The Annals of Mathematical Statistics
1970, Vol. 41, No. 6, 1992–1998

EXACT CONFIDENCE INTERVALS IN REGRESSION PROBLEMS WITH INDEPENDENT SYMMETRIC ERRORS[1]

BY J. A. HARTIGAN

Yale University and University of California, Los Angeles

0. Summary. Subsamples are used to generate confidence intervals for a parameter of a linear regression model, under the assumption that the error variables are independent, continuous and symmetric about 0 in distribution.

1. Introduction. Confidence intervals for the parameters of a linear regression model are usually based on the assumption that the errors are (1) independent, (2) have constant variance, (3) are normally distributed with mean zero. All three of these assumptions may be sufficiently violated in practice to make the confidence intervals significantly misleading. The model considered here assumes that the error variables are independent and symmetric about zero, but does not assume that they are identical. In other words, the normality and homogeneity of variance assumptions are relaxed.

Tests for zero constant and slope parameters in a straight line regression have been considered by Adichie [1] when the error variables are symmetric and identical, and by Daniels [3] with the assumption only that the error variables have zero median. A number of "distribution free" procedures are available for testing the hypothesis of all linear parameters zero in a general regression; for example, the median tests of Brown and Mood [2], the rank tests of Friedman [4], permutation tests, and others. These procedures may be used to generate joint confidence regions for the parameters.

Here a method is proposed for generating confidence intervals for a single parameter, which applies only to certain special but common regression models. The technique used is an extension of those used in Hartigan [6] which perform error analysis of a statistic t by recomputing t for selected subsamples of the data. There, a set of random variables $t_1, \cdots, t_k$ was defined to form a set of typical values for a parameter θ if the ordered variables $t_{(1)}, t_{(2)}, \cdots, t_{(k)}$ are such that the intervals $(-\infty, t_{(1)}), (t_{(1)}, t_{(2)}), \cdots, (t_{(k)}, \infty)$ each include θ with probability $1/(k+1)$. Given a set of typical values, a number of confidence intervals for θ of probability sizes $0, 1/(k+1), 2/(k+1), \cdots, k/(k+1), 1$ are available. For example suppose Y_1, Y_2, Y_3 are independent, continuous and symmetric about μ. Then Y_1, $(Y_2+Y_3)/2$, $(Y_1+Y_2+Y_3)/3$ form a set of typical values for μ; and Y_1, Y_2, Y_3, $(Y_1+Y_2)/2$, $(Y_1+Y_3)/2$, $(Y_2+Y_3)/2$, $(Y_1+Y_2+Y_3)/3$ form a set of typical values for μ. More generally if $Y_1, Y_2, Y_3, \cdots, Y_n$ are independent, continuous and symmetric about μ, if $S_1, S_2, \cdots, S_k$ are subsets of the set $Y_1, Y_2, \cdots, Y_n$ satisfying a certain group

Received April 2, 1969; revised June 9, 1970.

[1] This research supported by ONR grant N00014-67-A-0151-1107 and partly by NIH grant FR-3.

theoretic property, and if $\overline{Y}_{S_i}$ denotes the mean of the random variables in S_i, then $\overline{Y}_{S_1}, \overline{Y}_{S_2}, \cdots, \overline{Y}_{S_k}$ are typical values for μ.

To extend the subsample technique to linear regression, it is necessary to use only subsamples which are *fractions*, a property which depends on the linear model and the parameter of interest. The confidence intervals obtained in this way are not very much wider, on average, than the normal based intervals, if the normality assumption is valid. The existence of fractions, from which the typical values are computed, is demonstrated for certain common linear regression models, including estimation of a mean, estimation of a straight line, n-factor analysis of variance, and comparison of two means.

