The Annals of Statistics
1976, Vol. 4, No. 3, 625-628

ON A CRITERION FOR SIMULTANEOUS EXTRAPOLATION IN NONFULL RANK NORMAL REGRESSION

BY FEDERICO J. O'REILLY

Stanford University and

CIMAS, Universidad Nacional Autónoma de México

In recent work by O'Reilly, a necessary and sufficient condition for the existence of an unbiased estimate of the distribution function of a "future" observation was given. The result was obtained under the condition that the model had full rank. Here, the result is generalized to any number of future observations and the full rank condition is relaxed. The corresponding uniformly minimum variance unbiased estimator is identified from the density estimates given by Ghurye and Olkin.

1. Introduction and summary. In O'Reilly [1] the existence of an unbiased estimate of the distribution function of a "future" observation is proposed as a criterion to extrapolate at the associated design point. A necessary and sufficient condition to extrapolate, involving the design matrix X and the future design point $\mathbf{x}$, is derived; this condition is that $\mathbf{x}'(X'X)^{-1}\mathbf{x} \leqq 1$.

If one is interested in extrapolating at several design points, the unbiased estimation of each of the corresponding distribution functions might be possible, however, it might be the case that the unbiased estimation of the joint distribution function of the future observations is no longer possible.

In the present paper, allowing the number of future observations to be arbitrary and relaxing the full rankness of X, a necessary and sufficient condition for the existence of an unbiased estimate of the joint distribution function of the future observations is derived. This condition particularizes to the one given in [1], when the number of future observations is 1 and X is full rank.

In Section 2, the necessary notation is introduced. In Section 3, the condition is derived under full rankness. Finally, in Section 4, the results of the previous section are used in the nonfull rank case under reparameterization.

2. Definitions and notation. Let $\mathbf{Y}' = (Y_1, Y_2, \cdots, Y_n)$ be a vector of independent random variables with $Y_i \sim N(\mathbf{x}_i'\boldsymbol{\beta}, \sigma^2)$, where $\boldsymbol{\beta} \in R^p$, $\sigma^2 > 0$ are unknown and $\mathbf{x}_i'$ is the ith row of a known matrix X of rank $q \leqq p < n$.

Denote by T the statistic $(X'Y, Y'Y)$, by $\hat{\sigma}^2$ the usual unbiased estimate of σ^2 and by $\hat{\boldsymbol{\beta}}$ any solution to the normal equations.

Let $\mathbf{l} \sim N(U'\boldsymbol{\beta}, \sigma^2 I)$, be a $k \times 1$ vector of future observations, where U is arbitrary. Following [1], $\mathbf{l}$ will be said to have an *estimable distribution* iff there exists a vector valued measurable function $\mathbf{h}(Y_1, Y_2, \cdots, Y_n)$ distributed as $\mathbf{l}$,

Received July 1975.

AMS 1970 *subject classifications.* Primary 62F10, 62J05.

Key words and phrases. MVUE (minimum variance unbiased estimate), estimability of a distribution, validity of extrapolation.

and *extrapolation at U will be said to be valid* iff $\mathbf{l}$ has an estimable distribution.

3. Region of extrapolation (X full rank). Assume $q = p$, and define S as the set of $p \times k$ matrices U, where there exists a linear function of the observations distributed as $\mathbf{l}$, i.e.,

$$S = \{U_{p\times k};\ \exists\ A_{n\times k},\ A'A = I,\ U = X'A\}\ .$$

By construction, $U \in S$ is a sufficient condition to extrapolate at U, but to show that it is also a necessary condition, a characterization of S is given, which is a generalization of Lemma 3.1 of [1].

LEMMA 3.1.

$$S = \{U_{p\times k};\ I - U'(X'X)^{-1}U \text{ is positive semidefinite of rank } \leqq n - p\}\ .$$

PROOF. Let $U \in S$, since $I - X(X'X)^{-1}X'$ is p.s.d., it follows that $A'(I - X(X'X)^{-1}X')A \equiv I - U'(X'X)^{-1}U$ is p.s.d. of rank $\leqq n - p$.

