Exercise 1
(see STK2100, spring 2018: 1^{st} mandatory assignment)

Consider a linear regression with qualitative (categorical) explanatory variables. The data are in the form $(c_1, y_1), \ldots, (c_n, y_n)$, where $c_i \in \{1, ..., K\}$. For $j = 1, \ldots, K$, define

$$x_{i,j} = \begin{cases} 1 & \text{if } c_i = j \\ 0 & \text{otherwise} \end{cases}$$

(a) Show that the two models

$$Y_i = \beta_0 + \beta_2 x_{i,2} + \cdots + \beta_K x_{i,K} + \varepsilon_i \tag{1}$$

and

$$Y_i = \alpha_1 x_{i,1} + \cdots + \alpha_K x_{i,K} + \varepsilon_i \tag{2}$$

are equivalent. Write down the connection between $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$ explicitly. Give also an interpretation of the parameters.

In the following we will stick to model (2) as it is mathematically easier to deal with.

(b) Let $\boldsymbol{X}$ be the design matrix for model (2), i.e. the i-th row of X contains the values $x_{i,j}, j = 1, ..., K$. Show that $\boldsymbol{X}^T\boldsymbol{X}$ becomes a diagonal matrix with diagonal elements n_j, where n_j is the number of observations for which $c = j$.

Also show that $\boldsymbol{X}^T\boldsymbol{y}$ is a vector with the j-th element equal to $\sum_{i:c_i=j} y_i$.

Based on this, derive the least squares estimates for $\alpha_1, .., \alpha_K$. Discuss whether the estimates are reasonable.

(c) Based on the relation between $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$ also construct the estimates for $\boldsymbol{\beta}$.

Explain why these estimates also become the least squares estimates for $\boldsymbol{\beta}$.

(d) Another alternative model is

$$Y_i = \gamma_0 + \gamma_1 x_{i,1} + \cdots + \gamma_K x_{i,K} + \varepsilon_i \tag{3}$$

where $\sum_{j=1}^{K} \gamma_j = 0$.

What values must γ_j, $j = 1, \ldots, K$ have in order that this model becomes equivalent to the previous two?

What interpretation do the γ's have in this case?