SF 2930 Regression Analysis

Lecture 10

Diagnostics & Validation : PRESS, Cook's Distance

Timo Koski

KTH Royal Institute of Technology

2023

Learning outcomes

- Crossvalidation
- Crossvalidation by PRESS =predicted residual error sum of squares
  - LOOCV = leave-one-out-crossvalidation
  - Sherman-Morrison-Woodbury (S-M-W) theorem and formula : the problem of determining the change in the inverse matrix after a change in a column.
- Expectation of PRESS
- PRESS in Model Choice, Hjorth's Theorem.
- Cook's Distance for Detecting Data Points that are Influential for LSE $\widehat{\beta}$.
- Recursive Multiple Regression

Crossvalidation

Cross-validation or out-of-sample testing, is any of several similar model validation techniques for assessing how the results of a statistical analysis will generalize to an independent data set.
Cross-validation is a resampling method that uses different portions of the data to test and train a model on different iterations.
It is mainly used in settings where the goal is prediction, and one wants to estimate how accurately a predictive model will perform in practice.

Crossvalidation

The calculations of a summary measure of the fit of a model on the test set, a sample of observations that were not themselves used to estimate the model. This is asimple idea and easy to implement. But it has a few potential drawbacks

- Test error rate can be highly variable, depending on precisely which observations are included in the training set and which observations are included in the test set.
- Statistical methods tend to perform worse when trained on fewer observations (which is a result splitting data into tow sets).

Methods of crossvalidation (CV) are developed to overcome this problem, and do model validation by resampling without splitting the data.

LOOCV= Leave One Out Crossvalidation

The predicted residual error sum of squares (PRESS) is one of techniques of cross-validation used in linear regression analysis. A fitted model having been produced, each observation in turn is removed and the model is refitted using the remaining observations. The out-of-sample predicted value is calculated for the omitted observation in each case, and the PRESS statistic is calculated as the sum of the squares of all the resulting prediction errors.

Crossvalidation by PRESS

The training set is

$$\mathcal{D}_{tr} = \left\{ \left(\mathbf{x}_i^\mathsf{T}, y_i \right)_{j=1}^{n} \right\}.$$

We remove $(\mathbf{x}_i^\mathsf{T}, y_i)$ and the new training set is

$$\mathcal{D}_{tr,(i)} = \left\{ \left(\mathbf{x}_i^\mathsf{T}, y_i \right)_{j=1, j\neq i}^{n} \right\}$$

One finds the LSE $\widehat{\boldsymbol{\beta}}_{(i)}$ using $\mathcal{D}_{tr,(i)}$. Then the predictor of y_i is $\hat{y}_{(i)} = \mathbf{x}_{(i)}^\mathsf{T} \widehat{\boldsymbol{\beta}}_{(i)}$. We repeat this is for all $(\mathbf{x}_i^\mathsf{T}, y_i)$, $i = 1, \ldots, n$. The PRESS statistic is

$$\text{PRESS} = \sum_{i=1}^{n} (y_i - \hat{y}_{(i)})^2$$

Crossvalidation by PRESS. model choice

The PRESS statistic is

$$\text{PRESS} = \sum_{i=1}^{n} (y_i - \hat{y}_{(i)})^2$$

Given this procedure, the PRESS statistic can be calculated for a number of candidate model structures for the same dataset, with the lowest values of PRESS indicating the best structures.

Models that are over-parameterised (over-fitted) would tend to give small residuals for observations included in the model-fitting but large residuals for observations that are excluded.

Crossvalidation by PRESS

This is straighforward as an idea, but is this practical, as it seems that we must fit n multiple regressions to data.

However, there is a short way: only one LSE based on $\mathcal{D}_{tr}$ is needed.

Refresher

$$\mathbf{Y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

$$\mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_i \\ \vdots \\ y_n \end{pmatrix}, X = \begin{pmatrix} \mathbf{x}_1^\mathsf{T} \\ \mathbf{x}_2^\mathsf{T} \\ \vdots \\ \mathbf{x}_i^T \\ \vdots \\ \mathbf{x}_n^\mathsf{T} \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & \cdots & x_{1k} \\ 1 & x_{21} & \cdots & x_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{i1} & \cdots & x_{ik} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \cdots & x_{nk} \end{pmatrix}$$

$$\boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \end{pmatrix}, \quad \boldsymbol{\varepsilon} = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}.$$

LOOCV: Remove $(\mathbf{x}_i^T, y_i)$

$$\mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ \cancel{y_i} \\ \vdots \\ y_n \end{pmatrix}, X = \begin{pmatrix} \mathbf{x}_1^T \\ \mathbf{x}_2^T \\ \vdots \\ \cancel{\mathbf{x}_i^T} \\ \vdots \\ \mathbf{x}_n^T \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & \cdots & x_{1k} \\ 1 & x_{21} & \cdots & x_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ \cancel{1} & \cancel{x_{i1}} & \cancel{\cdots} & \cancel{x_{ik}} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \cdots & x_{nk} \end{pmatrix}$$

The (Ordinary) Multiple Linear Regression Model

$\beta \in \mathbb{R}^{k+1}$ and $n > k + 1$.

$$\mathbf{Y} = X\beta + \varepsilon. \tag{1}$$

The following assumptions hold :

1) $E[\varepsilon] = \mathbf{0} \in \mathbb{R}^n$
2) $C_\varepsilon = E\left[\varepsilon\varepsilon^T\right] = \sigma^2 \mathbb{I}_n$ (homoscedasticity)
3) X has full column rank

