The Hat Matrix and Regression Diagnostics

Paul Johnson

9th February 2004

1 OLS Review

Myers, Montgomery, and Vining explain the matrix algebra of OLS with more clarity than any other source I've found. Carefuly study p. 9-14 or so.

The only criticism I have of their style is that they don't use the hat symbol to differentiate a parameter estimate from the symbol that represents the true value. So if you compare what I write with what they write, you see I try to differentiate $\hat{\beta}$ from β, whereas I think they are inconsistent, sometimes using b for the estimates, but also sometimes β is either an estimate or a parameter.

Basically, the theory of OLS is that this linear relationship holds:

$$y = X\beta + e \tag{1}$$

The vectors y and e are $Nx1$, representing the observed dependent variable and the unobserved errors, respectively. X is an Nxp matrix, and the goal is to estimate β, which is a vector that is $p\,x\,1$.

If we have an estimate of β, say $\hat{\beta}$, we can calculate a predicted value, $\hat{y}_i$ for each case, $\hat{y}_i = X_i\hat{\beta}$ (let X_i refer to the i'th row of X), or, in matrix form:

$$\hat{y} = X\hat{\beta}, \tag{2}$$

as well as a "residual," the difference between the predicted and observed value: $\hat{e}_i = y_i - \hat{y}_i$.

The sum of squared residuals is represented in matrix terms as

$$\hat{e}' \cdot \hat{e} \tag{3}$$

Its always true: if you multiply the transpose of the vector by the vector, you end up with the sum of squared elements. Written out in full, it would look like:

$$\begin{bmatrix} y_1 - \hat{y}_1 & y_2 - \hat{y}_2 & y_3 - \hat{y}_3 & \cdots & y_{N-1} - \hat{y}_{N-1} & y_N - \hat{y}_N \end{bmatrix} \begin{bmatrix} y_1 - \hat{y}_1 \\ y_2 - \hat{y}_2 \\ y_3 - \hat{y}_3 \\ \vdots \\ y_{N-1} - \hat{y}_{N-1} \\ y_N - \hat{y}_N \end{bmatrix}$$

Please convince yourself that if you do this multiplication, it gives you the sum of squared residuals.

As you would expect from the regression model, of course, for each case the predicted value is calculated as $\hat{y}_i = X_i \cdot \hat{\beta}$, where the X_i is lazy notation I adopt to refer to the i'th row from the matrix X. The sem of squared residuals in matrix notation is

$$S(\hat{\beta}) = \hat{e}'\hat{e} = (y - X\hat{\beta})'(y - X\hat{\beta}) \tag{4}$$

The very famous **NORMAL EQUATIONS** result when 4 is differentiated with respect to each coefficient in the vector of estimates, $\hat{\beta}' = (\hat{\beta}_1, \hat{\beta}_2, \cdots, \hat{\beta}_p)$. Taking the partial derivatives of $S(\hat{\beta})$ with respect to each coefficient gives p equations, and to find the optimal estimates, the derivatives must be set equal to 0. That is, we need to find, in the end, that $\hat{\beta}$ is the right value so that:

$$\begin{array}{rcl} \frac{\partial S}{\partial \hat{\beta}_1} & = & 0 \\ \frac{\partial S}{\partial \hat{\beta}_2} & = & 0 \\ \frac{\partial S}{\partial \hat{\beta}_3} & = & 0 \end{array} \tag{5}$$

If you do the math one equation at a time, it gets very boring and tedious, but if you trust in the matrix algebra, it is quite concise. This amounts to

$$(X'X)\hat{\beta} - X'y = 0$$

or

$$(X'X)\hat{\beta} = X'y \tag{6}$$

Supposing that the $p\,x\,p$ matrix $(X'X)$ is *invertible* (review handout on matrices and inverses), then the "solution" to the problem is

$$\hat{\beta} = (X'X)^{-1}X'y \tag{7}$$

That is the "big kahoona" of OLS, of course. I once saw excellent t-shirts from ICPSR with that as the slogan.

