## Problems: STK4070-sp11

### Problem 1

**a):** Consider the data set sigarett and verify that if co and tar are used as covariates, the coefficient of tar can be obtaine either by

- including both regressors, or
- regressing nikot and tar on co, and then regress the residuals fron nikot on co on the residuals from tar on co.

b): Verify by formal calculations that the result must be true.

c): Let

$$Y = X\beta + Z\gamma + \epsilon$$

Generalize and verify the result from part b) to this setting.

### Problem 2

Verify that for the model

$$Y = X\beta + \epsilon,$$

$F = \frac{(n-(p+1))R^2}{p(1-R^2)}$. What null hypotesis is tested using this statistic?

### Problem 3

Consider the data set sigarett, and fit the model

$$nikot = \beta_0 + \beta_1 co + \beta_2 tar$$

a): Suppose only co is included as a regressor. Find the expectation of the estimator of the slope. Fit such a simple linear regression, and discuss and interpret the results.

b): Discuss whether it may be appropriate to include regressors as $co^2$, $tar^2$ and $co \times tar$.

Problem 4

The data set sleep2 contains information of the sleep lengths of 58 mammals. The column "TS" indicates the total sleep length in hours of the mammal, "BodyWt" is the body weight in kg and "D" is an index of danger, 0 is no danger and 1 is danger.

a): Fit a simple linear regression model of the form

```
lm(TS~log(BodyWt),x=T,data=sleep2)
```

and verify that the estimates of $\beta_0$, $\beta_1$ and $\sigma^2$ agree with what you get from the standard formulas.

b): Compute 95% confidence intervals for $\beta_0$ and $\beta_1$.

c): Discuss the model fit in this case and make a plot the observations and the fitted regression line.

d): Now we consider also including the danger index. Fit models of the type
- One common regression line
- Common intercept
- Paralell lines
- General situation

e): Compare the models using the function anova in R. Which model is most appropriate in this situation?

f): Apply the function vcov to the model you ended up with in part e) and discuss the result.

Problem 5

Consider the linear model

$$Y = X\beta + \epsilon$$

where $X$ is a $n \times (p+1)$ matrix of full rank. We assume that the elements $\epsilon_1, \ldots, \epsilon_n$ are independent and that $\epsilon_i \sim N(0, \sigma^2), i = 1, \ldots, n$.

Partition $X$ as $X = (X_1, X_2)$ where $X_1$ consists of the first $k$ columns of $X$.

Let $\hat{\beta}$ be the OLS-estimator for a model having design matrix $X$ and $\hat{\alpha}$ be the OLS-estimator for a model having design matrix $X_1$. Let $\hat{\beta}_H$ be the $p+1$-dimensional vector given by $\hat{\beta}_H' = (\hat{\alpha}', 0, \ldots, 0)'$. Let $RSS_0$ and $RSS_1$ be the residual sum of squares for the two models.

a) Show that

$$\hat{\beta}'(X'X)\hat{\beta} - \hat{\beta}_H'(X'X)\hat{\beta}_H = RSS_1 - RSS_0.$$

b) Write of the details for simple linear regression where we let $k = 1$ and $X_1 = (1, \ldots, 1)'$.

## Problem 6

The dataset jj contains quarterly earnings per share of the American company Johnson & Johnson for the period from first quarter 1960 to last quarter 1980. Let $y_t, t = 1, \ldots, 84$ be the log-transformed data.

a) Fit the regression model

$$y_t = \beta t + \alpha_1 D_1(t) + \alpha_2 D_2(t) + \alpha_3 D_3(t) + \alpha_4 D_4(t) + \epsilon_t$$

where $D_i(t) = 1$ if time t corresponds to quarter $i = 1, 2, 3, 4$, and zero otherwise. The $D_i(t)$'s are called dummy or indicator variables. We will assume for now that $\epsilon_t$ are i.i.d $N(0, \sigma^2)$. What is the interpretation of the parameters $\beta, \alpha_1, \alpha_2, \alpha_3$ and $\alpha_4$?

b) Explain why an intercept term cannot be included the model in (a)?

c) Fit two models where an intercept is included and where $D_1(t)$ is deleted in the first one and $D_2(t)$ is deleted in the second. What happens to the intercept?

