MAS6061 Project 2

student number 170138286

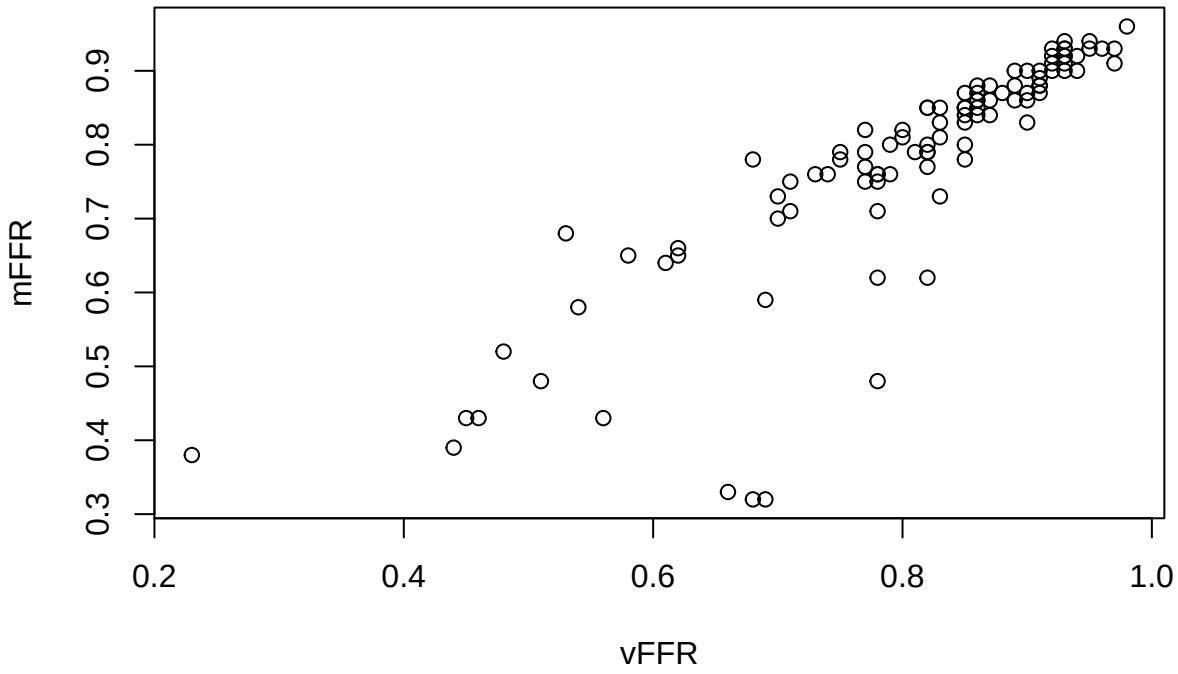
Question 1

For patients with stable heart disease, visual assessment of a coronary angiogram is a poor estimator of the physiological significance of coronary artery lesions. Fractional flow reserve (FFR) is an invasive medical procedure which measures the degree of obstruction to flow in conditions of hyperaemia. This measurement is reported as a proportion and its values must lie between 0 and 1. However, in clinical practice more than 95% of patients do not have an FFR measurement (mFFR). So research teams have developed computer models which estimate FFR (virtual, or vFFR) from angiographic images. One model (SISA) has been used on a consecutive series of 101 patients where the vFFR was also measured. (Project2.xls sheet 1) You are asked to analyse these data for agreement and clinical utility.

1a

What is your assessment of the vFFR SISA measurement as a possible alternative to the presumed best standard mFFR? A plot of mFFR and vFFR is shown below

```
FFR <- read.table("G:\\Masters Sheffield\\MAS6061\\Project_2\\project-2-1.csv",
                  header = T, sep=",")
plot(FFR$vFFR,FFR$mFFR, xlab="vFFR", ylab = "mFFR")
```

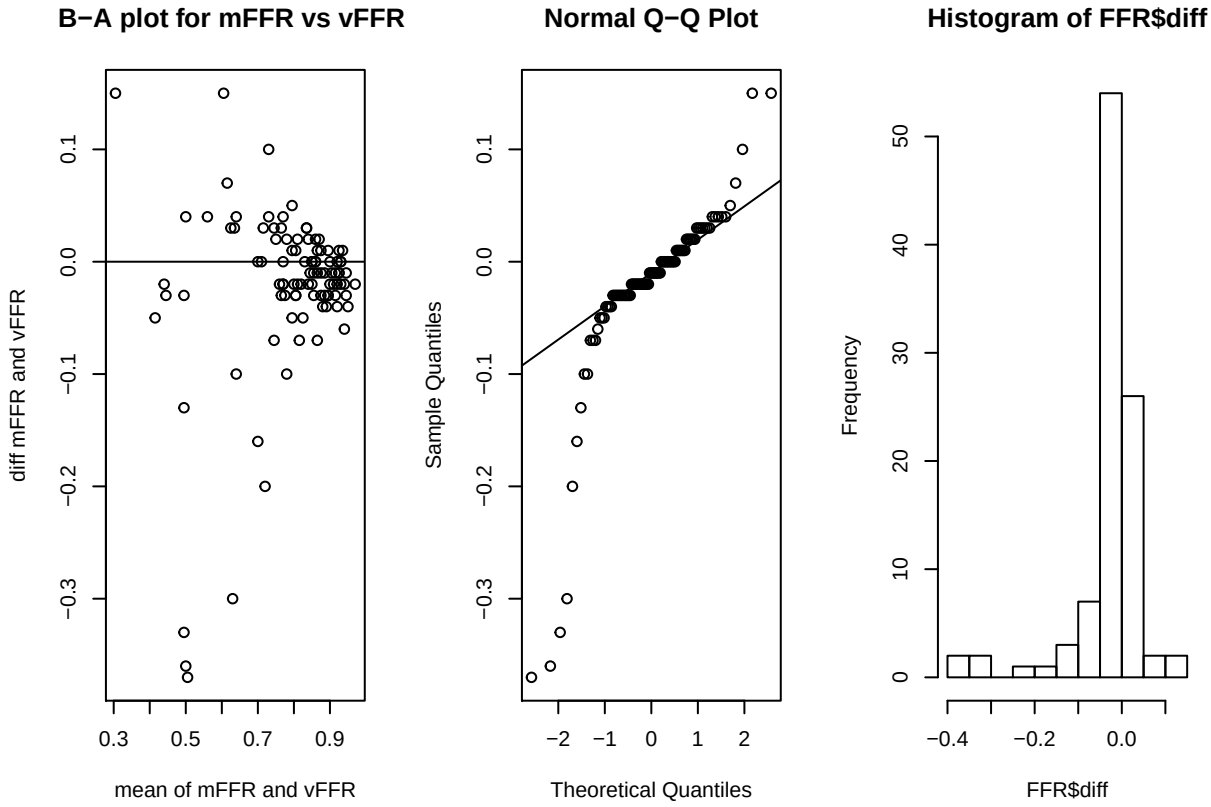


A linear relationship between mFFR and vFFR appears to exist on values of $vFFR > 0.7$, although several

data pairs above this value appear as outliers. These outliers have a much lower value of vFFR compared to mFFR. The range mFFR (0.32, 0.96) was less than vFFR (0.23, 0.98).

The Bland-Altman (B-A) plot for the FFR data (measured as a proportion), the associated normal plot and the R code are shown below

```
FFR$diff <- FFR$mFFR - FFR$vFFR
FFR$avg <- (FFR$mFFR + FFR$vFFR)/2
normlm <- lm(FFR$diff~ FFR$avg)
par(mfrow=c(1,3))
plot(FFR$avg, FFR$diff, xlab = "mean of mFFR and vFFR", ylab = "diff mFFR and vFFR",
     main = "B-A plot for mFFR vs vFFR ")
abline(h=0)
qqnorm(FFR$diff)
qqline(FFR$diff)
hist(FFR$diff)
```