2 H: The "hat" matrix.

Suppose you calculated the predicted value of y for all of the observations. Here's a vector:

$$\hat{y} = X \cdot \hat{\beta}$$

In that equation, replace $\hat{\beta}$ by the solution in 7. Then the predicted value is equal to

$$\hat{y} = X(X'X)^{-1}X'y$$

This says you can take input in the form of the OBSERVED y vector, and multiply y by that glob $X(X'X)^{-1}X'$, then you end up with the predicted values.

That glob is called the "hat matrix", H. Formally,

$$H = X(X'X)^{-1}X' \tag{8}$$

The hat matrix is NxN.

Obviously,

$$\hat{y} = Hy = X\hat{\beta}$$

Please note, the values in the hat matrix are directly tied to the observed values of y_i for all of the observations. You can't take "any old" vector of y and multiply by h to get meaningful the predicted values. Rather, the particular combination of observed X is used for the particular observed y .

3 R support

R has, in its base, a method called influence.measures(). It will calculate the results described in the following sections.

4 The hat matrix has magical properties.

There are many interesting properties of the hat matrix.

4.1 $\hat{e} = (I - H)y$

In words, the OLS residuals are equal to $(I - H)y$ (see Myers, Montgomery, Vining, p. 42.)

$$\begin{aligned} \hat{e} &= y - X\hat{\beta} \\ &= y - Hy \\ &= (I - H)y \end{aligned} \tag{9}$$

4.2 The matrix $(I - H)$ is symmetric.

A matrix X is symmetric if X is equal to its transpose, $X = X'$. If you wrote out a matrix with 4 elements, say, it would have the "same numbers" above and below the main diagonal,

$$\begin{bmatrix} x_{11} & x_{12} & x_{13} & x_{14} \\ x_{12} & x_{22} & x_{23} & x_{24} \\ x_{13} & x_{23} & x_{13} & x_{34} \\ x_{14} & x_{24} & x_{34} & x_{44} \end{bmatrix}$$

or, for instance,

$$\begin{bmatrix} 0 & 1 & 2 & 3 \\ 1 & 0 & 4 & 5 \\ 2 & 4 & 0 & 6 \\ 3 & 5 & 6 & 0 \end{bmatrix}$$

The only way I know of to convince myself that $(I-H)$ is symmetric is to proceed in two steps. First show the hat matrix itself is symmetric:

$$H = H' \tag{10}$$

If that is correct, then obviously $(I - H)$ is symmetric, because the subtraction $(I - H)$ only affects the magnitude diagonal elements of H, and it leaves all of the off-diagonal elements unchanged, except they become negative. So subtracting H from I cannot change its symmetry.

To convince yourself that $H = H'$, work it out! Recall the matrix handout point that "transpose of a product is the product of the transposes in reverse order". I'm going to use the word transpose instead of prime symbols when I fear ambiguity, especially when $(X'X)^{-1}$ has to be transposed. It would be very ugly to write $((X'X)^{-1})'$ Obviously, transpose(X)=X' and X'=transpose(X).

$$transpose[X(X'X)^{-1}X'] = X \cdot transpose[(X'X)^{-1}] \cdot X' \tag{11}$$

Further, we know that $(X'X)^{-1}$ is a symmetric matrix. We know that is true because $(X'X)$ is symmetric, and so, rather obviously, the inverse is symmetric. If you have trouble agreeing that $(X'X)$ is symmetric, get some paper and create an example X matrix for yourself. After a very very short amount of work, you will see that the off-diagonal elements of $(X'X)$ are perfectly symmetric.

If you believe that $(X'X)^{-1}$ is symmetric, then by definition

$$transpose[(X'X)^{-1}] = (X'X)^{-1}$$

and thus the problem is solved, because you can use that result to simplify 11. In fact, it generates the desired result:

$$transpose[X(X'X)^{-1}X'] = X(X'X)^{-1}X \tag{12}$$

4.3 (I-H) is idempotent.

