TMA4267 – Linear Statistical Models
Exercise 5 – V2015

10 March 2015

Comment: in Problem 4 β_0 is used for the intercept, so that with k additional covariates (not including intercept) $p = k+1$. This notation is used on previous exam questions in TMA4267. Problem 4 are exam questions from TMA4255 Applied statistics.

Problem 1: Simple calculations with the multivariate normal distribution

Let $\boldsymbol{X} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ with $\boldsymbol{\mu} = \begin{bmatrix} 2 \\ -3 \\ 1 \end{bmatrix}$ and $\boldsymbol{\Sigma} = \begin{bmatrix} 1 & 1 & 1 \\ 1 & 3 & 2 \\ 1 & 2 & 2 \end{bmatrix}$

a) Find the distribution of $3X_1 - 2X_2 + X_3$.

b) Find a 2×1 vector $\boldsymbol{a}$ such that X_2 and $X_2 - \boldsymbol{a}^T \begin{bmatrix} X_1 \\ X_3 \end{bmatrix}$ are independent.

c) Find the conditional distribution of X_1 given that $X_2 = x_2$ and $X_3 = x_3$.

Problem 2: Conditional expectation with the multivariate normal

Let

$$\mathbf{X} \sim N(\mathbf{0}, \Sigma) \tag{1}$$
$$\mathbf{Y} = \mathbf{X} + \mathbf{e} \tag{2}$$
$$\mathbf{e} \sim N(\mathbf{0}, \sigma^2 I) \tag{3}$$

where I is an $n \times n$ identity matrix and $\text{Cov}(\mathbf{X}, \mathbf{e}) = 0$.

a) Find an expression for the distribution of

$$\mathbf{Z} = \begin{bmatrix} \mathbf{X} \\ \mathbf{Y} \end{bmatrix} \tag{4}$$

and $[\mathbf{X} \mid \mathbf{Y}]$.

Realizations from $\mathbf{Z}$ can be sampled by

```
n = 100
d = 10
Sigma = exp(-as.matrix(dist(1:n)/d)^2)
sigma = 0.3                  # standard deviation
x = mvrnorm(1,rep(0,n),Sigma)
y = x + rnorm(n,0,sigma)

plot(x,type="l")
points(y)
```

b) Simulate a realization from $\mathbf{Z}$, try with different values for d and σ. Examine the covariance matrix with e.g. `image(Sigma)`.

The conditional expectation and 2 standard deviation confidence intervals for $\mathbf{X}$ given $\mathbf{Y}$ can in R be calculated by

```
condexp = 0 + Sigma%*%solve(Sigma + diag(rep(sigma^2,n)),y-0)
condvar = Sigma - Sigma%*%solve(Sigma + diag(rep(sigma^2,n)),Sigma)
lines(condexp,col=2)
lines(condexp - 2*sqrt(diag(condvar)),col=2,lty=2)
lines(condexp + 2*sqrt(diag(condvar)),col=2,lty=2)
```

c) Calculate the conditional expectation and 2 standard deviation confidence intervals. Comment the results.

Problem 3: Rat liver MLR

Description: The data describe an experiment conducted to investigate the amount of drug present in the liver of a rat. Nineteen rats were randomly selected, weighed, and placed under a light anesthetic and given an oral dose of the drug. Because it was thought that large livers would absorb more of a given dose than a small liver, the actual dose given was approximately determined as 40mg of the drug per kilogram of body weight. After a fixed length of time, each rat was sacrificed and the liver weighed, and the percent dose in the liver was determined.

There are nineteen observations on three covariates and one response:

`BodyWt`	The body weight of each rat in grams
`LiverWt`	The weight of each liver in grams
`Dose`	The relative dose of the drug given to each rat as a fraction of the largest dose
`DrugInLiver`	The proportion of the dose in the liver

a) Download and import the data:

```
data=
read.table("http://www.math.ntnu.no/~mettela/TMA4267/Data/ratliver.txt",
header=TRUE)
```

Make pairwise scatter plots of the data using `pairs`. Look at the relationship between the variables and compute the correlation matrix. Is a linear model appropriate? Comment on what you see.

Consider a model on the form

$$Y_i = \boldsymbol{\beta}^T \mathbf{X}_i + e_i, \quad i = 1, \ldots, n, \tag{5}$$

where $e_1, \ldots, e_n$ iid and $e_i \sim N(0, \sigma^2), \forall i$, Y_i is the response and $\mathbf{X}_i$ covariates.

b) We are now first going to fit a linear model to `BodyWt` only using the covariate `Dose` :

```
model1 = lm(BodyWt~Dose,data=data)
par(mfrow=c(1,2))
model1
plot(model1,which = c(1,2))
summary(model1)
```

Explore the distribution of the residuals and comment on your findings.

