Homework 1

STAT 117

Kent Codding

February 6, 2025

```
# install.packages('knitr') # if not installed, run this line for knitr installation.
library(knitr)
```

```
## Warning: package 'knitr' was built under R version 4.4.2
```

```
library(curatedOvarianData)
library(pROC)
```

```
## Warning: package 'pROC' was built under R version 4.4.2
```

```
data(TCGA_eset)
# Refer to Section 1 material if the above two lines have errors such as 'there is no
# package called 'curatedOvarianData'', as it indicates the required packages are not
# installed.
XX = as.matrix(cbind(exprs(TCGA_eset)))
YY = 1 * as.vector(pData(TCGA_eset)[, "debulking"] == "suboptimal")
XX = XX[, !is.na(YY)]
YY = YY[!is.na(YY)]
```

limit data to three specified biomarkers

```
CXCL12 <- XX[rownames(XX) == "CXCL12", ]
POSTN <- XX[rownames(XX) == "POSTN", ]
BRCA1 <- XX[rownames(XX) == "BRCA1", ]
```

1 Build an ROC curve for each and compare them on the same graph.

```
# use roc funct from pROC lib
plot(roc(YY, CXCL12), col = 'blue')
```

```
## Setting levels: control = 0, case = 1
```

```
## Setting direction: controls < cases
```

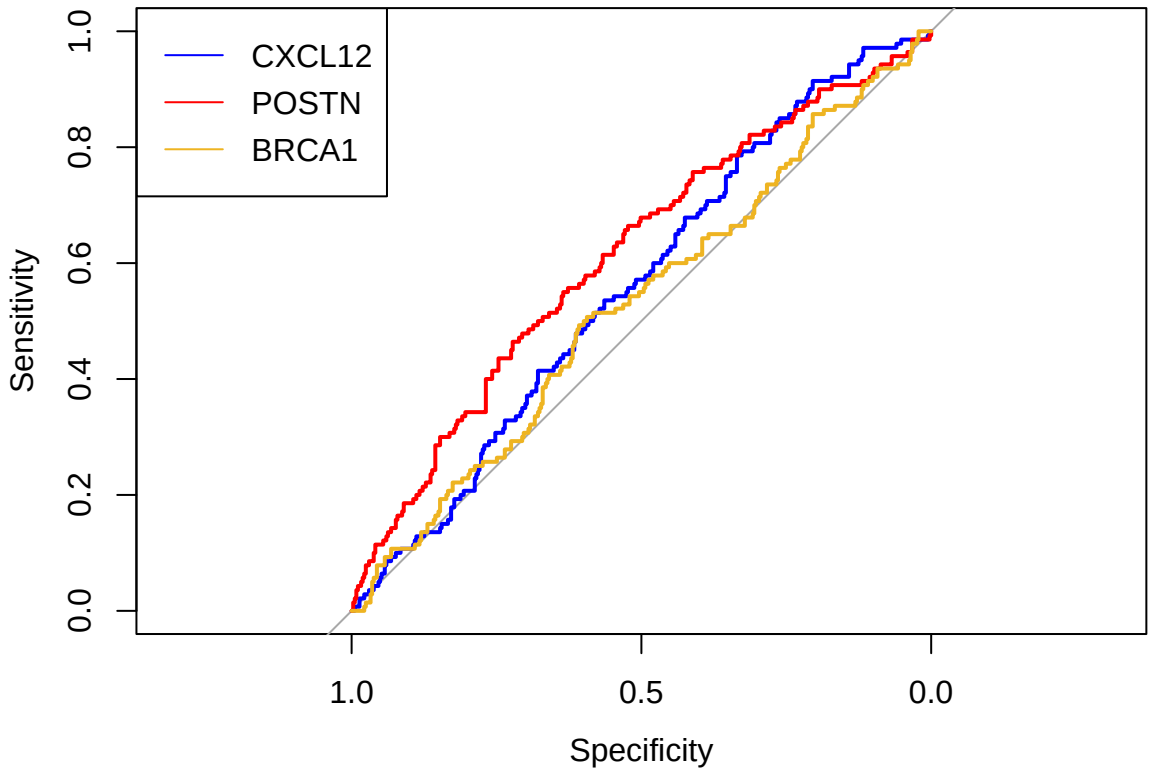
```
lines(roc(YY, POSTN), col = 'red')
```

```
## Setting levels: control = 0, case = 1
## Setting direction: controls < cases
```

```
lines(roc(YY, BRCA1), col = 'goldenrod2')
```

```
## Setting levels: control = 0, case = 1
## Setting direction: controls < cases
```

```
legend("topleft", legend = c("CXCL12", "POSTN", "BRCA1"),
       col = c("blue", "red", "goldenrod2"),
       lty = 1,
       cex = 1)
```



