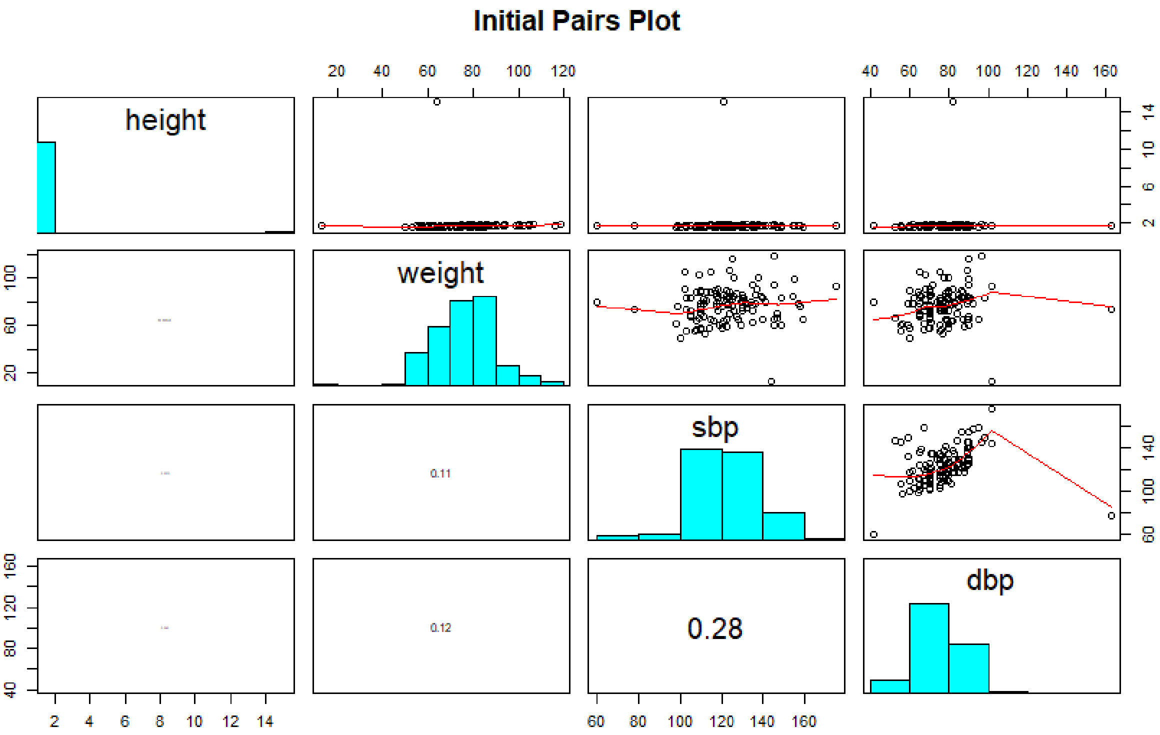
Stats 330 Assignment 1

Hasnain Cheena

16/03/2020

Data Cleaning

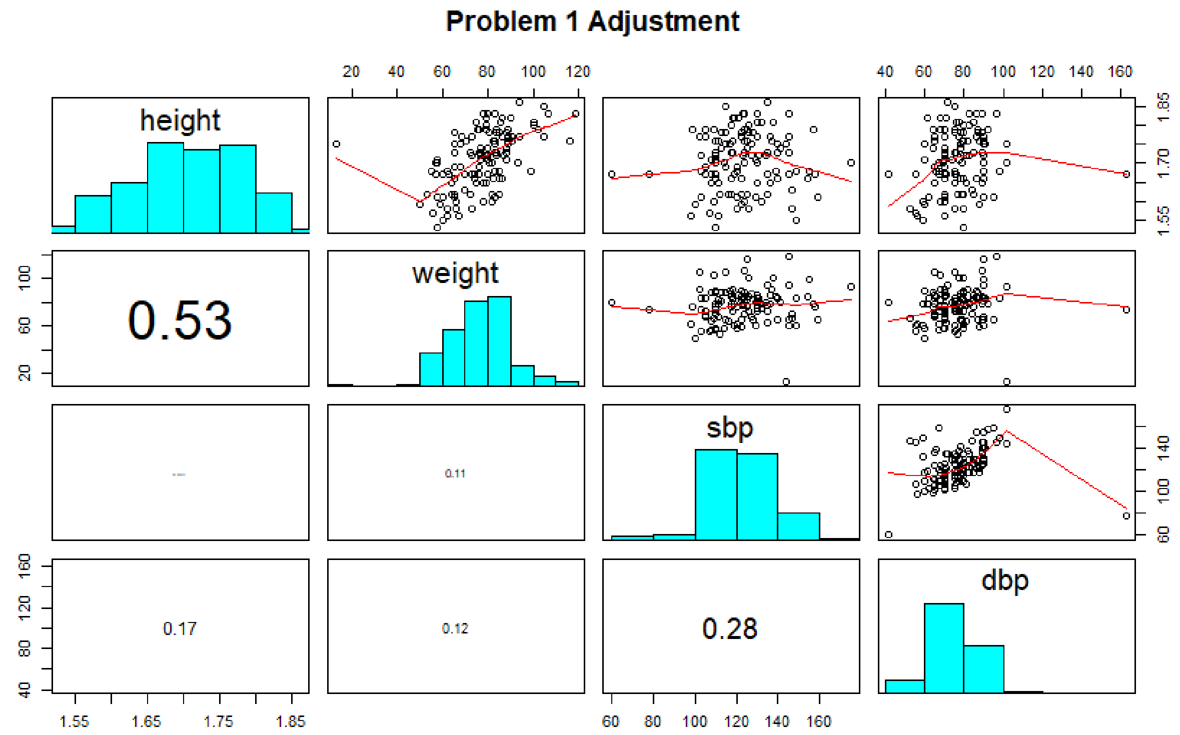
```r
#read in data
heartHealth.df = read.csv("hearthealth.csv")
#subset dataframe
heartHealth.cleaned.df = heartHealth.df[c("height", "weight", "sbp", "dbp")]

#initial pairs plot
pairs20x(heartHealth.cleaned.df, main="Initial Pairs Plot")
```



Problem 1- height of 15.1m

Looking at the height histogram in the pairs plot shows a single outlier (height = 15.1 metres). This height is not sensible, and I propose the decimal point was inputted in the wrong place. However, because I do not know how the data collection occurred, I have decided to change the outlier value to NA. The pairs plot below shows the effect of replacing the observation with NA.