Idempotent means that a matrix multiplied by itself is equal to itself! If X is idempotent, then $X \cdot X = X$. How peculiar! Obviously, if X is a symmetric matrix, and it is idempotent, then $X'X = X$ and $XX' = X$.

Try to reason through the argument about $(I - H)$ with me, because I've forgotten why I think that $(I - H)$ is idempotent. (Not really, of course, but I'm adding dramatic effect!). For starters, I believe the hat matrix itself is idempotent. It is convincing to write it out:

$$H \cdot H \tag{13}$$

$$X(X'X)^{-1}X' \cdot X(X'X)^{-1}X'$$
$$= X(X'X)^{-1}(X'X)(X'X)^{-1}X'$$

observe in the middle of this that $(X'X)(X'X)^{-1} = I$, so

$$= X(X'X)^{-1} \cdot I \cdot X'$$

and therfore the result is

$$= X(X'X)^{-1}X'$$

That's the answer we wanted, isn't it?

Now I just need a transition from the fact that H is idempotent to the claim that $(I - H)$ is idempotent. Well, think that through, its not so difficult.

$$(I - H) \cdot (I - H) = I(I - H) - H(I - H) = I - H - H + H \cdot H \tag{14}$$
$$= I - 2H + H \cdot H$$

However, because the hat matrix itself is idempotent, then $H \cdot H = H$, so that reduces to

$$I - 2H + H \tag{15}$$
$$= I - H$$

Where I come from, that means the proof is finished. I showed that $(I - H)(I - H) = (I - H)$.

4.4 $Var(\hat{e}) = \sigma^2(I - H)$

4.4.1 Apply the $Var()$ operator to begin.

The variance/covariance matrix of the residuals is what you get when you apply the $Var()$ operator to each side of 9:

$$Var(\hat{e}) = Var[(I - H)y] \tag{16}$$

Please recall from the matrix handout that, generally speaking, if v is a column and X is a matrix,

$$Var(X \cdot v) = X \cdot Var(v) \cdot X'.$$

So, putting $(I - H)$ in place of X and $\hat{e}$ in place of v in that general formula, 16 becomes:

$$Var(\hat{e}) = (I - H)Var(y)(I - H)' \tag{17}$$

4.4.2 Digression: $Var(\hat{e}) = Var(y)$.

Don't forget that $Var(\hat{e})$ and $Var(y)$ are both big, NxN symmetric matrices:

$$Var(y) = \begin{bmatrix} Var(y_1) & Cov(y_1, y_2) & Cov(y_1, y_3) & \cdots & & Cov(y_1, y_N) \\ Cov(y_2, y_1) & Var(y_2) & & & & \\ Cov(y_3, y) & & Var(y_3) & & & \\ & & & & & \\ \vdots & & & \ddots & & \\ & & & & Var(y_{N-1}) & Cov(y_{N-1}, y_N) \\ Cov(y_N, y_1) & & \cdots & & Cov(y_N, y_{N-1}) & Var(y_N) \end{bmatrix} \tag{18}$$

$Var(y)$ is the variance/covariance matrix of the observed y's, and $Var(\hat{e})$ is the variance/covariance matrix of the residuals.

If think for a while, it is obvious that $Var(y)$ equals $Var(e)$. I'm having trouble thinking of a way of justifying that claim with matrix algebra, but if you look at a particular observation, it is easy. Recall, we ASSUMED in the OLS description of the original model that, for an individual observation,

$$y_i = X_i\beta + e_i$$

Use the $Var()$ operator on y_i. (Recall the fundamental formula that $Var(aX+bY) = a^2Var(X)+ b^2Var(Y) + 2abCov(X,Y)$.)

$$Var(y_i) = Var(X_i\beta + e_i) = Var(X_i\beta) + Var(e_i) + 2Cov(X_i\beta, e_i)$$

Because $Var(\beta) = 0$ (the "true values" don't vary!) and because, deep in its guts OLS requires that X is uncorrelated with e,so $Cov(X_i\beta, e_i)$, then we have:

$$= Var(e_i)$$

$$= \sigma^2_{e_i}$$

Here, $\sigma^2_{e_i}$ the "true" variance of the error term for observation i. Homoskedasticity implies that error variance is the same for all observations. So $\sigma^2_{e_i} = \sigma^2_e$ for all i.