Based on the ROC curves above, the red line representing POSTN appears to have the highest AUC, suggesting that the Sensitivity to Specificity ratio is most generalizable across all thresholds and that POSTN could be an optimal biomarker to predict debulking.

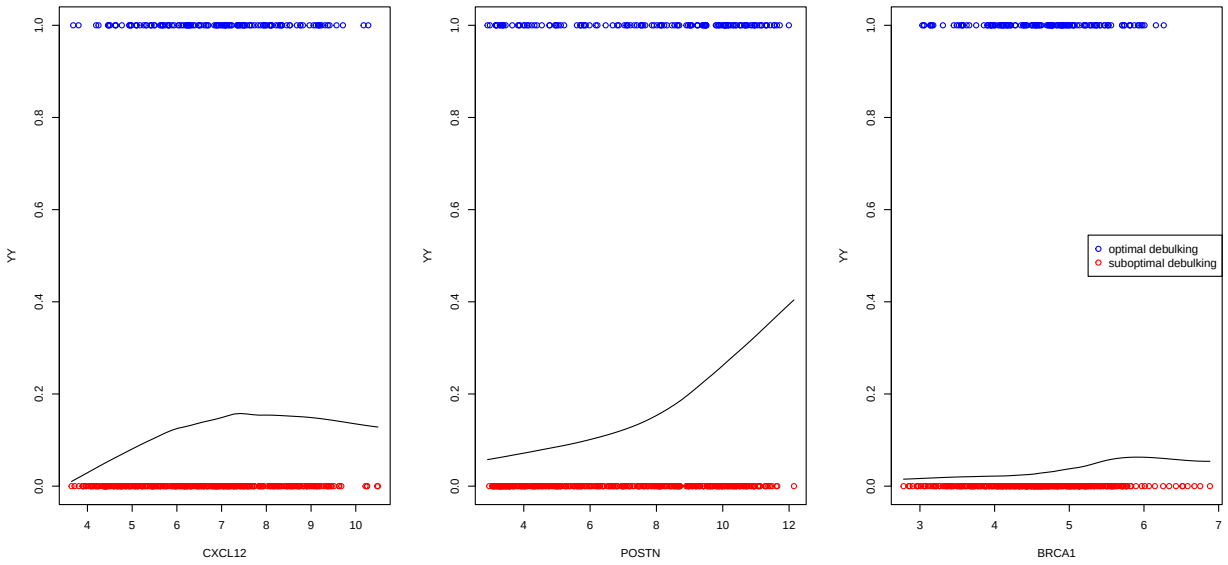
2) Generate scatterplots for all 3 pairs of biomarkers, coloring the "optimal" points differently from the "suboptimal" points. Comment based on the graph about whether combining these markers may provide better discrimination.

