TMA4267 Linear Statistical Models
Part 2: Linear regression
Recommended exercise 3 - V2017

February 7, 2017

Keywords: simple linear regression, multiple linear regression, residual plots, normal plots, requirements for linear regression, hat matrix, interpretation, weighted least squares.

- Problem 1 is an introduction to linear models - with the simple linear regression - in `R` using matrix algebra and also the built in R-function `lm`. We discuss the requirements of a linear model and study residual plots.
- Problem 2 takes this a bit further, moving on to multiple regressors, but also with focus on interpretation of plots and regression output.
- Problem 3 looks at important results for performing inference on regression parameters. Important ingredients are two symmetric and idempotent matrices; $\boldsymbol{H}$ hat matrix (projecting onto the column space of the design matrix $\boldsymbol{X}$) and the $\boldsymbol{I}-\boldsymbol{H}$ matrix used to define residuals (projecting onto the space orthogonal to the column space of the design matrix $\boldsymbol{X}$). And, then uses these matrices to show that estimators for regression parameters are independent from estimator for variance of error terms - identical to what was done for the mean and standard deviation in Compulsory Exercise 1, Problem 2.
- Problem 4 shows how the classical regression problem with uncorrelated errors with equal variances can be relaxed with the aid of weighted least squares. Only the square root matrix and results from Part 1 on multivariate normal distribution is needed.

Problem 1: Simple linear regression

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) and loaded by

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

a) Getting to know the data set. Check out the forbes data set (which is a data frame).

```
help(forbes)
mode(forbes)
names(forbes)
```

We will fit a simple linear model with boiling point as response and atmospheric pressure as the covariate. Let the boiling point (in degrees Celsius, converted from Fahrenheit) be the response variable and the pressure (in bar, converted from inches of mercury) be the explanatory variable, and construct the vector of reponses $\boldsymbol{Y}$ and the design matrix $\boldsymbol{X}$.

```
n = length(forbes$bp)
Y = matrix((forbes$bp-32)*5/9,ncol=1)
X = cbind(rep(1,n),forbes$pres*0.033863882)
```

What is the rank of $\boldsymbol{X}$?

b) Plot pressure vs boiling point.

```
plot(X[,2],Y,pch=20)
```

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

c) Calculate $\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y}$.

How would you explain to a layperson what these two numbers mean?

d) Plot the pressure, $\boldsymbol{X}[,2]$, against the raw residuals $\hat{\boldsymbol{\varepsilon}} = \boldsymbol{Y} - \hat{\boldsymbol{Y}}$, where $\hat{\boldsymbol{Y}} = \boldsymbol{X}\hat{\boldsymbol{\beta}}$. (We will look more into the topic of various types of residuals later in the course.) Comment on what you see.

e) Now we have looked at plots of covariate vs response and covariate vs residual. We may use these two plots to assess if the 4 assumptions (discussed in the lectures) for the classical linear model are fulfilled:

- Is the linear model appropriate?
- Are the error variances homoscedastic?
- Are the error terms uncorrelated?
- Does an additive model seem appropriate?

In addition we may also want to investigate if the errors are normally distributed. How can we do that? Comment on your findings.

In `R` we can use `lm` to fit linear models.

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

`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)

f) Fit a linear model with `lm`.

```
newds=data.frame(bp=(forbes$bp-32)*5/9,pres=forbes$pres*0.033863882)
lm1 = lm(bp~pres,data=newds)
```

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 **b)-e)**. Observe that summary gives many numbers – which we will look at the reasoning behind in the course.

Problem 2: Happiness

(Modified version of Exam TMA4255 V2011, Problem 4: TMA4255 Applied statistics is roughly speaking a non-mathematical version of TMA4267 for students at other programs than Industrial mathematics, and uses MINITAB instead of R and notation with sums for the regression instead of matrix notation. This version of the exam is shortened and adapted to R, and is suitable as an exam question in TMA4267.)

We will look at data collected from 39 individuals in a Master of Business Administration class for employed students at the University of Chicago Graduate School of Business. The reason for collecting the data was to test the hypothesis that love and work are the important factors in determining an individual's happiness. As alternatives, the variables money and sex (sexual activity) were included in the study. The five variables were codes as follows.

