Answer sheet 11

Assignment 1. We need to calculate the F_k's defined in slide 406 :

	df	decrease in RSS	MS	F	p-value
x_4	1	$\text{RSS}_0 - \text{RSS}_4$=1831.9	1831.9	(1831.9/5.98)=306.3	10^{-7}
x_3	1	$\text{RSS}_4 - \text{RSS}_{34}$=708.2	708.2	118.4	10^{-6}
x_2	1	$\text{RSS}_{34} - \text{RSS}_{234}$=101.89	101.89	17.04	0.003
x_1	1	$\text{RSS}_{234} - \text{RSS}_{1234}$=25.95	25.95	4.3	0.07
résidus	8	47.86	5.98		

The residual degrees of freedom is $n - p = 13 - 5 = 8$ and each difference of RSS has one degree of freedom, as we add one variable at a time. For the F-test we use the quantiles of $F_{1,8}$ distribution, and if the p-value is smaller than $\alpha = 0.05$ we add the variable to the model. The results are very different from those in slide 407. Here we include the variables x_4, x_3 and x_2 at level $\alpha = 0.05$, and even x_1 at level 0.1. In slide 407 the model only included x_1 and x_2. We see that the order matters in an analysis of variance.

Assignment 2. a) To decide whether to include the j-th variable or not in the model $y = \beta_0 + \sum_{i \in L} \beta_i x_i$ we use the test statistic

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

where $\hat{\beta}_{\text{full}}$ is the estimator of β in the complete model. Since $\text{RSS}(\hat{\beta}_{\text{L}}) - \text{RSS}(\hat{\beta}_{\text{L}\cup\{\text{j}\}}) \sim \sigma^2\chi^2_1$ under the null hypothesis $H_0 : \beta_j = 0$, and $\text{RSS}(\hat{\beta}_{\text{full}}) \sim \sigma^2\chi^2_{\text{n-p}}$ is independent of $\text{RSS}(\hat{\beta}_{\text{L}}) - \text{RSS}(\hat{\beta}_{\text{L}\cup\{\text{j}\}})$, we know that $F \sim F_{1,8}$ under H_0. In particular, the distribution of F does not depend on the size of L, and the critical value of the F-test at 5% is always 5.32.

Forward selection At each step we consider adding the variable that leads to the largest decrease of RSS.
- Initial model : $y = \beta_0 + \epsilon$
- Step 1 : $y = \beta_0 + \beta_4 x_4 + \epsilon$, $F = \frac{2715.8-883.9}{47.9/(13-5)} = 305.95 > 5.32$.
- Step 2 : $y = \beta_0 + \beta_4 x_4 + \beta_1 x_1 + \epsilon$, $F = 135.13 > 5.32$.
- Step 3 : $y = \beta_0 + \beta_4 x_4 + \beta_1 x_1 + \beta_2 x_2 + \epsilon$, $F = 4.47 < 5.32$.

We choose the model $y = \beta_0 + \beta_4 x_4 + \beta_1 x_1 + \epsilon$.

Backward selection At each step we consider removing the variable that would lead to the smallest increase in RSS.
- Initial model : $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \epsilon$
- Step 1 : $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_4 x_4 + \epsilon$, $F = \frac{48-47.9}{47.9/(13-5)} = 0.0167 < 5.32$.
- Step 2 : $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon$, $F = 1.65 < 5.32$.
- Step 3 : $y = \beta_0 + \beta_2 x_2 + \epsilon$, $F = 141.70 > 5.32$.

