DATASET DECSRIPTION

The dataset consist of 986 rows and 11 columns/features, namely:

1. Age: age of the customer
2. Diabetes: whether the customer has diabetes or not
3. BloodPressureProblems: whether the customer has abnormal blood pressure levels
4. AnyTransplants: whether customer has undergone any organ transplant
5. AnyChronicDiseases: if customer suffers from any chronic ailment
6. Height: height of customer
7. Weight: weight of customer
8. NumerOfMajorSurgeries: surgeries customer has had
9. PremiumPrice: amount to be paid for insurance policy

RESEARCH SCENARIO DESCRIPTION

Many factors contribute towards the price of health insurance premiums. Some of these factors that impact the amount you pay are not within your control. I try to understand the factors - age, height, weight and if the individual has a family history of Cancer, and how much they affect insurance costs.

1. Insurance companies consider one’s immediate family’s health history. Inherited conditions may cause these companies to increase premiums. In this project I explore if family history of Cancer affects insurance rates. I will be checking whether a family history of Cancer is associated with health care costs.
2. Exploring if age, height, weight significant predictors of health care costs based on a multiple linear regression model and analyzing the contribution of age, height, weight separately towards insurance costs.
3. Checking if premium prices paid by people of different age groups is the same? (using One way ANOVA, Anova and regression). Further exploring if mean premium price paid by each pair of age groups is different? (using Tukey procedure).

DATA CLEANING

A.Converting columns to required datatypes:

> sapply(df,class)

Age Diabetes BloodPressureProblems AnyTransplants

"integer" "integer" "integer" "integer"

AnyChronicDiseases Height Weight KnownAllergies

"integer" "integer" "integer" "integer"

HistoryOfCancerInFamily NumberOfMajorSurgeries PremiumPrice

"integer" "integer" "integer

> df$Diabetes<-factor(df$Diabetes, ordered=F, labels=c('No','Yes'))

> df$BloodPressureProblems<-factor(df$BloodPressureProblems, ordered=F,labels=c("No","Yes"))

> df$AnyChronicDiseases<-factor(df$AnyChronicDiseases, ordered=F,labels=c("No","Yes"))

> df$AnyTransplants<-factor(df$AnyTransplants, ordered=F,labels=c("No","Yes"))

> df$KnownAllergies<-factor(df$KnownAllergies, ordered=F,labels=c("No","Yes"))

> df$HistoryOfCancerInFamily<-factor(df$HistoryOfCancerInFamily, ordered=F,labels=c("No","Yes"))

> unique(df$NumberOfMajorSurgeries)

[1] 0 1 2 3

Levels: 0 < 1 < 2 < 3

> df$NumberOfMajorSurgeries<-factor(df$NumberOfMajorSurgeries, ordered=T,levels=c(0,1,2,3))

> sapply(df,class)

$Age

[1] "integer"

$Diabetes

[1] "factor"

$BloodPressureProblems

[1] "factor"

$AnyTransplants

[1] "factor"

$AnyChronicDiseases

[1] "factor"

$Height

[1] "integer"

$Weight

[1] "integer"

$KnownAllergies

[1] "factor"

$HistoryOfCancerInFamily

[1] "factor"

$NumberOfMajorSurgeries

[1] "ordered" "factor"

$PremiumPrice

[1] "integer"

1. Checking for null values

colSums(is.na(df))

Age Diabetes BloodPressureProblems AnyTransplants

0 0 0 0

AnyChronicDiseases Height Weight KnownAllergies

0 0 0 0

HistoryOfCancerInFamily NumberOfMajorSurgeries PremiumPrice

0 0 0

#no null values found

1. Randomly taking 500 observations from dataset

> set.seed(123)

> df2<-df[sample(nrow(df),500),]

> dim(df2) #500 observations, 6 rows

1. 500 11
2. Checking for outliers

> #for Height column

> lower.bound<-quantile(df2$Height,0.25)-1.5\*IQR(df2$Height)

> upper.bound<-quantile(df2$Height,0.75)+1.5\*IQR(df2$Height)

> df2$Height[df2$Height<lower.bound | df2$Height>upper.bound]

integer(0)

> #no outliers

> #for Weight column

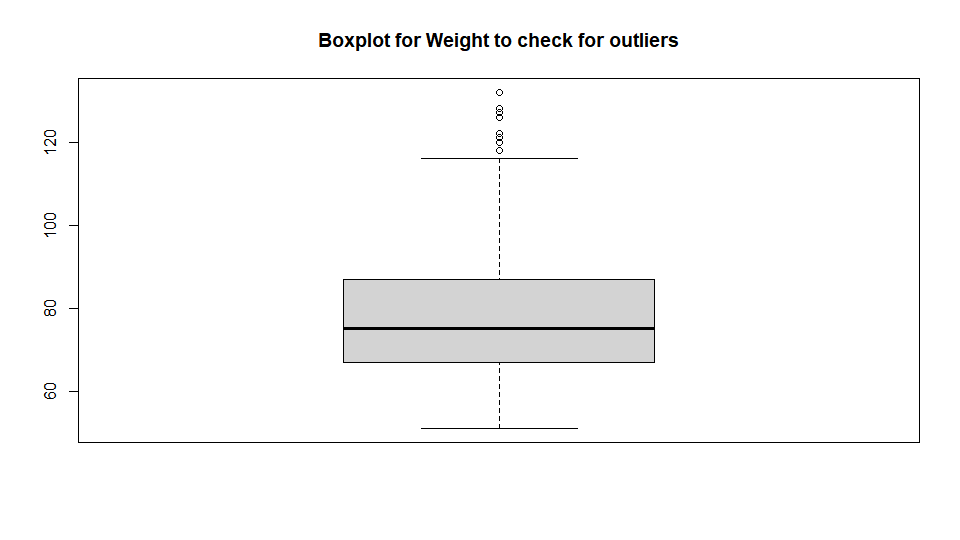
> lower.bound2<-quantile(df2$Weight,0.25)-1.5\*IQR(df2$Weight)

> upper.bound2<-quantile(df2$Weight,0.75)+1.5\*IQR(df2$Weight)

> df2$Weight[df2$Weight<lower.bound2 | df2$Weight>upper.bound2]

1. 122 121 118 132 120 126 118 128 128 121 127 120

> boxplot(df2$Weight,main="Boxplot for Weight to check for outliers")



#Removing outliers

> df2<-subset(df2,df2$Weight>lower.bound2 & df2$Weight<upper.bound2)

> dim(df2)

[1] 488 11

> #for charges column

> lower.bound3<-quantile(df2$PremiumPrice,0.25)-1.5\*IQR(df2$PremiumPrice)

> upper.bound3<-quantile(df2$PremiumPrice,0.75)+1.5\*IQR(df2$PremiumPrice)

> df2$PremiumPrice[df2$PremiumPrice<lower.bound3 | df2$PremiumPrice>upper.bound3]

[1] 39000 39000 39000

> df2<-subset(df2,df2$PremiumPrice>lower.bound3 & df2$PremiumPrice<upper.bound3)

> dim(df2)

1. 485 11

The cleaned data has been saved as “Medical\_Premiums\_cleaned.csv.”

