In The Name of God

Statistical Inference HW#7

Amin Asadi Sarijalou

810196410

Spring 1400

Contents

Problem 1

a.

Yes, we can. In order to do this, we can use “one-vs-all” classification, once for each class. In other words, for each class i, we train a LR classifier to predict the probability of $y = i$.

b.

$$odds_{heads} = \frac{P(Heads)}{P(Tails)} = \frac{\frac{1}{2}}{\frac{1}{2}} = 1$$

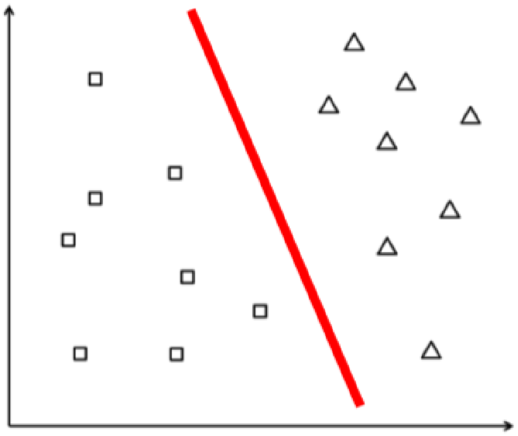
This means that in many trials, the ratio of the times we get heads to the times we get tails is equal to 1 (They have equal chance to occure).

c.

The model with blue curve is the best. Because for the same specificity (TN rate), sensitivity (TP rate) of blue model is higher. Also, for the same sensitivity, specificity of model A is higher (it’s FP is lower).

Problem 2

Yes, it can. The reason is that the samples of two classes are perfectly linearly separable. Therefore, after some iterations of the training process, we will find a linear decision rule (line) which is able to correctly classify all the samples, hence the error will be zero.



Problem 3

First we combine the two groups of data and calculate the ranks:

Combined: [315 317 316 316 295 318 317 316 269 314 321 319 267 242 324 323 284 258 257 322]

Ranks: [18.0 2.0 3.0 6.0 19.0 20.0 1.0 4.0 16.0 17.0 8.0 5.0 11.0 13.5 15.0 7.0 11.0 11.0 13.5 9.0]

$$U = \sum_{1}^{n} Rank(X_i) = 106$$

$$\mu = \frac{n \times (n + m + 1)}{2} = 105$$

$$\sigma = \frac{n \times (n + m + 1)}{2} \approx 13.22$$

$$pvalue = 2 \times \ P\left(U \geq \frac{|104 - 0.5 - 105|}{13.22}\right) \approx 0.97$$

Because p-value is very large and bigger that significance level (0.05) we can’t claim that these two distributions are different..

Problem 4

a.

$$sensitivity = P(TP|\ TP + FN) = \frac{867}{1000} = 0.867$$

b.

$$specifity = P(\frac{TN}{FP + TN}) = \frac{800 - 85}{800} \approx 0.89$$

c.

$$PPV = \frac{sensitivity \times prevalence}{sensitivity \times prevalence + (1 - specifity) \times (1 - prevalence)}$$

$$= \frac{0.867 \times 0.02}{0.867 \times 0.02 + 0.11 \times 0.98} \approx 0.143$$

Problem 5

a.

The response variable is whether the injury is fatal or nonfatal. Explanatory variable is (having or not having) safety equipment.

$$difference\ of\ proportions = \frac{510}{510 + 412{,}368} - \frac{1601}{1601 + 162{,}527} \approx -0.008$$

$$OR = \frac{\dfrac{P(Fatal\,|\,Seat\ Belt)}{1 - P(Fatal\,|\,Seat\ Belt)}}{\dfrac{P(Fatal\,|\,No\ Seat\ Belt)}{1 - P(Fatal\,|\,No\ Seat\ Belt)}} = \frac{\dfrac{\frac{510}{510 + 412{,}368}}{1 - \frac{510}{510 + 412{,}368}}}{\dfrac{\frac{1601}{1601 + 162{,}527}}{1 - \frac{1601}{1601 + 162{,}527}}} = \frac{0.001236}{0.00985} \approx 0.125$$

$$RR = \frac{P(fatal\,|\,Seat\ Belt)}{P(Fatal\,|\,No\ Seat\ Belt)} = \frac{\frac{510}{510 + 412{,}368}}{\frac{1601}{1601 + 162{,}527}} = \frac{0.001236}{0.00985} \approx 0.126$$

b.

Because proportions are approximately equal (their difference are almost equal to zero), so in the OR formula, the divisors in upper and lower part of the division can be cancelled out, hence OR and RR are almost equal.