- y, `happiness`. Happiness was measured on a 10-point scale, with 1 representing a suicidal state, 5 representing a feeling of «just muddling along», and 10 representing a euphoric state.
- x_1, `money`. Money was measured by annual family income in thousands of dollars.
- x_2, `sex`. Sex was measured as the values 0 or 1, with 1 indicating a satisfactory level of sexual activity.
- x_3, `love`. Love was measured on a 3-point scale, with 1 representing loneliness and isolation, 2 representing a set of secure relationships, and 3 representing a deep feeling of belonging and caring in the context of some family or community.
- x_4, `work`. Work was measured on a 5-point scale, with 1 indicating that an individual is seeking other employment, 3 indicating the job is OK, and 5 indicating that the job is enjoyable.

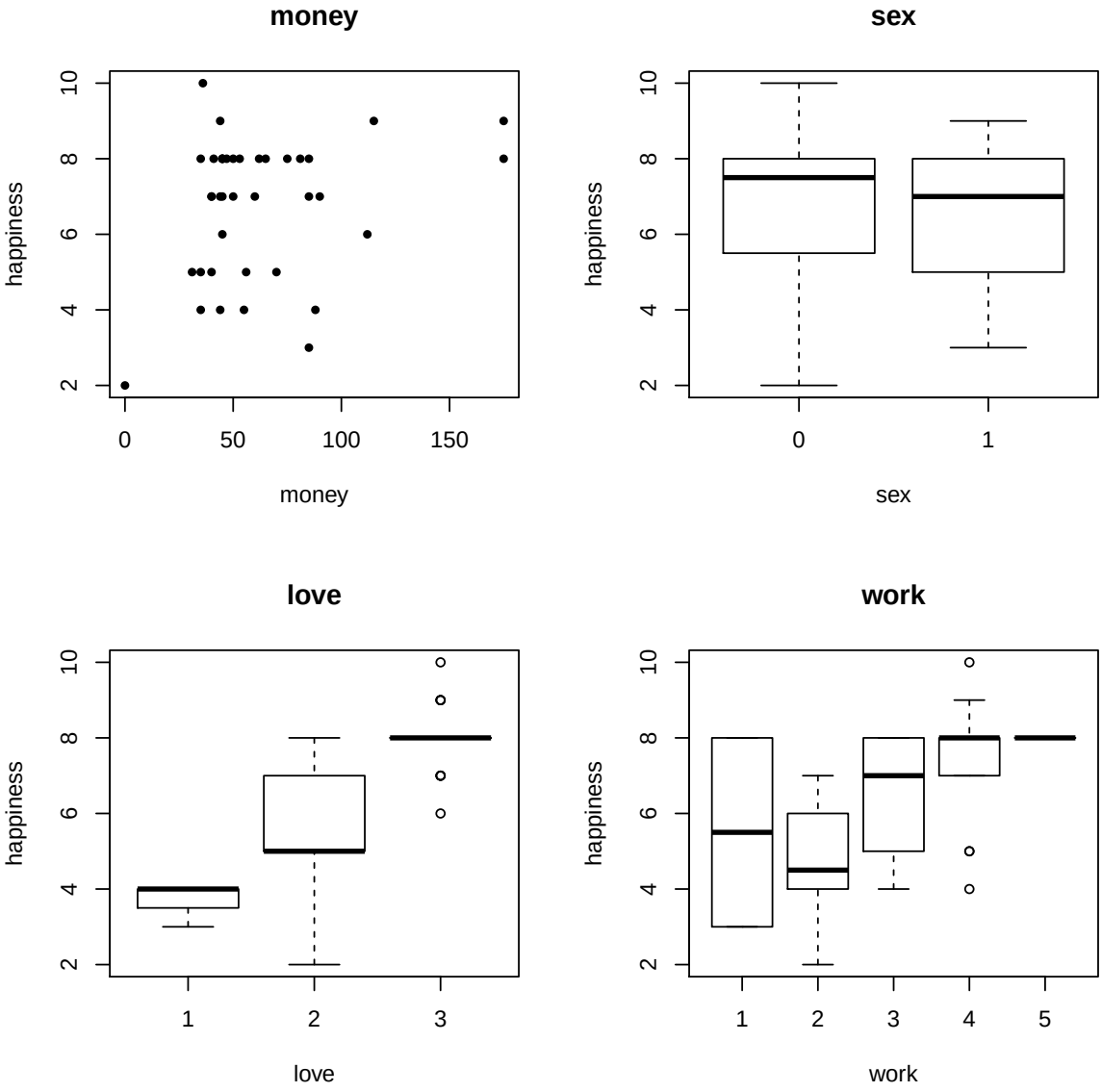
Boxplots and scatter plots are found on the next pages, and the data set is available from library "faraway" as data set "happy" (see the associated R-code at the course www-page).

