Data Science Assignment 5

Kishan Sarpangala

**Theme**:

Combining biomarkers Aneurysmal subarachnoid hemorrhage (aSAH) is a serious medical condition with bleeding occurring in the brain.

There are several causes. Head trauma is one of them. Neurosurgeons, to begin with, focus on stopping the bleeding and preventing its reoccurrence. Even if successful in stopping bleeding, brain damage with cognitive impairment could occur subsequently.

In the class I attempted to develop a diagnostic test to predict brain damage based on a single protein biomarker s100b. There are other ways to develop a prediction model. The one I have in mind in this exercise is combining several biomarkers. The biomarkers of interest are: s100b; log(ndka); gender; and age.

**Question 1**:

Write a short note on the biomarker ndka from internet sources (not more than 5 sentences). 2 points

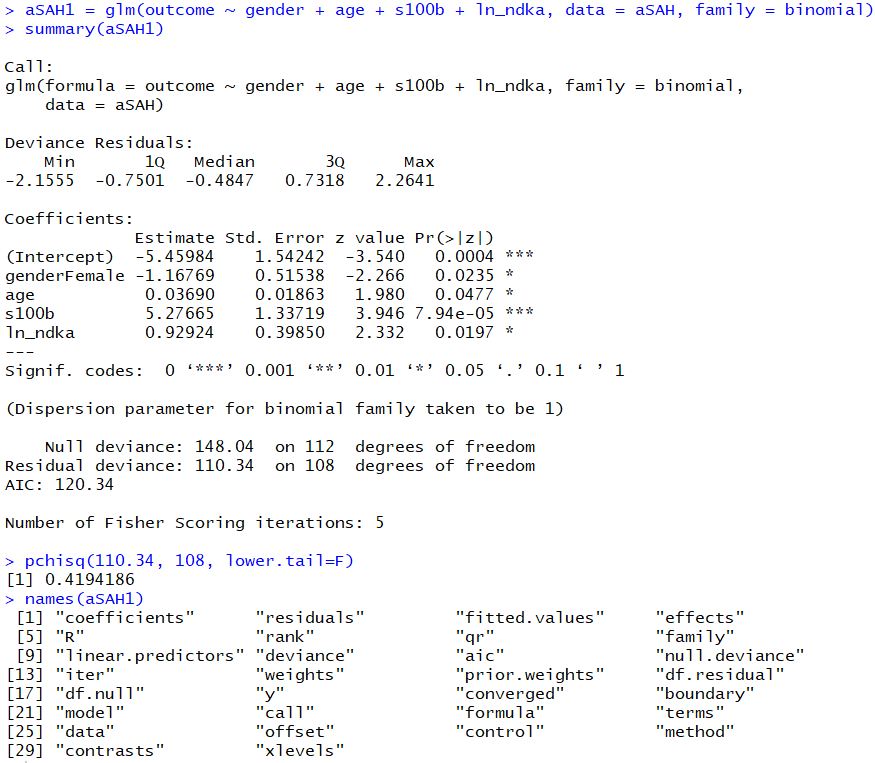
**Answer**:

The ATP gamma phosphate is transferred to the NDP beta phosphate via a ping-pong mechanism, using a phosphorylated active-site intermediate. Major role in the synthesis of nucleoside triphosphates other than ATP.

**Question 2**:

Fit a logistic regression model with response variable ‘outcome’ and predictors s100b, log(ndka), gender, and age. Identify significant predictors. Identify the most significant predictor. Check goodness-of-fit of the model. Write the prediction equation. 3 + 3 +1 + 2 + 2 points

**Answer**:



Significant predictors = gender/female, age, s100b, ln\_ndka

Most significant predictor = s100b

p-value of chi sq test = 0.4194 > 0.05, so we cannot reject the null hypothesis, therefore, model adequately summarizes the data.

**Prediction Equation**

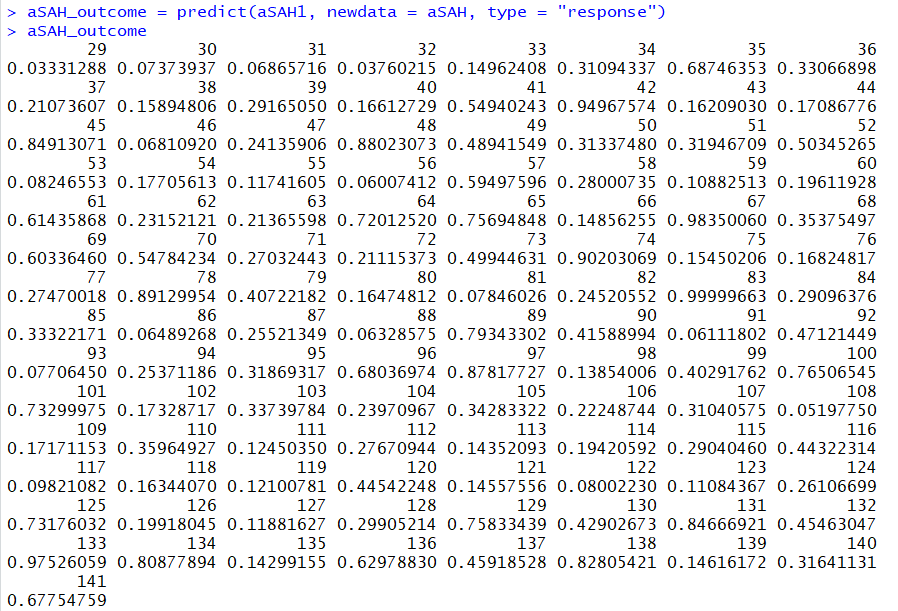
# Pr(Poor) =

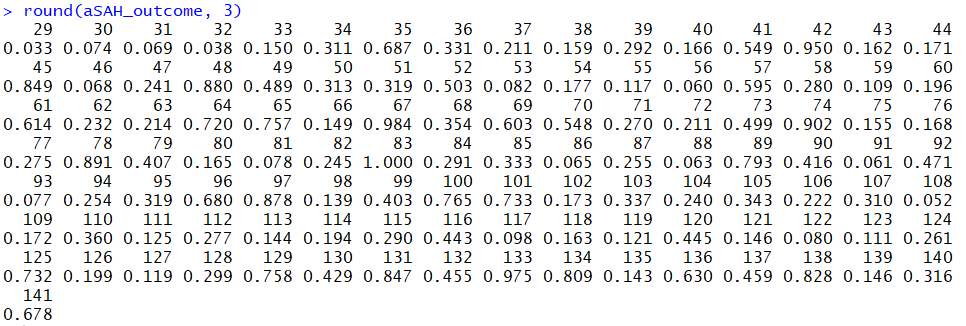
exp(-5.45 – 1.17\*genderFemale + 0.04\*age + 5.27\*s100b+ 0.93\*ln\_ndka)/ (1 + exp(-5.45 – 1.17\*genderFemale + 0.04\*age + 5.27\*s100b+ 0.93\*ln\_ndka)

