Norwegian University of Science and Technology
Department of Mathematical Sciences

Tentative Solutions to TMA4267 Linear Statistical Models 22 May 2014

Problem 1 Random vector

a)

$$\mathrm{E}(\boldsymbol{Y}) = \boldsymbol{A}\boldsymbol{\mu} = \begin{pmatrix} \frac{2}{3} & -\frac{1}{3} & -\frac{1}{3} \\ -\frac{1}{3} & \frac{2}{3} & -\frac{1}{3} \\ -\frac{1}{3} & -\frac{1}{3} & \frac{2}{3} \end{pmatrix} \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}$$

$$\begin{aligned} \mathrm{Cov}(\boldsymbol{Y}) &= \boldsymbol{A}\,\mathrm{Cov}(\boldsymbol{X})\boldsymbol{A}^T = \boldsymbol{A}\boldsymbol{I}\boldsymbol{A}^T = \boldsymbol{A}\boldsymbol{A}^T \\ &= \begin{pmatrix} \frac{2}{3} & -\frac{1}{3} & -\frac{1}{3} \\ -\frac{1}{3} & \frac{2}{3} & -\frac{1}{3} \\ -\frac{1}{3} & -\frac{1}{3} & \frac{2}{3} \end{pmatrix} \begin{pmatrix} \frac{2}{3} & -\frac{1}{3} & -\frac{1}{3} \\ -\frac{1}{3} & \frac{2}{3} & -\frac{1}{3} \\ -\frac{1}{3} & -\frac{1}{3} & \frac{2}{3} \end{pmatrix} \\ &= \begin{pmatrix} \frac{2}{3} & -\frac{1}{3} & -\frac{1}{3} \\ -\frac{1}{3} & \frac{2}{3} & -\frac{1}{3} \\ -\frac{1}{3} & -\frac{1}{3} & \frac{2}{3} \end{pmatrix} = \boldsymbol{A} \end{aligned}$$

The covariance between X_1 and X_2 is zero. This doesn't imply that X_1 and X_2 are independent, unless the vector with elements X_1 and X_2 are binormal. We have no information about the distribution of $\boldsymbol{X}$, and can't assume that zero covariance implies independence.

For Y_1 and Y_2 the covariance is $-\frac{1}{3}$, and Y_1 and Y_2 is dependent.

Find the expected value of $\boldsymbol{X}^T\boldsymbol{A}\boldsymbol{X}$.
We may use the trace-formula:

$$\mathrm{E}(\boldsymbol{X}^T\boldsymbol{A}\boldsymbol{X}) = \mathrm{tr}(\boldsymbol{A}\boldsymbol{\Sigma}) + \boldsymbol{\mu}^T\boldsymbol{A}\boldsymbol{\mu} = \mathrm{tr}\,\boldsymbol{A} + \boldsymbol{\mu}^T\boldsymbol{0} = 3 \cdot \frac{2}{3} = 2$$

b) $\boldsymbol{A}$ is clearly symmetric, which we can see by $\boldsymbol{A}^T = \boldsymbol{A}$. A projection matrix is an idempotent matrix, that is, $\boldsymbol{A}\boldsymbol{A} = \boldsymbol{A}$. We have already seen this in **a)**.

The rank of a symmetric idempotent matrix equals its trace, which we found in **a)** to be $\text{tr}(\boldsymbol{A}) = 2$.

Derive the distribution of $\boldsymbol{X}^T\boldsymbol{A}\boldsymbol{X}$.
From T3.26 in BF2010 we know that if $\boldsymbol{X} \sim N(\boldsymbol{0}, \sigma^2\boldsymbol{I})$ and $\boldsymbol{A}$ is a symmetric, idempotent matrix with rank r then $\boldsymbol{X}^T\boldsymbol{A}\boldsymbol{X} \sim \sigma^2\chi^2_r$.

We have that $\boldsymbol{X} \sim N(\boldsymbol{\mu}, \boldsymbol{I})$, and $\boldsymbol{A}$ is symmetric and idempotent with rank 2. We need to rewrite our expression so that we have a normally distributed random vector with mean zero and identity covariance matrix.

