USING SAS/IML® SOFTWARE TO FORECAST CORN AND SOYBEAN YIELDS

Thomas R. Birkett, National Agricultural Statistics Service, USDA

ABSTRACT

The United States Department of Agriculture uses SAS/IML® software to forecast corn and soybean yields. The statistical model used is an example of a linear model with parameters estimated under a linear restriction. It is also an example of combining several linear models into one global model so that the individual models satisfy a global restriction. The matrix representation of this model is programmed directly into SAS/IML software.

CORN AND SOYBEAN SURVEY VARIABLES

Introduction

The National Agricultural Statistics Service (NASS), an agency of the United States Department of Agriculture, conducts monthly field surveys in the late summer and fall to forecast corn and soybean yields. Data from the survey forms the input for a statistical model that predicts the current season final yield. The model NASS uses is more complicated than a simple regression, and they have found it convenient to implement this model with SAS/IML software. This paper gives a short description of the survey design and variables, and then presents the statistical model.

Description of the Survey

In June NASS conducts a very large survey of agricultural land use in the U.S. to estimate the current season's acreage planted to corn and soybeans. From the base generated by this survey NASS draws a random sample of corn and soybean plots. This is done through a two stage process, in which fields are selected and then random locations are designated within each selected field. The procedure is carried out in such way that a simple random sample is obtained, meaning that each planted area of corn or soybeans has an equal chance of being included in the sample. This simple random sample property is an important assumption for the statistical models that are applied to the survey data.

The randomly located plots are a few square feet in area. Within the plots enumerators count and measure variables that are positively correlated with final yield. Among the variables collected for soybeans are number of plants, number of nodes per plant, number of lateral branches per plant, number of blooms, dried flowers and pods per plant, and number of pods with beans per plant. For corn the NASS enumerators count the number of stalks, number of stalks with ears, number of ear shoots, and number of ears with kernels. They also husk a random sample of ears near the plot and measure the length of a typical kernel row on each ear. Just prior to farmer harvest of the corn or soybean field in which the sample is located, the enumerator harvests the plot and records the final yield.

Samples are laid out in all the major corn and soybean producing states. Starting in August and continuing through November, around the 10th of each month the USDA releases yield estimates for each state based on the survey.

Variables in the Regional Models

It was found that the best relationship between the survey data and final yield is found at the regional level, the region being the combined states in the survey. Consequently the plot level data is summarized to the state and then to the region level, where it is modeled against the region yield. Each monthly regional model normally has one independent variable X.

The form of the regional linear model is

$$Y = \alpha + \beta X + \epsilon$$

where

Y = regional yield and

α, β are the unknown model parameters.

X is the known independent variable

ϵ is the difference between Y and its expected value

The independent variable X in each model varies by month. For soybeans X is the following.

SOYBEAN VARIABLES

MONTH :	Independent variable
August	estimated number of lateral branches per square foot
September	estimated number of pods with beans per square foot
October-December	(estimated number of pods per square foot) X (estimated net weight per pod)

For corn the independent variables are the following.

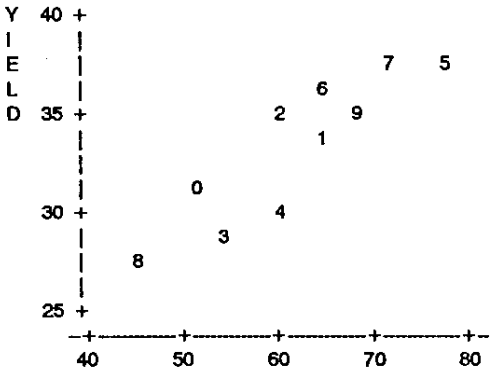
CORN VARIABLES

MONTH :	Independent variable
August	(stalks with ears per square foot) X (average kernel row length per ear)
September	(ears with kernels per square foot) X (average kernel row length per ear)
October-December	(ears with kernels per square foot) X (average grain weight per ear)