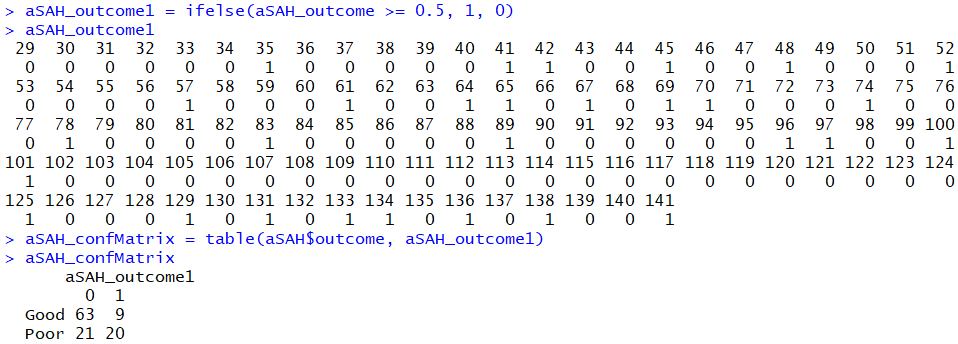
**Question 3**:

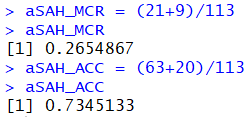
Obtain the confusion matrix of the model in Question 2. Spell out the accuracy of the prediction model. 3 + 1 points

**Answer**:







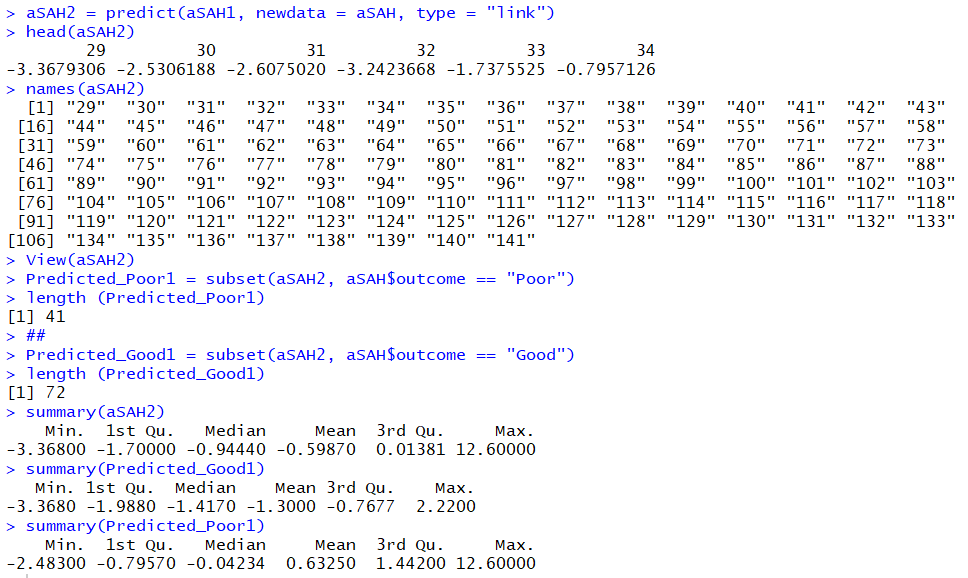


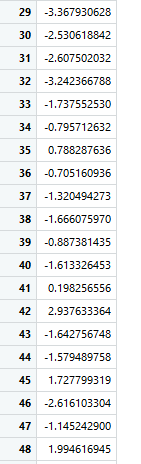
**Question 4**:

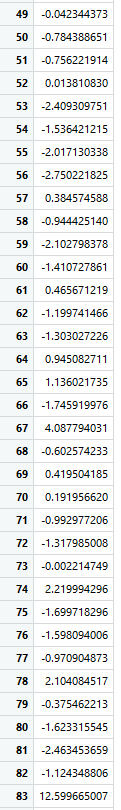
Write formally the link function of the model, which we will use as a single biomarker X.

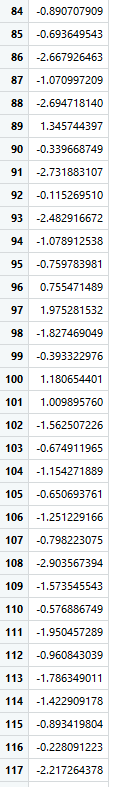
2 points

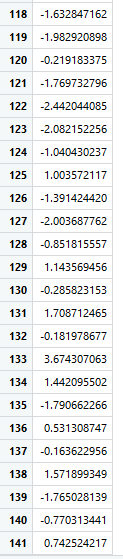
**Answer**:











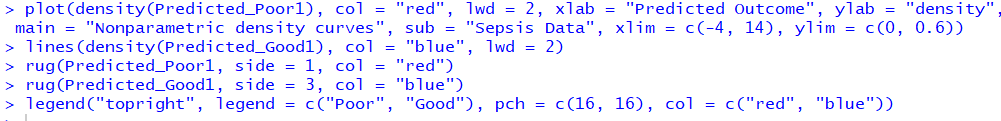


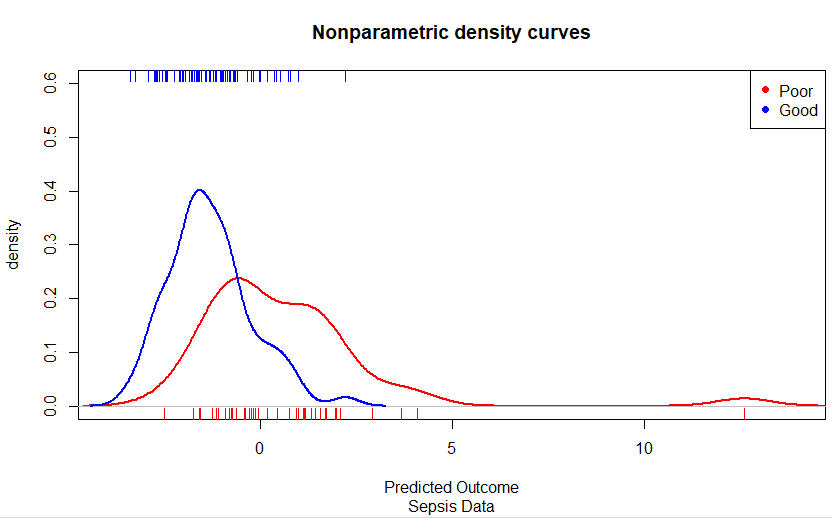
Link function = ln(Pr(Poor)/Pr(Good)) = linear combination of s100b, log(ndka), gender, and age.

