NESTED ERROR MODELS FOR THE OBJECTIVE YIELD SURVEY

Mark Reiser[1], Iowa State University, Ron Fecso, USDA, and Kathleen M. Taylor, AT&T
Mark Reiser, College of Business, Arizona State Univ., Tempe, Arizona 85287

The National Agricultural Statistics Service (NASS), formerly the Statistical Reporting Service, of the U. S. Department of Agriculture conducts surveys on many crops grown in the United States in order to make predictions about the current year's crop. Data are gathered each month during a crop's growing season. Data on corn are collected in the ten major corn producing states: Illinois, Indiana, Iowa, Michigan, Minnesota, Missouri, Nebraska, Ohio, South Dakota, and Wisconsin.

The data are collected through a two-phase process. In the first phase, information on land use is gathered during the June Enumerative Survey (JES). In the second phase, data from the JES are used to design the Objective Yield (OY) Survey, where information is recorded about crop growth. These surveys are discussed in Section II.

Based on measurements taken at certain times during the growing season, forecasting models are used to predict the number of ears of corn at harvest, grain weight per ear at harvest, and gross yield per acre at harvest. The models used by NASS are discussed in Section III.

In constructing the forecasting models, the relationships between early season plant characteristics such as length of the cob and final season values such as grain weight are established from historical data. Some important questions in establishing the relationship are (1) form of the models, (2) difference in models by maturity class and month, and (3) how many years of historical data to include. NASS currently uses the five years preceding the current crop year. Whenever observations are used from more than one year, a cluster effect is present, since observations within a year would show some degree of correlation due to common influences such as weather. A model which specifies an error structure with intrinsic error nested within years is presented in Section IV. Results are presented using data for corn in the state of Iowa for the years 1979 and 1984.

II. THE JUNE ENUMERATIVE AND OBJECTIVE YIELD SURVEYS

The June Enumerative Survey (JES) is conducted by NASS every year in each of the forty-eight contiguous states during late May and early June. The purpose of the survey is to collect information on land use. Area sampling frames are developed individually for each state based on the land use, economy, and agricultural practices of that state.

From the JES data, the number of acres within a state planted to corn or to be planted to corn can be estimated and locations of sample acres can be determined. This information is then used to develop a sampling frame for the Objective Yield Survey. Selection of the Objective Yield sample is a two-step process. In the first step, fields are chosen at random so that the probability of any given field being included in the sample is proportional to the size of that field. Larger fields have a higher probability of being selected and may be selected more than once in the sample. Selected fields are numbered from one to N, where N is the total number of samples in the state.

In the second step in selecting the Objective Yield sample, selected fields are visited and two plots are randomly located within each field. The plots are the sampling units for the Objective Yield Survey and each will be fifteen feet long and will contain two rows of corn. The size of these sampling units depends on the amount of space between the two rows of corn.

A maturity stage is assigned to each sample field. Maturity stage ranges from one to seven and indicates the stage of growth of the corn. Maturity stage one indicates stalks without silked ear shoots, and maturity stage seven indicates corn ready for harvest.

III. CURRENT METHODS - THE ORDINARY LEAST SQUARES MODEL

3.1 Introduction

Estimation in corn yield surveys involves two components: number of ears and average grain weight per ear. As the corn grows through the season, it is classified into various maturity stages. For each maturity stage, the two components are estimated separately and then combined to give an estimate of gross yield. In the forecast for gross yield, number of ears and grain weight may be taken from either a historical average, a predicted value, or an actual value, according to maturity stage of the corn.

3.4. Forecasting models

Let

$\hat{Y}_{Nijm}$ = predicted number of ears for sample location ij when the corn is at maturity stage m , and

$\hat{Y}_{Wijm}$ = predicted grain weight for sample location ij when the corn is at maturity stage m ,

A. Predicted number of ears (m = 1 to 4 only)

Two forecast models are used for number of ears:

$$\hat{Y}_{Nijm} = \hat{a}_m + \hat{b}_m X_{1ijm} \qquad (m = 1, 2, 3, 4)$$

and

$$\hat{Y}'_{Nijm} = \frac{X_{2ijm}}{\hat{a}'_m + \hat{b}'_m (X^{-1}_{1ijm} X_{3ijm})} \qquad (m = 2, 3, 4)$$

[1]Now at Arizona State University

where

X_1 = number of stalks

X_2 = number of ears and silked ear shoots

X_3 = number of stalks with ears or silked ear shoots

B. Predicted grain weight (M = 3 to 6 only).

Two forecast models are used for grain weight also:

$$\hat{Y}_{Wijm} = \hat{a}_m + \hat{b}_m X_{4ijm} \quad (m = 3, 4, 5, 6)$$

and

$$\hat{Y}'_{Wijm} = \hat{a}'_m + \hat{b}'_m X_{5ijm} \quad (m = 3, 4, 5, 6)$$

where

X_4 = total kernel row length.