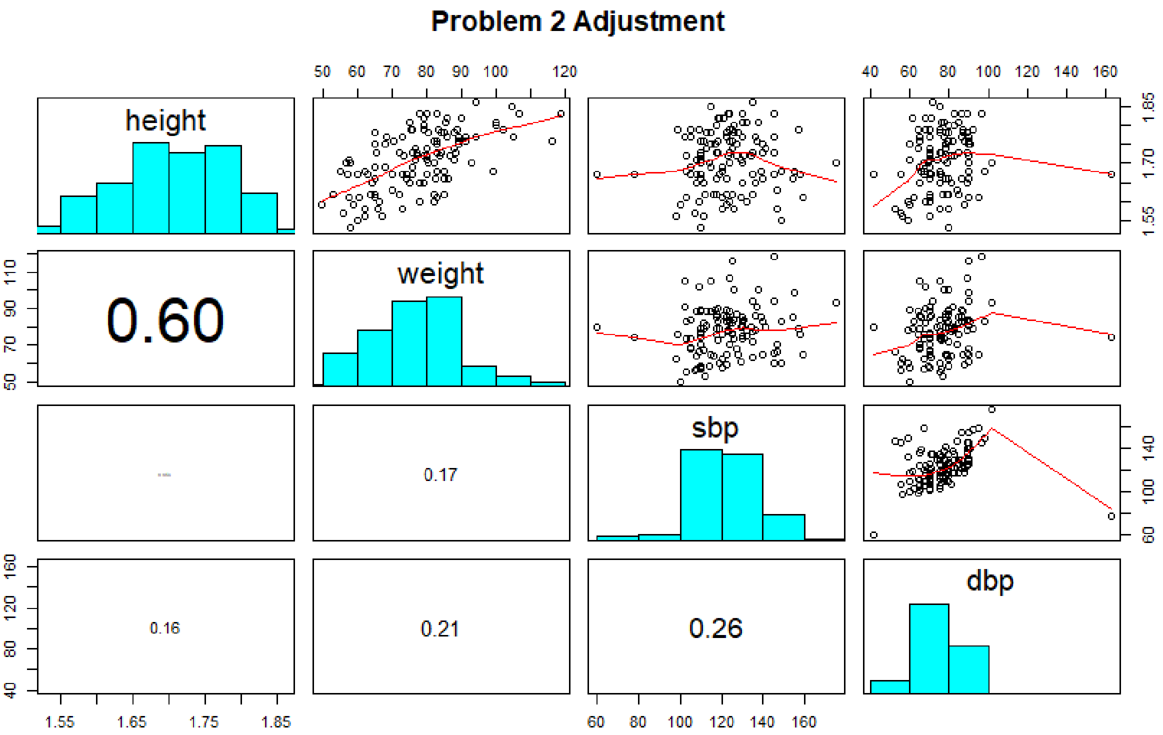
```
#adjustment 1
heartHealth.cleaned.df[heartHealth.cleaned.df["height"] == 15.1, "height"] = 
NA
pairs20x(heartHealth.cleaned.df, main="Problem 1 Adjustment")
```



Problem 2 – weight of 13.4kg

From the pairs plot you can see that weight has a single outlier (weight = 13.4kg). The subject is a European male of height 1.75m, therefore a weight of 13.4kg is infeasible. I believe this value is an error and have replaced it with NA. The pairs plot below shows the effect of replacing the observation with NA.

```
#adjustment 2 - add NA
heartHealth.cleaned.df[106,"weight"] = NA
pairs20x(heartHealth.cleaned.df, main="Problem 2 Adjustment")
```

