Notes on and Solutions to Selected Problems In: Residuals and Influence in Regression by R. Dennis Cook and Sanford Weisberg

John L. Weatherwax*

May 20, 1993

*wax@alum.mit.edu

Chapter 2 (Diagnostic methods using residuals)

Notes On The Text

Notes on the Hat matrix V

In this section of the text we derive many of the results presented and discussed in the book pertaining to the hat matrix V. Lets begin with the decomposition of the full measurement matrix X (where each row corresponds a feature/measurement) into two feature subset parts (X_1, X_2). Beginning with the first set of features represented by X_1 we can form the projection of Y onto the subspace spanned by these features using the matrix U defined by

$$U \equiv X_1(X_1^T X_1)^{-1} X_1^T .$$

After we have projected onto this initial subspace to utilize the information contained in the *second* set of features and represented by the matrix X_2 note that we don't gain any information from any component of X_2 that lie in the space already spanned by the features in X_1. Thus the "independent information" contained in X_2 is to be found in the orthogonal projection of X_2 onto X_1 or the space spanned by the columns of X_2^* defined as the reduction of X_2 by the projection of X_2 onto the span of the columns of X_1 or

$$X_2^* = X_2 - UX_2 = (I - U)X_2 . \tag{1}$$

Thus the correct subspace onto which we will project Y onto and which provided any additional information not already found in X_1 is given by

$$T^* = X_2^*(X_2^{*T} X_2^*)^{-1} X_2^{*T} .$$

We can put the definition of X_2^* from Equation 1 into the above expression to find an alternative expression for T^* in terms of U and X_2. Since U is symmetric we have that

$$T^* = ((I - U)X_2)(X_2^T(I - U)(I - U)X_2)^{-1} X_2^T (I - U) .$$

Note that since U is idempotent ($U^2 = U$) so is $I - U$ because

$$(I - U)(I - U) = I - 2U + U^2 = I - 2U + U = I - U ,$$

and the expression for T^* becomes

$$T^* = (I - U)X_2(X_2^T(I - U)X_2)^{-1} X_2^T (I - U) , \tag{2}$$

which is the books 2.1.5. This T^* is the projection matrix that projects onto the part of the column space of X that is orthogonal to the column space of X_1. Thus in the discussion above what we have done is to split the features in the total data matrix X into two parts X_1 and X_2 with projection matrices U to project onto the column space of X_1 and T^* to project onto the column space of X_2 and that is orthogonal to the column span of X_1. Thus the total transformation, onto into the column space of X and denoted by V is given by the sum of these two projections as

$$V = U + T^* . \tag{3}$$

This equation expresses the decompositional view of the affect of adding additional features to a linear regression in that the resulting total projection is the sum of individual features projections. We now use this relationship to derive some relationships about the hat matrix V and its elements v_{ij}

To begin we note that any symmetric and idempotent (i.e. $V^2 = V$) matrix V must have

$$v_{ii} = \sum_{j=1}^{n} v_{ij} v_{ji} = \sum_{j=1}^{n} v_{ij}^2 , \tag{4}$$

showing that $v_{ii} > 0$ since it is expressed as the sum of positive elements v_{ij}^2. Using this and Equation 3 which expresses that the total projection matrix, V, obtained when we add a new variable to an existing regression is equivalent to simply adding an appropriate symmetric idempotent projection matrix T^* to the current projection matrix U we see that as each new feature is added each adds another positive diagonal element so the diagonal elements of V are non-decreasing with respect to p the number of explanatory variables.

Consider the general result expressed by Equation 3 but for the specific case where we first split the feature matrix X into a column of ones denoted by $\mathbf{1}$ which will be X_1 and then take the matrix X_2 to be all the remaining predictors. Note that the projection onto the column vector of all ones, $\mathbf{1}$, is given by

$$U = \mathbf{1}(\mathbf{1}^T \mathbf{1})^{-1} \mathbf{1}^T = \frac{1}{n} \mathbf{1}\mathbf{1}^T .$$

From which we find that the reduced columns X_2^* is given by

$$X_2^* = (I - U) X_2 = \left(I - \frac{1}{n} \mathbf{1}\mathbf{1}^T \right) X_2 .$$