2. Definition of fractions and typical values. Let $\mathbf{Y}$ be an $n \times 1$ observation vector, let X be an $n \times m$ matrix of constants, let θ be an $m \times 1$ parameter vector. A regression model is expressed in the form

$$E(\mathbf{Y}) = X\theta.$$

An *estimable* linear function $l(\theta) = \sum_{j=1}^{m} a_j \theta_j$ is a linear combination of $\theta_1, \cdots, \theta_m$ which is the expectation of a linear combination of $Y_1, Y_2, \cdots, Y_n$. Let $\hat{\theta}$ be a least squares estimate of θ (minimizing $\sum_i (Y_i - \sum_j X_{ij}\theta_j)^2$). Let $\sum_{j=1}^{m} a_j \hat{\theta}_j = \sum_{i=1}^{n} \lambda_i Y_i$ be the corresponding estimate of $l(\theta)$; $l(\hat{\theta})$ is unique though $\hat{\theta}$ may not be; the constants $\{\lambda_i\}$ are functions of the constants $\{a_j\}$ and the matrix X. A subset S of the observations $Y_1, \cdots, Y_n$ is an *$l(\theta)$-fraction* if the least squares estimate of $l(\theta)$, using only the observations in S, is of form $\alpha \sum_{Y_i \in S} \lambda_i Y_i$. It may be shown that $\alpha = \sum \lambda_i^2 / \sum_{Y_i \in S} \lambda_i^2$. The fraction S is *relevant* if not all λ_i equal zero. A set Σ of subsets is *balanced* if $S_1 \in \Sigma_0$, $S_2 \in \Sigma_0$ implies $S_1 \circ S_2 = (S_1 - S_2) \cup (S_2 - S_1)$ lies in $\Sigma_0 = (\Sigma, \varphi)$; Σ_0 forms a group under the product, symmetric difference, with a unit element equal to the null set φ. Finally, the random variables $Z_1, Z_2, \cdots, Z_k$ form a *set of typical values* for the constant parameter $l(\theta)$ if, the probability that $l(\theta)$ lies in the intervals $(-\infty, Z_{(1)})$, $(Z_{(1)}, Z_{(2)}), \cdots, (Z_{(k)}, \infty)$ is $1/(k+1)$ for each interval; here $Z_{(1)}, Z_{(2)}, \cdots, Z_{(k)}$ denote the ordered values of $Z_1, \cdots, Z_k$.

3. The use of fractions as typical values. The logic of the main theorem follows Fisher's ([4] page 46) sign randomization test. Let $Y_1, Y_2, \cdots, Y_n$ be independent, continuous, symmetric about 0. Consider the 2^n variables $\{\pm Y_1 \pm Y_2 \pm \cdots \pm Y_n\}$. By symmetry the probability that $\sum_{i=1}^{n} Y_i$ is less than exactly k of these variables is 2^{-n}, $k = 0, 1, \cdots (2^n - 1)$. The event $\sum_{Y_i \in S} Y_i < 0$ is equivalent to the event $\sum_{Y_i \notin S} Y_i - \sum_{Y_i \in S} Y_i > \sum Y_i$. Thus the probability that exactly k of $\sum_{Y_i \in S} Y_i$ are less than zero is 2^{-n}; thus the $2^n - 1$ random variables $(\sum_{Y_i \in S} Y_i, S \neq \varphi\}$ form a typical set for 0. More generally,

LEMMA. *Let $Y_1, Y_2, \cdots Y_n$ be independent, continuous, symmetric about 0. Let $S_1, S_2, \cdots, S_k$ be a balanced set. Then $\{\sum_{Y_i \in S_j} Y_i, 1 \leqq j \leqq k\}$ is a typical set for 0.*

This is proved in Hartigan [6]. Or follow the above argument with a subgroup of the group of 2^n sign transformations.

THEOREM 1. Let $\mathbf{Y}$ *satisfy the regression model* $E(\mathbf{Y}) = X\theta$; *let* $l(\theta)$ *be estimable*; *let* $S_1, \cdots, S_k$ *be a balanced set of relevant fractions*; *let* l_i *denote the least squares estimate of* $l(\theta)$ *using the observations in* S_i; *let* $Y_1, Y_2, \cdots, Y_n$ *be independent, continuous, symmetric about their expected values. Then* $l_1, l_2, \cdots, l_k$ *form a set of typical values for* $l(\theta)$.