**Question 5**:

Get the non-parametric density curves of the biomarker X under each outcome in the frame. Comment on the densities. Comment. 3 + 1 points

**Answer**:



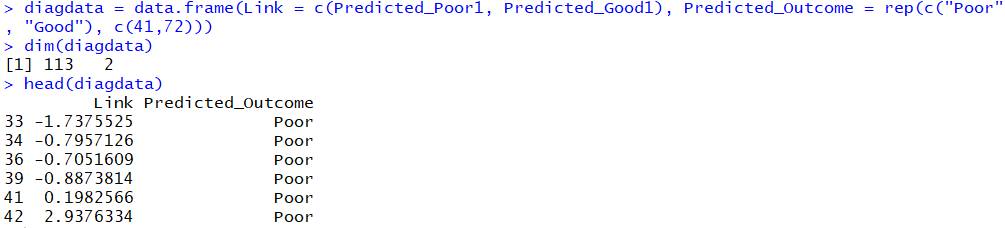


**Question 6**:

Outline how a diagnostic test looks like for the biomarker X. 2 points

**Answer**:

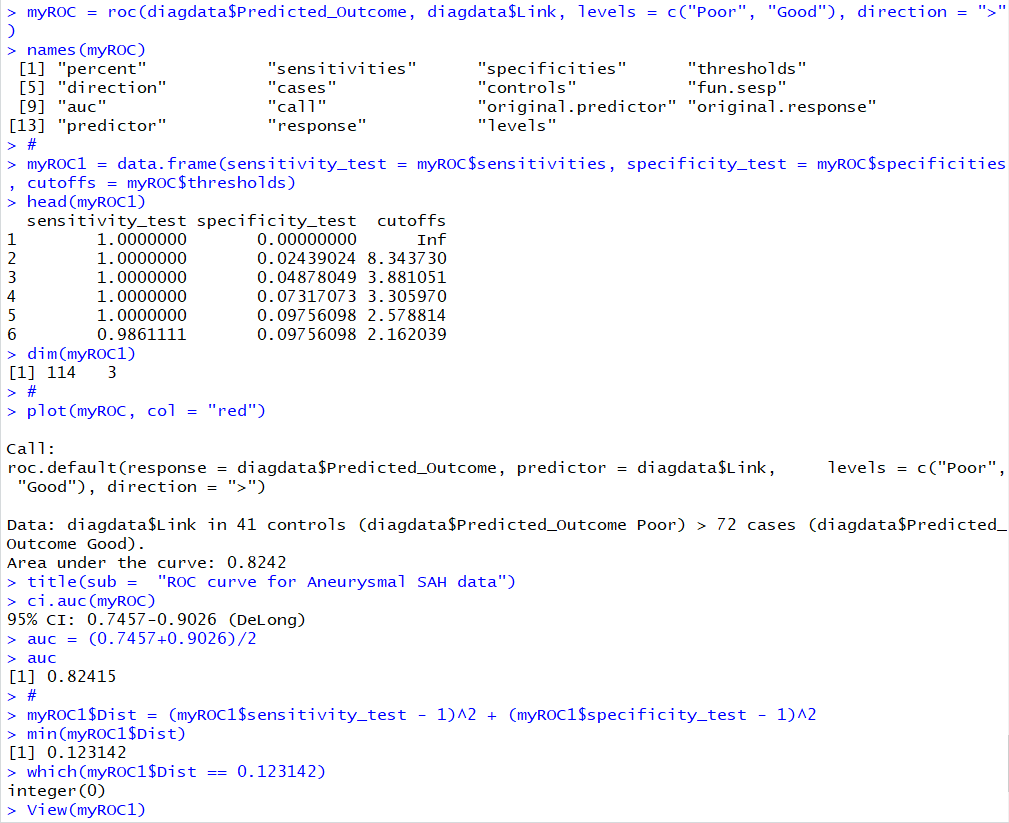
The diagnostic test will be the combination of predictors used in the logistic regression model, as one entity

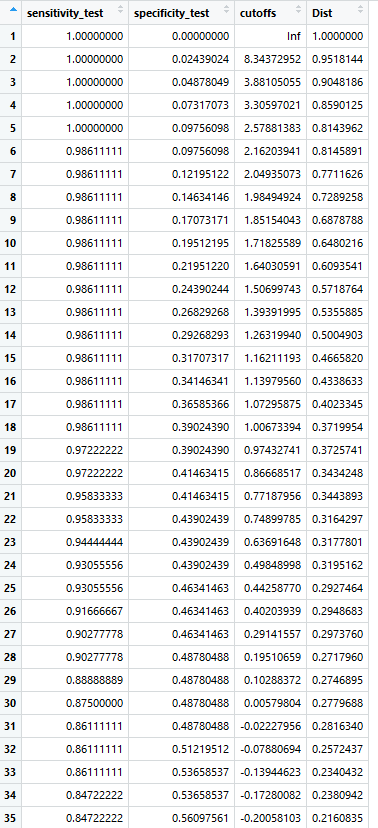


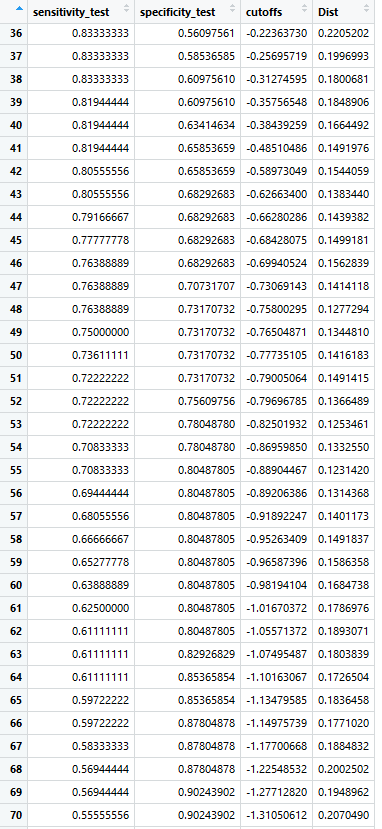
**Question 7**:

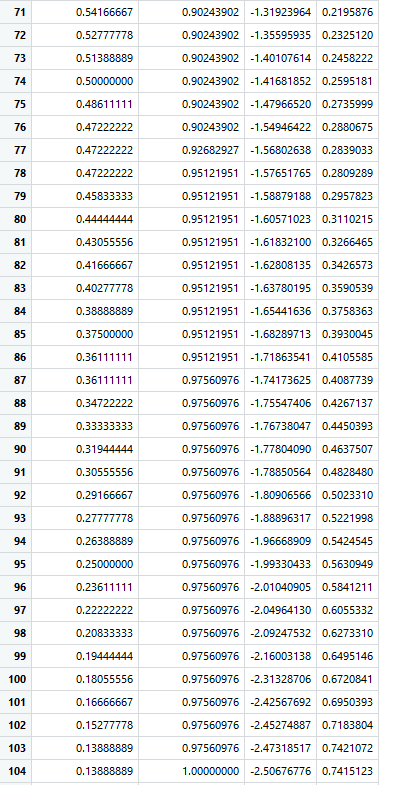
Invoke the ‘roc’ function with the link function X as a biomarker. Plot the ROC curve that comes with it. 8 points

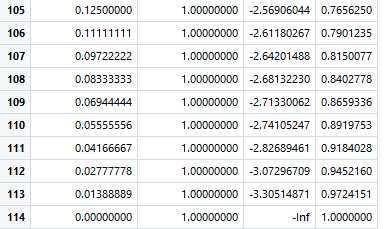
**Answer**:

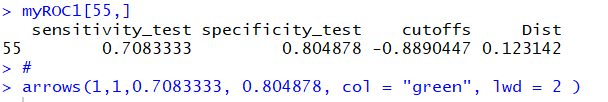


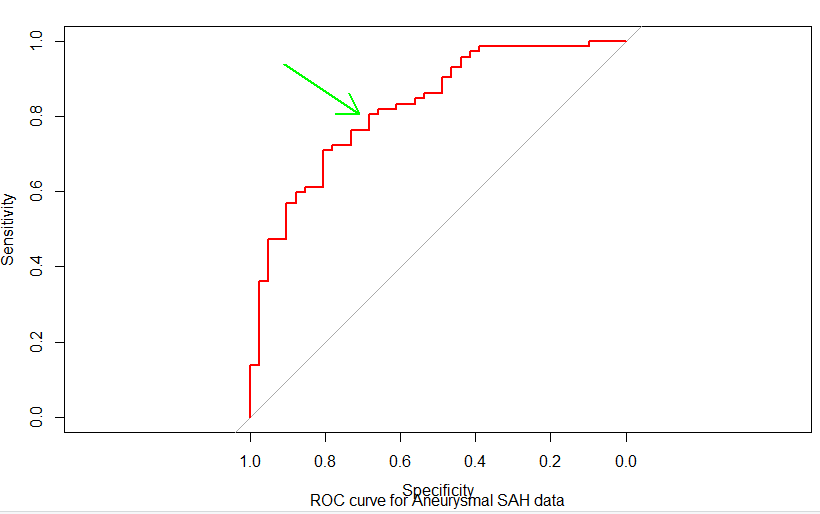








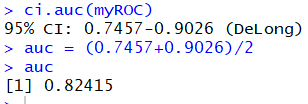




**Question 8**:

Find a 95% confidence interval for AUC. Comment on the utility of the biomarker ‘X.’ 5 points

**Answer**:



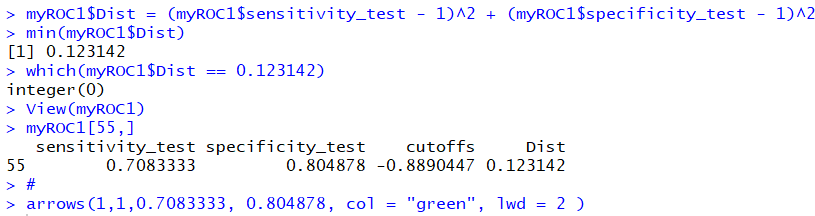
82% is a good value for the test

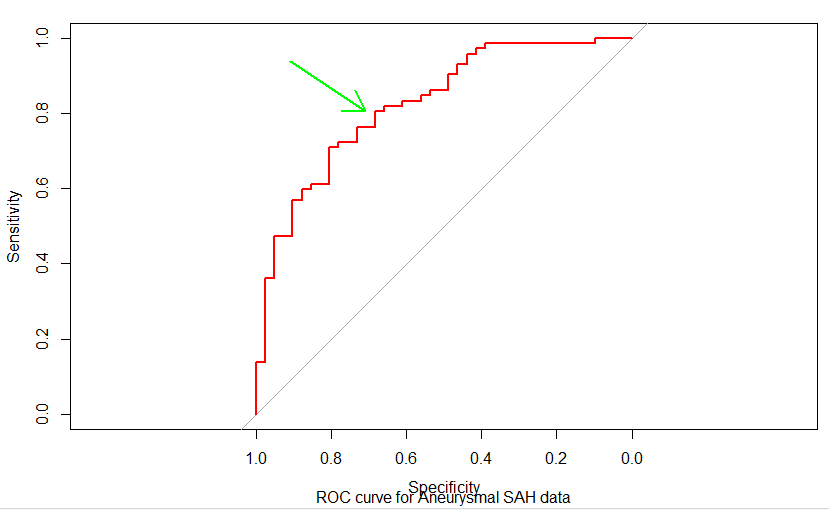
**Question 9**:

Develop a diagnostic test based on X.