Define $A_0 = X(X'X)^{-1}U$, it follows that $X'A_0 = U$ and $I - A_0'A_0$ is p.s.d. Let P be an orthogonal $k \times k$ matrix which diagonalizes $A_0'A_0$, i.e., $P'A_0'A_0P = \text{Diag}\,(\lambda_1, \lambda_2, \cdots, \lambda_k)$. Since $I - A_0'A_0$ is p.s.d., select $\boldsymbol{\xi}_1, \boldsymbol{\xi}_2, \cdots, \boldsymbol{\xi}_k$ as follows: For the $\lambda_i = 1$, set the corresponding $\boldsymbol{\xi}_i = \mathbf{0}$, and for the $\lambda_i < 1$, select the corresponding $\boldsymbol{\xi}_i$ to be any orthogonal set of vectors with $||\boldsymbol{\xi}_i||^2 = 1 - \lambda_i$ and orthogonal to $\mathscr{C}(X)$ (the column space of X). Let $C = [\boldsymbol{\xi}_1, \boldsymbol{\xi}_2, \cdots, \boldsymbol{\xi}_k]$, it follows that $A = A_0 + CP'$ is such that $A'A = I$, $X'A = U$.

THEOREM 3.1. *In the full rank normal regression model, extrapolation at U is valid* iff $U \in S$.

PROOF. Sufficiency is obvious. To show necessity assume $U \notin S$, and also assume $\exists\ \mathbf{h}(Y_1, Y_2, \cdots, Y_n) \sim N(U'\boldsymbol{\beta}, \sigma^2 I)$. Since $E(\mathbf{h}) = U'\boldsymbol{\beta}$, and because of the completeness of T, it follows that $E(\mathbf{h} \mid T) = U'\hat{\boldsymbol{\beta}}$ a.s., however $V(\mathbf{h}) - V\{E(\mathbf{h} \mid T)\} = \sigma^2(I - U'(X'X)^{-1}U)$, which by assumption is not p.s.d., and this contradicts the Rao–Blackwell theorem.

Ghurye and Olkin [2] exhibit the UMVUE of the corresponding $N(U'\boldsymbol{\beta}, \sigma^2 I)$ density, which exists iff $I - U'(X'X)^{-1}U$ is p.d. It is interesting to note that under the approach given here, that condition can be shown to be necessary, since if this was not the case and $I - U'(X'X)^{-1}U$ is p.s.d. but not p.d., then

$$\exists\ \boldsymbol{\lambda} \quad \text{s.t.} \quad \boldsymbol{\lambda}'(I - U'(X'X)^{-1}U)\boldsymbol{\lambda} = 0\ ,$$

and

$$\exists\ A \quad \text{s.t.} \quad \mathbf{h} \equiv A'\mathbf{Y} \sim N(U'\boldsymbol{\beta}, \sigma^2 I)\ ,$$

however $V(\boldsymbol{\lambda}'A'\mathbf{Y}) = V\{E(\boldsymbol{\lambda}'A\mathbf{Y} \mid T)\}$ which means that $\boldsymbol{\lambda}'A\mathbf{Y}$ is essentially a function of T, thus the UMVUE of the $N(U'\boldsymbol{\beta}, \sigma^2 I)$ distribution would assign probability 1 to the set $[\boldsymbol{\lambda}'A\mathbf{Y} = \boldsymbol{\lambda}'U'\hat{\boldsymbol{\beta}}]$ and thus will fail to be absolutely continuous a.s. with respect to the k-dimensional Lebesgue measure.

4. Region of extrapolation ($q < p$). Graybill [3], page 236, shows that there

exists a nonsingular $p \times p$ matrix W s.t. $W' = [A_1' : A_2']$, $W^{-1} = [B_1' : B_2']$ with A_1, B_1 $q \times p$ such that the rows A_1 (and of B_1) span $\mathscr{R}(X)$, and the rows of A_2 (and of B_2) span the orthogonal complement of $\mathscr{R}(X)$, where $\mathscr{R}(X)$ denotes the row space of the matrix X.

The mean vector $X\boldsymbol{\beta}$ can be written as $Z\boldsymbol{\gamma}$, where $Z = XB_1'$ is full rank and $\boldsymbol{\gamma} = A_1\boldsymbol{\beta}$ is the new vector of estimable parameters.