PROOF. We may assume all $\lambda_i \neq 0$. (Omitting all Y_i with $\lambda_i = 0$ leaves $S_1, \cdots, S_k$ as a balanced set of relevant fractions.) Now define

$$Z_i = \lambda_i(Y_i - EY_i).$$

Then $Z_1, Z_2, \cdots, Z_n$ are independent, continuous, symmetric about 0. From the lemma $\{\sum_{Y_i \in S_j} Z_i,\ 1 \leqq j \leqq k\}$ form a typical set for zero. Since the S_j are fractions, $l_j = \alpha_j \sum_{Y_i \in S_j} \lambda_i Y_i$ for some α_j. Taking expectations, $l_j - l(\theta) = \alpha_j \sum_{Y_i \in S_j} Z_i$. Since each S_j is relevant, $\alpha_j > 0$. So the probability that exactly r of $\alpha_j \sum_{Y_i \in S_j} Z_i$ are less than zero, is the probability that exactly r of $\sum_{Y_i \in S_j} Z_i$ are less than zero, which is $1/(k+1)$. Therefore $\{l_j - l(\theta)\}$ are typical values for 0, or $l_1, \cdots, l_k$ are typical values for $l(\theta)$ as required.

4. Generating balanced sets of fractions.

THEOREM 2. *Let* $S_1, S_2, \cdots, S_k$ *be a balanced set of fractions. Then the Boolean field generated from* $S_1, \cdots S_k$ *(by unions and complements) consists of fractions.*

PROOF. If $\sum_{i=1}^n \lambda_i Y_i$ is the least squares estimate of $l(\theta)$, define $u_i = \lambda_i / \sum_{i=1}^n \lambda_i^2$, and reparametrize the model

$$E[\mathbf{Y}] = l(\theta)\mathbf{u} + X^0\psi,$$

where ψ is a linear transform of θ, and X^0 is an $n \times p$ matrix of rank p with $\mathbf{u}'X^0 = 0$. Define

$$[\mathbf{u}'X^0]_S = \sum_{Y_i \in S} u_i X^0_{ij}.$$

The condition that S be an $l(\theta)$-fraction is equivalent to $[u'X^0]_S = 0$. (The constant α in the estimate $\alpha \sum_{Y_i \in S} \lambda_i Y_i$ is then given by $\sum_{i=1}^n \lambda_i^2 / \sum_{Y_i \in S} \lambda_i^2$.)

Now note that $[u'X^0]_{\bar{S}} = -[u'X^0]_S$

$$[u'X^0]_{S_1 \cap S_2} = [u'X^0]_{S_1} + [u'X^0]_{S_2} - [u'X^0]_{S_1 \circ S_2}$$

and generally, for any Boolean expression in $S_1, S_2, \cdots, S_k$, $[u'X^0]$ may be expressed as a linear combination of $[u'X]$ terms over products of $S_1, S_2, \cdots, S_k$. Since $S_1, S_2, \cdots, S_k$ and their products are fractions, $[u'X^0]_S = 0$ for any Boolean set S; the Boolean field consists of fractions.

This theorem indicates how to search for balanced sets. We will assume $\lambda_i \neq 0$, $1 \leqq i \leqq n$. Define a minimal fraction to be one which has no proper subset as a fraction. Define a base to be a partition of the set of observations into minimal fractions. Then the theorem guarantees that Σ is a balanced set of relevant fractions if and only if (Σ, φ) is a subgroup (under the product, symmetric difference) of the

group of all unions generated from a base. Balanced sets are therefore determined by enumerating bases, or more simply, all minimal fractions.

