

Series 5

1. The leave-one-out CV-score for the cubic smoothing spline and the least squares parametric estimator can be written in such a way that it depends only on the estimator $\hat{m}(\cdot)$ which is computed from the *full* dataset. To obtain the CV-score, it is therefore not necessary to calculate the leave-one-out estimators $\hat{m}_{n-1}^{(-i)}(\cdot)$. From the manuscript (Formula 4.5) we learn:

$$n^{-1} \sum_{i=1}^n \left(Y_i - \hat{m}_{n-1}^{(-i)}(X_i) \right)^2 = n^{-1} \sum_{i=1}^n \left(\frac{Y_i - \hat{m}(X_i)}{1 - S_{ii}} \right)^2,$$

where S is the *hat-matrix* of the linear estimator $\hat{m}(\cdot)$. In this exercise we are going to prove this formula step by step in the case of multiple-linear-regression $y_i = \mathbf{x}_i^T \beta + \epsilon_i$.

- a) Show that for an invertible $p \times p$ -matrix A and two p -vectors \mathbf{a} and \mathbf{b} with $\mathbf{b}^T A^{-1} \mathbf{a} \neq 1$ the matrix $A - \mathbf{a} \mathbf{b}^T$ is invertible too and that its inverse can be computed as follows:

$$(A - \mathbf{a} \mathbf{b}^T)^{-1} = A^{-1} + \frac{1}{1 - \mathbf{b}^T A^{-1} \mathbf{a}} \cdot A^{-1} \mathbf{a} \mathbf{b}^T A^{-1}.$$

- b) Show the following formula which describes the influence of omitting the i^{th} observation for the multiple-linear-regression estimator:

$$\hat{\beta}^{(-i)} - \hat{\beta} = -\frac{y_i - \mathbf{x}_i^T \hat{\beta}}{1 - S_{ii}} (X^T X)^{-1} \mathbf{x}_i.$$

Hints: Let $A := X^T X = \sum_{i=1}^n \mathbf{x}_i \mathbf{x}_i^T$, $\mathbf{c} := X^T y = \sum_{i=1}^n y_i \mathbf{x}_i$.

Now you might start as follows: $\hat{\beta}^{(-i)} = (A - \mathbf{x}_i \mathbf{x}_i^T)^{-1} (\mathbf{c} - y_i \mathbf{x}_i)$, then use **a**).

- c) From **b**) you can finally conclude the desired result:

$$y_i - \mathbf{x}_i^T \hat{\beta}^{(-i)} = \frac{1}{1 - S_{ii}} (y_i - \mathbf{x}_i^T \hat{\beta}).$$

2. Consider the **diabetes**-dataset from the lecture notes (Section 3.2) and the model

$$Y_i = m(X_i) + \epsilon_i,$$

where the response Y is a log-concentration of a serum (in connection with diabetes) and the predictor variable X is the age in months of children.

We want to know if a complicated nonparametric regression gives us valuable information and which one is the best. The following fits should be compared:

1. the kernel regression fit from `ksmooth`,
2. the local polynomial fit from `loess`,
3. a smoothing spline fit from `smooth.spline`, where you have chosen a fixed value for the parameter `df` in advance,
4. a smoothing spline fit from `smooth.spline` with the smoothing parameter selected automatically by cross-validation (consider the “Details”-section of `help(smooth.spline)` and the description of parameter `cv` and value `cv.crit`).
5. a constant “fit” by the overall mean of Y_i , simply ignoring the X_i -values.

- a) Perform a non-parametric regression on the `diabetes`-dataset using the five different methods. Calculate the corresponding leave-one-out CV-values and compare them. Make sure to use similar degrees of freedom in all methods. Note that the CV-value is calculated internally by the function `smooth.spline`, but not by the others; for this function, you may also compare the internally calculated value with your own calculation. For `smooth.spline`, you may take advantage of Formula (4.5) in the lecture notes; cf. Exercise 1 of this series.
- b) The comparison of the CV-value of method no. 4 with the others is not fair. Can you explain the problem?

R-hints: You may begin as follows:

```
diabetes <-  
  read.table("http://stat.ethz.ch/Teaching/Datasets/diabetes2.dat",  
            header = TRUE)  
reg <- diabetes[, c("Age", "C.Peptide")]  
names(reg) <- c("x", "y")  
## Sort the observations along x, for easier dealing of hat matrix:  
reg <- reg[sort.list(reg$x), ]
```

Then, `cdat <- reg[-i,]` is `reg` without point i .

Preliminary discussion: Friday, April 01.

Deadline: Friday, April 08.