X_5 = average length over husk.

When two forecast models exist for Y_N or Y_W, a single value is obtained from a weighted combination. The Final predictors for number of ears and grain weight for each sample location are as follows:

$$\hat{Y}^{(1)}_{Nijm} = \begin{cases} \hat{Y}_{Nijm} & \text{if } m=1 \\ C_1\hat{Y}_{Nijm} + C_2\hat{Y}'_{Nijm} & \text{if } m=2,\ 3,\ 4 \end{cases}$$

$$Y^{(1)}_{Wijm} = \begin{cases} Y_W & \text{if } m=1 \text{ or } 2 \\ C_1\hat{Y}_{Wijm} + C_2\hat{Y}'_{Wijm} & \text{if } m=3,\ 4,\ 5, \text{ or } 6 \end{cases}$$

where C_1 is a function of R^2 for model 1, and $C_2 = 1 - C_1$.

Then the predicted gross yield per sample unit is calculated as

$$\ddot{Y}_{ijm} = \begin{cases} \hat{Y}^{(1)}_{Nijm}\tilde{Y}_W & \text{if } m=1,\ 2 \\ \hat{Y}^{(1)}_{Nijm}\hat{Y}^{(1)}_{Wijm} & \text{if } m=3,\ 4 \\ Y_{Nij}\hat{Y}^{(1)}_{Wijm} & \text{if } m=5,\ 6\ , \end{cases}$$

and the predicted gross yield per sample field is given by

$$\hat{Y}_{ijm} = \hat{\ddot{Y}}_{ijm} K S^{-1}_{ij} \quad \text{for } m=1,\ldots,6\ .$$

where

S_{ij} = 8-row space for field ij.

The 8-row space is unit 1 four row width plus unit 2 four row width. K is a constant conversion factor used to adjust gross yield for the sample units to "Bushels per acre":

$$K = (CD)^{-1}AB = 103.714$$

where

A = the number of square feet per acre = 43560

B = conversion for 1 60-foot row measurement to 8 row equivalent

C = the combined length of 2 sample units (4 rows * 15 feet) = 60

D = bushels per pound = 56

C. Estimation

Under the procedures presently used by the National Agricultural Statistics Service, ordinary least squares is used to obtain the estimated intercept and slope parameters in the forecast equations. The forecasts for year i^* are based upon the pooled data from the previous five years, years $i^* - 5$ to $i^* - 1$.

The ordinary least squares error structure is also assumed, which implies that the n_i observation for year i and treated as independent. As will be discussed later, this is probably not a realistic assumption. We note the following points regarding the predictors for grain weight and number of ears:

(1) The predictor for Y_N under maturity classes 2, 3, and 4 is ill-advised because R^2 for $\hat{Y}_N$ is not comparably to R^2 for $\hat{Y}'_N$. The former is in units of number of ears, while the latter is in the units of the ratio of number of ears to number of stalks. A higher R^2 for one model does not mean it is a better predictor than the other.

(2) As mentioned previously, it is not realistic to assume that observations within years are independent, because they share many influences, such as weather, that are not shared by observations across years.

(3) $\hat{Y}^{(1)}_N$ for maturity classes 2 to 4, and $\hat{Y}^{(1)}_W$ for maturity classes 3 to 6 are not Best Linear Unbiased. The Best Linear Unbiased Estimator of grain weight from two predictors is

$$\hat{Y}^{(2)}_{Wijm} = \hat{a}_m + \hat{b}_{1m}X_{4ijm} + \hat{b}_{2m}X_{5ijm} + $$

$$+ \hat{b}_{3m}X_{4ijm}X_{5ijm} \quad .$$

A Best Linear Unbiased Estimator of number of ears that corresponds to $Y^{(1)}_N$ would be

$$\hat{Y}^{(2)}_{Nijm} = \hat{a}_m + \hat{b}_{1m}X_{1ijm} + \hat{b}_{2m}X_{2ijm} + $$

$$+ \hat{b}_{3m}X_{3ijm} + \hat{b}_{4m}X^{-1}_{1ijm}X_{3ijm} \quad .$$

IV. COMPONENTS OF VARIANCE MODEL WITH NESTED ERROR

The forecasting models described in the previous sections assumed an ordinary least squares error structure for observations that are pooled from the five years preceding the current crop year. A more realistic model would specify that observations within years are correlated, since they are subject to many of the same influences such as weather.