individual biomarkers

```
par(mfrow = c(1,3))
scatter.smooth(YY ~ CXCL12,
               col = ifelse(YY == 1, "blue", "red"))
scatter.smooth(YY ~ POSTN,
               col = ifelse(YY == 1, "blue", "red"),
               ylab = NULL)
scatter.smooth(YY ~ BRCA1,
               col = ifelse(YY == 1, "blue", "red"),
               ylab = NULL)

legend("right", legend = c("optimal debulking", "suboptimal debulking"),
```

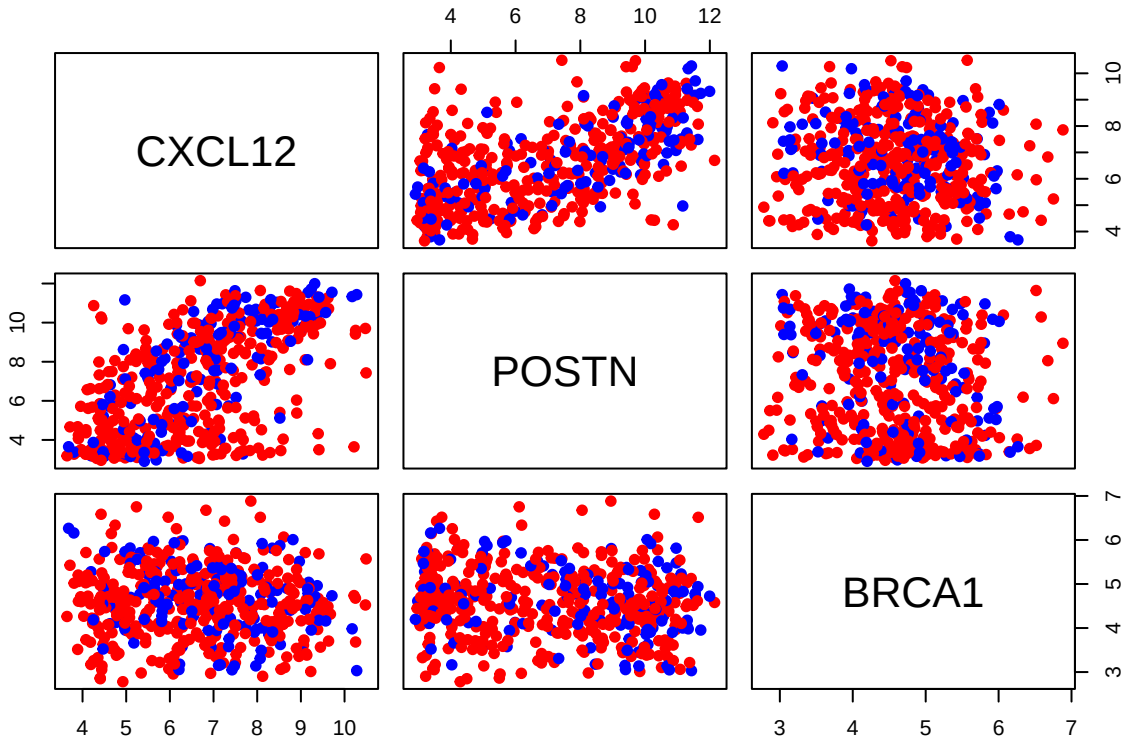
```
        col = c("blue", "red"),
        pch = 21,
        )
```



all 3 pairs of biomarkers

Note: Each of the 3 scatterplots should have the expression values on the y and x axis, so that it is a bivariate plot, NOT 3 separate dot plots for each gene. Hint: When might the average of "CXCL12" and "POSTN" predict better than each of the biomarkers alone?

```
pairs(
  data.frame(CXCL12, POSTN, BRCA1),
  bg = ifelse(YY == 1, "blue", "red"),
  col = ifelse(YY == 1, "blue", "red"),
  pch = 21,
)
```



based on the pairwise chart, combining biomarkers may provide better discrimination. For instance, the middle left chart with `CXCL12` on the x-axis and `POSTN` on the y-axis shows a clear separation of red points when expression of `CXCL12` is high and `POSTN` is low, suggests that suboptimal debulking is likely to occur. As follows, it may be better to consider both biomarkers when determining a threshold instead of a singular dot plot.