Explain the output from `summary(model1)`. Use `?summary.lm` and find:

- Estimation of β
- Standard Error
- Residual Standard Error
- R-Squared and Adjusted R-Squared
- F-statistic

c) Fit a linear model to `BodyWt` using the covariates `Dose` and `DoseInLiver`:

```
model2 = lm(BodyWt~Dose + DoseInLiver ,data=data)
```

Using all the predictor variables:

```
model3 = lm(BodyWt~.,data=data)
```

Compare the models and comment on your findings.

Which model is used :

```
model4 = lm(BodyWt~Dose -1 ,data=data)
```

Problem 4: Concrete

In an article in *Journal of Computing in Civil Engineering* the aim was to make a prediction model for the quantity of concrete, y, to be used in the construction of a silo complex. The prediction model will be used in the design stage of a construction job. A total of 23 possible explanatory variables were listed, and we will look at four of these. The following description is given.

- y, `concrete`. Quantity of concrete measured in m^3.
- x_1, `volume`. The volume of the silo complex.
- x_2, `perimeter`. The perimeter of the silo complex.

- x_3, `waste`. Waste percent in concrete.
- x_4, `steel`. The number of reinforcing steel crews.

Data are available from 28 construction jobs. Scatter plots are found in Figure 1.

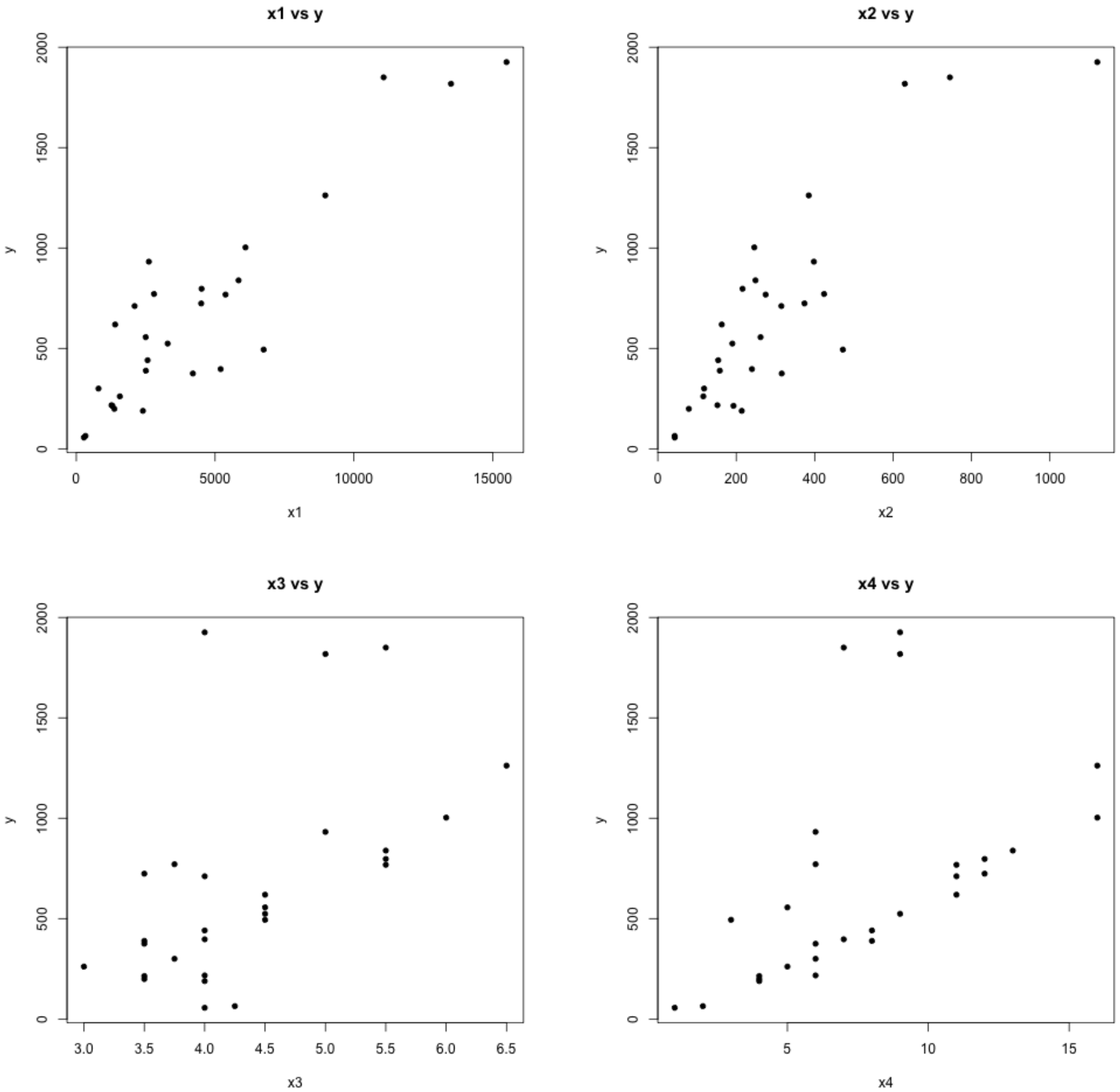

Figure 1: Scatter plots for the concrete quantities data set.

