TMA4267 - Linear Statistical Models
Exercise 1 - V2014

Problem 1: Covariance and correlation

We consider two random variables (RVs), X and Y, with expected values μ_X and μ_Y, respectively. We will study the covariance, defined as

$$\operatorname{Cov}(X,Y) = \mathrm{E}[(X-\mu_X)(Y-\mu_Y)].$$

Let further σ_X^2 and σ_Y^2 be the variances of X and Y. The correlation is

$$\operatorname{Corr}(X,Y) = \frac{\operatorname{Cov}(X,Y)}{\sigma_X\sigma_Y}.$$

Let a, b, c and d be constants.

a) Verify that

$$\operatorname{Cov}(aX+b, cY+d) = ac\operatorname{Cov}(X,Y)$$

b) Verify that

$$\operatorname{Corr}(aX+b, cY+d) = \operatorname{Corr}(X,Y)$$

c) Now we look at two more RVs, Z and V. Find an easy to use formula for $\operatorname{Cov}(aX+bY, cZ+dV)$.

d) Prove the following relationship

$$\operatorname{Cov}(X,Y)^2 \leq \operatorname{Var}(X)\operatorname{Var}(Y)$$

Hint: Look at $0 \leq \operatorname{Cov}(Z,Z)$ where and let $Z = X - Y \cdot \operatorname{Cov}(X,Y)/\operatorname{Var}(Y)$.

Problem 2: The bivariate normal distribution

Let X and Y be random variables with pdf $f(x,y)$ parameterized by $(\mu_X, \mu_Y, \rho, \sigma_X^2, \sigma_Y^2)$.

$$\begin{aligned}
f(x,y) &= ce^{-\frac{1}{2}Q(x,y)} \\
c &= \frac{1}{2\pi\sigma_X\sigma_Y\sqrt{1-\rho^2}} \\
Q(x,y) &= \frac{1}{(1-\rho^2)}[(\frac{x-\mu_X}{\sigma_X})^2 + (\frac{y-\mu_Y}{\sigma_Y})^2 - 2\rho(\frac{x-\mu_X}{\sigma_X})(\frac{y-\mu_Y}{\sigma_Y})]
\end{aligned}$$

a) We will study the quadratic form $Q(x, y)$. Show that $Q(x, y)$ can be written as

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

where

$$\boldsymbol{x} = \begin{bmatrix} x \\ y \end{bmatrix}$$
$$\boldsymbol{\mu} = \begin{bmatrix} \mu_X \\ \mu_Y \end{bmatrix}$$
$$\boldsymbol{\Sigma} = \begin{bmatrix} \text{Var}(X) & \text{Cov}(X,Y) \\ \text{Cov}(Y,X) & \text{Var}(Y) \end{bmatrix} = \begin{bmatrix} \sigma_X^2 & \rho\sigma_X\sigma_Y \\ \rho\sigma_X\sigma_Y & \sigma_Y^2 \end{bmatrix}$$

Remark: $\boldsymbol{\Sigma}$ is called the variance-covariance matrix of $\boldsymbol{X}$.

b) Now we focus on rewriting the pdf $f(x, y)$ using vectors and matrices. From a) we saw that $det(\boldsymbol{\Sigma}) = \sigma_X^2\sigma_Y^2(1 - \rho^2)$. Use this to rewrite the pdf $f(x, y)$ in a vector-matrix form, that is as a function of $\boldsymbol{x}$, with parameters $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Call the new pdf $f(\boldsymbol{x})$.

c) Now we turn to look at contours of $f(\boldsymbol{x})$, that is, the graphical form (in $\boldsymbol{x}$) given for constant values of $f(\boldsymbol{x})$.

Why can the contours be seen as solutions to the equation

$$(\boldsymbol{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{x} - \boldsymbol{\mu}) = d^2$$

for a given constant d^2?

Use the method of diagonalization (spectral decomposition) to explain that the contours are ellipses with center in $\boldsymbol{\mu}$, axes in the direction of the eigenvectors of $\boldsymbol{\Sigma}$, with halflengths $\sqrt{\lambda_1}d$ and $\sqrt{\lambda_2}d$, where λ_1 and λ_2 are the eigenvalues of $\boldsymbol{\Sigma}$.

Remark: remember that there is a simple connection between the eigenvectors and eigenvalues of $\boldsymbol{\Sigma}$ and $\boldsymbol{\Sigma}^{-1}$.

d) For the special case that $\sigma_X = \sigma_Y$ find the eigenvalues and eigenvectors of $\boldsymbol{\Sigma}$. Make a drawing of the contours by hand.

e) Now we turn to R to draw ellipses. This can be done using

```
library(ellipse)
```

First look at $\mu_X = \mu_Y = 1$, and $\sigma_X = 1$, $\sigma_y = 1$ and $\rho = 0.5$.

```
mu1=mu2=1
sigma1=1
sigma2=1
rho=0.5
plot(ellipse(rho, scale=c(sigma1, sigma2),centre=c(mu1,mu2)),type="l")
```

Try varying the parameters in $\boldsymbol{\Sigma}$ and observe.