Problem 6

$$H_0: median = 45$$

$$H_A: median \neq 45$$

$$Conditions\ of\ sign\ test: n\ =\ 30 \geq\ 20 \rightarrow large\ enough$$

$$S_{obs} = \#\ of\ observations\ greater\ thatn\ 45$$

$$\rightarrow S_{obs} = 13$$

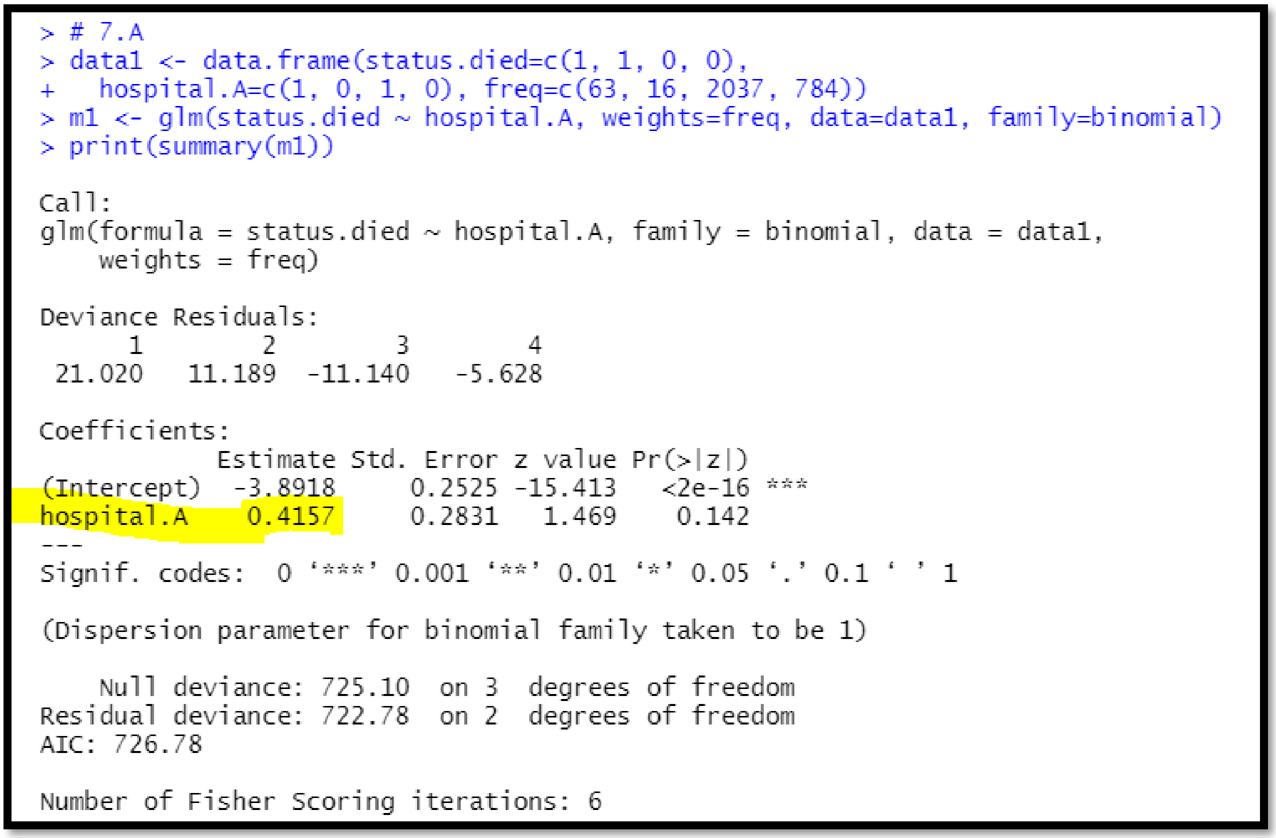
$$z = \frac{S_{obs} + 0.5 - \frac{n}{2}}{\sqrt{n}/2} = \frac{13.5 - 15}{\frac{5.48}{2}} = \frac{-1.5}{2.74} \approx -0.55$$

$$pvalue = 2 \times P(Z >\ |-0.72|) \approx 0.58$$

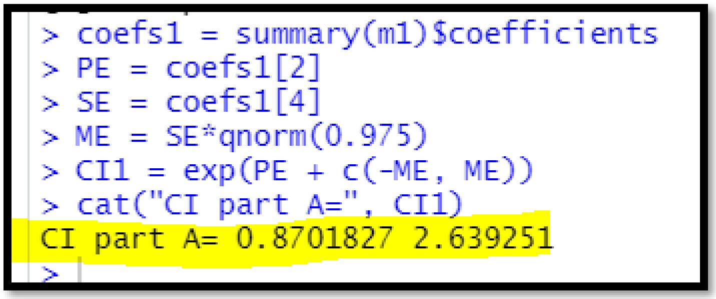
Because 0.58 >> 0.05 we don’t have enough evidence to claim that median of BC id different from 45.

Problem 7

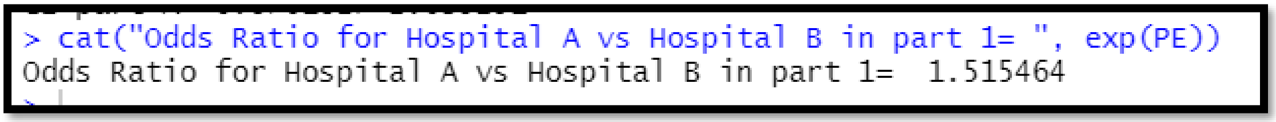
a.



```
> # 7.A
> data1 <- data.frame(status.died=c(1, 1, 0, 0),
+   hospital.A=c(1, 0, 1, 0), freq=c(63, 16, 2037, 784))
> m1 <- glm(status.died ~ hospital.A, weights=freq, data=data1, family=binomial)
> print(summary(m1))

Call:
glm(formula = status.died ~ hospital.A, family = binomial, data = data1,
    weights = freq)

Deviance Residuals:
      1        2        3        4
 21.020   11.189  -11.140   -5.628

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -3.8918     0.2525 -15.413   <2e-16 ***
hospital.A    0.4157     0.2831   1.469    0.142
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 725.10  on 3  degrees of freedom
Residual deviance: 722.78  on 2  degrees of freedom
AIC: 726.78

Number of Fisher Scoring iterations: 6
```

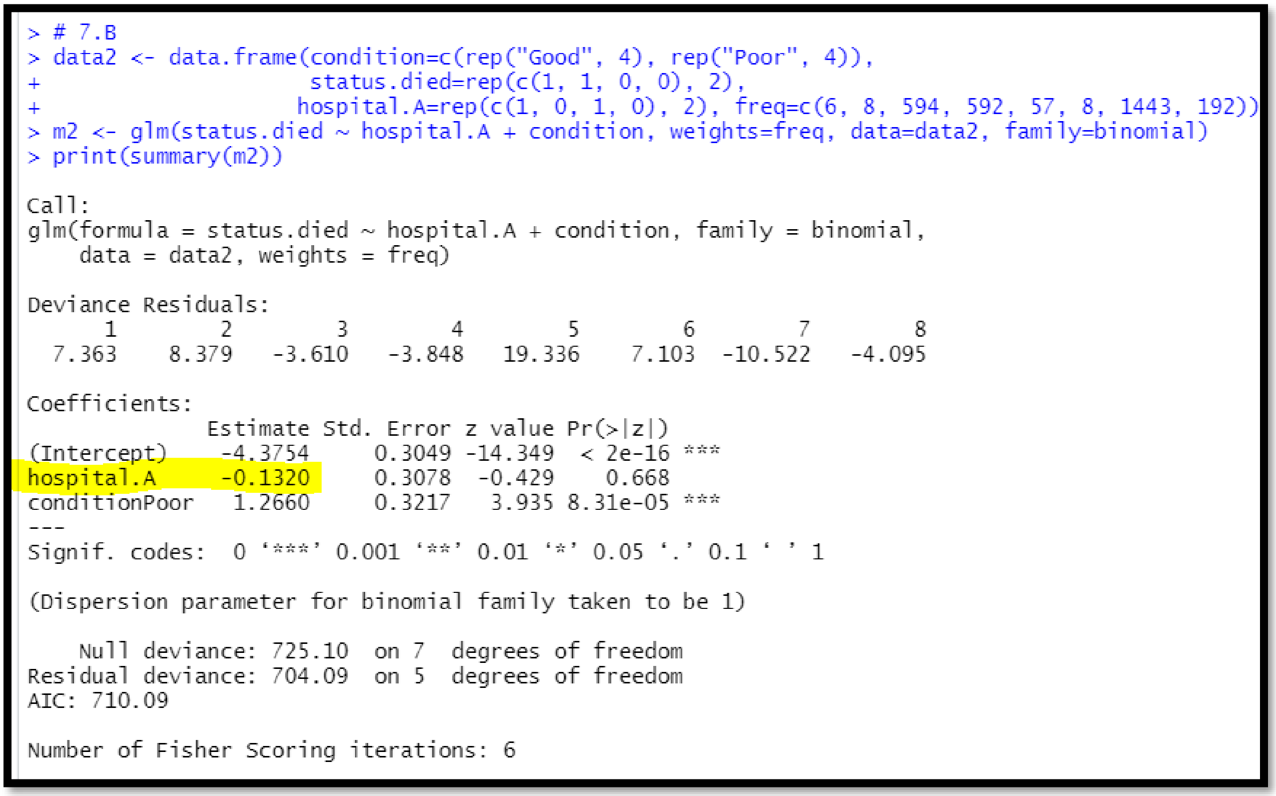


```
> coefs1 = summary(m1)$coefficients
> PE = coefs1[2]
> SE = coefs1[4]
> ME = SE*qnorm(0.975)
> CI1 = exp(PE + c(-ME, ME))
> cat("CI part A=", CI1)
CI part A= 0.8701827 2.639251
>
```