Maturity Adjustment

While NASS conducts the survey during the last ten days of each month, the overall maturity of the crop at that time will vary from year to year, depending on when it was planted, subsequent weather, etc. The forecasting power of the model is enhanced by classifying each plot by stage of maturity and basing the independent variable calculations on data from samples in pre-selected stages. This adjustment allows the independent variables to be more comparable across years. Variables not used directly in **X** (such as nodes and blooms, dried flowers and pods) are used for maturity classification. Consequently the predictor variable is not a function of all the data, but only those plots that are at a stage that has exhibited good predictive power for final yield. This criteria normally means the exclusion of very immature samples in the first month of the survey. After that the vast majority of the samples are used directly in X.

A plot of the data in one of the regional models is given below (the soybean model for August 1). (The digits plotted represent the years 1980-1989).



lateral branches per 18 square feet

LINEAR MODEL FRAMEWORK FOR THE STATE MODELS

State Models

As mentioned the USDA issues state estimates each month. To convert the best forecasting model, the regional model, into state estimates, a global model of all the state models is created. The parameters of this model are restricted so that the state forecasts weight to the regional forecast in the current year. The variables in the state models are the state level analogues to the regional variables (although in this general development states can have models with variables that are different from each other and the regional model). This general linear model is implemented with SAS/IML software.

The Restricted State Linear Models

Definitions

The individual state models have the following matrix notation.

Let

$\mathbf{X}_i$ = the data matrix for the linear model for state i, i=1,...,q, the states in the survey.

$\mathbf{X}_i$ will generally be $N_i \times 2$, where N_i is the number of years in the data base. In most cases $\mathbf{X}_i$ will have a column of 1's and one column with independent variable values, x_i.

$$\mathbf{X}_i = [\underline{1}\ \underline{x}_i]$$

This also indicates that the state parameter vector $\underline{\beta}_i$ is generally 2×1.

For the dependent variable let

$\underline{Y}_i$ = the dependent vector of net yield for state i.

Analogously, for the regional model let

$\mathbf{X}$ = the data matrix for the regional model, and

$\underline{Y}$ = the dependent vector of final net yield for the region.

The rows of the X's and $\underline{Y}$'s represent each of the years in the data base.

The Global Model

To create the global state model vertically concatenate the $\underline{Y}_i$'s and create a block diagonal matrix of the $\mathbf{X}_i$'s to form

$$\begin{bmatrix} \underline{Y}_1 \\ \cdot \\ \cdot \\ \cdot \\ \underline{Y}_q \end{bmatrix} = \begin{bmatrix} \mathbf{X}_1 & & & & \\ & \mathbf{X}_2 & & & \\ & & \cdot & & \\ & & & \cdot & \\ & & & & \mathbf{X}_q \end{bmatrix} \begin{bmatrix} \underline{\beta}_1 \\ \cdot \\ \cdot \\ \cdot \\ \underline{\beta}_q \end{bmatrix} + \begin{bmatrix} \underline{\varepsilon}_1 \\ \cdot \\ \cdot \\ \cdot \\ \underline{\varepsilon}_q \end{bmatrix}$$

$$\underline{Y}_g = \mathbf{X}_g \underline{\beta}_g + \underline{\varepsilon}_g$$

To make inferences we will assume

$$\underline{\varepsilon}_s \sim N(\underline{0}, \sigma^2 \underline{I})$$

Predicting the Current Year

In the current year the values of the independent variables are observed and the unknown value of the future Y_{fi} is predicted. For this terminology let

> $\underline{x}_{fi}'$ = the observed $\underline{x}'$ for the current year for state i, before Y_{fi} is known (the f stands for future). Normally $\underline{x}_{fi}'$ = [1 x_{fi}], where x_{fi} is the value of the independent variable from the just completed survey for state i.

Another variable needed as part of the restriction is a_i, the current acres for harvest for each state i.

