Norwegian University of Science and Technology
Department of Mathematical Sciences

TMA4267 Linear statistical models
Recommended exercises 5

Problem 1 Simple linear regression

James Forbes measured the atmospheric pressure and boiling point of water at 17 locations in the Alps. The data set `forbes` is available in the R package MASS (a companion package to the book *Modern applied statistics with S; fourth edition*, 2002, by Venables and Ripley). To install (only needed once) and load:

```
install.packages("MASS")
library(MASS)
```

a) Check out the data set (which is a data frame).

```
help(forbes) # or ?forbes
names(forbes)
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 $\boldsymbol{Y}$ of reponses and the design matrix X.

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

b) What is the rank of X?

c) Plot pressure versus boiling point.

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

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

d) Calculate $\hat{\boldsymbol{\beta}} = (X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}\boldsymbol{Y}$. How would you explain to a layperson what these two numbers mean?

e) Plot the pressure, the second column of X, against the raw residuals $\hat{\boldsymbol{\epsilon}} = \boldsymbol{Y} - \hat{\boldsymbol{Y}}$, where $\hat{\boldsymbol{Y}} = X\hat{\boldsymbol{\beta}}$. (We will look more into various types of residuals later in the course.) Comment on what you see.

f) Use the two plots, covariate versus response, and covariate versus residual, to assess if linearity of covariate effects, homoscedasticity of errors, uncorrelated errors and additivity of errors are satisfied.

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` is a symbolic description of the model to be fit. Note that the intercept term is included by default in the regression model. You can exclude it by using e.g. `lm(y ~ x - 1)`, where `x` is the covariate you want to include. `data` is name of the data frame (optional).

g) 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)
# or lm1 <- lm((forbes$bp - 32) * 5 / 9 ~ I(forbes$pres * 0.033863882))
# I() must be used to inhibit interpretation of "*" as formula operator
summary(lm1)
```

Check that you get the same results as in (c)–(f).

To plot residuals versus fitted values and Q–Q-plot of residuals:

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

Problem 2 Results on $\hat{\beta}$ and SSE in multiple linear regression

(Exam 2014 spring, Problem 4)

The classical multiple linear regression model can be written in matrix notation as $\boldsymbol{Y} = X\boldsymbol{\beta}+\boldsymbol{\epsilon}$, where $\boldsymbol{Y}$ is an n-dimensional random column vector, X is a fixed design matrix with n rows and p columns, $\boldsymbol{\beta}$ is an unknown p-dimensional vector of regression coefficients and $\boldsymbol{\epsilon}$ is an n-dimensional vector of random errors. Assume that $n > p$ and that X has rank p. Define the matrix $H = X(X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}$.

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

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

Further, assume that the vector $\boldsymbol{\epsilon}$ of random errors is multivariate normal with mean $E\boldsymbol{\epsilon} = \mathbf{0}$ and covariance matrix $\operatorname{Cov}\boldsymbol{\epsilon} = \sigma^2 I$, where I is the $n \times n$ identity matrix. Let $\mathrm{SSE} = \boldsymbol{Y}^{\mathrm{T}}(I - H)\boldsymbol{Y}$.

b) Derive the distribution of SSE. Use this to suggest an unbiased estimator for σ^2, and call the estimator $\widehat{\sigma^2}$. Find the variance of $\widehat{\sigma^2}$.

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

c) What are the dimensions of the matrices A and B? Show that $A\boldsymbol{Y}$ and $B\boldsymbol{Y}$ 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?