So, with the additional conditions imposed in the typical OLS model:

$$Var(y) = \begin{bmatrix} \sigma^2_e & 0 & 0 & 0 & 0 \\ 0 & \sigma^2_e & 0 & 0 & 0 \\ 0 & 0 & \cdots & 0 & 0 \\ 0 & 0 & 0 & \sigma^2_e & 0 \\ 0 & 0 & 0 & 0 & \sigma^2_e \end{bmatrix} = \sigma^2_e \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix} \tag{19}$$

That matrix

- is diagonal because of the "no autocorrelation" assumption on the error terms, meaning $Cov(e_i, e_j) = 0$.
- has the same value in each diagonal element because of homoskedasticity.

4.4.3 Now proceed to finish the argument.

As a result of the preceeding analysis, equation 17 is equal to

$$Var(\hat{e}) \quad = \quad \sigma^2_e(I - H)(I - H)' \tag{20}$$

and, because $(I - H)$ is idempotent and symmetric,

$$= \sigma^2_e(I - H)$$

That's why the work on "symmetric and idempotent" had to be done in the earlier part of this handout.

We never get to know the true variance of the error term, but we estimate it from the data as the MSE (mean square error), and so in a formula like this, we replace σ^2_e with the estimate $\hat{\sigma}^2_e = MSE$. So our best estimate of the Variance/Covariance matrix of the residuals is

$$Var(\hat{e}) = \hat{\sigma}^2_e(I - H) \tag{21}$$

4.4.4 Estimate the variance of a particular observation's residual

Myers, Montgomery, and Vining adopt a custom that I've seen used elsewhere of referring to the individual elements of the hat matrix by h_{ij}, for the $i'th$ row and $j'th$ column of H.

The elements on the diagonal of H are the important ones in many cases, because you can take, say, the 10'th observation, and you calculate the variance of the residual for that obervation:

$$Var(\hat{e}_{10}) = \hat{\sigma}_e^2(1 - h_{10,10})$$

That means, if you just look at the diagonal values of $(I - H)$ you are seeing numbers that indicate how precise the estimate of y is likely to be for a particular value of X.

5 Regression diagnostics with the hat matrix

5.1 You can use the hat matrix to "standardize" –err, "studentize"– the residuals!

Consider the importance of the result stated in 21. If you wonder to yourself, "Is the value of this particular residual, say for the 10'th case, $\hat{e}_{10}$, extremely large?" you can then answer yourself by comparing that value against the variance of that particular residual, $Var(\hat{e}_{10})$.

Do you remember the idea of a standardized Normal variable, one for which the expected value is 0 and the standard deviation is 1? A variable y divided by its standard deviation σ gives a pleasant standardized variable. If you could get a "standardized residual" you could easily gauge outliers.

Your natural instinct might be to divide the residual by the RMSE. Just about everybody has that idea. That would tell you, roughly, how extreme the error term is. Myers, Montgomery, and Vining, and a few others I've found, call this a standardized residual.

$$standardized\,residual_i = \frac{\hat{e}_i}{\sqrt{\hat{\sigma}_e^2}} = \frac{\hat{e}_i}{\hat{\sigma}_e} \tag{22}$$

I don't know who proposed that in the first place, or if many people follow it, but is it a mistake to call that "standardized." We know that the standard deviation of $\hat{e}_i = \hat{\sigma}\sqrt{1 - h_{ii}}$. So usage of this denominator is wrong. It is a mistake to call that a standardized residual because the Root Mean Squared Error (RMSE), $\sqrt{\hat{\sigma}_e^2}$is not really and truly the standard deviation of $\hat{e}_i$. It is, in fact, an OVERSTATEMENT, since $h_{ii} < 1$.