The following results hold if $\mathscr{C}(U) \subset \mathscr{R}(X)$:

(i) there is a 1 : 1 relationship between U and $U^* \equiv B_1U$,

(ii) $U^{*\prime}(Z'Z)^{-1}U^* = U'(X'X)^-U$, for $(X'X)^-$ any generalized inverse,

(iii) $A_1\hat{\boldsymbol{\beta}}$ is the unique solution to $(Z'Z)\boldsymbol{\gamma} = Z'\mathbf{Y}$.

LEMMA 4.1.

$$S = \{U_{p\times k};\ \mathscr{C}(U) \subset \mathscr{R}(X),\ I - U'(X'X)^-U \text{ is p.s.d. of rank } \leqq n - q\}\,.$$

PROOF. If $U \in S$, it follows that $\mathscr{C}(U) \subset \mathscr{R}(X)$ and clearly $S = \{U^*_{q\times k};\ \exists\, A,\ A'A = I,\ U^* = Z'A\}$.

THEOREM 4.1. *In the normal regression model, extrapolation at U is valid* iff $U \in S$.

PROOF. The only implication that needs to be shown is that $\mathscr{C}(U)$ must be a subspace of $\mathscr{R}(X)$. If this was not true, then it would be possible to have an estimate of the $N(U'\boldsymbol{\beta}, \sigma^2I)$ distribution with a nonestimable mean, and this is absurd. (It can be shown that a nonestimable (linearly), linear function of $\boldsymbol{\beta}$ is nonestimable.)

The form of the UMVUE of the $N(U'\boldsymbol{\beta}, \sigma^2I)$ distribution can be obtained from the UMVUE of the corresponding density given in [2], page 1268, which in this case exists iff $\mathscr{C}(U) \subset \mathscr{R}(X)$, $I - U'(X'X)^-U$ is p.d. and $k < n - q$. This estimate is

$$\begin{aligned}\tilde{f}_n(u) &= 0 \qquad \text{if} \quad (\mathbf{u} - U'\hat{\boldsymbol{\beta}})'(I - U'(X'X)^-U)^{-1}(\mathbf{u} - U'\hat{\boldsymbol{\beta}}) \geqq (n-q)\hat{\sigma}^2 \\ &= \frac{\Gamma(\frac{1}{2}(n-q))\{(n-q)\hat{\sigma}^2\}^{-\frac{1}{2}(n-q-2)}}{\pi^{k/2}\Gamma(\frac{1}{2}(n-k-q))}\{(n-q)\hat{\sigma}^2 - (\mathbf{u} - U'\hat{\boldsymbol{\beta}})'(I - U'(X'X)^-U)^{-1} \\ &\qquad \times (\mathbf{u} - U'\hat{\boldsymbol{\beta}})\}^{\frac{1}{2}(n-q-k-2)}|I - U'(X'X)^-U|^{-\frac{1}{2}} \quad \text{elsewhere.}\end{aligned}$$

This last expression, when positive, is just a multivariate t density with $n - q - 1$ d.f. obtained by the transformation

$$\mathbf{t} = \frac{(I - U'(X'X)^-U)^{-\frac{1}{2}}(\mathbf{u} - U'\hat{\boldsymbol{\beta}})}{\{(n-q)\hat{\sigma}^2 - (\mathbf{u} - U'\hat{\boldsymbol{\beta}})'(I - U'(X'X)^-U)^{-1}(\mathbf{u} - U'\hat{\boldsymbol{\beta}})\}^{\frac{1}{2}}}\,.$$

If $k = n - q$, the UMVUE of the $N(U'\boldsymbol{\beta}, \sigma^2I)$ distribution assigns probability 1 to the set where $(A'\mathbf{Y} - U'\hat{\boldsymbol{\beta}})'(I - U'(X'X)^-U)^{-1}(A'\mathbf{Y} - U'\hat{\boldsymbol{\beta}}) = (n-q)\hat{\sigma}^2$ and therefore does not have a density with respect to k-dimensional Lebesgue measure.

REFERENCES

[1] O'REILLY, F. J. (1975). On a criterion for extrapolation in normal regression. *Ann. Statist.* **3** 219–222.

[2] GHURYE, S. G. and OLKIN, I. (1969). Unbiased estimation of some multivariate probability densities and related functions. *Ann. Math. Statist.* **40** 1261–1271.

[3] GRAYBILL, F. A. (1961). *An Introduction to Linear Statistical Models.* McGraw-Hill, New York.

CIMAS
APARTADO POSTAL 20-726
MEXICO 20, D. F.
MEXICO