We choose the model $y = \beta_0 + \beta_2 x_2 + \beta_1 x_1 + \epsilon$.

b) i) One uses Mallows' C_p like AIC : choose the model with the smallest value of C_p. In order to calculate the missing C_p values, we need to find s^2. This can be done using any model for which C_p is given. Alternatively, we can use it's very definition :

$$s^2 = \frac{\|e_{\text{full}}\|^2}{n-p} = \frac{\text{RSS}_{\text{full}}}{13-5} = \frac{47.9}{8} = 5.99.$$

Here is the table with all C_p values :

model	RSS	C_p	model	RSS	C_p	model	RSS	C_p
- - - -	2715.8	442.58	1 2 - -	57.9	2.67	1 2 3 -	48.1	3.03
			1 - 3 -	1227.1	197.94	1 2 - 4	48.0	3.02
1 - - -	1265.7	202.39	1 - - 4	74.8	5.49	1 - 3 4	50.8	3.48
- 2 - -	906.3	142.37	- 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

ii) With forward selection, we choose the model $y = \beta_0 + \sum_{i\in\{1,2,4\}} \beta_i x_i$. With backward selection we choose the model $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon$. This is also the model with the smallest value of C_p.

Assignment 3.

a) Both forward and backward selection choose the model `Y ~ X1 + X3` with AIC value of 118.

b) Now both methods choose the model `Y ~ X1` with BIC value of 121.43.

Remark 1. BIC chooses a smaller model because of the heavier penalisation for the number of variables, in comparison with AIC ($\log(30) > 2$).

Remark 2. For this dataset, forward and backward choose the same model. This is not always the case, as we have already seen. The only way to verify that the chosen model has the smallest AIC/BIC value is by an exhaustive search. In R, this can be done using the function `regsubsets` in the package `leaps`.

Assignment 4.

We have

$$\frac{1}{1-R_{0,j}^2} = \frac{\|X_j\|^2}{\|X_j\|^2 - \|H_{-j}X_j\|^2}.$$

Since $H_{-j}X_j$ is the orthogonal projection of X_j onto $M(X_{-j})$, we have $H_{-j}X_j \perp X_j - H_{-j}X_j$. Pythagoras gives

$$\|X_j\|^2 - \|H_{-j}X_j\|^2 = \|X_j - H_{-j}X_j\|^2 = \|(I-H_{-j})X_j\|^2 = X_j^T(I-H_{-j})X_j.$$

Thus

$$\frac{1}{1-R_{0,j}^2} = \frac{\|X_j\|^2}{X_j^T(I-H_{-j})X_j}.$$

Let $\Pi = [e_0, \ldots, e_{j-1}, e_{p-1}, e_j, \ldots, e_{p-2}]$ be the matrix that sends the last column of a matrix to the j-th position when multiplying from right. (Recall that the indices run from 0 to $p-1$. When $j = p-1$, Π is simply the identity.) In other words, $X = [X_{-j}X_j]\Pi$. A straightforward calculation shows that $\Pi^T\Pi = I$, so that Π is orthogonal. We obtain

$$\begin{aligned}(X^TX)^{-1} &= (([X_{-j}X_j]\Pi)^T([X_{-j}X_j]\Pi))^{-1} \\ &= \Pi^T\left(\begin{bmatrix}X_{-j}^T \\ X_j^T\end{bmatrix}\begin{bmatrix}X_{-j} & X_j\end{bmatrix}\right)^{-1}\Pi \\ &= \Pi^T\begin{bmatrix}X_{-j}^TX_{-j} & X_{-j}^TX_j \\ X_j^TX_{-j} & X_j^TX_j\end{bmatrix}^{-1}\Pi.\end{aligned}$$

Using the hint now gives

$$\begin{bmatrix} X_{-j}^T X_{-j} & X_{-j}^T X_j \\ X_j^T X_{-j} & X_j^T X_j \end{bmatrix}^{-1} = \begin{bmatrix} \# & \# \\ \# & (X_j^T X_j - X_j^T X_{-j}(X_{-j}^T X_{-j})^{-1} X_{-j}^T X_j)^{-1} \end{bmatrix}.$$

Since Π sends the last column to the j-th position and Π^T sends the last row to the j-th position, we have

$$\begin{aligned} \left[(X^T X)^{-1}\right]_{jj} &= (X_j^T X_j - X_j^T X_{-j}(X_{-j}^T X_{-j})^{-1} X_{-j}^T X_j)^{-1} \\ &= \frac{1}{X_j^T X_j - X_j^T X_{-j}(X_{-j}^T X_{-j})^{-1} X_{-j}^T X_j} \\ &= \frac{1}{X_j^T (I - H_{-j}) X_j}. \end{aligned}$$

