SF_JUN

2022-06-15

The gene ERBB2 (also known as HER2 or Neu) is amplified in ~30% of breast cancers. The region of actual amplification varies among patients and can include several other genes. SRCIN1 (also known as P140) is a gene located ~1 Mbp away from ERBB2 on chromosome 17 and sometimes, but not always, co-amplified with ERBB2.

The data in the file "test_2106.RData" concern 1071 breast cancer patients. For each patient the data contain:

- the expression levels of the genes ERBB2 and SRCIN1 measured in TPM (transcripts per million), not in log scale
- the ERBB2 genomic amplification status (not available for all patients)

```
load("C:/Users/seren/Desktop/test_2106.RData")
head(data)
```

```
##              ERBB2_expr SRCIN1_expr   her2_status
## TCGA-3C-AAAU   36.77080    3.033554      Negative
## TCGA-3C-AALI  972.58669   17.174349      Positive
## TCGA-3C-AALJ   54.89144    5.288192 Indeterminate
## TCGA-3C-AALK  165.80549   14.136236      Positive
## TCGA-4H-AAAK   60.51770    4.198549     Equivocal
## TCGA-5L-AAT0   46.80810    3.792047      Negative
```

```
dim(data)
```

```
## [1] 1071    3
```

Before getting to the questions, let's manipulate this data frame properly:

1. logarithmic transformation of expression data (ERBB2_expr, SRCIN1_expr):

```
logtransf <- function(x, base = 2, a =1) {
  log(x + a, base = base)
}

datalog <- data
datalog$ERBB2_expr <- logtransf(data$ERBB2_expr)
datalog$SRCIN1_expr <- logtransf(data$SRCIN1_expr)

head(datalog)
```

```
##              ERBB2_expr SRCIN1_expr   her2_status
## TCGA-3C-AAAU   5.239199    2.012051      Negative
## TCGA-3C-AALI   9.927166    4.183832      Positive
## TCGA-3C-AALJ   5.804556    2.652645 Indeterminate
## TCGA-3C-AALK   7.382023    3.919935      Positive
## TCGA-4H-AAAK   5.942930    2.378109     Equivocal
## TCGA-5L-AAT0   5.579183    2.260642      Negative
```

2. shrinking of the data frame (as "Indeterminate" and "Equivocal" her2_status are not informative):

```
datalog_red <- datalog[datalog$her2_status == "Positive" | datalog$her2_status == "Negative", ]
dim(datalog_red)
```

```
## [1] 709   3
```

3. work on T and F (logical) values instead of "Positive" and "Negative" (character) values:

```
#we add a new variable for this purpose:
datalog_red$her2_num <- FALSE
datalog_red[datalog_red$her2_status == "Positive", "her2_num"] <- TRUE
head(datalog_red)
```

```
##              ERBB2_expr SRCIN1_expr her2_status her2_num
## TCGA-3C-AAAU   5.239199    2.012051    Negative    FALSE
## TCGA-3C-AALI   9.927166    4.183832    Positive     TRUE
## TCGA-3C-AALK   7.382023    3.919935    Positive     TRUE
## TCGA-5L-AAT0   5.579183    2.260642    Negative    FALSE
## TCGA-A1-A0SB   5.059755    1.742386    Negative    FALSE
## TCGA-A1-A0SD   6.057911    1.764140    Negative    FALSE
```

Use the appropriate regression methods to answer the following questions:

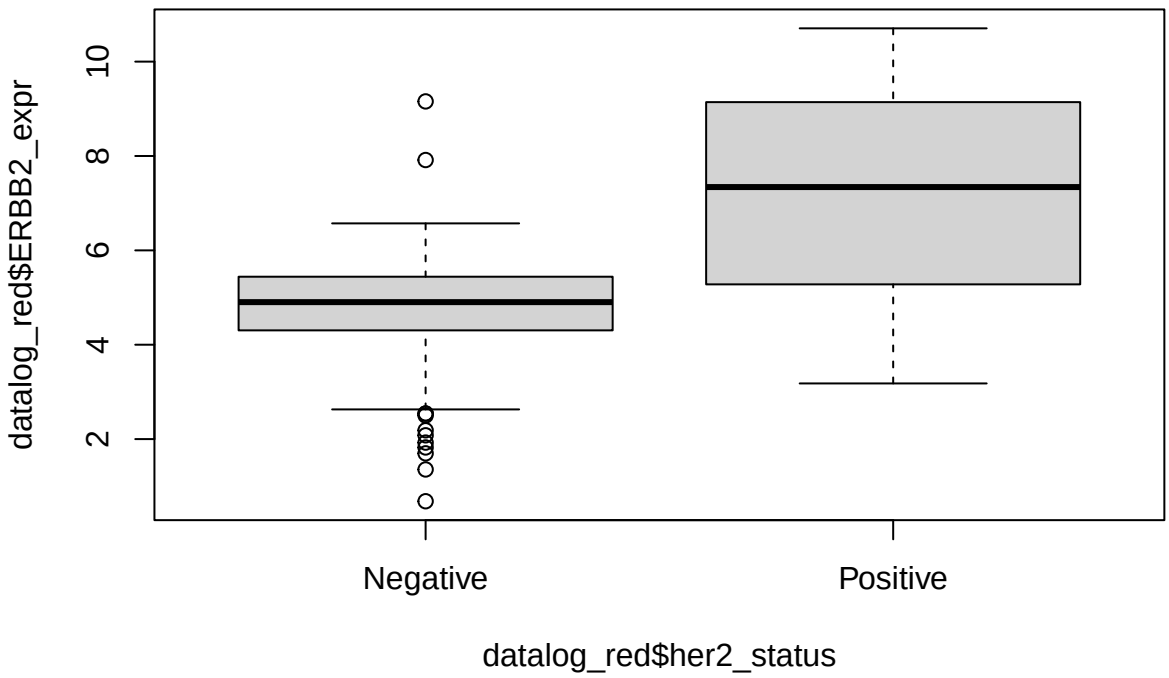
1. Is ERBB2 expression a predictor of ERBB2 amplification?

Let's see graphically the relationship between these two variables:

```
datalog_red$her2_status <- factor(datalog_red$her2_status)
class(datalog_red$her2_status)
```

```
## [1] "factor"
```

```
boxplot(datalog_red$ERBB2_expr ~ datalog_red$her2_status)
```



Since we need to predict a categorical variable (her2_num), let's use logistic regression:

```
ERBB2expr_amp <- glm(datalog_red$her2_num ~ datalog_red$ERBB2_expr, family = "binomial")
summary(ERBB2expr_amp)$coefficients
```

```
##                           Estimate Std. Error   z value     Pr(>|z|)
## (Intercept)             -7.785368  0.6053691 -12.86053 7.505958e-38
## datalog_red$ERBB2_expr   1.164315  0.1059346  10.99088 4.227908e-28
```

ERBB2 expression is a good predictor of HER2 amplification status (P-value = 4.227908e-28).

2. Can you improve such prediction by using also the expression of SRCIN1, together with that of ERBB2, as predictor?

To address this point, a multivariable linear regression is needed:

```
ERBSRCIN1expr_amp <- lm(datalog_red$her2_num ~ datalog_red$ERBB2_expr + datalog_red$SRCIN1_expr)
summary(ERBSRCIN1expr_amp)$coefficients
```

```
##                            Estimate  Std. Error    t value     Pr(>|t|)
## (Intercept)              -0.6245271 0.044469772 -14.043856 1.088646e-39
## datalog_red$ERBB2_expr    0.1731503 0.009169245  18.883809 1.079051e-64
## datalog_red$SRCIN1_expr -0.0458438 0.017379065  -2.637875 8.526470e-03
```

Yes, the prediction is improved when SRCIN1_expr is added, as at fixed ERBB2_expr the increase in SRCIN1 expression by 1 unit is correlated to a decrease of the probability to have HER2 amplification by 0.046 (and the associated P-value is pretty low, so good prediction).

3. Using a suitable training set, develop a model to predict SRCIN1 expression from ERBB2 expression and amplification status. Then evaluate its performance on the testing set composed of all patients that were not included in the training set.