We subtract the mean and write

$$(\boldsymbol{X}-\boldsymbol{\mu})^T\boldsymbol{A}(\boldsymbol{X}-\boldsymbol{\mu}) = \boldsymbol{X}^T\boldsymbol{A}\boldsymbol{X} - \boldsymbol{\mu}^T\boldsymbol{A}\boldsymbol{X} - \boldsymbol{X}^T\boldsymbol{A}\boldsymbol{\mu} + \boldsymbol{\mu}^T\boldsymbol{A}\boldsymbol{\mu} = \boldsymbol{X}^T\boldsymbol{A}\boldsymbol{X}$$

since $\boldsymbol{A}\boldsymbol{\mu} = \boldsymbol{0}$. Define $\boldsymbol{Z} = \boldsymbol{X} - \boldsymbol{\mu}$, where $\boldsymbol{Z} \sim N(\boldsymbol{0}, \boldsymbol{I})$. We may thus use the above theorem, to find that $\boldsymbol{X}^T\boldsymbol{A}\boldsymbol{X} = \boldsymbol{Z}^T\boldsymbol{A}\boldsymbol{Z} \sim \chi^2_2$, that is χ^2 with 2 degrees of freedom.

Tabeller og formler i statistikk, page 5, we see that 6 is the critical value in the χ^2-distribution with 2 degrees of freedom and probability 0.05 (value in table is 5.991). The probability that the quadratic form is smaller than 6 is thus 95%.

Problem 2 Galapagos species

a) The fitted regression model is:

$$\widehat{\texttt{Species}} = 7.07 - 0.02\cdot\texttt{Area} + 0.32\cdot\texttt{Elevation} + 0.009\cdot\texttt{Nearest} - 0.24\cdot\texttt{Scruz} - 0.75\cdot\texttt{Adjacent}$$

This model explains 77% of the variability in the data. The regression is significant (the hypothesis that all regression coefficients are zero is rejected) and t-tests claim that `Elevation` and `Adjacent` are significant covariates.

The residual plots: The plot of studentized residuals vs. fitted values hints to heteroscedasticity in the errors (differing variances), and the qq-plot shows deviance from the normal distribution in the tails. The latter is also observed by looking at the Anderson-Darling normality test, which gives a p-value of 0.0002 (reject the null hypothesis that the errors are normal). The Box-Cox plot doesn't include 1 in the 95% confidence interval (dotted lines in the plot), and suggests that the cube root transform ($\lambda = 1/3$) may be suitable as a variance stabilizating transform.

b) Let us assume that we have p covariates plus the intercept (notation from Part 6-7 of the course, in Parts 1-5 we have used p to include the intercept).

Estimate (Intercept): t value*Std. Error=7.365· 0.305=2.25
Meaning: estimate for the regression coefficient. Intercept associated with first column of design matrix (first column of ones) $\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y}$.

p-value of `Area`:two tails of t-distribution with 24 degrees of freedom. Can't find precise value, but from table 4 of Tabeller og formler i statistikk we see that the critical value in the t-distribution with 24 degrees of freedom is 2.064 for $\alpha = 0.025$. This means that the p-value will approximately be 0.05.
Meaning: Test the null hypothesis that $\beta_{Area} = 0$ vs. $\beta_{Area} \neq 0$, with the other four covariates present in the model, and produce a p-value of the test.

Std.Error of `Nearest`:estimate/tobs=0.012/0.7=0.017
Meaning: the estimated standard deviation for the regression estimate. Mathematically the corresponding (`Nearest`) diagonal element of the square root of $(\boldsymbol{X}^T\boldsymbol{X})^{-1}s^2$, where s^2 is the estimate for the regression variance σ^2.

Adjusted R-squared: $1-(1-R^2)(n-1)/(n-p-1) = 1-(1-0.7543)\cdot 29/24 = 0.7032$, or in a two stage process by first observing SSE=$s^2\cdot(n-p) = 0.9716^2\cdot 24 = 22.65$ and then finding SST from R^2=1-SSE/SST, SST=SSE/(1-R^2), and finally using $R^2_{\text{adj}} = 1 - \frac{\frac{\text{SSE}}{n-p-1}}{\frac{\text{SST}}{n-1}}$.