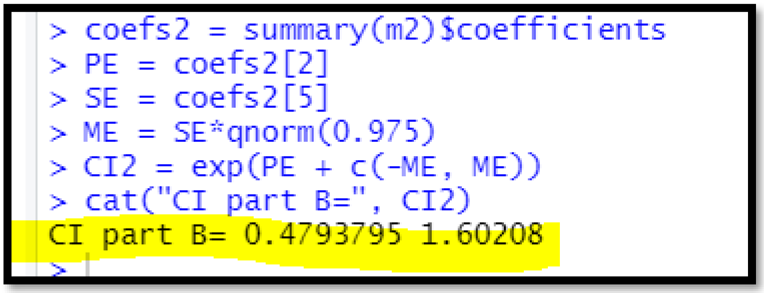


```
> cat("Odds Ratio for Hospital A vs Hospital B in part 1= ", exp(PE))
Odds Ratio for Hospital A vs Hospital B in part 1=  1.515464
```

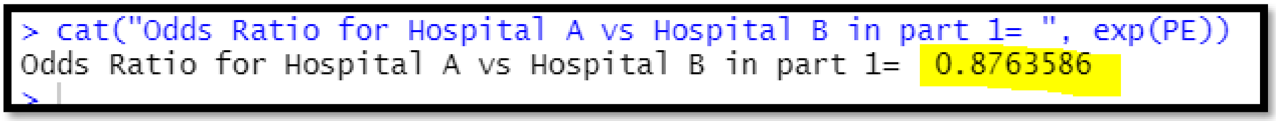
b.



```
> # 7.B
> data2 <- data.frame(condition=c(rep("Good", 4), rep("Poor", 4)),
+                     status.died=rep(c(1, 1, 0, 0), 2),
+                    hospital.A=rep(c(1, 0, 1, 0), 2), freq=c(6, 8, 594, 592, 57, 8, 1443, 192))
> m2 <- glm(status.died ~ hospital.A + condition, weights=freq, data=data2, family=binomial)
> print(summary(m2))

Call:
glm(formula = status.died ~ hospital.A + condition, family = binomial,
    data = data2, weights = freq)

Deviance Residuals:
      1        2        3        4        5        6        7        8
  7.363    8.379   -3.610   -3.848   19.336    7.103  -10.522   -4.095

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)    -4.3754     0.3049 -14.349  < 2e-16 ***
hospital.A     -0.1320     0.3078  -0.429    0.668
conditionPoor   1.2660     0.3217   3.935 8.31e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 725.10  on 7  degrees of freedom
Residual deviance: 704.09  on 5  degrees of freedom
AIC: 710.09

Number of Fisher Scoring iterations: 6
```



```
> coefs2 = summary(m2)$coefficients
> PE = coefs2[2]
> SE = coefs2[5]
> ME = SE*qnorm(0.975)
> CI2 = exp(PE + c(-ME, ME))
> cat("CI part B=", CI2)
CI part B= 0.4793795 1.60208
>
```



```
> cat("Odds Ratio for Hospital A vs Hospital B in part 1= ", exp(PE))
Odds Ratio for Hospital A vs Hospital B in part 1=  0.8763586
>
```

c.

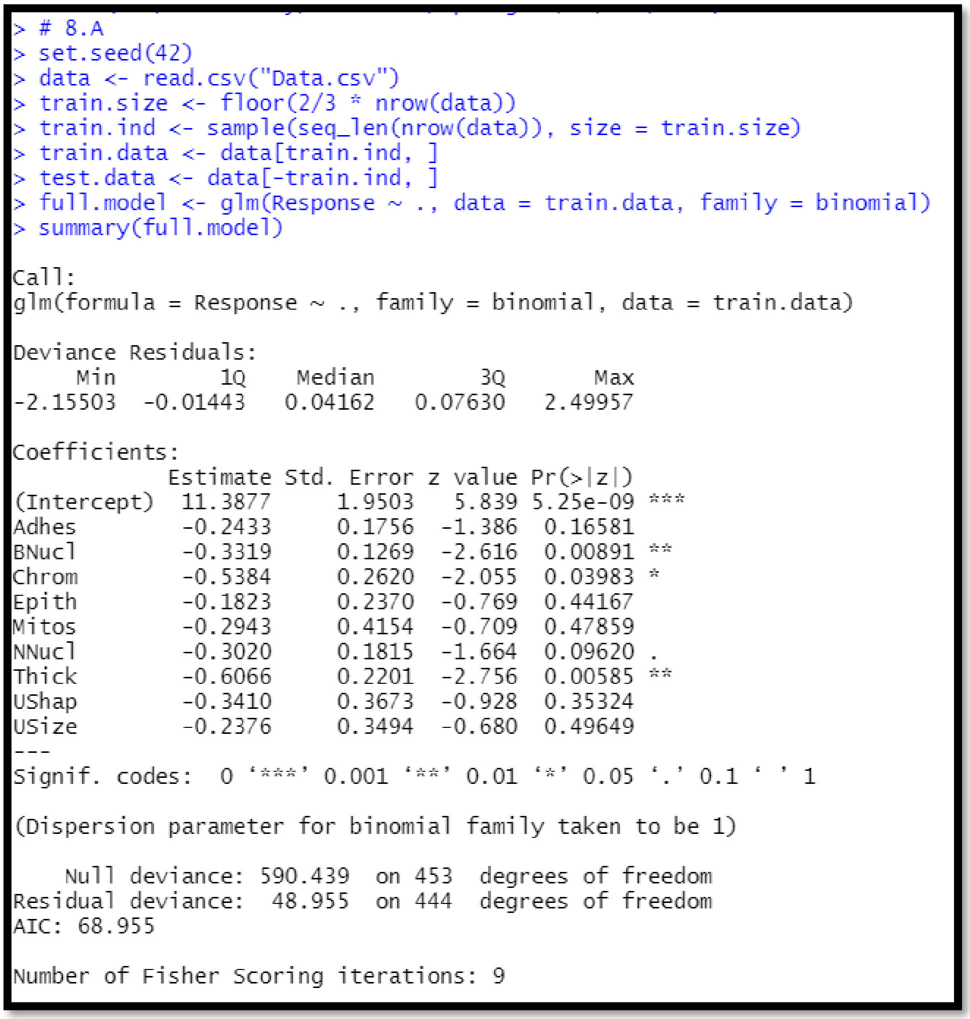
In part a, the odds ratio of death in hospital A relative to hospital B was greater than 1 but In part b this ratio was smaller than 1.

