Arab Gulf J. Scient. Res., Math. Phys. Sci., **A6** (3), pp. 329-335 (1988)

Prediction Limits in the Balanced Random One-Way Model

Mohamed M.T. Limam

Department of Agricultural Economics and Rural Sociology,
College of Agriculture, King Saud University,
P.O. Box 2460 Riyadh 11451, Saudi Arabia

ABSTRACT. Prediction limits are developed for a future observation of the random one-way model. If the variance ratio is known, one may obtain exact prediction limits. For settings where the variance ratio is unknown, we describe a procedure based on the Satterthwaite approximation. The exact confidence level of the approximate limits has been studied by computer simulation experiments. This study shows that this approximation is acceptable for a wide range of cases.

1. Introduction

Prediction limits in regression analysis are discussed in many regression textbooks [see for example, Neter, Wasserman and Kutner (1983)]. Prediction limits for a univariate normal distribution are treated by Whitmore (1986) to provide easier transition to prediction limits in regression, and as a useful method in its own right.

Here we discuss a procedure for obtaining prediction limits for an observation from an infinite population on the basis of a sample of m independent observations from each of n randomly chosen units. It is assumed that this sample satisfies the random one-way model

$$Y_{ij} = \mu + a_i + e_{ij}, \; i = 1, \ldots, n, \; j = 1, \ldots, m, \tag{1.1}$$

where Y_{ij} denotes the j^{th} observation from the i^{th} unit, μ the overall mean, $\mu + a_i$ the mean of the i^{th} unit, and e_{ij} is a random deviation. We assume that the a_i's and e_{ij}'s are independently distributed as normal variates with means zero and

variances σ_a^2 and σ_e^2, respectively. The random variables Y_{ij} are distributed $N(\mu, \sigma_a^2 + \sigma_e^2)$, and the grand average $\hat{\mu} = \Sigma \Sigma Y_{ij}/nm$ is distributed $N(\mu, [m\sigma_a^a + \sigma_e^2]/nm)$. Let MS_b denote the between units mean square and MS_w the within units mean square, then $(MS_b - MS_w)/m$ and MS_w are unbiased estimators of σ_a^2 and σ_e^2 respectively.

As a practical example, we consider the following situation. Some agricultural surveys are conducted over the years to estimate the yield of a given crop. For planning purposes, an inference of interest to the agronomist is the prediction of future yield of the yearly crop. In the statistics division, yield data is classified by varieties over many years. A two-stage sampling is designed to select n varieties, as primary sampling units, and m yields, as secondary sampling units. Model (1.1) is assumed and prediction limits are required for a future yield, y_0, drawn from $N(\mu, \sigma_a^2 + \sigma_e^2)$.

In Section 2, we derive exact prediction limits for y_0, when the variance ratio $R = \sigma_a^2/\sigma_e^2$ is known, and approximate limits when R is unknown. Section 3 investigates the exact confidence level of the approximate prediction limits through a simulation study.

2. Development of the Prediction Limits

The variance of the prediction error is

$$\sigma_p^2 = \sigma_e^2 \, k(R, n, m),$$

where $k(R, n, m) = (1 + 1/nm) + R(1 + 1/n)$. If R is known, $\hat{\sigma}_p^2/\sigma_p^2$, where $\hat{\sigma}_p^2 = \hat{\sigma}_e^2 \, k(R, n, m)$, is distributed independently of μ, and as a known multiple of a $\chi^2_{n(m-1)}$ variate. Then, the statistic

$$T = (\hat{\mu} - y_0)/\hat{\sigma}_p$$

has a t-distribution with $n(m-1)$ degrees of freedom (df). Hence, the $(1 - \alpha)$ prediction limits for y_0 are

$$\hat{\mu} - t_{(n(m-1),\ 1-\alpha/2)}\hat{\sigma}_p, \ \hat{\mu} + t_{(n(m-1),\ 1-\alpha/2)}\hat{\sigma}_p,$$

where $t_{(n(m-1),\ 1-\alpha/2)}$ is the $100(1 - \alpha/2)$ percentile of a t-distribution with $n(m-1)$ df.