Yes, I would prefer model B to A. The plot of standardized residuals vs fitted values shows no clear structure, and the qq-plot looks much better for B and A. The Anderson-Darling normality test doesn't reject the null hypothesis of normal data.

c) Let SSE be the sum-of-squares of error, SSR be the regression sum-of-squares, and SST be the total som of squares. Then R^2: coefficient of multiple determination is defined as

$$1 - \text{SSE/SST} = \text{SSR/SST}$$

and is interpreted as the amount of variability in the data that is accounted for by the regression. R^2 will increase when a regressors are added to the model, even if the new regressors are independent of the response. Why? The least squares estimator will minimize SSE and if the regression coefficient for the new regressor is estimated to be a value different from zero, this means that the SSE of this larger model will be smaller than the SSE of the smaller model.

The R^2_{adj} is constructed to also include information about the number of parameters estimated and the number of observations in the data set. Assume we have 1 intercept and in addition p regression parameters, then

$$R^2_{\text{adj}} = 1 - \frac{\frac{\text{SSE}}{n-p-1}}{\frac{\text{SST}}{n-1}}$$

R^2 will always increase then new covariates are added to the model, so R^2 can only be used to select the best model among models with the same number of covariates. This

is done when in best subset selection one model is reported for each total number of covariates. To choose between these models a criterion taking into account the number of covariates in the model need to be used, and one such criterion is R^2_{adj}. We therefore use R^2_{adj} to choose between the best models of each size.

In our example the best model is according to this strategy the model with four covariates. These are all covariates except `Nearest`. To write down the estimated regression equation we need to refit this model.

Lasso regression adds a penalty term to the least squares criterion to make the model more sparse. This may give a robust fit and avoid overfitting. The penalty term for lasso is the sum of the absolute value of the regression coefficients, and the optimization procedure is to minimize with respect to $\boldsymbol{\beta}$ the following quantity

$$(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})^T(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}) + \lambda \sum_{j=1}^{p} \mid \beta_j \mid$$

The difficulty with using R^2_{adj} is that the number of parameters that enter into the formula is not defined since the lasso regression shrinks parameter estimates - some by a certain amount and some all the way down to zero. The model complexity for the lasso is defined from λ and we can't put λ into the R^2_{adj}-formula.

Instead cross-validation is used for model choice.

The following was not asked for: SST is the same for all models, and thus R^2 is proportional to SSE. SSE can therefore be used as a model selection criterion in the cross-validation. In short - this involves: 10-fold, training set, test set, fit, predict, SSE on the test set, and sum to produce SSE on the whole set. Choose the penalty parameter with the smallest SSE.

The fitted lasso regression model is:

$$\widehat{\texttt{Species}^{1/3}} = 3.54 + 0.00028 \cdot \texttt{Elevation}$$

Problem 3 Design of experiments

A column for BC is added to the design,

	A	B	C	D	BC
1	-1	-1	-1	1	1
2	1	-1	-1	1	1
3	-1	1	-1	-1	-1
4	1	1	-1	-1	-1
5	-1	-1	1	-1	-1
6	1	-1	1	-1	-1
7	-1	1	1	1	1
8	1	1	1	1	1

a) What type of experiment is this?
We see that we have a full factorial design in the factors A, B, C, but there is a fourth factor D added. This is a half fraction of a 2^4 design, also called a 2^{4-1}-design.

What is the generator and the defining relation for the experiment?
The generator for the design is D=BC (which is seen from the table above after the BC column is added). The defining relation is then I=BCD.

What is the resolution of the experiment?
The resolution of the design equals the number of letters in the defining relation, thus the resolution is III.