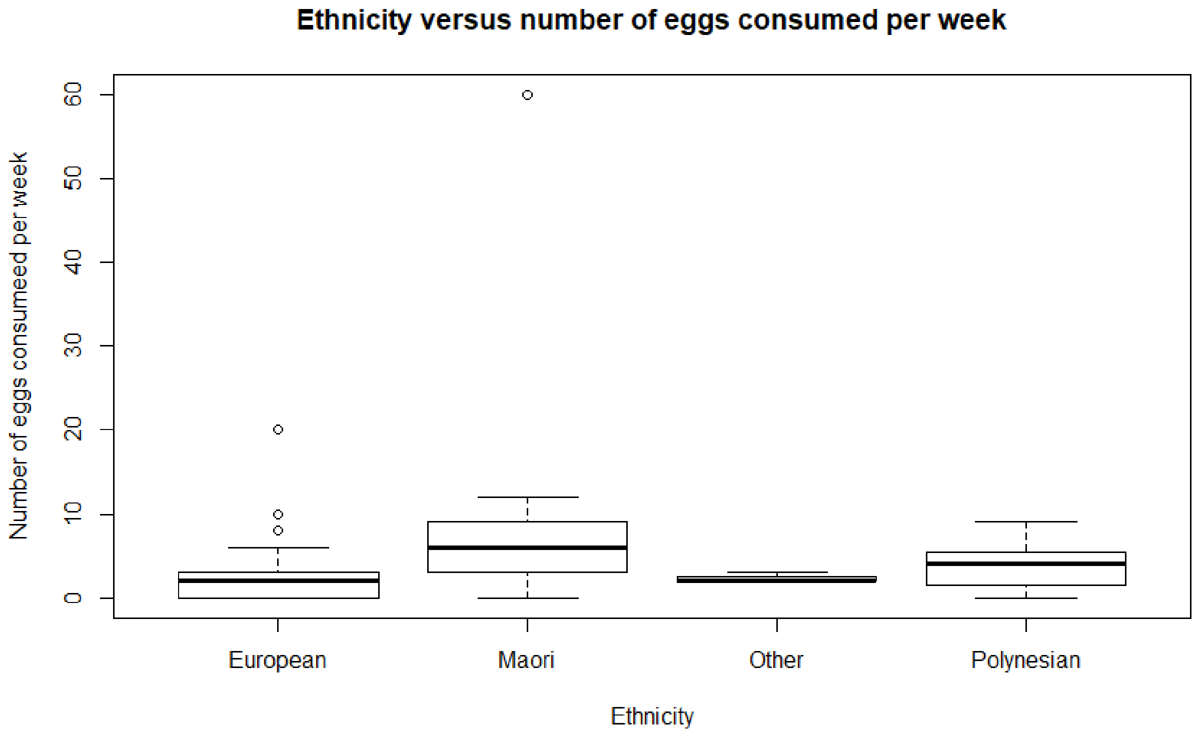


Problem 3 – dbp of 163 mm Hg

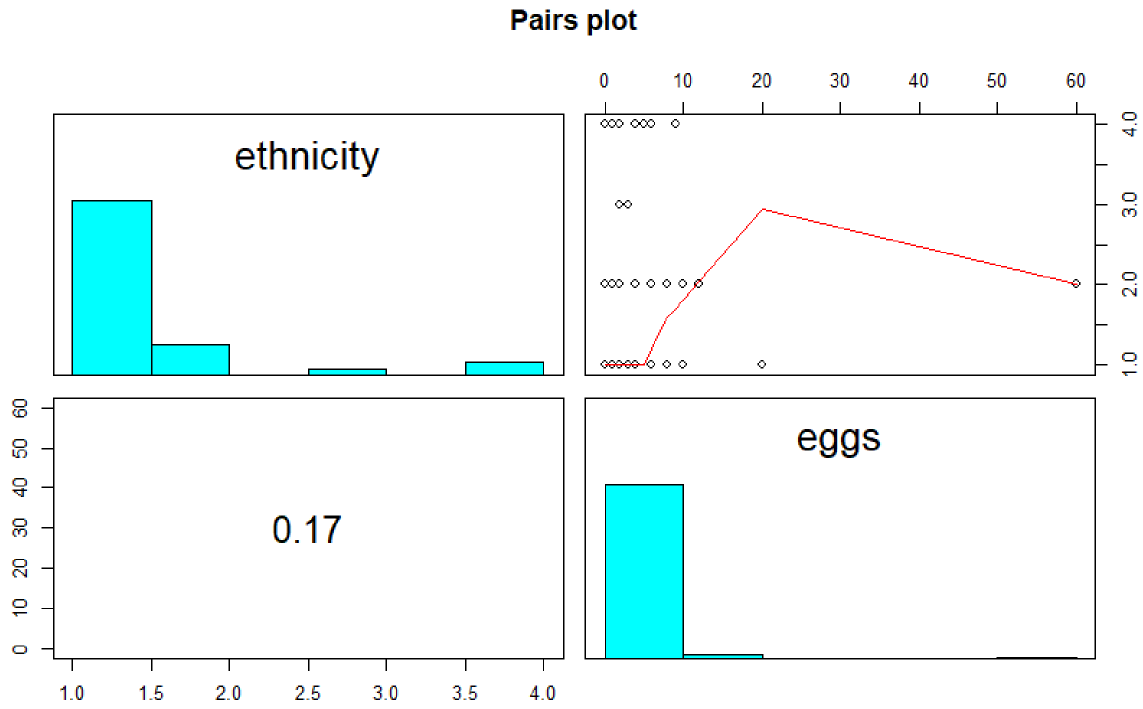
Looking at the histogram of dbp in the pairs plot above, there's a clear outlier of 163 mm Hg. I believe the dbp value of the participant was incorrectly entered during the data collection stage. However, because I cannot confirm how the data was collected nor am I a healthcare expert (no knowledge about reasonable dbp values and the value could relate to medical condition the participant had) I have elected to retain the value.

Eggs

```
#extract relevant fields and remove any NA's
eggs.df = heartHealth.df[,c("ethnicity", "eggs")]
eggs.df = eggs.df[complete.cases(eggs.df[, "eggs"]),]

#box plot
plot(eggs.df$ethnicity, eggs.df$eggs, main="Ethnicity versus number of eggs c
onsumed per week",
     xlab="Ethnicity", ylab="Number of eggs consumeed per week")
```



```
#pairs plot
pairs20x(eggs.df, main="Pairs plot")
```



Exploratory Analysis

Firstly, I did some initial data exploration creating a box plot showing ethnicity versus number of eggs consumed per week. The plot showed a clear outlier where a Maori participant typically consumed 60 eggs per week. I have decided to keep this outlier as even though it is quite large it is definitely within the realm of possibility. Furthermore, a pairs plot was created for exploratory analysis. The pairs plot showed that the typical number of eggs consumed per week is right skewed. It also showed that many there were many more European participants than other ethnicities.

```
#fit model
eggs.fit = glm(eggs ~ ethnicity, data=eggs.df, family="poisson")

#anova
anova(eggs.fit, test = "Chisq")

## Analysis of Deviance Table
##
## Model: poisson, link: log
##
## Response: eggs
##
## Terms added sequentially (first to last)
##
##
##            Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
## NULL                         117     591.34
## ethnicity  3   142.71       114     448.63 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

#summary
summary(eggs.fit)

##
## Call:
## glm(formula = eggs ~ ethnicity, family = "poisson", data = eggs.df)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -4.2426  -1.8742  -0.1699   0.4755  11.2096
##
## Coefficients:
##                      Estimate Std. Error z value Pr(>|z|)
## (Intercept)           0.81093    0.06950  11.667  < 2e-16 ***
## ethnicityMaori        1.38629    0.10851  12.776  < 2e-16 ***
## ethnicityOther        0.03637    0.38430   0.095  0.92461
## ethnicityPolynesian   0.53900    0.20462   2.634  0.00843 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 591.34  on 117  degrees of freedom
## Residual deviance: 448.63  on 114  degrees of freedom
## AIC: 727.5
##
## Number of Fisher Scoring iterations: 5
```

```
#confidence intervals
100*(exp(confint(eggs.fit)) -1)
```

```
## Waiting for profiling to be done...
```

```
##                          2.5 %   97.5 %
## (Intercept)           95.72387 157.0573
## ethnicityMaori       222.86037 394.2273
## ethnicityOther       -55.86501 103.7081
## ethnicityPolynesian   12.25882 151.1673
```