6 points

**Answer**:

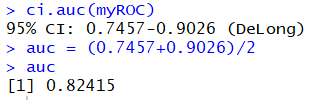


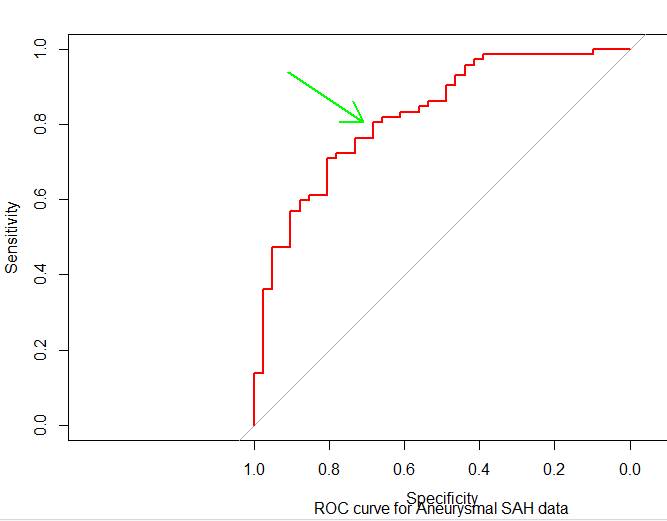


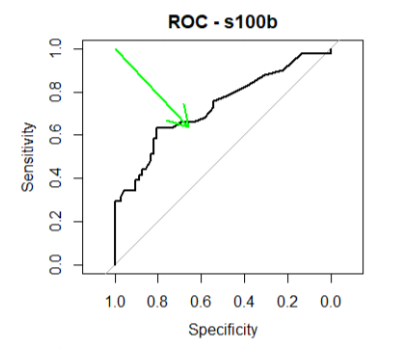
**Question 10**:

Compare the biomarkers s100b and X. 6 points

**Answer**:

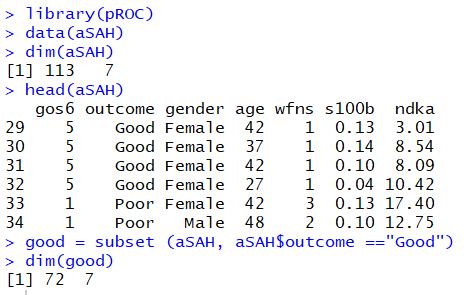


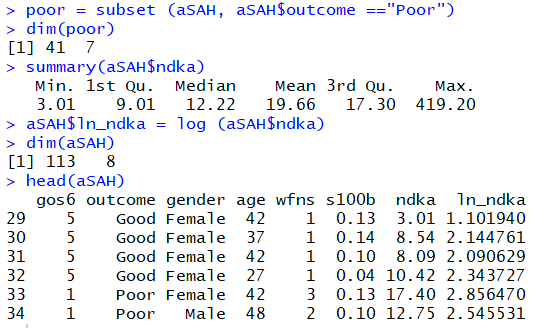


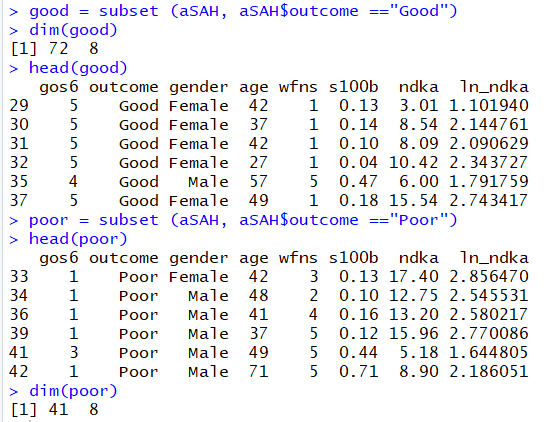


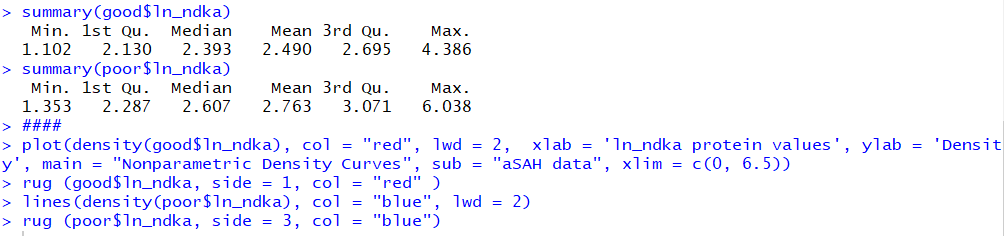
Based on the observed AUCs, the combined biomarker seems better than s100b alone as a diagnostic test