Write down the alias structure of the experiment.
A=ABCD, B=CD, C=BD, D=BC.
AB=ACD, AC=ABD, AD=ABC
I=BCD

Problem 4 Multiple Linear Regression

a) Let $(\lambda_i, \boldsymbol{e}_i)$, $i = 1, ..., p$ be the eigenvalues and eigenvectors of $\boldsymbol{V}$. Let $\boldsymbol{P}$ be the $(p \times p)$ matrix of eigenvectors,

$$\boldsymbol{P} = [\boldsymbol{e}_1 \boldsymbol{e}_2 \cdots \boldsymbol{e}_p]$$

and $\boldsymbol{\Lambda}$ be a diagonal matrix with the eigenvalues $\lambda_1, \lambda_1, ..., \lambda_p$ on the diagonal. Then $\boldsymbol{V}^{-\frac{1}{2}}$ is defined as

$$\boldsymbol{V}^{-\frac{1}{2}} = \boldsymbol{P}\boldsymbol{\Lambda}^{-\frac{1}{2}}\boldsymbol{P}^T$$

Observe that $\boldsymbol{V}^{-\frac{1}{2}}$ is symmetric, and that $\boldsymbol{V}^{-\frac{1}{2}}\boldsymbol{V}^{-\frac{1}{2}} = \boldsymbol{V}^{-1}$.

$$\begin{aligned}\boldsymbol{Y} &= \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \\ \boldsymbol{V}^{-\frac{1}{2}}\boldsymbol{Y} &= \boldsymbol{V}^{-\frac{1}{2}}\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{V}^{-\frac{1}{2}}\boldsymbol{\varepsilon} \\ \boldsymbol{Y}^* &= \boldsymbol{X}^*\boldsymbol{\beta} + \boldsymbol{\varepsilon}^*\end{aligned}$$

where $\boldsymbol{\varepsilon}^* \sim N(\boldsymbol{0}, \sigma^2\boldsymbol{I})$. To see this calculate $\mathrm{Cov}(\boldsymbol{\varepsilon}^*) = \boldsymbol{V}^{-\frac{1}{2}}\,\mathrm{Cov}(\boldsymbol{\varepsilon})\boldsymbol{V}^{-\frac{1}{2}} = \boldsymbol{V}^{-\frac{1}{2}}\sigma^2\boldsymbol{V}\boldsymbol{V}^{-\frac{1}{2}} = \sigma^2\boldsymbol{I}$.

We have now the ordinary least squares problem in the new quantities $\boldsymbol{Y}^*$, $\boldsymbol{X}^*$ and $\boldsymbol{\varepsilon}^*$, and know that the least squares solution is

$$\begin{aligned}\tilde{\boldsymbol{\beta}} &= (\boldsymbol{X}^{*T}\boldsymbol{X}^*)^{-1}\boldsymbol{X}^{*T}\boldsymbol{Y}^* \\ &= (\boldsymbol{X}^T\boldsymbol{V}^{-\frac{1}{2}}\boldsymbol{V}^{-\frac{1}{2}}\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{V}^{-\frac{1}{2}}\boldsymbol{V}^{-\frac{1}{2}}\boldsymbol{Y} \\ &= (\boldsymbol{X}^T\boldsymbol{V}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{V}^{-1}\boldsymbol{Y}\end{aligned}$$

Mean:

$$\begin{aligned}\mathrm{E}(\tilde{\boldsymbol{\beta}}) &= \mathrm{E}((\boldsymbol{X}^{*T}\boldsymbol{X}^*)^{-1}\boldsymbol{X}^{*T}\boldsymbol{Y}^*) = (\boldsymbol{X}^{*T}\boldsymbol{X}^*)^{-1}\boldsymbol{X}^{*T}\,\mathrm{E}(\boldsymbol{Y}^*) \\ &= (\boldsymbol{X}^{*T}\boldsymbol{X}^*)^{-1}\boldsymbol{X}^{*T}\boldsymbol{X}^*\boldsymbol{\beta} = \boldsymbol{\beta}\end{aligned}$$

since $\mathrm{E}(\boldsymbol{Y}^*) = \boldsymbol{X}^*\boldsymbol{\beta}$.

