DELETING AN OBSERVATION FROM A LINEAR REGRESSION

by

R. L. Eubank

Technical Report No. SMU-DS-TR-183
Department of Statistics ONR Contract

February 1984

Research sponsored by the Office of Naval Research
Contract N00014-82-K-0207
Project NR 042-479

Department of Statistics
Southern Methodist University
Dallas, Texas 75275

Deleting an Observation From a Linear Regression

by

R. L. Eubank
Southern Methodist University

SUMMARY

An elementary proof is given for the form of regression coefficient estimates and fitted values when a particular observation has been deleted from the data. The proof requires minimal use of matrix algebra and, consequently, provides an approach to the derivation of various "leave-one-out" diagnostic measures which bypasses the matrix manipulations utilized in most texts that address the subject. Applications to the derivation of jackknife estimator and variance formulae for linear models are also discussed.

1. INTRODUCTION

Most courses in regression analysis employ a variety of model diagnostics based on residuals and fits obtained from the entire sample as well as from reduced data sets where one of the observations has been deleted. The rationale behind such "leave-one-out" measures is that, to assess an observations impact, both fitted models which include and exclude an observation should be examined. It is a remarkable fact that the computation of these types of diagnostic measures can be accomplished using only the information available from the fit to the complete data set so that no refitting of the

model to data subsets is required. Previous proofs of this fact have utilized the Sherman-Morrison-Woodbury Theorem (see Rao (1973, pg. 33)) and somewhat tedious matrix algebra. In this note an alternative proof is provided that keeps matrix manipulations to a minimum. Use of this approach may provide time savings in regression courses and would be particularly appropriate in instances where detailed matrix algebra is to be avoided. Moreover, many inference courses include the jackknife as a topic, for which a natural application is the linear model. The results in this paper provide a simple method of deriving jackknife estimator and variance formulae for this setting that does not require the background development necessary for more matrix oriented proofs.

Consider the model

$$\underline{y} = X\underline{\beta} + \underline{\varepsilon}$$

where $\underline{y} = (y_1,\ldots,y_n)'$ is the vector of observations, $\underline{\beta} = (\beta_1,\ldots,\beta_p)'$ is a vector of unknown parameters, X is an $n\times p$ matrix of rank p having ith row $\underline{x}_i'$ and $\underline{\varepsilon}$ is a vector of zero mean uncorrelated errors with common variance σ^2. Define the matrix of catchers by

$$C = \{c_{ij}\} = (X'X)^{-1}X'$$

and the hat matrix by

$$H = \{h_{ij}\} = XC$$

(see Hoaglin and Welsch (1978) and Velleman and Welsch (1981) for discussions of these matrices). Then, the least squares coefficient estimates and fitted values are given by

$$\hat{\underline{\beta}} = (\hat{\beta}_1,\ldots,\hat{\beta}_p)' = C\underline{y}$$

and

$$\hat{\underline{y}} = (\hat{y}_1,\ldots,\hat{y}_n)' = H\underline{y}.$$

Now let $\hat{\underline{\beta}}^{(i)} = (\hat{\beta}_1^{(i)},\ldots,\hat{\beta}_p^{(i)})'$ denote the coefficient estimates obtained using only the observations $y_1,\ldots,y_{i-1},y_{i+1},\ldots,y_n$ and define

$$\hat{y}_j^{(i)} = \underline{x}_j'\hat{\underline{\beta}}^{(i)} \qquad i,j = 1,\ldots,n.$$

Using the Sherman-Morrison-Woodbury formula, viz.

$$(A + \underline{u}\,\underline{v}')^{-1} = A^{-1} - \frac{(A^{-1}\underline{u})(\underline{v}'A^{-1})}{1-\underline{v}'A^{-1}\underline{u}}, \tag{1.1}$$

which holds for A a nonsingular p×p matrix and $\underline{u}$, $\underline{v}$, p×1 vectors, it is possible to show that

$$\hat{\beta}_j^{(i)} = \hat{\beta}_j - c_{ji}(y_i - \hat{y}_i)/(1-h_{ii}) \tag{1.2}$$

and, hence, that

$$y_i - \hat{y}_i^{(i)} = (y_i - \hat{y}_i)/(1-h_{ii}). \tag{1.3}$$

Proofs of (1.2)-(1.3) are outlined, for example, in Hoaglin and Welsch (1978) and Belsley, Kuh and Welsch (1980). A detailed proof of (1.3) can be found in Gunst and Mason (1980, pg. 258).

Formulas (1.2)-(1.3) are the important identities for the derivation of diagnostics such as studentized residuals and DFITS (see Velleman and Welsch (1981) for this terminology) and (1.2) is the key to obtaining jackknife coefficient and variance estimates for linear regression (see Miller (1974) and Efron (1982, pg. 18)). In the next section we present a simple direct method of deriving these relations which does not require (1.1).

2. THE RESULT

The objective of this section is to prove the following theorem.

<u>Theorem.</u> <u>Let</u> $\hat{\underline{\beta}}^{(i)}(z_i)$ <u>solve</u>

$$\min_{\underline{\beta}} \left\{ \sum_{\substack{j=1\\ j\neq i}}^{n} (y_j - \underline{x}_j'\underline{\beta})^2 + (z_i - \underline{x}_i'\underline{\beta})^2 \right\}.$$

Then,

$$\hat{y}_i^{(i)} = y_i - (y_i - \hat{y}_i)/(1-h_{ii}) \tag{2.1}$$

and

$$\hat{\underline{\beta}}^{(i)}(\hat{y}_i^{(i)}) = \hat{\underline{\beta}}^{(i)} . \tag{2.2}$$

Before proving this result let us pause to interpret what it says. First note that (2.1) is precisely (1.3). Equation (2.2) has the implication that to obtain $\hat{\underline{\beta}}^{(i)}$ we need only multiply the vector $(y_1,\ldots,y_{i-1}, \hat{y}_i^{(i)}, y_{i+1},\ldots,y_n)'$ by the matrix C. In view of (2.1), (1.2) is an immediate consequence.