RESEARCH QUESTION 1

Exploring whether or not customers with family history of Cancer pay higher insurance costs. ( using a significance level alpha of 0.05).

I will be exploring this question using a **2 sample t test.**

A.**Summarizing the data based on groups**- splitting into subets of customers with/without family history of Cancer and calculating summary for each group.

aggregate(df$PremiumPrice, by =list(df$HistoryOfCancerInFamily), FUN=summary)

Group.1 x.Min. x.1st Qu. x.Median x.Mean x.3rd Qu. x.Max.

1 No 15000.00 21000.00 23000.00 23914.35 28000.00 38000.00

2 Yes 15000.00 21000.00 28000.00 25962.26 31000.00 38000.00

#or

#Comparing mean and median of the two groups

df2%>%group\_by(HistoryOfCancerInFamily)%>%summarise(mean=mean(PremiumPrice))

# A tibble: 2 x 2

HistoryOfCancerInFamily mean

\* <fct> <dbl>

1 No 23914.

2 Yes 25962.

> df2%>%group\_by(HistoryOfCancerInFamily)%>%summarise(median=median(PremiumPrice))

# A tibble: 2 x 2

HistoryOfCancerInFamily median

\* <fct> <dbl>

1 No 23000

2 Yes 28000

> df2%>%group\_by(HistoryOfCancerInFamily)%>%summarise(stddev=sd(PremiumPrice))

# A tibble: 2 x 2

HistoryOfCancerInFamily stddev

\* <fct> <dbl>

1 No 5900.

2 Yes 6875.

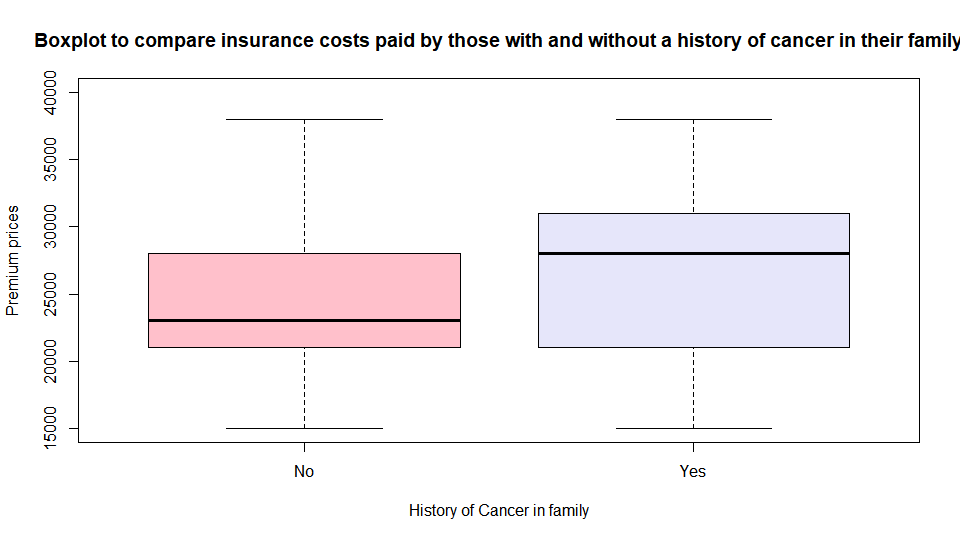
#The range for the two groups appears to be the same. However, mean and median prices both are higher for customers with a family history of Cancer.

1. **Boxplot to compare the 2 groups**

par(mfrow=c(1,1))

range(df2$PremiumPrice)

boxplot(df2$PremiumPrice~df2$HistoryOfCancerInFamily,ylim=c(15000,40000),xlab="History of Cancer in family", ylab="Premium prices", main="Boxplot to compare insurance costs paid by those with and without a history of cancer in their family",col=c('pink','lavender'))



Median price paid by customers with a family history of Cancer is higher than those without a family history of Cancer. Variability of insurance costs paid is also higher for those with a history of Cancer in their family.

1. **Performing hypothesis test**

#dividing into 2 groups

df.a<-subset(df2,HistoryOfCancerInFamily=="No")

df.b<-subset(df2,HistoryOfCancerInFamily=="Yes")

#Step 1: Stating hypothesis.

Null hypothesis, H0:mean insurance cost for customers without family history of Cancer=mean hospital insurance cost for customers with family history of Cancer

Alternate hypothesis, Ha:mean insurance cost for customers with family history of Cancer>insurance cost for customers without family history of Cancer

#Step 2: Select test statistic: t statistic for 2 sample t test

#t=xbar1-xbar2/(sqrt((s1\*\*2)/n1 + (s2\*\*2)/n2))

#Step 3: critical t value

degree.of.freedom<-min(nrow(df.a)-1,nrow(df.b)-1)

degree.of.freedom

1. 52

critical.t<-qt(0.975, degree.of.freedom)

critical.t

[1] 2.006647

#Step 4:Compute statistic

t.test(df.b$PremiumPrice,df.a$PremiumPrice,conf.level=0.95,alternative="greater")

Welch Two Sample t-test

data: df.b$PremiumPrice and df.a$PremiumPrice

t = **2.0767**, df = 61.761, p-value = 0.021

alternative hypothesis: true difference in means is greater than 0

95 percent confidence interval:

401.1926 Inf

sample estimates:

mean of x mean of y

25962.26 23914.35

#or

>t<-(mean(df.b$PremiumPrice)-mean(df.a$PremiumPrice))/(sqrt( ((sd(df.a$PremiumPrice)^2)/nrow(df.a))+((sd(df.b$PremiumPrice)^2)/nrow(df.b))))

> t

[1] 2.076743

#Step 4:Conclusion

#Since t > critical.t, reject null hypothesis. At 0.05 significance level, we can conclude that customers with a family history of Cancer pay higher for insurance as comapred to customers without a family history of Cancer.

Here, p value=0.021 (<0.05).

**D.Checking if assumptions of the performed 2 sample t test are met**:

-Independence: Both samples, insurance prices for customers with family history of Cancer and customers without family history of Cancer, are independent of each other. They do not influence each other.

-Data values for insurance policy costs are randomly selected from entire population.

-Same Measurement: The variable of interest i.e. mean of both samples is measured in the same manner by adding all values and dividing by number of values.

-No outliers in the data. They have been removed during data cleaning.

-Continuous data: Since data values represent policy costs, they are continuous measurements.

-No strong skew observed.

