Assignment sheet 11

November 28, 2018

Assignment 1 (efficient computation of Cook's distance)**.** We have seen a measure of the influence of the k-th observation over the regression coefficient. This measure, *Cook's distance*, is defined as

$$C_k = \frac{1}{ps^2}\|\hat{y} - \hat{y}_{-k}\|^2,$$

where $\hat{y}_{-k} = X\hat{\beta}_{-k}$ and $\hat{\beta}_{-k}$ is the estimator of β without the k-th observation. It seems like one would need $n+1$ regressions in order to calculate $C_1, \ldots, C_n$. We shall see that one can get the C_k's *using only the complete regression on* (y, X) by means of the formula

$$C_k = \frac{r_k^2 h_{kk}}{p(1 - h_{kk})}, \tag{1}$$

where r_k is the k-standardised residual and h_{kk} is the k-th diagonal element of the hat matrix $H = X(X^TX)^{-1}X^T$.
Let x_k^T be the k-th row of X, so that $x_k \in \mathbb{R}^p$ and

$$X^T = (x_1, \ldots, x_n)_{p\times n}.$$

Denote X_{-k} the $n \times p$ matrix whose l-th row is x_l^T if $l \neq k$ and whose kth row is $0 \in \mathbb{R}^p$. In symbols

$$X_{-k}^T = (x_1, \ldots, x_{k-1}, 0, x_{k+1}, \ldots, x_n).$$

In this exercise, you can use the identity

$$(x_1, \ldots, x_n)\begin{pmatrix} z_1^T \\ \vdots \\ z_n^T \end{pmatrix} = \sum_{i=1}^{n} x_i z_i^T \in \mathbb{R}^{p\times q},$$

where $x_i \in \mathbb{R}^p, z_i \in \mathbb{R}^q, i = 1, \ldots, n.$
Moreover, for compatible matrices A, B and C,

$$\mathrm{row}_j(AB) = \mathrm{row}_j(A)\cdot B,$$

$$\mathrm{col}_k(AB) = A\cdot\ \mathrm{col}_k(B)$$

$$(ACB)_{j,k} = \mathrm{row}_j(A)\cdot C\cdot \mathrm{col}_k(B),$$

where $\mathrm{row}_j(A)$ represents the j-th row of A, *as a row (rather than column) vector*, $\mathrm{col}_k(B)$ represnts the k-th column of B, as a column vector, and "$\cdot$" is the usual matrix product.

(i). Show that $X_{-k}^T X_{-k} = X^TX - x_k x_k^T$.

(ii). (a) Show the Sherman–Morrison formula

$$\left(A + uv^T\right)^{-1} = A^{-1} - \frac{A^{-1}uv^TA^{-1}}{1 + v^TA^{-1}u},$$

where $A_{n\times n}$ is invertible and $u, v \in \mathbb{R}^n$ satisfy $v^TA^{-1}u \neq -1$.

(b) Deduce that

$$(X_{-k}^T X_{-k})^{-1} = \left(I + \frac{1}{1-h_{kk}} \left(X^T X\right)^{-1} x_k x_k^T\right) \left(X^T X\right)^{-1}.$$

(iii). Show that

(a) $X_{-k}^T y = X^T y - y_k x_k$,

(b) $x_k^T (X^T X)^{-1} X_{-k}^T y = (1-h_{kk}) y_k - e_k$,

and conclude that

$$\hat\beta_{-k} = \hat\beta - \frac{e_k \left(X^T X\right)^{-1} x_k}{1-h_{kk}}.$$

(iv). Lastly, show that $\|\hat y - \hat y_{-k}\|^2 = h_{kk} e_k^2/(1-h_{kk})^2$, and conclude (1).

Assignment 2. Consider the cement data ($n = 13$). The residual sum of squares (RSS) for all the models containing the intercept are given below.

model	RSS	model	RSS	model	RSS
− − − −	2715.8	1 2 − −	57.9	1 2 3 −	48.11
1 − − −	1265.7	1 − 3 −	1227.1	1 2 − 4	47.97
− 2 − −	906.3	1 − − 4	74.8	1 − 3 4	50.84
− − 3 −	1939.4	− 2 3 −	415.4	− 2 3 4	73.81
− − − 4	883.9	− 2 − 4	868.9		
		− − 3 4	175.7	1 2 3 4	47.86

Calculate the analysis of variance table when adding x_4, x_3, x_2 and x_1 to the model in this order and test which terms should be included in the model at significance level $\alpha = 0.05$. Are the conclusions the same as in slide 407 ?

Assignment 3 (automatic model selection)**.** Consider again the cement data from the course. The residual sum of squares (RSS) as well (some of !) the values of Mallows' C_p for the models containing the intercept are as follows :

model	RSS	C_p	model	RSS	C_p	model	RSS	C_p
- - - -	2715.8	442.58	1 2 - -	57.9		1 2 3 -	48.1	
			1 - 3 -	1227.1	197.94	1 2 - 4	48.0	
1 - - -	1265.7	202.39	1 - - 4	74.8	5.49	1 - 3 4	50.8	
- 2 - -	906.3		- 2 3 -	415.4	62.38	- 2 3 4	73.8	7.325
- - 3 -	1939.4	314.90	- 2 - 4	868.9	138.12			
- - - 4	883.9	138.62	- - 3 4	175.7	22.34	1 2 3 4	47.9	5

a) Use *forward selection* and *backward elimination* to choose a model for the data. Include significant variable at 5% using the F-test

$$F = \frac{\mathrm{RSS}(\hat\beta_{\mathrm{L}}) - \mathrm{RSS}(\hat\beta_{\mathrm{L}\cup\{\mathrm{j}\}})}{\mathrm{RSS}(\hat\beta_{\mathrm{full}})/(13-5)}$$

in order to decide whether the j-th variable is significant.

b) Mallows C_p is defined as (see slide 423)

$$C_p = \frac{\mathrm{RSS}_p}{s^2} + 2p - n.$$

Note that s^2 is the estimator of the variance σ^2 *under the full model.*

i) Calculate the missing values of C_p in the table, and explain how one uses this criterion for model selection.

ii) Which models would be chosen by *forward selection*, *backward elimination*, and Mallows' C_p ? Are the three models same ?

Assignment 4 (AIC and Gaussian linear models)**.**
Show that the AIC criterion for a gaussian linea model and a response vector of size n with p covariates can be written as

$$\mathrm{AIC} = n \log \hat{\sigma}^2 + 2p + \mathrm{const},$$

where σ^2 is the unknown variance of the model and $\hat{\sigma}^2 = RSS_p/n$ is the MLE estimator for σ^2.

Assignment 5 (Cross validation and number of parameters)**.**
Using the fact that

$$\hat{\beta}_{-j} = \hat{\beta} - \frac{(y_j - \hat{y}_j)(X^t X)^{-1} x_j}{1 - h_{jj}},$$

show that

$$\mathrm{CV} = \sum_{j=1}^{n} (y_j - x_j^t \hat{\beta}_{-j})^2 \tag{2}$$

can be written as

$$\mathrm{CV} = \sum_{j=1}^{n} \frac{(y_j - x_j^t \hat{\beta})^2}{(1 - h_{jj})^2}. \tag{3}$$

What is the advantage of using the formula (3) over the formula (2) ?