In the preceding discussion, Y_W was used as a symbol for grain weight, Y_N was the symbol for number of ears, and Y was the symbol for gross yield. We now use Y as a generic dependent variable, so that it may represent grain weight, number of ears, or gross yield. A model with dependence among observations within years can then be written as follows:

$$Y_{ijm} = X_{ijm}\underset{\sim}{\beta}_m + u_{ijm} \; , \quad i=1, \ldots, h \; ; \quad j=1, \ldots, n_i$$

$$u_{ijm} = v_{im} + e_{ijm}$$

where

Y_{ijm} = the value for the dependent variable at sample location ij when the corn is at maturity stage m .

X_{ijm} = a row vector of independent variables for the j-th field in year i at maturity stage m .

e_{ijm} = intrinsic error, and

v_{im} = random effect for year i , maturity stage m .

The error in the equation for the ij-th observation is the sum of the random year effect, v_{im} , and the intrinsic error, e_{ijm} . The v_{im} are assumed to be distributed $NID(0, \sigma^2_{vm})$, the e_{ij} are distributed $NID(0, \sigma^2_{em})$, and the e_{ijm}'s are independent of the v_{im}'s for all ij .

The covariance structure of the u_{ijm}'s has the form

$$E(u_{ijm}u_{\ell km}) = \begin{cases} \sigma^2_{vm} + \sigma^2_{em} & \text{if } i=\ell \text{ and } j=k \\ \sigma^2_{vm} & \text{if } i=\ell \text{ and } j\neq k \; . \\ 0 & \text{if } i\neq\ell \end{cases}$$

Under the ordinary least squares models of the preceding sections, the v_{im} are all assumed to be zero, which gives an error covariance matrix of $\sigma^2_{em}I$. For the nested error model, it follows that the error covariance matrix for corn at maturity stage m is block diagonal, with diagonal element $V_{im}(i=1, \ldots, h)$ given by:

$$V_{im} = \sigma^2_{em}I_{n_{im}} + \sigma^2_{vm}J_{n_{im}}$$

where $I_{n_{im}}$ is an $n_{im} \times n_{im}$ identity matrix and $J_{n_{im}}$ is an $n_{im} \times n_{im}$ matrix of ones.

This nested error model is closely related to models developed by Battese and Fuller (1982) and Battese, Harter, and Fuller (1986) for small area estimation.

4.1. Prediction for Cluster Means

The prediction of cluster means is often of interest. For a particular time in year i we may be interested in the prediction for corn that is at maturity stage m . The mean for such corn would be given by:

$$\bar{Y}_{i.m} = \bar{X}_{im}\underset{\sim}{\beta}_m + v_{im} \; , \quad i=1, 2, \ldots, h$$

where $\bar{X}_i = n^{-1}_{im}(\Sigma^{n_{im}}_{j=1}X_{ijm})$ is a k x 1 vector.

n_{im} = number of sample locations at maturity stage m at the particular time of interest

$\underset{\sim}{\beta}_m$ = the k x 1 vector of coefficients from a GLS regression

and

v_{im} = the year effect.

If $\bar{X}_{im}$, $\underset{\sim}{\beta}_m$, σ^2_{vm} and σ^2_{em} are assumed known,

$\hat{v}_{im}$ can be written as

$$\hat{v}_{im} = \delta_{im}\bar{u}_{i.m} \; , \quad 0 < \delta_{im} < 1 \; .$$

where

$$\bar{u}_{i.m} = \bar{Y}_{i.m} - \bar{X}_{i.m}\underset{\sim}{\beta}_m \; .$$

The mean squared error for this predictor is

$$E[(v_{im} - \hat{v}_{im})^2] = (1 - \delta_{im})^2\sigma^2_{vm} +$$

$$+ (\delta^2_{im})(\sigma^2_{em}/n_{im}) .$$

Using this mean squared error and the usual method of differentiation, one obtains the following value of δ_{im} that will give the minimum mean squared error:

$$\hat{\delta}_{im} = \frac{\sigma^2_{vm}}{\sigma^2_{vm} + (\sigma^2_{em}/n_{im})} .$$

The preceding derivations assumed that $\underset{\sim}{\beta}_m$, σ^2_{vm}, and σ^2_{em} were known. In practice, they usually are not known and must be estimated.