LSE

$$\mathcal{D}_{tr} = \{\left(\mathbf{x}_i^{\mathsf{T}}, y_i\right)_{i=1}^{n}\}$$

We use the training set to compute

$$\widehat{\boldsymbol{\beta}} = (X^T X)^{-1} X^T \mathbf{y} = \mathbf{argmin}_{\boldsymbol{\beta}} \parallel \mathbf{y} - X\boldsymbol{\beta} \parallel^2$$

Hat Matrix

$$H = X(X^T X)^{-1} X^T. \tag{2}$$

The predicted values are $\widehat{\mathbf{y}} = X\widehat{\boldsymbol{\beta}} = X(X^T X)^{-1} X^T \mathbf{y} = H\mathbf{y}$.
The elements on the main diagonal of H, h_{ii}, are called the leverages of the respective data point.
The $n \times 1$ random vector $\widehat{\varepsilon}$ of observed residuals

$$\widehat{\varepsilon} = \mathbf{Y} - X\widehat{\boldsymbol{\beta}} = \mathbf{Y} - H\mathbf{Y}. \tag{3}$$

We have found that the covariance matrix of $\widehat{\varepsilon}$ is

$$C_{\widehat{\varepsilon}} = \sigma^2 (\mathbb{I}_n - H) \tag{4}$$

The Short Way

$$\text{PRESS} = \sum_{i=1}^{n} (y_i - \hat{y}_{(i)})^2$$

We shall establish that

$$\hat{y}_i = h_{ii} y_i + (1 - h_{ii}) \hat{y}_{(i)}, \tag{5}$$

where h_{ii} is the element on the main diagonal of H.

The Short Way

From (5)

$$y_i - \hat{y}_i = (1 - h_{ii})y_i - (1 - h_{ii})\hat{y}_{(i)},$$

which gives

$$\frac{y_i - \hat{y}_i}{(1 - h_{ii})} = y_i - \hat{y}_{(i)}$$

Hence

$$\text{PRESS} = \sum_{i=1}^{n} \left(\frac{y_i - \hat{y}_i}{1 - h_{ii}}. \right)^2. \tag{6}$$

This requires, of course, that $h_{ii} < 1$, which is checked in an Appendix. This is a consequence of idempotency of H. MPV on p. 134, p. 151, p. 592 does not pay attention to this detail.

Exam Generator

Mahalanobis distance d_M^2 between $\mathbf{x}_i$ and the arithmetic mean $\bar{\mathbf{x}}$ of the rows of X (the first 1 excluded). S is the empirical covariance matrix of $\mathbf{x}_i$.

$$d_M^2(\mathbf{x}_i, \bar{\mathbf{x}}; S) = (\mathbf{x}_i - \bar{\mathbf{x}})^T S^{-1} (\mathbf{x} - \bar{\mathbf{x}}).$$

The (possible) exam question is to show that

$$(n-1)\left(h_{ii} - \frac{1}{n}\right) = d_M^2(\mathbf{x}_i, \bar{\mathbf{x}}; S).$$

Thus

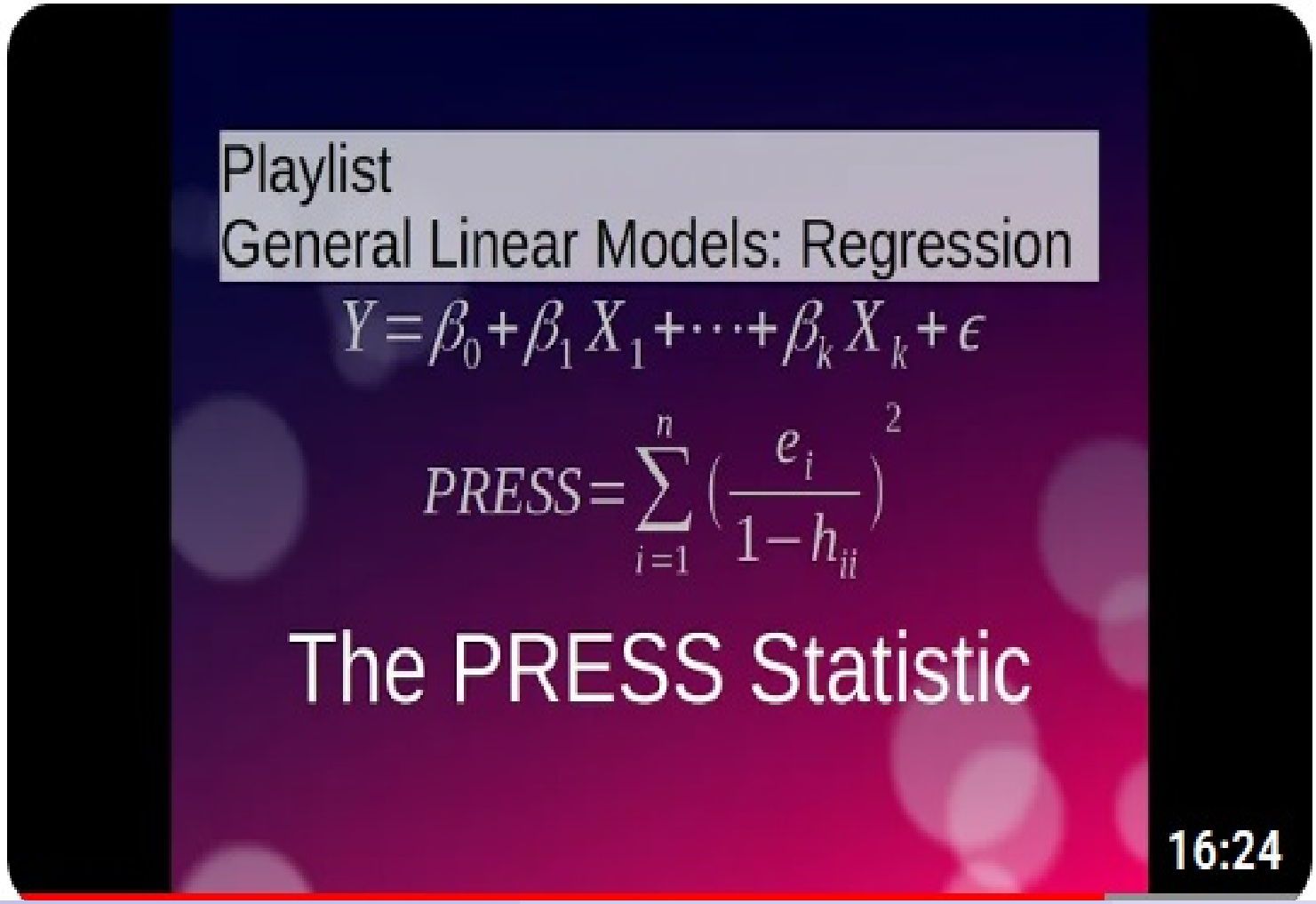
$$h_{ii} = \frac{1}{n-1} d_M^2(\mathbf{x}_i, \bar{\mathbf{x}}; S) + \frac{1}{n}$$