Consider so-called *centered seasonal dummies*, $CS_i(t) = 3/4$ if time t corresponds to quarter $i = 1, 2, 3, 4$, and equal to $-1/4$ otherwise.

d) Repeat part a)-c) but with the $CS_i(t)$ instaed of the $D_i(t)$

e) What is now a reasonable interpretation of the intercept?

## Problem 7

Consider a situation with 12 observations where the covariates are two factors or categorical variables with two and three levels respectively. Let the observations be so-called lexicographically ordered, i.e. $y = (y_{111}, y_{112}, y_{121}, \ldots, y_{232})$.

a) Find the design matrix $X$, i.e. express the model in the form $Y = X\beta + \epsilon$, using the corner point paraneterization. Then

$$E[Y_{ijk}] = \mu_0 + \delta_i + \gamma_j + (\delta\gamma)_{ij}, i = 1,2\; j = 1,2,3\; k = 1,2$$

satisfying the constraints $\delta_1 = \gamma_1 = 0$ and $(\delta\gamma)_{ij} = 0$ if $i = 1$ or $j = 1$.

b) Do the same as in part a) using sum contrasts. Then

$$E[Y_{ijk}] = \alpha + \alpha_i^{(1)} + \alpha_j^{(2)} + \alpha_{ij}^{(12)}$$

satisfying $\sum \alpha_i^{(1)} = \sum \alpha_j^{(2)} = \sum_i \alpha_{ij}^{(12)} = \sum_j \alpha_{ij}^{(12)} = 0$

c) Express $\mu_0, \delta_2, \gamma_2, \gamma_3, (\delta\gamma)_{22}, (\delta\gamma)_{23}$ in terms of the $\alpha$'s and vice versa. [Hint: Here you should use R. To invert a matrix A, write solve(A). To multiply two matrices A and B, write A%*%B.]

Problem 8

We consider the linear model

$$Y = X\beta + \epsilon$$

where the $n\times(p+1)$ matrix is the design matrix. Let $x_i'$ denote the i'th row of $X$. Let $Y_{(i)}$ be the $n-1$ vector where the i'th element of $Y$ is deleted. Let $X_{(i)}$ be the $(n-1)\times(p+1)$ matrix where the i'th row of $X$ is deleted.

a) Explain why $X'X = X_{(i)}'X_{(i)} + x_i x_i'$.

b) Use part a) to show that

$$(X_{(i)}'X_{(i)})^{-1} = (X'X)^{-1} + \frac{(X'X)^{-1}x_i x_i'(X'X)^{-1}}{1-h_{ii}}$$

where $h_{ii}$ is the leverage of observation i.

c) Show that the OLS-estimator $\hat{\beta}_{(i)}$ based on $n-1$ observations, where the i'th is deleted, can be expressed as

$$\hat{\beta}_{(i)} = \hat{\beta} - \frac{(X'X)^{-1}x_i\hat{e}_i}{1-h_{ii}}$$

where $e_i$ is the residual of the i'th observation.

d) Show that

$$(\hat{\beta}_{(i)} - \hat{\beta})'(X'X)(\hat{\beta}_{(i)} - \hat{\beta}) = \frac{h_{ii}}{(1-h_{ii})^2}\hat{e}_i^2$$

and discuss the relation to Cook's distance. Also consider the standarized residual vs leverage plot implemented for models fitted by lm() in R.

Problem 9

Consider the setup in problem 7. A nested structure has the form

$$\mu_{ij} = E(Y_{ijk}) = \mu + \alpha_i + \beta_{ij}, i = 1, 2, j = 1, 2, 3, k = 1, 2$$

a) Discuss situations where this is a sensible parameterization.
b) Formulate restrictions so that such models are well specified.
c) What are the corresponding design matrices?

Problem 10

Do Problem 12.5.20 in John. Rice (2007): Mathematical Statistics and Data Analysis, thrd. ed, Duxbury Press. The data can be found on the course home page as ''dye''.