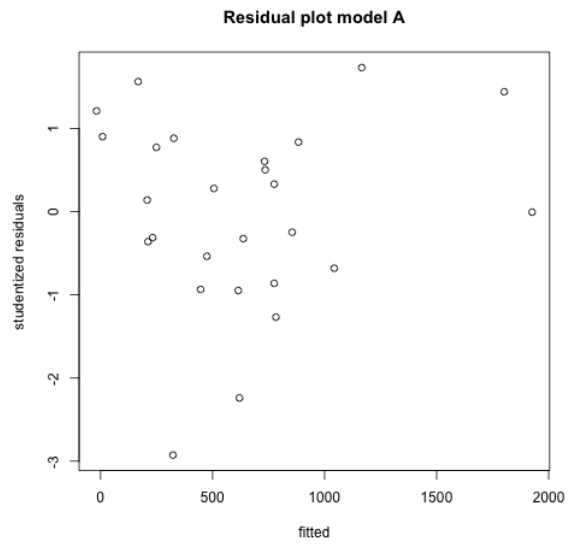
A multiple linear regression was fitted to the data with y as response and x_1, x_2, x_3 and x_4 as explanatory variables. Let $(y_i, x_{i1}, x_{i2}, x_{i3}, x_{i4})$ denote the observations from individual i, where $i = 1, \ldots, 39$. Define the full model (model A):

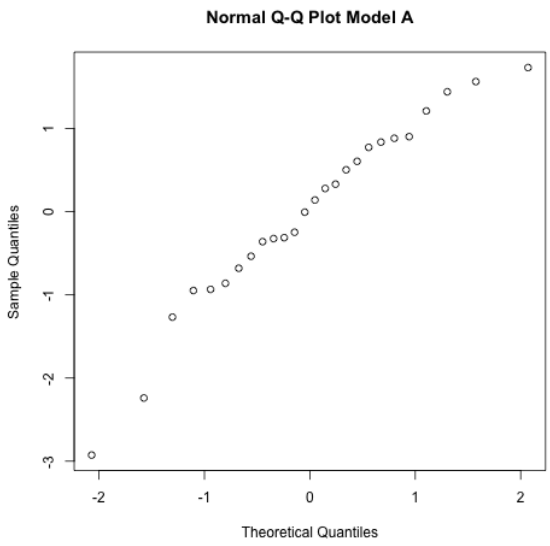
$$\text{Model A: } y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4} + \varepsilon_i$$

where the ε_i's are i.i.d. $N(0, \sigma^2)$ for $i = 1, \ldots, 39$.

Printout from R and plots of (studentized) residuals are found on the next pages. Three of the numerical values in the R printout have been replaced by question marks.








```
> fitA <- lm(happy~.,data=happy)
> summary(fitA)

Call:
lm(formula = happy ~ ., data = happy)
Residuals:
    Min      1Q  Median      3Q     Max
-2.7186 -0.5779 -0.1172  0.6340  2.0651
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.072081   0.852543  -0.085   0.9331
money        0.009578   0.005213   1.837   0.0749 .
sex         -0.149008   0.418525  -0.356   0.7240
love         1.919279   0.295451   6.496 1.97e-07 ***
work         0.476079   0.199389   ? ?
---
Residual standard error: 1.058 on 34 degrees of freedom
Multiple R-squared:  0.7102,Adjusted R-squared:  0.6761
F-statistic: 20.83 on 4 and 34 DF,  p-value: ?
```

a) What is the estimated regression coefficient for x_4, `work`? How would you explain this number to the common man (that does not know linear regression)? Is the effect of x_4, `work`, significant in this model? You decide yourself which significance level you choose.