This can be confirmed using:

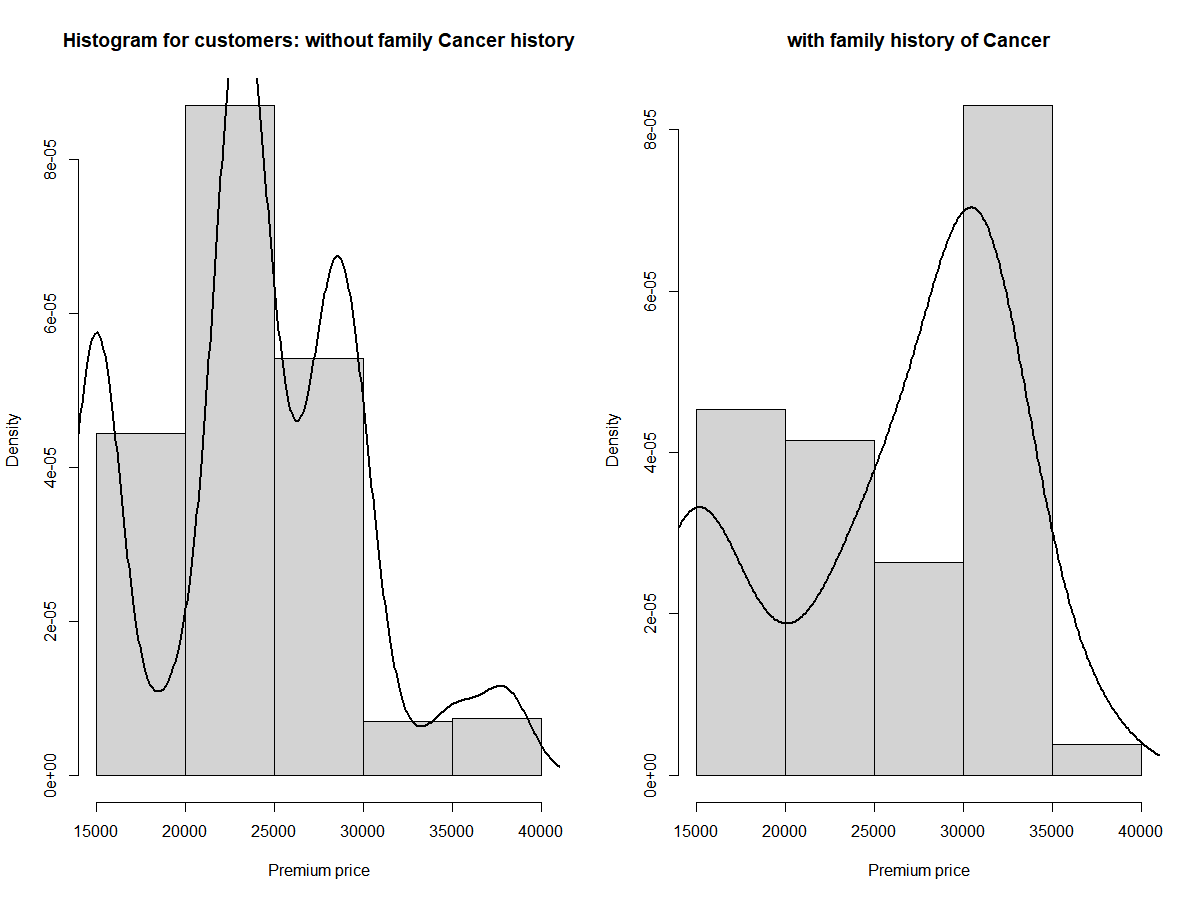
par(mfrow=c(1,2))

range(df.a$PremiumPrice)

range(df.b$PremiumPrice)

hist(df.a$PremiumPrice,xlim=c(15000,40000),main="Histogram for customers without family history of Cancer",xlab="Premium price")

hist(df.b$PremiumPrice,xlim=c(15000,40000),main="Histogram for customers with family history of Cancer", xlab="Premium price")



Range for premium price for customers with and without family cancer history is the same (15000 to 38000). The histogram for premium price for customers without family Cancer history is slightly right skewed (mean marginally greater than median).The median is around the centre of bell curve , i.e. around 23000. The mean however is pushed slightly to the right. There are more number of smaller observations.

> mean(df.b$PremiumPrice)

[1] 25962.26

> median(df.b$PremiumPrice)

[1] 28000

> mean(df.a$PremiumPrice)

[1] 23914.35

> median(df.a$PremiumPrice)

[1] 23000

The histogram for premium price for customers with family Cancer history is moderately left skewed. Median is higher than the mean. There are more number of larger observations.

> library(“moments”)

> skewness(df.a$PremiumPrice)

1. 0.1524325

Skewness of premium price for customers without family Cancer history is 0.15(almost symmetrical).

> skewness(df.b$PremiumPrice)

1. -0.5717226

Skewness of premium price for customers with family Cancer history is -0.57(moderately skewed). Left skew observed but it is not strong.

**E.95% Confidence interval**:

(x1bar-x2bar)-sqrt(s1^2/n1+s2^2/n2)

> (mean(df.a$PremiumPrice)-mean(df.b$PremiumPrice))-(critical.t\*sqrt((sd(df.a$PremiumPrice^2)/nrow(df.a))+ (sd(df.b$PremiumPrice^2)/nrow(df.b))))

[1] -7336.975

> (mean(df.a$PremiumPrice)-mean(df.b$PremiumPrice))+(critical.t\*sqrt((sd(df.a$PremiumPrice^2)/nrow(df.a))+ (sd(df.b$PremiumPrice^2)/nrow(df.b))))

[1] 3241.151

With 95% confidence we can say that mean difference in premium prices for customers with and without cancer family history is between -7336.975 and 3241.151.

RESEARCH QUESTION 2

Examining association between age, weight, height and premium costs using correlation tests. How well does a combined model with age, height, weight predict insurance costs using Multiple Linear Regression. Are age, height, weight significant predictors of health care costs using Global F test and subsequent t tests?

I have performed Multiple Linear Regression to predict premium costs based on age,height and weight. I have examined if each of these 3 features contribute significantly towards insurance costs (using Global F and subsequent t tests). I also performed correlation tests between each of the features and premium costs to see their associations. Both the methods (correlation test and MLR t tests) gave me comparable results.

**A.Correlation tests**

#For age

#Set up the hypotheses and select the alpha level

#H0:ρ=0 (there is no linear association)

#H1:ρ≠0 (there is a linear association)

#α=0.05

#Select the appropriate test statistic

#t=r\*sqrt(n−2/1−r^2)

degf=nrow(df2)-2 #n-2

#State the decision rule

qt(0.975,df=degf)

#Decision Rule: Reject H0 if |t|≥1.96, otherwise accept H0.

#Compute the test statistic

r<-cor(df2$Age,df2$PremiumPrice)

t2<-r\*(sqrt(483/(1-r^2)))

t2

[1] 22.53

#Conclusion

#Reject H0 since 22.53>1.96. We have significant evidence at the α=0.05 level that ρ≠0. That is, there is evidence of a significant linear association between age and premium costs.

#For height

#Set up the hypotheses and select the alpha level

#H0:ρ=0 (there is no linear association)

#H1:ρ≠0 (there is a linear association)

#α=0.05

#Select the appropriate test statistic

#t=r\*sqrt(n−2/1−r^2)

degf=nrow(df2)-2 #n-2

#State the decision rule

qt(0.975,df=degf) #1.96

#Decision Rule: Reject H0 if |t|≥1.96, oherwise accept H0.