In addition:

c) Find a $90\%$ confidence interval for $\sigma_\epsilon^2$ and $\sigma_A^2/\sigma_\epsilon^2$ .

Problem 11
Do Problem 2.1 on page 44-45 in Fitzmaurice, Laird and Ware. The data set can be found on the course home page as lead.

Problem 12

Consider the situations where the observations $Y_i, i = \ldots, N$ are independent $n$-dimensional vectors where

$$Y_i = X_i\beta + \epsilon_i$$

with $\epsilon_i \sim MN(0, \Sigma)$.

Let $W_i, i = 1, \ldots, N$ be $n \times n$ symmetric, positive definite matrices.

a) Show that the estimator of $\beta$ minimizing $\sum_{i=1}^N (Y_i - X_i\beta)' W_i (Y_i - X_i\beta)$ has the form

$$\hat\beta_W = (\sum_{i=1}^N X_i' W_i X_i)^{-1} \sum_{i=1}^N X_i' W_i Y_i$$

b) Show that $E[\hat\beta_W] = \beta$ and that the covariance of $\hat\beta_W$, $E[(\hat\beta_W - \beta)(\hat\beta_W - \beta)'] = (\sum_{i=1}^N X_i' W_i X_i)^{-1} \sum_{i=1}^N X_i' W_i \Sigma W_i X_i (\sum_{i=1}^N X_i' W_i X_i)^{-1}$.
c) What is the distribution of $\hat\beta_W$?

Problem 13

Consider the situations where the observations $Y_i, i = \ldots, N$ are independent $n$-dimensional vectors where

$$Y_i = X_i\beta + \epsilon_i$$

with $\epsilon_i \sim MN(0, \Sigma)$.

This can be considered as a system of $n$ regression equations, where the expected response in equation $j, j = 1, \ldots, n$ is $\mu_{ij} = \beta_0 + x_{ij1}\beta_1 + \cdots + \beta_p x_{ijp}$. Here we include a constant term so the design matrices have $p+1$ columns.

Let $\hat{e}_{ij}, i = 1, \ldots, N$ be the residuals from the $j$'th regression, let

$$\hat{\sigma}_{j_1 j_2} = \frac{1}{N-(p+1)} \sum_{i=|}^{N} \hat{e}_{ij_1}\hat{e}_{ij_2}$$

and let the estimator for the covarince matrix $\Sigma$ be $\hat{\Sigma} = \{\hat{\sigma}_{j_1 j_2}\}$.

a) Find $E[\hat{\Sigma}]$.

b) Let $\hat{\beta}_W$ be the weighted least squares estimator using $\hat{\Sigma}^{-1}$ as weighting matrix. When is the distribution of $\hat{\beta}_W$ approximately multivariate normal for $N$ large? What are the expectation and covariance of the approximate distribution?

c) Compare the estimated covariance matrix in FLW, page 116, with the $\hat{\Sigma}$ as defined above. The data set can be found on the textbook home page.

Problem 14

Let $X$ be a $n \times p$ matrix of rank $p$.

a) If $A = I - X(X'X)^{-1}X'$, use the spectral theorem to show that there is a $Nm \times (n-p)$ matrix $B$ such that $B'B = I$ and $BB' = A$.[Hint: Show that $A$ is idempotent, i.e. $A^2 = A$. Then the eigenvalues of $A$ must be equal to 1 or 0.]

b) Show that $B'X = 0$

c) Show that for a $n \times n$ non-singular matrix $\Sigma^{-1} - \Sigma^{-1}X(X'\Sigma^{-1}X)^{-1}X'\Sigma^{-1} = B(B'\Sigma B)^{-1}B'$.

Problem 15

Consider the situation where the observations $Y_i, i = \ldots, N$ are independent $n$-dimensional vectors

$$Y_i = X_i\beta + \epsilon_i$$

with $\epsilon_i \sim MN(0, \Sigma_i)$. The covariance matrices $\Sigma_i, i = 1; \ldots, N$ can be different and will generally be of the form $\Sigma_i = \Sigma_i(\theta)$ where $\theta$ denotes one or more unknown parameters.