Is the regression found to be significant? You decide yourself which significance level you choose. Comment briefly on the residual plots.

We now want to compare the full regression model (model A), with a reduced model (called model B) where x_1 (`money`) and x_2 (`sex`) are excluded from the model:

$$\text{Model B: } y_i = \beta_0 + \beta_3 x_{i3} + \beta_4 x_{i4} + \varepsilon_i$$

The results from fitting model B are as follows.

```
> fitB<- lm(happy~work+love,data=happy)
```

```
> summary(fitB)
Call:
lm(formula = happy ~ work + love, data = happy)
Residuals:
    Min      1Q  Median      3Q     Max
-3.1454 -0.6365 -0.1259  0.8333  1.8741
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   0.2057     0.7757   0.265  0.79241
work          0.5106     0.1874   2.725  0.00987 **
love          1.9592     0.2954   6.633 9.99e-08 ***
---
Residual standard error: 1.08 on 36 degrees of freedom
Multiple R-squared:  0.6808,Adjusted R-squared:  0.6631
F-statistic: 38.39 on 2 and 36 DF,  p-value: 1.182e-09
```

b) The estimate $\hat{\beta}_3$ (`love`) is 1.919 for model A and 1.959 for model B. Explain why these two estimates differ.

Problem 3: Results on $\hat{\boldsymbol{\beta}}$ and SSE in multiple linear regression

(Exam K2014, Problem 4)

The classical multiple linear regression model can be written in matrix notation as

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon},$$

where $\boldsymbol{Y}$ is a n-dimensional random column vector, $\boldsymbol{X}$ is a fixed design matrix with n rows and p columns, $\boldsymbol{\beta}$ is an unknown p-dimensional vector of regression coefficients and $\boldsymbol{\varepsilon}$ is a n-dimensional vector of random errors.

Assume that $n > p$ and that $\boldsymbol{X}$ has rank p.

Define the matrix $\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T$.

a) What type of matrix is $\boldsymbol{H}$? Justify your answer.
Find the rank of $\boldsymbol{H}$.
How would you graphically interpret the vector $\boldsymbol{HY}$?

Answer the same three questions for the matrix $\boldsymbol{I} - \boldsymbol{H}$, using the findings you already have for $\boldsymbol{H}$. Here $\boldsymbol{I}$ is the $n \times n$ identity matrix.

Further, assume that the vector of random errors $\boldsymbol{\varepsilon}$ is multivariate normal with mean $\mathrm{E}(\boldsymbol{\varepsilon}) = \boldsymbol{0}$ and covariance matrix $\mathrm{Cov}(\boldsymbol{\varepsilon}) = \sigma^2\boldsymbol{I}$, where $\boldsymbol{I}$ is the $n \times n$ identity matrix.

b) Let $\mathrm{SSE} = \boldsymbol{Y}^T(\boldsymbol{I} - \boldsymbol{H})\boldsymbol{Y}$. Derive the distribution of SSE.
Use this to suggest an unbiased estimator for σ^2, and call the estimator $\hat{\sigma}^2$.
Find the variance of $\hat{\sigma}^2$.

Define two constant matrices $\boldsymbol{A} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T$ and $\boldsymbol{B} = (\boldsymbol{I} - \boldsymbol{H})$.

c) What are the dimensions of the matrices $\boldsymbol{A}$ and $\boldsymbol{B}$?
Show that $\boldsymbol{AY}$ and $\boldsymbol{BY}$ are independent random vectors.
Use this to prove that the least squares estimator $\hat{\boldsymbol{\beta}}$ and SSE are independent random variables. What is the use of this result in multiple linear regression?