Since we think that the standard deviation of the residual is

$$\hat{\sigma}_e\sqrt{1 - h_{ii}} \tag{23}$$

it only seems natural to use that in the denominator instead. Myers, Montgomery, and Vining use the term **studentized residual** for the following value, r_i. I have found at least one source that refers to simply as a standardized residual:

$$r_i = \frac{\hat{e}_i}{\hat{\sigma}_e\sqrt{1 - h_{ii}}} \tag{24}$$

It seems only right to at least make the known correction to the denominator.

One might object against 24 because it uses the RMSE from the full model, the one with all of the observations is used, as an estimate of $\hat{\sigma}_e$. This is often referred to as the "internal estimate" of σ^2. Suppose, instead, we could use an external estimate of σ_e^2, one that did not depend on

observation i. If we recalculate the RMSE after omitting observation i, then, we have a new variabt of the studentized residual that has a stronger property: it follows the Student's t distribution. Myers, Montgomery, and Vining call this the R-student residual. Here, the term $\hat{\sigma}_{e(-i)}$ means the RMSE from the model that has the i'th observation deleted. The R-student residual (which the SAS manual calls the studentized residual) is then:

$$R_i = \frac{\hat{e}_i}{\sqrt{\hat{\sigma}^2_{e(-i)}(1-h_{ii})}} = \frac{\hat{e}_i}{\hat{\sigma}_{e(-i)}\sqrt{1-h_{ii}}} \tag{25}$$

I think either version of the "studentized residuals" is likely to be fine. The key is not to use the flat-out-wrong version of standardized residuals. The studentized residual r_i is approximately distributed as a Student's t statistic, and the R-student residual is exactly distributed as a t statistic. Since the t is so similar to the Normal, of course, can scan the values of r_i to look for "extreme cases" or "outliers" with the Normal in mind. If a value of r_i or R_i is greater than 2 or 2.5, then you know you have isolated a truly atypical observation.

Now what does this all have to do with the hat matrix? You don't have to run regressions over and over, dropping observations. The hat matrix can be used to form the estimate of the external MSE, based on all observations except i:

$$\hat{\sigma}^2_{e(-i)} = \frac{(N-p)\hat{\sigma}^2_e - \frac{e_i^2}{(1-h_{ii})}}{N-p-1} \tag{26}$$

Note that the multiplication by $(N-p)$ in the numerator just converts the original MSE back into a sum of squares. Recall that

$$\hat{\sigma}^2_e = \frac{1}{N-p}\sum_{i=1}^{N}\hat{e}^2 = \frac{1}{N-p}\hat{e}' \cdot \hat{e}. \tag{27}$$

Recall we are dividing by $N-p$ to convert the "sum of squares" into the "mean square." We divide by $N-p$, rather than just N, as a correction. Use of $N-p$ makes the estimate of the error term's variance unbiased and consistent.

Anyway, you could write 26 as:

$$\hat{\sigma}^2_{e(-i)} = \frac{\sum e_i^2 - \frac{\hat{e}_i^2}{(1-h_{ii})}}{N-p-1} = \frac{1}{N-p-1}[\hat{e}'\hat{e} - \frac{\hat{e}_i^2}{(1-h_{ii})}] \tag{28}$$

Its just the Sum of Squared Errors with a correction element subtracted, and then divided by the degrees of freedom. Hm. As I write this, I fear I'm not clarifying anything for you. But, don't worry, it is doing wonders for me. Plus, I saved a bundle on my car insurance :)

5.2 PRESS residuals

Suppose you drop the i'th observation from the data set and then recalculate the regression model. With that model, use X_i to predict y_i. That number, which Myers, Montgomery, and Vining (p. 42) refer to as the PRESS residual, ("Prediction Error Sum of Squares"), might be referred to as

$$\hat{e}_{(i)} \; or \; PRESS_i$$

If that procedure is repeated for each observation, then the statistic proposed by Allen (1971) is the sum of squares:

$$PRESS = \sum_{i=1}^{N} \hat{e}_{(i)}^2$$

The $PRESS$ estimate is sometimes useful as a summary measure of a model's ability to predict new observations.

$$R^2_{prediction} = 1 - \frac{PRESS}{Total\ Sum\ of\ Squares} \tag{29}$$

If you calculate that number, it can be thought of as the ability to explain the variability in predicting new observations. The ordinary R^2 is higher than this PRESS statistic.

