The Annals of Statistics
1975, Vol. 3, No. 1, 219-222

ON A CRITERION FOR EXTRAPOLATION IN NORMAL REGRESSION[1]

BY FEDERICO J. O'REILLY

CIMAS, Universidad Nacional Autónoma de México

In a full rank normal regression model, the existence of an unbiased estimate of the distribution associated to any "future" observation is studied. A necessary and sufficient condition in terms of the point at which the "future" observation will be taken is given for such an estimate to exist. This criterion is introduced to consider the validity of extrapolation.

1. Introduction and summary. In the usual normal regression model, there is no formal answer to the question of whether it is valid or not to extrapolate the adjusted regression in the sense of the validity of the postulated model at the point where one wishes to extrapolate.

In this paper, extrapolation is studied under the assumption of validity of the model, and the criterion used to decide whether extrapolation is valid at a certain point is based on the existence of an unbiased estimate of the distribution function associated to the "future" observation. In this sense extrapolation is understood as the unbiased estimation of the random behavior of the future observation.

In Section 2, the necessary notation is introduced. In Section 3, restricting the analysis to the full rank model, it is shown that extrapolation is restricted to an ellipsoid centered at the origin which turns out to be a function of the regression matrix X. Moreover, within the ellipsoid the minimum variance unbiased estimate (MVUE) of the associated distribution is identified. Finally, in Section 4 some general comments are given.

2. Definitions and notation. Let $\mathbf{Y}' = (Y_1, Y_2, \cdots, Y_n)$ be a vector of n independent rv's 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 full rank matrix X.

Also denote by T the statistic $(X'\mathbf{Y}, \mathbf{Y}'\mathbf{Y})$ and by $\hat{\boldsymbol{\beta}}$, $\hat{\sigma}^2$ the usual unbiased estimates of $\boldsymbol{\beta}$ and σ^2 respectively.

In order to restrict estimability to unbiased estimability and to define the criterion for extrapolation, the following definitions are given.

DEFINITION 2.1. A rv Y_0 is said to have an estimable distribution if there exists a measurable function $h(Y_1, Y_2, \cdots, Y_n)$ distributed as Y_0.

DEFINITION 2.2. Extrapolation at $\mathbf{x}_0$ is said to be valid if the rv $Y_0 \sim N(\mathbf{x}_0'\boldsymbol{\beta}, \sigma^2)$, has an estimable distribution.

Received August 1973; revised January 1974.

[1] Presented at an Alumni meeting at CIENES, Santiago, Chile (August 1973).

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

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

3. Region of extrapolation. Let $S \subset R^p$ be defined by $S = \{\mathbf{x};\, \exists \boldsymbol{\alpha} \in R^n,\ \|\boldsymbol{\alpha}\| = 1 \text{ and } \mathbf{x} = X'\boldsymbol{\alpha}\}$. Obviously if $\mathbf{x}_0 \in S$, extrapolation at $\mathbf{x}_0$ is valid since $h(Y_1, \cdots, Y_n) = \boldsymbol{\alpha}'\mathbf{Y} \sim N(\mathbf{x}_0'\boldsymbol{\beta}, \sigma^2)$. However, it is not obvious that in order to have a valid extrapolation at $\mathbf{x}_0$, the point must lie in S. To show that this is the case, an alternative characterization of S is needed.

LEMMA 3.1. $S = \{\mathbf{x};\, \mathbf{x}'(X'X)^{-1}\mathbf{x} \leqq 1\}$.

PROOF. Let $\mathbf{x} \in S$, thus $\exists \boldsymbol{\alpha}$ with norm 1 such that $\mathbf{x} = X'\boldsymbol{\alpha}$. Denote by $C(X)$ the column space of X. Let $\boldsymbol{\alpha}_0$ be the orthogonal projection of $\boldsymbol{\alpha}$ in $C(X)$. It follows that $\|\boldsymbol{\alpha}_0\| \leqq 1$ and $\exists \mathbf{u} \in R^p$ such that $\boldsymbol{\alpha}_0 = X\mathbf{u}$. Also, since $\boldsymbol{\alpha} - \boldsymbol{\alpha}_0$ is orthogonal to $C(x)$, this means that $\mathbf{x} = X'\boldsymbol{\alpha}_0$.