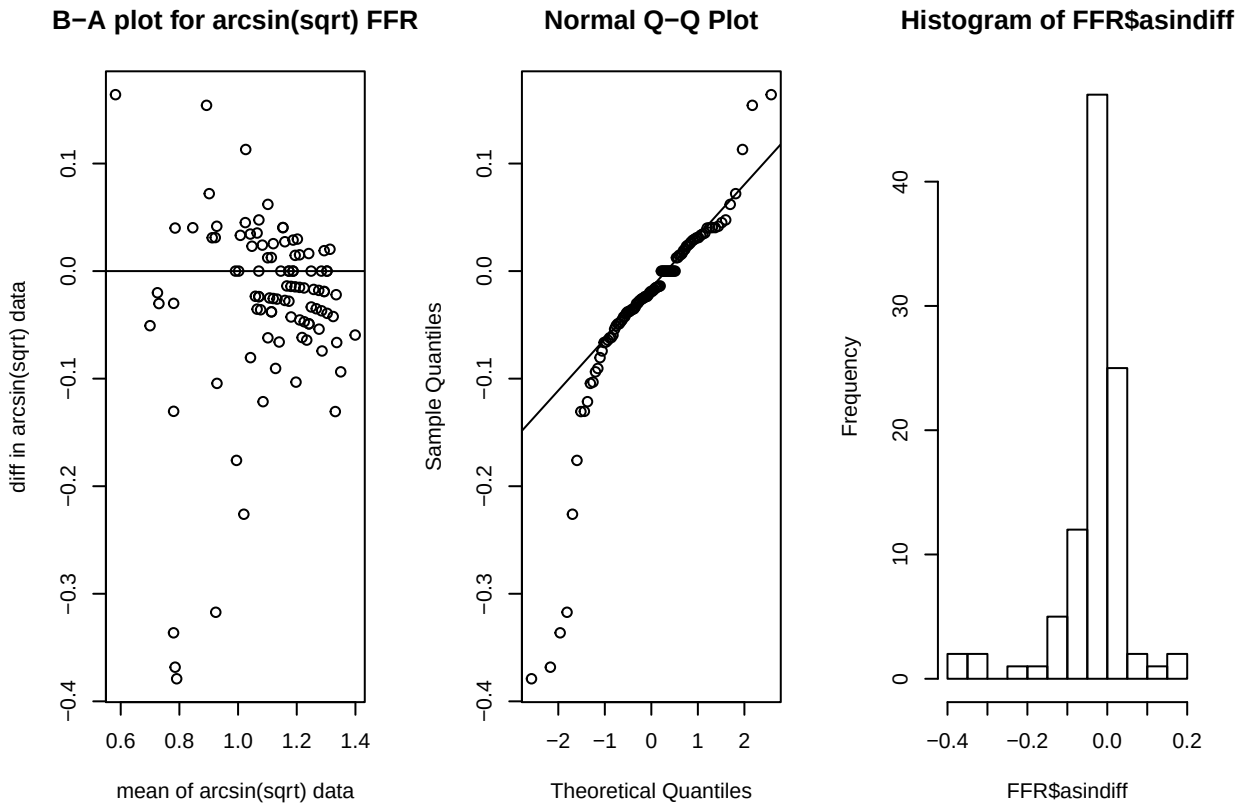


In order to substitute one method of measurement for another, we require that the bias and variability between the two methods are constant (Bland and Altman, 1999). It can be seen that the variance appears to decrease on the B-A plot with increasing average value of the two methods. A normal plot of the residuals confirms that the data are not normally distributed, as there are many points at both tails of the plot away from the line of normality. A quantitive test for normality on the differences of paired data was performed using the Shapiro-Wilk test and found to be statistically significant ($W=0.72$, $p<0.001$), confirming the conclusion that the data are not normally distributed. As the variability between the two methods is not constant, assessment for bias (agreement) is inconsequential.

As the data are in the form of proportions, and considering that the variance appears to decrease with increasing mean value of the methods, an arcsine transformation on the data was performed, using the

formula $y = sin^{-1}(\sqrt{(x)})$. The B-A plot of the transformed mFFR and vFFR values and R code is shown below.

```
FFR$asinmFFR <- asin(sqrt(FFR$mFFR))
FFR$asinvFFR <- asin(sqrt(FFR$vFFR))
FFR$asindiff <- FFR$asinmFFR - FFR$asinvFFR
FFR$asinavg <- (FFR$asinmFFR + FFR$asinvFFR)/2
par(mfrow=c(1,3))
plot(FFR$asinavg, FFR$asindiff, xlab = "mean of arcsin(sqrt) data",
     ylab = "diff in arcsin(sqrt) data", main = "B-A plot for arcsin(sqrt) FFR")
abline(h=0)
qqnorm(FFR$asindiff)
qqline(FFR$asindiff)
hist(FFR$asindiff)
```



The B-A plot for the transformed data is similar to that of the untransformed data. The decreasing variance of the residuals with increasing mean of the transformed FFR measurements has changed little and the data are still not normally distributed (Shapiro-Wilks, W=0.80, p<0.001).

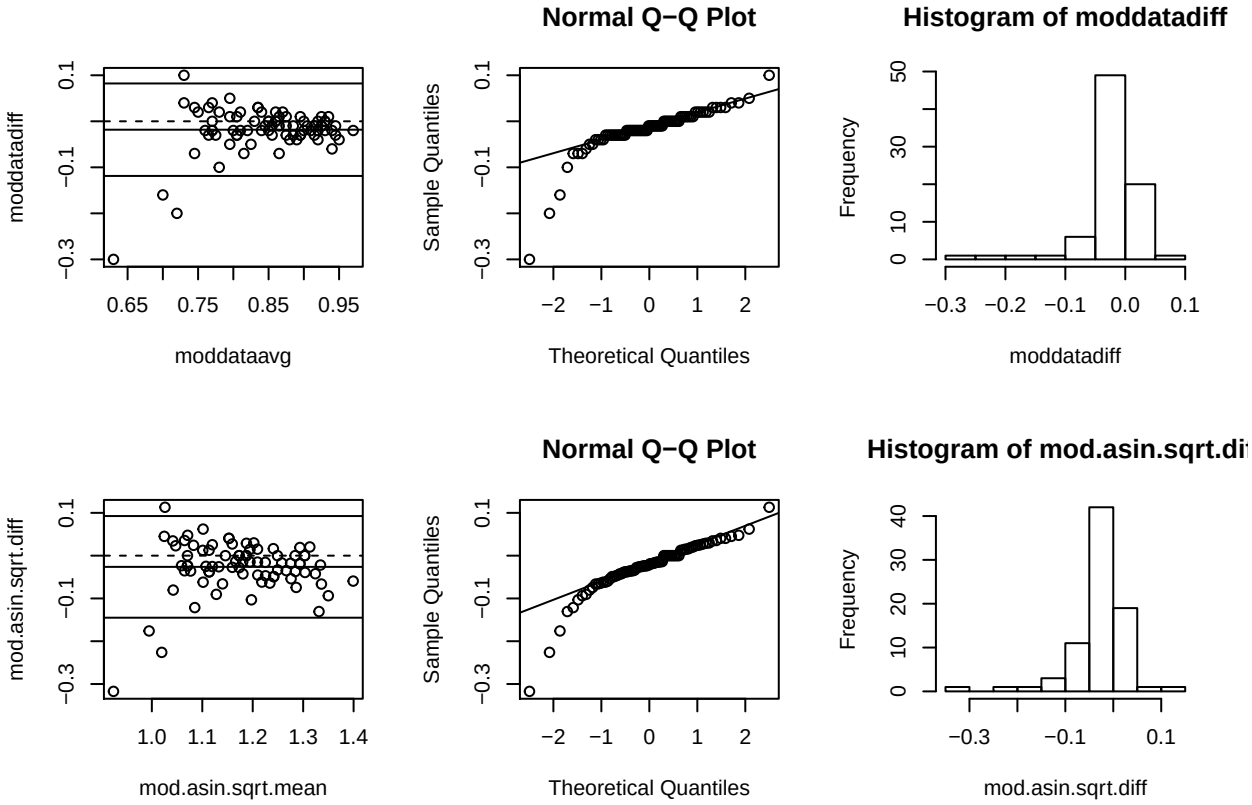
1b

In reality the mFFR is used to make clinical decisions. In prior clinical evaluations a mFFR above 0.75 would make a patient likely to benefit from a STENT operation. Hence the critical threshold of 0.75 is often used in practice. In the light of this new information examine the impact (positives and negatives) of using the vFFR SISA where the mFFR is not available to guide clinical decisions to performing a STENT operation.

