A REMARK ON ADJUSTING FOR COVARIATES IN MULTIPLE REGRESSION

BY

M.L. EATON

AND

D.A. FREEDMAN

TECHNICAL REPORT NO. 11

SEPTEMBER 1982

RESEARCH PARTIALLY SUPPORTED

BY

NATIONAL SCIENCE FOUNDATION GRANT MCS 81 00762

AND

NATIONAL SCIENCE FOUNDATION GRANT MCS 80 02535

DEPARTMENT OF STATISTICS

UNIVERSITY OF CALIFORNIA

BERKELEY, CALIFORNIA

August 1982

A REMARK ON ADJUSTING FOR COVARIATES IN MULTIPLE REGRESSION

by

M. L. Eaton[1]

Department of Theoretical Statistics
University of Minnesota, Minneapolis

and

D. A. Freedman[2]

Department of Statistics
University of California, Berkeley

Abstract. A formula is given to determine the impact of adjusting for covariates on the accuracy of estimates in a multiple regression model.

Key words and phrases. Regression, covariates

Running head. Adjusting for covariates

[1]Research partially supported by National Science Foundation Grant MCS 81 00762

[2]Research partially supported by National Science Foundation Grant MCS 80 02535

1. Introduction

Breiman and Freedman (1982) consider the problem of determining the optimal number of explanatory variables in a multiple regression equation, in order to minimize prediction error; that paper has a review of the literature. Using similar techniques, Freedman and Moses (1982) determine the optimal number of covariates in a clinical trial to measure a treatment effect. The model considered there is

$$(1) \qquad Y_i = \alpha\xi_i + \beta\zeta_i + \sum_{j=1}^{\infty} \gamma_j X_{ij} + \varepsilon_i$$

where

Y_i is the response of the i^{th} subject

ξ_i is 1 for all subjects

ζ_i is 1 for subjects in treatment, and 0 for subjects in control

X_{ij} is covariate j measured on subject i

In this equation, α and γ_j are nuisance parameters; the object is to minimize the variance of the regression estimate of β. The covariates are considered as observed values of random variables. In principle, there are infinitely many covariates that could be entered into the equation, and a decision must be made as to when to stop. The order for entering the covariates is pre-determined. Thus, β will be estimated from the regression of Y_i on $\alpha\xi_i + \beta\zeta_i + \sum_{j=1}^{p} \gamma_j X_{ij}$, for $i = 1,\ldots,n$. The problem is to choose p.

This paper will consider a slightly more general model, namely

$$(2) \qquad Y_i = \sum_{j=1}^{k} \zeta_{ij}\beta_j + \sum_{j=1}^{\infty} X_{ij}\gamma_j + \varepsilon_i \text{ for } i = 1,\ldots,n$$

Here, ζ_{ij} is deterministic, and ζ has rank $k < n$; the X_{ij} are nonsingular multivariate gaussian, with mean 0; the infinite vectors $\{X_{ij}: j = 1,2,\ldots\}$ are independent and identically distributed in i; the ε's are independent of the X's, having mean 0 and variance σ^2. The assumptions on X may be relaxed to orthogonal invariance of the joint distribution of the rows, but we do not pursue this.

Let c be a fixed k-vector. The object is to estimate the contrast $c'\beta$, in a regression of Y_i on

$$\sum_{j=1}^{k} \zeta_{ij}\beta_j + \sum_{j=1}^{p} X_{ij}\gamma_j$$

Let $\hat{\theta}_{npc}$ denote this estimator of $c'\beta$. How is p to be chosen to minimize $\operatorname{var}\hat{\theta}_{npc}$? To determine the answer, let

$$\sigma_p^2 = \operatorname{var}\{\sum_{j=p+1}^{\infty} X_{ij}\gamma_j \mid X_{i1},\ldots,X_{ip}\}$$

By our assumption, σ_p^2 is deterministic and does not depend on i. The main result of this paper can now be stated; the proof is given in the next section.

<u>Theorem</u>. Let $V_{npc} = \operatorname{var}\{\hat{\theta}_{npc} \mid X_{ij} \text{ for } 1 \leqq i \leqq n \text{ and } 1 \leqq j \leqq p\}$. Then V_{npc} is distributed as

$$(\sigma^2+\sigma_p^2)c'(\zeta'\zeta)^{-1}c[1 + \chi_p^2/\chi_{n-p-k+1}^2]$$

the chi-squared variables being independent.

In particular, the optimal p minimizes

$$(\sigma^2+\sigma_p^2)(1 + \frac{p}{n-p-k-1})$$

The quantity $\sigma^2+\sigma_p^2$ may be estimated from the data. For more details, see Breiman and Freedman (1982).

2. Proof of theorem

We begin with a special case of an identity due to Woodbury (1950). Let C be an arbitrary $k \times p$ matrix. Notice that $C'C$ and CC' are non-negative definite. Let I_k and I_p be the $k \times k$ and $p \times p$ identity matrices.