Model Overview

The response variable in the model fitted is the typical number of eggs consumed per week and the explanatory variable was the ethnicity of participants. The response variable is highly right skewed. Further the response variable is a count and therefore a Poisson regression model was fitted.

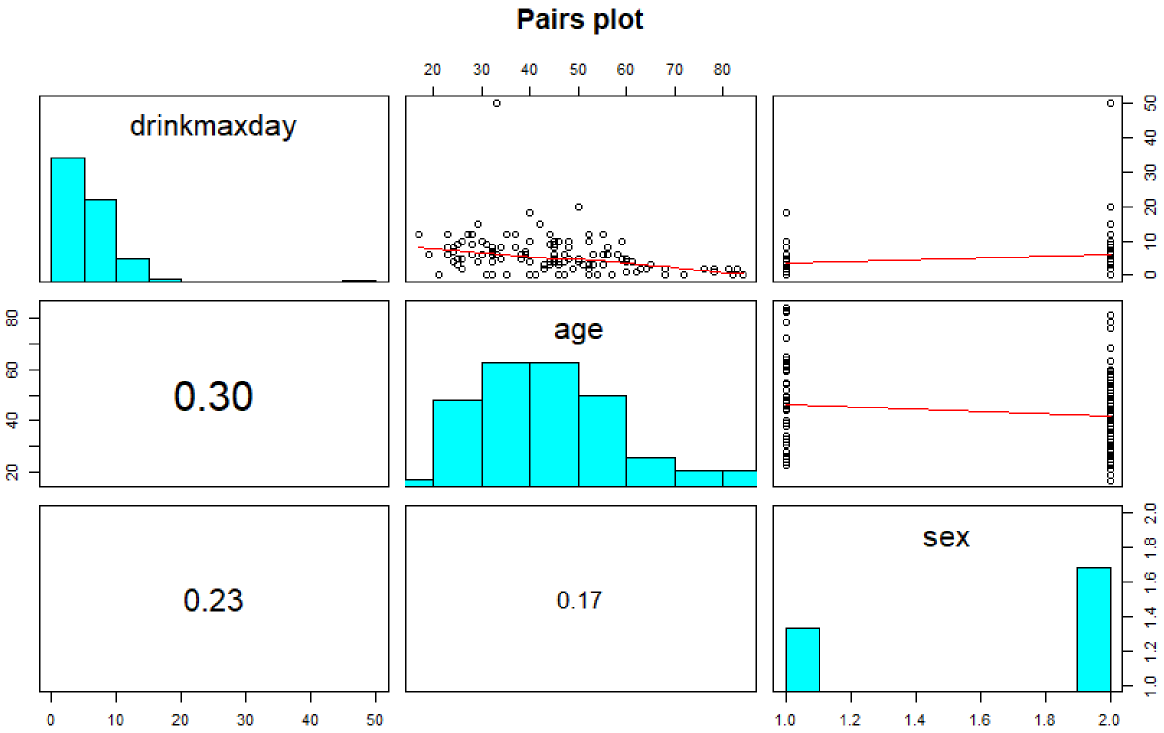
Model Interpretation

We were interested to answer the question whether the typical number of eggs consumed per week varied according to a person's ethnicity. We have strong evidence (p-value ≈ 0) to suggest that the typical number of eggs consumed per week does vary by ethnicity.

We have strong evidence (p-value ≈ 0) that the expected number of eggs consumed per week by Maori people is higher than European people. We estimate that the expected number of eggs consumed per week by a Maori person is 222.9% to 394.2% higher than a European person. We have evidence (p-value ≈ 0.008) that the expected number of eggs consumed per week by a Polynesian person is higher than a European person. We estimate that the expected number of eggs consumed per week by a Polynesian person is 12.3% to 151.2% higher than European person. We have no evidence (p-value ≈ 0.9) for a difference in the expected number of eggs consumed per week between European people and people of Other ethnicities.

Drinking

```
#subset dataframe extracting relevant field
drinking.df = heartHealth.df[, c("drinkmaxday", "age", "sex")]

pairs20x(drinking.df, main="Pairs plot")
```



Explanatory and Response Variables
The response variable for this analysis is the maximum number of alcoholic drinks a person consumes in a day (within the last 3 months) and the explanatory variables are the person's age and sex.