#Compute the test statistic

r2<-cor(df2$Height,df2$PremiumPrice)

t3<-r2\*(sqrt(483/(1-r2^2)))

t3

[1] -1.03

#Conclusion

#Fail to reject H0 since -1.03<1.96. We do not have significant evidence at the α=0.05 level that ρ≠0. That is, there is not enough evidence to conclude significant linear association between height and premium costs.

#For weight

#Set up the hypotheses and select the alpha level

#H0:ρ=0 (there is no linear association)

#H1:ρ≠0 (there is a linear association)

#α=0.05

#Select the appropriate test statistic

#t=r\*sqrt(n−2/1−r^2)

degf=nrow(df2)-2 #n-2

#State the decision rule

qt(0.975,df=degf)

#Decision Rule: Reject H0 if |t|≥1.96, oherwise accept H0.

#Compute the test statistic

r3<-cor(df2$Weight,df2$PremiumPrice)

t4<-r3\*(sqrt(483/(1-r3^2)))

t4

1. 3.39

#Conclusion

#Reject H0 since 3.39>1.96. We have significant evidence at the α=0.05 level that ρ≠0. That is, there is evidence of a significant linear association between weight and premium costs.

**B.Combined model with age,height,weight to predict insurance premium costs (Multiple Linear Regression)**

#Here response variable (dependent variable or y) is premium price and explanatory variables (independent variables or x) are age, weight, height.

mdl<-lm(data=df2,PremiumPrice~Height+Age+Weight)

summary(mdl)

Call:

lm(formula = PremiumPrice ~ Height + Age + Weight, data = df2)

Residuals:

Min 1Q Median 3Q Max

-11860.3 -2714.6 -634.4 1297.7 19803.0

Coefficients:

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

(Intercept) 6951.60 3419.64 2.033 0.0426 \*

Height -11.94 19.07 -0.626 0.5317

Age 318.16 13.70 23.217 < 2e-16 \*\*\*

Weight 79.82 14.63 5.456 7.81e-08 \*\*\*

---

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

Residual standard error: 4106 on 481 degrees of freedom

Multiple R-squared: 0.5409, Adjusted R-squared: 0.538

F-statistic: 188.9 on 3 and 481 DF, p-value: < 2.2e-16

Coefficient of determination, R2= 0.54, hence 54% of the variation in Premium prices is explained by the linear model. Model fits data moderately and approximately half the variation y is explained by all x.

Adjusted R squared=0.538.

#y=6951 - 11.94x1 + 318.16x2 + 79.82x3, where y=premium price, x1=height, x2=age and x3=weight.

#B0 (intercept): value of y when x =0. In this case value of insuarnce cost when age, height and weight are 0 will be 6951.60. B0 interpretation is not relevant here since these parametsrs cannot be zero.

#B1(coefficient): Premium price decreases by 11.94 when there is a unit increase in height, after controlling for age and weight. Premium price increases by 318.16 when there is a unit increase in age, after controlling for height and weight. Premium price increases by 79.82 when there is a unit increase in weight, after controlling for age and height.

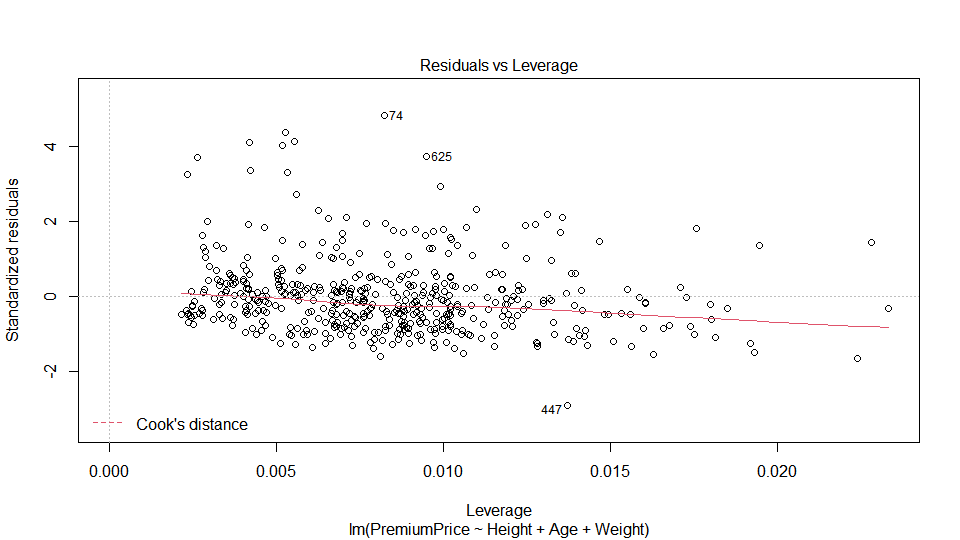
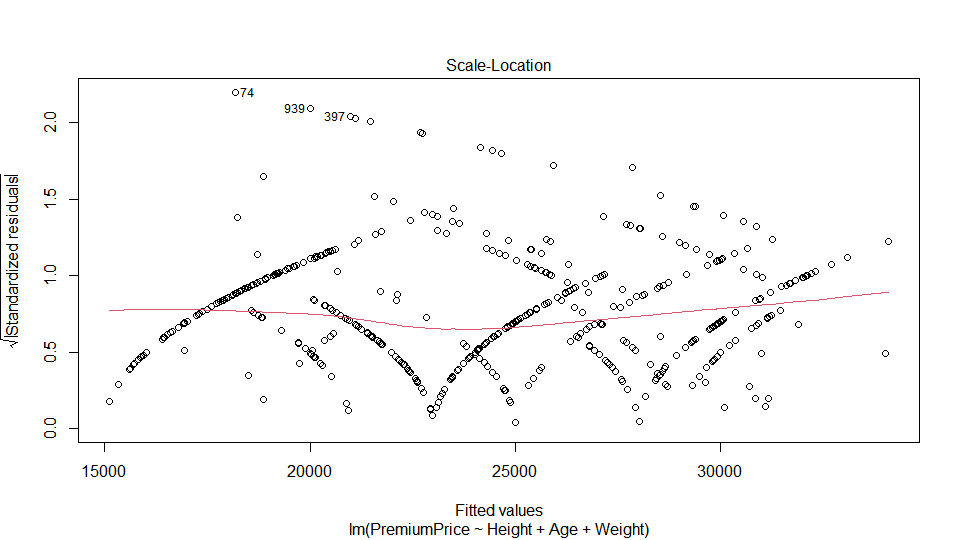
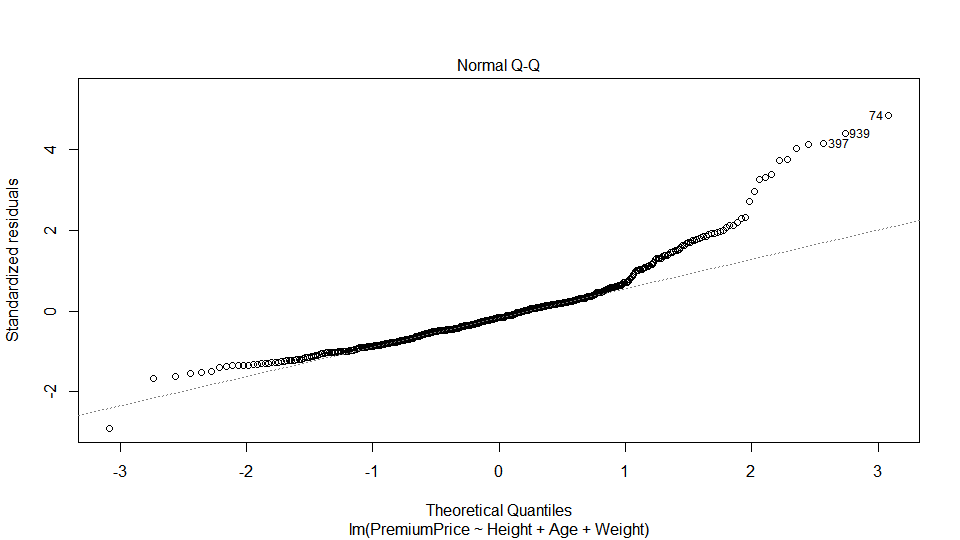
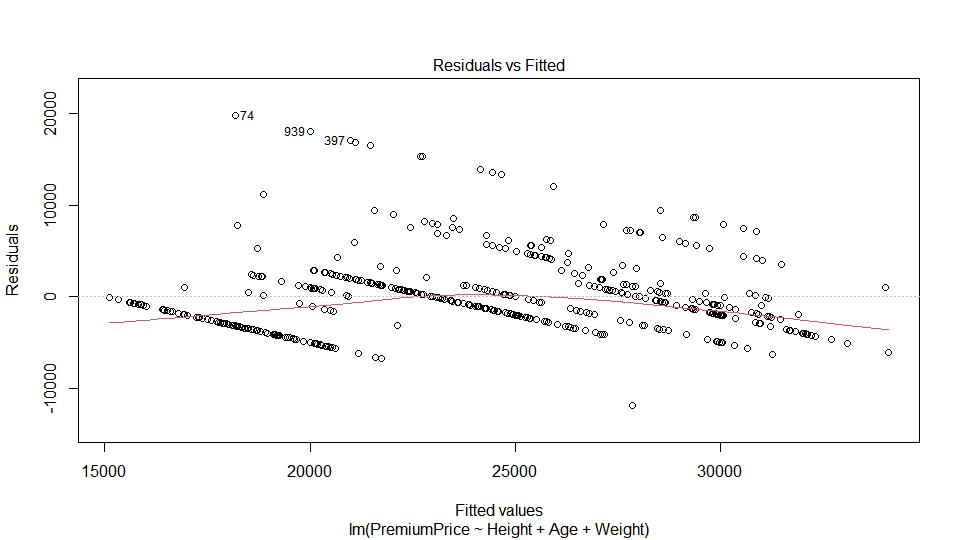
#Root mean square error (rmse)