and then the PRESS statistic becomes

$$\text{PRESS} = \sum_{i=1}^{n} \left(\frac{y_i - \hat{y}_i}{1 - \frac{1}{n-1} d_M^2(\mathbf{x}_i, \bar{\mathbf{x}}; S) - \frac{1}{n}} \right)^2 \tag{7}$$

YouTube

https://www.youtube.com/watch?v=SBYO0dPbA



We prove

$$\hat{y}_i = h_{ii} y_i + (1 - h_{ii}) \hat{y}_{(i)}, \tag{8}$$

by deriving the following.

$$\widehat{\boldsymbol{\beta}}_{(i)} = \widehat{\boldsymbol{\beta}} + \frac{(\widehat{y}_i - y_i)}{1 - h_{ii}} (X^T X)^{-1} \mathbf{x}_i. \tag{9}$$

Simple Steps on the Short Way after Removing $(\mathbf{x}_i^\mathsf{T}, y_i)$: Step One

$$X^T X = \begin{pmatrix} \mathbf{x}_1^\mathsf{T} \\ \mathbf{x}_2^\mathsf{T} \\ \vdots \\ \mathbf{x}_n^\mathsf{T} \end{pmatrix}^T \begin{pmatrix} \mathbf{x}_1^\mathsf{T} \\ \mathbf{x}_2^\mathsf{T} \\ \vdots \\ \mathbf{x}_n^\mathsf{T} \end{pmatrix} = (\mathbf{x}_1 \mathbf{x}_2 \dots \mathbf{x}_n) \begin{pmatrix} \mathbf{x}_1^\mathsf{T} \\ \mathbf{x}_2^\mathsf{T} \\ \vdots \\ \mathbf{x}_n^\mathsf{T} \end{pmatrix} = \sum_{j=1}^{n} \mathbf{x}_j \mathbf{x}_j^\mathsf{T}$$

$$= \sum_{j=1, j\neq i}^{n} \mathbf{x}_j \mathbf{x}_j^\mathsf{T} + \mathbf{x}_i \mathbf{x}_i^\mathsf{T} = X_{(i)}^T X_{(i)} + \mathbf{x}_i \mathbf{x}_i^\mathsf{T} \tag{10}$$

$$X^T \mathbf{y} = (\mathbf{x}_1 \mathbf{x}_2 \dots \mathbf{x}_n) \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = X_{(i)}^T \mathbf{y}_{(i)} + \mathbf{x}_i y_i \tag{11}$$

: Step Two

From (11)

$$\widehat{\boldsymbol{\beta}} = (X^T X)^{-1} X^T \mathbf{y} = (X^T X)^{-1} X_{(i)}^T \mathbf{y}_{(i)} + (X^T X)^{-1} \mathbf{x}_i y_i \tag{12}$$

From (10)

$$\widehat{\boldsymbol{\beta}}_{(i)} = (X_{(i)}^T X_{(i)})^{-1} X_{(i)}^T \mathbf{y} = \left(X^T X - \mathbf{x}_i \mathbf{x}_i^\mathsf{T} \right)^{-1} X_{(i)}^T y_i. \tag{13}$$

This is the ith regression due to removal of $(\mathbf{x}_i^\mathsf{T}, y_i)$, but will not be computed along the short way.

Step 3

Sherman-Morrison-Woodbury (S-M-W) theorem and formula : the problem of determining the change in the inverse matrix after a change in a column. In the current situation we are determining the change in the inverse matrix $X^T X$, when one row has been removed in X.

S-M-W

Suppose A is is an invertible square $n \times n$ matrix and $\mathbf{u}, \mathbf{v} \in \mathbb{R}^n$ are column vectors. Then $A + \mathbf{u}\mathbf{v}^\mathsf{T}$ is invertible iff $1 + \mathbf{v}^\mathsf{T} A^{-1}\mathbf{u} \neq 0$. In this case,

$$\left(A + \mathbf{u}\mathbf{v}^\mathsf{T}\right)^{-1} = A^{-1} - \frac{A^{-1}\mathbf{u}\mathbf{v}^\mathsf{T}A^{-1}}{1 + \mathbf{v}^\mathsf{T}A^{-1}\mathbf{u}}. \tag{14}$$

We apply this with $A = X^T X$, $\mathbf{u}\mathbf{v}^\mathsf{T} = -\mathbf{x}_i\mathbf{x}_i^\mathsf{T}$.