So, in the first part we conclude that deaths are more in hospital A but in the second part we have the inverse conclusion.

So, when the data are combined, the direction of the association is reversed, hence the Simpson's paradox, this is because the number of **patients in poor condition are much more in hospital A** than that of hospital B.

Problem 8

a.



```
> # 8.A
> set.seed(42)
> data <- read.csv("Data.csv")
> train.size <- floor(2/3 * nrow(data))
> train.ind <- sample(seq_len(nrow(data)), size = train.size)
> train.data <- data[train.ind, ]
> test.data <- data[-train.ind, ]
> full.model <- glm(Response ~ ., data = train.data, family = binomial)
> summary(full.model)

Call:
glm(formula = Response ~ ., family = binomial, data = train.data)

Deviance Residuals: 
     Min        1Q    Median        3Q       Max  
-2.15503  -0.01443   0.04162   0.07630   2.49957  

Coefficients:
            Estimate Std. Error z value Pr(>|z|)    
(Intercept)  11.3877     1.9503   5.839 5.25e-09 ***
Adhes        -0.2433     0.1756  -1.386  0.16581    
BNucl        -0.3319     0.1269  -2.616  0.00891 ** 
Chrom        -0.5384     0.2620  -2.055  0.03983 *  
Epith        -0.1823     0.2370  -0.769  0.44167    
Mitos        -0.2943     0.4154  -0.709  0.47859    
NNucl        -0.3020     0.1815  -1.664  0.09620 .  
Thick        -0.6066     0.2201  -2.756  0.00585 ** 
UShap        -0.3410     0.3673  -0.928  0.35324    
USize        -0.2376     0.3494  -0.680  0.49649    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 590.439  on 453  degrees of freedom
Residual deviance:  48.955  on 444  degrees of freedom
AIC: 68.955

Number of Fisher Scoring iterations: 9
```

Formula:

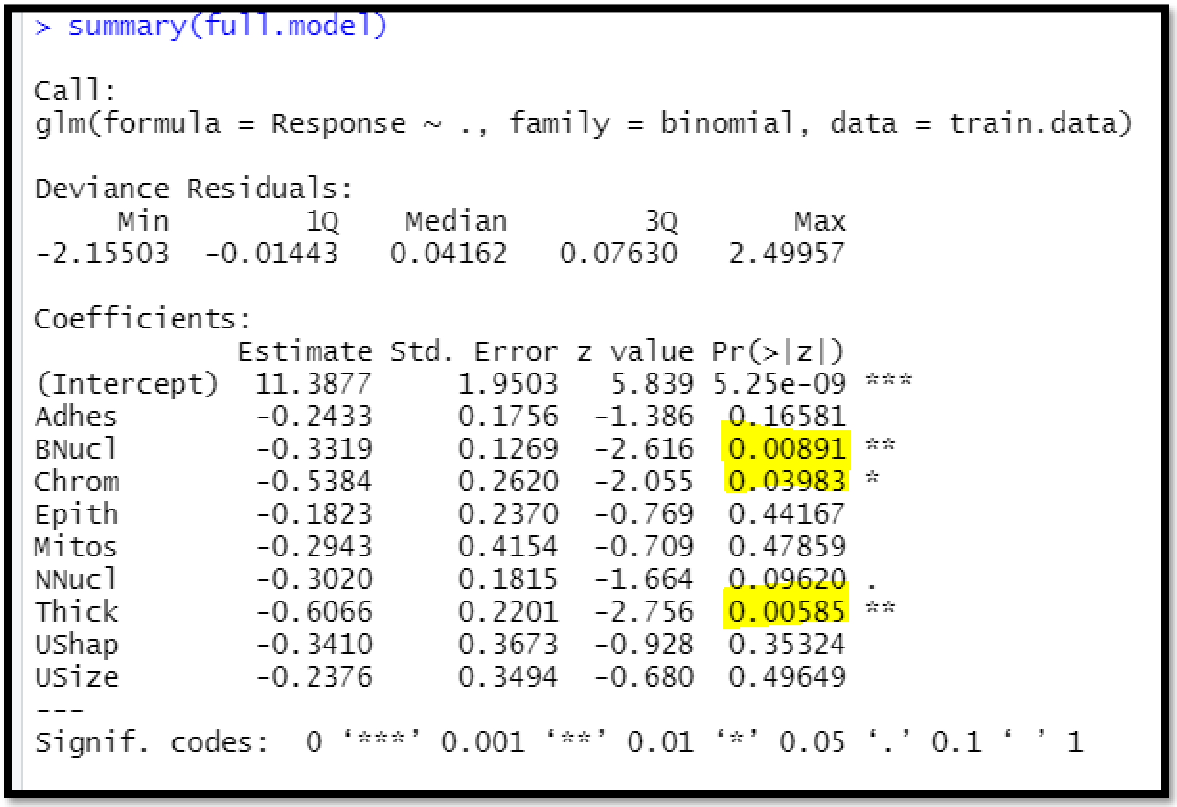
$$Logit(Response) = -0.24 \times Adhes - 0.33 \times BNucl - 0.53 \times Chrom - 0.18 \times Epith$$
$$-0.29 \times Mitos - 0.30 \times NNucl - 0.60 \times Thick - 0.34 \times UShap - 0.23 \times USize$$

b.

for each predictor P we have:

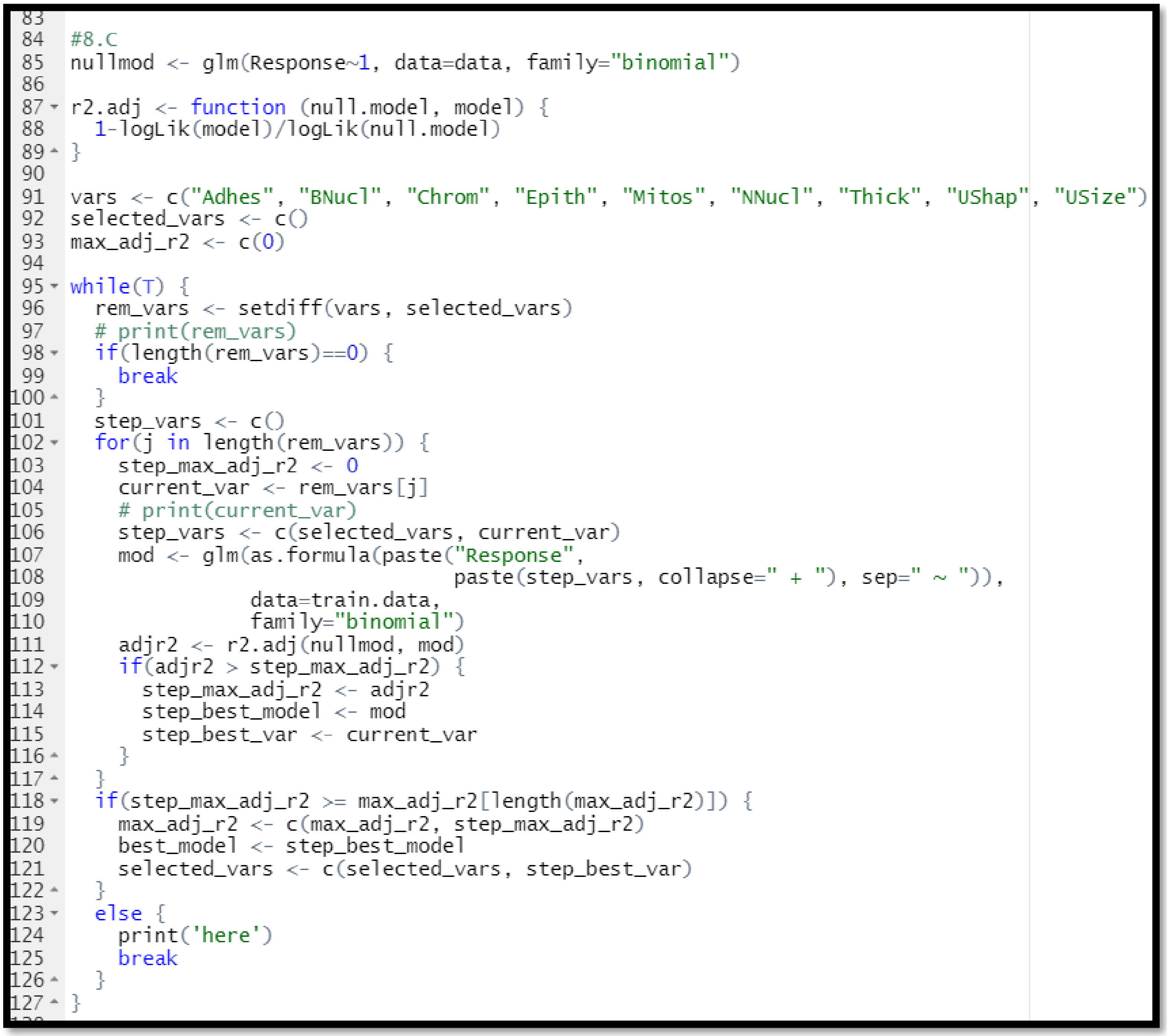
$$\boldsymbol{H_0: \beta_P = 0}$$
$$\boldsymbol{H_A: \beta_P \neq 0}$$

Therefore, as it is apparent, from figure below, **BNucl**, **Chrom**, **Thick** are significant because their P-value are less than $\alpha = 0.05$

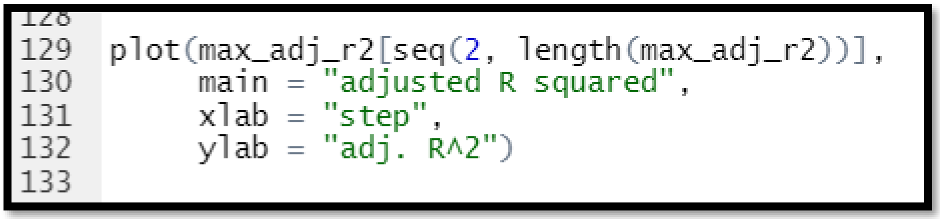


```
> summary(full.model)

Call:
glm(formula = Response ~ ., family = binomial, data = train.data)

Deviance Residuals: 
     Min        1Q    Median        3Q       Max  
-2.15503  -0.01443   0.04162   0.07630   2.49957  

Coefficients:
            Estimate Std. Error z value Pr(>|z|)    
(Intercept)  11.3877     1.9503   5.839 5.25e-09 ***
Adhes        -0.2433     0.1756  -1.386  0.16581    
BNucl        -0.3319     0.1269  -2.616  0.00891 ** 
Chrom        -0.5384     0.2620  -2.055  0.03983 *  
Epith        -0.1823     0.2370  -0.769  0.44167    
Mitos        -0.2943     0.4154  -0.709  0.47859    
NNucl        -0.3020     0.1815  -1.664  0.09620 .  
Thick        -0.6066     0.2201  -2.756  0.00585 ** 
UShap        -0.3410     0.3673  -0.928  0.35324    
USize        -0.2376     0.3494  -0.680  0.49649    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

c.

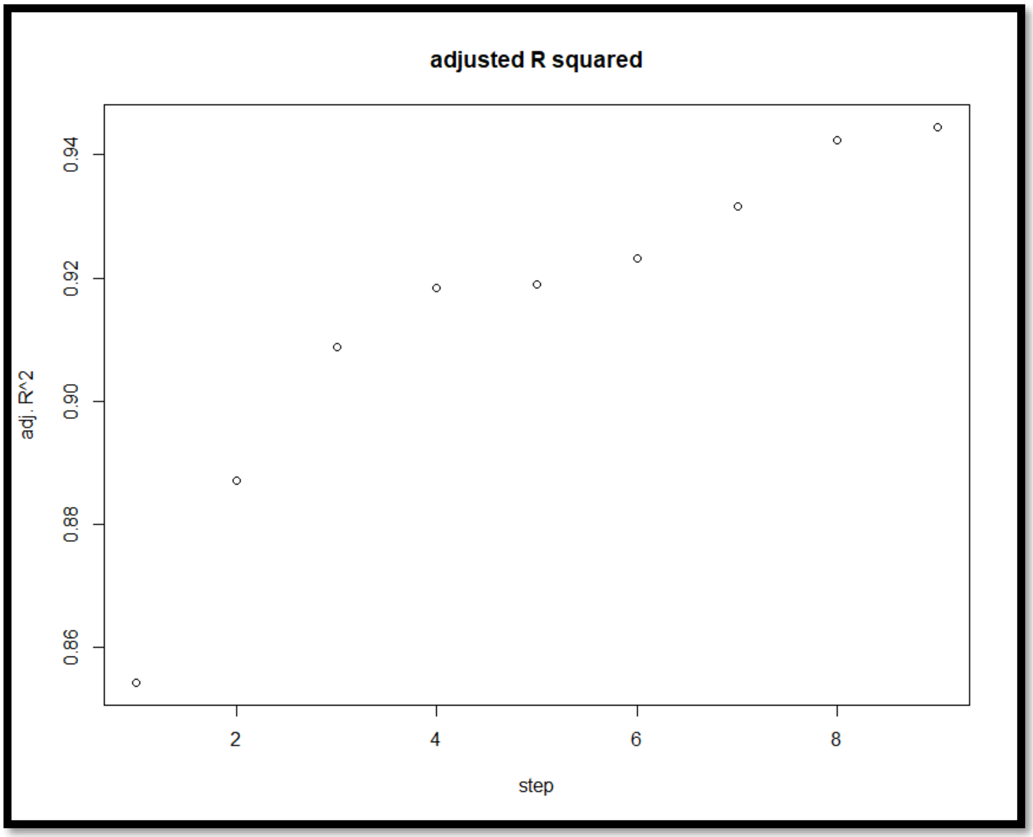