We can simplify the second term above as

$$\frac{1}{n} \mathbf{1}\mathbf{1}^T X_2 = \mathbf{1} \left(\frac{1}{n} \mathbf{1}^T X_2 \right) = \mathbf{1}(\bar{x}^T) ,$$

where $\bar{x}$ is is the mean vector and then write X_2^* as

$$X_2^* = X_2 - \mathbf{1}\bar{x}^T = \mathcal{X} .$$

We have defined $\mathcal{X}$ as the mean centered $n \times p$ matrix of explanatory variables. Then T^* the projection onto X_2^* is given by

$$T^* = X_2^* (X_2^{*T} X_2^*)^{-1} X_2^{*T} = \mathcal{X} (\mathcal{X}^T \mathcal{X})^{-1} \mathcal{X}^T .$$

Using all of this we put everything back into Equation 3 to find that

$$V = \frac{1}{n} \mathbf{1}\mathbf{1}^T + \mathcal{X} (\mathcal{X}^T \mathcal{X})^{-1} \mathcal{X}^T \tag{5}$$

which is the books equation 2.17.

We can use this expression to derive some more results involving the diagonal elements of V. Take e_i to be a vector of all zeros except with a single one in the ith spot $1 \le i \le n$. Using this v_{ii} is expressed simply as $v_{ii} = e_i^T V e_i$ and from Equation 5 we see that v_{ii} is given by

$$\begin{aligned} v_{ii} &= e_i^T V e_i = \frac{1}{n} + e_i^T \mathcal{X}(\mathcal{X}^T\mathcal{X})^{-1}\mathcal{X}^T e_i \\ &= \frac{1}{n} + (\mathcal{X}^T e_i)^T(\mathcal{X}^T\mathcal{X})^{-1}\mathcal{X}^T e_i \,. \end{aligned}$$

Now $\mathcal{X}^T e_i$ is another expression for the ith centered feature vector x_i or the ith row of $\mathcal{X}$. Thus

$$v_{ii} = \frac{1}{n} + x_i^T(\mathcal{X}^T\mathcal{X})^{-1}x_i \,, \tag{6}$$

which is the books equation 2.1.8. In the case of simple linear regression the matrix $\mathcal{X}$ is really a vector given by

$$\mathcal{X} = \left[\begin{array}{c} x_1 - \bar{x} \\ x_2 - \bar{x} \\ \vdots \\ x_n - \bar{x} \end{array}\right],$$

so $\mathcal{X}^T\mathcal{X} = \sum_{i=1}^n (x_i - \bar{x})^2$ and we get from Equation 6 that

$$v_{ii} = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2} \,.$$

From which we see that $v_{ii} > 1/n$. This result holds in the multidimensional case also. Note that in the multidimensional case $\mathcal{X}^T\mathcal{X}$ is positive definite, since if we have a vector v such that $v \neq 0$ then

$$v^T\mathcal{X}^T\mathcal{X}v = (\mathcal{X}v)^T(\mathcal{X}v) = ||\mathcal{X}v||^2 > 0 \,.$$

Since $\mathcal{X}^T\mathcal{X}$ is positive definite the inverse of $\mathcal{X}^T\mathcal{X}$ is also positive definite and so $x_i^T(\mathcal{X}^T\mathcal{X})^{-1}x_i > 0$ and

$$v_{ii} = \frac{1}{n} + x_i^T(\mathcal{X}^T\mathcal{X})^{-1}x_i > \frac{1}{n} \,, \tag{7}$$

is a lower bound on v_{ii}. An upper bound can be obtained and depends on the number of *repeated* feature vectors. If several feature vectors x_j all equal the same value, say x_i, then since

$$v_{ij} = e_i^T V e_j = e_i^T X(X^TX)^{-1}X^T e_j \,.$$

As we have repeated features vectors since $X^T e_j = x_j$ and $x_j = x_i$ we have $X^T e_j = X^T e_i$, thus

$$v_{ij} = e_i^T X(X^TX)^{-1}X^T e_i = v_{ii} \,.$$

Since V is idempotent and symmetric we know that Equation 4 holds true. If in the sum, $\sum_{j=1}^n v_{ij}^2$ we sum only over the values of j for which the rows of X are equal to the value x_i and for which $v_{ij} = v_{ii}$ we have

$$v_{ii} = \sum_{j=1}^n v_{ij}^2 \ge c v_{ii}^2 \,,$$

assuming that there are c such rows. From this inequality dividing both sides by the positive v_{ii} we are left with $v_{ii} \le 1/c$. This expression combined with Equation 7 gives the bounds

