Norwegian University of Science and Technology
Department of Mathematical Sciences

Corrections 25 May 2018
Corrections and additions 8 June 2018

TMA4267 2018V – Solutions

Problem 1

a) If $\left(\begin{smallmatrix}\boldsymbol{X}_1\\ \boldsymbol{X}_2\end{smallmatrix}\right) \sim N\big(\left(\begin{smallmatrix}\boldsymbol{\mu}_1\\ \boldsymbol{\mu}_2\end{smallmatrix}\right), \left(\begin{smallmatrix}\boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12}\\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22}\end{smallmatrix}\right)\big)$, the conditional distribution of $\boldsymbol{X}_2$ given $\boldsymbol{X}_1 = \boldsymbol{x}_1$ is $N(\boldsymbol{\mu}_2+\Sigma_{21}\Sigma_{11}^{-1}(\boldsymbol{x}_1-\boldsymbol{\mu}_1), \Sigma_{22}-\Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12})$. Here, $\boldsymbol{\mu}_1 = 0,\ \boldsymbol{\mu}_2 = 0,\ \Sigma_{11} = 1,\ \Sigma_{12} = -0.8,$ $\Sigma_{21} = -0.8,\ \Sigma_{22} = 2$, and $\boldsymbol{x}_1 = x$, so $\boldsymbol{\mu}_2 + \Sigma_{21}\Sigma_{11}^{-1}(\boldsymbol{x}_1 - \boldsymbol{\mu}_1) = 0 + (-0.8)\cdot 1^{-1}(x-0) = -0.8x$ and $\Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12} = 2 - (-0.8)\cdot 1^{-1}\cdot(-0.8) = 1.36$, so the distribution is univariate normal $N(-0.8x, 1.36)$.

b) Since Σ is symmetric, there exists an orthogonal matrix P such that $\Sigma = P\Lambda P^{\mathrm{T}}$, where Λ is diagonal with the eigenvalues of Σ on the diagonal. Since Σ is positive semidefinite, all eigenvalues are non-negative. Denote by $\Lambda^{1/2}$ the matrix obtained from Λ by replacing all entries by their square roots, and define $\Sigma^{1/2} = P\Lambda^{1/2}P^{\mathrm{T}}$. Then $(\Sigma^{1/2})^2 = P\Lambda^{1/2}P^{\mathrm{T}}P\Lambda^{1/2}P^{\mathrm{T}} = P\Lambda^{1/2}\Lambda^{1/2}P^{\mathrm{T}} = P\Lambda P^{\mathrm{T}} = \Sigma$.

The definition $\Sigma^{1/2} = P\Lambda^{1/2}P^{\mathrm{T}}$ is actually an orthogonal diagonalization of $\Sigma^{1/2}$, so the diagonal of $\Lambda^{1/2}$ contains the eigenvalues of $\Sigma^{1/2}$. Since Σ is non-singular, zero is not an eigenvalue (and Σ is actually positive definite). The eigenvalues of $\Sigma^{1/2}$ – the square roots of the positive eigenvalues of Σ – are then also all non-zero, so that $\Sigma^{1/2}$ is non-singular. It is also easy to show explicitly that $P(\Lambda^{1/2})^{-1}P^{\mathrm{T}}$ is an inverse of $\Sigma^{1/2}$.

Since $\boldsymbol{Y}$ is multivariate normal, so is $(\Sigma^{1/2})^{-1}(\boldsymbol{Y} - \boldsymbol{\mu})$ (it can be rewritten on the form $A\boldsymbol{Y} + \boldsymbol{c}$), and is determined by its mean vector and covariance matrix. $E((\Sigma^{1/2})^{-1}(\boldsymbol{Y}-\boldsymbol{\mu})) = (\Sigma^{1/2})^{-1}E(\boldsymbol{Y}-\boldsymbol{\mu}) = (\Sigma^{1/2})^{-1}(E\boldsymbol{Y}-\boldsymbol{\mu}) = (\Sigma^{1/2})^{-1}(\boldsymbol{\mu}-\boldsymbol{\mu}) = \mathbf{0}$ and $\mathrm{Cov}((\Sigma^{1/2})^{-1}(\boldsymbol{Y}-\boldsymbol{\mu})) = (\Sigma^{1/2})^{-1}\,\mathrm{Cov}(\boldsymbol{Y}-\boldsymbol{\mu})((\Sigma^{1/2})^{-1})^{\mathrm{T}} = (\Sigma^{1/2})^{-1}(\mathrm{Cov}\,\boldsymbol{Y})(\Sigma^{1/2})^{-1} = (\Sigma^{1/2})^{-1}\Sigma(\Sigma^{1/2})^{-1} = (\Sigma^{1/2})^{-1}\Sigma^{1/2}\Sigma^{1/2}(\Sigma^{1/2})^{-1} = II = I$, so $(\Sigma^{1/2})^{-1}(\boldsymbol{Y}-\boldsymbol{\mu}) \sim N(\mathbf{0}, I)$, with $\mathbf{0}$ a p-dimensional zero vector and I a $p\times p$ identity matrix. This is the *Mahalanobis transform* applied to the multivariate normal $\boldsymbol{Y}$.