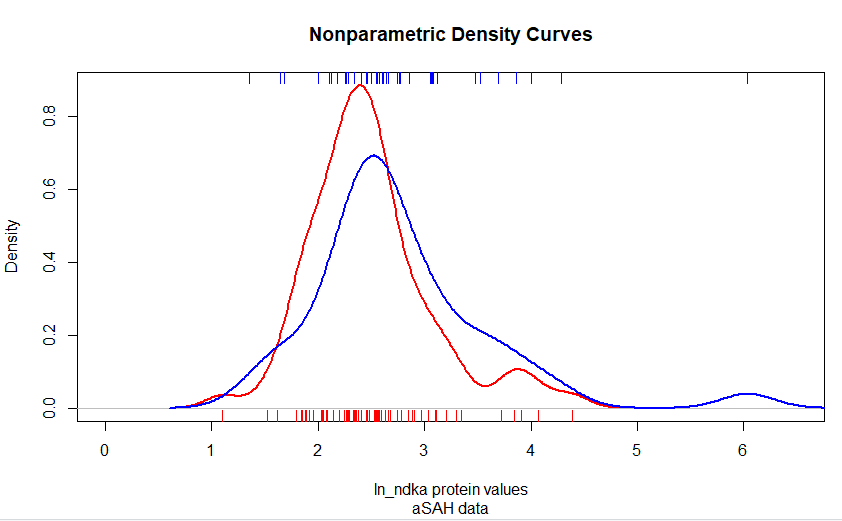
**Step vise Execution:**

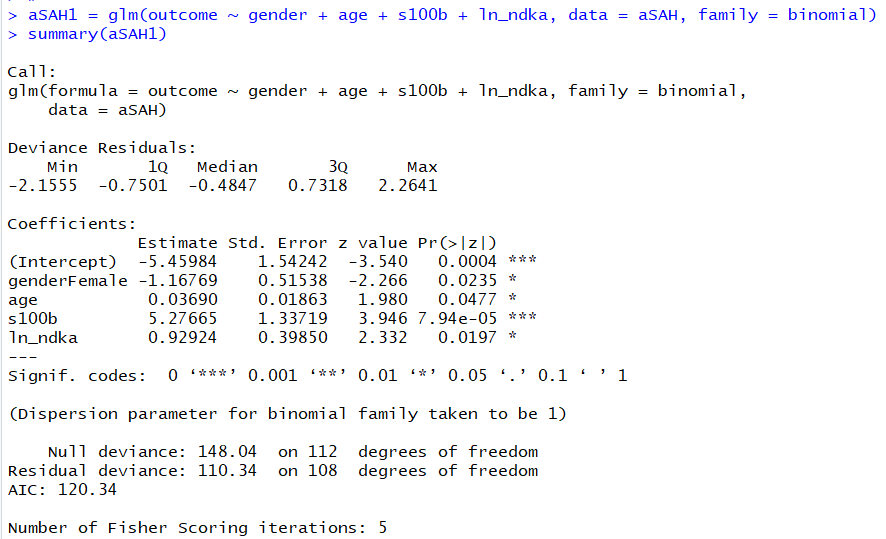


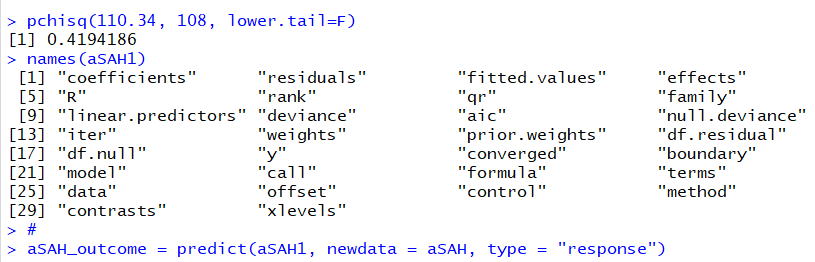


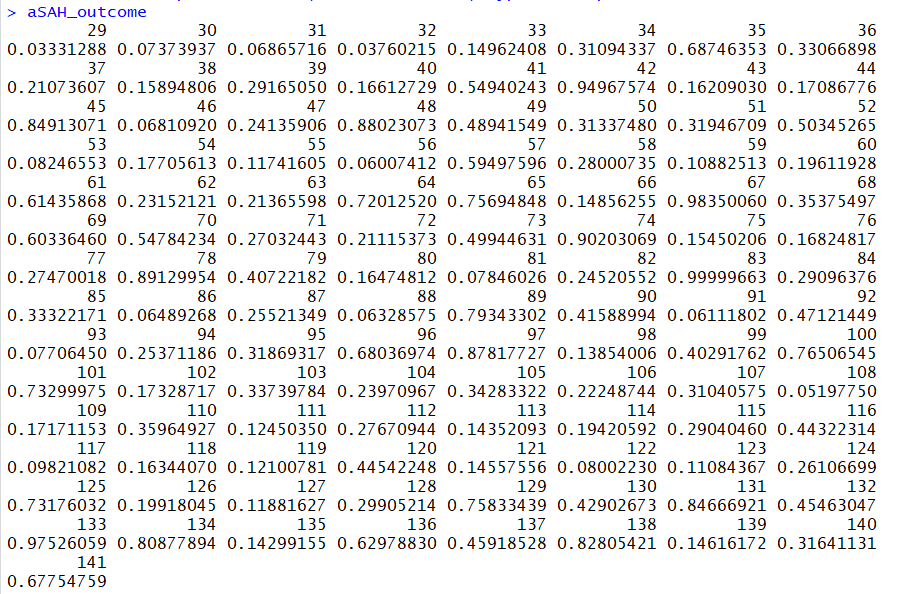


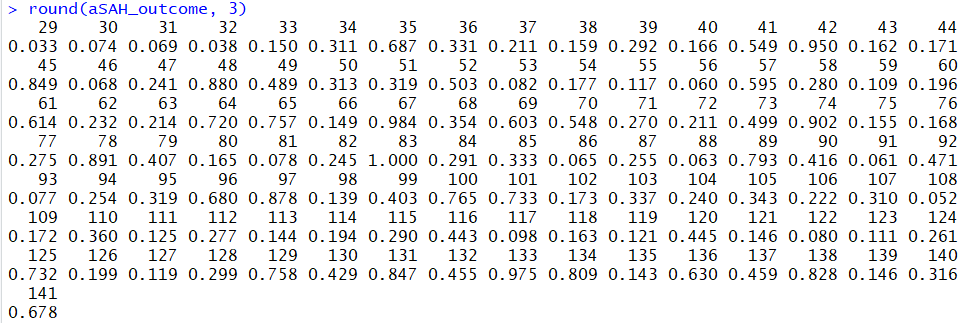


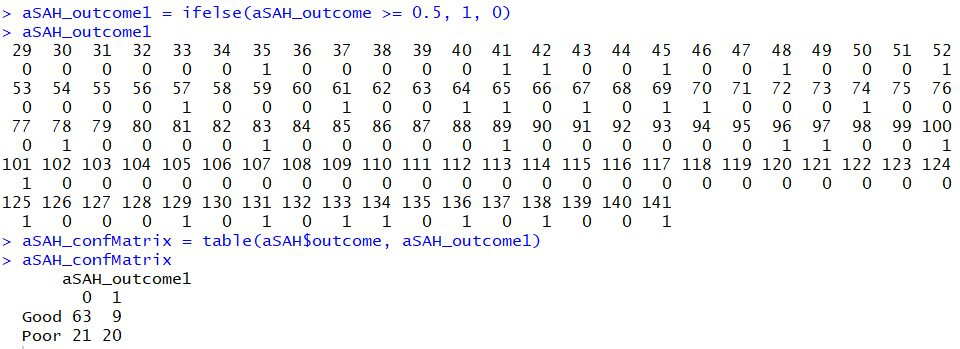


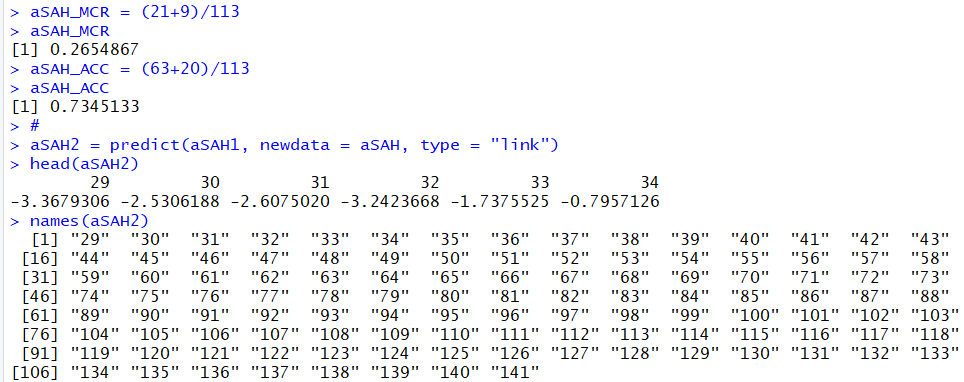






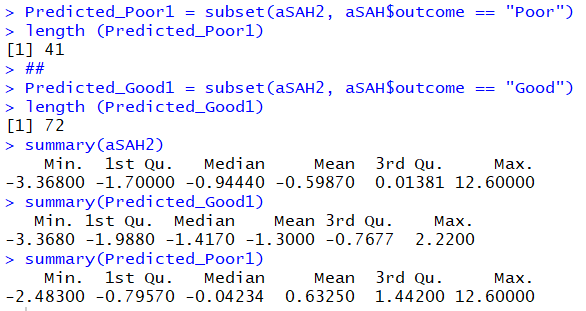


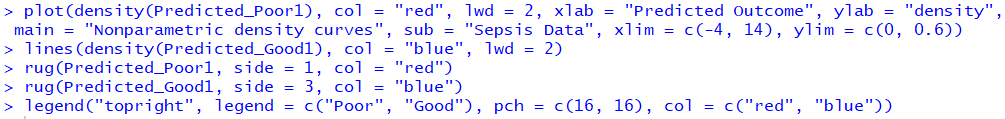


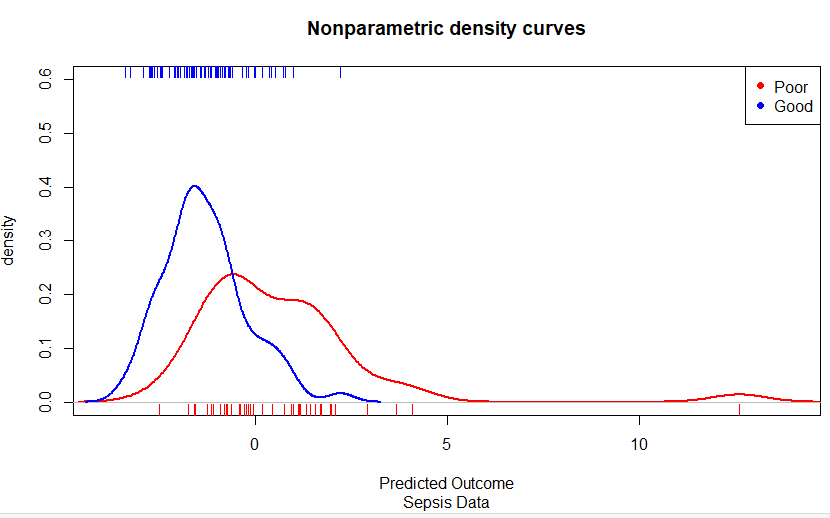


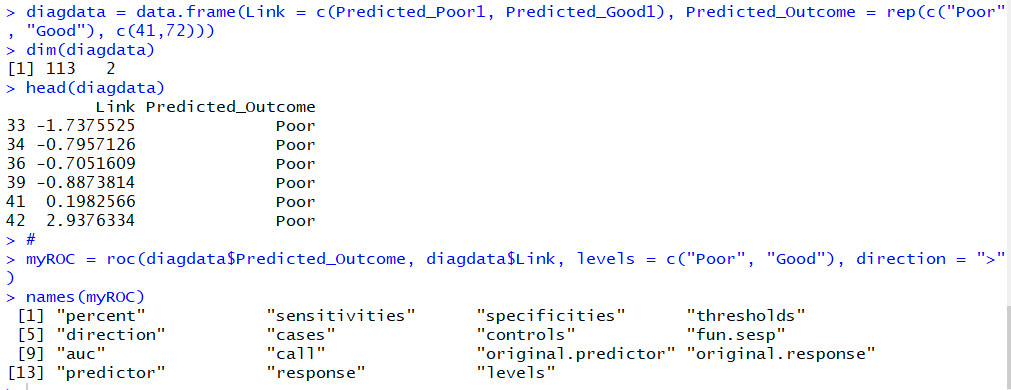


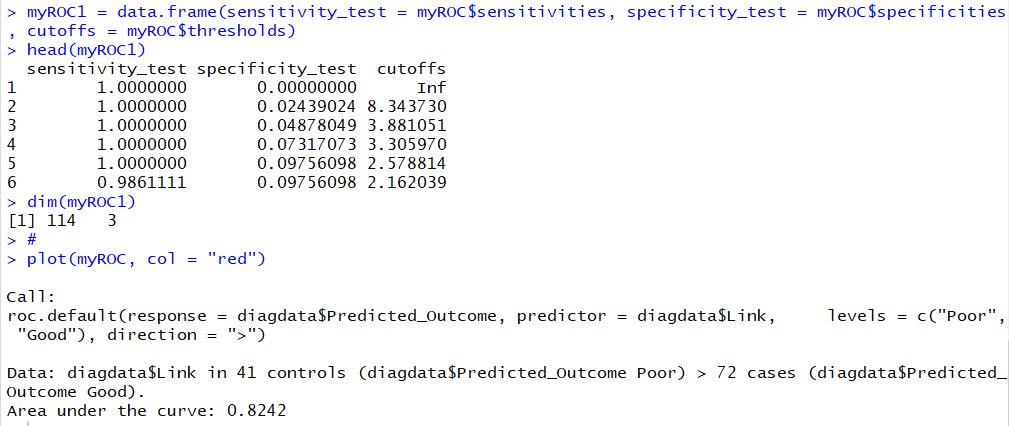
This step generates a huge output

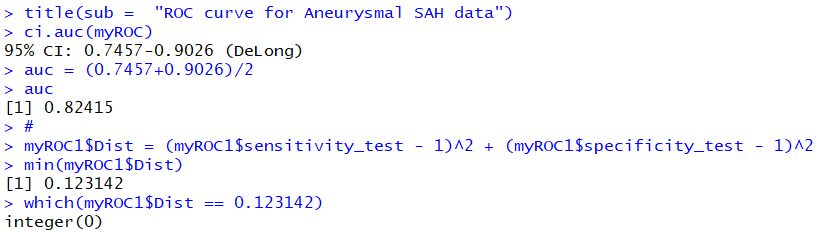






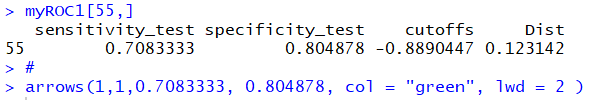


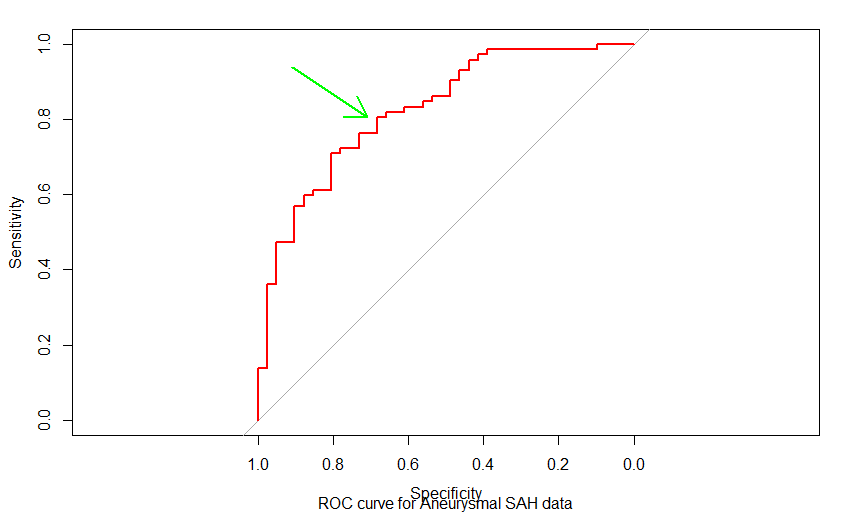






Note: Huge table is generated after you run this





**Entire Script:**

library(pROC)

data(aSAH)

dim(aSAH)

head(aSAH)

good = subset (aSAH, aSAH$outcome =="Good")

dim(good)

poor = subset (aSAH, aSAH$outcome =="Poor")

dim(poor)

summary(aSAH$ndka)

aSAH$ln\_ndka = log (aSAH$ndka)

dim(aSAH)

head(aSAH)

good = subset (aSAH, aSAH$outcome =="Good")

dim(good)

head(good)

poor = subset (aSAH, aSAH$outcome =="Poor")

head(poor)

dim(poor)

summary(good$ln\_ndka)

summary(poor$ln\_ndka)