$$\frac{1}{n} \le v_{ii} \le \frac{1}{c} \,, \tag{8}$$

which is the books equation 2.1.9. In the most common case if there are *no* repeated feature vectors for x_i then $c = 1$ and the above gives $v_{ii} \le 1$. If v_{ii} achieve this maximum value of 1 then from Equation 4 we can factor out the single term v_{ii}^2 from the sum $\sum_{j=1}^n v_{ij}^2$ on the right-hand-side and bringing it to the left-hand-side to get the expression

$$\sum_{j=1;j\neq i}^{n} v_{ij}^2 = v_{ii} - v_{ii}^2 \,.$$

If $v_{ii} = 1$ the right-hand-side of the above vanishes and we have $\sum_{j=1;j\neq i}^n v_{ij}^2 = 0$ which means that each term v_{ij}^2 must vanish which in tern means that $v_{ij} = 0$ for all $j \neq i$. Then from the error-residual relationship $e = (I - V)\varepsilon$ written in component form

$$e_i = \varepsilon_i - \sum_{j=1}^{n} v_{ij}\varepsilon_j = \varepsilon_i - v_{ii}\varepsilon_i = \varepsilon_i - \varepsilon_i = 0 \,.$$

Now since the ith residual e_i is given by $e_i = y_i - \hat{y}_i$ we conclude that $\hat{y}_i = y_i$ or that in this case the prediction $\hat{y}_i$ exactly equals the data y_i.

Starting from the result presented in Equation 6 we will now derive an alternative expression for v_{ii} that will show examples of what type of properties an inputs x_i will need to have to produce extreme values of v_{ii}. Since the matrix $\mathcal{X}^T\mathcal{X}$ is symmetric it has an eigenvector decomposition that we can write as

$$\mathcal{X}^T\mathcal{X}P = P\Lambda \,,$$

where P is an orthogonal matrix with columns given by the eigenvectors of $\mathcal{X}^T\mathcal{X}$ and Λ is a diagonal matrix matrix with the eigenvalues $\mu_i \ge 0$ on the diagonal. Taking the inverse of $\mathcal{X}^T\mathcal{X}$ using this expression we see that

$$(\mathcal{X}^T\mathcal{X})^{-1} = P\Lambda^{-1}P^T \,.$$

Using this in the expression $x_i^T(\mathcal{X}^T\mathcal{X})^{-1}x_i$ we find

$$x_i^T(\mathcal{X}^T\mathcal{X})^{-1}x_i = x_i^T(P\Lambda^{-1}P^T)x_i = (P^Tx_i)^T\Lambda^{-1}(P^Tx_i) \,.$$

Note that we have

$$P^Tx_i = \begin{bmatrix} p_1^T \\ p_2^T \\ \vdots \\ p_p^T \end{bmatrix} x_i = \begin{bmatrix} p_1^Tx_i \\ p_2^Tx_i \\ \vdots \\ p_p^Tx_i \end{bmatrix} ,$$

so the product P^Tx_i gives the vector that has components $p_l^Tx_i$ for $l = 1, 2, \cdots, p$. Then

$$(P^Tx_i)^T\Lambda^{-1}(P^Tx_i) = \sum_{l=1}^{p} \frac{(p_l^Tx_i)^2}{\mu_l} \,.$$

If we put the μ_l inside of square of the above we see that v_{ii} can be written as

$$v_{ii} = \frac{1}{n} + \sum_{l=1}^{p} \left(\frac{p_l^Tx_i}{\sqrt{\mu_l}} \right)^2 , \tag{9}$$

which is the books equation. Since p_i has unit length we define θ_{li} as

$$\cos(\theta_{li}) = \frac{p_l^T x_i}{||p_l||\,||x_i||} = \frac{p_l^T x_i}{(x_i^T x_i)^{1/2}}\,.$$

Thus using this expression for $p_l^T x_i$ we find

$$v_{ii} = \frac{1}{n} + (x_i^T x_i)\sum_{l=1}^{p}\left(\frac{\cos(\theta_{li})}{\sqrt{\mu_l}}\right)^2\,,$$

which is the books equation 2.1.10. From this expression we see that one way for v_{ii} to be large will happen if $x_i^T x_i$ is large. Since x_i is the mean removed ith sample this inner product $x_i^T x_i$ will be large if this sample is very far from the mean $\bar{x}$. Another way for the value of v_{ii} to be large is to have $\cos(\theta_{pi})^2 \approx 1$. This is equivalent to x_i having a significant component in the same direction as the eigenvector, p_p, with the smallest eigenvalue μ_p.