Model Overview
The response variable (drinkmaxday) is a count and thus cannot be negative. Furthermore, as can be seen from the pairs plot it is right skewed and so I have chosen to use Poisson regression. There is a outlier in the response variable of 50 drinks consumed in a day (within the last 3 months). However, I have decided to keep this outlier in my analysis as it could be feasible because we do not have any additional information on key factors such as size or alcohol type of the drinks consumed.

```
#fit model
drinks.fit1 = glm(drinkmaxday ~ age*sex, data=drinking.df, family="poisson")
#model summary
summary(drinks.fit1)

##
## Call:
## glm(formula = drinkmaxday ~ age * sex, family = "poisson", data = drinking
.df)
##
## Deviance Residuals:
##     Min       1Q    Median       3Q      Max
## -4.4430  -1.1539  -0.3180   0.5838  10.0326
##
## Coefficients:
```

```
##              Estimate Std. Error z value Pr(>|z|)    
## (Intercept)  2.474042   0.247327  10.003  < 2e-16 ***
## age         -0.024327   0.005548  -4.385 1.16e-05 ***
## sexM         0.212089   0.282507   0.751    0.453    
## age:sexM     0.005440   0.006463   0.842    0.400    
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## (Dispersion parameter for poisson family taken to be 1)
## 
##     Null deviance: 569.26  on 119  degrees of freedom
## Residual deviance: 474.67  on 116  degrees of freedom
## AIC: 844.69
## 
## Number of Fisher Scoring iterations: 5
```

```
#fit model
drinks.fit2 = glm(drinkmaxday ~ age+sex, data=drinking.df, family="poisson")
#model summary
summary(drinks.fit2)
```

```
## 
## Call:
## glm(formula = drinkmaxday ~ age + sex, family = "poisson", data = drinking
.df)
## 
## Deviance Residuals: 
##     Min       1Q   Median       3Q      Max  
## -4.5017  -1.1361  -0.2448   0.6126   9.9960  
## 
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)    
## (Intercept)  2.304301   0.145300  15.859  < 2e-16 ***
## age         -0.020351   0.002839  -7.167 7.66e-13 ***
## sexM         0.438832   0.090249   4.862 1.16e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## (Dispersion parameter for poisson family taken to be 1)
## 
##     Null deviance: 569.26  on 119  degrees of freedom
## Residual deviance: 475.39  on 117  degrees of freedom
## AIC: 843.4
## 
## Number of Fisher Scoring iterations: 5
```

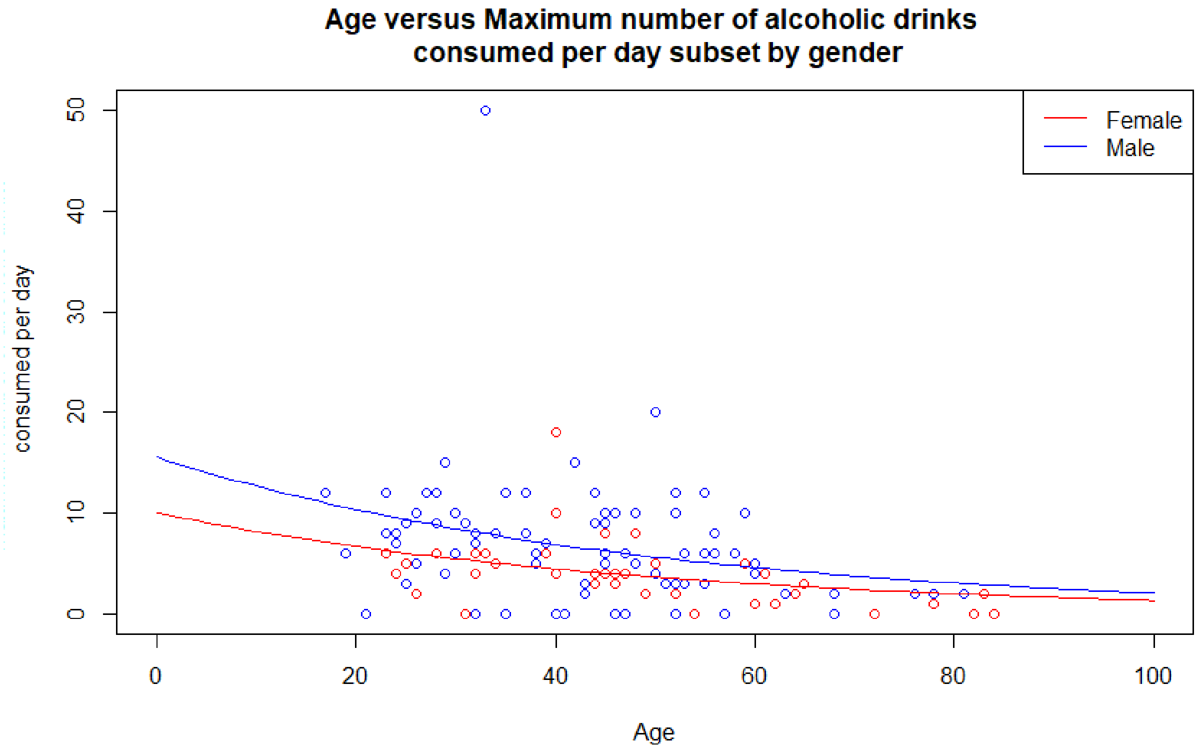
```
#confidence intervals
100*(exp(confint(drinks.fit2))-1)
```

```
## Waiting for profiling to be done...
##                  2.5 %      97.5 %
## (Intercept) 651.913430 1229.154739
## age          -2.562935   -1.472227
## sexM         30.281393   85.621145
```

Model Interpretation
We were interested assessing whether there is evidence to suggest that age and/or gender are related to the maximum number of alcoholic drinks consumed per day (within the last 3 months). We have strong evidence to suggest that age is related to the maximum number of drinks consumed per day (within the last 3 months). Further we have strong evidence that gender had an effect on the maximum number of drinks consumed per day (within the last 3 months). However, there was no evidence that the relationship between age and the maximum number of drinks consumed per day (within the last 3 months) depended on gender (and vice versa).

