HW4: Multiple Testing and Prediction Modelling (Due : Wed Oct 27 on D2L)

For this HW you will use a prostate cancer data set originally made available in Stamey et al. (J. of Urology, 1989) which is also one of the datasets analyzed in the book *The Elements of Statistical Learning* (Hastie, Tibshirani, and Friedman, 2009).

You can read the data set into an R-environment using the following code

```
fname='https://raw.githubusercontent.com/gdlc/STAT_COMP/refs/heads/master/DATA/prostate.csv'
DATA=read.csv(fname,header=TRUE,row.names=1)

head(DATA)
```

```
##       lcavol  lweight age      lbph svi       lcp gleason pgg45       lpsa
## 1 -0.5798185 2.769459  50 -1.386294   0 -1.386294       6     0 -0.4307829
## 2 -0.9942523 3.319626  58 -1.386294   0 -1.386294       6     0 -0.1625189
## 3 -0.5108256 2.691243  74 -1.386294   0 -1.386294       7    20 -0.1625189
## 4 -1.2039728 3.282789  58 -1.386294   0 -1.386294       6     0 -0.1625189
## 5  0.7514161 3.432373  62 -1.386294   0 -1.386294       6     0  0.3715636
## 6 -1.0498221 3.228826  50 -1.386294   0 -1.386294       6     0  0.7654678
##   train
## 1  TRUE
## 2  TRUE
## 3  TRUE
## 4  TRUE
## 5  TRUE
## 6  TRUE
```

You can find more information about each of the variables included in the data set here, and here.

We have two objectives: (i) Identify clinical variables that are significantly associated with log-PSA, and (ii) Developing a prediction model for log-PSA.

The column `train` will be used to split the data set into a training and a testing data set. We will use the observations with `train=TRUE` to develop models and those with `train=FALSE` to evaluate the prediction accuracy of each of the models evaluated.data

```
DATA.TRN=DATA[DATA$train, -ncol(DATA)]
DATA.TST=DATA[!DATA$train,-ncol(DATA)]
```

Linear model using all the predictors

Fitting the model to the traning data set

```
fm0=lm(lpsa~.,data=DATA.TRN)
summary(fm0)
```

```
##
## Call:
## lm(formula = lpsa ~ ., data = DATA.TRN)
```

```
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.64870 -0.34147 -0.05424  0.44941  1.48675
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.429170   1.553588   0.276  0.78334
## lcavol       0.576543   0.107438   5.366 1.47e-06 ***
## lweight      0.614020   0.223216   2.751  0.00792 **
## age         -0.019001   0.013612  -1.396  0.16806
## lbph         0.144848   0.070457   2.056  0.04431 *
## svi          0.737209   0.298555   2.469  0.01651 *
## lcp         -0.206324   0.110516  -1.867  0.06697 .
## gleason     -0.029503   0.201136  -0.147  0.88389
## pgg45        0.009465   0.005447   1.738  0.08755 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7123 on 58 degrees of freedom
## Multiple R-squared:  0.6944, Adjusted R-squared:  0.6522
## F-statistic: 16.47 on 8 and 58 DF,  p-value: 2.042e-12
```

Q1: Determining significance, accounting fro multiple testing

1.1) What variables are significantly associated with `lpsa` after accounting for multiple testing using Bonferroni's method (use a family-wise error rate of 0.05).

1.2) Controlling error rate in multiple testing using permutations

Use permutation analysis to determine the p-value threshold that you should used to control the Family Wise Error Rate at a level < 0.05. Report the p-value threshold you selected and the variables that are significant when using a family-wise error rate of 0.05.

Q2: Evaluation of prediction accuracy

The prediction mean-squared error is defined as $PMSE = \frac{\Sigma_{i=1}^{n}(y_i - \hat{y}_i)^2}{n}$ where y_i is an observation (in a testing data set) and $\hat{y}_i$ is the prediction made for that observation derived using a model fitted using a traning data set.

Estimate the prediction mean-squared error (in testing data) of each of the following models

- M1: `lpsa~lcavol`
- M2: `lpsa~lcavol+lweight,`
- M3: `lpsa~lcavol+lweight+svi`
- M4: `lpsa~lcavol+lweight+svi+lbph`
- M5: `lpsa~lcavol+lweight+svi+lbph+age`
- M6: `lpsa~lcavol+lweight+svi+lbph+age+lcp`
- M7: `lpsa~lcavol+lweight+svi+lbph+age+lcp+gleason`
- M8: `lpsa~lcavol+lweight+svi+lbph+age+lcp+gleason+ pgg45`

Hint: Once you fit a mode to the training data, you can derive predictions using `predict(fittedModel,newdata=DATA.TST)`.

Report a table with PMSE for each of the models.

What model would you recommend?

Q3: Conclussions

- For which predictors you have strong evidence of association with lpsa?
- What model would you recommend if the objective is to predict lpsa of future patients based on the other clinical variables included in the data set?