Sherman-Morrison-Woodbury Theorem

We apply S-M-W Formula (14) with $A = X^T X$, $\mathbf{u}\mathbf{v}^\mathsf{T} = -\mathbf{x}_i \mathbf{x}_i^\mathsf{T}$. We note that $X^T X$ is invertible, since $X_{(i)}$ has full column rank, a standing assumption under $n > k + 1$. Then we must check that $1 + \mathbf{v}^\mathsf{T} A^{-1} \mathbf{u} \neq 0$, i.e.,

$$1 - \mathbf{x}_i^\mathsf{T} (X^T X)^{-1} \mathbf{x}_i \neq 0.$$

This detail is omitted/overlooked in MVP p. 591. However, $h_{ii} = \mathbf{x}_i^\mathsf{T} X^T X \mathbf{x}_i$ is the ith element on the main diagonal of the hat matrix

$$H = X(X^T X)^{-1} X^T$$

It is shown in an Appendix that $h_{ii} < 1$.

Sherman-Morrison-Woodbury Theorem

$$\left(X^T X - \mathbf{x}_i \mathbf{x}_i^\intercal\right)^{-1} X_{(i)}^T y_i = \left(X^T X\right)^{-1} X_{(i)}^T y_i$$
$$- \frac{\left(X^T X\right)^{-1} \mathbf{x}_i \mathbf{x}_i^\intercal \left(X^T X\right)^{-1}}{1 - h_{ii}} X_{(i)}^T y_i.$$

Thus

$$\widehat{\boldsymbol{\beta}}_{(i)} = \underbrace{\left(X^T X\right)^{-1} X_{(i)}^T y_i}_{\equiv A} - \underbrace{\frac{\left(X^T X\right)^{-1} \mathbf{x}_i \mathbf{x}_i^\intercal \left(X^T X\right)^{-1}}{1 - h_{ii}} X_{(i)}^T y_i}_{\equiv B}. \tag{15}$$

Now we re-write the right hand side of (15) by means of (12), that is, by insertion of

$$A = (X^T X)^{-1} X_{(i)}^T \mathbf{y}_{(i)} = \widehat{\boldsymbol{\beta}} - (X^T X)^{-1} \mathbf{x}_i y_i. \tag{16}$$

We substitute A in B to get

$$B = \frac{\left(X^T X\right)^{-1} \mathbf{x}_i \mathbf{x}_i^{\mathsf{T}} \left(X^T X\right)^{-1}}{1 - h_{ii}} X_{(i)}^T y_i = \frac{\left(X^T X\right)^{-1} \mathbf{x}_i \mathbf{x}_i^{\mathsf{T}} \left(\widehat{\boldsymbol{\beta}} - (X^T X)^{-1} \mathbf{x}_i y_i\right)}{1 - h_{ii}}$$

Here

$$\left(X^T X\right)^{-1} \mathbf{x}_i \mathbf{x}_i^{\mathsf{T}} \left(\widehat{\boldsymbol{\beta}} - (X^T X)^{-1} \mathbf{x}_i y_i\right) = \left(X^T X\right)^{-1} \mathbf{x}_i \underbrace{\mathbf{x}_i^{\mathsf{T}} \widehat{\boldsymbol{\beta}}}_{=\widehat{y}_i}$$

$$- \left(X^T X\right)^{-1} \mathbf{x}_i \underbrace{\mathbf{x}_i^{\mathsf{T}} (X^T X)^{-1} \mathbf{x}_i}_{=h_{ii}} y_i$$

Hence

$$B = \frac{\left(X^T X\right)^{-1} \mathbf{x}_i \left(\widehat{y}_i - h_{ii} y_i\right)}{1 - h_{ii}}. \tag{17}$$

Final Steps

When we insert (16) and (17) in (15) in we get

$$\begin{aligned}\widehat{\boldsymbol{\beta}}_{(i)} &= A - B \\ &= \widehat{\boldsymbol{\beta}} - (X^T X)^{-1}\mathbf{x}_i y_i \\ &\quad - \frac{(X^T X)^{-1}\mathbf{x}_i\left(\widehat{y}_i - h_{ii}y_i\right)}{1-h_{ii}}\end{aligned}$$

and this equals

$$= \widehat{\boldsymbol{\beta}} + \frac{(X^T X)^{-1}\mathbf{x}_i\widehat{y}_i}{1-h_{ii}} + \frac{-(X^T X)^{-1}\mathbf{x}_i(1-h_{ii})y_i - (X^T X)^{-1}\mathbf{x}_i h_{ii}y_i}{1-h_{ii}}$$

$$= \widehat{\boldsymbol{\beta}} + \frac{(X^T X)^{-1}\mathbf{x}_i\widehat{y}_i}{1-h_{ii}} - \frac{(X^T X)^{-1}\mathbf{x}_i y_i}{1-h_{ii}}$$

$$= \widehat{\boldsymbol{\beta}} + \frac{(\widehat{y}_i - y_i)}{1-h_{ii}}(X^T X)^{-1}\mathbf{x}_i.$$

We have thus shown how to compute $\widehat{\boldsymbol{\beta}}_{(i)}$ by of one LSE.

$$\widehat{\boldsymbol{\beta}}_{(i)} = \widehat{\boldsymbol{\beta}} + \frac{(\widehat{y}_i - y_i)}{1 - h_{ii}}(X^T X)^{-1}\mathbf{x}_i. \tag{18}$$

