## Assignment sheet 11

December 6, 2017

**Assignment 1.** Consider the cement data ($n = 13$). The residual sum of squares (RSS) for all the models containing the intercept are given below.

| model | RSS | model | RSS | model | RSS |
|---|---|---|---|---|---|
| – – – – | 2715.8 | 1 2 – – | 57.9 | 1 2 3 – | 48.11 |
| 1 – – – | 1265.7 | 1 – 3 – | 1227.1 | 1 2 – 4 | 47.97 |
| – 2 – – | 906.3 | 1 – – 4 | 74.8 | 1 – 3 4 | 50.84 |
| – – 3 – | 1939.4 | – 2 3 – | 415.4 | – 2 3 4 | 73.81 |
| – – – 4 | 883.9 | – 2 – 4 | 868.9 | | |
| | | – – 3 4 | 175.7 | 1 2 3 4 | 47.86 |

Calculate the analysis of variance table when adding $x_4$, $x_3$, $x_2$ and $x_1$ to the model in this order and test which terms should be included in the model at significance level $\alpha = 0.05$. Are the conclusions the same as in slide 407 ?

**Assignment 2** (automatic model selection)**.** Consider again the cement data from the course. The residual sum of squares (RSS) as well (some of !) the values of Mallows' $C_p$ for the models containing the intercept are as follows :

| model | RSS | $C_p$ | model | RSS | $C_p$ | model | RSS | $C_p$ |
|---|---|---|---|---|---|---|---|---|
| - - - - | 2715.8 | 442.58 | 1 2 - - | 57.9 | | 1 2 3 - | 48.1 | |
| | | | 1 - 3 - | 1227.1 | 197.94 | 1 2 - 4 | 48.0 | |
| 1 - - - | 1265.7 | 202.39 | 1 - - 4 | 74.8 | 5.49 | 1 - 3 4 | 50.8 | |
| - 2 - - | 906.3 | | - 2 3 - | 415.4 | 62.38 | - 2 3 4 | 73.8 | 7.325 |
| - - 3 - | 1939.4 | 314.90 | - 2 - 4 | 868.9 | 138.12 | | | |
| - - - 4 | 883.9 | 138.62 | - - 3 4 | 175.7 | 22.34 | 1 2 3 4 | 47.9 | 5 |

**a)** Use *forward selection* and *backward elimination* to choose a model for the data. Include significant variable at 5% using the $F$-test

$$F = \frac{\mathrm{RSS}(\hat{\beta}_{\mathrm{L}}) - \mathrm{RSS}(\hat{\beta}_{\mathrm{L}\cup\{\mathrm{j}\}})}{\mathrm{RSS}(\hat{\beta}_{\mathrm{full}})/(13-5)}$$

in order to decide whether the $j$-th variable is significant.

**b)** Mallows $C_p$ is defined as (see slide 423)

$$C_p = \frac{\mathrm{RSS}_p}{s^2} + 2p - n.$$

Note that $s^2$ is the estimator of the variance $\sigma^2$ *under the full model.*

i) Calculate the missing values of $C_p$ in the table, and explain how one uses this criterion for model selection.

ii) Which models would be chosen by *forward selection*, *backward elimination*, and Mallows' $C_p$ ? Are the three models same ?

**Assignment 3** (model selection with R)**.** a) Use AIC, *backward stepwise* and *forward stepwise* in order to choose a model for the data "Supervisor Performance" that can be found on the website `http://www1.aucegypt.edu/faculty/hadi/RABE4/Data4/P056.txt`. (For a description of the dataset, you can see Section 3.3 in the book *Regression Analysis by Example* by S. Chatterjee et A. S. Hadi, Wiley, 5th ed., 2012). Here is the R code

```
perf <- read.table("P056.txt",header=TRUE)

m1 <- lm(Y ~ ., data = perf)
m.backward <- step(m1, direction = "backward")

m0 <- lm(Y ~ 1, data = perf)
my.scope <- formula(perf)
m.forward <- step(m0, scope = my.scope, direction = "forward", data = perf)
```

Which model has the best AIC value?

b) Repeat the analysis using BIC. Are the resuls same? *Hint :* Use `k = log(n)` in `step`.