```
#8.C
nullmod <- glm(Response~1, data=data, family="binomial")

r2.adj <- function (null.model, model) {
  1-logLik(model)/logLik(null.model)
}

vars <- c("Adhes", "BNucl", "Chrom", "Epith", "Mitos", "NNucl", "Thick", "UShap", "USize")
selected_vars <- c()
max_adj_r2 <- c(0)

while(T) {
  rem_vars <- setdiff(vars, selected_vars)
  # print(rem_vars)
  if(length(rem_vars)==0) {
    break
  }
  step_vars <- c()
  for(j in length(rem_vars)) {
    step_max_adj_r2 <- 0
    current_var <- rem_vars[j]
    # print(current_var)
    step_vars <- c(selected_vars, current_var)
    mod <- glm(as.formula(paste("Response",
                                paste(step_vars, collapse=" + "), sep=" ~ ")),
               data=train.data,
               family="binomial")
    adjr2 <- r2.adj(nullmod, mod)
    if(adjr2 > step_max_adj_r2) {
      step_max_adj_r2 <- adjr2
      step_best_model <- mod
      step_best_var <- current_var
    }
  }
  if(step_max_adj_r2 >= max_adj_r2[length(max_adj_r2)]) {
    max_adj_r2 <- c(max_adj_r2, step_max_adj_r2)
    best_model <- step_best_model
    selected_vars <- c(selected_vars, step_best_var)
  }
  else {
    print('here')
    break
  }
}
```

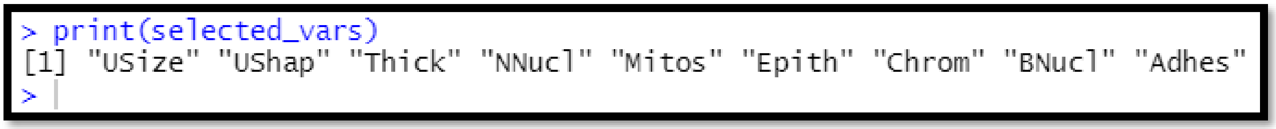