We estimate that for each additional year of age the expected number of the maximum number of alcoholic drinks consumed per day (within the last 3 months) decreases by 1.5% to 2.6%, regardless of gender. We estimate that, for the same age, the expected number of the maximum number of alcoholic drinks consumed per day (within the last 3 months) for male is 30.3% to 85.6% higher than a female.

```r
#create dataframe for predictiom
ages <- data.frame(age=seq(0,100), sex="M")
malePredictedValues <- predict(drinks.fit2, newdata = ages)
malePredictedValues <- exp(malePredictedValues)
ages <- data.frame(age=seq(0,100), sex="F")
femalePredictedValues <- predict(drinks.fit2, newdata = ages)
femalePredictedValues <- exp(femalePredictedValues)

#plot response variable
plot(drinkmaxday ~ age, data = heartHealth.df,col = ifelse(heartHealth.df$sex
== "M" , "blue", "red"), xlim=c(0, 100),
     main="Age versus Maximum number of alcoholic drinks \n consumed per day
subset by gender", xlab="Age", ylab="Maximum number of alcoholic drinks \n co
nsumed per day")
#plot model
lines(ages$age, femalePredictedValues, col="red")
lines(ages$age, malePredictedValues, col="blue")
legend("topright", legend=c("Female", "Male"),
       col=c("red", "blue"), lty=1)
```



Plot Interpretation
The plot shows that the expected value of maximum number of alcoholic drinks consumed per day (within the last 3 months) is higher for males than females for the same age. Further, the plot shows that as a person gets older the maximum number of alcoholic drinks consumed per day (within the last 3 months) is decreases for both males and females. This decreasing relationship is the same for both males and females.
The plot also shows us a clear outlier who is male, around 33 years old and has drunk a maximum of 50 alcoholic drinks per day (within the last 3 months).

```
#Point estimate males of ages 30 and 50
pred1.df = data.frame(age=c(30,50), sex=c("M", "M"))
prediction1.pred = predict(drinks.fit2, newdata = pred1.df, type = "response"
)

#point estimates male and female of age 40
pred2.df = data.frame(age=c(40,40), sex=c("M", "F"))
prediction2.pred = predict(drinks.fit2, newdata = pred2.df, type = "response"
)

prediction1.pred

##        1        2
## 8.436875 5.615880

prediction2.pred
```

```
##        1        2 
## 6.883348 4.438309
```

Estimate Interpretation
We estimate that the expected value of maximum number of alcoholic drinks consumed per day (within the last 3 months) for a 30-year-old male is 8.4 drinks. In comparison we estimate that the expected value of maximum number of alcoholic drinks consumed per day (within the last 3 months) for a 50-year-old male is 5.6 drinks.

We estimate the that the expected value of maximum number of alcoholic drinks consumed per day (within the last 3 months) for a 40-year-old male is 6.9 drinks. In comparison we estimate that the expected value of maximum number of alcoholic drinks consumed per day (within the last 3 months) for a 40-year-old female is 4.4 drinks.

Exercise

```
#subset dataframe to extract relevant fields
exercise.df = heartHealth.df[c("exermin", "exerhour", "exerday", "smoke", "ag
e")]

#create exerHours field
exerMinsInHours = exercise.df$exermin/60
exerHoursPerDay =  exercise.df$exerhour + exerMinsInHours
exerHoursPerWeek = exercise.df$exerday * exerHoursPerDay
exercise.df$exerHoursPerWeek = exerHoursPerWeek
```

Explanatory and Response Variables
In this situation the explanatory variables are age of the participant and smoking history of the participant. The response variable is the typical number of exercise hours per week. This variable is a combination of exermin, exerhour and exerday variables. The method of calculation is shown in the code chunk above.

```
#fit model
exer.fit1 = glm(exerHoursPerWeek ~ age*smoke, data=exercise.df, family="gauss
ian")
#model summary
summary(exer.fit1)

## 
## Call:
## glm(formula = exerHoursPerWeek ~ age * smoke, family = "gaussian", 
##     data = exercise.df)
## 
## Deviance Residuals: 
##     Min       1Q   Median       3Q      Max  
## -10.068   -4.035   -2.286    0.686   44.109  
## 
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)  
## (Intercept)     6.36621    3.76211   1.692   0.0933 .
## age            -0.09067    0.08471  -1.070   0.2867
## smokeYes      7.77353    5.06345   1.535   0.1275
## age:smokeYes -0.08635    0.11038  -0.782   0.4357
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 75.38409)
##
##     Null deviance: 9500.7  on 117  degrees of freedom
## Residual deviance: 8593.8  on 114  degrees of freedom
##   (2 observations deleted due to missingness)
## AIC: 850.87
##
## Number of Fisher Scoring iterations: 2
```

```
#fit model
exer.fit2 = glm(exerHoursPerWeek ~ age+smoke, data=exercise.df, family="gauss
ian")
#model summary
summary(exer.fit2)
```

```
##
## Call:
## glm(formula = exerHoursPerWeek ~ age + smoke, family = "gaussian",
##     data = exercise.df)
##
## Deviance Residuals:
##    Min      1Q  Median      3Q     Max
## -9.267  -4.315  -2.272   1.007  44.875
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  8.49572    2.59233   3.277  0.00139 **
## age         -0.14152    0.05422  -2.610  0.01025 *
## smokeYes     4.02595    1.63675   2.460  0.01539 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 75.12972)
##
##     Null deviance: 9500.7  on 117  degrees of freedom
## Residual deviance: 8639.9  on 115  degrees of freedom
##   (2 observations deleted due to missingness)
## AIC: 849.5
##
## Number of Fisher Scoring iterations: 2
```

```
#confidence intervals
confint(exer.fit2)
```

```
## Waiting for profiling to be done...
##                  2.5 %      97.5 %
## (Intercept)  3.4148570 13.57658732
## age         -0.2477918 -0.03525431
## smokeYes     0.8179667  7.23392379
```