5. Efficiency of sets of typical values.

THEOREM 3. *Let $S_1, S_2, \cdots, S_k$ be a balanced set of fractions. Let the observations $Y_1, Y_2, \cdots, Y_n$ be independent normal variables with known constant variance, and suppose that each observation appears in at least one fraction. Let l_j, the estimate of $l(\theta)$ based on the observations in S_j, have the same variance, $1 \leqq j \leqq k$. The interval $(l_{(p)}, l_{(k-p+1)})$ is a confidence interval for $l(\theta)$ of probability size $1-2p/(k+1)$. The ratio of length of this interval to the length of the interval based on standard normal theory, is distributed as*

$$(Z_{(k-p+1)}-Z_{(p)})[k/(k+1)]^{\frac{1}{2}}/(z_{k-p+1}-z_p)$$

where $Z_{(1)}, Z_{(2)}, \cdots, Z_{(k)}$ denote the order statistics of a sample $Z_1, Z_2, \cdots, Z_k$ from a unit normal variable Z, and where $P(Z \leqq z_p) = p/(k+1)$.

PROOF. Let $l_j = \alpha_j \sum_{Y_i \in S_j} \lambda_i Y_i$, where $\alpha_j = \sum \lambda_i^2 / \sum_{Y_i \in S_j} \lambda_i^2$. Since l_j has the same variance all j, it follows that α and also $\sum_{Y_i \in S_j} \lambda_i^2$ are the same for all j. If $j \neq m$, $\sum_{Y_i \in S_j \circ Sm} \lambda_i^2 = \sum_{Y_i \in S_j} \lambda_i^2 = \sum_{Y_i \in S_m} \lambda_i^2$, so that $\sum_{Y_i \in S_j \bigcap S_m} \lambda_i^2 = \frac{1}{2} \sum_{Y_i \in S_j} \lambda_i^2$. It follows that l_j and l_m have correlation $\frac{1}{2}$.

If an observation Y_i appears in S_j, it appears in $S_j \circ S_m$ if and only if it does not appear in S_m. Therefore every observation appears $(k+1)/2$ times in the set $S_1, \cdots, S_k$. Thus $\sum_j \sum_{Y_i \in S_j} \lambda_i^2 = (k+1) \sum \lambda_i^2/2$ and $\operatorname{Var} l_j = 2k \sum \lambda_i^2 \operatorname{Var} Y_i/(k+1) = 2k\sigma_1^2/(k+1)$ where σ_1^2 denotes the variance of the estimate of $l(\theta)$ based on all observations. We now set $l_j = \sigma_l(k/(k+1))^{\frac{1}{2}}(Z_j+Z)+l(\theta)$ where $Z_1, Z_2, \cdots, Z_k$, Z are independent unit normal variables. This is justified by noting that $l_1, l_2, \cdots, l_k$ are normal with mean $l(\theta)$, variances $2\sigma_1^2 k/(k+1)$ and correlations 0.5.

The interval $(l_{(p)}, l_{(k-p+1)})$ has length $(Z_{(k-p+1)}-Z_{(p)})\sigma_1(k/(k+1))^{\frac{1}{2}}$; the interval based on normal theory has length $\sigma_1(z_{k-p+1}-z_p)$. The ratio of the lengths is as stated in the theorem.

Exact values of the expected relative length, for small numbers of fractions, are given in Table 1. The subsample intervals may be as much as ten per cent longer than normal intervals. For large k, with fixed probability size $1-2p/(k+1) = \alpha$

$$1+(\pi \exp(z_p^2)(1-\alpha^2)/4-\tfrac{1}{2})k^{-1}+O(k^{-\frac{3}{2}});$$

is an asymptotic expression for the expected relative length, which approaches one

TABLE 1

Expected length of confidence interval of size α based on k typical values, divided by length of normal confidence interval

k \ α	$\frac{1}{2}$	$\frac{3}{4}$	$\frac{7}{8}$	$\frac{15}{16}$
3	1.0858			
7	1.0503	1.0996		
15	1.0262	1.0504	1.0956	
31	1.0133	1.0251	1.0468	1.0866

as $k \to \infty$. The relative length is $1+O(k^{-\frac{1}{2}})$ and approaches one as $k \to \infty$ with probability one.