Problem 3: Simple linear regression, basic theory and R

Suppose that we are given a sequence $(x_1, y_1), \ldots, (x_n, y_n)$ that is described by the model:

$$y_i = \beta_0 + \beta_1 x_i + e_i, \qquad i = 1, \ldots, n, \tag{1}$$

and $e_1, \ldots, e_n$ are iid $N(0, \sigma^2)$.

The likelihood is

$$L = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2\sigma^2}\left(y_i - (\beta_0 + \beta_1 x_i)\right)^2\right\} \tag{2}$$

and the log-likelihood is

$$\begin{aligned} l &= \log L && (3)\\ &= -\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}\left(y_i - (\beta_0 + \beta_1 x_i)\right)^2 && (4) \end{aligned}$$

a) Show that the maximum likelihood estimates of β_0 and β_1 are:

$$\begin{aligned} \hat{\beta}_1 &= \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} && (5)\\ &= \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2} && (6)\\ \hat{\beta}_0 &= \bar{y} - \hat{\beta}_1\bar{x} && (7) \end{aligned}$$

and

$$\begin{aligned} \mathrm{E}(\hat{\beta}_1) &= \beta_1 && (8)\\ \mathrm{Var}(\hat{\beta}_1) &= \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2} && (9)\\ \mathrm{E}(\hat{\beta}_0) &= \beta_0 && (10)\\ \mathrm{Var}(\hat{\beta}_0) &= \sigma^2\left[\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right] && (11) \end{aligned}$$

James Forbes measured the atmospheric pressure and boiling point of water at 17 locations in the Alps. The dataset `forbes` is available in the library `MASS` and installed (only needed once) loaded by

```
install.packages("MASS") # then select the nearest CRAN mirror
library(MASS)
```

Let the boiling point be the response variable and the pressure be the explanatory variable:

```
y = forbes$bp
x = forbes$pres
```

b) Plot boiling point against pressure

```
plot(bp ~ pres,data=forbes)
```

Does it look like there is a linear relationship between boiling point and pressure?

c) Calculate Expression 6 and 7:

```
n = length(x);
beta1 = (sum(x*y)-n*mean(x)*mean(y))/(sum(x^2)-n*mean(x)^2);
beta0 = mean(y)-beta1*mean(x);
```

d) Plot the raw residuals (we will look more into the topic of residuals later in the course).

```
r = (y-(beta0+beta1*x))              # Residuals

plot((beta0+beta1*x),r,
xlab="Fitted values",ylab="Residuals",main="Residuals vs Fitted")

qqnorm(r)                            # qq-plot
qqline(r)

plot(r,ylab="Residuals")
```

- Is the linear model appropriate?
- Are the error terms independent?
- Are the error terms approximately normally distributed?
- Do the error terms have a common variance?

It can be shown that

$$t = \frac{\hat{\beta}_1 - \beta_1}{s/\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}} \tag{12}$$

is distributed as $t(n-2)$, t-distributed with $n-2$ degrees of freedom and

$$s^2 = \frac{\sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2}{n-2}. \tag{13}$$

e) Show that t can be used to construct a $100(1-\alpha)\%$ confidence interval for β_1:

$$\hat{\beta}_1 \pm t_{\alpha/2,n-2}\frac{s}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}}, \tag{14}$$

where $t_{\alpha/2,n-2}$ is t-quantiles.

Calculate the $100(1-0.05)\%$ confidence interval in R:

```
s = sqrt(sum(r^2)/(n-2))
alpha = 0.05
beta1 + qt(alpha/2, n-2)*s/(sqrt(sum((x-mean(x))^2)))         # lower
beta1 - qt(alpha/2, n-2)*s/(sqrt(sum((x-mean(x))^2)))         # upper
```

f) The t statistic can be used to test the hypothesis $H_0 : \beta_1 = 0, \quad H_1 : \beta_1 \neq 0$. Why should we reject H_0 if

$$\left| \frac{\hat{\beta}_1}{s/\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}} \right| > t_{\alpha/2,n-2} \tag{15}$$

In R:

```
t = beta1/(s/(sqrt(sum((x-mean(x))^2))))     # t-statistic
qt(alpha/2, n-2,low=F)                       # t-quantile
2*pt(t, n-2,low=F)                             # p-value
```

What would be the result of the test?

In R we can use lm to fit linear models.

```
lm(formula,data)
```

where

`formula` a symbolic description of the model to be fit. Note that the intercept term is included by default in the regression model, if you want to exclude it use the command `lm(y~x-1)` where `x` is the covariate you want to include

`data` name of the data frame (optional)

g) Fit a linear model with `lm` by

```
lm1 = lm(bp~pres,data=forbes)
```

The results:

```
summary(lm1)
confint(lm1)
```

Plot residuals

```
par(mfrow=c(1,2))        # change number of subplots in a window
plot(lm1,which=c(1,2))
```

Check that you get the same results as in **c)-f)**.