#RMSE

sqrt(mean((df2$PremiumPrice-fitted(mdl))^2))

[1] 4088.564

plot(mdl)



#Applying log transform on y to improve model

y<-sqrt(df2$PremiumPrice)

mdllog<-lm(data=df2,y~Age+Height+Weight)

summary(mdllog)

Call:

lm(formula = y ~ Age + Height + Weight, data = df2)

Residuals:

Min 1Q Median 3Q Max

-0.56479 -0.11504 -0.01118 0.08024 0.75554

Coefficients:

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

(Intercept) 9.3207394 0.1420182 65.631 < 2e-16 \*\*\*

Age 0.0146228 0.0005691 25.694 < 2e-16 \*\*\*

Height -0.0005757 0.0007920 -0.727 0.468

Weight 0.0030375 0.0006076 4.999 8.08e-07 \*\*\*

---

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

Residual standard error: 0.1705 on 481 degrees of freedom

Multiple R-squared: 0.5871, Adjusted R-squared: 0.5845

1. statistic: 228 on 3 and 481 DF, p-value: < 2.2e-16

Approximately 59% of variation in log of Premium prices can now be explained by the age, height and weight of the customer. Adjusted R squared=0.584.

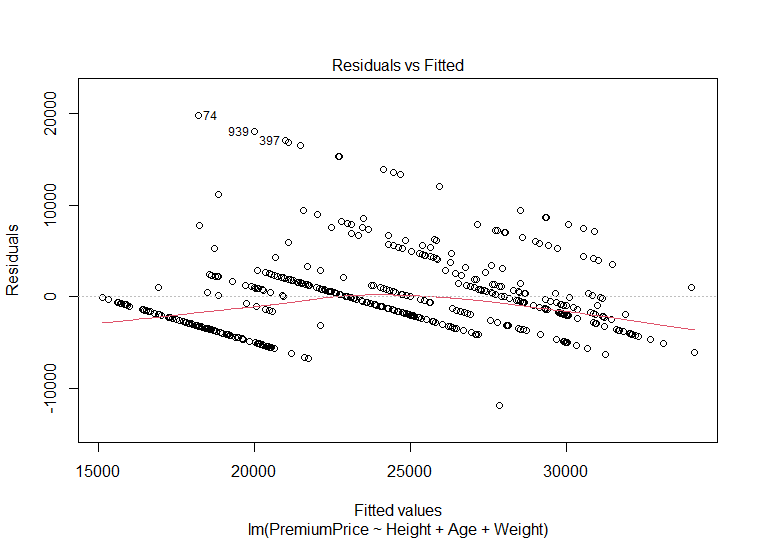
#RMSE

> sqrt(mean((df2$PremiumPrice-fitted(mdllog))^2))

[1] 24871.16

**D.Checking for assumptions of MLR**

# Assumption1: Linearity between y(premium price) and each independent variable.



Here the relationship Is not completely linear. It is slightly curved.

#Additivity assumption is satisfied as well since there is no categorical variable.

