MathALEA Mathieu Ribatet — École Centrale de Nantes

Linear models (`v1`)

Remark. This exercise sheet puts emphasis on the theoretical properties of the linear model

$$Y_i = x_i^\top \beta + \varepsilon_i, \qquad \varepsilon_i \overset{iid}{\sim} N(0, \sigma^2), \quad i = 1, \dots, n.$$

Don't expect much concrete stuffs here... No worries, we'll cover that in due time during Labs!

Further, the framework assumed throughout this document is simplified as it completely ignores the randomness of the covariates x_i. A more rigorous treatment should do all the computations conditionned on those x_i as emphasized by the alternative definition covered in the lecture

$$\mathbb{E}[Y \mid x] = x^\top \beta.$$

Fortunately, results presented here are still valid (but we need mutual independence between the x_i and ε_i though)

Exercise 1 (The hat matrix).

a) Show that the "hat matrix" is $H = X\left(X^\top X\right)^{-1} X^\top$, and that we have $H = H^\top = H^2$.

b) Show that $H(\mathrm{Id} - H) = \mathbf{0}$ and $HX = X$.

c) Give a closed form expression for the prediction of $\hat{Y}_i$ in terms of the matrix H and the response vector $Y = (Y_1, \dots, Y_n)^\top$. Deduce why the i-th diagonal element of H, say h_{ii}, can be seen as a leverage measure?

d) Show that $\mathrm{tr}(H) = p$, where p is the number of parameter to be estimated in our linear model. Note: We will thus say that the model has p degrees of freedom.

✎ ✎ ✎

Exercise 2 (Variance estimation).

a) Find the maximum likelihood estimator for σ^2.

b) Is it biased? If yes, give an unbiased version.

✎ ✎ ✎

Exercise 3 (Residual properties). In this exercise we focus on the residuals $R_i = Y_i - \hat{Y}_i$ and we let $R = (R_1, \dots, R_n)^\top$.

a) Show that $X^\top R = \mathbf{0}$ and deduce that $\sum_{i=1}^n R_i = 0$.

b) What do you think about people claiming that their fitted linear model (from least squares) is good since the mean error is 0.

c) Show that $\sum_{i=1}^n \hat{Y}_i R_i = 0$. Give a geometrical interpretation.

d) Show that $R = (\mathrm{Id} - H)\varepsilon$, where Id is the identity matrix (with desired dimension). Deduce that $\mathrm{Var}(R) = \sigma^2(\mathrm{Id} - H)$.

e) What can you say about the residuals R_i?

✎ ✎ ✎

Exercise 4 (Leave one out ☑). We often need to evaluate the performance of a predictive model, for example to do model selection. In a regression setting, it is common pradctice to use the residual sum of square (RSS in the lecture) as performance measure. However, strictly speaking it is not a predictive performance measure as we use Y_i in the fitting stage to get a prediction $\hat{Y}_i$. It would be much more sensible to see what happens when we predict Y_i without using it in the fitting stage. This is the **leave one out** principle. Hence we rather focus on the predictive residual sum of squares

$$\mathrm{PRSS} = \sum_{i=1}^{n} \left(Y_i - \hat{Y}_{[i]}\right)^2,$$

where $\hat{Y}_{[i]}$ is the prediction of Y_i obtained when the model is fitted without Y_i.

a) Let $\hat{\beta}_{[i]}$ the MLE where the i-th observation has been discarded from the fitting stage. Show that

$$\hat{\beta}_{[i]} = \hat{\beta} + \left(\sum_{\substack{j=1 \\ j\neq i}}^{p} h_{ij} Y_j\right) \frac{(X^\top X)^{-1} x_i}{1 - h_{ii}},$$

where X is the design matrix, x_i its i–th row and $\hat{\beta}$ the MLE (obtained from all observations).

Hint: You'll probably need to use the Sherman–Morisson formula

$$(A + uv^\top)^{-1} = A^{-1} - \frac{A^{-1}uv^\top A^{-1}}{1 + v^\top A^{-1} u}.$$

b) Deduce that

$$\hat{Y}_{[i]} = \frac{\sum_{j=1}^{p} h_{ij} Y_j 1_{\{j\neq i\}}}{1 - h_{ii}}.$$

c) Finally show that

$$PRSS = \sum_{i=1}^{n} \left(\frac{Y_i - \hat{Y}_i}{1 - h_{ii}}\right)^2$$

and explain how one can benefit from the above expression compare to the original one?

✎ ✎ ✎

Exercise 5 (Linear smoother and Ridge regression). A predictive model is said to be a linear smoother if $\hat{Y} = SY$ where S is the smoother matrix. Clearly, the linear model is a linear smoother with $S = H$.

Let's focus on the following optimization problem

$$J_\lambda(\beta) = (Y - X\beta)^\top (Y - X\beta) + \lambda \beta^\top \beta,$$

for some fixed value of $\lambda > 0$.

a) Show that the solution of this optimization problem is a linear smoother and give an expression for its smoothing matrix (that we will denote $S(\lambda)$).

b) What is $S(\lambda)$ as $\lambda = 0$? And when $\lambda \to \infty$?

c) Show that $S(\lambda) = U\tilde{D}U^\top$ where $\tilde{D}$ is a diagonal matrix and U is the orthogonal matrix of the SVD of X, i.e., $X = UDV^\top$.

d) Deduce that

$$\mathrm{tr}\{S(\lambda)\} = \sum_{j=1}^{p} \frac{d_j^2}{d_j^2 + \lambda}.$$

and that the degrees of freedom decrease as λ increases. Comment.

✎ ✎ ✎