To make a training and testing set in a unbiased, random way, we do this:

```
random <- sample(1:709, 709)

train_set <- datalog_red[random[1:473], ]
test_set <- datalog_red[random[474:709], ]
dim(train_set)
```

```
## [1] 473   4
```

```
dim(test_set)
```

```
## [1] 236   4
```

Let's fit a regression model on the training set:

```
lm_SRCIN1expr <- lm(SRCIN1_expr ~ ERBB2_expr + her2_num, data = train_set)
summary(lm_SRCIN1expr)$r.squared
```

```
## [1] 0.2253925
```

Let's evaluate this model on the testing set now:

```
pred_SRCIN1expr <- predict(lm_SRCIN1expr, newdata = test_set)
```

To see the R^2 associated to this prediction, we need the following function:

```
r2 <- function(y, y_pred) {
  1 - sum((y - y_pred)^2)/sum((y - mean(y))^2)
}

r2(test_set$SRCIN1_expr, pred_SRCIN1expr)
```

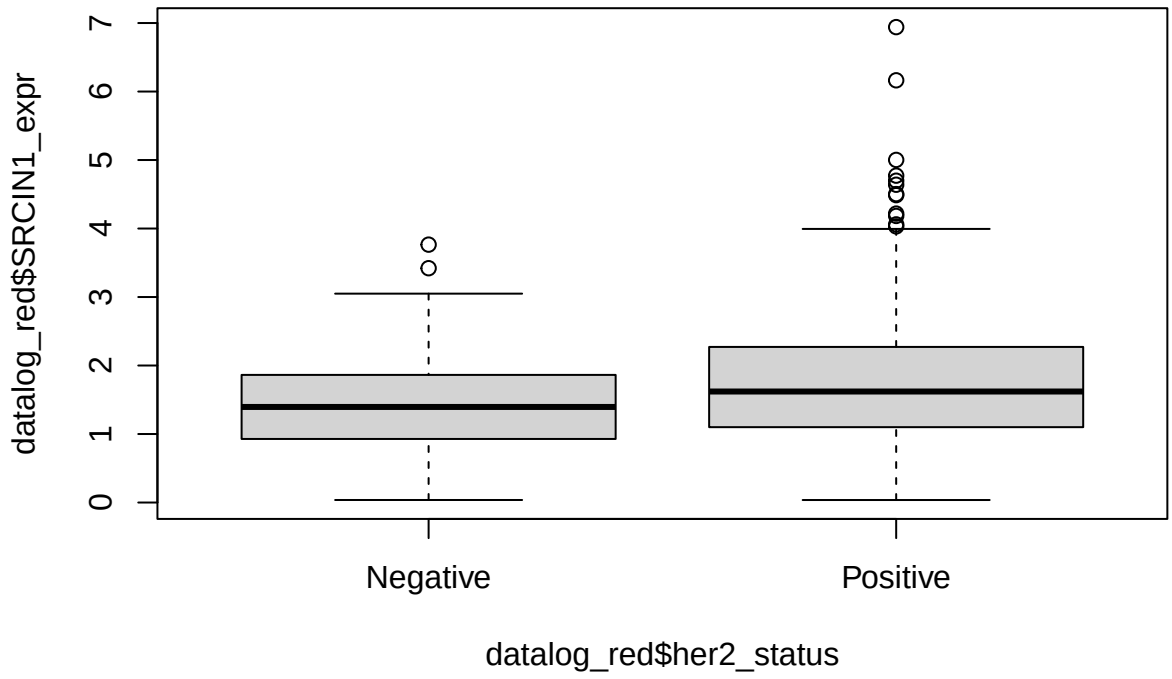
```
## [1] 0.2802597
```

If the R^2 comes out smaller than that associated with the model in the training set, it means that the model was just fiting noise in the training dataset (where it was built), and obviously in another dataset noise will be different so the model won't fit well. The model is not good.

4. *Using regression* answer the following question: is the expression of SRCIN1 in patients with ERBB2 amplification significantly different from that in patients without such amplification?

Let's first look at the aspect of this relationship:

```
boxplot(datalog_red$SRCIN1_expr ~ datalog_red$her2_status)
```



From this first look, it seems not. Let's look at the statistics though:

```
SRCIN1exp_amp <- lm(datalog_red$SRCIN1_expr ~ datalog_red$her2_status)
summary(SRCIN1exp_amp)$coefficients
```

```
##                                  Estimate Std. Error   t value      Pr(>|t|)
## (Intercept)                     1.4012634 0.03451555 40.598034 6.489914e-187
## datalog_red$her2_statusPositive 0.4444377 0.07220718  6.155034  1.257365e-09
```

Even though there is a only a slight increase in SRCIN1 expression when the status is HER2+, the associated P-value is pretty low so we're confident that this difference in SRCIN1 expression between HER+ and HER- patients is significant.