Lemma. $(I_k+CC')^{-1} = I_k - C(I_p+C'C)^{-1}C'$

Proof. This is almost a computation:

$$\begin{aligned} I_p &= (I_p+C'C)^{-1}(I_p+C'C) \\ &= (I_p+C'C)^{-1} + (I_p+C'C)^{-1}C'C \\ &= (I_p+C'C)^{-1} + C'C(I_p+C'C)^{-1} \end{aligned}$$

Multiply on the left by C and on the right by C' and juggle:

$$(I_k+CC')[I_k - C(I_p+C'C)^{-1}C'] = I_k . \qquad \square$$

Turn now to the theorem. We may assume without loss of generality that the X_{ij} are all independent $N(0,1)$ variables, as argued in Breiman and Freedman (1982). By redefining ε and σ^2, we may also assume that $\gamma_j = 0$ for $j > p$. Thus, we may restrict attention to the model

$$\underset{n\times 1}{Y} = \underset{n\times k}{\zeta}\ \underset{k\times 1}{\beta} + \underset{n\times p}{X}\ \underset{p\times 1}{\gamma} + \underset{n\times 1}{\epsilon} \tag{3}$$

where the X_{ij} are independent $N(0,1)$ variables; the components of ϵ are independent of X, with mean 0 and variance σ^2. As usual, introduce the matrix $H = X(X'X)^{-1}X'$, which is the projection into the column space of X.

Lemma. In the model (3), the least squares estimate $\hat{\beta}$ of β is given by the formula

$$\hat{\beta} = (W'W)^{-1}W'Y$$
$$W = (I-H)\zeta$$

Proof. As usual, $\hat{\beta}$ may be obtained by the regression of $\tilde{Y}$ on $\tilde{\zeta}$, where $\tilde{Y}$ is the part of Y orthogonal to the columns of X, and likewise for $\tilde{\zeta}$. Formally, this is the regression of $(I-H)Y$ or even Y itself on $(I-H)\zeta$, since HY is orthogonal to $(I-H)\zeta$. □

In particular, since $I-H$ is idempotent,

$$\text{Cov}\{\hat{\beta}|X\} = \sigma^2(W'W)^{-1} = \sigma^2(\zeta'\zeta-\zeta'H\zeta)^{-1} \tag{4}$$

Using for example the Gram-Schmidt process, write $\zeta = \psi N$ where ψ is $n \times k$ and $\psi'\psi = I_k$, while N is $k \times k$ and nonsingular. Now extend ψ to a full $n \times n$ orthonormal matrix; that is, create an $n \times (n-k)$ matrix Φ such that the concatenation $M = [\psi,\Phi]$ is orthonormal. Let

$$U = \psi'X \quad \text{and} \quad V = \Phi'X$$

Then U is a $k \times p$ matrix, V is an $(n-k) \times p$ matrix; the entries of U and V are all independent $N(0,1)$ variables.

Proposition. $\text{cov}\{\hat{\beta}|X\} = \sigma^2[(\zeta'\zeta)^{-1} + N^{-1}FN'^{-1}]$ where

$$F = U(V'V)^{-1}U' .$$

Proof. We pick up the argument from (4). Put $\zeta = \psi N$ and recall that $N'N = \zeta'\zeta$ to see

$$(\zeta'\zeta-\zeta'H\zeta)^{-1} = (\zeta'\zeta)^{-1} + N^{-1}TN'^{-1}$$

where

$$T = (I_k - \psi' H \psi)^{-1} - I_k$$

Recall that the concatenation $M = [\psi, \Phi]$ is orthonormal, so $(X'X) = (MX)'(MX) = U'U + V'V$. Of course, $\psi' H \psi = (\psi' X)(X'X)^{-1}(X\psi)$. Thus

$$T = [I_k - U(U'U+V'V)^{-1}U']^{-1} - I_k$$

Let $S = V'V$ and $C = US^{-1/2}$. Then

$$\begin{aligned} T &= [I_k - C(C'C+I_p)^{-1}C']^{-1} - I_k \\ &= CC' = U(V'V)^{-1}U \end{aligned}$$

by the Lemma. □

Remark. The proof shows that $\mathrm{cov}\{\hat{\beta}|X\} = \sigma^2 N^{-1}(I_k + F)N'^{-1}$. The random matrix F has a matrix F-distribution -- see Dawid (1981) for a discussion and properties of such distributions.

Proof of the Theorem for the model (3). Plainly, $\mathrm{var}\{c'\hat{\beta}|X\}$ is

(5) $$\sigma^2[c'(\zeta'\zeta)^{-1}c + R] \quad \text{where} \quad R = c'N^{-1}FN'^{-1}c$$

Since the law of $F = U(V'V)^{-1}U'$ is invariant under rotations of U, the distribution of R in (5) depends only on the squared length of $N'^{-1}c$, which is

$$d^2 = c'N^{-1}N'^{-1}c = c'(N'N)^{-1}c = c'(\zeta'\zeta)^{-1}c \; .$$

Moreover, the distribution of R/d^2 coincides with that of $U_1(V'V)^{-1}U_1'$, where U_1 is the first row of U. This is Hotelling's T^2-statistic. □

References

L. Breiman and D. Freedman (1982). How many variables should be entered in a regression equation? Technical report no. 1, Department of Statistics, University of California, Berkeley.

Dawid, A.P. (1981). Some matrix-variate distribution theory: Notational considerations and a Bayesian application. Biometrika, **68**, 265-274.

D. Freedman and L. Moses (1982). Adjusting for covariates in clinical trials. Technical report, in preparation, Department of Statistics, Stanford University.

M. Woodbury (1950). Princeton Technical Report No. 42.