The proof of this theorem is an adaptation of work by Craven and Wahba (1979) for smoothing splines and proceeds as follows. Set $z_i^* = \underline{x}_i'\hat{\underline{\beta}}^{(i)} = \hat{y}_i^{(i)}$. Then, (2.2) is a result of the inequalities

$$\sum_{\substack{j=1\\ j\neq i}}^{n} (y_j - \underline{x}_j'\hat{\underline{\beta}}^{(i)})^2 + (z_i^* - \underline{x}_i'\hat{\underline{\beta}}^{(i)}) = \sum_{\substack{j=1\\ j\neq i}}^{n} (y_j - \underline{x}_j'\hat{\underline{\beta}}^{(i)})^2$$

$$< \sum_{\substack{j=1\\ j\neq i}}^{n} (y_j - \underline{x}_j'\underline{\beta})^2 \quad (\text{since } \hat{\underline{\beta}}^{(i)} \text{ minimizes } \sum_{\substack{j=1\\ j\neq i}}^{n} (y_j - \underline{x}_j'\underline{\beta})^2)$$

$$\leq \sum_{\substack{j=1\\ j\neq i}}^{n} (y_j - \underline{x}_j'\underline{\beta})^2 + (z_i^* - \underline{x}_i'\underline{\beta})^2 .$$

To verify (2.1) note that $\underline{x}_i'\hat{\underline{\beta}}^{(i)}(z_i)$ is linear in z_i and, by expanding about y_i, can be written as

$$\underline{x}_i'\hat{\underline{\beta}}^{(i)}(z_i) = \hat{y}_i + h_{ii}(z_i - y_i).$$

Taking $z_i = \hat{y}_i^{(i)}$ and using (2.2) gives the desired result.

To conclude we note that, for example, the formula for a studentized residual can be derived as in Belsley, Kuh and Welsch (1980, pg. 64) once (1.2)-(1.3) have been established.. To derive the jackknife estimator of β_j, $\tilde{\beta}_j$, observe that (c.f. Miller (1974) or Efron (1982))

$$\begin{aligned}\tilde{\beta}_j &= n^{-1} \sum_{i=1}^{n} \{n\hat{\beta}_j - (n-1)\hat{\beta}_j^{(i)}\} \\ &= \hat{\beta}_j + \frac{n-1}{n} \sum_{i=1}^{n} c_{ji} \frac{(y_i - \hat{y}_i)}{1-h_{ii}}\end{aligned}$$

with expressions for their variances and covariances following similarly.

REFERENCES

Belsley, D. A., Kuh, E. and Welsch, R. E. (1980). Regression Diagnostics: Identifying Influential Data and Sources of Collinearity, John Wiley: New York.

Craven, P. and Wahba, G. (1979), Smoothing noisy data with spline functions. Numer. Math, 31, 377-403.

Efron, B. (1983), The Jackknife, The Bootstrap and Other Resampling Plans, SIAM, monograph no. 38, CBMS-NSF.

Gunst, R. F. and Mason R. L. (1980), Regression Analysis and Its Applications: A Data-Oriented Approach, Marcel-Dekker, New York.

Hoaglin, D. C. and Welsch, R. E. (1978), "The Hat Matrix in Regression and ANOVA", The American Statistician, 32, 17-22 and Corrigenda, 32, 146.

Miller, R. G. (1974), "An Unbalanced Jackknife", Annals of Statistics, 2, 880-891.

Rao, C. R. (1965), Linear Statistical Inference and Its Applications (2nd ed.), John Wiley, New York.

Velleman, P. F. and Welsch, R. E. (1981), "Efficient Computing of Regression Diagnostics," The American Statistician, 35, 234-242.

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE

READ INSTRUCTIONS BEFORE COMPLETING FORM

1. REPORT NUMBER SMU/DS/TR-183	2. GOVT ACCESSION NO.
3. RECIPIENT'S CATALOG NUMBER	
4. TITLE (and Subtitle) DELETING AN OBSERVATION FROM A LINEAR REGRESSION	5. TYPE OF REPORT & PERIOD COVERED Technical Report
	6. PERFORMING ORG. REPORT NUMBER SMU/DS/TR-183
7. AUTHOR(s) R. L. Eubank	8. CONTRACT OR GRANT NUMBER(s) N00014-82-K-0207
9. PERFORMING ORGANIZATION NAME AND ADDRESS Southern Methodist University Dallas, Texas 75275	10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS NR 042-479
11. CONTROLLING OFFICE NAME AND ADDRESS Office of Naval Research Arlington, VA 22217	12. REPORT DATE February 1984
	13. NUMBER OF PAGES 6
14. MONITORING AGENCY NAME & ADDRESS(if different from Controlling Office)	15. SECURITY CLASS. (of this report)
	15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

This document has been approved for public release and sale; its distribution is unlimited. Reproduction in whole or in part is permitted for any putpose of The United States Government.

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

An elementary proof is given for the form of regression coefficient estimates and fitted values when a particular observation has been deleted from the data. The proof requires minimal use of matrix algebra and, consequently, provides an approach to the derivation of various "leave-one-out" diagnostic measures which bypasses the matrix manipulations utilized in most texts that address the subject. Applications to the derivation of jackknife estimator and variance formulae for linear models are also discussed.

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE

S/N 0102-LF-014-6601

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)