Technichal University of Munich
Department of Mathematics

MA4401 Applied Regression, Homework problem 4

Prof. Donna Ankerst, Stephan Haug (December 12, 2017)

Problem H.4

Consider the linear model $\mathbf{y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, with X an $n \times (p+1)$ matrix with rank $p+1$ and $\boldsymbol{\varepsilon} \in \mathbb{R}^n$ a vector of uncorrelated errors with mean $\mathbf{0}$ and covariance matrix $\sigma^2 I_n$. Further let $\widehat{\boldsymbol{\mu}} = X\widehat{\boldsymbol{\beta}}$ be the fitted values, where $\widehat{\boldsymbol{\beta}}$ is the vector of least squares estimates, and $H = X(X'X)^{-1}X'$ denotes the hat matrix.

a) Find the mean vector and covariance matrix of $\widehat{\boldsymbol{\mu}}$.

b) Show that

$$\frac{1}{n}\sum_{i=1}^{n} \mathrm{Var}(\widehat{\boldsymbol{\mu}}_i) = \sigma^2 \frac{p+1}{n}$$

Hint: Find the trace of $\mathrm{Cov}(\widehat{\boldsymbol{\mu}})$ and use the fact that $\mathrm{tr}(AB) = \mathrm{tr}(BA)$ for matrices A and B, whenever the product is well-defined.

c) Show that H is a symmetric and idempotent matrix (https://en.wikipedia.org/wiki/Idempotent_matrix). Further show that the diagonal elements h_{ii} must lie between zero and one.

Hint: Consider $\mathbf{a}_i' H \mathbf{a}_i$, where $\mathbf{a}_i \in \mathbb{R}^n$ is a vector with all components equal to 0 except for the i-th, which is 1.

d) Assume that the linear model contains a constant term. Show that the diagonal elements h_{ii} of the hat matrix satisfy $h_{ii} \geq \frac{1}{n}$.

Hint: Parametrise the model by centering the predictor variables, i.e. consider $x_{ij} - \overline{x}_j$, $j = 1, \ldots, p$, as predictor variables instead of x_{ij}.

e) Read in the weightloss data set available on moodle. The response variable is `Loss` (weight loss in pounds after 1 month of diet). The predictor variables are `Diet` (type of diet), and `Before` (weight in pounds before the diet).
Use `ggplot()` for a scatterplot of `Loss` against `Before`. Determine the hat matrix for the model `Loss ~ Before`. Based on the hat matrix, compute the leverage for all data points. Mark the data points with high leverage in a different colour in the scatterplot. Does this approach catch all outliers?