Model Overview

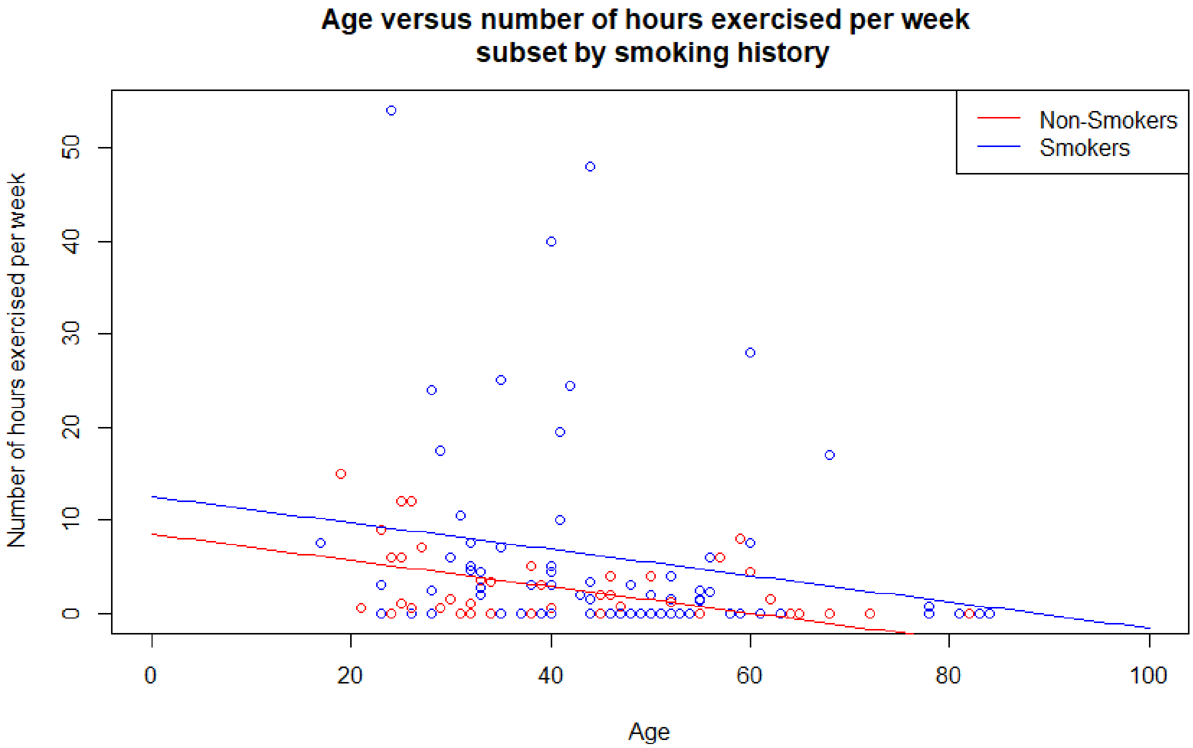
The response variable is numeric measure and thus linear regression has been fitted. One limitation of the modelling is that the response variable cannot be negative, but the model doesn't have this constraint.

Model Interpretation

We were interested in whether there is evidence that age and/or smoking history are related to the typical number of hours of exercise undertaken in a week. We have evidence to suggest that age is related to the typical number of hours of exercise undertaken in a week. Further, we have evidence that a person's smoking history is related to the typical number of hours of exercise undertaken in a week. However, we have no evidence that the relationship between age and the number of hours of exercise undertaken in a week depended on smoking history (and vice versa).

We estimate that for each additional year of age, the expected number of hours exercised per week decreases by between 0.04 hours and 0.25 hours, regardless of smoking history. We estimate that the expected number of hours exercised per week for people who have smoked once a week or more in the past, is between 0.8 to 7.2 hours greater than people with no smoking history, for the same age.

```
#create predictions
pred.df <- data.frame(age=seq(0,100), smoke="Yes")
smokersPredictions <- predict(exer.fit2, newdata = pred.df)
pred.df <- data.frame(age=seq(0,100), smoke="No")
nonsmokersPredictions <- predict(exer.fit2, newdata = pred.df)

#plot response
plot(exerHoursPerWeek ~ age, data = exercise.df, col = ifelse(exercise.df$smo
ke == "Yes" , "blue", "red"),xlim=c(0, 100),
     main="Age versus number of hours exercised per week \n subset by smoking
history", xlab="Age",
     ylab = "Number of hours exercised per week")
#plot model
lines(pred.df$age, nonsmokersPredictions, col="red")
lines(pred.df$age, smokersPredictions, col="blue")
legend("topright", legend=c("Non-Smokers", "Smokers"),
       col=c("red", "blue"), lty=1)
```



Plot Interpretation
The plot shows that the expected value of response variable is higher for people with a smoking history than people without a smoking history, for the same age. Further, the plot shows that as a person gets older, the expected number of hours exercised per week decreases for both people with/without a history of smoking. The plot also shows the rate of decrease is the same for non-smokers and smokers. The plot shows that there are a few large outliers, which are predominately people with a smoking history. This indicates that the model may not be a good fit. The plot also shows that the model can predict negative hours exercised per week under certain conditions. For example, if a person has no smoking history and is greater than 78 years old, our model will predict negative hours exercised per week.

Heart Attacks

```
#subset relevant data
heartattack.df = heartHealth.df[c("heartattack", "chol", "age")]

#fit model
heartattack.fit1 = glm(heartattack ~ age*chol, data=heartattack.df, family="b
inomial")
#model summary
summary(heartattack.fit1)

##
## Call:
## glm(formula = heartattack ~ age * chol, family = "binomial",
##     data = heartattack.df)
##
```

```
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -1.10547  -0.25130  -0.14950  -0.09596   2.53261
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -12.9650     9.1288  -1.420    0.156
## age           0.2063     0.1560   1.323    0.186
## cholLow       4.7798     9.4066   0.508    0.611
## age:cholLow  -0.1151     0.1597  -0.720    0.471
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 53.366  on 119  degrees of freedom
## Residual deviance: 34.657  on 116  degrees of freedom
## AIC: 42.657
##
## Number of Fisher Scoring iterations: 7
```

```
#fit model
heartattack.fit2 = glm(heartattack ~ age+chol, data=heartattack.df, family="binomial")
#model summary
summary(heartattack.fit2)
```

```
##
## Call:
## glm(formula = heartattack ~ age + chol, family = "binomial",
##     data = heartattack.df)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -1.00569  -0.23428  -0.13777  -0.07972   2.57403
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -6.89477    2.07903  -3.316 0.000912 ***
## age          0.10205    0.03351   3.046 0.002322 **
## cholLow     -1.99378    0.96266  -2.071 0.038349 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 53.366  on 119  degrees of freedom
## Residual deviance: 35.467  on 117  degrees of freedom
## AIC: 41.467
##
## Number of Fisher Scoring iterations: 7
```

```r
#confidence interval
100*(exp(confint(heartattack.fit2))-1)

## Waiting for profiling to be done...

##                  2.5 %     97.5 %
## (Intercept) -99.999128 -96.063922
## age           4.383572  19.667131
## cholLow     -98.216632  -7.210025
```