If we let $\bar{Y}_*$ be the mean for corn that is observed at maturity stage m at a certain time during the current year, then we know that

$$\bar{Y}_* = \bar{X}_{*m}\underset{\sim}{\beta}_m + v_{*m} .$$

Since the mean of v_{*m} is assumed to be zero, the best linear predictor for $\bar{Y}_*$ is given by

$$\hat{\bar{Y}}_* = \bar{X}_{*m}\hat{\underset{\sim}{\beta}}_m$$

where $\bar{\underset{\sim}{\beta}}_m$ is the best linear unbiased estimator of $\underset{\sim}{\beta}_m$.

The predictor has, approximately, the following mean squared error:

$$E[\bar{Y}_* - \hat{\bar{Y}}_*)^2] = \bar{X}_{*m}V(\hat{\underset{\sim}{\beta}}_m)\bar{X}'_{*m} + \sigma^2_{vm} .$$

V. Results

Tables 3 and 4 show the generalized least squares estimates for the nested error model. There is little difference in the parameters or their standard errors under ordinary least squares estimation. Tables 3 and 4 also show estimates for the components of variance. We can use these estimates to estimate ${}_{im}$, the multiplier of $\bar{\underset{\cdot}{u}}$ that gives the year effect. These estimates are given in Table 2 For grain weight, the multiplier is nearly 1.0 except in maturity category three, where frequencies are relatively low. For number of ears, the multiplier is lower, indicating less year to year variation in the equation error. The same information can be presented as the intraclass correlation, which gives the percent of the total error variance that is due to year to year variation:

$$\rho_m = \frac{\sigma^2_{vm}}{\sigma^2_{vm} + \sigma^2_{em}} .$$

The intraclass correlations are given in Tables 3 and 4 Once again, the year to year effect is much more substantial for grain weight than number of ears. $\hat{\rho}$ ranges from 0.14 to 0.43 for grain weight, and from 0.02 to 0.10 for number of ears. They are directly comparable for maturity category four, where 27% of the equation error for grain weight is due to year variation, while only 2% is year variation for number of ears.

The error in a forecast for the current crop year is approximately,

$$E[(\bar{Y}_* - \hat{\bar{Y}}_*)^2] = \bar{X}_{*m}V(\hat{\underset{\sim}{\beta}}_m)\bar{X}'_* + \sigma^2_{vm} .$$

Since we can always reduce $V(\hat{\underset{\sim}{\beta}}_m)$ by including more years or more sample locations, σ^2_{vm} represents a lower bound on the error of the forecast. If we use $2\hat{\sigma}_{vm}$ has the half-width of an interval, we can compare the mean actual values to the mean predicted values given in Table 1 With this procedure, we find that the mean predicted value is already within this distance of the true value except for number of ears at both maturity values two and three. Although these forecasting results are for only one year, it appears that the sampling of additional locations would not improve the forecast for grain weight based on plant characteristics - the error of the forecast is already within the immutable limit determined by year to year variation.

In the forecast for number of ears from the August Objective Yield Survey, only one-half of the sample locations are visited (approximately 120). This means that there is a lower number of observations for maturity classes one and two, and since $\hat{\bar{Y}}_N$ is further than $2\sigma_{v2}$ from $\bar{Y}_N$, a significant improvement in the precision of the forecast at maturity stage two should result if the number of sample locations is increased.

REFERENCES

Battese, G. E., and W. A. Fuller. 1982. An error components model for prediction of county crop areas using survey and satellite data. Working papers in econometrics and applied statistics. Department of Economic Statistics, University of New England, Armidale, N.S.W., Australia.

Battese, G. E., R. M. Harter, and W. A. Fuller. 1986. An error components model for prediction of county crop areas using survey and satellite data. Submitted for publication.

Hidiroglou, M. A., W. A. Fuller, and R. D. Hickman. 1980. SUPER CARP. 6th ed. Survey Section, Statistical Laboratory, Iowa State University, Ames, Iowa.