Data was included in the analysis if either pair of mFFR or vFFR had a value of 0.75 or greater. The B-A

and Q-Q plots for the restricted model with and without mFFR and vFFR arcsine transformation and the R code are shown below.

```
par(mfrow=c(2,3))
rows <-  which((FFR$mFFR>= 0.75)|(FFR$vFFR>= 0.75))
moddata <- FFR[c(rows),]
moddatadiff<- moddata$mFFR-moddata$vFFR
moddataavg <- (moddata$mFFR+moddata$vFFR)/2
plot(moddataavg,moddatadiff)
abline(h=0, lty=2)
abline(h=mean(moddatadiff))
abline(h=mean(moddatadiff)+1.96*sd(moddatadiff))
abline(h=mean(moddatadiff)-1.96*sd(moddatadiff))
qqnorm(moddatadiff)
qqline(moddatadiff)
hist(moddatadiff)
mod.asinmFFR <- asin(sqrt(moddata$mFFR))
mod.asinvFFR <- asin(sqrt(moddata$vFFR))
mod.asin.sqrt.diff<- mod.asinmFFR-mod.asinvFFR
mod.asin.sqrt.mean <- (mod.asinmFFR+mod.asinvFFR)/2
plot(mod.asin.sqrt.mean, mod.asin.sqrt.diff)
abline(h=0, lty=2)
abline(h=mean(mod.asin.sqrt.diff))
abline(h=mean(mod.asin.sqrt.diff)+1.96*sd(mod.asin.sqrt.diff))
abline(h=mean(mod.asin.sqrt.diff)-1.96*sd(mod.asin.sqrt.diff))
qqnorm(mod.asin.sqrt.diff)
qqline(mod.asin.sqrt.diff)
hist(mod.asin.sqrt.diff)
```



The B-A plot of the restricted data appears to demonstrate a variance that is more constant. The histogram of the untransformed data demonstrates a negatively skewed histogram, and the normal plot indicates that the data are not normally distributed due to the three extreme points on the left hand tail. The Shapiro-Wilk test is significant (W=0.74, p<0.001) indicating the data is not normally distributed. The mean difference is -0.02 with a 95% confidence interval (-0.15, 0.08). The bias appears independent of mean mFFR/vFFR measurements.

The data for the transformed data appears similar to untransformed data for the restricted dataset. The data is not normally distributed (Shapiro-Wilk test, W=0.85, p<0.05). The bias is -0.03 with a 95% confidence interval (-0.15, 0.09). The bias appears to be decrease with increasing mean mFFR/vFFR.

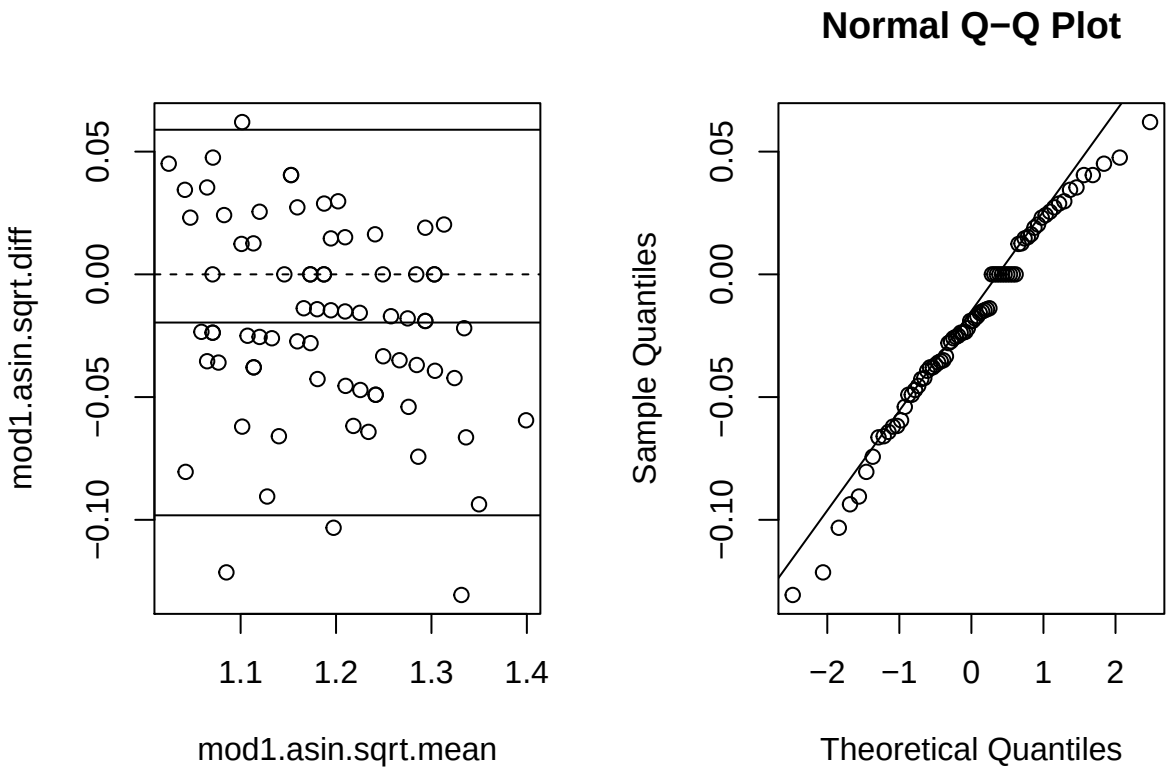
Of these two datasets, I would accept the use of the untransformed dataset, because the bias is independent of the mean mFFR/vFFR. Apart from the few outliers at low values of mean mFFR/vFFR, the data is relatively homoscedastic and normally distributed. Ideally I would investigate the four outliers shown on the untransformed B-A plot to determine if they can be dropped from the B-A plot. These outliers have the coordinates (0.48, 0.78), (0.62, 0.78), (0.62, 0.82) and (0.78, 0.68).

The variation between mFFR and vFFR can be quite large up to vFFR values of 0.82, and as low as 0.48 for mFFR values. Using vFFR as a surrogate measurement for mFFR can therefore result in occasionally inaccurate mFFR values. Our measuremnents on untransformed data demonstrate that mFFR values will be within (-0.15, 0.09) of vFFR measurements on 95% of occasions. For vFFR values of greater than 0.82, the data suggests this 95% confidence interval will be narrower producing more precise estimates of mFFR values from vFFR measurements. If this level of reproducibility is acceptable then the technique can be implemented. The negatives of not using this technique include a lack of mFFR estimation, an important measurement to determine clinical outcomes of patients.

To further the analysis of the agreement between mFFR and vFFR the B-A and Q-Q plots after the outliers were dropped for the transformed data, and are shown below with the R code.

```r
rows <-  which((FFR$mFFR>= 0.75)|(FFR$vFFR>= 0.75))
moddata <- FFR[c(rows),]
moddata1 <- moddata[-c(1,11,12,28),] # remove 4 points to produce normally distributed B-A plot
par(mfrow=c(1,2))
mod1.asinmFFR <- asin(sqrt(moddata1$mFFR))
mod1.asinvFFR <- asin(sqrt(moddata1$vFFR))
mod1.asin.sqrt.diff<- mod1.asinmFFR-mod1.asinvFFR
mod1.asin.sqrt.mean <- (mod1.asinmFFR+mod1.asinvFFR)/2
plot(mod1.asin.sqrt.mean, mod1.asin.sqrt.diff)
abline(h=mean(mod1.asin.sqrt.diff))
abline(h=0, lty=2)
abline(h=mean(mod1.asin.sqrt.diff)+1.96*sd(mod1.asin.sqrt.diff))
abline(h=mean(mod1.asin.sqrt.diff)-1.96*sd(mod1.asin.sqrt.diff))

qqnorm(mod1.asin.sqrt.diff)
qqline(mod1.asin.sqrt.diff)
```



Having removed the three outliers from the transformed dataset with mFFR and/or vFFR values$\geq$ 0.75, the data appears normally distributed (W=0.99 p>0.05, Shapiro-Wilks test). The bias is -0.02 and the 95% confidence interval is (-0.1, 0.06). The 95% confidence interval has reduced from (-0.15, 0.09), confirming that the four outliers should be dropped from the data if justified.

Question 2

2a

a) What factors (variables) seem to influence which patients are offered a CT scan on arrival? Is the data consistent with hospitals adhering to the recommendation using CGS<14 and a report of vomiting?