Next we establish (8).

$$\hat{y}_{(i)} = \mathbf{x}_i^T \widehat{\boldsymbol{\beta}}_{(i)}.$$

We use (18) to find

$$\hat{y}_{(i)} = \mathbf{x}_i^T \widehat{\boldsymbol{\beta}} + \frac{(\widehat{y}_i - y_i)}{1 - h_{ii}} \mathbf{x}_i^T (X^T X)^{-1} \mathbf{x}_i.$$

Here $\mathbf{x}_i^T \widehat{\boldsymbol{\beta}} = \widehat{y}_i$ and $h_{ii} = \mathbf{x}_i^T (X^T X)^{-1} \mathbf{x}_i$. Hence

$$\hat{y}_{(i)} = \widehat{y}_i + \frac{h_{ii} (\widehat{y}_i - y_i)}{1 - h_{ii}}$$

and

$$(1 - h_{ii}) \hat{y}_{(i)} = (1 - h_{ii}) \widehat{y}_i + h_{ii} (\widehat{y}_i - y_i).$$

After little simplification this becomes (8).

Properties of PRESS

In view of (3) we have

$$\text{PRESS} = \sum_{i=1}^{n} \left(\frac{y_i - \hat{y}_i}{1 - h_{ii}}. \right)^2 = \sum_{i=1}^{n} \frac{\widehat{\epsilon}_i^2}{(1 - h_{ii})^2}$$

Therefore

$$E\left[\text{PRESS}\right] = \sum_{i=1}^{n} \frac{E\left[\widehat{\epsilon}_i^2\right]}{(1 - h_{ii})^2}$$

By (4) we have that $E\left[\widehat{\epsilon}_i^2\right] = \sigma^2(1 - h_{ii})$. Hence

$$E\left[\text{PRESS}\right] = \sigma^2 \sum_{i=1}^{n} \frac{1}{(1 - h_{ii})^2}$$

$$\text{PRESS} = \sum_{i=1}^{n} \frac{\widehat{\epsilon}_i^2}{(1 - h_{ii})^2}$$

Interpretation as follows: for a fair estimate of the error in predicting y_i, we should inflate the square training error $\widehat{\epsilon}_i^2$ by a factor of $\frac{1}{(1 - h_{ii})^2}$. Hence, the closer the value of h_{ii} to1, the more we should inflate the training error. This makes sense when we recall (see Exam questions) that h_{ii} is the leverage of the ith data point, which is a measure of how much our prediction $\hat{y}_i$ depends on the actual value y_i. The more $haty_i$ depends on y_i, the more $\hat{y}_i$ is an overfit, and the greater we should inflate the error squared error $\widehat{\epsilon}_i^2$.

Cook's distance

Data points with large residuals (outliers) and/or high leverage may distort the outcome and accuracy of a regression. Cook's distance, denoted by D_i, measures the effect of deleting a given observation. Hence this involves an application of (18). Points with a large Cook's distance are considered to merit closer examination in the analysis. Cook (1977) introduced D_i by

$$D_i := \frac{\left(\widehat{\boldsymbol{\beta}}_{(i)} - \widehat{\boldsymbol{\beta}}\right)^T X^T X \left(\widehat{\boldsymbol{\beta}}_{(i)} - \widehat{\boldsymbol{\beta}}\right)}{(k+1)\widehat{\sigma}^2} \tag{19}$$

Here we have the familiar $\widehat{\sigma}^2 = \frac{\mathbf{e}^T\mathbf{e}}{n-k-1}$. Next Cook substitutes for $\widehat{\boldsymbol{\beta}}_{(i)} - \widehat{\boldsymbol{\beta}}$ by (18) and obtains

$$D_i = \left(\frac{(\widehat{y}_i - y_i)}{\sqrt{(1-h_{ii})}\widehat{\sigma}}\right)^2 \frac{\mathbf{x}_i^T (X^T X)^{-1} \mathbf{x}_i}{(k+1)(1-h_{ii})}$$

Cook's distance

But, see slides 25-29 of Lecture 4 or MPV p. 99, $\widehat{y}_i = \mathbf{x}_i^{\mathsf{T}}\widehat{\boldsymbol{\beta}}$ is the minimal MSE predictor of y_i and that $\mathbf{Var}\left[\widehat{y}_i\right] = \sigma^2\mathbf{x}_i^{\mathsf{T}}\left(X^TX\right)^{-1}\mathbf{x}_i$. In addition, by (4), $\mathbf{Var}\left[\widehat{\varepsilon}_i\right] = \sigma^2 = (1 - h_{ii}) = (1 - \mathbf{x}_i^{\mathsf{T}}\left(X^TX\right)^{-1}\mathbf{x}_i)$. Thus we write

$$D_i = \left(\frac{(\widehat{y}_i - y_i)}{\sqrt{(1-h_{ii})}\widehat{\sigma}}\right)^2 \frac{\mathbf{x}_i^T(X^TX)^{-1}\mathbf{x}_i}{(k+1)(1-h_{ii})} = \frac{t_i^2}{(k+1)}\frac{\mathbf{Var}\left[\widehat{y}_i\right]}{\mathbf{Var}\left[\widehat{\varepsilon}_i\right]}$$

where

$$t_i = \frac{(\widehat{y}_i - y_i)}{\sqrt{(1-h_{ii})}}, \quad i = 2, \ldots, n$$

Note that $\mathbf{Var}[t_i] = 1$ for every i.