These exact and asymptotic results show that the subsample intervals are not much longer than the usual intervals if the errors are independent normal with constant variance. The most important requirement is that the fractions used give estimators with constant variance. This requirement is met approximately in many cases; for example, in estimating a mean from a set of n observations, all 2^n-1 subsamples are fractions; and for large n almost all subsample means have variance nearly equal to twice the variance of the sample mean.

6. Examples of fractions and balanced sets.

A. A location parameter.

$$\text{MODEL}: \qquad E(Y_i) = \mu.$$

The least squares estimate of μ is $\hat{\mu} = \bar{Y}$. Any subset of the observations is a fraction. For example, suppose given observations (1.3, 6.2, 1.4, 2.7, 4.3); the 31 subsample means are (1.3, 6.2, 1.4, 2.7, 4.3, 3.8, 1.4, 2.0, 2.8, 3.8, 4.5, 5.3, 2.1, 2.9, 3.5, 2.8, 4.4, 4.0, 3.4, 2.8, 2.3, 1.8, 3.9, 3.4, 3.0, 3.7, 2.4, 3.6, 3.3, 2.9, 3.2). Confidence intervals for μ are SIZE $\frac{1}{16}$—(1.3, 6.2), SIZE $\frac{7}{8}$—(1.4, 5.3), SIZE $\frac{3}{4}$—(1.4, 4.5).

B. Estimation of a straight line.

$$\text{MODEL}: \qquad E(Y_i) = \alpha+\beta x_i\,.$$

The least squares estimates are

$$\hat{\alpha} = \bar{Y}-\hat{\beta}\bar{x}$$

$$\hat{\beta} = \sum(x_i-\bar{x})Y_i/\sum(x_i-\bar{x})^2.$$

The subset S is a β-fraction if $\sum_{Y_i \in S}(x_i-\bar{x}) = 0$. For example with $x = (-5, -4, -3, -2, -1, 0, 1, 2, 3, 4, 5)$, a set of disjoint fractions (which may be combined to give 31 fractions) are $(-5, 5)$, $(-4, 4)$, $(-3, 3)$, $(-2, 2)$, $(-1, 0, 1)$.

C. Two-way analysis of variance.

$$\text{MODEL}: \qquad E[Y_{ij}] = \mu_{i.}+\mu_{.j}, \qquad 1 \leqq 1 \leqq n_I, \qquad 1 \leqq j \leqq n_J.$$

Linear contrasts $\sum a_i\mu_{i.}$ and $\sum b_j\mu_{.j}$ are estimable if $\sum a_j = 0$, $\sum b_j = 0$. A subset S forms a $\sum a_i\mu_{i.}$-fraction if for each $a_i \neq 0$, $Y_{ij} \in S$ implies $Y_{kj} \in S$ for all $a_k \neq 0$. The rows form a set of disjoint fractions from which typical values for column contrasts may be generated—and similarly columns for row contrasts.

D. Comparing two means.

$$\text{MODEL}: \qquad E[Y_i] = \mu_1,\ i = 1, \cdots, n_1$$

$$E[Y_i] = \mu_2,\ i = n_1+1, \cdots, n_2.$$

The least squares estimate of $\mu_1-\mu_2$ is $\sum_{i=1}^{n_1} Y_i/n_1-\sum_{i=n_1+1}^{n_2} Y_i/n_2$.

Let n_1 and n_2 have greatest common denominator d; minimal fractions are subsets containing (n_1/d) observations with expectation μ_1 and n_2/d observations with expectation μ_2. Any disjoint set of k minimal fractions is suitable as a base for a balanced set.