Let

$$A = \sum_I a_i$$

the sum of the state acres, which is the current regional acreage.

The Current Year Restriction on the Parameters

The constraint that the state estimates weight to the regional estimate will take the form of a linear restriction on $\underline{\beta}_s$, specifically $\underline{k}'\underline{\beta}_s = m$.

In particular,

$$\underline{k}' = \frac{\left[a_1 \underline{x}'_{f1} \quad a_2 \underline{x}'_{f2} \quad \ldots\ldots \quad a_q \underline{x}'_{fq}\right]}{A}$$

Parameter Estimation For the Restricted Model

The regional model is

$$\underline{Y} = \boldsymbol{X}\underline{\beta} + \underline{\varepsilon}$$

We estimate $\underline{\beta}$ from the regional model with OLS with

$$\underline{b} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\underline{Y}$$

Then the current year predicted value m for the region is

$$m = \underline{x}'_f \, \underline{b}$$

where $\underline{x}_f'$ is the regional $\underline{x}'$ from the current survey.

This is the m that completes the specification of the linear restriction $\underline{k}'\underline{\beta}_s = m$.

For the state models the unrestricted OLS estimate of $\underline{\beta}_s$ is

$$\underline{b}_s = (\boldsymbol{X}'_s \boldsymbol{X}_s)^{-1}\boldsymbol{X}'_s \underline{Y}_s$$

The estimate of $\underline{\beta}_s$ subject to $\underline{k}'\underline{\beta}_s = m$ is

$$\underline{b}_{rs} = \underline{b}_s - (\boldsymbol{X}'_s \boldsymbol{X}_s)^{-1} \underline{k} \left(\underline{k}' (\boldsymbol{X}'_s \boldsymbol{X}_s)^{-1} \underline{k}\right)^{-1} (\underline{k}' \underline{b}_s - m)$$

Estimation of σ^2

σ^2 is estimated under the restricted model as

$$\hat{\sigma}^2 = \frac{\underline{r}'_{rs} \underline{r}_{rs}}{(N_s - p - 1)}$$

where

$$\underline{r}_{rs} = \underline{Y}_s - \boldsymbol{X}_s \underline{b}_{rs}$$

the vector of residuals under the restricted model.

In addition,

N_s = the number of rows in $\boldsymbol{X}_s$ and

p = the number of columns in $\boldsymbol{X}_s$.

$\underline{X}_s$ is always constructed to have full column rank, so that all parameters are identifiable. The additional 1 is subtracted from the denominator because of the 1 degree of freedom restriction on the parameter estimates.

The Constrained Predicted Values and Variances

The constrained predicted values for the future Y_{fi}'s then become

$$\underline{\hat{Y}}_{fs} = \begin{bmatrix} \underline{x}'_{f1} \\ \cdot \\ \cdot \\ \underline{x}'_{fq} \end{bmatrix} \underline{b}_{rs}$$

The estimated variances of the state predicted values can be found from the diagonal elements of

$$\left\{ \begin{bmatrix} \underline{x}'_{f1} \\ \cdot \\ \cdot \\ \underline{x}'_{fq} \end{bmatrix} \left(\underline{x}'_s \underline{x}_s\right)^{-1} \begin{bmatrix} \underline{x}_{f1} & \cdot & \cdot & \underline{x}_{fq} \end{bmatrix} + I \right\} \hat{\sigma}^2$$

The preceding matrix equations have been translated into SAS/IML software to produce the forecasts and their standard errors.

References

Birkett, T.R. (1990), "The New Objective Yield Models for Corn and Soybeans", National Agricultural Statistics Service, SMB-90-02, Washington, DC, 20250.

Searle, S.R. (1971), Linear Models, New York: John Wiley & Sons, Inc.

The author can be contacted at

NASS/USDA, Room 5819,
14th and Independence, SW,
Washington,DC, 20250

202-447-5359
