In‌ ‌The‌ ‌Name‌ ‌of‌ ‌God‌ ‌

Statistical‌ ‌Inference‌ ‌HW#7

Amin‌ ‌Asadi‌ Sarijalou‌

810196410‌ ‌

Spring 1400

Contents

[Problem 1 3](#_Toc75806311)

[a. 3](#_Toc75806312)

[b. 3](#_Toc75806313)

[c. 3](#_Toc75806314)

[Problem 2 3](#_Toc75806315)

[Problem 3 4](#_Toc75806316)

[Problem 4 4](#_Toc75806317)

[a. 4](#_Toc75806318)

[b. 4](#_Toc75806319)

[c. 4](#_Toc75806320)

[Problem 5 5](#_Toc75806321)

[a. 5](#_Toc75806322)

[b. 5](#_Toc75806323)

[Problem 6 5](#_Toc75806324)

[Problem 7 6](#_Toc75806325)

[a. 6](#_Toc75806326)

[b. 7](#_Toc75806327)

[c. 8](#_Toc75806328)

[Problem 8 8](#_Toc75806329)

[a. 8](#_Toc75806330)

[b. 9](#_Toc75806331)

[c. 10](#_Toc75806332)

[d. 12](#_Toc75806333)

[e. 13](#_Toc75806334)

# 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 .

## b.

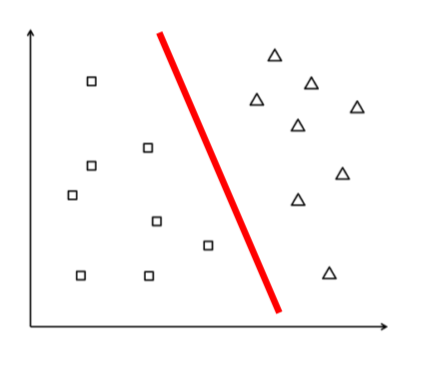
This means that in many trials, the ratio of the times we get heads to the times we get tails is equal to 1 (They have equal chance to occure).

## 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]

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.

## b.

## c.

# Problem 5

## a.

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

## b.

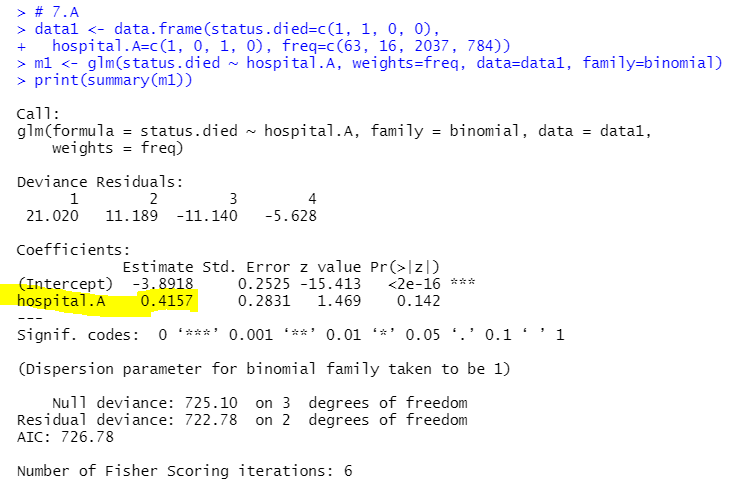
Because proportions are approximately equal (their difference are almost equal to zero), so in the OR formula, the divisors in upper and lower part of the division can be cancelled out, hence OR and RR are almost equal.