Problem 4: Weighted least squares

The classical multiple linear regression (MLR) model can be written in matrix notation as

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon},$$

where $\boldsymbol{Y}$ is an n-dimensional random column vector, $\boldsymbol{X}$ is a fixed design matrix with n rows and p columns, $\boldsymbol{\beta}$ is an unknown p-dimensional vector of regression coefficients and $\boldsymbol{\varepsilon}$ is an n-dimensional column vector of random errors.

Further, in the classical multiple linear model we generally assume that the vector of random errors $\boldsymbol{\varepsilon}$ is multivariate normal with mean $\mathrm{E}(\boldsymbol{\varepsilon}) = \boldsymbol{0}$ and covariance matrix $\mathrm{Cov}(\boldsymbol{\varepsilon}) = \sigma^2\boldsymbol{I}$, where $\boldsymbol{I}$ is the $n \times n$ identity matrix.

We will now study slightly different situation. Assume that $\boldsymbol{\varepsilon}$ is multivariate normal with mean $\mathrm{E}(\boldsymbol{\varepsilon}) = \boldsymbol{0}$ and covariance matrix $\mathrm{Cov}(\boldsymbol{\varepsilon}) = \sigma^2\boldsymbol{V}$, where $\boldsymbol{V}$ is a known positive definite $n \times n$ matrix. The unknown parameters in this model are the regression coefficients $\boldsymbol{\beta}$ and the variance parameter σ^2.

a) Write down and explain the definition of the inverse square root matrix $\boldsymbol{V}^{-\frac{1}{2}}$.

Use the inverse square root matrix to define three new quantities

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

Use these new quantities together with the method of least squares to derive an unbiased estimator for $\boldsymbol{\beta}$, in terms of $\boldsymbol{X}$, $\boldsymbol{V}$ and $\boldsymbol{Y}$.

Show that the estimator is unbiased.

Is the ordinary least square estimator $\hat{\beta} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y}$ unbiased in this model? Justify your answer. Comment on your findings.

We go back to the classical MLR with identically normally distributed random errors, $\mathrm{Cov}(\boldsymbol{\varepsilon}) = \sigma^2\boldsymbol{I}$, but now look at misspecification of $\mathrm{E}(\boldsymbol{Y})$. Suppose that the true model is

$$\begin{aligned}\boldsymbol{Y} &= \boldsymbol{X}_1\boldsymbol{\beta}_1 + \boldsymbol{X}_2\boldsymbol{\beta}_2 + \boldsymbol{\varepsilon},\\ \boldsymbol{\varepsilon} &\sim N_n(\boldsymbol{0}, \sigma^2\boldsymbol{I}),\end{aligned} \tag{1}$$

where we have partitioned the design matrix into two parts $\boldsymbol{X}_1$ $(n \times p_1)$ and $\boldsymbol{X}_2$ $(n \times p_2)$ and $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$ are unknown p_1- and p_2-dimensional vectors of regression coefficients $(p = p_1 + p_2)$.

Assume that we ignore the covariates in $\boldsymbol{X}_2$ and fit the model

$$\begin{aligned}\boldsymbol{Y} &= \boldsymbol{X}_1\boldsymbol{\alpha}_1 + \boldsymbol{\delta},\\ \boldsymbol{\delta} &\sim N_n(\boldsymbol{0}, \tau^2\boldsymbol{I}).\end{aligned} \tag{2}$$

Here $\boldsymbol{\alpha}_1$ is used in place of $\boldsymbol{\beta}_1$ to emphasize that $\boldsymbol{\alpha}_1$ (and estimates thereof) will in general be different from $\boldsymbol{\beta}_1$ in the true model.

The least squares estimator for model (2) is $\hat{\boldsymbol{\alpha}_1} = (\boldsymbol{X}_1^T \boldsymbol{X}_1)^{-1} \boldsymbol{X}_1^T \boldsymbol{Y}$.

b) Find the mean and covariance matrix of $\hat{\boldsymbol{\alpha}_1}$ under the true model (1).

Under which conditions is $\hat{\boldsymbol{\alpha}_1}$ an unbiased estimator of $\boldsymbol{\beta}_1$? Justify your answer.