ANSWER SHEET 11

Assignment 1. (i). $X^TX = (x_1, \ldots, x_n)\begin{pmatrix} x_1^T \\ \vdots \\ x_n^T \end{pmatrix} = \sum_{i=1}^n x_i x_i^T = X_{-k}^T X_{-k} + x_k x_k^T$.

(ii). (a) It suffices to verify that

$$(A + uv^T)\left[B - \frac{Buv^TB}{1 + v^TBu}\right] = I,$$

where we denote $B = A^{-1}$ to simplify notation. We have

$$\begin{aligned}(A + uv^T)\left[B - \frac{Buv^TB}{1 + v^TBu}\right] &= I - \frac{uv^TB}{1 + v^TBu} + uv^TB - \frac{u\{v^TBu\}v^TB}{1 + v^TBu} \\ &= I + uv^TB - \frac{uv^TB}{1 + v^TBu}(1 + v^TBu) \\ &= I.\end{aligned}$$

We used that $AB = I$, and that the expression $\{v^TBu\}$ is a scalar and thus commutes with any matrix.

(b) Write $C = X^TX$. and use (a) :

$$\begin{aligned}(X_{-k}^T X_{-k}^T)^{-1} &= (C - x_k x_k^T)^{-1} \\ &= C^{-1} + \frac{C^{-1}x_k x_k^T C^{-1}}{1 - x_k^T C^{-1} x_k} \\ &= \left(I + \frac{C^{-1}x_k x_k^T}{1 - h_{kk}}\right) C^{-1} \\ &= \left(I + \frac{(X^TX)^{-1}x_k x_k^T}{1 - h_{kk}}\right)(X^TX)^{-1},\end{aligned}$$

where we have used $x_k^T C^{-1} x_k = (X(X^TX)^{-1}X^T)_{k,k} = h_{kk}$.

(iii). Recall that $y = (y_1, \ldots, y_n)^T$ with $y_j \in \mathbb{R}$ and $e = (e_1, \ldots, e_n)^T$ is the residual vector.

(a) $X^Ty = (x_1, \ldots, x_n)y = \sum_{i=1}^n x_i y_i = X_{-k}^T y + x_k y_k$.

(b)

$$\begin{aligned}x_k^T(X^TX)^{-1}X_{-k}^T y &= x_k^T(X^TX)^{-1}(X^Ty - x_k y_k) \\ &= \hat{y}_k - h_{kk}y_k \\ &= y_k - e_k - h_{kk}y_k \\ &= (1 - h_{kk})y_k - e_k.\end{aligned}$$

We have

$$\begin{aligned}\hat{\beta}_{-k} &= \left(\sum_{i \neq k} x_i x_i^T\right)^{-1}\left(\sum_{i \neq k} x_i y_i\right) \\ &= (X_{-k}^T X_{-k})^{-1} X_{-k}^T y \\ &= \left(I + \frac{(X^TX)^{-1}x_k x_k^T}{1 - h_{kk}}\right)(X^TX)^{-1}X_{-k}^T y \\ &= (X^TX)^{-1}(X^Ty - y_k x_k) + (1 - h_{kk})^{-1}(X^TX)^{-1}x_k x_k^T (X^TX)^{-1} X_{-k}^T y\end{aligned}$$

and using (b),

$$\begin{aligned}\hat\beta_{-k} &= \hat\beta - (X^TX)^{-1}x_ky_k + (1-h_{kk})^{-1}(X^TX)^{-1}x_k[(1-h_{kk})y_k - e_k]\\ &= \hat\beta - (1-h_{kk})^{-1}e_k(X^TX)^{-1}x_k.\end{aligned}$$

(iv). We have

$$\hat y - \hat y_{-k} = X\hat\beta - X\hat\beta_{-k} = X(\hat\beta - \hat\beta_{-k}) = e_k(1-h_{kk})^{-1}X(X^TX)^{-1}x_k,$$

and so

$$\begin{aligned}\|\hat y - \hat y_{-k}\|^2 &= (\hat y - \hat y_{-k})^T(\hat y - \hat y_{-k})\\ &= e_k^2(1-h_{kk})^{-2}x_k^T(X^TX)^{-1}(X^TX)(X^TX)^{-1}x_k = e_k^2(1-h_{kk})^{-2}h_{kk}.\end{aligned}$$

Finally, recall that $r_k = \frac{e_k}{s\sqrt{1-h_{kk}}}$.

Assignment 2. We need to calculate the F_k's defined in slide 406 :

	df	decrease in RSS	MS	F	p-value
x_4	1	$\mathrm{RSS}_0 - \mathrm{RSS}_4$=1831.9	1831.9	(1831.9/5.98)=306.3	10^{-7}
x_3	1	$\mathrm{RSS}_4 - \mathrm{RSS}_{34}$=708.2	708.2	118.4	10^{-6}
x_2	1	$\mathrm{RSS}_{34} - \mathrm{RSS}_{234}$=101.89	101.89	17.04	0.003
x_1	1	$\mathrm{RSS}_{234} - \mathrm{RSS}_{1234}$=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 3. 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

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

where $\hat\beta_{\mathrm{full}}$ is the estimator of β in the complete model. Since $\mathrm{RSS}(\hat\beta_{\mathrm{L}}) - \mathrm{RSS}(\hat\beta_{\mathrm{L}\cup\{\mathrm{j}\}}) \sim \sigma^2\chi^2_1$ under the null hypothesis $H_0 : \beta_j = 0$, and $\mathrm{RSS}(\hat\beta_{\mathrm{full}}) \sim \sigma^2\chi^2_{\mathrm{n-p}}$ is independent of $\mathrm{RSS}(\hat\beta_{\mathrm{L}}) - \mathrm{RSS}(\hat\beta_{\mathrm{L}\cup\{\mathrm{j}\}})$, 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 : $y = \beta_0 + \beta_4x_4 + \epsilon$, $F = \frac{2715.8-883.9}{47.9/(13-5)} = 305.95 > 5.32$.
— Step 2 : $y = \beta_0 + \beta_4x_4 + \beta_1x_1 + \epsilon$, $F = 135.13 > 5.32$.
— Step 3 : $y = \beta_0 + \beta_4x_4 + \beta_1x_1 + \beta_2x_2 + \epsilon$, $F = 4.47 < 5.32$.

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 4.

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

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

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 5.
We have that

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

Hence we have

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

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.