If R is unknown, an unbiased estimator of σ_p^2 is

$$\hat{\sigma}_p^2 = MS_w(m - 1)/m + MS_b(n + 1)/nm.$$

The distribution of $\hat{\sigma}_p^2$ is that of a linear combination of the independent chi-squared variables, MS_w and MS_b. Satterthwaite (1946) suggested that, if the linear combination of the mean squares MS_i, $i = 1, \ldots, r$ is $\widetilde{MS} = \Sigma\, c_i MS_i$, such that $E(MS_i) = \sigma_i^2$, $E(\widetilde{MS}) = \sigma^2$, and f_i are the df of MS_i for $i = 1, \ldots, r$, then the distribution of $\widetilde{MS}/\sigma^2$ is approximated by χ_f^2/f, where

$$f = \sigma^4/[\Sigma\, c_i^2\, E(MS_i)^2/f_i)].$$

Using this result, $\hat{\sigma}_p^2/\sigma_p^2$ is approximately distributed as a χ_f^2/f variate where

$$f = [k(R, n, m)]^2/[(m - 1)/nm^2 + (n + 1)^2\,(1 + mR)^2/n^2m^2(n - 1)] \tag{2.1}$$

Hence, the distribution of T is approximated by a t-distribution with df, f, given in (2.1). Then, the $(1 - \alpha)$ prediction limits for y_0 are

$$\hat{\mu} - t_{(f,\,1-\alpha/2)}\,\hat{\sigma}_p,\ \hat{\mu} + t_{(f,\,1-\alpha/2)}\hat{\sigma}_p \tag{2.2}$$

It is informative to investigate t_f as a function of R. As R tends to infinity, t_f approaches $t_{(n-1)}$, and the prediction limits correspond to those of a random sample of size n. In this case, essentially all the variation is between units, and sampling within a unit provides no additional information. As $R \to 0$, we have $t_f = t_{(nm-1-\delta)}$, where $\delta = 4n(m-1)/[n(n-1)(m-1) + (n+1)^2]$. Then, except for the small term δ, at $R = 0$, the prediction limits in (2.2) correspond to those of a random sample of size nm, discussed by Whitmore (1986). Therefore, for $0 < R < \infty$, we have $t_{(nm-1)} < t_f < t_{(n-1)}$.

Since the df, f, in (2.1) depend on the unknown variance ratio R, one might consider an estimate of R,

$$\hat{R} = \text{maximum } \{0,\ (F - 1)/m\},$$

where $F = MS_b/MS_w$. If $F < 1$, $\hat{R} = 0$, we proceed as thought the nm observations are a simple random sample from the population and use prediction limits in (2.2) with $f = nm - 1$. When f is not integer, either it is rounded to the nearest integer and t-values obtained directly from t-tables, or by interpolation which is also considered to be acceptable.

3. Simulation study

We computed the probabilities

$$\beta(R) = \Pr\{|T| < t_{(f,\,1-\alpha/2)}\,|\hat{\mu}, \hat{R}\}, \tag{3.1}$$

for a variety of cases to investigate the exact confidence level of the approximate prediction limits in (2.2). Since the t-distribution depends only on f, which depends on R, without loss of generality, we take $\mu = 0$ in (1.1), $\sigma_e^2 = 1$ and different values for R. The combinations used here are n = 10, 20, and R = 0,25, 1, 4. They are suggested by Seeger and Thorsson (1973) in their Monte Carlo studies for tolerance intervals in the random one-way model.

A Fortran Program was written with the use of IMSL subroutines, and for each combination, 10,000 independent samples were generated. Table 1 gives the probabilities ß(R), defined in (3.1), for the nominal confidence levels ß = .90, .95, and .99. The variability of estimates of the true confidence levels is $1.96[ß(1-ß)/10{,}000]^{1/2}$ = .006, .004, .002 for ß = .90, .95, and .99, respectively. It can be seen from Table 1 that ß(R) are reasonably accurate over all the cases.

Table 1. Exact confidence levels for the approximate prediction limits

n	m	R	ß		
			.90	.95	.99
10	2	.25	.905	.952	.990
		1	.899	.949	.989
		4	.897	.947	.987
	5	.25	.904	.947	.989
		1	.898	.946	.987
		4	.895	.946	.988
	10	.25	.903	.950	.988
		1	.898	.947	.988
		4	.896	.945	.988
20	2	.25	.904	.949	.990
		1	.899	.948	.987
		4	.900	.949	.989
	5	.25	.903	.949	.989
		1	.900	.949	.988
		4	.899	.951	.988
	10	.25	.902	.951	.989
		1	.901	.949	.987
		4	.898	.950	.988

Along those presented in Table 1, we investigated ß(R) for a wide range of values of R. We found that ß(R) approaches ß as R tends to infinity and exceeds ß for R sufficiently small. However, for intermediate values of R, ß(R) is less than ß. Eventhough a correction for this situation is not needed here one might use an upper η confidence bound for R (Searle 1971, p. 414),

$$R^* = \text{maximum } \{0, (FF_\eta - 1)/m\}$$

where F_η is 100η percentile of an F-distribution with $n(m-1)$ and $(n-1)$ df, and a value of η chosen to guarantee $\beta(R) \geq \beta$, for all n, m, and R. Obviously, it will result in a conservative prediction limits.