A general linear model was fitted to the data and a summary of the model and R code is displayed below.

```
df <- read.csv("G:\\Masters Sheffield\\MAS6061\\Project_2\\project-2-2.csv")
CT <- df
CT$gender <- factor(CT$gender, labels = c("M","F"))
CT$amnesia <- factor(CT$amnesia, levels=c("2", "1"), labels = c("N", "Y"))
CT$vomiting <- factor(CT$vomiting, levels=c("2","1"), labels = c("N","Y"))
CT$consciousness_loss <- factor(CT$consciousness_loss, levels=c("2", "1"), labels = c("N","Y"))
CT$headache <- factor(CT$headache, levels=c("2", "1"), labels = c("N","Y"))
CT$CT.performed <- factor(CT$CT.performed, levels= c("0", "1"), labels = c("N","Y") )
CT$OutcomeCode <- factor(CT$OutcomeCode, levels= c("0", "1"), labels = c("Good","Poor"))
###full model  . denotes all independent vars ####
CTfm <- glm(CT.performed~age+gender+inr+GCS.admission+amnesia+vomiting+consciousness_loss+headache,
            family = binomial, data = CT)
summary(CTfm) # inr, GCS.admission, amnesia, vomiting, consciousness_loss : p<0.05
```

```
##
## Call:
## glm(formula = CT.performed ~ age + gender + inr + GCS.admission +
##     amnesia + vomiting + consciousness_loss + headache, family = binomial,
##     data = CT)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -4.1302  -1.2547   0.6932   1.0347   1.2155
##
## Coefficients:
##                       Estimate Std. Error z value Pr(>|z|)
## (Intercept)           8.664523   1.868703   4.637 3.54e-06 ***
## age                  -0.007413   0.003980  -1.862 0.062548 .
## genderF               0.084248   0.084582   0.996 0.319228
## inr                   0.120074   0.036255   3.312 0.000927 ***
## GCS.admission        -0.545626   0.122551  -4.452 8.50e-06 ***
## amnesiaY              0.434651   0.131345   3.309 0.000936 ***
## vomitingY             0.443371   0.183202   2.420 0.015515 *
## consciousness_lossY   0.817092   0.124891   6.542 6.05e-11 ***
## headacheY             0.106982   0.107512   0.995 0.319699
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 3545.9  on 2724  degrees of freedom
## Residual deviance: 3341.1  on 2716  degrees of freedom
## AIC: 3359.1
##
## Number of Fisher Scoring iterations: 6
```

In the above general linear model, CT.performed takes a value of 1 if performed and 0 if not performed. Significant variables which increase the probability that a CT will be performed are an increase in INR, a decrease in GCS.admission value, the patient having amnesia, the patient having vomited, and the patient having loss of consciousness.

An exposure-specific risk table was calculated for patients having a CT with a GCS<14.

```
#recodes GCS.admissions{<14} and vomiting{"Y"} <- 1 otherwise 0
recode.ct <- function(x,y){
  ifelse((x<14 & y=="Y"),"CT_recommended","No_CT_recommended")
}

CT$CT.performed <- factor(CT$CT.performed, levels= c("Y", "N"), labels = c("Y","N") )
CT$combined <- factor(recode.ct(CT$GCS.admission, CT$vomiting))
CTROC <- table((CT$CT.performed), CT$combined)

dimnames(CTROC) = list(c("CT  performed" , "CT not performed "),
                       c("vomiting & GCS<14", "not(vomiting & GCS<14)"))
CTROC
```

```
##                   vomiting & GCS<14 not(vomiting & GCS<14)
## CT  performed                    24                   1733
## CT not performed                  1                    967
```

```
sens <- CTROC[1,1]/sum(CTROC[,1])
spec <- CTROC[2,2]/sum(CTROC[,2])
cat("\n")
```

```
cat("Sensitivity is ", sens,"\n")
```

```
## Sensitivity is  0.96
```

```
cat("Specificity is ", round(spec, digits = 3),"\n")
```

```
## Specificity is  0.358
```

It can be seen that 96% of patients with vomiting and GCS<14 score received a CT scan. Therefore this hospital was adhering to the recommendation to perform a CT scan on patients with GCS<14 score and having an incidence of vomiting (sensitivity =0.96). However, around 2 out of 3 patients underwent a CT scan when there wasn't both an incidence of vomiting and GCS<14 score (sensitivity = 0.36). This suggests that doctors often believed a patient could have experienced a stroke even if the patient was not vomiting and had a GCS<14 score.

2b

Which variables are associated with a poor outcome within 6 weeks? For the continuous/numerical variables is it fair to assume they show a linear relationship with risk of the poor outcome?

Performing a binary logit regression using "Outcome_code" as the dependent variable and age, gender, INR, GCS, amnesia, vomiting, loss-consciousness and headache as co-variates produced the following results

```
OUTCOME <- df[,c(1:8, 10)]
OUTCOME$gender <- factor(OUTCOME$gender, labels = c("M","F"))
OUTCOME$amnesia <- factor(OUTCOME$amnesia, level=c(2,1), labels = c("N", "Y"))
OUTCOME$vomiting <- factor(OUTCOME$vomiting, levels=c("2","1"), labels = c("N","Y"))
OUTCOME$consciousness_loss <- factor(OUTCOME$consciousness_loss, levels=c("2", "1"),
                                     labels = c("N","Y"))
```

```r
OUTCOME$headache <- factor(OUTCOME$headache, levels=c("2", "1"), labels = c("N","Y"))
#outcome.interaction <-
#   glm(OutcomeCode~age*gender*inr*GCS.admission*amnesia*vomiting*consciousness_loss*headache
                          # , family = binomial, data = df)
outcome.fm <- glm(OutcomeCode~., family = binomial, data = OUTCOME)
summary(outcome.fm) # only gender and inr not significant variables in model
```

```
##
## Call:
## glm(formula = OutcomeCode ~ ., family = binomial, data = OUTCOME)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.1266  -0.3668  -0.3062  -0.2732   2.6926
##
## Coefficients:
##                    Estimate Std. Error z value Pr(>|z|)
## (Intercept)       -0.641001   0.855737  -0.749   0.4538
## age                0.015836   0.007464   2.122   0.0339 *
## genderF           -0.245527   0.161725  -1.518   0.1290
## inr                0.067583   0.050422   1.340   0.1801
## GCS.admission     -0.260171   0.038771  -6.711 1.94e-11 ***
## amnesiaY           0.570500   0.196572   2.902   0.0037 **
## vomitingY          0.458183   0.229226   1.999   0.0456 *
## consciousness_lossY 0.452065   0.191585   2.360   0.0183 *
## headacheY          0.457389   0.181954   2.514   0.0119 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1399.1  on 2724  degrees of freedom
## Residual deviance: 1252.3  on 2716  degrees of freedom
## AIC: 1270.3
##
## Number of Fisher Scoring iterations: 5
```

The results demonstrate that the variables which are associated with a poor outcome within 6 weeks are increasing age, decreasing GCS score and presence of amnesia, vomiting, consciousness loss and headache.

The continuous/numerical variables are age, inr and GCS score. INR is not associated with a poor outcome within 6 weeks.
The results of a Hosmer and Lemeshow GOF test for age is

```r
#install ResourceSelection package
age.fm <- glm(OutcomeCode~age, family = binomial, data = OUTCOME)
OUTCOME$age.pred.outc <- predict(age.fm, type = c("response"))
#hoslem.test(OUTCOME$OutcomeCode, OUTCOME$age.pred.outc)
```

As p>0.05, there is no evidence to suggest that age is not linearly related to the risk of a poor outcome. Likewise, the results of the Hosmer and Lemeshow GOF test for GCS score is

```r
#install ResourceSelection package
GCS.fm <- glm(OutcomeCode~GCS.admission, family = binomial, data = OUTCOME)
OUTCOME$GCS.pred.outc <- predict(GCS.fm, type = c("response"))
#hoslem.test(OUTCOME$OutcomeCode, OUTCOME$GCS.pred.outc)
```

p>0.05, there no evidence to suggest that GCS score is not linearly related to the risk of a poor outcome.

2c

Build a multivariable predictive model (using the Likelihood Ratio Test to select optimal model with alpha=5%) to predict a poor outcome for patients presenting at an emergency department following a head injury when taking anticoagulants. In this application we are keen to find an optimal clinical decision tool which will identify those patients who will not likely benefit from a CT scan. Hence we would weight sensitivity more valuable than specificity in this context. Explore the calibration of your 'best' model and report the sensitivity and specificity.