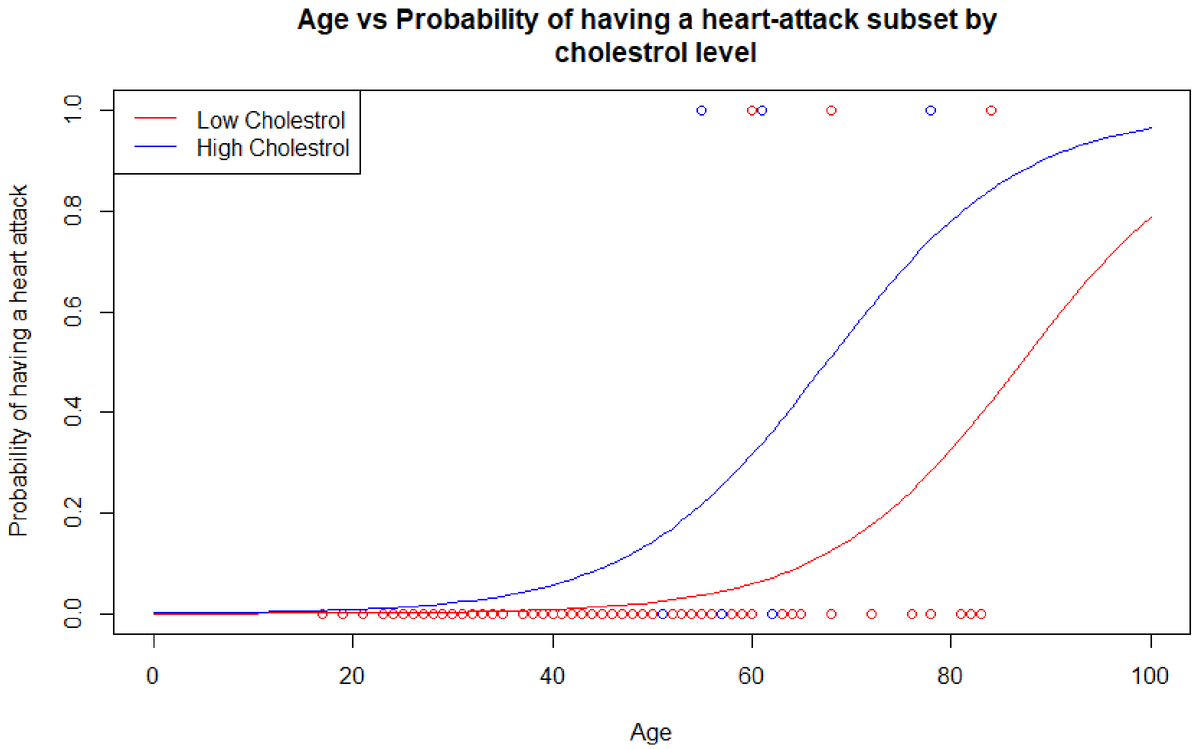
Explanatory and Response Variables
The response variable for this analysis is whether people have had heart attacks and the explanatory variables are age and cholesterol level of the person.

Model Overview
The response variable is a categorical variable with two levels; thus, it is binary. As a result of having a binary response variable a logistic regression model has been fit.

Model Interpretation
We were interested in whether there is evidence to suggest that age and/or cholesterol level status influence whether someone had a heart-attack or not. We have strong evidence that age had an effect on whether a person had a heart-attack. However, we had no evidence that this relationship depended on cholesterol level status. Likewise, we had strong evidence that cholesterol level status had an effect on whether a person had a heart-attack. However, we have no evidence that this relationship depended on age.

We estimate that for every additional year of age, the odds of having a heart-attack increase between 4.4% to 19.7%, regardless of cholesterol level status. We estimate that the odds of people who have low cholesterol having a heart-attack are 7.2% to 98.2% lower than people who have high cholesterol, for the same age.

```r
#high cholestrol predictions
pred.df <- data.frame(age=seq(0,100), chol="High")
highCholPredictions <- predict(heartattack.fit2, newdata = pred.df, type="res
ponse")
#low cholestrol predictions
pred.df <- data.frame(age=seq(0,100), chol="Low")
lowCholPredictions <- predict(heartattack.fit2, newdata = pred.df, type="resp
onse")

plot(heartattack ~ age, data = heartattack.df, col = ifelse(heartattack.df$ch
ol == "High" , "blue", "red"),xlim=c(0,100), ylim=c(0,1),
     main="Age vs Probability of having a heart-attack subset by \n cholestro
l level", xlab="Age", ylab="Probability of having a heart attack")
lines(pred.df$age, lowCholPredictions, col="red")
lines(pred.df$age, highCholPredictions, col="blue")
legend("topleft", legend=c("Low Cholestrol", "High Cholestrol"),
       col=c("red", "blue"), lty=1)
```



Plot Interpretation

The plot shows that the probability of having a heart-attack for people with high cholesterol is higher than people with low cholesterol at the same age. Further, the plot shows that typically as a person gets older, the probability of having a heart-attack increases, regardless of cholesterol level. The plot also highlights that there are very few people in the study (6 observations out of 120 in the sample) that had heart attacks.