Cook's distance

$$D_i = \frac{t_i^2}{(k+1)} \tag{20}$$

- t_i^2 is a measure of the degree to which the ith observation can be considered as an outlier from the assumed model. Any point i with with a large $\widehat{y}_i - y_i$ and a large h_{ii} is potentially highly influential on the least squares fit.
- The ratios $\mathbf{Var}\left[\widehat{y}_i\right] / \mathbf{Var}\left[\widehat{\varepsilon}_i\right]$ measure the relative sensitivity of the estimate, $\widehat{\boldsymbol{\beta}}$, to potential outlying values at each data point.
- The use of D_i is possible only if $\mathbf{Var}\left[\widehat{\varepsilon}_i\right] > 0$.

Cook's Distance

Cooke (1977) discusses a regression analysis relating six economic variables to total derived employment for the years 1947 to 1962 in USA. Cook points out first that there are considerable differences between residuals e_i and t_i. Second, the point with the largest D_i value corresponds to 1951. In the table $R_i/s \leftrightarrow \widehat{\varepsilon}_i/\widehat{\sigma}$, t_i, $V(Y_i)/V(R_i)) \leftrightarrow \mathbf{Var}\left[\widehat{y}_i\right]/\mathbf{Var}\left[\widehat{\varepsilon}_i\right]$.

Cook's Distance

TABLE I—*Longley Data*

Year	R_i/s	$\lvert t_i \rvert$	$V(\hat{Y}_i)/V(R_i)$	D_i
1947	0.88	1.15	0.74	0.14
48	-0.31	0.48	1.30	0.04
49	0.15	0.19	0.57	*
50	-1.34	1.70	0.59	0.24
51	1.02	1.64	1.60	0.61
52	-0.82	1.03	0.59	0.09
53	-0.54	0.75	0.97	0.08
54	-0.04	0.06	1.02	*
55	0.05	0.07	0.84	*
56	1.48	1.83	0.49	0.23
57	-0.06	0.07	0.56	*
58	-0.13	0.18	0.93	*
59	-0.51	0.64	0.60	0.04
60	-0.28	0.32	0.30	*
61	1.12	1.42	0.59	0.17
62	-0.68	1.21	2.21	0.47

*: smaller than 5×10^{-3}

TECHNOMETRICS©, VOL. 19, NO. 1, FEBRUARY 1977

Studentized Residuals

The quantities in Cook's Distance

$$t_i = \frac{(\widehat{y}_i - y_i)}{\sqrt{(1-h_{ii})}} = \frac{(\widehat{\epsilon}_i)}{\sqrt{(1-h_{ii})}}, \quad i = 2, \ldots, n$$

are a form of a Student's t-statistic, are therefore known as the studentized residuals, see MVP p. 133.
The key reason for studentizing in regression analysis is that the variances of the residuals at different data points may differ, even if the variances of the errors at these data points are equal.

$$\mathbf{Var}\,[\epsilon_i] = \sigma^2, \quad \mathbf{Var}\,[\widehat{\epsilon}_i] = (1-h_{ii}),$$

where h_{ii} is the leverage of ith data point.

PRESS and Model Choice

Models $\mathcal{M}_k$, $k = 1, 2, \ldots, M$.

$$\text{PRESS}\left(\mathcal{M}_k\right) = \sum_{i=1}^{n} \left(\frac{y_i - \hat{y}_i\left(\mathcal{M}_k\right)}{1 - h_{ii}\left(\mathcal{M}_k\right)} \right)^2$$

Choose $\mathcal{M}_{k_*}$, if

$$k_* = \mathbf{argmin}_{k \in \{1,\ldots,M\}} \text{PRESS}\left(\mathcal{M}_k\right)$$

Urban Hjorth's Bias Theorem

Proposition

Models $\mathcal{M}_k$, $k = 1, 2, \ldots, M$.

$$\mathcal{M}_{k_*} = \min_{k \in \{1,\ldots,M\}} \text{PRESS}\left(\mathcal{M}_k\right)$$

Then

$$E\left[\text{PRESS}\left(\mathcal{M}_{k_*}\right)\right] \leq \min_{k \in \{1,\ldots,M\}} E\left[\text{PRESS}\left(\mathcal{M}_k\right)\right]. \tag{21}$$

Proof: Define the r.v. $X_k := \text{PRESS}\left(\mathcal{M}_k\right) - \textit{PRESS}\left(\mathcal{M}_{k_*}\right)$ for every outcome of $X, \mathbf{Y}$. Then $X_k \geq 0$. A result in probabilty calculus says that

$$\text{If } X \geq 0, \text{ and } P(X = 0) < 1, \text{ then } E[X] > 0.$$

Hence $E\left[X_k\right] > 0$, and (21) follows. □

Urban Hjorth's Bias Theorem

The proposition above is on p. 35 of J.S. Urban Hjorth: *Computer Intensive Statistical Methods*, Chapman & Hall, London, 1994.

It says that there is a bias that depends on the model selection method. The expected predictive value of any single PRESS $(\mathcal{M}_k)$ is larger, than the expected predictive value of the best model. Hence the best chosen model underestimates the expected predictive value of any PRESS $(\mathcal{M}_k)$, i.e., the best models give a too optimistic view.