A full model was constructed using the variables associated with a poor outcome within 6 weeks, as described in question 2b, with the addition of vomiting, as recommended in the guidelines. This model was compared to the full model and found that the deviance change on 2df was not significant (p>0.05). Therefore the smaller model was considered more appropriate. Furthermore, as age had a p value of 0.051 in the smaller model, the effect of removing age from the smaller model was investigated. The results of the two likelihood tests performed are shown below

```
outcome.alpha_05<- glm(OutcomeCode~age+GCS.admission+amnesia+vomiting+consciousness_loss+headache,
                       family = binomial, data = OUTCOME)
summary(outcome.alpha_05)
```

```
##
## Call:
## glm(formula = OutcomeCode ~ age + GCS.admission + amnesia + vomiting +
##     consciousness_loss + headache, family = binomial, data = OUTCOME)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.0846  -0.3640  -0.3012  -0.2811   2.7209
##
## Coefficients:
##                       Estimate Std. Error z value Pr(>|z|)
## (Intercept)         -0.283678   0.828238  -0.343  0.73197
## age                  0.014094   0.007244   1.946  0.05170 .
## GCS.admission       -0.270359   0.038362  -7.048 1.82e-12 ***
## amnesiaY             0.572245   0.195775   2.923  0.00347 **
## vomitingY            0.446289   0.228748   1.951  0.05106 .
## consciousness_lossY  0.469870   0.191022   2.460  0.01390 *
## headacheY            0.432107   0.180910   2.389  0.01692 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1399.1  on 2724  degrees of freedom
## Residual deviance: 1256.1  on 2718  degrees of freedom
## AIC: 1270.1
##
## Number of Fisher Scoring iterations: 5
```

```
anova(outcome.alpha_05, outcome.fm, test="Chi")
```

```
## Analysis of Deviance Table
```

```
##
## Model 1: OutcomeCode ~ age + GCS.admission + amnesia + vomiting + consciousness_loss +
##     headache
## Model 2: OutcomeCode ~ age + gender + inr + GCS.admission + amnesia +
##     vomiting + consciousness_loss + headache
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1      2718     1256.1
## 2      2716     1252.3  2    3.855   0.1455
```

```
outcome.alpha_05.noage<- glm(OutcomeCode~GCS.admission+amnesia+vomiting+consciousness_loss+headache,
                             family = binomial, data = OUTCOME)
summary(outcome.alpha_05.noage)
```

```
##
## Call:
## glm(formula = OutcomeCode ~ GCS.admission + amnesia + vomiting +
##     consciousness_loss + headache, family = binomial, data = OUTCOME)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.0279  -0.3527  -0.2854  -0.2854   2.5381
##
## Coefficients:
##                      Estimate Std. Error z value Pr(>|z|)
## (Intercept)           0.88349    0.57101   1.547  0.12181
## GCS.admission        -0.27092    0.03813  -7.105  1.2e-12 ***
## amnesiaY              0.57458    0.19496   2.947  0.00321 **
## vomitingY             0.43500    0.22785   1.909  0.05624 .
## consciousness_lossY   0.43426    0.18967   2.290  0.02205 *
## headacheY             0.40482    0.17926   2.258  0.02392 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1399.1  on 2724  degrees of freedom
## Residual deviance: 1260.2  on 2719  degrees of freedom
## AIC: 1272.2
##
## Number of Fisher Scoring iterations: 5
```

```
anova(outcome.alpha_05.noage, outcome.alpha_05, test="Chi")
```

```
## Analysis of Deviance Table
##
## Model 1: OutcomeCode ~ GCS.admission + amnesia + vomiting + consciousness_loss +
##     headache
## Model 2: OutcomeCode ~ age + GCS.admission + amnesia + vomiting + consciousness_loss +
##     headache
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1      2719     1260.2
## 2      2718     1256.1  1   4.0695  0.04366 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

From the likelihood ratio test it can be seen that removal of INR and gender from the full model doesn't

make a statistically significant contribution (at the 5% level) to the fit, allowing these to covariates to be dropped. However, removing the "age" covariate does produce a significant result (p=0.04) and therefore the age covariate must remain in the model. The optimal model to predict outcome for patients presenting at an emergency department following a head injury while on anticoagulants is therefore

glm(OutcomeCode~age+GCS.admission+amnesia+vomiting+consciousness_loss+headache, family = binomial).

The AUROC plot for the aforementioned optimal model are shown below.

```
#install "pROC and "ResourceSelection" packages
library(pROC)
```

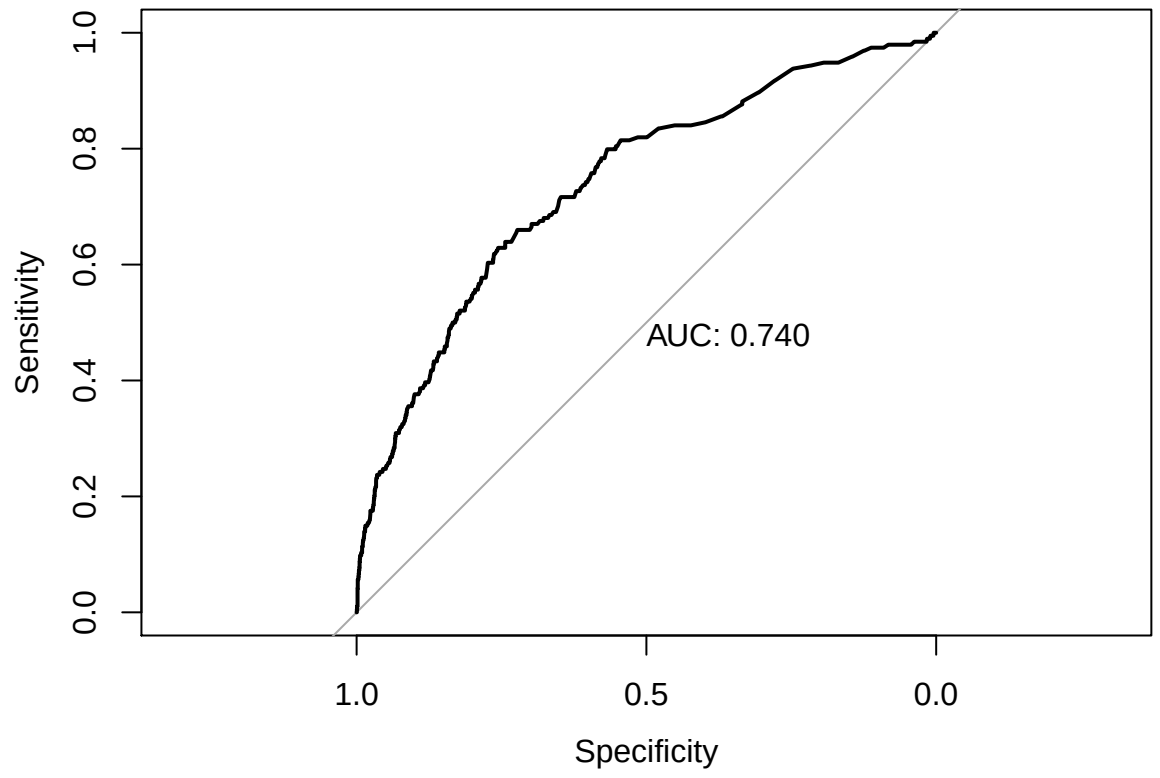
```
## Warning: package 'pROC' was built under R version 3.4.3
```

```
## Type 'citation("pROC")' for a citation.
```

```
##
## Attaching package: 'pROC'
```

```
## The following objects are masked from 'package:stats':
##
##     cov, smooth, var
```

```
OUTCOME$prob1 <- predict(outcome.alpha_05, type = c("response")) # calculates predicted response for ea
roc_plot <- plot(roc(OUTCOME$OutcomeCode, OUTCOME$prob1), print.auc=TRUE)
```




```
roc_plot
```

```
##
## Call:
```

```
## roc.default(response = OUTCOME$OutcomeCode, predictor = OUTCOME$prob1)
##
## Data: OUTCOME$prob1 in 2531 controls (OUTCOME$OutcomeCode 0) < 194 cases (OUTCOME$OutcomeCode 1).
## Area under the curve: 0.74
```

```
roc_head_injury <- roc(OUTCOME$OutcomeCode, OUTCOME$prob1)
Youden <- coords(roc_head_injury, "best")
closest.top.left <- coords(roc_head_injury, "best", best.method = "closest.topleft")
#Spielhalter test statistic
num <- (OUTCOME$OutcomeCode-OUTCOME$prob1)*(1-2*OUTCOME$prob1)
dem <- (OUTCOME$OutcomeCode-2*OUTCOME$prob1)^2*OUTCOME$prob1*(1-OUTCOME$prob1)
Spiel.optimal <- sum(num)/sqrt(sum(dem))
```

The optimal threshold for the ROC curve using the Youden method is

```
Youden
```

```
##   threshold specificity sensitivity
##  0.06624903  0.75582774  0.62886598
```

The optimal threshod for the ROC curve using the closest.top.left method is

```
closest.top.left
```

```
##   threshold specificity sensitivity
##  0.06395979  0.72263927  0.65979381
```

From these methods the closest.top.left method produces the highest sensitivity value of 0.66 at a threshold 0.064, with a specificity of 0.72.

The calibration for the optimal model ws performed using the Hosmer and Lemeshow GOF test. The results of which are shown below.