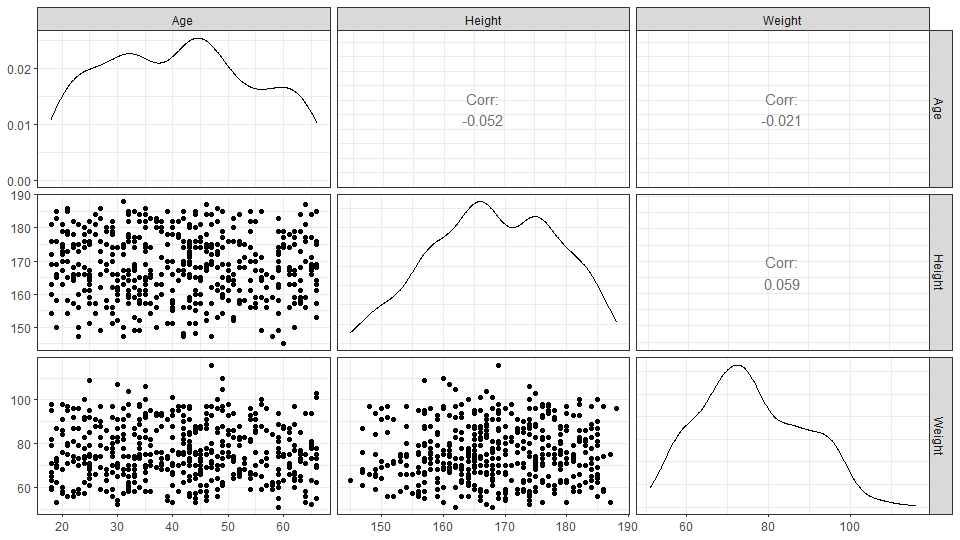
#Assumption2: Independence- Each row represents a unique customer, hence this assumption is valid. The rows are not dependent on each other. This can also be verified using a pairplot to make sure that the various independent variables are not correlated.

install.packages("GGally")

require(GGally)

df3<-df2[,c(1,6,7)]

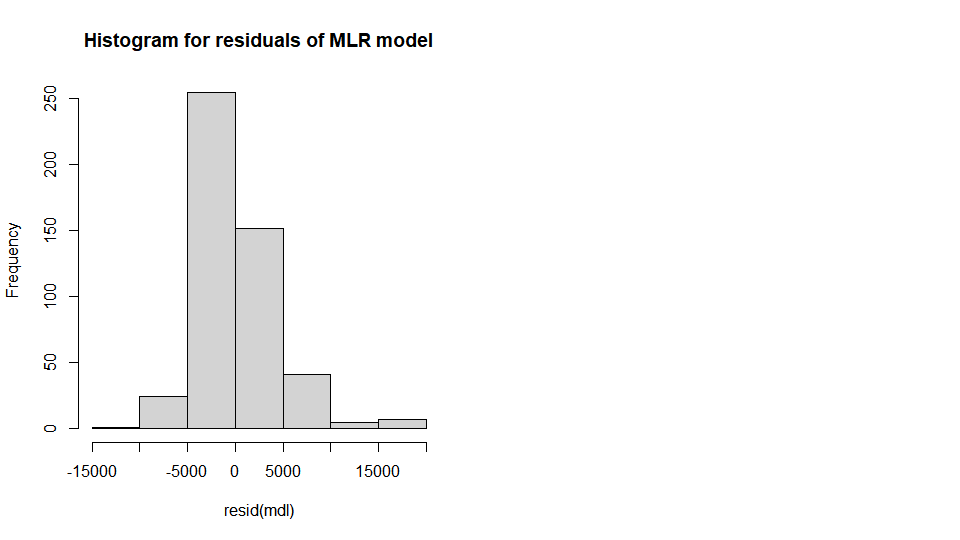
ggpairs(df3) + theme\_bw()



#Assumption3: Constant variance. This can be verified using the scale location plot above. The red line here is mostly horizontal, there are no obvious bumps/patterns. Thus variability of residuals is constant across the regression line. Variance of residuals does not depend on y.

#Assumption4: Nomality-Residuals should follow a normal distribution. This can be verified from the Q-Q plot above. Distribution is close to normal with light deviation in the region of higher quantiles. This can also be verified using:

hist(resid(mdl),main="Histogram for residuals of MLR model")



#Residuals vs Leverage plot- Leverage on the x axis( change in predicted value of y on minor change in actual y) and standardised resiauls on y axis. Cook’s distance indiactes how much the reression line changes on deleting observation i. Cook’s distance contours in a resuals vs leverage plot indicates possible outliers. If these points have a higher leverage as compared to all other points, they are probably influential.

There are no points around Cook’s distance contours. Thus, there are no outliers and hence no influence points in this case.

#Checking for influence points- examining points with id 74,625 and 447

Creating a model by removing each of these points individually and checking if it made any significant difference.

df3<-df2[-74,]

mdl2<-lm(df3$PremiumPrice~df3$Height+df3$Weight+df3$Age)

summary(mdl)

summary(mdl2)

df3b<-df2[-447,]

mdl2b<-lm(df3b$PremiumPrice~df3b$Height+df3b$Weight+df3b$Age)

summary(mdl2b)

df3c<-df2[-625,]

mdl2c<-lm(df3c$PremiumPrice~df3c$Height+df3c$Weight+df3c$Age)

summary(mdl2c)

Model parameters (R2,B1,B0) remain almost the same even when these points were removed. Slight increase in B0 is seen on removing point 447, however it does not seem to be an influence point.

Thus, there are no influence points.

**E.Global F test to check if atleast one out of the three independent variables contribute significantly towards prediction of insurance costs**.

#Global F test

#Step 1: Set hypothesis

#H0: B1=0 i.e. B(age)=0,B(height)=0 and B(weight)=0 (no association)

#Ha: B1!=0 i.e. B(age)!=0 and/or B(height)!=0 and/or B(women)!=0 (atleast one of the predictors is significant)

#Step 2: Decide test statistic

#F=Regression mean square/Residual mean square

degree.of.freedom1=3 #k

dim(df2)

[1] 485 11

degree.of.freedom2=485-4 #n-k-1

degree.of.freedom2

1. 481

#Step 3:Decision rule

alpha=0.05

qf(0.95,df1=degree.of.freedom1,df2=degree.of.freedom2)

#If f>= critical f value, i.e. f>=2.623, reject null hypothesis. Otherwise, accept null hypothesis.

[1] 2.623442

#Step 4: Calculate test statistic

summary(mdl)

Call:

lm(formula = PremiumPrice ~ Height + Age + Weight, data = df2)

Residuals:

Min 1Q Median 3Q Max

-11860.3 -2714.6 -634.4 1297.7 19803.0

Coefficients:

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

(Intercept) 6951.60 3419.64 2.033 0.0426 \*

Height -11.94 19.07 -0.626 0.5317

Age 318.16 13.70 23.217 < 2e-16 \*\*\*

Weight 79.82 14.63 5.456 7.81e-08 \*\*\*

---

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

Residual standard error: 4106 on 481 degrees of freedom

Multiple R-squared: 0.5409, Adjusted R-squared: 0.538

F-statistic: 188.9 on 3 and 481 DF, p-value: < 2.2e-16

#F statistic=188.9

#Step 5: Conclusion

#Since F> critical , reject null hypothesis that B=0.

#Thus at 0.05 significance level we can say that atleast one out of age, height and weight are significant predictors of insurance cost.

**F.t tests to check contribution of each age, height and weight towards premium costs.**

#For age

#Step 1: Set hypothesis

#H0: B(age)=0(after controlling for height,weight)

#Ha: B(age)!=0 (after controlling for height,weight)

#Step 2: select test statistic

#one sample 2 sided t test

#t=B/SE(B)

degree.of.freedom4<-nrow(df2)-4 #n-k-1

degree.of.freedom4

[1] 481

#Step 3: State decision rule

#If |t|>= critical t value, reject null hypothesis, otherwise accept H0.

qt(0.975,df=degree.of.freedom4)

[1] 1.964

#Step 4: Compute test statistic

summary(mdl)

#t=23.217

#or

318.16/13.70

#Step 5: Conclusion

#Since |t|> critical t, reject null hypothesis.