**Assignment 4.** Define the $j$-th *variance inflation factor* (slide 434) by

$$\mathrm{VIF}_j = \frac{\mathrm{Var}\,\hat{\beta}_j \|X_j\|^2}{\sigma^2} = \|X_j\|^2 \left[(X^T X)^{-1}\right]_{jj}, \quad j = 0, \ldots, p-1,$$

where $X_j$ is the $j$-th column of the design matrix $X_{n\times p}, n \geq p$, assume of full rank. (Note that the indices are from 0 to $p-1$.) Show that

$$\mathrm{VIF}_j = \frac{1}{1 - R^2_{0,j}},$$

where

$$R^2_{0,j} = \frac{\|H_{-j} X_j\|^2}{\|X_j\|^2}$$

is the the coefficient of determination of the model

$$X_{i,j} = \beta_0 + \beta_1 X_{i,1} + \ldots + \beta_{j-1} X_{i,j-1} + \beta_{j+1} X_{i,j+1} + \ldots + \beta_{p-1} X_{i,p-1} + \varepsilon_i, \quad i = 1, \ldots, n,$$

and $H_{-j}$ is hat matrix constructed without the $j$-th variable in the design matrix.
*Hint :* show that both sides are equal to $\|X_j\|^2/(X_j^T(I - H_{-j})X_j)$. For the $R^2_{0,j}$ use orthogonality of $X_j$ and $X_j - H_{-j}X_j$. For the VIF, assume that $j = p-1$, write $X = [X_{-j}\ X_j]$ and use the following result : if $A$ and $D$ are square and if $A$ and $D - CA^{-1}B$ are both invertible, then

$$\begin{bmatrix} A & B \\ C & D \end{bmatrix}^{-1} = \begin{bmatrix} A^{-1} + A^{-1}B(D - CA^{-1}B)^{-1}CA^{-1} & -A^{-1}B(D - CA^{-1}B)^{-1} \\ -(D - CA^{-1}B)^{-1}CA^{-1} & (D - CA^{-1}B)^{-1} \end{bmatrix}.$$

Argue (heuristically) why the same idea works for $j \neq p-1$. For the formal proof one needs to introduce a permutation matrix $\Pi$, you do not have to do this rigorously.

**Assignment 5** (conditioning)**.** Let $X_{n\times p}, n \geq p$, be a matrix of full rank. Define the *condition number* of $X$ by

$$\kappa(X) = \frac{\sigma_1}{\sigma_p},$$

where $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_p > 0$ are the singular values of $X$. In class (slides 436) the condition number was defined by

$$\mathrm{Cond}_X = \sqrt{\frac{\lambda_1}{\lambda_p}},$$

where $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p > 0$ are the eigenvalues of $X^T X$.

a) Show that $\text{Cond}_X = \kappa(X)$.

b) Show that for all $j$, $\text{VIF}_j \leq \kappa(X)^2 = \kappa(X^T X)$, where $\text{VIF}_j$ is the $j$-th *variance inflation factor* of $X$. Show that this bound is tight, and can be attained. Explain the usefullness of the bound.

*Hint :* the matrix norm of $A_{m\times n}$

$$\|A\| = \sup_{\|x\|=1} \|Ax\|$$

equals $\sigma_1$, the largest singular value of $A$.

**Assignment 6** (Orthogonal variables)**.**
Consider the regression model

$$y = X\beta + \varepsilon = (X_1,\, X_2) \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix} + \varepsilon,$$

where $X = (X_1, X_2)$, $\beta^t = (\beta_1^t, \beta_2^t)$, $X_1$ is $n \times p_1$, $X_2$ is $n \times p_2$ (both injective) such that

$$X_1^t X_2 = 0_{p_1 \times p_2}.$$

Let $H_i$ the hat matrix associated with $X_i$.

(i). What is the geometrical interpretation of $X_1^t X_2 = 0$ ?