E. Three-factor analysis of variance.

$$\text{MODEL:} \qquad E[Y_{ijk}] = \mu_{i..} + \mu_{.j.} + \mu_{..k} + \mu_{ij.} + \mu_{i.k} + \mu_{.jk}.$$

For an I contrast, the minimal fractions (ignoring subsets irrelevant to that contrast) are the observations $\{Y_{ijk} \mid i = 1, 2, \cdots, n_I\}$, one minimal fraction for each JK interaction. Similarly for an IJ contrast, the minimal fractions are the observations $\{Y_{ijk} \mid i = 1, \cdots, n_I, j = 1, \cdots, n_J\}$ one for each K effect. This pattern extends to n factor models.

7. Concluding remarks. The general purpose of the method is to provide valid confidence intervals for a regression parameter under weak assumptions about the error model.

(1) It will be noted that least squares estimates are used in generating the confidence intervals; these estimates will not be *optimal* under the weak error assumptions, but nevertheless the confidence intervals using them will be *valid.*

(2) Choice of good or best balanced sets of fractions has not been settled. Under the usual normal assumptions, I surmise that the expected length of confidence intervals is smaller for a given balanced set than for any balanced subset of this set. This suggests that only maximal balanced sets should be considered. Computation may be excessive if very large balanced sets exist; if so, a set of typical values may be obtained by selecting at random (without replacement) from a given balanced set, and computing estimates for these randomly selected subsets.

(3) Sets of typical values have a Bayes interpretation; suppose that θ is uniformly distributed a priori and the usual normal assumptions hold with σ fixed. A sufficient statistic for θ is $\hat{\theta}$, the least squares estimate of θ. If $l_1, l_2, \cdots, l_k$ are typical values for $l(\theta)$, given θ fixed, then *conditionally on* $\hat{\theta}$ it may be shown that they form typical values for the random variable $l(\theta)$. The quantities $l_1, l_2, \cdots, l_k$ behave like a random sample from the posterior distribution of $l(\theta)$ given $\hat{\theta}$, in the sense that $l(\theta)$ is less than exactly r of them with the same probability, all r.

(4) Confidence intervals are based on the fact that the error distribution is invariant under sign changes; other invariance relations on the error will produce different sets of typical values. For example, the permutation group, the full orthogonal group. The role of fractions is to specify certain transformations in the groups as relevant to the parameter of interest. The full orthogonal group should recover the usual t-type confidence intervals.

(5) Here is another method, not efficient, but simple to use and valid under the same assumptions about error as the fraction method. Divide the data into k disjoint sets, let $L_1, L_2, \cdots, L_k$ be the corresponding least squares estimates of $l(\theta)$. For efficiency, each L_i should have approximately the same variance. Let $L_{(1)}, L_{(2)}, \cdots, L_{(k)}$ be the ordered estimates. Then $(L_{(i)}, L_{(i+1)})$ contains $l(\theta)$ with

probability $\binom{n}{i}2^{-n}$. Since the L_i are independent with median $l(\theta)$, this is nothing but the nonparametric method for generating confidence intervals for the median due to Thompson [7].

REFERENCES

[1] ADICHIE, J. N. (1957). Asymptotic efficiency of a class of non-parametric tests for regression parameters. *Ann. Math. Statist.* **38** 884–893.

[2] BROWN, G. W. and MOOD, A. M. (1950). On median tests for linear hypotheses. *Proc. Second Berkeley Symp. Math. Statist. Prob.* 159–166. Univ. of California.

[3] DANIELS, H. E. (1954). A distribution free test for regression parameters. *Ann. Math. Statist.* **25** 499–513.

[4] FISHER, R. A. (1935). *The Design of Experiments*, 8th ed. Oliver and Boyd, London.

[5] FRIEDMAN, M. (1937). Use of ranks to avoid the assumption of normality in analysis of variance. *J. Amer. Statist. Assoc.* **32** 675–701.

[6] HARTIGAN, J. A. (1969). Using subsample values as typical values. *J. Amer. Statist. Assoc.* **64** 1303–1317.

[7] THOMPSON, W. R. (1936). On confidence ranges for the median and other expectation distributions for populations of unknown distribution form. *Ann. Math. Statist.* **7** 122–128.