TMA4267 – Linear Statistical Models
Exercise 4 – V2015

Last changes: Reformulated question 4b for clarity.

Problem 1: F test and partial F test in multiple linear regression

Consider a multiple linear regression model $\boldsymbol{Y} = X\boldsymbol{\beta} + \boldsymbol{\epsilon}$, with $\boldsymbol{\beta} = [\beta_1 \ \cdots \ \beta_p]$ and $\boldsymbol{\epsilon} \sim N(\mathbf{0}, \sigma^2 I)$. We want to test the null hypothesis $\beta_{r+1} = \beta_{r+2} = \cdots = \beta_p = 0$. In effect, we want to compare the full model, involving p covariates and p parameters, with a reduced model, involving only r covariates and r parameters. Let X_0 be the design matrix consisting only of the first r columns of X. Let $H = X(X^TX)^{-1}X^T$ and $H_0 = X_0(X_0^TX_0)^{-1}X_0^T$ be the projection matrices of the full and the reduced model, respectively.

a) Show that in the partition $I = (I - H) + (H - H_0) + H_0$, all three terms on the right are idempotent. (Hint for $H - H_0$: View H and H_0 as projections onto the column space of X and X_0, respectively, to see that $HH_0 = H_0$ and that $H_0(I - H) = O$.)

b) Show that under the null hypothesis, $(I - H)(\boldsymbol{y} - X\boldsymbol{\beta}) = (I - H)\boldsymbol{y}$ and that $(H - H_0)(\boldsymbol{y} - X\boldsymbol{\beta}) = (H - H_0)\boldsymbol{y}$.

c) What are the ranks of $I - H$ and $H - H_0$ (assuming full rank of X)? Show that

$$F = \frac{\boldsymbol{Y}^T(H - H_0)\boldsymbol{Y}/(p - r)}{\boldsymbol{Y}^T(I - H)\boldsymbol{Y}/(n - p)} \sim F(p - r, n - p)$$

under the null hypothesis.

$\boldsymbol{y}^T(I - H)\boldsymbol{y}$ is the error sum of squares. $\boldsymbol{y}^T(H - H_0)\boldsymbol{y}$ is sometimes called a *sequential sum of squares*. A large sequential sum of squares compared to the error sum of squares, thus a large F, indicates that the null hypothesis is false.

d) Use the acid rain data (`http://www.math.ntnu.no/~bakke/TMA4267/2015V/acidrain2.r`) to test whether all coefficients except the intercept and the coefficients of `x1` and `x3` are zero.

e) Show, that if the model includes an intercept, and the null hypothesis is that all other parameters are zero, the projection matrix $H_0 = \frac{1}{n}\mathbf{1}\mathbf{1}^T$, so that the sequential sum of squares is what is called (not by B/F) the regression sum of squares. For the acid rain data, test whether all coefficients except the intercept are zero.

Problem 2: Inference about a new observation in multiple linear regression

Consider again the multiple linear regression model $\boldsymbol{Y} = X\boldsymbol{\beta} + \boldsymbol{\epsilon}$ with $\boldsymbol{\epsilon} \sim N(\mathbf{0}, \sigma^2 I)$. Assume that $Y_0 = \boldsymbol{X}_0^T\boldsymbol{\beta} + \epsilon_0$ is a new observation, with ϵ_0 independent of $\boldsymbol{\epsilon}$.

a) Show that $\mathbf{X_0}\hat{\boldsymbol{\beta}}$ is an unbiased estimator of EY_0, where $\hat{\boldsymbol{\beta}}$ is the least-square estimator of β. Find the distribution of the estimator.

b) Find a $100(1-\alpha)$ confidence interval for EY_0.

c) Find a $100(1-\alpha)$ prediction interval for Y_0, that is, an interval that will contain Y_0 with probability $1-\alpha$.

d) Use the acid rain data (`http://www.math.ntnu.no/~bakke/TMA4267/2015V/acidrain2.r`) to find a confidence interval for the expected value of a new observation having covariates $(\texttt{x1}, \ldots, \texttt{x7}) = (1, 3, 50, 1, 50, 2, 1, 0)$. Also find a prediction interval for such a new observation.

e) From the theory of simple linear regression, you know that the bounds of the confidence interval are

$$\hat{y}_0 \pm t_{\alpha/2}\sqrt{\frac{\text{SSE}}{n-2}\left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2)}{\sum_i (x_i - \bar{x})^2}\right)},$$

where $\hat{y}_0$ is the estimator of EY_0, $-t_{\alpha/2}$ the $\alpha/2$-quantile of a $t(n-2)$ variable, n the number of observations, x_i the covariates and x_0 the new covariate. Show that this is the same confidence interval as found above.