A multiple linear regression was fitted to the data with y as response and x_1, x_2, x_3 and x_4 as explanatory variables. Let $(y_i, x_{1i}, x_{2i}, x_{3i}, x_{4i})$ denote the observations from job i, where $i = 1$, $\ldots$, 28. Define the full model (model A):

$$\text{Model A: } y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \beta_4 x_{4i} + \varepsilon_i$$

where the ε_i's are i.i.d. $N(0, \sigma^2)$ for $i = 1, \ldots, 28$. Printout from a statistical analysis is found in Figure 2 and plots of studentized residuals are found in Figure 3. Two of the numerical values in the printout have been replaced by question marks.

```
Predictor      Coef  SE Coef       T      P
Constant     -574,2    190,1  -3,02  0,006
x1          0,02670  0,02142   1,25  0,225
x2           1,3612   0,3241   4,20  0,0003
x3           124,08    49,39   2,51  ?
x4            23,22    10,28   2,26  0,034

S = 158,231   R-Sq = ? %   R-Sq(adj) =  90,56 %

Analysis of Variance

Source          DF       SS       MS      F      P
Regression       4  6586535  1646634  65,77  0,000
Residual Error  23   575852    25037
Total           27  7162388
```

Figure 2: Printout from statistical analyses for Model A of the concrete data set.

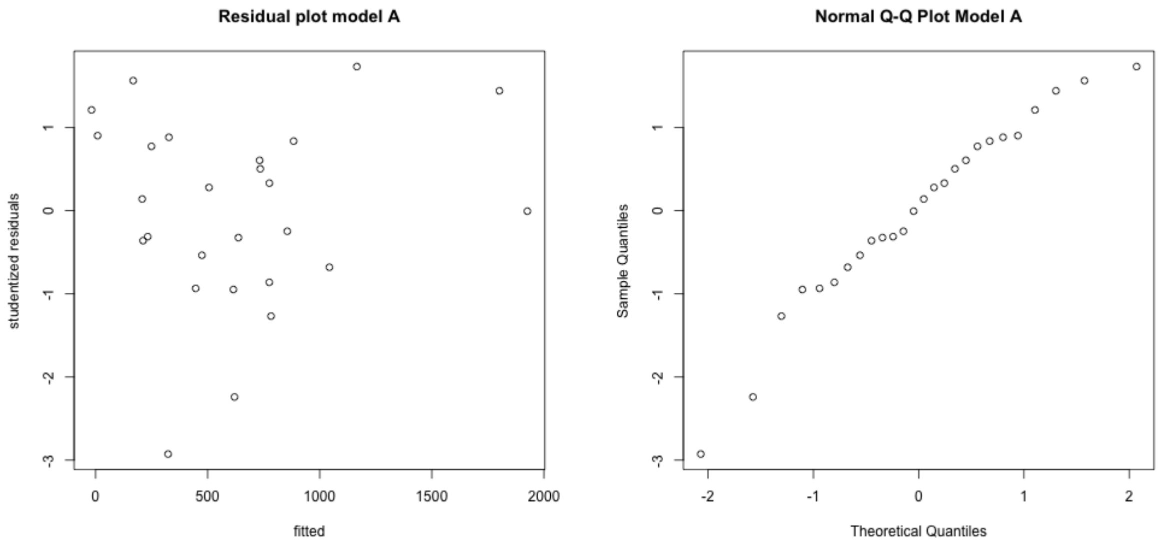

Figure 3: Residual plots (studentized residual versus fitted values in the left panel, normal plot based on studentized residuals in the right panel) for Model A for the concrete quantities data set.

```
Predictor        Coef   SE Coef         T           P
Constant    -815,6      174,1    -4,68      8,45e-05
x2           1,7575    0,1521    11,55      1,62e-11
x3           219,65     40,09     5,48      1,09e-05

S = 176,717   R-Sq = 89,1%   R-Sq(adj) = 88,2%

Analysis of Variance
Source          DF       SS       MS       F      P
Regression       2  6381667  3190833  102,18  0,000
Residual Error  25   780721    31229
Total               27  7162388
```

Figure 4: Printout from statistical analyses for Model B of the concrete data set.

a) Write down the estimated regression equation.

Now turn to the estimated regression coefficient for x_3, `waste`, in this model. How would you explain this number to the common man (that does not know linear regression)?
Is the effect of x_3, `waste`, significant in this model?