####

plot(density(good$ln\_ndka), col = "red", lwd = 2, xlab = 'ln\_ndka protein values', ylab = 'Density', main = "Nonparametric Density Curves", sub = "aSAH data", xlim = c(0, 6.5))

rug (good$ln\_ndka, side = 1, col = "red" )

lines(density(poor$ln\_ndka), col = "blue", lwd = 2)

rug (poor$ln\_ndka, side = 3, col = "blue")

####

# P(death)/P(Survival)= e^(beta0+ beta1V1 + .....)

# ln P(death)/P(Survival)= beta0+ beta1V1 + ..... = X = link function of the biomarkers

#

aSAH1 = glm(outcome ~ gender + age + s100b + ln\_ndka, data = aSAH, family = binomial)

summary(aSAH1)

pchisq(110.34, 108, lower.tail=F)

names(aSAH1)

#

aSAH\_outcome = predict(aSAH1, newdata = aSAH, type = "response")

aSAH\_outcome

round(aSAH\_outcome, 3)

aSAH\_outcome1 = ifelse(aSAH\_outcome >= 0.5, 1, 0)

aSAH\_outcome1

aSAH\_confMatrix = table(aSAH$outcome, aSAH\_outcome1)

aSAH\_confMatrix

aSAH\_MCR = (21+9)/113

aSAH\_MCR

aSAH\_ACC = (63+20)/113

aSAH\_ACC

#

aSAH2 = predict(aSAH1, newdata = aSAH, type = "link")

head(aSAH2)

names(aSAH2)

View(aSAH2)

#

Predicted\_Poor1 = subset(aSAH2, aSAH$outcome == "Poor")

length (Predicted\_Poor1)

##

Predicted\_Good1 = subset(aSAH2, aSAH$outcome == "Good")

length (Predicted\_Good1)

summary(aSAH2)

summary(Predicted\_Good1)

