TP2 : Model selection

1 Subset selection, AIC, BIC, Cp criteria

1. Open filedataset poumon describe in class with command `poumon <- read.table(file=file.choose(), sep=" ", header= TRUE, dec=" ")` with the good choice of separators. View the data and verify that they are well imported.
2. We want to predict TLCO by the covariates Origine, Sexe, Age, Taille, POIDS, BMI. Build the good table with these last seven variables only and the rownames. What are the values n and p in these data ? What is the number of possibilities for the subset of data in the model ?
3. Some remind about command `lm`. Try to understand what are the objects given by the command.

```
modelp <- lm(TLCO~ . , data=poum)
res<-summary(modelp)
res
```

```
##
## Call:
## lm(formula = TLCO ~ ., data = poum)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.1850 -3.4867 -0.2514  3.1729 13.4053
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -56.77701   36.35075  -1.562  0.11936
## origineG     -2.96015    0.63700  -4.647 5.05e-06 ***
## SexeM         4.32350    0.78303   5.522 7.28e-08 ***
## AGE          -0.22580    0.01656 -13.632  < 2e-16 ***
## TAILLE_EN_M  56.57009   21.77255   2.598  0.00983 **
## POIDS        -0.26902    0.26077  -1.032  0.30308
## BMI           0.87932    0.72463   1.213  0.22590
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.602 on 300 degrees of freedom
## Multiple R-squared:  0.7349, Adjusted R-squared:  0.7296
## F-statistic: 138.6 on 6 and 300 DF,  p-value: < 2.2e-16
```

```
head(res$residuals)
```

```
##         32         94        244        103        197         86
##  9.5130870 -7.2484727 -2.0982773 -3.2807845  6.0259039  0.6863815
```

```
RSS<-sum((res$residuals)^2)
RSS
```

```
## [1] 6354.346
```

```
#autre methode
pc<- predict(modelp, data=poum)
RSS2<-sum((pc-poum$TLCO)^2)
RSS2
```

```
## [1] 6354.346
```

```
sig<-(res$sigma)^2
res$r.squared
```

```
## [1] 0.7349047
```

```
res$adj.r.squared
```

```
## [1] 0.7296027
```

```
################
#pour obtenir AIC et BIC
#############
 Ak<-AIC(modelp)
Ak
```

```
## [1] 1817.453
```

```
#calcul manuel
RSS2/(sig)+n*log(sig)+n*log(2*pi)+2*(dim(poum)[2]+1)
```

```
## [1] 1817.534
```

```
BIC<-AIC(modelp, k=log(dim(poumon)[1]))
BIC
```

```
## [1] 1847.267
```

4. Build a R function that compute the Cp in linear regression model with gaussian noise
5. Manually, apply the Backward stepwise selection, by eliminating one by one the variable whose p-value is the minimum untill all the p-values<0.05.
6. In R, command `step` allows to perform forward and backward selection with criterion AIC (by default or choose k=2) and BIC (k=log(n)). (See below the command for forward selection and AIC). Do it for the AIC and the BIC. Compare the results

```
#
reg0<-lm(TLCO~ 1 , data=poum) #modele de départ (sans variable)
reg1<-lm(TLCO~ . , data=poum) #modele plein (avec toutes les variables)
select<-step(reg0, scope=formula(reg1), direction="forward",k=2)
```

```
## Start:  AIC=1339.82
## TLCO ~ 1
##
##                Df Sum of Sq   RSS    AIC
## + TAILLE_EN_M   1   12219.7 11750 1123.0
## + AGE           1    9738.5 14232 1181.8
## + Sexe          1    7475.1 16495 1227.1
## + POIDS         1    3366.3 20604 1295.4
```

```
## + origine      1    1468.2 22502 1322.4
## + BMI          1     295.3 23675 1338.0
## <none>                     23970 1339.8
##
## Step:  AIC=1122.95
## TLCO ~ TAILLE_EN_M
##
##           Df Sum of Sq     RSS    AIC
## + AGE      1    3905.0  7845.3 1000.9
## + origine  1    1050.7 10699.6 1096.2
## + Sexe     1     271.9 11478.4 1117.8
## <none>                 11750.3 1123.0
## + POIDS    1      66.9 11683.4 1123.2
## + BMI      1      55.6 11694.7 1123.5
##
## Step:  AIC=1000.93
## TLCO ~ TAILLE_EN_M + AGE
##
##           Df Sum of Sq    RSS     AIC
## + Sexe     1    973.46 6871.9  962.26
## + origine  1    519.19 7326.1  981.91
## + BMI      1    251.81 7593.5  992.92
## + POIDS    1    231.05 7614.3  993.76
## <none>                 7845.3 1000.93
##
## Step:  AIC=962.26
## TLCO ~ TAILLE_EN_M + AGE + Sexe
##
##           Df Sum of Sq    RSS    AIC
## + origine  1    441.74 6430.1 943.86
## <none>                 6871.9 962.26
## + BMI      1     26.27 6845.6 963.09
## + POIDS    1     19.34 6852.5 963.40
##
## Step:  AIC=943.86
## TLCO ~ TAILLE_EN_M + AGE + Sexe + origine
##
##         Df Sum of Sq    RSS    AIC
## + BMI    1    53.241 6376.9 943.31
## + POIDS  1    44.594 6385.5 943.73
## <none>               6430.1 943.86
##
## Step:  AIC=943.31
## TLCO ~ TAILLE_EN_M + AGE + Sexe + origine + BMI
##
##         Df Sum of Sq    RSS    AIC
## <none>               6376.9 943.31
## + POIDS  1    22.542 6354.3 944.22

select
```

```
##
## Call:
## lm(formula = TLCO ~ TAILLE_EN_M + AGE + Sexe + origine + BMI,
##     data = poum)
##
## Coefficients:
## (Intercept)  TAILLE_EN_M          AGE        SexeM     origineG          BMI
##    -20.1211      34.5247      -0.2262       4.2997      -2.9928       0.1371
```

```
#to have elements of the final model
select$formula
```

```
## NULL
```

```
#don't hesitate to use other select$...
```