3) For the gene "POSTN", graph the probability of optimal debulking as a function of gene expression, at a prevalence of .25.

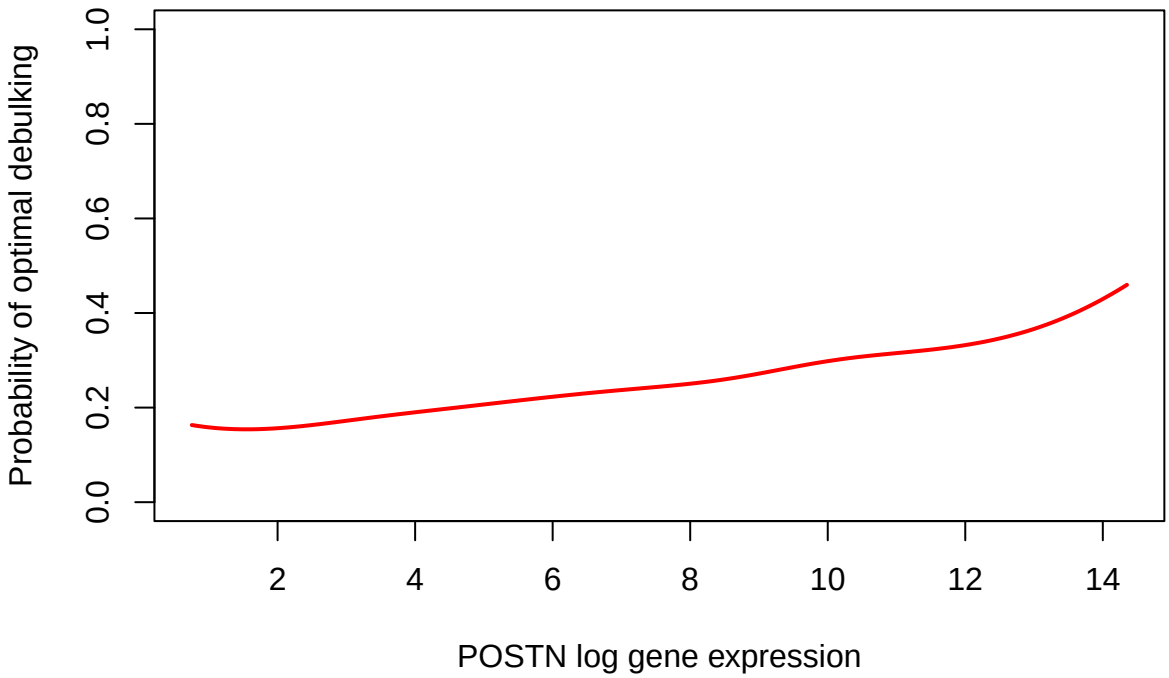
```
# get sequence of x values from gene exp range for POSTN
ffx = density(POSTN[YY==FALSE])$x
              #from=range(POSTN)[1],
              #to=range(POSTN)[2])$x

# get density of suboptimal debulking: 0 at above x-values
ff0 = density(POSTN[YY==FALSE],from=range(POSTN)[1],
              to=range(POSTN)[2])$y
ff1 = density(POSTN[YY==TRUE],from=range(POSTN)[1],
              to=range(POSTN)[2])$y

# calculate ppv from Bayes rule
prev = .25
ppv = prev * ff1 / (prev * ff1 + (1-prev) * ff0)

plot(ffx,ppv,ylim=c(0,1),
     xlab=c("POSTN log gene expression"),
```

```
    ylab="Probability of optimal debulking",
    type = "l",
    col = "red",
    lwd=2)
```



The probability of optimal debulking increases as the log-transformed `POSTN` gene expression increases according to Bayes rule with a prevalence set to *0.25*.