summary(Predicted\_Poor1)

#

plot(density(Predicted\_Poor1), col = "red", lwd = 2, xlab = "Predicted Outcome", ylab = "density", main = "Nonparametric density curves", sub = "Sepsis Data", xlim = c(-4, 14), ylim = c(0, 0.6))

lines(density(Predicted\_Good1), col = "blue", lwd = 2)

rug(Predicted\_Poor1, side = 1, col = "red")

rug(Predicted\_Good1, side = 3, col = "blue")

legend("topright", legend = c("Poor", "Good"), pch = c(16, 16), col = c("red", "blue"))

#

diagdata = data.frame(Link = c(Predicted\_Poor1, Predicted\_Good1), Predicted\_Outcome = rep(c("Poor", "Good"), c(41,72)))

dim(diagdata)

head(diagdata)

#

myROC = roc(diagdata$Predicted\_Outcome, diagdata$Link, levels = c("Poor", "Good"), direction = ">")

names(myROC)

#

myROC1 = data.frame(sensitivity\_test = myROC$sensitivities, specificity\_test = myROC$specificities, cutoffs = myROC$thresholds)

head(myROC1)

dim(myROC1)

#

plot(myROC, col = "red")

title(sub = "ROC curve for Aneurysmal SAH data")

ci.auc(myROC)

auc = (0.7457+0.9026)/2

auc

#

myROC1$Dist = (myROC1$sensitivity\_test - 1)^2 + (myROC1$specificity\_test - 1)^2

min(myROC1$Dist)

which(myROC1$Dist == 0.123142)

View(myROC1)

myROC1[55,]

#

arrows(1,1,0.7083333, 0.804878, col = "green", lwd = 2 )