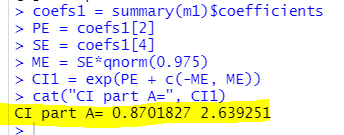
# Problem 6

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

# Problem 7

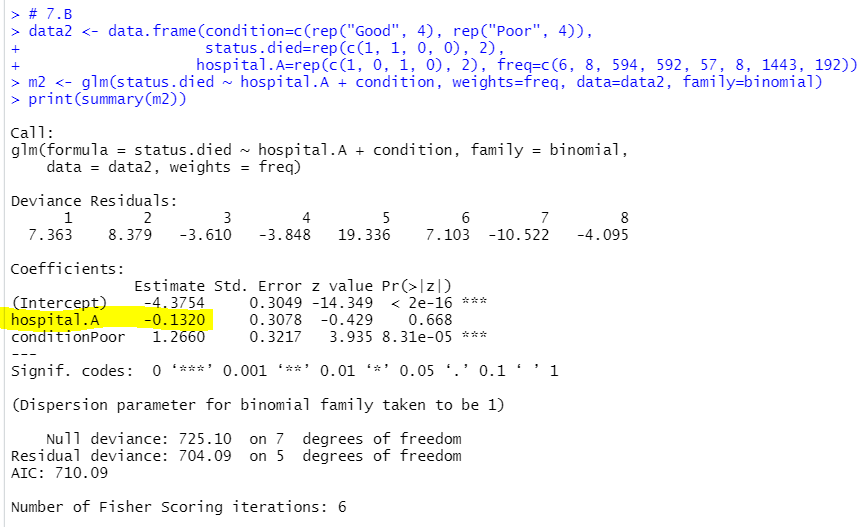
## a.

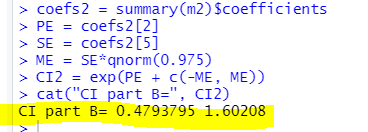






## b.







## c.

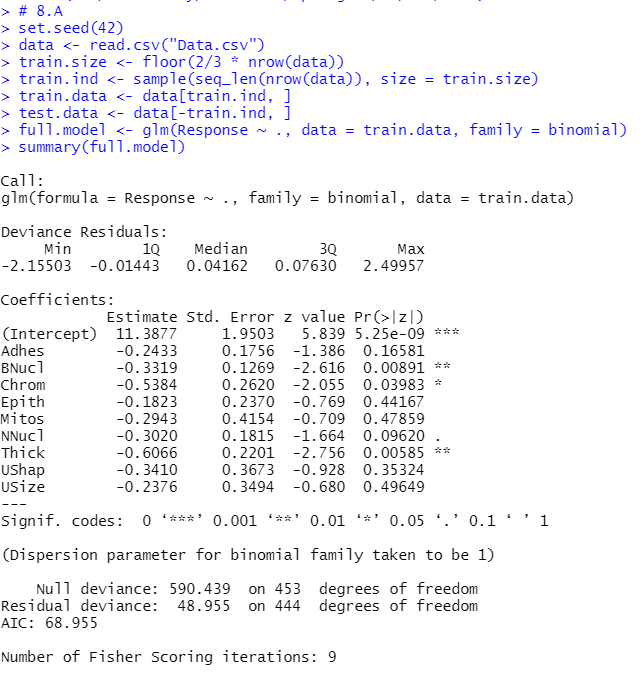
In part a, the odds ratio of death in hospital A relative to hospital B was greater than 1 but In part b this ratio was smaller than 1.

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.