```
library(ResourceSelection)
```

```
## Warning: package 'ResourceSelection' was built under R version 3.4.3
```

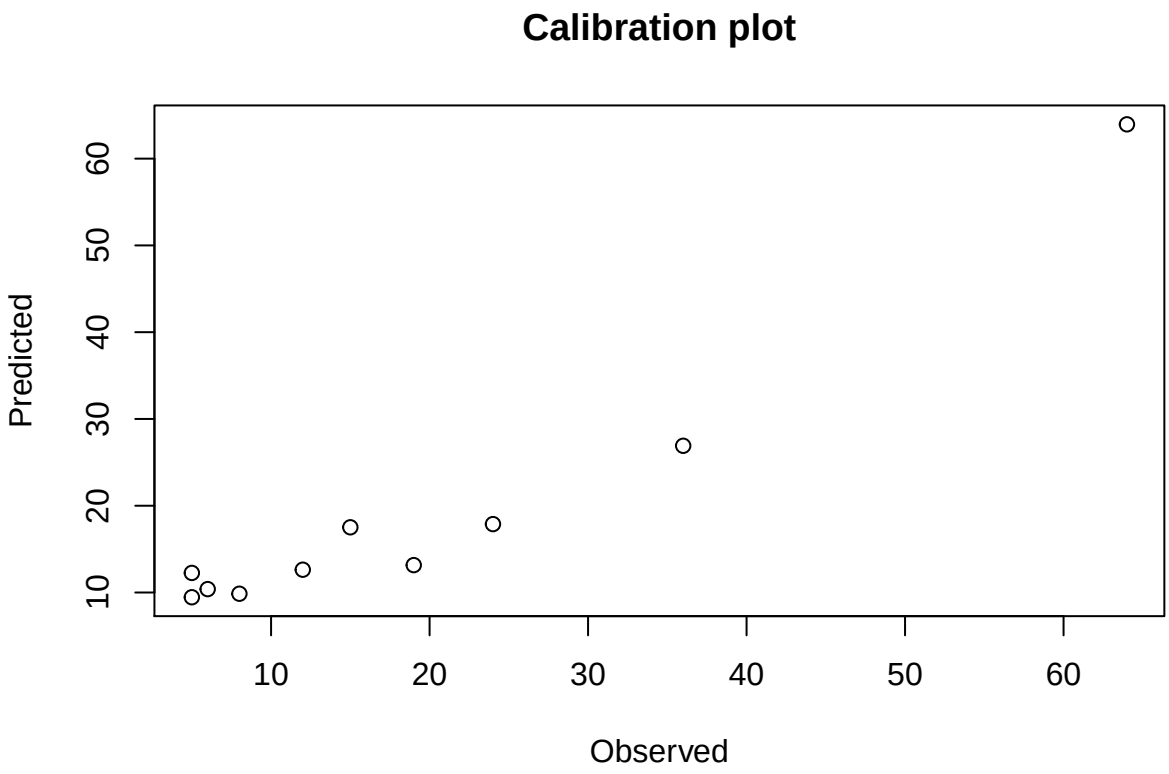
```
## ResourceSelection 0.3-2   2017-02-28
```

```
htest <- hoslem.test(OUTCOME$OutcomeCode, fitted(outcome.alpha_05), g=10)
htest
```

```
##
##  Hosmer and Lemeshow goodness of fit (GOF) test
##
## data:  OUTCOME$OutcomeCode, fitted(outcome.alpha_05)
## X-squared = 17.774, df = 8, p-value = 0.02298
```

The value of the Hosmer and Lemeshow goodness of fit (GOF) test is significant ($p<0.05$), suggesting the data is no following a linear fit. The calibration plot is shown below

```
obs <- htest$observed[,2]
exp<- htest$expected[,2]
plot(obs, exp,  main="Calibration plot", xlab="Observed", ylab="Predicted")
```



The calibration plot appears to follow a linear relationship apart from the outlier at observation point 63. The overall calibration using the spielhalter test (normally distributed under H_0) is calculated as 0.82, which is not statistically significant (P>0.05), suggesting there is no evidence that overall calibration is poor.

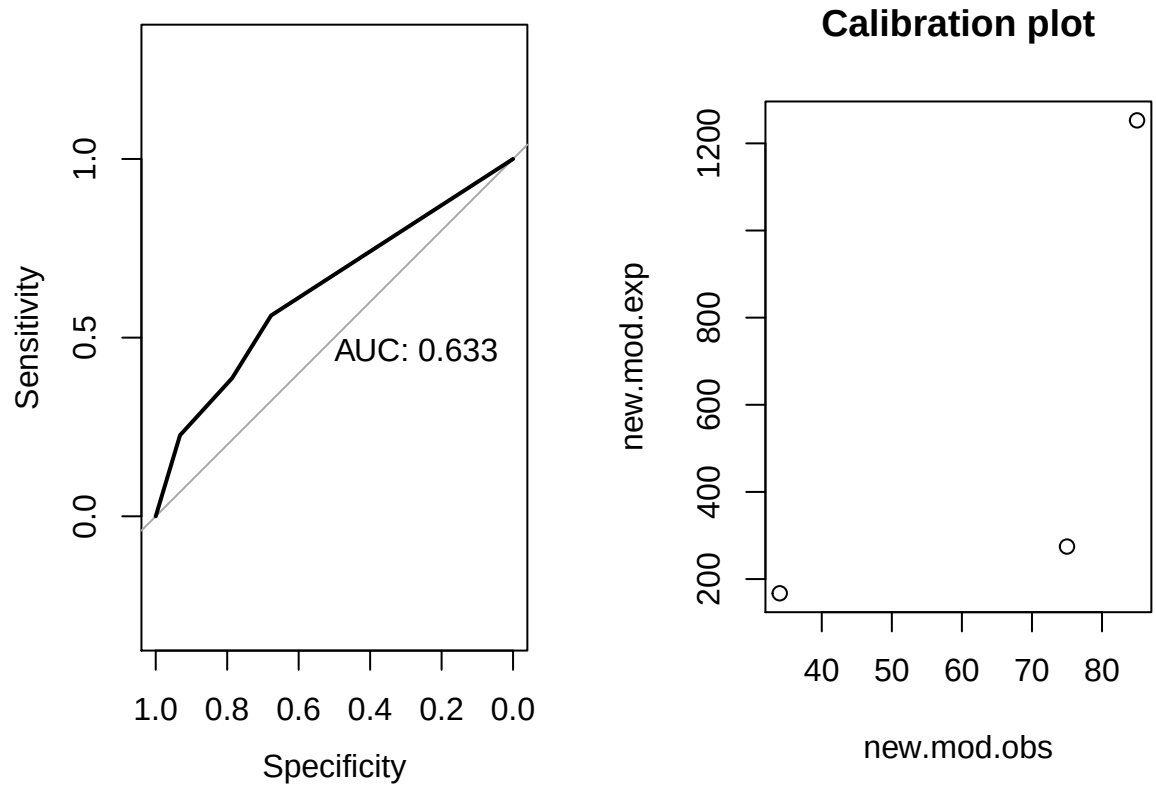
Q2.4

Another research group has developed a predictive model for a similar situation but limited to predict risk of a poor outcome where GCS score is high (i.e. 15). Their equation for the logit(risk)=0.83 - 0.82 x headache -0.67 x amnesia. Their result is somewhat controversial as they do not include 'vomiting' as a predictor in this model. How does this predictor perform (calibration and discrimination) on the relevant subset of the study in sheet 2?

The predictors amnesia and headache in study sheet 2 were entered into a new database and converted so that the prescence of amnesia or headache were given a score of 1, otherwise the score was set at 0. The logit predictor outcome was calaculated, and from this the probability of the outcome was calculated (predict variable).