#Thus at 0.05 significance level we can say that B(age)!=0 after controlling for height and weight.

#For height

#Step 1: Set hypothesis

#H0: B(height)=0(after controlling for age,weight)

#Ha: B(height)!=0 (after controlling for age,weight)

#step 2: select test statistic

#one sample 2 sided t test

#t=B/SE(B)

degree.of.freedom5<-nrow(df2)-4 #n-k-1

#Step 3: State decision rule

#If |t|>= critical t value, reject null hypothesis, otherwise accept H0.

qt(0.975,df=degree.of.freedom5)

[1] 1.964

#Step 4: Compute test statistic

summary(mdl)

#t=-0.626

#Step 5: Conclusion

#Since |t|< critical t, fail to reject null hypothesis.

#Thus at 0.05 significance level we do not have enough evidence to conclude that B(height)!=0 after controlling for age and weight.

#For weight

#Step 1: Set hypothesis

#H0: B(weight)=0(after controlling for age,height)

#Ha: B(weight)!=0 (after controlling for age,height)

#step 2: select test statistic

#one sample 2 sided t test

#t=B/SE(B)

degree.of.freedom6<-nrow(df2)-4 #n-k-1

#Step 3: State decision rule

#If |t|>= critical t value, reject null hypothesis, otherwise accept H0.

qt(0.975,df=degree.of.freedom6)

[1] 1.964

#Step 4: Compute test statistic

summary(mdl)

#t=5.456

#Step 5: Conclusion

#Since |t|> critical t, reject null hypothesis.

#Thus at 0.05 significance level we can say that B(weight)!=0 after controlling for age and height.

#Thus age,weight are significant contributors towards premium costs after controlling for other variables.

Model is significant (from Global F test). On further t tests, we found that two out of the 3 features (age, weight) are significant.

**G.95% Confidence intervals**

confint(mdl,level=0.95)

2.5 % 97.5 %

(Intercept) 232.31895 13670.88141

Height -49.40711 25.53529

Age 291.23198 345.08497

Weight 51.07491 108.57061

95% of the time B1(age) lies between 291.23 to 345.08. 95% of the time B1(weight) lies between 51.07 to 108.57.

RESEARCH QUESTION 3:

Are premium prices paid by young people, middle aged people and seniors the same?

1. **One Way ANOVA**

I try to explore this using One Way ANOVA. Two measures used in this procedure are:

* Within grop variability (variance between each point and respective group mean.
* Between group variability(variance between each group mean and overall mean).

df2$Agegroup[df2$Age >= 18 & df2$Age < 34] = "Young"

df2$Agegroup[df2$Age >= 34 & df2$Age < 50] = "Middle.aged"

df2$Agegroup[df2$Age >=50] = "Senior"

df2$Agegroup<-factor(df2$Agegroup)

#Using One way ANOVA (Analysis of Variance) to compare means of more than 2 groups.

1. Set up the hypotheses and select the alpha level

H0 ∶ u(young) = u(middleaged) =u(µsenior) (All underlying population means are equal)

H1 ∶ ui != uj for some i and j (Not all of the underlying population means are equal)

alpha = 0.05

2. Select the appropriate test statistic

F =Mean Square Between/Mean Square Within

k − 1 = 2 and n − k = 485 − 3 = 482 degrees of freedom

nrow(df2)-3

[1] 482

3. State the decision rule

qf(0.95, df1=2, df2=482)

[1] 3.014429

#critical F=3.0144

Decision Rule: Reject H0 if F ≥ 3.0144. Otherwise, do not reject H0.

4. Compute the test statistic

m<-aov(PremiumPrice~Agegroup,data=df2)

summary(m)

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

Agegroup 2 8.891e+09 4.446e+09 244.4 <2e-16 \*\*\*

Residuals 482 8.769e+09 1.819e+07

---

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

#F=244.4

#or

m2<-lm(PremiumPrice~Agegroup,data=df2)

anova(m2)

5. Conclusion

Reject H0 since 244.4 > 3.0144.

We have significant evidence at alpha = 0.05 level that there is a difference in Premium prices among young, middle aged and senior people.

1. **How many people are in each group? Graphical and numerical summary for premium prices across 3 age groups.**

#Count of people in each group

> table(df2$Agegroup)

Middle.aged Senior Young

182 139 164

There are 164 young people, 182 middle aged people and 139 senior people.

#Summarising price based on age groups

> aggregate(df2$PremiumPrice, by=list(df2$Agegroup), summary)

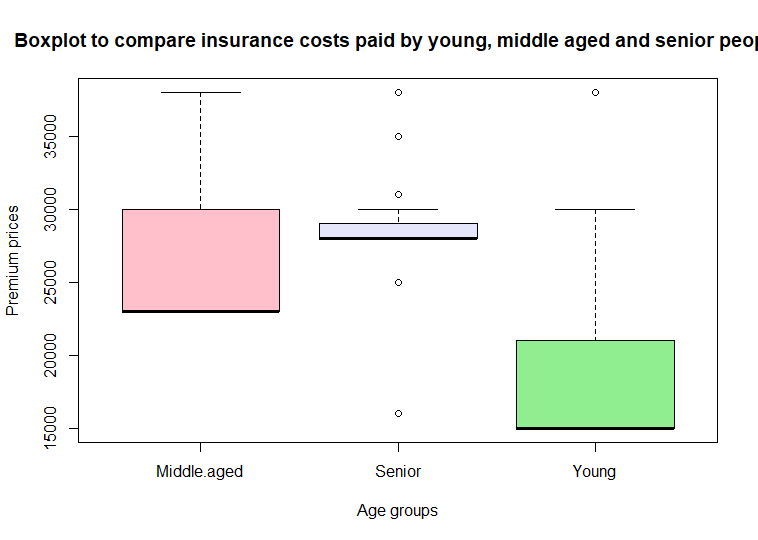
Group.1 x.Min. x.1st Qu. x.Median x.Mean x.3rd Qu. x.Max.

1 Middle.aged 23000.00 23000.00 23000.00 25961.54 30000.00 38000.00

2 Senior 16000.00 28000.00 28000.00 28597.12 29000.00 38000.00

3 Young 15000.00 15000.00 15000.00 18335.37 21000.00 38000.00

> boxplot(df2$PremiumPrice~df2$Agegroup,xlab="Age groups", ylab="Premium prices", main="Boxplot to compare insurance costs paid by young, middle aged and senior people",col=c('pink','lavender',"light green"))



1. **Subsequent pairwise comparisons using Tukey procedure to adjust for multiple comparisons.**

#Pairwise comparisons

#Step 1: State hypothesis

#H0, null hypothesis: ui=uj, means of both groups are equal

#Ha, alternate hypothesis, ui !=uj, groups means are not equal

#Step 2: Decide test statistic

#t = B/ SE

#two sided t test

#degree of freedom=n-k, where n=number of observations in each group and k=number of groups

#Step 3: State decision rule

#If |t|>= critical t value, reject null hypothesis. Otherwise accept H0.