Problem 3: The square root matrix and the Mahalanobis transform

Let the expectation (mean) and covariance matrix for a p-variate random vector $\boldsymbol{X}$ be $\boldsymbol{\mu} = \mathrm{E}(\boldsymbol{X})$ and $\boldsymbol{\Sigma} = \mathrm{Cov}(\boldsymbol{X})$.

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

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

and $\boldsymbol{\Lambda}$ be a diagonal matrix with the eigenvalues $\lambda_1, \lambda_1, ..., \lambda_p$ on the diagonal.

a) Show that if $\boldsymbol{\Sigma}$ is symmetric and positive definite then all eigenvalues of $\boldsymbol{\Sigma}$ are positive. (Hint: a symmetric matrix $\boldsymbol{A}$ is positive definite if $\boldsymbol{z}^T\boldsymbol{A}\boldsymbol{z} > 0$ for all vectors $\boldsymbol{z} \neq \mathbf{0}$).

What can you say about the eigenvalues and eigenvectors to the inverse matrix of $\boldsymbol{\Sigma}$? Justify the answer.

b) Define the matrices $\boldsymbol{\Sigma}^{\frac{1}{2}}$ and $\boldsymbol{\Sigma}^{-\frac{1}{2}}$ by

$$\begin{aligned} \boldsymbol{\Sigma}^{\frac{1}{2}} &= \boldsymbol{P}\boldsymbol{\Lambda}^{\frac{1}{2}}\boldsymbol{P}^T \\ \boldsymbol{\Sigma}^{-\frac{1}{2}} &= \boldsymbol{P}\boldsymbol{\Lambda}^{-\frac{1}{2}}\boldsymbol{P}^T \end{aligned}$$

Show that both matrices are symmetric and that the following is true:

$$\begin{aligned} \boldsymbol{\Sigma}^{\frac{1}{2}}\boldsymbol{\Sigma}^{\frac{1}{2}} &= \boldsymbol{\Sigma} \\ \boldsymbol{\Sigma}^{-\frac{1}{2}}\boldsymbol{\Sigma}^{-\frac{1}{2}} &= \boldsymbol{\Sigma}^{-1} \\ \boldsymbol{\Sigma}^{\frac{1}{2}}\boldsymbol{\Sigma}^{-\frac{1}{2}} &= \boldsymbol{I} \end{aligned}$$

where $\boldsymbol{I}$ is the identity matrix.

c) The transform

$$\boldsymbol{Y} = \boldsymbol{\Sigma}^{-\frac{1}{2}}(\boldsymbol{X} - \boldsymbol{\mu})$$

is called the Mahalanobis transform. Show that $\mathrm{E}(\boldsymbol{Y}) = 0$ and $\mathrm{Cov}(\boldsymbol{Y}) = \boldsymbol{I}$.

d) Assume now that $\boldsymbol{X} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ where $\boldsymbol{\Sigma}$ is symmetric and positive definite. Derive the distribution of

$$(\boldsymbol{X} - \boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{X} - \boldsymbol{\mu})$$

e) Write down an expression for the contours of the pdf of $\boldsymbol{X}$. What is the graphical intepretation of these contours?

f) Let $\boldsymbol{X} \sim N_2(\boldsymbol{0}, \boldsymbol{\Sigma})$, where

$$\boldsymbol{\Sigma} = \begin{bmatrix} 1 & 0.8 \\ 0.8 & 1 \end{bmatrix}$$

Draw contours for which 90% of the probability mass lies within the contours.

The following printout from R may be of help.

```
> Sigma <- matrix(c(1,0.8,0.8,1),2,2)
> eigen(Sigma)
$values
[1] 1.8 0.2

$vectors
          [,1]       [,2]
[1,] 0.7071068  0.7071068
[2,] 0.7071068 -0.7071068

> sqrt(qchisq(0.1,2,lower.tail=F))
[1] 2.145966
```

Problem 4: Sampling multivariate normal data

We will look at two different ways of simulating data sets from the multivariate normal distribution (you will learn more about simulation techniques in TMA4300).

Using a library:

We will first use the function `mvrnorm` that is available from the library `MASS`.

Start R. To load the library `MASS` you write `library(MASS)`

If a library is not installed at the system, you may either request it, or install it *locally*; do

```
install.packages("ellipse")
```

and then use

```
library("ellipse")
```

to load it.