We shall consider how the restricted maximum likelihood (REML) estimators for $\theta$ can be derived. Remember that the main idea behind the REML is to base the estimators of $\theta$ on a set of linear combinations of the original observations.

For notational convenience we introduce the stacked $Nn \times 1$ vector $Y = (Y_1', \ldots, Y_N')'$, the $Nn \times p$ matrix$X = (X_1', \ldots, X_N')'$ and the $Nn \times Nn$ matrix $\Sigma$ where the diagonal blocks equal $\Sigma_1, \ldots, \Sigma_N$, and the others are equal to 0. Thus $Y \sim MN(X\beta, \Sigma)$.

Remember that for fixed $\theta$ the generalized least squares estimator has the form $\hat{\beta}(\theta) = (X'\Sigma(\theta)^{-1}X)^{-1}X'\Sigma(\theta)^{-1}Y = GY$ where $G$ is a $p \times Nn$ matrix.

a) Define $Z = B'Y$, where the matrix $B$ satisfies $B'B = I$ and $BB' = I - X(X'X)^{-1}X'$. Show that $E(Z) = 0$. What is $E(ZZ')$?

b) Why is the distribution of $Z$ multinormal?

c) Show that $Z$ and $\hat{\beta}(\theta)$ are independent.

We now want to use the distribution of $Z$ for the estimation of $\theta$.

d) Let $L = (G', B)$ and show that $det(B'\Sigma B) = det(L'L)det(\Sigma)det(X'\Sigma^{-1}X)$

e) Use part c) of Problem 14 to show that $Z'(B\Sigma B')^{-1}Z = Y'\Sigma^{-1}Y - \hat{\beta}(\theta)'X'\Sigma^{-1}X\hat{\beta}(\theta)$.

f) Conclude that the logarithm of the part of a likelihood based on $Z$ may be written as

$$-\frac{1}{2}\log(\Sigma) - \frac{1}{2}\log(X'\Sigma^{-1}X) - \frac{1}{2}(Y - X\hat{\beta}(\theta))'\Sigma^{-1}(Y - X\hat{\beta}(\theta))$$

Problem 16

Consider the "Treatment of Lead-Exposed children" case discussed in section 5.4 in the textbook by Fitzmaurice, Laird and Ware. The model is

$$Y_i = X_i\beta + e_i, i = 1, \ldots, 100,$$

where $Y_i$ and $e_i$ are 4-dimensional vectors. The error terms $e_i, i = 1, \ldots, 100$ are independent multivariate normal, $e_i \sim MN(0, \Sigma)$ where $\Sigma$ is an unstructured covariance matrix. In Table 5.5 estimates of the regression coefficients can be found.

a) Explain how the design matrices $X_i$ must be to produce the estimates in Table 5.5.

b) What are the estimates of $\mu_{ij} = E(Y_{ij}), j = 1, \ldots, 4, i = 1, \ldots, 100$.

c) Now, consider design matrices of the form

$$X_i^1 = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 \end{pmatrix}$$

if the child belongs to the placebo group, and

$$X_i^1 = \begin{pmatrix} 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 1 & 1 & 0 & 0 \\ 1 & 1 & 1 & 0 & 1 & 1 & 1 & 0 \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \end{pmatrix}$$

otherwise. Find the estimates of $\gamma$ in the model

$$Y_i = X_i^1\gamma + e_i, i = 1, \ldots, 100.$$

d) What are the interpretations of the coefficients in $\gamma$?

e) As explained in the textbook, the covariance matrix of the estimators $\hat{\beta}$ can be estimated as $\hat{C} = (\sum_{i=1}^{100} X_i'\hat{\Sigma}^{-1}X_i)^{-1}$. How can the covariance matrix of the estimators of $\gamma$ be estimated using $\hat{C}$?

Problem 17

Do problem 5.1 in FLW using the R-procedure gls. The data sets can be found on the textbook home page or the course webpage as ''cholesterol''. For R it is best to indicate missing observations with NA.