#Or if p<0.05, reject null hypothesis. Otherwise accept H0.

#Step 4: Calculate test statistic

TukeyHSD(m) #using tukey procedure to adjust for family wise type 1 error rate.

Tukey multiple comparisons of means

95% family-wise confidence level

Fit: aov(formula = PremiumPrice ~ Agegroup, data = df2)

$Agegroup

diff lwr upr p adj

Senior-Middle.aged 2635.584 1506.043 3765.125 2e-07

Young-Middle.aged -7626.173 -8705.798 -6546.547 0e+00

Young-Senior -10261.756 -11417.828 -9105.685 0e+00

Step 5: Conclusion

With 95% confidence we can say that mean difference between premium prices of senior and middle aged people lies between 1506.043 and 3765.125.

With 95% confidence we can say that mean difference between premium prices paid by yound and middle aged people lies between -8705.798 and -6546.547.

With 95% confidence we can say that mean difference between premium prices paid by yound and senior people lies between -11417.828 and -9105.685.

All pairwise comparisons were significant at the alpha=0.05 level. Since p value<0.05, reject H0 that mean of the two groups is equal. Thus at 0.05 significance level we can say that mean premium prices are different for Young and Middle aged, Middle Aged and Seniors, and Young and Senior people.

Difference in group means for Young and Middle aged, Seniors and Middle aged, and Young and Senior people

are -7626.173,2635.584 and -10261.756 respectively.

1. **One Way ANOVA and Regression**

#creating dummy variables

df2$young<-ifelse(df2$Agegroup=="Young",1,0)

df2$middle<-ifelse(df2$Agegroup=="Middleaged",1,0)

df2$old<-ifelse(df2$Agegroup=="Old",1,0)

m3<-lm(PremiumPrice~young+middle,data=df2) #setting Seniors as reference group (using k-1 dummy variables in linear model to avoid multicollinearity)

summary(m3)

Call:

lm(formula = PremiumPrice ~ young + middle, data = df2)

Residuals:

Min 1Q Median 3Q Max

-12597.1 -3335.4 -961.5 2664.6 19664.6

Coefficients:

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

(Intercept) 28597.1 361.8 79.047 < 2e-16 \*\*\*

young -10261.8 491.7 -20.868 < 2e-16 \*\*\*

middle -2635.6 480.5 -5.486 6.67e-08 \*\*\*

---

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

Residual standard error: 4265 on 482 degrees of freedom

Multiple R-squared: 0.5035, Adjusted R-squared: 0.5014

F-statistic: **244.4** on 2 and 482 DF, p-value: < 2.2e-16

anova(m3)

Analysis of Variance Table

Response: PremiumPrice

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

young 1 8343581975 8343581975 458.631 < 2.2e-16 \*\*\*

middle 1 547437858 547437858 30.092 6.666e-08 \*\*\*

Residuals 482 8768724496 18192374

---

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

Here F=244.4 (same as one way anova). Since F> critical F value (3.0144), reject null hypothesis that mean of prices of all age groups is equal. Thus at 0.05 significance level we can say that mean of atleast one pair of groups is different. These results are same as One way ANOVA.

B0: Mean of premium prices of Seniors= 28597.1

Coefficient interpretation: mean of premium prices for young people=28597.1-10261.8 =18335.3 or difference between mean premium prices of senior and young people is 10261.8 .

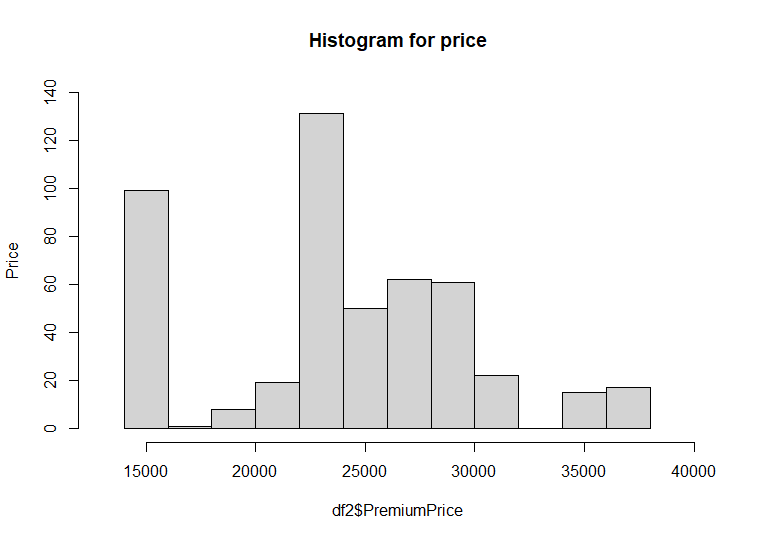
Mean of premium prices of middle aged people=28597.1-2635.6=25961.5 or difference between mean premium prices of senior and middle aged people is 2635.6.

Equation: y=28597.1- 10261.8 .group(young)- 2635.6.group(middle.aged)

1. **Assumptions of One Way ANOVA.**

1. Each sample is an independent random sample. Here each row (representing an individual) is independent.

2. Distribution of the response variable follows a normal distribution.



mean(df2$PremiumPrice)

[1] 24138.14

> median(df2$PremiumPrice)

1. 23000

Mean>median, slight right skew observed.

3. Each group has equal population variance for the response variable.

Largest sample variance divided by smallest sample variance is not greater than two.

> aggregate(df2$PremiumPrice, by=list(df2$Agegroup), var)

Group.1 x

1 Middle.aged 17992988

2 Senior 9937963

3 Young 25402177

This assumption is not satisfied.

CONCLUSION

* Customers with a family history of Cancer pay higher for insurance as comapred to customers without a family history of Cancer (p value=0.021 (<0.05)).
* Significant linear association was found between weight and premium costs, and age and premium costs.

(not height and premium costs)

* The Multiple Regression model predicts premium costs moderately (half of the variation in insurance costs can be explained by age, weight and height). Applying log transformation helps improve the model’s performance (multiple R2=0.59).
* Age and weight are significant predictors of insurance costs. (and not height)

[Using Global F test, I concluded that atleast one out of age, height and weight are significant predictors of insurance cost.

Using t tests, I found that two out of the 3 features (age, weight) are significant predictors of insurance costs.

* Using One Way ANOVA, I confirmed that there is a difference in Premium prices paid by young, middle aged and senior people.
* Mean premium prices are different for Young and Middle aged, Middle Aged and Seniors, and Young and Senior people (based on results of pairwise comparisons using Tukey procedure).
* I corroborated the results of One way ANOVA using Regression (Mean price paid by atleast one pair of age groups is different).