TMA4267 2019 spring – Solutions

Problem 1

a) A covariance matrix of a random vector is symmetric and positive semidefinite. We also usually want the matrix to be positive definite – otherwise there is a non-trivial linear combination of the components having variance zero.

Σ_X is symmetric. Its eigenvalues are the solution of the equation $0 = \det(\lambda I - \Sigma_X) = \left|\begin{smallmatrix} \lambda-1 & -a \\ -a & \lambda-1 \end{smallmatrix}\right| = (\lambda-1)^2 - a^2$, that is $\lambda - 1 = \pm a$, $\lambda = 1 \pm a$. All eigenvalues are positive, and thus Σ_X positive definite, if and only if $-1 < a < 1$.

b) $\mathrm{Cov}(X_1+X_2, X_1-X_2) = \mathrm{Cov}\big((1\ 1)\boldsymbol{X}, (1\ -1)\boldsymbol{X}\big) = (1\ 1)\Sigma_X(1\ -1)^{\mathrm{T}} = (1\ 1)\left(\begin{smallmatrix} 1 & a \\ a & 1 \end{smallmatrix}\right)\left(\begin{smallmatrix} 1 \\ -1 \end{smallmatrix}\right) = (1\ 1)\left(\begin{smallmatrix} 1-a \\ a-1 \end{smallmatrix}\right) = 1 - a + a - 1 = 0$. Since $\boldsymbol{X}$ is bivariate normal, $(1\ 1)\boldsymbol{X}$ and $(1\ -1)\boldsymbol{X}$ are independent if and only if their covariance is zero, that is, for all a, $-1 < a < 1$.

$$\begin{aligned}
\mathrm{Cov}(Y_1+Y_2+Y_3, Y_1-Y_2-Y_3) &= \mathrm{Cov}\big((1\ 1\ 1)\boldsymbol{Y}, (1\ -1\ -1)\boldsymbol{Y}\big) \\
&= (1\ 1\ 1)\Sigma_Y(1\ -1\ -1)^{\mathrm{T}} = \begin{pmatrix} 1 & 1 & 1 \end{pmatrix}\begin{pmatrix} 1 & a & 0 \\ a & 1 & b \\ 0 & b & 1 \end{pmatrix}\begin{pmatrix} 1 \\ -1 \\ -1 \end{pmatrix} \\
&= \begin{pmatrix} 1 & 1 & 1 \end{pmatrix}\begin{pmatrix} 1-a \\ a-1-b \\ -b-1 \end{pmatrix} = 1 - a + a - 1 - b - b - 1 = -2b - 1
\end{aligned}$$

Since $\boldsymbol{Y}$ is trivariate normal, $(1\ 1\ 1)\boldsymbol{Y}$ and $(1\ -1\ -1)\boldsymbol{Y}$ are independent if and only if their covariance is zero, that is, for $b = -\frac{1}{2}$.

Further, Σ_Y is symmetric for all a and b. Its eigenvalues are the solutions of the equation

$$\begin{aligned}
0 = \det(\lambda I - \Sigma_Y) &= \begin{vmatrix} \lambda-1 & -a & 0 \\ -a & \lambda-1 & -b \\ 0 & -b & \lambda-1 \end{vmatrix} = (\lambda-1)\begin{vmatrix} \lambda-1 & -b \\ -b & \lambda-1 \end{vmatrix} + a\begin{vmatrix} -a & -b \\ 0 & \lambda-1 \end{vmatrix} \\
&= (\lambda-1)\big((\lambda-1)^2 - b^2\big) - a^2(\lambda-1) = (\lambda-1)\big((\lambda-1)^2 - a^2 - b^2\big),
\end{aligned}$$

that is, 1 and $1 \pm \sqrt{a^2+b^2}$. For Σ_Y to be positive definite, we need all three to be positive, that is, $\sqrt{a^2+b^2} < 1$, or $a^2 < 1 - b^2$, which means $|a| < \sqrt{3}/2$ for $b = -\frac{1}{2}$. The answer to the question is $b = -\frac{1}{2}$ and $|a| < \sqrt{3}/2$.

Problem 2

The resolution of a fractional factorial experiment is the minimum number of factors in the defining relation. We want it high to avoid aliasing between main effects and low-order interactions.

We have generators $D = AB$ and $E = AC$. Then we have the defining relation

$$\begin{aligned} 1 &= ABD && \text{(first generator)} \\ &= ACE = BCDE && \text{(second generator and product with previous)} \end{aligned}$$

The minimum length of the words is three, which means that the design is of resolution III.

From the defining relation we find the following aliasing of main effects with 2-factor interactions: $A = BD = CE$, $B = AD$, $C = AE$, $D = AB$ and $E = AC$.

Problem 3

a) In best subset selection, all 2^k models including various subsets of the $k = p-1$ covariates in addition to the intercept is considered. Among models including j covariates, the one with the best fit (highest R^2 or smallest SSE) is chosen. The best models with 1, 2 and 3 covariates are shown in the R printout and the upper figure. Among those candidate model, the best one – the one with the smallest Mallows' C_P was chosen. The reduced model included `lactic` and `logh2s`.

The residual plot shows no particular pattern except one observation for which the residual was large (approximately 3). This observation also stands out in the Q–Q plot, otherwise it is OK. The F-test shows that the model is significant on the 0.05 level.