Problem 18

Consider a situation where $N$ individuals or units are observed at the same $n$ occasions. A possible model for the measurements $y_{ij}, j = 1, \ldots, n, i = 1, \ldots, N$ is

$$Y_i = \beta_1 1_n + \beta_2 t + b_i 1_n + \epsilon_i, i = 1, \ldots, N$$

where $b_i, i = 1, \ldots, N$ and the elements of $\epsilon_i, i = 1, \ldots, N$ are assumed to be independent, normally distributed random variables with expectation $0$ and an variance $\tau^2$ and $\sigma^2$ respectively. The n-dimensional vector where all elements are equal to $1$ is denoted by $1_n$ and $t$ is the vector where the elements are $t_1, \ldots, t_n$.

a) Let $\Sigma$ be the covariance matrix of $Y_i$. Verify that $\Sigma^{-1}$ is the matrix where all the diagonal elements equal $\frac{\sigma^2+(n-1)\tau^2}{\sigma^2(\sigma^2+n\tau^2)}$ and all the off-diagonal elements equal $\frac{-\tau^2}{\sigma^2(\sigma^2+n\tau^2)}$.

b) Find the empirical BLUP , $\hat{b}_i$, for the random effect $b_i$.

c) Explain why the within-individual variation can be measured by $\sigma^2$ and the between-individual variation can be measured by $\tau^2$. Write the empirical BLUP $\hat{b}_i$ on a form where it is easy to see the effect of $\sigma^2$ and $\tau^2$.

d) What happens when $\hat{\sigma}^2$ is large compared to $\hat{\tau}^2$?

e) What happens when $\hat{\tau}^2$ is large compared to $\hat{\sigma}^2$?

Problem 19

The data set Sitka in the R-library MASS contains observations on 79 Sitka spruce trees which are divided into two groups. The first group consists of 54 trees which were grown in an ozone enriched environment, and the other group consists of 25 trees which were controls and grown under natural conditions. Ozone pollution is common in urban areas, and the impact of increased ozone concentrations on tree growth is of considerable interest.

The size of the trees were measured at five occasions roughly at monthly intervals. The time is given in days since 1 January 1988 and is indicated as $t_{ij}, j = 1, \ldots, 5, i = 1, \ldots, 79$.

In this problem we will analyze the size, over the observation period, of the 79 trees, and compare the two groups.

a) Compute the mean values of the sizes of the trees in each group and plot the means in the two groups as a function of time. Describe the main features of the plot.

Consider a model of the form

$$Y_i = X_i \beta + e_i, i = 1, \ldots, 79$$

where the error terms $e_i, i = 1, \ldots, 79$ are assumed to be independent, multivariate normally distributed with expectation $0$ and an unstructured covariance matrix $\Sigma$. The observations of the control group at the first occasion are used as reference group.

b) What must the design matrices $X_i$ look like for estimating the response profiles? Explain how $\Sigma$ can be estimated. Find estimators for $\beta$ and $\Sigma$ and compute the estimated expected response profiles.

c) Estimate two models: in the first model $E[Y_{ij}] = \beta_1 + \beta_2 \times t_{ij} + \beta_3 \times grp_i + \beta_4 \times grp_i \times t_{ij}$, where $grp_i$ is equal to $0$ if the tree belong to the control group and equal to $1$ if it is grown in an ozone enriched environment, in the second model quadratic terms are added, i.e, $E[Y_{ij}] = \beta_1 + \beta_2 \times t_{ij} + \beta_3 \times grp_i + \beta_4 \times grp_i \times t_{ij} + \beta_5 \times t_{ij}^2 + \beta_6 \times grp_i \times t_{ij}^2$. What are the estimated standard errors of $\beta_5$ and $\beta_6$?

d) Plot the estimated expected values for the model containing the quadratic terms, i.e. estimates of the expected response profiles.

e) Test whether the quadratic terms are significant. What is the p-value?

