Series 8

1. In the lecture we saw the Westfall Young permutation procedure. In this exercise we compare it with the Bonferroni correction in a synthetic example that you can construct with the help of the following R code (`x` is the matrix of covariates and `y` the response vector):

```
# for replication
set.seed(1)
n <- 30
p <- 50
# relevant covariate
x_true <- sample(c(0:1),size = n,replace = T)
# noise covariates
x <- matrix(sample(c(0:1),size=n*p,replace = T),ncol=p,nrow=n)
# combination of the two
x <- cbind(x_true, x)
# response
y <- ifelse(x[,1]==0, 0, sample(c(0:1), size = n, replace = T))
```

a) First, use the Chi-squared test (R function `chisq.test`) to obtain p-values when testing for individual association of the covariates in x with y and apply a Bonferroni correction. What do you see?

b) Now use the Westfall Young permutation procedure (with 1000 replications) for multiple testing. Do the results change?

2. Let us consider the dataset `Hitters` that you can find in the R-package `ISLR`. This exercise is an extension of lab Session 6.5.1 in *An introduction to statistical learning* (G. James, D. Witten, T. Hastie, and R. Tibshirani, *An Introduction to Statistical Learning - with Applications in R*, 2017, p. 244). There are different approaches/criteria to select a model based on its performance and complexity. We can select a model based on adjusted training error criteria like:

- BIC
- Mallows's Cp
- Adjusted R-squared

or based on a direct estimate of the expected test MSE using cross-validation. In this exercise we want to compare these different methodologies using cross-validation and assess their performance.

a) First load the dataset `Hitters` and remove the entries containing the missing values.
Hint:
- you can use the R function `na.omit`.

b) Then, in order to perform cross-validation we should:

1. Split the dataset into 10 folds.
2. Remove one of the folds (called *hold-out* fold) and on the remaining 9 folds run the best-subset model algorithm over all potential numbers of covariates to obtain one optimal model (in terms of RSS). To obtain this optimal model, you can either: (1) directly use BIC, Mallow's Cp or adjusted R-squared to select the optimal model across the different numbers of covariates, or (2) estimate the expected test MSE using a second level cross-validation on the current 9 folds. The best model selected by each procedure can be used at the end to get predictions on the hold-out fold.
3. Finally, repeat this last step for each hold-out fold to get estimates of the expected test MSE under each model selection method. These final estimates will be used to compare and assess the performances of the model selection approaches.

4. Discuss your results and fit the best model to the full data.

3. In this exercise, we want to compare the lasso and ridge regression on the dataset `riboflavin` contained in the R package `hdi` (consult the help page of `riboflavin` to get more information about this dataset).

a) Perform ridge regression as well as lasso regression of y on x with help of the R package `glmnet` for a grid of lambda values provided below, and then compare the coefficient paths. What do you observe?

```
grid <- 10^seq(10,-2, length = 100)
```

b) Perform a nested cross-validation (in a similar way as in the second exercise) to compare the performances (in terms of expected test MSE) of ridge and lasso regression with their optimal lambda values.

Hint: You can use the function `cv.glmnet()` to perform the lambda parameter selection in the inner cross-validation.

c) Fit the best model on the whole dataset (selecting one more time the optimal lambda value by cross-validation).

This exercise is inspired by Section 6.6 of *An Introduction to Statistical Learning* (*An Introduction to Statistical Learning - with Applications in R*, 2017). Note that you can use the prediction method implemented in this lab exercise.

4. Do the conceptual exercices 1 (p. 259, except point b) and 3 (p. 260) of ISL (G. James, D. Witten, T. Hastie, and R. Tibshirani, *An Introduction to Statistical Learning - with Applications in R*, 2017) freely available online `http://www-bcf.usc.edu/~gareth/ISL/ISLR%20Seventh%20Printing.pdf`.

Preliminary discussion: Friday, May 03.

Deadline: Friday, May 10.