c) $(\boldsymbol{Y}-\boldsymbol{\mu})^{\mathrm{T}}\Sigma^{-1}(\boldsymbol{Y}-\boldsymbol{\mu}) = ((\Sigma^{1/2})^{-1}(\boldsymbol{Y}-\boldsymbol{\mu}))^{\mathrm{T}}(\Sigma^{1/2})^{-1}(\boldsymbol{Y}-\boldsymbol{\mu}) = \sum_{j=1}^{p} Z_j^2$, where Z_j are the components of $(\Sigma^{1/2})^{-1}(\boldsymbol{Y}-\boldsymbol{\mu})$. Since all pairwise covariances of the Z_j are zero, they are independent (they are components of a multivariate normal vector). The sum of squares of p independent standard normal variables has the chi-squared distribution with p degrees of freedom.

Problem 2

a) The residual plot shows no particular pattern. The normal Q–Q plot might indicate some deviation from normality – most middle points lie above the line and upper right points below the line. The F-test of the model shows that it is significant on the 0.05 level.

It is true that $R^2 = 0.4805$ is not a particularly good fit, but this does not indicate, as many had written in their papers, that the model is not good. In fact, even if the multiple regression model is perfect, R^2 may be far from 1 due to a high variance σ^2 of the responses.

The larger number of points in the (right) mid-range of the residual plot does not indicate any deviation from the model assumptions. It simply means that a majority of the predictions (fitted values) lie in the mid-range, which is influenced by what the values of the covariates happened to be. Also, it is hard to read any heteroscedasticy (non-constant variance) out of the plot, as there are so few points at the low and high ends of the range of the predictions. Finally, the sum of the (raw) residuals is *always* 0, although one could argue whether there are more residuals of one sign than the other (with a tendency to greater absolute values for the latter).

From the R output, we see that $\hat{\sigma} = 5.956$. $\hat{\sigma}^2 = \text{SSE}/(n-p)$, where $n = 20$ is the number of observations and $p = 4$ is the number of coefficients. So $\text{SSE} = (n-p)\hat{\sigma}^2 = (20-4) \cdot 5.956^2 = 567.6$. Next, $0.4805 = R^2 = 1 - \text{SSE}/\text{SST}$, so $\text{SST} = \text{SSE}/(1-R^2) = 567.58/(1-0.4805) = 1092.6$. Finally, $\text{SSR} = \text{SST} - \text{SSE} = 1092.56 - 567.58 = 525.0$.

b) In best subset selection, all 2^k models including various subsets of the $k = p-1$ covariates in addition to the intercept are considered. Among models including j covariates, the one with the best fit (highest R^2 or smallest SSE) is chosen, $j = 0, 1, \ldots, k$. Among these $k+1$ candidate models, the best, according to some model choice criterion (such as large adjusted R^2, small Mallows' C_P, small AIC or small BIC) is finally selected.

The philosophy of model choice criteria is that they not only reward good fit, but also penalize complexity (number of covariates). R^2 and SSE do not penalize complexity – in fact, R^2 can only increase and SSE only decrease if new covariates are added to a model.

For the current data set, a model including weight and protein, but not age, in addition to an intercept, has the lowest C_P among the best models of each size.

Problem 3

a) X has one column for each of the four dummy covariates of the second formulation. So

$$X = \begin{pmatrix} 1 & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 1 \end{pmatrix},$$

a 60×4 matrix. The first 15 rows are identical, so are the 15 next, etc. It is easy to verify that

$$X^{\mathrm{T}}X = \begin{pmatrix} 15 & 0 & 0 & 0 \\ 0 & 15 & 0 & 0 \\ 0 & 0 & 15 & 0 \\ 0 & 0 & 0 & 15 \end{pmatrix} = 15I,$$

so that $(X^{\mathrm{T}}X)^{-1} = \frac{1}{15}I$.

b) H_0 can be written $C\boldsymbol{\beta} = \mathbf{0}$, where $C = (0\ 0\ 1\ -1)$ and $\boldsymbol{\beta} = (\beta_1\ \beta_2\ \beta_3\ \beta_4)^{\mathrm{T}}$. If H_0 is true,

$$F = \frac{(C\hat{\boldsymbol{\beta}})^{\mathrm{T}}(C(X^{\mathrm{T}}X)^{-1}C^{\mathrm{T}})^{-1}C\hat{\boldsymbol{\beta}}/r}{\mathrm{SSE}/(n-p)},$$

where $\hat{\boldsymbol{\beta}} = (\hat{\beta}_1\ \hat{\beta}_2\ \hat{\beta}_3\ \hat{\beta}_4)^{\mathrm{T}}$ is the vector of least-squares estimators, $r = \operatorname{rank} C = 1$, $n = 60$ is the number of observations and $p = 4$ the number of columns in the design matrix, has an F-distribution with $r = 1$ and $n - p = 56$ degrees of freedom. Here, $C\hat{\boldsymbol{\beta}} = \hat{\beta}_3 - \hat{\beta}_4$ and $C(X^{\mathrm{T}}X)^{-1}C^{\mathrm{T}} = \frac{1}{15}CC^{\mathrm{T}} = \frac{2}{15}$. The value of F becomes

$$\frac{\frac{15}{2}(\hat{\beta}_3 - \hat{\beta}_4)^2}{\mathrm{SSE}/(60-4)} = \frac{\frac{15}{2}(1.0902858 - 0.1752633)^2}{43.04524/(60-4)} = 8.17.$$