Besides the simulation study, using normality asumptions of model (1.1), it is also informative to check the robustness of the proposed prediction limits. Hahn (1970) developed prediction limits for s future observations from a normal population, and mentioned that the correctness of the limits is likely to be sensitive to deviations from normality. Here, we used 10,000 replications to investigate the prediction limits in (2.2) when e_{ij}'s are χ^2_4, and Y_{ij} are sampled from χ^2_8, for $n = 10$, $m = 5$, and $R = .25, 1, 4$. The exact confidence levels of the limits in (2.2) with $Y_{ij} \sim \chi^2_8$ do not depend on R. Table 2 gives probabilities $\beta(R)$ when $e_{ij} \sim \chi^2_4$ and $Y_{ij} \sim \chi^2_8$ with $n = 10$ and $m = 5$. These results show that prediction limits in (2.2) are sensitive to the violations investigated for either a large or a small confidence level. For the nominal confidence level $\beta = .95$ the results are accurate. However this short study is not a clear answer to the robustness property of the t-based prediction limits, a thorough investigation is required.

Table 2. Exact confidence levels for the approximate prediction limits where $e_{ij} \sim \chi^2_4$, and $Y_{ij} \sim \chi^2_8$ with $n = 10$, $m = 5$

	R	ß		
		.90	.95	.99
$e_{ij} \sim \chi^2_4$	.25	.928	.952	.978
	1	.923	.952	.980
	4	.908	.949	.984
$Y_{ij} \sim \chi^2_8$	–	.916	.951	.981

This study has been performed because it is difficult to construct exact prediction limits for the random one-way model as there is no exact distribution of $\hat{\sigma}^2_p$ when the variance ratio R is unknown.

4. Extensions

The extension of prediction limits in (2.2) to include the case of predicting the mean of s new observations from the $N(\mu, \sigma^2_a + \sigma^2_e)$ population is straightforward.

For the example discussed in Section 1, we might consider the case where there is one (or more) covariate measured with each yield, such as rainfall. The random one-way model with one covariate,

$$Y_{ij} = \mu + a_i + b_i x_{ij} + e_{ij},\ i = 1, \ldots, n,\ j = 1, \ldots, m$$

is assumed, where x_{ij} is the rainfall associated with variety i in year j. It is of interest to extend prediction limits in (2.2) to the case of predicting y_o, and the mean of s new observations, given $x_{ij} = x_o$.

References

Hahn, G.J. (1969) Factors for calculating two-sided prediction intervals for samples from a normal distribution, *Journal of the American Statistical Association*, **64:** 878-888.

Neter, J., Wasserman, W. and **Kutner, M.H.** (1983) Applied linear regression models, first edition, Richard D. Irwin, Inc.

Satterthwaite, F.E. (1946) An approximate distribution of estimates of variance components, *Biometrics*, **3:** 110-114.

Searle, S.R. (1971) Linear Models, New York, Wiley.

Seeger, P. and **Thorsson, U.** (1973) Two-sided tolerance limits with two-stage sampling from normal populations - Monte Carlo studies of the distribution of coverages, *Journal of Royal Statistical Society*, **22**(33): 292,300.

Whitmore, G.A. (1986) Prediction limits for a Univariate normal distribution, *The American Statistician*, **40**(2): 141-143.

(Received 12/10/1987;
in revised form 22/10/1988)

فترات التنبؤ للنموذج العشوائي المتزن ذو اتجاه واحد

محمد محمد الطاهر الإمام

قسم الاقتصاد الزراعي والمجتمع الريفي ـ كلية الزراعة ـ جامعة الملك سعود
ص. ب ٢٤٦٠ الرياض ١١٤٥١ ـ المملكة العربية السعودية

تم ايجاد فترات التنبؤ لمشاهدة مستقبلية للنموذج العشوائي المتزن ذو اتجاه واحد. عندما تكون نسبة التباين معروفة بالإمكان أن نحصل على فترات تنبؤ صحيحة، وفي حالة عدم معرفة نسبة التباين نصف طريقة ترتكز على عملية Satterthwaite التقريبية. وقعت دراسة مستوى الثقة لهذه الفترات التقريبية بطريقة المحاكاة واثبتت هذه الدراسة أن العملية التقريبية مقبولة في عدة حالات.