From the R output we see that $\hat{\sigma} = 9.942$, and $\hat{\sigma}^2 = \text{SSE}/(n-p)$, where $n = 30$ is the number of observations and $p = 3$ is the number of coefficients. So $\text{SSE} = (n-p)\hat{\sigma}^2 = 27 \cdot 9.942^2 = 2669$.

b) $\text{Var}\,\hat{\beta}_j$ will be smaller the smaller the variance σ^2 of the errors is, and smaller the smaller the linear dependence between x_j and the other covariates is (small R_j^2), and smaller the larger the variability of x_j is (measured by $\sum(x_{ij} - \bar{x}_j)^2$ or the sample variance of x_j).

Problem 4

a) X has one column for each of three dummy covariates – if we let the level 1 observations be the ten first, the level 2 observations the ten next, and the level three observations the ten last, then

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

a 30×3 matrix. The first 10 rows are identical, so are the 10 next, etc., and it is given, and also easy to verify, that

$$X^{\mathrm{T}}X = \begin{pmatrix} 10 & 0 & 0 \\ 0 & 10 & 0 \\ 0 & 0 & 10 \end{pmatrix} = 10I,$$

so that $(X^{\mathrm{T}}X)^{-1} = \frac{1}{10}I$.

The null hypothesis can be written $C\boldsymbol{\beta} = \mathbf{0}$, where $C = (0\ 1\ -1)$ and $\boldsymbol{\beta} = (\mu_1\ \mu_2\ \mu_3)^{\mathrm{T}}$. If the null hypothesis 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 $\hat{\boldsymbol{\beta}} = (\hat{\mu}_1\ \hat{\mu}_2\ \hat{\mu}_3)^{\mathrm{T}}$ is the vector of least-squares estimators, $r = \operatorname{rank} C = 1$, $n = 30$ is the number of observations and $p = 3$ the number of columns in the design matrix, has an F-distribution with $r = 1$ and $n - p = 27$ degrees of freedom. Here, $C\hat{\boldsymbol{\beta}} = \hat{\mu}_2 - \hat{\mu}_3$ and $C(X^{\mathrm{T}}X)^{-1}C^{\mathrm{T}} = \frac{1}{10}CC^{\mathrm{T}} = \frac{1}{5}$. The value of F becomes

$$\frac{5(\hat{\mu}_2 - \hat{\mu}_3)^2}{\mathrm{SSE}/(30-3)} = \frac{5(0.2488 - 1.1663)^2}{24.00/(30-3)} = 4.74.$$

In the statistical tables of critical values corresponding to 0.05 for the F-distribution, we find 4.21 for 1 and 27 degrees of freedom. So at the 0.05 level we reject the null hypothesis and conclude that $\mu_2 \neq \mu_3$.

b) If m hypothesis tests are performed, then Bonferroni's method is to use significance level α/m for each test. Then the family-wise error rate (FWER) – the probability of making at least one type I error (reject a true null hypothesis) – is less than or equal to α: If m_0 of the m null hypotheses are true, then

$$\begin{aligned} &P(\text{at least 1 type I error}) \\ &\quad = P(\text{reject true null hypothesis } 1 \cup \cdots \cup \text{reject true null hypothesis } m_0) \\ &\quad \leq P(\text{reject true null hypothesis } 1) + \cdots + P(\text{reject true null hypothesis } m_0) \\ &\quad \leq m_0 \frac{\alpha}{m} \leq \alpha \end{aligned}$$

(the proof is not required in your papers).

Here we have three tests, so by the Bonferroni method, we should use significance level $0.05/3 = 0.0167$ for each individual test to keep FWER below 0.05. None of the null hypotheses are rejected.

c) If *all three null hypotheses are true*, then also the null hypothesis $\mu_1 = \mu_2 = \mu_3$ is true. None of the three original null hypotheses are rejected unless also the null hypothesis $\mu_1 = \mu_2 = \mu_3$ is rejected. Since the test of the latter has level 0.05, the probability of rejection of any of the original hypotheses, and thus of making a type I error, is less than or equal to 0.05:

$$\begin{aligned} &P(\text{at least 1 type I error}) \\ &\quad = P\big(\mu_1 = \mu_2 = \mu_3 \text{ rejected} \cap (\mu_1 = \mu_2 \text{ or } \mu_1 = \mu_3 \text{ or } \mu_2 = \mu_3 \text{ rejected})\big) \\ &\quad \leq P(\mu_1 = \mu_2 = \mu_3 \text{ rejected}) \leq 0.05 \end{aligned}$$

If *two of the three null hypotheses are true*, then in the present situation the third is also true (e.g., if $\mu_1 = \mu_2$ and $\mu_2 = \mu_3$, then $\mu_1 = \mu_3$), and we have the previous case.

If *one of the three null hypotheses is true*, then the only possibility of making a type I error is to reject the only true null hypothesis. But this is only done if it is rejected at the 0.05 level, meaning that the probability is less than or equal to 0.05. E.g., if $\mu_1 = \mu_2$ is true and the two other null hypotheses false, then

$$\begin{aligned} P(\text{at least 1 type I error}) &= P(\mu_1 = \mu_2 = \mu_3 \text{ rejected} \cap \mu_1 = \mu_2 \text{ rejected}) \\ &\leq P(\mu_1 = \mu_2 \text{ rejected}) \leq 0.05. \end{aligned}$$

If *none of the null hypotheses are true*, then a type I error cannot be done.

By this method, $\mu_1 = \mu_3$ and $\mu_2 = \mu_3$ is rejected, since both are rejected at the 0.05 level and additionally $\mu_1 = \mu_2 = \mu_3$ is rejected at the 0.05 level.