In the statistical tables of critical values corresponding to 0.05 for the F-distribution, we find 4.03 for 1 and 50 degrees of freedom, and 4.00 for 1 and 60 degrees of freedom. So at the 0.05 level we reject H_0 and conclude that $\beta_3 \neq \beta_4$. (We would arrive at the same conclusion even for level 0.01.)

c) The family-wise error rate (FWER) is the probability of making at least one type I error (reject a true null hypothesis). Using the Bonferroni method, if α/m is used as significance level for each of m tests, FWER $\leq \alpha$ even if all null hyptheses are true.

Here there are six tests, so by the Bonferroni method, we should use significance level $0.05/6 = 0.0083$ for each individual test to keep FWER below 0.05. The null hypotheses $\beta_2 = \beta_3$ and $\beta_3 = \beta_4$ are rejected.

Problem 4

a) According to the general rule $\operatorname{Cov}(A\boldsymbol{Y}) = A(\operatorname{Cov}\boldsymbol{Y})A^{\mathrm{T}}$, we have

$$\begin{aligned}\operatorname{Cov}\hat{\boldsymbol{\beta}} = \operatorname{Cov}((X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}\boldsymbol{Y}) &= (X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}(\operatorname{Cov}\boldsymbol{Y})((X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}})^{\mathrm{T}} \\ &= \sigma^2(X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}X(X^{\mathrm{T}}X)^{-1} = \sigma^2(X^{\mathrm{T}}X)^{-1}\end{aligned}$$

and $\operatorname{Cov}\tilde{\boldsymbol{\beta}} = \operatorname{Cov}(B\boldsymbol{Y}) = B(\operatorname{Cov}\boldsymbol{Y})B^{\mathrm{T}} = \sigma^2 BB^{\mathrm{T}}$.

For all $\boldsymbol{\beta}$, $\boldsymbol{\beta} = E\tilde{\boldsymbol{\beta}} = E(B\boldsymbol{Y}) = BE\boldsymbol{Y} = BX\boldsymbol{\beta}$, so that $BX = I_p$ (consider $\boldsymbol{\beta} = (1\ 0\ \cdots\ 0)^{\mathrm{T}}$ to show that the first column of BX is $(1\ 0\ \cdots\ 0)^{\mathrm{T}}$ and so on).

b)

$$\begin{aligned}MM^{\mathrm{T}} &= \sigma^2(B - (X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}})(B - (X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}})^{\mathrm{T}} \\ &= \sigma^2(B - (X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}})(B^{\mathrm{T}} - X(X^{\mathrm{T}}X)^{-1}) \\ &= \sigma^2(BB^{\mathrm{T}} - (X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}B^{\mathrm{T}} - BX(X^{\mathrm{T}}X)^{-1} + (X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}X(X^{\mathrm{T}}X)^{-1}) \\ &= \sigma^2(BB^{\mathrm{T}} - (X^{\mathrm{T}}X)^{-1} - (X^{\mathrm{T}}X)^{-1} + (X^{\mathrm{T}}X)^{-1}) \qquad \text{since } BX = I_p \\ &= \sigma^2 BB^{\mathrm{T}} - \sigma^2(X^{\mathrm{T}}X)^{-1} \\ &= \operatorname{Cov}\tilde{\boldsymbol{\beta}} - \operatorname{Cov}\hat{\boldsymbol{\beta}}.\end{aligned}$$

On the diagonal of $\operatorname{Cov}\tilde{\boldsymbol{\beta}} - \operatorname{Cov}\hat{\boldsymbol{\beta}}$, we find the differences of the variances of the components of $\tilde{\boldsymbol{\beta}}$ and the variances of the corresponding components of $\hat{\boldsymbol{\beta}}$. Specifically, if $\tilde{\beta}_j$ and $\hat{\beta}_j$ denote the jth component of $\tilde{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\beta}}$, respectively, then the jth diagonal entry is $\operatorname{Var}\tilde{\beta}_j - \operatorname{Var}\hat{\beta}_j$. But, denoting the jk entry of M m_{jk}, the jth diagonal entry of MM^{T} is $\operatorname{Var}\tilde{\beta}_j - \operatorname{Var}\hat{\beta}_j = \sum_{k=1}^{p} m_{jk}m_{jk} = \sum_{k=1}^{p} m_{jk}^2 \geq 0$, where the dimensions of MM^{T} are $p \times p$.

The conclusion is that the variances of the components of any unbiased estimator of $\boldsymbol{\beta}$ of form $B\boldsymbol{Y}$, are always greater than or equal to the variances of the least-squares estimator. This is called the *Gauss–Markov Theorem.*