Recursive Estimation for Multiple Linear Models

A Second Application of S-M-W: Recursive estimation algorithms for LSE

$$\mathcal{D}_{tr,(i)} = \left\{ \left(\mathbf{x}_j^{\mathsf{T}}, y_j \right)_{j=1, j \neq i}^{n} \right\}$$

New data $(\mathbf{x}_{n+1}, y_{n+1})$ is received and the training set is augmented by it. The least-squares estimate $\widehat{\boldsymbol{\beta}}$ for the parameters is updated to reflect the new data. By S-M-V formula we show that the updated LSE $\widehat{\boldsymbol{\beta}}_{n+1}$ is given recursively as

$$\widehat{\boldsymbol{\beta}}_{n+1} = \widehat{\boldsymbol{\beta}} + C\left((\mathbf{x}_{n+1}, y_{n+1}), (X^T X)^{-1} \right)$$

where $C\left((\mathbf{x}_{n+1}, y_{n+1}), (X^T X)^{-1} \right)$ is a $(k+1) \times 1$ the random vector that depends only on $(\mathbf{x}_{n+1}, y_{n+1})$ and the current estimate of σ^2 and a storing of current $(X^T X)^{-1}$. Hence, we do not need to solve the normal equations again to get the updated LSE.

New data $\mathbf{x}_{n+1}, y_{n+1}$ is received and the training set is augmented by it.

$$\mathbf{y}_{n+1} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_i \\ \vdots \\ y_n \\ y_{n+1} \end{pmatrix}, X_{n+1} = \begin{pmatrix} \mathbf{x}_1^T \\ \mathbf{x}_2^T \\ \vdots \\ \mathbf{x}_i^T \\ \vdots \\ \mathbf{x}_n^T \\ \mathbf{x}_{n+1}^T \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & \cdots & x_{1k} \\ 1 & x_{21} & \cdots & x_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{i1} & \cdots & x_{ik} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \cdots & x_{nk} \\ 1 & x_{(n+1)1} & \cdots & x_{(n+1)k} \end{pmatrix}$$

The current desing matrix X is found inside the augmented matrix as

$$X_{n+1} = \begin{pmatrix} X \\ \mathbf{x}_{n+1}^{\mathsf{T}} \end{pmatrix}.$$

Hence we get

$$\begin{aligned} X_{n+1}^T X_{n+1} &= \begin{pmatrix} X \\ \mathbf{x}_{n+1}^{\mathsf{T}} \end{pmatrix}^T \begin{pmatrix} X \\ \mathbf{x}_{n+1}^{\mathsf{T}} \end{pmatrix} \\ &= \begin{pmatrix} X^T & \mathbf{x}_{n+1} \end{pmatrix} \begin{pmatrix} X \\ \mathbf{x}_{n+1}^{\mathsf{T}} \end{pmatrix} = X^T X + \mathbf{x}_{n+1} \mathbf{x}_{n+1}^T. \end{aligned} \tag{22}$$

As new data $(\mathbf{x}_{n+1}, y_{n+1})$ is received, the least-squares estimate for the parameters is updated to reflect the new data. The up-dated LSE $\widehat{\boldsymbol{\beta}}_{n+1}$ is

$$\widehat{\boldsymbol{\beta}}_{n+1} = \left(X_{n+1}^T X_{n+1}\right)^{-1} X_{n+1}^T \mathbf{y}_{n+1} \tag{23}$$

Now can apply the S-H-W formula (14) and (22) for the inversion

$$\begin{aligned} \left(X_{n+1}^T X_{n+1}\right)^{-1} &= \left(X^T X + \mathbf{x}_{n+1}\mathbf{x}_{n+1}^T\right)^{-1} \\ &= \left(X^T X\right)^{-1} - \frac{\left(X^T X\right)^{-1}\mathbf{x}_{n+1}\mathbf{x}_{n+1}^{\mathsf{T}}\left(X^T X\right)^{-1}}{1 + \mathbf{x}_{n+1}^{\mathsf{T}}\left(X^T X\right)^{-1}\mathbf{x}_{n+1}}. \end{aligned} \tag{24}$$

Let us note that $1 + \mathbf{x}_{n+1}^{\mathsf{T}}\left(X^T X\right)^{-1}\mathbf{x}_{n+1} \neq 0$, since $\mathbf{x}_{n+1}^{\mathsf{T}}\left(X^T X\right)^{-1}\mathbf{x}_{n+1} \geq 0$, because $\left(X^T X\right)^{-1}$ is positive definite.

In other words, if the old inverse of $(X^T X)$ is available, then the new inverse of $X_{n+1}^T X_{n+1}$ is found by adding a rank 1 correction to $(X^T X)$. In addition

$$\begin{aligned} X_{n+1}^T \mathbf{y}_{n+1} &= \begin{pmatrix} X \\ \mathbf{x}_{n+1}^{\mathsf{T}} \end{pmatrix}^T \begin{pmatrix} \mathbf{y} \\ y_{n+1} \end{pmatrix} \\ &= \begin{pmatrix} X^T & \mathbf{x}_{n+1} \end{pmatrix} \begin{pmatrix} \mathbf{y} \\ y_{n+1} \end{pmatrix} = X^T \mathbf{y} + \mathbf{x}_{n+1} y_{n+1}. \end{aligned} \tag{25}$$

We start now by using the S-H-W formula (24) in (23)

$$\widehat{\boldsymbol{\beta}}_{n+1} = \left(X_{n+1}^T X_{n+1}\right)^{-1} X_{n+1}^T \mathbf{y}_{n+1} =$$

$$= \left(X^T X\right)^{-1} X_{n+1}^T \mathbf{y}_{n+1} - \frac{\left(X^T X\right)^{-1} \mathbf{x}_{n+1} \mathbf{x}_{n+1}^{\mathsf{T}} \left(X^T X\right)^{-1}}{1 + \mathbf{x}_{n+1}^{\mathsf{T}} \left(X^T X\right)^{-1} \mathbf{x}_{n+1}} X_{n+1}^T \mathbf{y}_{n+1}$$