In the rest of the problem we consider a linear mixed model of the form

$$Y_i = X_i\beta + Z_i b_i + \epsilon_i, i = 1, \ldots, 79$$

where $b_i, i = 1, \ldots, 79$ and $\epsilon_i, i = 1, \ldots, 79$ are assumed to be independent, multivariate normally distributed random vectors with expectation $0$ and an unstructured covariance matrix $G$ and $\sigma^2 I_5$ respectively.

f) Fit a model where $E[Y_{ij}|b_i] = \beta_1 + \beta_2 \times t_{ij} + \beta_3 \times grp_i + \beta_4 \times grp_i \times t_{ij} + \beta_5 \times t_{ij}^2 + \beta_6 \times grp_i \times t_{ij}^2 + b_{1i} + b_{2i} \times t_{ij}$.

g) Test whether the quadratic terms are significant using a Wald and a likelihood ratio test. What are the p-values?

h) Compute the estimated covariance matrix of the observations in this model and compare the estimate to what you found in part b).

i) Plot the residuals against the fitted values. Comment on the result.

[ R-hint: To get hold of the data:

```
library(MASS)
attach(Sitka)
```

In part f) you need the R-function lme. To get hold of it use:

```
library(nlme)
```

Problem 20

Situations where models involving multiple components of variation naturally arise are in the so-called *split-plot designs*. Originating in agricultural field trials, they have proved useful also in many other context. In the textbook by Fitzmaurice et al. the relevance for biological and medical studies is discussed in several places, as you can see from the index at the end of the book. In this problem we will consider an industrial application.

The typical feature of experiment of this kind is that there are I blocks available consisting of J *whole plots* each. To each whole plot within a block a different whole plot treatment is applied. Then, each whole plot is divided into K *subplots* with a different subplot treatment to each subplot.

The table below [1] shows the outcome of an industrial experiment to investigate the corrosion resistance in steel bars with four different coatings, $C_1, C_2, C_3, C_4$ produced at furnace temperatures 360, 370 and 380 centigrades. There are two important sources of variation: the temperature at which the measurements were taken and the position of the steel bars within the furnace, i.e how close the bars were placed to the center of the furnace where the temperature is highest. Here the whole plots were the six furnace heats at which the measurements were taken, and the subplots the four pour positions in the furnace at which the steel bars with coatings $C_1, C_2, C_3, C_4$ could be placed. These were randomly allocated. Thus, there are two associated variances: $\sigma_W^2$ for the whole plots describing the variation from one heat to another and the subplot variance $\sigma_S^2$ measuring variation from position to position of the steel bars within the same furnace heat.

TABLE 1. Corrosion resistance of steel bars treated with four different coatings and produced at three different temperatures.

| | $C_1$ | $C_2$ | $C_3$ | $C_4$ |
|---|---|---|---|---|
| 360 | 67 | 73 | 83 | 89 |
| | 33 | 8 | 46 | 54 |
| 370 | 65 | 91 | 87 | 86 |
| | 140 | 142 | 121 | 150 |
| 380 | 155 | 127 | 147 | 212 |
| | 108 | 100 | 90 | 153 |

With $IJ = 2 \times 3$ whole plots and $K = 4$ subplots a possible model for these measurements $y_{ijk}, i = 1, \ldots, I, j = 1, \ldots, J, k = 1, \ldots, K$ is

$$Y_{ijk} = \mu + \beta_j + \kappa_{ij} + \gamma_k + (\gamma\beta)_{jk} + \epsilon_{ijk}, i = 1, \ldots, I, j = 1, \ldots, J, k = 1, \ldots, K$$

with constraints $\sum_{j=1}^{J} \beta_j = 0, \sum_{k=1}^{K} \gamma_k = 0$ and $\sum_{j=1}^{J} (\gamma\beta)_{jk} = \sum_{k=1}^{K} (\gamma\beta)_{jk} = 0$.

The random variables $\kappa_{ij}, i = 1, \ldots, I, j = 1, \ldots J$ are assumed to be independent and identically distributed $N(0, \sigma_W^2)$ and independent of $\epsilon_{ijk}, i = 1, \ldots, I, j = 1, \ldots J, k = 1, \ldots, K$ which are independent and identically distributed $N(0, \sigma_S^2)$.

a) Indicate what the numerical values of the estimates for $\mu, \beta_j, \gamma_k, (\gamma\beta)_{jk}, j = 1, \ldots, J, k = 1, \ldots, K$ and the variances $\sigma_W^2$ and $\sigma_S^2$ are. Comment on the size of the estimated variances. Here you can use the R-procedure lme() to find the estimates.