The ordinary least square estimator $\hat{\beta} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y}$ is unbiased in this model since the mean of $\boldsymbol{Y}$ doesn't depend on $\boldsymbol{V}$.

$$\begin{aligned}\mathrm{E}(\hat{\beta}) &= \mathrm{E}((\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y}) = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\,\mathrm{E}(\boldsymbol{Y}) \\ &= (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{\beta} = \boldsymbol{\beta}\end{aligned}$$

If we just look at unbiasedness it may appear that the two estimators are equally good. However, since $\tilde{\boldsymbol{\beta}}$ is the least squares estimator (from looking at transformed quantities) we may conclude using the Gauss-Markov Theorem (T3.13 in Bingham and Fry 2010) that $\tilde{\boldsymbol{\beta}}$ has the minimum variance in each component among all the unbiased estimators, BLUE. If we had calculated the covariance matrices of $\tilde{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\beta}}$, we should see this. Thus, $\tilde{\boldsymbol{\beta}}$ should be preferred. Another issue is the fact that $\boldsymbol{V}$ seldom is known, and need to be estimated. The concept of BLUE is handled in detail in our Statistical Inference course.

b) Find the expected value and covariance matrix of $\hat{\boldsymbol{\alpha}}_1$ under the true model

$$\begin{aligned}
\mathrm{E}(\hat{\boldsymbol{\alpha}}_1) &= \mathrm{E}((\boldsymbol{X}_1^T\boldsymbol{X}_1)^{-1}\boldsymbol{X}_1^T\boldsymbol{Y}) \\
&= (\boldsymbol{X}_1^T\boldsymbol{X}_1)^{-1}\boldsymbol{X}_1^T\,\mathrm{E}(\boldsymbol{Y}) == (\boldsymbol{X}_1^T\boldsymbol{X}_1)^{-1}\boldsymbol{X}_1^T(\boldsymbol{X}_1\boldsymbol{\beta}_1 + \boldsymbol{X}_2\boldsymbol{\beta}_2) \\
&= \boldsymbol{\beta}_1 + (\boldsymbol{X}_1^T\boldsymbol{X}_1)^{-1}\boldsymbol{X}_1^T\boldsymbol{X}_2\boldsymbol{\beta}_2
\end{aligned}$$

Thus, $\hat{\boldsymbol{\alpha}}_1$ is a biased estimator for $\boldsymbol{\beta}_1$.

$$\begin{aligned}
\mathrm{Cov}(\hat{\boldsymbol{\alpha}}_1) &= \mathrm{Cov}((\boldsymbol{X}_1^T\boldsymbol{X}_1)^{-1}\boldsymbol{X}_1^T\boldsymbol{Y}) \\
&= (\boldsymbol{X}_1^T\boldsymbol{X}_1)^{-1}\boldsymbol{X}_1^T\,\mathrm{Cov}(\boldsymbol{Y})\boldsymbol{X}_1(\boldsymbol{X}_1^T\boldsymbol{X}_1)^{-1} \\
&= (\boldsymbol{X}_1^T\boldsymbol{X}_1)^{-1}\boldsymbol{X}_1^T\sigma^2\boldsymbol{I}\boldsymbol{X}_1(\boldsymbol{X}_1^T\boldsymbol{X}_1)^{-1} \\
&= \sigma^2(\boldsymbol{X}_1^T\boldsymbol{X}_1)^{-1}
\end{aligned}$$

Observe, $\mathrm{Cov}(\hat{\boldsymbol{\alpha}}_1)$ is not dependent on $\boldsymbol{\beta}_2$.

We see that the bias term for $\hat{\boldsymbol{\alpha}}_1$ is $(\boldsymbol{X}_1^T\boldsymbol{X}_1)^{-1}\boldsymbol{X}_1^T\boldsymbol{X}_2\boldsymbol{\beta}_2$. When is the bias term equal to zero?

When $\boldsymbol{\beta}_2 = \boldsymbol{0}$ there is no bias (but that is not so exciting). The bias is also zero when $\boldsymbol{X}_1^T\boldsymbol{X}_2 = \boldsymbol{0}$. This will happen if the two matrices are orthogonal. If we think back to Part 6: DOE this is true for DOE, and is why the interpretation of the DOE coefficients is easy and the model chosen doesn't influence the unbiasedness of the coefficients - but influences the variance thereof.