We can thus conclude that

$$\mathrm{VIF}_j = \|X_j\|^2 \left[(X^T X)^{-1}\right]_{jj} = \frac{\|X_j\|^2}{X_j^T (I - H_{-j}) X_j} = \frac{1}{1 - R_{0,j}^2}.$$

Assignment 5.

a) Let $X = U\Sigma V^T$ be the singular value decomposition of X. Then

$$X^T X = (U\Sigma V^T)^T (U\Sigma V^T) = V\Sigma^T U^T U \Sigma V^T = V\Sigma^T \Sigma V^T,$$

where

$$\Sigma^T \Sigma = \begin{bmatrix} \mathrm{diag}(\sigma_1, \ldots, \sigma_p) & 0 \end{bmatrix} \begin{bmatrix} \mathrm{diag}(\sigma_1, \ldots, \sigma_p) \\ 0 \end{bmatrix} = \mathrm{diag}(\sigma_1^2, \ldots, \sigma_p^2).$$

Consequently the eigenvalues of $X^T X$ are $\lambda_i = \sigma_i^2$ and we obtain

$$\mathrm{Cond}_X = \sqrt{\frac{\lambda_1}{\lambda_p}} = \frac{\sigma_1}{\sigma_p} = \kappa(X).$$

b) Denote $\Lambda := \Sigma^T \Sigma = \mathrm{diag}(\lambda_1, \ldots, \lambda_p)$. Then

$$(X^T X)^{-1} = (V \Lambda V^T)^{-1} = V \Lambda^{-1} V^T,$$

where $\Lambda^{-1} = \mathrm{diag}(\lambda_1^{-1}, \ldots, \lambda_p^{-1})$. We can write the (i,j)-th element of $(X^T X)^{-1}$ as

$$\left[(X^T X)^{-1}\right]_{jj} = e_j^T (X^T X)^{-1} e_j = e_j^T V \Lambda^{-1} V^T e_j = (Se_j)^T \Lambda^{-1} S e_j = s_j^T \Lambda^{-1} s_j = \sum_{i=1}^p \frac{s_{ij}^2}{\lambda_i},$$

where $S = V^T$ and s_j is the j-th column of S. Since S is orthogonal, we obtain the bound

$$\left[(X^T X)^{-1}\right]_{jj} \leq \sum_{i=1}^p \frac{s_{ij}^2}{\lambda_p} = \frac{1}{\lambda_p} \sum_{i=1}^p s_{ij}^2 = \frac{1}{\lambda_p} \|s_j\|^2 = \frac{1}{\lambda_p}.$$

For $\|X_j\|^2$, one has the bound

$$\|X_j\|^2 = \|Xe_j\|^2 \leq \sup_{\|c\|=1} \|Xc\|^2 = \left(\sup_{\|c\|=1} \|Xc\| \right)^2 = \|X\|^2 = \sigma_1^2 = \lambda_1.$$

Thus for all j

$$\mathrm{VIF}_j = \underbrace{\|X_j\|^2}_{\leq \lambda_1} \underbrace{\left[(X^T X)^{-1}\right]_{jj}}_{\leq \frac{1}{\lambda_p}} \leq \frac{\lambda_1}{\lambda_p} = \kappa(X)^2 = \kappa(X^T X).$$

The bound is attained when $X^T X = I$. It is useful, because it shows that if the condition number of X is small, then all the variance inflation factors must be small. We can say something about the multicolinearity of the model using one number $\kappa(X)$.

Assignment 6.

(i). This means that all the columns of X_1 are orthogonal to the columns of X_2. In other words $\mathcal{M}(X_1) \perp \mathcal{M}(X_2)$.