Let

$$\mathbf{Y} = A\mathbf{X} \tag{1}$$

where

$$\mathbf{X} \sim N(0, I) \tag{2}$$

$$I = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \tag{3}$$

$$A = \begin{bmatrix} 1 & a \\ a & 1 \end{bmatrix} \tag{4}$$

a) Find $\boldsymbol{\mu} = \mathrm{E}(\mathbf{Y})$, $\Sigma = \mathrm{Cov}(\mathbf{Y})$ and $\mathrm{Corr}(\mathbf{Y})$.

b) Simulating with `mvrnorm` Code for doing this (with one specific choice of $\boldsymbol{\mu}$=`meanvec` and Σ=`covarmat`):

```
binorm.sample <- mvrnorm(1000,meanvec,covarmat)
x <- binorm.sample[,1]
y <- binorm.sample[,2]
plot(x,y)
```

Choose different values for a (e.g. $a = \{-0.9, -1/4, 0, 1/4, 0.9\}$, and what happens when $a \to \pm\infty$?), and simulate 1000 observations from the distribution of $\mathbf{Y}$ and plot the observations.

Estimate the mean vector, covariance matrix and the correlation matrix of your binormal sample.
Command: `mean, cov, cor`.

Also make boxplots, histograms and normal-qq-plot for for each variable. What does a normal-qq-plot show?
Command: `plot, boxplot, hist, qqplot, qqline`.

c) Plotting with `dmvnorm` We will first look at the function `dmvnorm` that is available from the library `mvtnorm`. To load the library you write `library(mvtnorm)`.

What are the inputs to the function (scalar, vector, matrix?) and the output from the function?

d) Plotting with `persp` We would like to plot the binormal density in a 3D-plot. Try the following commands:

```
x <- seq(-5,5,length=100)
y <- seq(-5,5,length=100)
xymat <- expand.grid(x,y)
z <- matrix(dmvnorm(xymat,mean=meanvec,sigma=covarmat),ncol=100)
persp(x,y,z,theta=10,phi=30,col="lightblue",shade=0.3,expand=0.8)
```

Use different values for a. Try out different choices of angles `theta` and `phi` in the `persp`-function.

3D plot can also be done with the `rgl` library:

```
library(rgl)
zlim = range(z)
ztemp = z/zlim[2]*10
zlen = zlim[2] - zlim[1] + 1
colorlut = heat.colors(100) # height color lookup table
col = colorlut[ ztemp*10-zlim[1]+1 ]
open3d()
surface3d(x, y, ztemp, color=col,back="lines")
```

e) Plotting with the `ellipse` library (this has also been done on Exercise 1, so you may already know this - new thing - the connection between the contours and a probability distribution.) We will now plot contours with equal density in the binormal distribution using the `library(ellipse)`.

Start by plotting the data you simulated in b) with `plot(x,y)`. An equal density ellipse can be added to the current plot using the command

```
lines(ellipse(covarmat,centre=meanvec))
```

The argument `level` can be given to `ellipse`, giving the ellipse where the probability of being inside the ellipse equals this argument (default is 0.95). Plot an ellipse of probability 0.95.

Then we will examine the contour plotted. We know that the ellipse is centered at the mean (which here is at `meanvec`), and has axes $\pm c\sqrt{\lambda_i}\boldsymbol{e}_i$, where $(\lambda_i, \boldsymbol{e}_i)$ is an egenvalue, eigenvector pair of the covariance matrix ($i = 1, ..., p$), and $c^2 = \chi^2_{0.05}$ (since level 0.95 was chosen). make a vector...)

Find the eigenvalues and eigenvectors of the covariance matrix. Plot the axes of the ellipse, and check that this is in correspondence with the ellipse plotted. Also check that the ellipse axes half length is correct.

```
eig = eigen(covarmat)
const = sqrt(qchisq(0.95,2))
for (i in 1:2){
     lambda = eig$values[i]
     e = eig$vectors[,i]
     pkt = const*sqrt(lambda)*e
     points(c(-pkt[1],pkt[1]),c(-pkt[2],pkt[2]),col=3,pch=20)
     lines(c(-pkt[1],pkt[1]),c(-pkt[2],pkt[2]),lwd=2,col=3)
}
```

See what happens when you change the covariance matrix (it may be easier to relate to the standard deviations and the correlation). Try covariance matrices giving negative correlations between the two variables.