   left son=26 (316 obs) right son=27 (49 obs)
##   Primary splits:
##       Weight         < 94.5  to the left,  improve=0.26233410, (0 missing)
##       BMI            < 28.8  to the left,  improve=0.17985150, (0 missing)
##       Age            < 65.5  to the left,  improve=0.01585263, (0 missing)
##       CancerInFamily splits as  LR,        improve=0.01407977, (0 missing)
##       Height         < 148.5 to the left,  improve=0.01192298, (0 missing)
##   Surrogate splits:
##       BMI < 36.05 to the left,  agree=0.912, adj=0.347, (0 split)
```

As we can see, BMI can be used as a surrogate at node number 13, when the primary split of Age is missing. Node number 6 and 27 also uses BMI as a surrogate variable. This means that BMI is important in determining Premium, and why it appears in the `cart2$variable.importance` call despite not appearing in the optimal tree.

Gender however does not appear even as a surrogate, hence unimportant in CART.

Linear regression

BMI and Gender both do not appear in the optimal linear regression model. This means the two variables are not statistically significant enough to be considered when performing linear regression.

There are no concepts of surrogates in linear regression, therefore both BMI and Gender are unimportant in linear regression