(ii). Compute $H$ as a function of $X_i$ and $H_i$, then compute the products

$$H_1 H_2, H_2 H_1, H H_1, H_1 H.$$

Comment. What is their geometric interpretation ?

(iii). Show that each of the following quantities is equal to $Hy$ :

(a) $H_1 y + H_2 y$ ;

(b) $H_1 y + H_2 e_1$, avec $e_1 = (I - H_1)y$ ;

(c) $H_1 y + H e_1$.

Finish by observing that the above equalities imply that for the model

$$y = X\beta + \varepsilon \qquad (M)$$

the fitted values under the full model $M$ equal

(a) the sum of the fitted values under $(M_1)$ and $(M_2)$ (where the model $M_i$ corresponds to the pair $(y, X_i)$.

(b) The sum of the fitted values under $(M_1)$ (with input data $(y, X_1)$) and of the residuals of $(M_1)$ computed under $(M_2)$ (with variables $(e_1, X_2)$).

(c) The sum of the fitted values under $(M_1)$ (with variables $(y, X_1)$) and of the residuals of $(M_1)$ computed under $(M)$ (with variables $(e_1, X)$).

**Assignment 7** (Orthogonal variables and ANOVA)**.**
Consider the regression model

$$y = X\beta + \varepsilon = (X_1, \dots, X_k) \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_k \end{pmatrix} + \varepsilon$$

where $X_i$ is $n \times p_i$, all the $X_i$ are injectives and

$$i \neq j \implies X_i^t X_j = 0.$$

Let $H$ be the hat matrix associated with $X$, $H_i$ the hat matrix associated with $X_i$ and $\hat{\beta} = (X^t X)^{-1} X^t y = (\hat{\beta}_1^t, \dots, \hat{\beta}_k^t)^t$. Denote by $\delta_{ij}$ the Kronecker delta : $\delta_{ij} = 1$ if $i = j$, and 0 otherwise. For a set $L \subset \{1, \dots, k\}$ we define $X_L = (X_i : i \in L)$ and $\hat{\beta}_L = (\hat{\beta}_i^t : i \in L)^t$. For example if $L = \{1, 2, 4\}$, $X_L = (X_1, X_2, X_4)$ et

$$\hat{\beta}_L = \begin{pmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \\ \hat{\beta}_4 \end{pmatrix}.$$

Define $RSS_L = \left\| y - H_L y^2 \right\|$, où $H_L = X_L (X_L^t X_L)^{-1} X_L^t$.

(i). Show that $H = H_1 + \cdots + H_k$ and that $H_L = \sum_{i \in L} H_i$.

(ii). Show that $H_i H_j = \delta_{ij} H_i$.

(iii). Show that $\hat{\beta}_j = (X_j^t X_j)^{-1} X_j^t y$.

(iv). For $j \notin L$, compute

$$RSS_L - RSS_{L \cup \{j\}},$$

and show that such expression doesn't depend on $L$.

(v). What is the interpretation of point 4 w.r.t. ANOVA ?

**Assignment 8** (AIC and Gaussian linear models)**.**
Show that the AIC criterion for a gaussian linea model and a response vector of size $n$ with $p$ covariates can be written as

$$\text{AIC} = n \, \log \hat{\sigma}^2 + 2p + \text{const},$$

where $\sigma^2$ is the unknown variance of the model and $\hat{\sigma}^2 = RSS_p / n$ is the MLE estimator for $\sigma^2$.

**Assignment 9** (Cross validation and number of parameters)**.**
Using the fact that

$$\hat{\beta}_{-j} = \hat{\beta} - \frac{(y_j - \hat{y}_j)(X^t X)^{-1} x_j}{1 - h_{jj}},$$

show that

$$\text{CV} = \sum_{j=1}^{n} (y_j - x_j^t \hat{\beta}_{-j})^2 \tag{1}$$

can be written as

$$\text{CV} = \sum_{j=1}^{n} \frac{(y_j - x_j^t \hat{\beta})^2}{(1 - h_{jj})^2}. \tag{2}$$

What is the advantage of using the formula (2) over the formula (1) ?