```
plot(max_adj_r2[seq(2, length(max_adj_r2))],
     main = "adjusted R squared",
     xlab = "step",
     ylab = "adj. R^2")
```

We can see that adjusted R squared has strictly increased at each step:

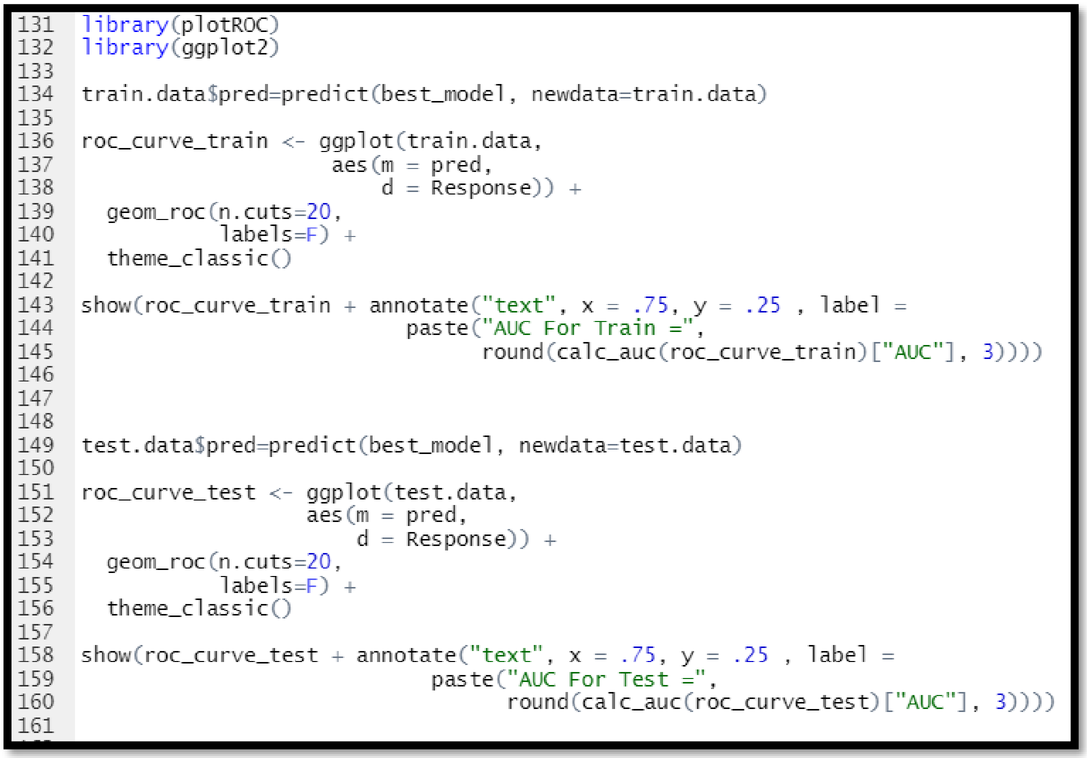


As we can see below, the best model includes 9 variables:



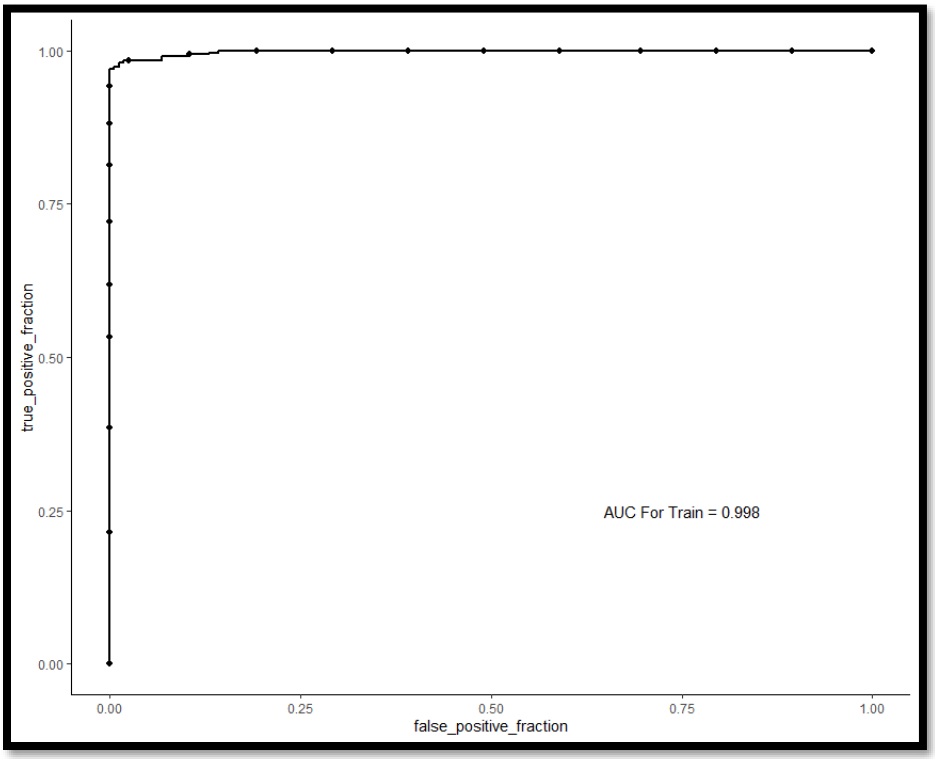
```
> print(selected_vars)
[1] "USize" "UShap" "Thick" "NNucl" "Mitos" "Epith" "Chrom" "BNucl" "Adhes"
> 
```

d.

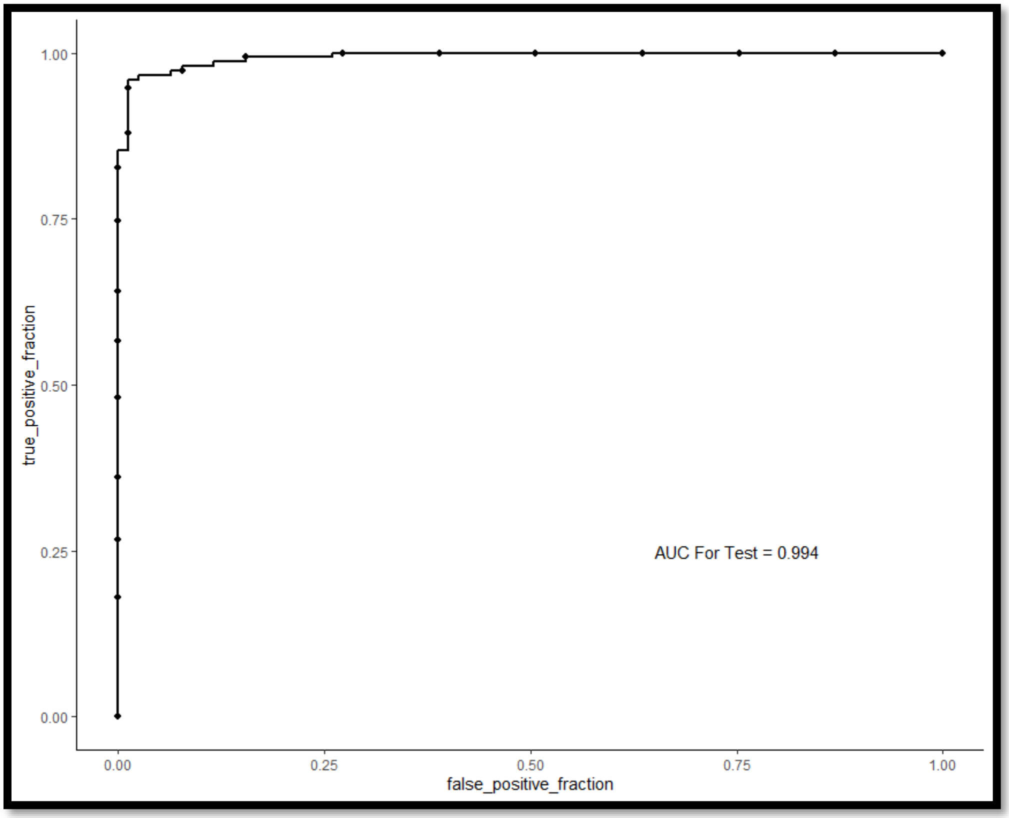


```
library(plotROC)
library(ggplot2)

train.data$pred=predict(best_model, newdata=train.data)

roc_curve_train <- ggplot(train.data,
                    aes(m = pred,
                        d = Response)) +
  geom_roc(n.cuts=20,
           labels=F) +
  theme_classic()

show(roc_curve_train + annotate("text", x = .75, y = .25 , label =
                          paste("AUC For Train =",
                                round(calc_auc(roc_curve_train)["AUC"], 3))))



test.data$pred=predict(best_model, newdata=test.data)

roc_curve_test <- ggplot(test.data,
                  aes(m = pred,
                      d = Response)) +
  geom_roc(n.cuts=20,
           labels=F) +
  theme_classic()