7. *To go further :* Package leaps is another way to make automatic selection. Have a look at the documentation and perform stepwise selection with other criteria
8. *To go further* Code the Backward stepwise selection by yourself! (you can try new criteria) Did different criteria give the same best model ?
9. *To go further* Code the Backward stepwise selection, did different criteria give the same best model ?

2 Crossvalidation

This section is dedicated to different cross-validation methods.

2.1 Train error - test error

We will compute different kind of error for the 6 best models given by the forward method ((the best one with 1 variable, the best one with 2 variables, ...). For each of the models :

1. Use the whole dataset to fit each of the 6 best models and compute the RMSE. What is the best model if we use this criteria, was this fact expected ?
2. Cut randomly the dataset in a training set of size 200 and a test set of size 107.
   (`tr <- sample(1:nrow(poumon),200)`
   `train <- poum[tr,]`
   `test <- poum[-tr,].`
   Estimate the model with train set and compute the RMSE for test set (Use the command `predict(model, newdata=test)`). Compute the RMSE of the six best models (1 with 1 variable, 1 with 2, ...).
3. Do it 100 times and draw the boxplot of the RMSE. Can we distinguish a model that is better than the other ?

3 K-fold method

1. **5-Fold method** Make a program that compute the errors with the K Folds method for $K = 5$ of the three best models among the 6 evalueted in the previous section.
2. Do the method 100 times and make the boxplot of the errors to see the variability of the estimator of the errors.
3. **10-Fold method** do the same

4 Bootstrap method

Compute the bootstrap error for B=1000 bootstrap samples.
Compare the results of the different methods used above and propose a final model.

5 Classification model

We propose to use a logistic model to predict the survival at 12 months in the datacancer file. As the number of variables is very large, we will use the variables selected by the Benjamini Hochberg method at the 0.05 threshold (see previous tutorial). Build a data frame with the Surv12 variable and the covariates selected by Benjamini-Hochberg. We will discuss later the problem of missing data, so for simplicity, we will remove all individuals who have at least 1 missing data on these selected variables. (Use code `ind=which(apply(is.na(frame),1,sum)==0)` and `frame<-frame[ind,])` where frame is the name of the dataset.

1. Apply forward and backward methods to select variables in the logistic model (command `reg0<-glm(Surv12  1 , data=frame,family=binomial(link=logit))` gives the empty model with only a constant and `reg0<-glm(Surv12 . , data=frame,family=binomial(link=logit))` gives the model with all the variables.
2. Construct the confusion matrix and the roc curves of the two resulting models (use `library(pROC)` and command `roc(Y,p)` where `p` is the vector of the $P(Y_i = 1|x_i)$). Compute Area Under the Curve (AUC) for both (command `auc(Y,p)` where `p` is the vector of the $P(Y_i = 1|x_i)$)
3. Compare these two models by cross-validation using the rate of well classified as a criterion. Do the same with AUC as criteria.
4. Choose the best model to predict the survival.
5. *To go further* Code by yourself the elements to draw the ROC curve