The role of V in data analysis

Recall that the residual vector e is related to the true error vector ε by $e = (I - V)\varepsilon$. If the true errors are distributed as $\varepsilon \sim N(0, \sigma^2 I)$ then using their relationship we see that $E(e) = 0$ and the variance of e can be computed as

$$\begin{aligned}
\mathrm{Var}(e) &= (I - V)\mathrm{Var}(\varepsilon)(I - V)^T \\
&= \sigma^2(I - V)(I - V) = \sigma^2(I - 2V + V^2) \\
&= \sigma^2(I - V)\,,
\end{aligned}$$

since $V^2 = V$ and V is symmetric. This last result is useful since it states that the variance of the observed residuals e will depend on the hat matrix V.

The use of the ordinary residuals: bias in the model

Statistics of the residuals e can be used to suggest errors in the functional specification of the linear model. One way in which this can be seen is with the following example. If the *true* linear model (the relationship that actually generates the observed $(\mathbf{x}_i, y_i)$ data) is really given by

$$Y = X\beta + B + \varepsilon\,, \tag{10}$$

that is the functional representation between X and Y contains an unmodeled bias term B. Assume then as a modeler we make a "mistake" and assuming that the relationship between X and Y is in fact given by

$$Y = X\beta + \varepsilon\,, \tag{11}$$

then the residuals e will demonstrate this error with a bias in their expectation. The bias to the residuals that results is stated without proof in the book but we can derive the explicit bias representation as follows. Express the ith residual using the hat matrix with elements v_{ij} as

$$e_i = y_i - \hat{y}_i = y_i - \sum_{j=1}^{N} v_{ij} y_j$$

The expectation of e_i is then simply

$$E(e_i) = E(y_i) - \sum_{j=1}^{N} v_{ij} E(y_j) \,.$$

Since we are told that the true model is given by Equation 10 we see that

$$E(y_i) = \beta^T x_i + b_i \,,$$

since $E(\varepsilon_i) = 0$. Using this we have that $E(e_i)$ becomes

$$\begin{aligned} E(e_i) &= \beta^T x_i + b_i - \sum_{j=1}^{N} v_{ij} (\beta^T x_j + b_j) \\ &= \beta^T \left[x_i - \sum_{j=1}^{N} v_{ij} x_j \right] + b_i - \sum_{j=1}^{N} v_{ij} b_j \,. \end{aligned} \quad (12)$$

We will now consider the expression in brackets on the right-hand-side of the above expression and show that it is in fact zero. To do this recall that from the definition of the hat matrix V we have

$$X - VX = X - X(X^T X)^{-1} X^T X = 0 \,.$$

Taking the transpose of this equation and using symmetry of V gives

$$X^T = X^T V \,.$$

Lets write out this matrix equation in terms of its columns. We see that it is equivalent to

$$\begin{aligned} \begin{bmatrix} x_1 & x_2 & \cdots & x_N \end{bmatrix} &= \begin{bmatrix} x_1 & x_2 & \cdots & x_N \end{bmatrix} \begin{bmatrix} v_{11} & v_{12} & \cdots & v_{1N} \\ \vdots & \vdots & & \vdots \\ v_{N1} & v_{N2} & \cdots & v_{NN} \end{bmatrix} \\ &= \begin{bmatrix} \sum_{j=1}^N v_{j1} x_j & \sum_{j=1}^N v_{j2} x_j & \cdots & \sum_{j=1}^N v_{jN} x_j \end{bmatrix} . \end{aligned}$$

Thus the ith column of this expression gives

$$x_i - \sum_{j=1}^{N} v_{ji} x_j = 0 \,.$$

Since V is symmetric $v_{ij} = v_{ji}$ so this last expression is what is needed to make the term in brackets in Equation 12 vanish and we are left with

$$E(e_i) = (1 - v_{ii}) b_i - \sum_{j=1; j \neq i}^{N} v_{ij} b_j \,, \quad (13)$$

when we bring the b_i term out of the summation. This is the book's equation 2.1.13.