```
par(mfrow=c(1,2))
new.model.df <- df[,c(5,8,10)]
new.model.df$amnesia[new.model.df$amnesia==2] <- 0
new.model.df$headache[new.model.df$headache==2] <- 0
new.model.df$pred.logit <- 0.83 - 0.82*new.model.df$headache - 0.67*new.model.df$amnesia
new.model.df$exp.lf<- exp(0.83 - 0.82*new.model.df$headache - 0.67*new.model.df$amnesia)
new.model.df$predict <- new.model.df$exp.lf/(1+new.model.df$exp.lf)
#pROC and ResourceSelection packages required
plot(roc(new.model.df$OutcomeCode, new.model.df$predict), print.auc=TRUE)
```

```
htest.new <- hoslem.test(new.model.df$OutcomeCode, new.model.df$predict, g=10)
#create calibration plot
new.mod.exp <- htest.new$expected[,2]
new.mod.obs <- htest.new$observed[,2]
plot(new.mod.obs, new.mod.exp, main="Calibration plot")
```




```
#plot(htest$observed, htest$expected)
```

Because the two variables used in the model are binary, there are only four possible values for the predictor variable. The Hosmer and Lemeshow GOF test creates three "bins" of data. Therefore only three points for the calibration plot are created and, as one can see, the plot is not linear. The result of the Hosmer and Lemeshow GOF test show a statistical significance, indicating that the observed values of outcomes do not follow the expected values of outcomes ($p<0.001$)

```
htest.new
```

```
##
##  Hosmer and Lemeshow goodness of fit (GOF) test
##
## data:  new.model.df$OutcomeCode, new.model.df$predict
## X-squared = 4077.6, df = 8, p-value < 2.2e-16
```

Because of the lack of "bins" used for the Hosmer and Lemeshow GOF test, the Spielhalter calibration test was also performed.

```
new.model.df$spiel.num <- (new.model.df$OutcomeCode-new.model.df$predict)*
  (1-2*new.model.df$predict)
new.model.df$spiel.dem <- (new.model.df$OutcomeCode-2*new.model.df$predict)^2*
```

```
  new.model.df$predict*(1-new.model.df$predict)
spiel.controv <- sum(new.model.df$spiel.num)/sqrt(sum(new.model.df$spiel.dem))
```

The Speilhalter statistic was calculated as 15.37, which is statistically significant (p<0.001), indicating that there is a lack of overall calibration.

The AUC of the "controversial" model is less (0.633) than for the optimal model in question 2c (0.74), indicating that the optimal model's better ability to correctly identify patients with poor and good outcomes.

In conclusion, the controversial model performs worse than the optimal model in question 2, and the controversial model displays a lack of overall calibration. Therefore the optimal model should be used in favour of the controversial model.

Question 3

Perform standard quality control checking on the genotyping of the 7 candidate genes measured in this cross sectional study. State all of your assumptions and declare which (if any) patient samples or SNPs need to be discarded for quality control reasons.

Rows with call rate errors less than 0.6 (row 21) where identified and gt fields with call rate errors per SNP less than 0.95 (gt6) were removed. These rows and columns were removed before determining where the distribution of genotypes is not consistent with HWE (with two-sided significance=0.001).

```
gdf <- read.csv("G:\\Masters Sheffield\\MAS6061\\Project_2\\project-2-3.csv")
#idenitfy call rate errors per sample to be removed
row.error.prop <- apply(gdf[7:13], 1, function(x)  sum(!is.na(x))/(length(x)))
cat("Call rate error sample <0.6 is row", which(row.error.prop<0.6),"\n")
```

```
## Call rate error sample <0.6 is row 21
```

```
#identify call rate errors per SNP to be removed
col.error.prop <- apply(gdf[7:13], 2, function(x)  sum(!is.na(x))/(length(x)))
cat("Call rate error SNP <0.95 is column", which(col.error.prop<0.95))
```

```
## Call rate error SNP <0.95 is column 6
```

```
#remove rows and columns with poor quality data
gdf <- gdf[-21,] #row 21 removed
gdf <- gdf[,-12] # gt6 removed
```

Next the SNPs were determined where the distribution of genotypes is not consistent with HWE (with two-sided significance=0.001).

```
#function to determine if SNP follows HWE distribtuion
test.stat <- function(fname){
  #calculates genotype frequency per field
  fvector <- fname[!is.na(fname)]
  obs <- matrix(c(0,0,0), nrow=3)
  obs[1] <- length(fvector[fvector==0])
  obs[2] <- length(fvector[fvector==1])
  obs[3] <- length(fvector[fvector==2])

  #estimate pop allele freq
  total.genotypes <- sum(obs)
  total.a <- total.genotypes*2
  prop.big.a <- (2*obs[1]+obs[2])/total.a
```

```
  prop.small.a <- (2*obs[3]+obs[2])/total.a

  #expectation v0-v2
  exp <- matrix(c(0,0,0), nrow=3)
  exp[1] <- prop.big.a^2*total.genotypes
  exp[2] <- 2*prop.big.a*prop.small.a*total.genotypes
  exp[3] <- prop.small.a^2*total.genotypes

  oe.table <- as.table(cbind(obs,exp))

  chi.stat <- chisq.test(oe.table, correct = FALSE)

  print(exp) #demonstrate that expected cell value >5 to ensure Chi-test is appropriate
  print(chi.stat)
}
#determine SNP's that don't follow HWE distribution
gt.analyses <- apply(gdf[7:12], 2, test.stat)
```

```
##          [,1]
## [1,] 955.3063
## [2,] 765.3874
## [3,] 153.3063
##
##  Pearson's Chi-squared test
##
## data:  oe.table
## X-squared = 0.00060045, df = 2, p-value = 0.9997
##
##          [,1]
## [1,] 443.0972
## [2,] 935.8057
## [3,] 494.0972
##
##  Pearson's Chi-squared test
##
## data:  oe.table
## X-squared = 79.473, df = 2, p-value < 2.2e-16
##
##          [,1]
## [1,] 473.0134
## [2,] 935.9733
## [3,] 463.0134
##
##  Pearson's Chi-squared test
##
## data:  oe.table
## X-squared = 4.6637, df = 2, p-value = 0.09712
##
##          [,1]
## [1,] 651.6668
## [2,] 905.6664
## [3,] 314.6668
##
##  Pearson's Chi-squared test
```

```
##
## data:  oe.table
## X-squared = 4.7568, df = 2, p-value = 0.0927
##
##        [,1]
## [1,] 566.17
## [2,] 926.66
## [3,] 379.17
##
##  Pearson's Chi-squared test
##
## data:  oe.table
## X-squared = 5.9148, df = 2, p-value = 0.05195
##
##           [,1]
## [1,] 1235.9649
## [2,]  571.0702
## [3,]   65.9649
##
##  Pearson's Chi-squared test
##
## data:  oe.table
## X-squared = 4.0592, df = 2, p-value = 0.1314
```

```
#removed gt2 as pvalue<0.001
gdf <- gdf[,-8]
```

It can be seen that only gt2 SNP doesn't follow a HWE distribution and is therefore removed from further analyses.

3.2

Examine the association between the SNPs and for their association with RA severity. Adjust the analysis for the known risk factors, age, duration of disease and smoking. Calculate the population attributable fraction/risk (based on the equation 3.23 on page 112, Woodward textbook) for each SNP that is statistically significantly associated with severity. State any assumptions you have made in your analysis and comment on your results.

Each remaining SNP was modelled for association with RA severity, adjusting for age, duration of disease and smoking.

```
#gt1.lm <- glm(severity~gt1, family = binomial, data=gdf)
#summary(gt1.lm)
gt1.adj.lm <- glm(severity~gt1+age+duration+smoke, family = binomial, data=gdf)
summary(gt1.adj.lm)
```

```
##
## Call:
## glm(formula = severity ~ gt1 + age + duration + smoke, family = binomial,
##     data = gdf)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.3423  -1.1281   0.6597   0.9846   1.8256
##
```

```
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -0.597864   0.253995  -2.354   0.0186 *
## gt1         -0.778865   0.086632  -8.991  < 2e-16 ***
## age          0.005663   0.005410   1.047   0.2952
## duration     0.061979   0.005923  10.464  < 2e-16 ***
## smoke        0.511933   0.110084   4.650 3.31e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 2538.1  on 1873  degrees of freedom
## Residual deviance: 2304.8  on 1869  degrees of freedom
##   (4 observations deleted due to missingness)
## AIC: 2314.8
##
## Number of Fisher Scoring iterations: 3
```

```
#gt3.lm <- glm(severity~gt3, family = binomial, data=gdf)
#summary(gt3.lm)
gt3.adj.lm <- glm(severity~gt3+age+duration+smoke, family = binomial, data=gdf)
summary(gt3.adj.lm)
```

```
##
## Call:
## glm(formula = severity ~ gt3 + age + duration + smoke, family = binomial,
##     data = gdf)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.2273  -1.1608   0.7138   1.0354   1.4499
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -0.810718   0.257745  -3.145  0.00166 **
## gt3          0.014307   0.071818   0.199  0.84210
## age          0.005062   0.005289   0.957  0.33857
## duration     0.058901   0.005768  10.211  < 2e-16 ***
## smoke        0.120486   0.098141   1.228  0.21956
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 2534.5  on 1871  degrees of freedom
## Residual deviance: 2386.7  on 1867  degrees of freedom
##   (6 observations deleted due to missingness)
## AIC: 2396.7
##
## Number of Fisher Scoring iterations: 4
```