Evaluating $\mathbf{x}'(X'X)^{-1}\mathbf{x}$, this is equal to $\|\boldsymbol{\alpha}_0\|^2$. Thus, $S \subset \{\mathbf{x};\, \mathbf{x}'(X'X)^{-1}\mathbf{x} \leqq 1\}$.

Now, assume that $\mathbf{x}'(X'X)^{-1}\mathbf{x} \leqq 1$, and define $\boldsymbol{\alpha}_0 = X(X'X)^{-1}\mathbf{x}$. The result follows, letting $\boldsymbol{\alpha}$ be any vector with $\|\boldsymbol{\alpha}\| = 1$ whose orthogonal projection in $C(X)$ is $\boldsymbol{\alpha}_0$.

THEOREM 3.1. *In the full rank normal regression model, extrapolation at* $\mathbf{x}_0$ *is valid iff* $\mathbf{x}_0 \in S$.

PROOF. Sufficiency is obvious. To show necessity, assume $\mathbf{x}_0 \notin S$ and also assume $\exists h(Y_1, \cdots, Y_n) \sim N(\mathbf{x}_0'\boldsymbol{\beta}, \sigma^2)$. Since $E(h) = \mathbf{x}_0'\boldsymbol{\beta}$ and since T is complete and sufficient it follows that

$$E(h \mid T) = \mathbf{x}_0'\hat{\boldsymbol{\beta}}\,.$$

However $V\{E(h \mid T)\} > V(h)$, which is a contradiction.

In order to build the MVUE of a $N(\mathbf{x}_0'\boldsymbol{\beta}, \sigma^2)$ distribution function when $\mathbf{x}_0 \in S$ the results of Ghurye and Olkin [1], and O'Reilly and Quesenberry [2] are used.

In [1], page 1268, a density and its MVUE related to the $N(\mathbf{x}_i'\boldsymbol{\beta}, \sigma^2)$ distribution is given from which in [2], page 80, for $n > p + 1$, the MVUE of the $N(\mathbf{x}_i'\boldsymbol{\beta}, \sigma^2)$ distribution is obtained as:

$$\begin{aligned}\tilde{F}_{y_i}(y) &= 0 \quad \text{if } \ y - \mathbf{x}_i'\hat{\boldsymbol{\beta}} < -\{(1 - \mathbf{x}_i'(X'X)^{-1}\mathbf{x}_i)(n-p)\hat{\sigma}\}^{\frac{1}{2}}\,,\\ &= 1 \quad \text{if } \ y - \mathbf{x}_i'\hat{\boldsymbol{\beta}} \geqq \{(1 - \mathbf{x}_i'(X'X)^{-1}\mathbf{x}_i)(n-p)\hat{\sigma}\}^{\frac{1}{2}}\,,\\ &= G_{n-p-1}(U_i) \quad \text{elsewhere,}\end{aligned}$$

where

$$U_i = \frac{(n-p-1)^{\frac{1}{2}}(y - \mathbf{x}_i'\hat{\boldsymbol{\beta}})}{\{(1 - \mathbf{x}_i'(X'X)^{-1}\mathbf{x}_i)(n-p)\hat{\sigma}^2 - (y - \mathbf{x}_i'\hat{\boldsymbol{\beta}})^2\}^{\frac{1}{2}}}$$

and $G_\nu(z)$ is a Student's t distribution function with ν degrees of freedom evaluated at z.