The hat matrix enters this discussion because it saves a lot of calculation. One need not actually re-calculate the regression results N different times. Rather, it is true that the PRESS residual is equal to the ordinary residual divided by 1 minus the diagonal of the hat matrix.

$$\hat{e}_{(i)} = \frac{\hat{e}_i}{1 - h_{ii}} \tag{30}$$

Furthermore, since the hat matrix has already been calculated, one can simply cycle through the values in H and calculate PRESS.

I've seen at least one stats manual that refers to this PRESS value as DRESID, short for "Deleted Residual".

5.2.1 Digression: Standardized PRESS equals Studentized residual

Here's an interesting fact I just recently learned from reading a professor's class notes for a Biostatistics class (at http://www.sph.umich.edu/class/bio650/2001/LN_Nov05.pdf).

Fact: a "standardized" PRESS residual is identical to a studentized residual.

Begin by noting that the variance of $PRESS_i = \sigma_e^2/(1 - h_{ii})$, and if we use the square root of that as the standardizing value, watch what happens:

$$\frac{PRESS_i}{\sigma_e/\sqrt{1 - h_{ii}}} = \frac{\hat{e}_i/(1 - h_{ii})}{\sigma_e/\sqrt{1 - h_{ii}}} = \frac{\hat{e}_i}{\sigma_e\sqrt{1 - h_{ii}}} \tag{31}$$

So, if you aren't concerned about the prediction R^2, then perhaps you can dispense with the $PRESS$ concept altogether and stick with studentized residuals.

5.3 Inspect H for "leverage points"

Worrying about outliers is often justified, but just about as often, it is a waste of time. Suppose you sit and worry that observation 19 is extreme. Then you consider excluding that observation, and you wonder if you are doing the right thing.

It may be that observation does not influence the regression estimates. You'd rather find out if the observation is distorting the predicted values of the other cases.

As it turns out, the hat matrix, H, offers terriffic evidence in this regard. The hat matrix is an NxN square. Suppose you calculate predicted values, thus:

$$\begin{bmatrix} \hat{y}_1 \\ \hat{y}_2 \\ \\ \vdots \\ \\ \hat{y}_{N-1} \\ \hat{y}_N \end{bmatrix} = \begin{bmatrix} h_{11} & h_{12} & h_{13} & & & h_{1N} \\ h_{21} & & & & & h_{2N} \\ h_{31} & & & \ddots & & \vdots \\ \vdots & & & & & \\ & & & & h_{N-1N} & \\ h_{N1} & & & & h_{NN-1} & h_{NN} \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ \\ \vdots \\ y_{N-1} \\ y_N \end{bmatrix} \tag{32}$$

$$= \begin{bmatrix} y_1h_{11} & +y_2h_{12} & +y_3h_{13} & + & \cdots & & +y_Nh_{1N} \\ y_1h_{21} & + & & & & & +y_Nh_{2N} \\ y_1h_{31} & + & & \ddots & & & \vdots \\ \vdots & & & & & & \\ & & & & & & +y_Nh_{N-1N} \\ y_1h_{N1} & + & \cdots & & +y_{N-1}h_{NN-1} & & +y_Nh_{NN} \end{bmatrix}$$

If you look at this for a while, it becomes apparent that the element, h_{ij} gives the influence of the j'th observation on the i'th predicted value, $\hat{y}_i$.

If you compare across row i in the hat matrix, and some values are huge, it means that some observations are exercising a disproportionate influence on the prediction for the i'th observation.

If you concentrate on the diagonal elements, h_{ii}, you are focusing on the effects that observations have on their own predicted values. If a model estimated without observation i offers a grossly different predicted value for y_i than a model that includes i, then you know that observation i is having a pretty dramatic effect on the fitted model.