We will now consider how the estimators for the variances $\sigma_W^2$ and $\sigma_S^2$ can be explicitly found by considering the appropriate sum of squares. We indicate sums over an index with a " · ", e.g $Y_{ij\cdot} = \sum_{k=1}^{K} Y_{ijk}$, and the corresponding averages as $\bar{Y}_{ij\cdot} = \frac{1}{K} \sum_{k=1}^{K} Y_{ijk}$.

[1] The data can be found on `furnace.txt` on the course web page

b) Find the expected value of

$$SSERR_1 = K \sum_{i=1,j=1}^{IJ} (\bar{Y}_{ij\cdot} - \bar{Y}_{i\cdot\cdot} - \bar{Y}_{\cdot j\cdot} + \bar{Y}_{\cdot\cdot\cdot})^2$$

c) Explain how another unbiased estimate of the expected value from part b), $E[SSERR_1]$, can be found from a one-way layout analysis of variance using the sums

$$\bar{y}_{ij\cdot}, i = 1, \ldots, I, j = 1, \ldots, J$$

d) Compute the numerical estimate from the previous part, and compare it with the estimate that can be deduced from the fitted model in part a).

e) Find the expected value of

$$SSERR_2 = \sum_{i=1,j=1,k=1}^{IJK} (Y_{ijk} - \bar{Y}_{ij\cdot} - \bar{Y}_{\cdot jk} + \bar{Y}_{\cdot j\cdot})^2$$

f) Use the result from part b) and e) to find unbiased estimators for the variances $\sigma_W^2$ and $\sigma_S^2$ based on $SSERR_1$ and $SSERR_2$.

g) Consider the sum of squares $SSERR_1$ and $SSERR_2$. Discuss the properties that are needed, in addition to what you have already found, for constructing 95% confidence intervals for $\sigma_S^2$ and $\sigma_W^2/\sigma_S$ based on the $\chi^2$ and Fisher distributions. What would such intervals look like? [ NB. In part f) and g) you are not asked to do any numerical computations.]

Problem 21

To study the effect of a drug on epileptic seizures the following experiment is conducted. The study group is divided into two groups of size $m_1$ and $m_2$. One group is chosen at random and the members receive the drug. The members of the other group are given a placebo. For each participant the number of epileptic seizures in a number of months is recorded, as $y_{ij}, i = 1, \ldots, m_1 + m_2, j = 1, \ldots, n_i$.

Consider the model where the observations are considered as realizations of random variables, $Y_{ij}$ where

$$\begin{aligned} Y_{ij}|u_i \text{ are independent Poisson}, &\quad i = 1, i = \ldots, m_1 + m_2, j = 1, \ldots, n_i, \\ E[Y_{ij}|u_i] = exp(\beta_0 + u_i), &\quad i = 1, \ldots, m_1, j = 1, \ldots, n_i \\ E[Y_{ij}|u_i] = exp(\beta_0 + \beta_1 + u_i), &\quad i = m_1 + 1, \ldots, m_1 + m_2, j = 1, \ldots, n_i, \\ u_i \sim i.i.d. N(0, \sigma^2) & \end{aligned}$$

a) Find an expression for the logarithm of the likelihood.

b) Discuss how the likelihood can be maximized and what problems one faces.

Problem 22

Assume that $Y_1,\dots,Y_n$ are independent $N(\mu,\sigma^2)$ distributed.

Show that the maximum likelihood estimator for $\sigma^2$ based on the (transformed) variables $V_1 = Y_1 - \bar{Y},\dots,V_{n-1} = Y_{n-1} - \bar{Y}$ is $\sum_{i=1}^{n}(Y_i - \bar{Y})^2/(n-1)$ [ Hint: Express the covariance matrix of $V_1,\dots,V_{n-1}$ as $\sigma^2(I_{n-1} - \frac{1}{n}\underline{1}_{n-1}\underline{1}_{n-1})$ and use that the determinant of this matrix can be found from the formula $\det(A + \underline{b}\underline{b}') = \det(A)(1 + \underline{b}'A^{-1}\underline{b})$.]