THEOREM 3.2. *In the full rank normal regression model, for* $\mathbf{x}_0 \in S$ *and for* $n > p + 1$, *the* MVUE *for the* $N(\mathbf{x}_0'\boldsymbol{\beta}, \sigma^2)$ *distribution function is* $\tilde{F}_{y_0}(y)$ *where,*

$$\begin{aligned}\tilde{F}_{y_0}(y) &= 0 \quad \textit{if } \ y - \mathbf{x}_0'\hat{\boldsymbol{\beta}} < -\{(1 - \mathbf{x}_0'(X'X)^{-1}\mathbf{x}_0)(n-p)\hat{\sigma}\}^{\frac{1}{2}}\,,\\ &= 1 \quad \textit{if } \ y - \mathbf{x}_0'\hat{\boldsymbol{\beta}} \geqq \{(1 - \mathbf{x}_0'(X'X)^{-1}\mathbf{x}_0)(n-p)\hat{\sigma}\}^{\frac{1}{2}}\,,\\ &= G_{n-p-1}(U_0) \quad \textit{elsewhere,}\end{aligned}$$

and

$$U_0 = \frac{(n-p-1)^{\frac{1}{2}}(y - \mathbf{x}_0'\hat{\boldsymbol{\beta}})}{\{(1 - \mathbf{x}_0'(X'X)^{-1}\mathbf{x}_0)(n-p)\hat{\sigma}^2 - (y - \mathbf{x}_0'\hat{\boldsymbol{\beta}})^2\}^{\frac{1}{2}}} .$$

PROOF. Since $\mathbf{x}_0 \in S$, $\exists\, \boldsymbol{\alpha} \in R^n$ such that $||\boldsymbol{\alpha}|| = 1$ and for which $\boldsymbol{\alpha}'\mathbf{Y} \sim N(\mathbf{x}_0'\boldsymbol{\beta}, \sigma^2)$.

Let P be any orthogonal matrix whose first row is $\boldsymbol{\alpha}'$. For $\mathbf{Z} = P\mathbf{Y}$, $\mathbf{Z} \sim N(PX\boldsymbol{\beta}, \sigma^2 I)$, PX is still full rank and T is still the sufficient and complete statistic. The result follows.

Seheult and Quesenberry [3] show that an unbiased estimator of the density of a rv exists iff an unbiased estimator of the corresponding distribution function which happens to be absolutely continuous a.s. exists.

COROLLARY. *In the full rank normal regression model, for* $n > p + 1$ *the* $N(\mathbf{x}_0'\boldsymbol{\beta}, \sigma^2)$ *density is unbiasedly estimable iff* $\mathbf{x}_0'(X'X)^{-1}\mathbf{x}_0 < 1$.

It is interesting to note that for $\mathbf{x}_0 = \mathbf{x}_i$ $i = 1, \cdots, n$, unbiased estimation of the corresponding distributions is always possible; however, it is not necessarily the case that unbiased estimation of the densities is always possible.

In order to characterize existence of unbiased density estimation a result is given:

LEMMA 3.2. $\mathbf{x}_i'(X'X)^{-1}\mathbf{x}_i < 1$ *if the matrix* X *diminished by its* ith *row is still full rank.*

4. Some comments. It is a well-known result that the variance with which one estimates in a regression increases as one moves away from the "center of the design." In the previous result, this fact plays an important role in the sense that the variance increases up to the point where unbiased estimation of the distribution is no longer possible.

It is interesting to note that at $\mathbf{x}_0 = \mathbf{0}$, extrapolation is always valid and, that the corresponding distribution is that of the error.

Finally, for purpose of illustration let $\mathbf{x}_i = (1, z_i)$, i.e. suppose one is dealing with a simple regression.

Note that when extrapolating at $\mathbf{x}_0$, it is not necessary that its first component be equal to 1. If that were the case the values of z for which extrapolation is valid are those obeying:

$$(z - \bar{z})^2 < \frac{n-1}{n} \sum_{i=1}^{n} (z_i - \bar{z})^2 .$$

REFERENCES

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

[2] O'REILLY, F. and QUESENBERRY, C. P. (1973). The conditional probability integral transformation and applications to obtain composite chi-square goodness of fit tests. *Ann. Statist.* **1** 74-83.

[3] SEHEULT, A. H. and QUESENBERRY, C. P. (1971). On unbiased estimation of density functions. *Ann. Math. Statist.* **42** 1434–1438.

CENTRO DE INVESTIGACION EN MATEMATICAS
APLICADAS Y EN SISTEMAS
APARTADO POSTAL 20–726
MEXICO 20, D. F.