Consider at the diagonal of the hat matrix:

$$\begin{matrix} h_{11} & & & & \\ & h_{12} & & & \\ & & \ddots & & \\ & & & h_{N-1,N-1} & \\ & & & & h_{NN} \end{matrix} \tag{33}$$

the most "pleasant" result would be that all of the elements are the same. It would mean that the positioning of an observation in the X space, as indicated by h_{ii}, is not exerting an extraordinary influence for any observation. Since being an "outlier" is a matter of an observation's position in the X space as well as the y space, that is meaningful.

On p. 47 in Myers, Montgomery, and Vining, there is a discussion of this that is not entirely clear to me. They claim that the sum of the diagonal elements h_{ii} is equal to p, the number of parameters to be estimated. If so, that gives us a good standard against which to evaluate values of h_{ii}. If all of the h_{ii} were exactly the same, then they would be equal to p/N. If an element in the hat matrix diagonal is twice as great as that average value, then that observation should be considered a leverage point and one ought to be cautious about it.

5.4 DFFITs

I have seen this question treated slightly differently in several statistics programs and books. This is really just a re-statement of the previous point, but in a slightly different vocabulary. Suppose we calculate the change in predicted value of the j'th observation due to the deletion of observation j from the dataset. Call that the DFFIT:

$$DFFIT_j = \hat{y}_j - \hat{y}_{(-j)} \tag{34}$$

Keep in mind that $\hat{y}_j$ is the predicted value for observation j from the whole model and the predicted value for observation j based on parameters estimated after deleting observation j is $\hat{y}_{(-j)}$.

In and of itself, this value is difficult to interpret. A standardizing approach has often been proposed that employs the hat matrix. The estimate of the RMSE when the j'th observation is deleted is referred to as $\hat{\sigma}_{e(-j)}$. That value is

$$DFFITS_j = \frac{\hat{y}_j - \hat{y}_{(-j)}}{\hat{\sigma}_{e(-j)}\sqrt{h_{jj}}} \tag{35}$$

If $DFFITS_j$ is large, of course, it means that the j'th observation is influential on the model's predicted value for the j'th observation. In other words, the model does not fit observation j particularly well.

Probably because of the p/N reasoning I discuss in the previous section (drawn from Myers, Montgomery, and Vining), it is widely recommended that one should be cautious of observations for which $DFFITS > 2\sqrt{p/N}$.

5.5 DFBETA

Apply the "drop-one-observation-at-a-time" approach to find out if an observation influences the estimate of a slope parameter. Let

$$\hat{\beta}_{(-j)}$$

represent the vector of estimates based on the dataset with observation j omitted. As usual, $\hat{\beta}$refers to the estimate obained using all data.

The DFBETA value, a measure of influence of observation j on the parameter estimate, is

$$d_j = \hat{\beta} - \hat{\beta}_{(-j)} \tag{36}$$

If an element in this vector is huge, it means you should be cautious about including observation j in your analysis.

Of course, "huge" is difficult to define, so one idea is to standardize. A standardized variant, where you divide the difference by the standard error of the estimated coefficient. This is called DFBETAS. DFBETAS is meaningful one-variable-at-a-time, so consider the estimate of parameter i when observation j is omitted. The notation is getting tedious here, but let's use $d[i]_j$. Standardize the impact of the j'th observation on the i'th parameter estimate as:

$$d[i]_j* = \frac{d[i]_j}{\sqrt{Var(\hat{\beta}_{i(-j)})}} \tag{37}$$

The denominator is the standard error of the estimated coefficient when j is omitted. There is a rule of thumb that is often brought to bear: If the DFBETAS value for a particular coefficient is greater than $2/\sqrt{N}$ then the influence is large.

Of course, you are wondering why I introduced DFBETA in the middle of a section on the hat matrix. Well, it can be shown that:

$$d[i]_j = \frac{\hat{e}(X'X)^{-1}X_j}{1 - h_{ii}} \tag{38}$$

5.6 Cook's distance: Integrating the DFBETA results

The DFBETA analysis is unsatisfying because we can calculate a whole vector of DFBETAS, one for each parameter, but we only analyze them one-by-one. Can't we combine all of those parameters?

