Lineaariset mallit, Spring 2012, Exercise 3, week 14

1. Consider the linear model $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, $\boldsymbol{\varepsilon} \sim \mathsf{N}\left(\mathbf{0}, \sigma^2\mathbf{I}_n\right)$ ($\boldsymbol{\beta} \in \mathbb{R}^p$, $\sigma^2 > \mathbf{0}$, $\mathsf{r}(\mathbf{X}) = p$). Show that the residual vector $\hat{\boldsymbol{\varepsilon}} = \mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}$ satisfies $\hat{\boldsymbol{\varepsilon}} = (\mathbf{I}_n - \mathbf{P})\,\boldsymbol{\varepsilon}$, where $\mathbf{P} = \mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'$ (an orthogonal projector) and $\hat{\boldsymbol{\beta}} = \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{Y}$ (the least squares estimator). Use this result to calculate $\mathsf{E}\left(\hat{\boldsymbol{\varepsilon}}\right)$, $\mathsf{Cov}\left(\hat{\boldsymbol{\varepsilon}}\right)$ and $\mathsf{Cov}(\hat{\boldsymbol{\varepsilon}}, \hat{\boldsymbol{\beta}})$.

2. Continuation of the preceding exercise. Show that $\hat{\mathbf{y}}'\hat{\mathbf{y}} = \mathbf{y}'\mathbf{X}\hat{\boldsymbol{\beta}}$ and that $\hat{\beta}_1 = \bar{y} - \hat{\beta}_2\bar{x}_2 - \cdots - \hat{\beta}_p\bar{x}_p$, when there is a constant term in the model or when $x_{i1} = 1$ for all $i = 1, ..., n$. Here $\bar{x}_j = (x_{1j} + \cdots + x_{nj})/n$, where x_{ij} is a general element of the matrix $\mathbf{X}$ ($i = 1, ..., n$, $j = 1, ..., p$), $\hat{\boldsymbol{\beta}} = [\hat{\beta}_1 \ \cdots \ \hat{\beta}_p]' = \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{y}$, and $\hat{\mathbf{y}} = \mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{P}\mathbf{y}$. (Note also the result $\mathbf{X}'\hat{\boldsymbol{\varepsilon}} = \mathbf{0}$.)

3. In exercise 1.4 is shown that $\sum_{i=1}^n (y_i - \bar{y})^2 = \sum_{i=1}^n y_i^2 - n\bar{y}^2 = \mathbf{y}'\left(\mathbf{I}_n - \mathbf{J}\right)\mathbf{y}$, where $\mathbf{y} = [y_1 \ \cdots \ y_n]'$ and $\mathbf{J} = \mathbf{1}_n\left(\mathbf{1}_n'\mathbf{1}_n\right)^{-1}\mathbf{1}_n'$, or that the total sum of squares SST $= \sum_{i=1}^n (y_i - \bar{y})^2$ can be written as SST $= \mathbf{y}'\left(\mathbf{I}_n - \mathbf{J}\right)\mathbf{y}$. Show that the regression sum of squares SSR $= \sum_{i=1}^n (\hat{y}_i - \bar{y})^2$ and the residual sum of squares SSE $= \sum_{i=1}^n \hat{\varepsilon}_i^2$ can be written as SSR $= \mathbf{y}'\left(\mathbf{P} - \mathbf{J}\right)\mathbf{y}$ and SSE $= \mathbf{y}'\left(\mathbf{I}_n - \mathbf{P}\right)\mathbf{y}$, respectively. Here $\mathbf{P} = \mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'$ with the first column of the matrix $\mathbf{X}$ equal to $\mathbf{1}_n = [1 \ \cdots \ 1]'$ $(n \times 1)$, and $\hat{y}_i$ and $\hat{\varepsilon}_i$ typical components of the vectors $\hat{\mathbf{y}}$ and $\hat{\boldsymbol{\varepsilon}}$. This gives you one way to see the result SST = SSR + SSE used to define the coefficient of determination $R^2 = 1-$ SSE/SST = SSR/SST.

4. Consider the model of exercise 1 in the special case $p = 1$ where the model can be expressed in component form as $Y_i = \beta x_i + \varepsilon_i$, $i = 1, ..., n$, with $\varepsilon_1, ..., \varepsilon_n$ independent and $\varepsilon_i \sim \mathsf{N}\left(0, \sigma^2\right)$ and the matrix $\mathbf{X}$ equal to the vector $\mathbf{x} = [x_1 \ \cdots \ x_n]'$. (i) Show that the least squares estimator of the parameter β is

$$\hat{\beta} = \frac{\sum_{i=1}^n x_i Y_i}{\sum_{i=1}^n x_i^2}.$$

(ii) Suppose now that the fixed regressor x_i in the model and least squares estimator is replaced by the random variable X_i. Show that the least squares estimator is unbiased or that $\mathsf{E}(\hat{\beta}) = \beta$, when the independence $(X_1, ..., X_n) \perp\!\!\!\perp (\varepsilon_1, ..., \varepsilon_n)$ holds. Explain further why unbiasedness does not necessarily hold without this independence. (*Hint*: In part (ii) you may use the equation $Y_i = \beta X_i + \varepsilon_i$ in the expression of $\hat{\beta}$ and end up considering the expectation of $\hat{\beta} - \beta$. You may assume that all needed expectations are finite. In the last part an exact mathematical proof is not required.)

5. Consider the model of exercise 1 in the special case of a simple linear regression so that in the component form of the model reads $y_i = \beta_1 + \beta_2 x_i + \varepsilon_i$, $i = 1, ..., n$, with $\varepsilon_1, ..., \varepsilon_n$ independent and $\varepsilon_i \sim \mathsf{N}\left(0, \sigma^2\right)$. Denote $\hat{\boldsymbol{\beta}} = [\hat{\beta}_1 \ \hat{\beta}_2]'$ (the least squares estimator) and $\bar{x} = (x_1 + \cdots + x_n)/n$. (i) Using the theorem below, form $\mathsf{Cov}(\hat{\boldsymbol{\beta}})$ and furthermore $\mathsf{Var}(\hat{\beta}_2)$ and $\mathsf{Var}(\hat{\beta}_1 + \hat{\beta}_2\bar{x})$. (ii) Suppose that the values of the regressors

$x_1, ..., x_n$ can be chosen freely from the interval $[c, d]$. How should they be chosen if the aim is to minimize $\mathsf{Var}(\hat{\beta}_2)$? Is this choice reasonable otherwise? (*Note*: In part (ii) an exact mathematical proof is not required and you may also assume that n is even.)

Theorem. *Consider the linear model* $\mathbf{Y} \sim \mathsf{N}\left(\mathbf{X}\boldsymbol{\beta}, \sigma^2 \mathbf{I}_n\right)$, $\boldsymbol{\beta} \in \mathbb{R}^p$, $\sigma^2 > 0$, *where* $\mathsf{r}(\mathbf{X}) = p$. *Then, the maximum likelihood estimators of the parameters* $\boldsymbol{\beta}$ *and* σ, $\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$ *and* $\hat{\sigma}^2 = \frac{1}{n}(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}})'(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}})$, *satisfy*

(i) $\hat{\boldsymbol{\beta}} \sim \mathsf{N}(\boldsymbol{\beta}, \sigma^2(\mathbf{X}'\mathbf{X})^{-1})$

(ii) $n\hat{\sigma}^2/\sigma^2 = (\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}})'(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}})/\sigma^2 \sim \chi^2_{n-p}$

(iii) $\hat{\boldsymbol{\beta}} \perp\!\!\!\perp \hat{\sigma}^2$.