Calculate the R^2 and explain how you can interpret this value.

Do you think Model A is a good model for the data? To answer this you need to comment on important features in the printout from the statistical analysis and the residual plots in Figure 3.

We now want to compare the full regression model (model A), with a reduced model (called model B) with only x_2 (`perimeter`) and x_3 (`waste`).

$$\text{Model B: } y_i = \beta_0 + \beta_2 x_{2i} + \beta_3 x_{3i} + \varepsilon_i$$

The results from fitting model B are found in Figure 4.

b) Comment on the most important differences between model A and model B.

Model A and model B can be compared by testing the following hypotheses.

$$H_0\colon \beta_1 = \beta_4 = 0 \text{ vs. } H_1\colon \beta_1 \text{ and } \beta_4 \text{ are not both zero}$$

Perform the hypothesis test and conclude.

Would you prefer to use Model A or B?

c) We will now use Model B.

Assume that a new construction job is planned and that we expect that $x_2 = 300$ and $x_3 = 4.4$ for this new job. What would be the best prediction for the quantity of concrete to be used?

Let X denote the design matrix used to fit Model B, and let $\boldsymbol{x}_0$ be the covariate vector we want to use in the prediction, $\boldsymbol{x}_0^T = [1\ 300\ 4.4]$. Is is given that then $\boldsymbol{x}_0^T(X^TX)^{-1}\boldsymbol{x}_0 = 0.0357$. Use this information to construct a 95% prediction interval for the quantity of concrete to be used when $x_2 = 300$ and $x_3 = 4.4$.
What is the interpretation of this interval?

Problem 5: ANOVA and MLR

In this problem we will consider a data set where income is explained by the two factors place and sex.

	Place A	Place B	Place C
Male	300 350 370 360	400 370 420 390	400 430 420 410
Female	300 320 310 305	350 370 340 355	370 380 360 365

Hence two categorical predictors: sex with 2 levels (Male and Female) and place with 3 levels (named A, B and C).

a) Enter the data in R as a data.frame with three columns, income, gender and place. Make gender and place to be factors. Call this `data`.

Examine the data visually with e.g.

```
pairs(data)
plot(income~place,data=data)
plot(income~gender,data=data)
interaction.plot(data$gender, data$place, data$income)
plot.design(income~place+gender, data = data)
```

b) (One-way anova).
Consider first a model with a factor α_i occurring at $i = 1, \ldots, I$ levels, with K observations per level. Use the model

$$y_{ik} = \mu + \alpha_i + e_{ik}, \quad i = 1, \ldots, I \quad k = 1, \ldots, K \tag{6}$$

where $e_{11}, \ldots, e_{IK}$ iid and $e_{ik} \sim N(0, \sigma^2)$, $\forall i, k$. Assume place is the only factor. The design matrix X can be found by:

```
X = cbind(rep(1,length(data$income)),data$place=="A",
   data$place=="B",data$place=="C")
```

What is the rank of X^TX? Why do we need X^TX to have full rank. How can we solve rank problems?

Hint: `qr(matrix)$rank`.

c) Fit the model:

```
model = lm(income~place-1,data=data,x=TRUE)
```

where `x=TRUE` tells the function to calculate the design matrix X, which is stored as `model$x`. Examine the results with `summary` and `anova`.

What parameterization is used? What is the interpretation of the parameters?

d) Fit the models:

```
options(contrasts=c("contr.treatment","contr.poly"))
model1 = lm(income~place,data=data,x=TRUE)

options(contrasts=c("contr.sum","contr.poly"))
model2 = lm(income ~ place,data=data,x=TRUE)
```

What are the parameterizations? What is the interpretation of the parameters and how does the parameter interpretations differ between the models in b) and c)?

(You can read about `contr.treatment` by typing `?contr.treatment`, and ignore `contr.poly`.)

e) (two-way anova) Suppose now that there are two factors: α_i at I levels and β_j at J levels, with K observations per level. Then a general model is

$$y_{ijk} = \mu + \alpha_i + \beta_j + e_{ijk}, \quad i = 1, \dots, I \quad j = 1, \dots, J \tag{7}$$

where $e_{111}, \dots, e_{IJK}$ iid and $e_{ijk} \sim N(0, \sigma^2)$, $\forall i, j, k$.

Fit the model without interactions effects:

```
options(contrasts=c("contr.treatment","contr.poly"))
model3 = lm(income~place+sex,data=data,x=TRUE)
summary(model3)
anova(model3)

options(contrasts=c("contr.sum","contr.poly"))
model4 = lm(income~place+sex,data=data,x=TRUE)
summary(model4)
anova(model4)
```

What are the parameterizations? What is the interpretation of the parameters? Does the ANOVA table look diffent for the two parameterizations?

Finally, fit a model with interactions for both the contrasts and check if the interaction effect is significant.