show(roc_curve_test + annotate("text", x = .75, y = .25 , label =
                            paste("AUC For Test =",
                                  round(calc_auc(roc_curve_test)["AUC"], 3))))

```

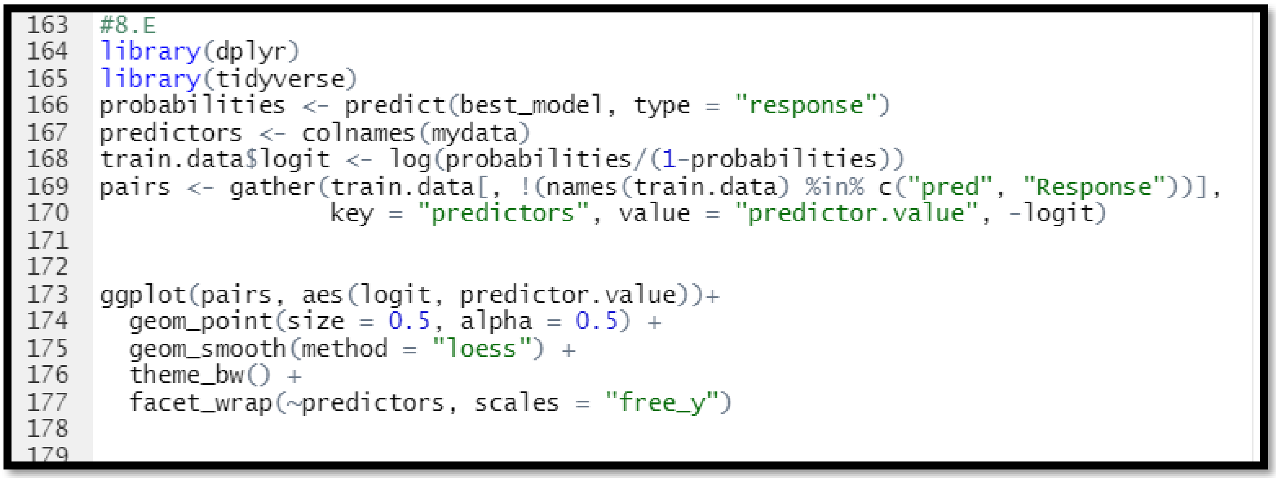
ROC plot for Train Data (AUC=0.998):



ROC plot for Test Data (AUC=0.994):

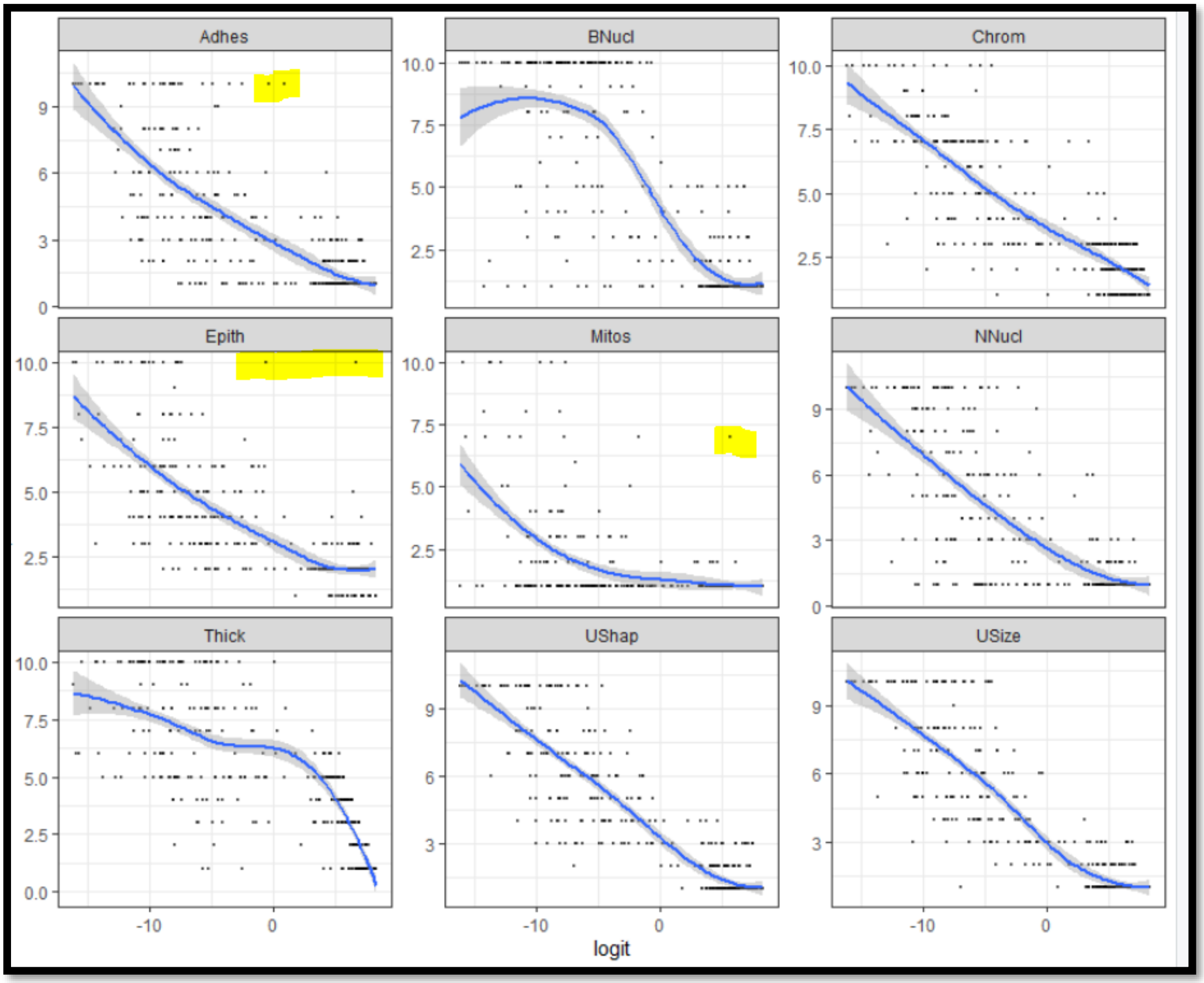


e.



```
#8.E
library(dplyr)
library(tidyverse)
probabilities <- predict(best_model, type = "response")
predictors <- colnames(mydata)
train.data$logit <- log(probabilities/(1-probabilities))
pairs <- gather(train.data[, !(names(train.data) %in% c("pred", "Response"))],
                key = "predictors", value = "predictor.value", -logit)


ggplot(pairs, aes(logit, predictor.value))+
  geom_point(size = 0.5, alpha = 0.5) +
  geom_smooth(method = "loess") +
  theme_bw() +
  facet_wrap(~predictors, scales = "free_y")
```

As we can see below, all the explanatory variables except “Thick” and “BNucl” has linear association with logit. Outliers are highlighted in Yellow.


Adhes
BNucl
Chrom
Epith
Mitos
NNucl
Thick
UShap
USize
logit