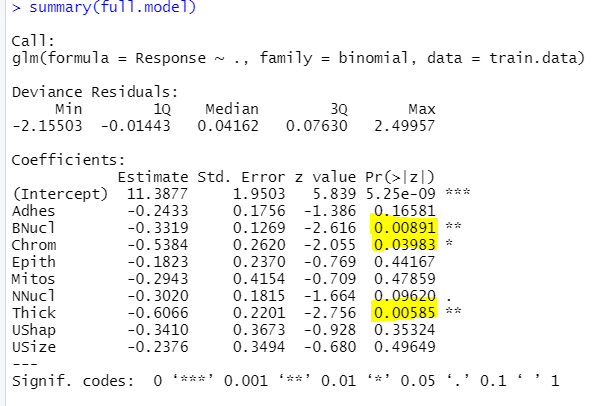


***Formula:***

## b.

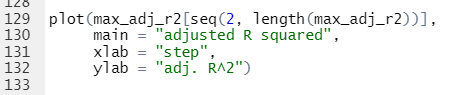
**for each predictor P we have:**

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

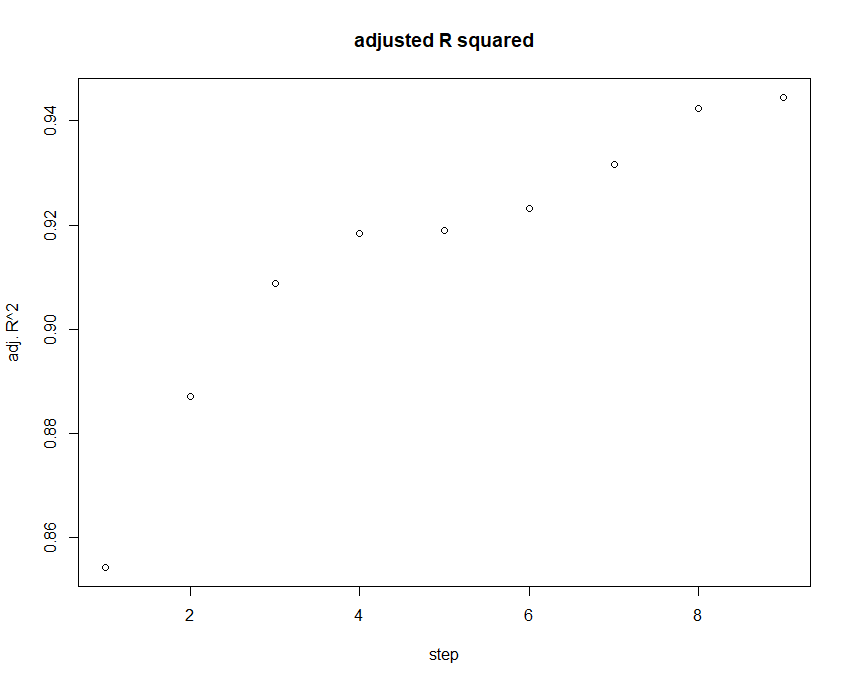


## c.





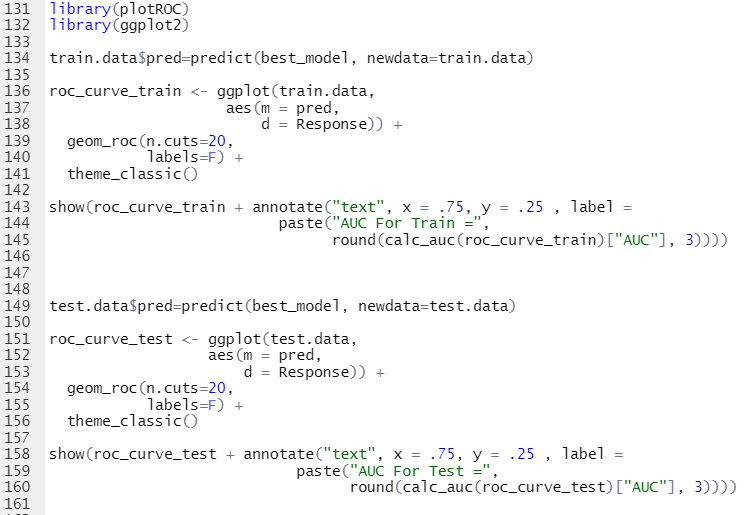
We can see that adjusted R squared has strictly increased at each step:



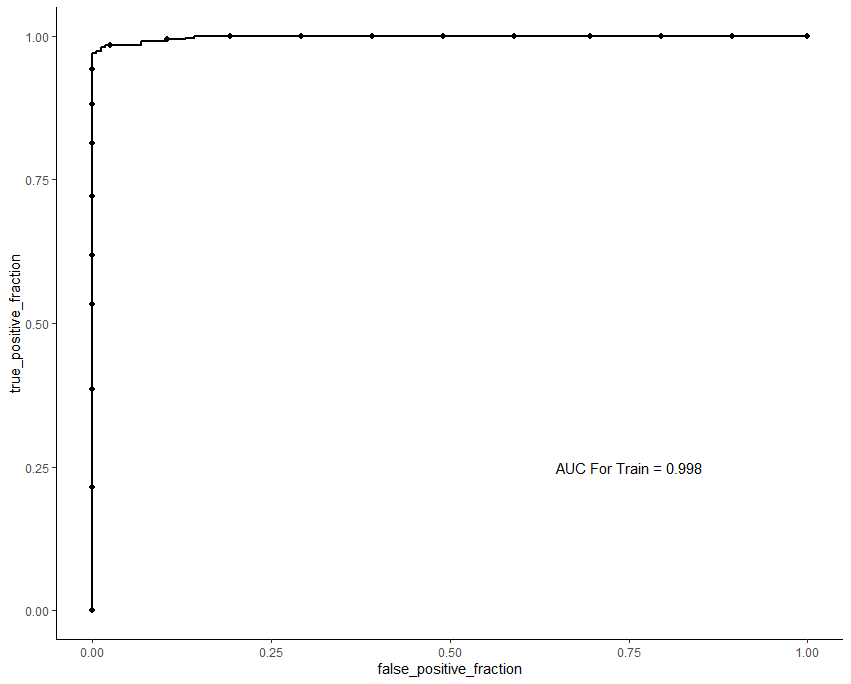
As we can see below, the best model includes 9 variables:



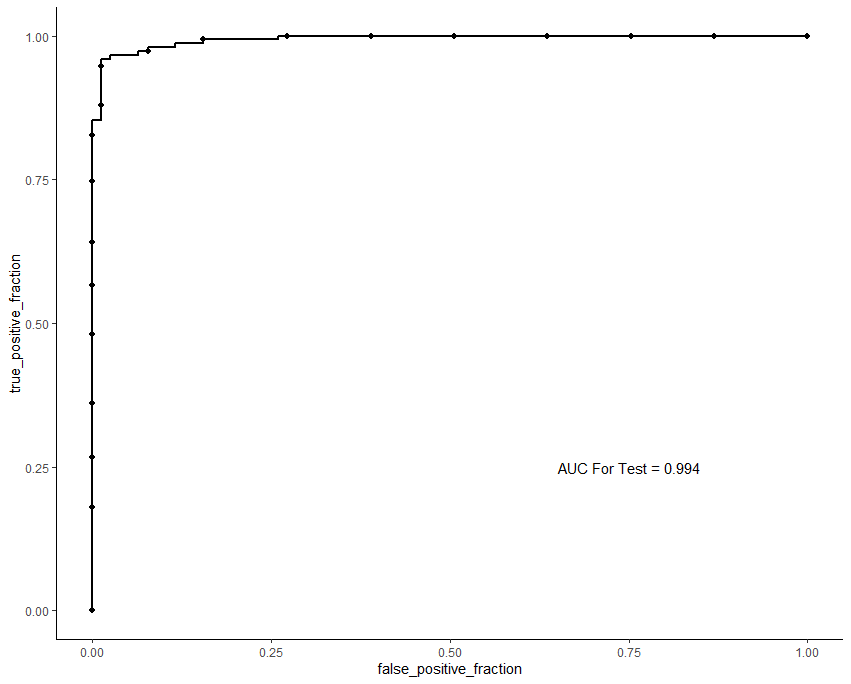
## d.



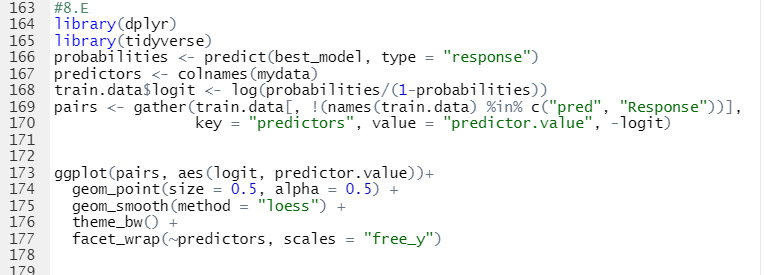
**ROC plot for Train Data (AUC=0.998):**



**ROC plot for Test Data (AUC=0.994):**



## e.



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