(ii). Remember first that

$$X^t X = \begin{pmatrix} X_1^t X_1 & 0 \\ 0 & X_2^t X_2 \end{pmatrix},$$

thus

$$\begin{aligned} H &= (X_1, X_2) \begin{pmatrix} (X_1^t X_1)^{-1} & 0 \\ 0 & (X_2^t X_2)^{-1} \end{pmatrix} (X_1, X_2)^t \\ &= X_1 (X_1^t X_1)^{-1} X_1^t + X_2 (X_2^t X_2)^{-1} X_2^t = H_1 + H_2. \end{aligned}$$

Moreover as $X_1^t X_2 = 0$, we have $H_1 H_2 = 0$. And thus $H_2 H_1 = H_2^t H_1^t = (H_1 H_2)^t = 0$,

$$H H_1 = (H_1 + H_2) H_1 = H_1^2 = H_1$$

and $H_1 H = H_1^t H^t = (H H_1)^t = H_1^t = H_1$.

Interpretation : $H_1 H_2 = 0$ comes from the fact that the columns of X_1 et X_2 are orthogonal, hence if one projects on $\mathcal{M}(X_2)$ and then on $\mathcal{M}(X_1)$, will obtain the vector 0 as a result. The interpretation for $H_2 H_1 = 0$ is similar. $H H_1 = H_1$ comes from projecting on $\mathcal{M}(X_1)$ and then projecting on $\mathcal{M}(X)$ is equivalent to project uniquely on $\mathcal{M}(X_1)$, as $\mathcal{M}(X_1)$ is a subspace of $\mathcal{M}(X)$. For the same reason, $H_1 H = H_1$ because we project on $\mathcal{M}(X)$ and after that on $\mathcal{M}(X_1)$, which is like if we were projecting only on $\mathcal{M}(X_1)$. In tuitively we remark that even if $X_1^t X_2 \neq 0$, we still have $H H_1 = H_1 = H_1 H$, but $H_1 H_2 \neq 0$ and $H_2 H_1 \neq 0$.

(iii). Using the fact that $Hy = (H_1 + H_2)y$,

(a) immediate

(b) follows from $H_2 H_1 = 0$;

(c) follows from $H(I - H_1) = H - H_1 = H_2$.

Assignment 7.

(i). This questions is very similar to Ex. 1. since

$$(X^t X)^{-1} = \begin{pmatrix} (X_1^t X_1)^{-1} & 0 & \dots & 0 \\ 0 & (X_2^t X_2)^{-1} & 0 & \vdots \\ \vdots & 0 & \ddots & 0 \\ 0 & \dots & 0 & (X_k^t X_k)^{-1} \end{pmatrix}$$

and

$$(X_L^t X_L)^{-1} = \text{diag}((X_i^t X_i)^{-1} : i \in L).$$

Hence

$$H = X_1(X_1^t X_1)^{-1} X_1^t + \cdots + X_k(X_k^t X_k)^{-1} X_k^t = H_1 + \cdots + H_k$$

and

$$H_L = \sum_{i \in L} X_i (X_i^t X_i)^{-1} X_i^t = \sum_{i \in L} H_i.$$

(ii). If $i = j$, $H_i H_j = H_i^2 = H_i$ and if $i \neq j$, $H_i H_j = X_i(X_i^t X_i)^{-1} X_i^t X_j (X_j^t X_j)^{-1} X_j^t = 0$ so that $X_i^t X_j = 0$.

(iii).

$$\hat{\beta} = (X^t X)^{-1} X^t y = \begin{pmatrix} (X_1^t X_1)^{-1} & 0 & \dots & 0 \\ 0 & (X_2^t X_2)^{-1} & 0 & \vdots \\ \vdots & 0 & \ddots & 0 \\ 0 & \dots & 0 & (X_k^t X_k)^{-1} \end{pmatrix} \begin{pmatrix} X_1^t \\ X_2^t \\ \vdots \\ X_k^t \end{pmatrix} y = \begin{pmatrix} (X_1^t X_1)^{-1} X_1^t y \\ (X_2^t X_2)^{-1} X_2^t y \\ \vdots \\ (X_k^t X_k)^{-1} X_k^t y \end{pmatrix}.$$