```
#gt4.lm <- glm(severity~gt4, family = binomial, data=gdf)
#summary(gt4.lm)
gt4.adj.lm <- glm(severity~gt4+age+duration+smoke, family = binomial, data=gdf)
```

```r
summary(gt4.adj.lm)
```

```
##
## Call:
## glm(formula = severity ~ gt4 + age + duration + smoke, family = binomial,
##     data = gdf)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.2326  -1.1590   0.7119   1.0384   1.4525
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -0.786160   0.253640  -3.100  0.00194 **
## gt4         -0.012032   0.067897  -0.177  0.85934
## age          0.004869   0.005279   0.922  0.35636
## duration     0.059389   0.005774  10.286  < 2e-16 ***
## smoke        0.114599   0.098137   1.168  0.24291
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 2536.0  on 1871  degrees of freedom
## Residual deviance: 2386.5  on 1867  degrees of freedom
##   (6 observations deleted due to missingness)
## AIC: 2396.5
##
## Number of Fisher Scoring iterations: 4
```

```r
#gt5.lm <- glm(severity~gt5, family = binomial, data=gdf)
#summary(gt5.lm)
gt5.adj.lm <- glm(severity~gt5+age+duration+smoke, family = binomial, data=gdf)
summary(gt5.adj.lm)
```

```
##
## Call:
## glm(formula = severity ~ gt5 + age + duration + smoke, family = binomial,
##     data = gdf)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.2498  -1.1641   0.7147   1.0378   1.4840
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -0.720479   0.254129  -2.835  0.00458 **
## gt5         -0.075267   0.066868  -1.126  0.26033
## age          0.005007   0.005293   0.946  0.34411
## duration     0.058848   0.005774  10.191  < 2e-16 ***
## smoke        0.105757   0.098151   1.077  0.28126
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 2535.2  on 1871  degrees of freedom
## Residual deviance: 2386.4  on 1867  degrees of freedom
##   (6 observations deleted due to missingness)
## AIC: 2396.4
##
## Number of Fisher Scoring iterations: 4
```

```
#gt7.lm <- glm(severity~gt7, family = binomial, data=gdf)
#summary(gt7.lm)
gt7.adj.lm <- glm(severity~gt7+age+duration+smoke, family = binomial, data=gdf)
summary(gt7.adj.lm)
```

```
##
## Call:
## glm(formula = severity ~ gt7 + age + duration + smoke, family = binomial,
##     data = gdf)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.2281  -1.1598   0.7143   1.0377   1.4479
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -0.774585   0.249874  -3.100  0.00194 **
## gt7         -0.005754   0.091197  -0.063  0.94969
## age          0.004537   0.005286   0.858  0.39075
## duration     0.059348   0.005772  10.282  < 2e-16 ***
## smoke        0.117332   0.098242   1.194  0.23235
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 2536.3  on 1872  degrees of freedom
## Residual deviance: 2387.6  on 1868  degrees of freedom
##   (5 observations deleted due to missingness)
## AIC: 2397.6
##
## Number of Fisher Scoring iterations: 4
```

The results of the analyses, after adjusting for age, duration of disease and smoking, demonstrate that only gt1 of all the SNP's was statistically significant at $p<0.05$.

As this is a The population attributable fraction/risk for gt1 is calculated below

```
#The PAR formula is
#PAR = Pe (RRe-1)  /  [1 + Pe (RRe-1)],
#where  Pe is the prevalence of the exposure (e.g., proportion who are
#overweight) and RRe is the relative risk of disease
#due to that exposure.

#create table frequency of disease vs exposure.
#include in dataframe gender, severity and gt1 fields
gdf.mod.df <- gdf[,5:7]
```

```r
#classify responses gt1(2) as exposed and gt1(1) or gt(0) as non-exposed
gdf.mod.df$gt1 <- ifelse(gdf.mod.df$gt1==2, "exposed", "non-exposed")
#classify response severity(1) as severe and severity(0) as not severe
gdf.mod.df$severity <- ifelse(gdf.mod.df$severity==1, "severe", "not severe")
#define column order in table function
gdf.mod.df$severity <- factor(gdf.mod.df$severity, levels = c("severe", "not severe"))
#gender(0) is male and gender(1) is female
gdf.mod.df$gender <- ifelse(gdf.mod.df$gender==0, "male", "female")
par.table <- table(gdf.mod.df$gt1, gdf.mod.df$severity)

n <- sum(par.table)
A <- par.table[1]
C <- par.table[2]
D <- par.table[4]

PAR <- 1/(A+C)*((A+C)-(n*C)/(C+D))
paste("The population attributable risk (PAR) for gt1 is", round((PAR), digit=3))
```

```
## [1] "The population attributable risk (PAR) for gt1 is -0.034"
```

If the relationship is causal, then the risk of the population of having severe RA would be increased by 3.4% if the gt1 gene was not present. This result assumes there is a causal relationship between the presence of gt1 and severity of RA and not just an association. Also, it has been assumed that the results of the cohort study can be applied to the population in general, and that there is no bias in the estimation of PAR. Finally, it has been assumed that the elimination of gt1 has no effect on the distribution of any other risk factors.

3c

Could any of your conclusions be affected by the presence of a confounding or modifying variable? Explore if this could be the case in this cross sectional study. Discuss your results and modify your conclusions to b) if necessary.

We investigated the marginal (crude) odds ratio (OR) and the conditional OR dependent on gender, to determine if gender was a confounder or effects modifier. The results are shown below.

```r
#install vcd packages
library(vcd)
```

```
## Warning: package 'vcd' was built under R version 3.4.3
```

```
## Loading required package: grid
```

```r
tab_crude <- xtabs(~gt1+severity, data= gdf.mod.df) #check out data.table package
OR.crude <- oddsratio(tab_crude, log=FALSE)
CI.crude.OR <- confint(OR.crude) #vcd package

tab.strat.gender <- xtabs(~gt1+severity+gender, data= gdf.mod.df)
OR.strat.gender <- oddsratio(tab.strat.gender, log=FALSE)
CI.OR <- confint(OR.strat.gender) #vcd package

cat("For the crude OR \n")
```

```
## For the crude OR
```

```r
OR.crude
```

```
##  odds ratios for gt1 and severity
```

```
##
## [1] 0.3707405
```

```
cat("\n")
```

```
cat("95% conf.int of crude OR is (",round((CI.crude.OR[1]), digit=2),",",
    round((CI.crude.OR[2]),digit=2),") \n" )
```

```
## 95% conf.int of crude OR is ( 0.26 , 0.52 )
```

```
cat("\n")
```

```
cat("\n")
```

```
cat("The OR accounting for gender \n")
```

```
## The OR accounting for gender
```

```
OR.strat.gender
```

```
##  odds ratios for gt1 and severity by gender
##
##    female      male
## 0.4333333 0.3341909
```

```
cat("\n")
```

```
cat("95% conf.int OR for females is (",round((CI.OR[1]), digit=2),",",
    round((CI.OR[3]),digit=2),") \n" )
```

```
## 95% conf.int OR for females is ( 0.25 , 0.74 )
```

```
cat("95% conf.int OR for males is (",round((CI.OR[2]), digit=2),",",
    round((CI.OR[4]),digit=2),") \n" )
```

```
## 95% conf.int OR for males is ( 0.21 , 0.52 )
```

```
#test homogeniety between strata using woolf test- if not homogenous then there
#are effect modifiers
woolf_test(tab.strat.gender) #pvalue = 0.469. conditional OR on gender homogeneous.
```

```
##
##  Woolf-test on Homogeneity of Odds Ratios (no 3-Way assoc.)
##
## data:  tab.strat.gender
## X-squared = 0.52402, df = 1, p-value = 0.4691
```

The crude OR value (0.37) is between the OR value for males (0.33) and females (0.43). The Confidence intervals of both male and female conditional tables and of the marginal (crude) table all overlap. The test for homogeniety between levels of the gender modifying variable is non-significant (p>0.05), indicating that gender is not a modifying variable. Therefore in terms of OR and the collapsibility definition of confounding, there are no confounding or modifying effects of gender on the severity of disease for those with genotype 2 of gt1. As all 95% confidence intervals are less than one, it can be concluded that odds of having severe RA with genotype 2 in gt1 compared with having genotype 1 or 0 is less than 1.