The Cook distance approach is a way of integrating all of the information in a single test to answer the following question:

> Is the vector of estimates obtained with observation j omitted, $\hat{\beta}_{(-j)}$, meaningfully differerent from the vector obtained when all observations are used?

Cook's distance measure is, in essence, a way to evaluate the overall distance between the point $\hat{\beta} = (\hat{\beta}_1, \hat{\beta}_2, ..., \hat{\beta}_p)$ and the point $\hat{\beta}_{(-j)} = (\hat{\beta}_{1(-j)}, \hat{\beta}_{2(-j)}, ..., \hat{\beta}_{p(-j)})$. (There are deep thoughts awaiting if you start considering the vector of parameter estimates as a point in a p-dimensional space, incidentally.)

If we were interested only in raw, unstandardized distance, we could use the usual "straight line between two points" measure of distance. (Recall the pythagorean theorem? $a^2 + b^2 = c^2$. That is for distances in 2 dimensions, but the idea easily generalizes to p dimensions. The distance is the square root of the sum of the squared differences of the individual elements). The square of the total distance would be

$$(\hat{\beta}_{(-j)} - \hat{\beta})'(\hat{\beta}_{(-j)} - \hat{\beta})$$

Again, I'd urge you to convince yourself that this is indeed a sum of squares, i.e., a distance measure. Write it out, in other words.

It seems to me that the clever part of Cook's scheme was to weight the distance calculations in order to bring them into a meaningful scale. For a weight, Cook proposed the cross product matrix divided by the number of parameters that are estimated and the MSE.

$$\frac{X'X}{p \cdot \hat{\sigma}_e^2}$$

The measure D_j indicates the magnitude of the difference in parameter estimates when j is omitted.

$$D_j = \frac{(\hat{\beta}_{(-j)} - \hat{\beta})'X'X(\hat{\beta}_{(-j)} - \hat{\beta})}{p \cdot \hat{\sigma}_e^2}$$

If you think of the change in predicted value as $X(\hat{\beta}_{(-j)} - \hat{\beta})$, then you can look at the above index as the squared change in predicted value divided by a normalizing factor. To see that, regroup as

$$D_j = \frac{[X(\hat{\beta}_{(-j)} - \hat{\beta})]'[X(\hat{\beta}_{(-j)} - \hat{\beta})]}{p \cdot \hat{\sigma}_e^2}$$

Clearly, the top is the sum of squared prediction changes. The denominator includes p because

there are p parameters that can change and $\hat{\sigma}_e^2$ is, of course, your friend, the MSE, the estimate of the variance of the error term.

Myers, Montgomery, and Vining, citing Cook (1997), suggest that D_j is distributed as an F variable with the p for the numerator degrees of freedom and $(n-p)$for the denominator. They recommend that we think of the "extreme" outcomes as the ones that would happen less than half of the time, so the significance level in the F table is 0.5. That makes it convenient because the critical value of F is 1.

Of course, since I introduce this under the hat matrix section, you know what's coming. Cook's distance can be calculated as:

$$D_j = \frac{r_j^2}{p} \frac{h_{jj}}{(1-h_{jj})} \tag{39}$$

6 Alternatives to diagnostics

In case the diagnostics give you a headache, you might investigate further into the topic of "robust" regression. With alternative estimation techniques, one can often automatically overcome the impact of outliers simply by estimating according to an alternative criterion. Then one does not need to worry if an observation is an outlier or not. The algorithm will do it for you.

7 Conclusion

The ideas here have never been very useful to me. I think that's because my work has mostly been done with large samples and categorical variables. If one had continuous data, especially small datasets, these ideas could be very important.

These ideas, however, are very useful building blocks for ideas that are indeed very useful. In the Generalized Linear Model, one encounters a sequence of different kinds of residuals, standardized and not, and they are very important in that context.