(iv). First of all notice that

$$e_L := y - H_L y = y - \sum_{i \in L} H_i y$$

and that

$$e_{L \cup \{j\}} := y - H_{L \cup \{j\}} y = y - \sum_{i \in L \cup \{j\}} H_i y.$$

Moreover

$$\begin{aligned} (I - H_{L \cup \{j\}}) e_L &= (I - H_{L \cup \{j\}})(I - H_L) y \\ &= (I - H_L - H_{L \cup \{j\}} + H_{L \cup \{j\}} H_L) y \\ &= (I - H_{L \cup \{j\}}) y \\ &= e_{L \cup \{j\}}. \end{aligned}$$

Then $e_{L \cup \{j\}}$ is an orthogonal projection of e_L, where $e_L - e_{L \cup \{j\}} \perp e_{L \cup \{j\}}$ and

$$\|e_{L \cup \{j\}}\|^2 + \|e_L - e_{L \cup \{j\}}\|^2 = \|e_L\|^2.$$

Hence

$$RSS_L - RSS_{L \cup \{j\}} = \|e_L\|^2 - \|e_{L \cup \{j\}}\|^2 = ||e_L - e_{L \cup \{j\}}||^2 = ||H_j y||^2$$

is independent from L.

(v). The interpretation wrt ANOVA is that in this case, adding one variable X_j does not depend on the variables that are already in the model. **This is not true in general!**

Assignment 8.

For the Gaussian linear model $y \sim N(X\beta, \sigma^2 I_n)$, the likelihood of (β, σ^2) is given by

$$L(\beta, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left(-\frac{1}{2\sigma^2} (y - X\beta)^t (y - X\beta) \right).$$

Then the log likelihood is

$$l(\beta, \sigma^2) = -\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}(y - X\beta)^t(y - X\beta).$$

We have that the m.l.e. for β and σ^2 are

$$\hat{\beta} = (X^tX)^{-1}X^ty, \quad \hat{\sigma}^2 = \frac{1}{n}(y - X\hat{\beta})^t(y - X\hat{\beta}).$$

Hence the maximum for the likelihood is achieved at

$$l(\hat{\beta}, \hat{\sigma}^2) = -\frac{n}{2}\log(2\pi\hat{\sigma}^2) - \frac{1}{2\hat{\sigma}^2}\underbrace{(y - X\hat{\beta})^t(y - X\hat{\beta})}_{=n\hat{\sigma}^2} = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log\hat{\sigma}^2 - \frac{n}{2}.$$

By definition of AIC, we obtain that

$$\text{AIC} = -2l(\hat{\beta}, \hat{\sigma}^2) + 2p = n\log(2\pi) + n\log\hat{\sigma}^2 + n + 2p = n\log\hat{\sigma}^2 + 2p + \text{const.}$$

Assignment 9.
We have that

$$\hat{\beta}_{-j} = \hat{\beta} - \frac{(y_j - \hat{y}_j)\left(X^tX\right)^{-1}x_j}{1 - h_{jj}}.$$

Hence we have

$$\begin{aligned} x_j^t\hat{\beta}_{-j} &= x_j^t\hat{\beta} - (1 - h_{jj})^{-1}x_j^t(X^tX)^{-1}x_j(y_j - \hat{y}_j) \\ &= \hat{y}_j - \frac{h_{jj}}{1 - h_{jj}}(y_j - \hat{y}_j) \\ &= \hat{y}_j + \left(1 - \frac{1}{1 - h_{jj}}\right)(y_j - \hat{y}_j) \\ &= \hat{y}_j + y_j - \hat{y}_j - \frac{1}{1 - h_{jj}}(y_j - \hat{y}_j) \end{aligned}$$

where

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

If we use formula (1), we have to estimate all the $\hat{\beta}_{-j}$, $j = 1, \ldots, n$, hence proceed to n adjustements. Instead formula (2), only the fitting of the full model is required.