In order to simplify writing (actually $K = K_n$ and $\alpha = \alpha_n$), let us set

$$K := \left(X^T X\right)^{-1} \mathbf{x}_{n+1}, \quad \alpha := \frac{1}{1 + \mathbf{x}_{n+1}^{\mathsf{T}} \left(X^T X\right)^{-1} \mathbf{x}_{n+1}}$$

Thus

$$\widehat{\boldsymbol{\beta}}_{n+1} = \left(X^T X\right)^{-1} X_{n+1}^T \mathbf{y}_{n+1} - \alpha K \mathbf{x}_{n+1}^{\mathsf{T}} \left(X^T X\right)^{-1} X_{n+1}^T \mathbf{y}_{n+1}$$

Now (25) in the first term in the right hand side entails

$$= \underbrace{\left(X^T X\right)^{-1} X^T \mathbf{y}}_{=\widehat{\boldsymbol{\beta}}} + \underbrace{\left(X^T X\right)^{-1} \mathbf{x}_{n+1}}_{=K} y_{n+1} - \alpha K \mathbf{x}_{n+1}^{\mathsf{T}} \left(X^T X\right)^{-1} X_{n+1}^T \mathbf{y}_{n+1}$$

Now we have come to

$$\widehat{\boldsymbol{\beta}}_{n+1} = \widehat{\boldsymbol{\beta}} + y_{n+1}K - \alpha K \mathbf{x}_{n+1}^{\mathsf{T}} \left(X^T X\right)^{-1} X_{n+1}^T \mathbf{y}_{n+1}. \tag{26}$$

We use again (25) to get

$$\alpha K \mathbf{x}_{n+1}^{\mathsf{T}} \left(X^T X\right)^{-1} X_{n+1}^T \mathbf{y}_{n+1} = \alpha K \mathbf{x}_{n+1}^{\mathsf{T}} \underbrace{\left(X^T X\right)^{-1} X^T \mathbf{y}}_{\widehat{\boldsymbol{\beta}}}$$

$$+\alpha K \mathbf{x}_{n+1}^{\mathsf{T}} \left(X^T X\right)^{-1} \mathbf{x}_{n+1} y_{n+1}$$

Again by slides 25-29 of Lecture 4 or MPV p. 99 we get that $\widehat{y}_{n+1} = \mathbf{x}_{n+1}^{\mathsf{T}} \widehat{\boldsymbol{\beta}}$ is the minimal MSE predictor of y_{n+1} and that $\mathbf{Var}\left[\widehat{y}_{n+1}\right] = \sigma^2 \mathbf{x}_{n+1}^{\mathsf{T}} \left(X^T X\right)^{-1} \mathbf{x}_{n+1}$.

When we insert in (26) we find

$$\begin{aligned}\widehat{\boldsymbol{\beta}}_{n+1} &= \widehat{\boldsymbol{\beta}} + y_{n+1}K \\ &\quad -\alpha\widehat{y}_{n+1}K - \frac{\alpha}{\sigma^2}\mathrm{Var}\left[\widehat{y}_{n+1}\right]K.\end{aligned} \tag{27}$$

Here

$$\frac{\alpha}{\sigma^2} = \frac{1}{\sigma^2\left(1+\mathbf{x}_{n+1}^{\mathsf{T}}\left(X^TX\right)^{-1}\mathbf{x}_{n+1}\right)} = \frac{1}{\sigma^2\left(1+\frac{\mathrm{Var}[\widehat{y}_{n+1}]}{\sigma^2}\right)}$$

$$= \frac{1}{\sigma^2+\mathrm{Var}\left[\widehat{y}_{n+1}\right]}$$

When we insert in (27) we get

$$\widehat{\boldsymbol{\beta}}_{n+1} = \widehat{\boldsymbol{\beta}} + c(n+1)K_{n+1}, \tag{28}$$

where the scalar $c(n+1)$ is

$$c(n+1) = y_{n+1} - \frac{\sigma^2}{\sigma^2 + \mathbf{Var}\left[\widehat{y}_{n+1}\right]}\widehat{y}_{n+1} - \frac{1}{\sigma^2 + \mathbf{Var}\left[\widehat{y}_{n+1}\right]}\mathbf{Var}\left[\widehat{y}_{n+1}\right]$$

and the $(k+1) \times 1$ vector K_n is

$$K_{n+1} = \left(X^T X\right)^{-1} \mathbf{x}_{n+1}.$$

The idea for this recursion is from
Hager, William W: Updating the inverse of a matrix. *SIAM Review*, 31 , 2, pp. 221−239, 1989
But the final recursion in (28) is different from the one found loc.cit..

SIAM =Society for Industrial and Applied Mathematics

Appendix: $h_{ii} < 1$

h_{ii} is on the main diagonal of the hat matrix H. We take a look at the ith element on the main diagonal of H^2. Let $\mathbf{h}_{ii}$ denote column i in H. Then the ith element on the main diagonal of H^2 is found as

$$\mathbf{h}_i^T \mathbf{h}_i = \sum_{j=1}^{n} h_{ij}^2.$$

But $H^2 = H$. Hence

$$h_{ii} = \sum_{j=1}^{n} h_{ij}^2 = \sum_{j=1, j\neq i}^{n} h_{ij}^2 + h_{ii}^2$$

. Since $\sum_{j=1, j\neq i}^{n} h_j^2 > 0$ we get

$$h_{ii} > h_{ii}^2 \Leftrightarrow 1 > h_{ii}.$$

Here we could divide both sides by h_{ii}, since $h_{ii} > 1/n > 0$, see one of the problems in SF2930Examgenerator 2023.