Table 1 Predicted and Actual Means by Maturity Class

		Grain weight		Number of ears		Gross yield	
m	n	Actual mean	Predicted mean	Actual mean	Predicted mean	Actual mean	Predicted mean
1	26	$\bar{Y}_W = 0.30$	$\tilde{Y}_W = 0.37$	$\bar{Y}_N = 82.62$	$\bar{\hat{Y}}_{N1} = 79.71$	$\bar{Y} = 109.02$	$\bar{\hat{Y}} = 128.94$
2	71	$\bar{Y}_W = 0.34$	$\tilde{Y}_W = 0.37$	$\bar{Y}_N = 85.46$	$\bar{\hat{Y}}_N^{(1)} = 79.31$	$\bar{Y} = 132.37$	$\bar{\hat{Y}} = 134.67$
3	4	$\bar{Y}_W = 0.29$	$\bar{\hat{Y}}_N^{(1)} = 0.33$	$\bar{Y}_N = 72.00$	$\bar{\hat{Y}}_N^{(1)} = 73.62$	$\bar{Y} = 103.72$	$\bar{\hat{Y}} = 117.79$
4	104	$\bar{Y}_W = 0.33$	$\bar{\hat{Y}}_N^{(1)} = 0.36$	$\bar{Y}_N = 83.66$	$\bar{\hat{Y}}_N^{(1)} = 82.70$	$\bar{Y} = 125.06$	$\bar{\hat{Y}} = 136.82$
5	71	$\bar{Y}_W = 0.33$	$\bar{\hat{Y}}_N^{(1)} = 0.35$	----		$\bar{Y} = 127.94$	$\bar{\hat{Y}} = 135.14$
6	79	$\bar{Y}_W = 0.33$	$\bar{\hat{Y}}_N^{(1)} = 0.35$	----		$\bar{Y} = 121.03$	$\bar{\hat{Y}} = 129.53$

Table 2 δ_{im} by year

	1979	1980	1981	1982	1983	1984
Grain Weight						
m = 3	0.90	0.79	0.84	0.94	0.79	0.75
m = 4	0.98	0.96	0.98	0.98	0.96	0.98
m = 5	0.96	0.97	0.97	0.96	0.96	0.96
m = 6	0.96	0.92	0.95	0.96	0.92	0.94
Number of Ears						
m = 1	0.76	0.19	0.19	0.83	0.75	0.76
m = 2	0.86	0.90	0.90	0.88	0.88	0.87
m = 3	0	0	0	0	0	0
m = 4	0.76	0.62	0.75	0.74	0.64	0.72

Table 3 Nested error regression models for number of ears by maturity stage

Maturity stage	Number of observations	Parameter Estimates with Standard Errors							
		β_0	X_1	X_2	X_3	$(X_1/X_3)X_2$	$\hat{\sigma}^2_v$	$\hat{\sigma}^2_e$	$\hat{\rho}$
1	84	15.3* (4.1)	0.771* (0.048)				5.7432 (2.39)	50.181	0.103
2	409	8.6* (2.6)	0.655* (0.083)	-0.016 (0.086)	0.05 (0.10)	0.148* (0.067)	7.0824	80.201	0.08
3	44	6.5* (3.1)	-0.24 (0.85)	-0.10 (0.75)	0.99 (0.90)	0.23 (0.71)	0	19.409	0
4	483	3.3* (1.3)	-0.35 (0.31)	0.01 (0.26)	1.09* (0.31)	0.16 (0.25)	0.59428 (.77)	24.531	0.0236

*Significant at the 5% level.

Table 4 Nested error regression models for grain weight per ear by maturity stage

Maturity stage	Number of observations	Parameter estimates with standard errors						
		β_0	X_4	X_5	$(X_4)(X_5)$	$\hat{\sigma}^2_v$	$\hat{\sigma}^2_e$	$\hat{\rho}$
1	967	0.3681 (0.0029)						
2	967	0.3681 (0.0029)						
3	42	0.45 (0.81)	-0.008 (0.023)	-0.011 (0.095)	0.0009 (0.0027)	0.0026808	0.0035247	0.432
4	469	-0.0001 (0.0989)	0.0046 (0.0030)	0.026* (0.012)	-0.00006 (0.00035)	0.0015378	0.0041404	0.2708
5	401	-0.32* (0.12)	0.0119* (0.0034)	0.065* (0.014)	0.00093* (0.00040)	0.00091701	0.0030041	0.2339
6	495	0.09 (0.13)	-0.0007 (0.0036)	0.016 (0.015)	0.00060 (0.00043)	0.00056895	0.0034008	0.1433

*Significant at the 5% level.