United States
Department of
Agriculture

National
Agricultural
Statistics
Service

Estimates
Division

SMB Staff Report
Number SMB-90-02

August 1990

The New Objective Yield Models for Corn and Soybeans

Thomas R. Birkett

The New Objective Yield Models for Corn and Soybeans. Thomas R. Birkett, Estimates Division, National Agricultural Statistics Service, U. S. Department of Agriculture, Washington D. C. 20250, August 1990, Staff Report No. SMB-90-02.

Abstract

New objective yield regression models for soybeans and corn were adopted in January, 1990 for implementation in August 1990. The new models have as input the same data as the old models, but they use this data more efficiently to make a many fold increase in accuracy over the old models. The new models are state level models, rather than plot level models. In addition, a regional model sets a regional estimate, and then the state models are constrained to produce forecasts that are consistent with the regional model. This paper presents the linear model framework for this model and defines the soybean and corn objective yield variables to which it is applied.

I. CORN AND SOYBEAN OBJECTIVE YIELD VARIABLES

a. General

In 1990 new operational models will be introduced for forecasting soybean and corn net yield and will become the official operational models in 1991. The older operational models will be phased out. The new models are many times more accurate than the old models and will greatly increase the value of the objective yield survey to each State's estimating program. The new models are also less complicated and consequently easier to analyze and present. They were built using ideas from multiple sources, combined in a classical linear model framework.

b. Variables in the Regional Models

For the region made up of all the states in the corn or soybean objective yield surveys, a model for the regional yield is constructed for each month. In general each monthly model has one independent variable X. The form is

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

where

Y = regional ASB yield

α, β are the unknown model parameters, to be estimated from historical data.

The independent variable X in each model varies by month. For soybeans X is a function of the following objective yield variables.

SOYBEAN VARIABLES

MONTH :	Independent variable
August	estimated number of lateral branches per 18 square foot
September	estimated number of pods with beans per 18 square foot
October-December	estimated net yield from current objective yield harvested samples and estimated number of pods with beans per 18 square foot from unharvested samples

For corn the independent variable is a function of the following variables.

CORN VARIABLES

MONTH :	Independent variable
August	stalks with ears and ears with kernels per square foot, average kernel row length per ear
September	ears with kernels per square foot, average kernel row length per ear
October-December	estimated net yield from current objective yield harvested samples, and ears with kernels per square foot and average kernel row length per ear from unharvested samples

Details on the estimation of net yield from harvested samples and the definition of forecasting categories for unharvested samples (see below) is presented in the S & E sections 8.4 and 8.6.

Soybeans

For soybeans only samples from maturity categories 2-6 (1-6 in the southern states) are used in the construction of X in August. In September only samples falling in categories 6-9 are used. From October on only category 6-10 samples are used. For the period 1980-1989 the following table shows the percentage of status 1 samples that were included in the models.

% SAMPLES USED IN MODELS (1980-1989)

	August	September
Arkansas		49
Illinois	88	96
Indiana	79	94
Iowa	93	99
Kansas		68
Louisiana		71
Minnesota	88	98
Mississippi		51
Missouri	77	75
Nebraska		98
Ohio	75	93
Region	85	88

From October on virtually all status 1 (sampled) and status 4 (harvested) samples are included.

Algebraic definitions for X

August

In August the algebraic formula for estimated number of lateral branches per 18 square feet is

$$lat_i = \frac{1}{n_i} \sum_{J_i} pl_{ij} lat_{ij}$$

where

J_i *is the subset of samples classified in maturity categories* 2-6 *(or* 1-6 *in the southern states), state* i.

n_i = *the number of samples* $\in J_i$

pl_{ij} = *plants per* 18 *square feet for sample* j, *state* i

lat_{ij} = *lateral branches per plant, sample* j, *state* i

lat_i = *state* i *estimate lateral branches per* 18 *square feet.*

The state level estimates are combined to the regional level with current ASB acres for harvest as the weight.

$$X = \frac{\sum_I a_i \, lat_i}{\sum_I a_i}$$

where a_i *is the ASB acres for harvest for state* i *and* I *is the set of states in the survey.*

September

The formula for estimated number of pods per 18 square feet is

$$pod_i = \frac{1}{n_i} \sum_{J_i} pl_{ij} pod_{ij}$$

where

J_i *is the subset of samples classified in maturity categories* 6-9 *, state* i.

n_i = *the number of samples* $\in J_i$

pl_{ij} = *plants per* 18 *square feet for sample* j, *state* i

pod_{ij} = *pods with beans per plant, sample* j, *state* i

pod_i = *state* i *estimate pods with beans per* 18 *square feet.*

The regional level estimate is

$$X = \frac{\sum_I a_i \, pod_i}{\sum_I a_i}$$

where a_i *is the ASB acres for harvest for state* i *and* I *is the set of states in the survey.*

October-December

After September X is the average of the yields from the harvested samples and the predicted yields from the unharvested samples. At this time the predicted yields are those from the current operational models. The models in categories 6-10 have pods with beans as the independent variable.

$$X_i = \frac{1}{n_{ih}+n_{ik}} \left(\sum_{J_{ih}} Y_{ij} + \sum_{J_{ik}} \hat{Y}_{ij} \right)$$

where

J_{ih} *is the subset of samples classified in maturity category* 10 *(harvested), state i.*

J_{ik} *is the subset of samples classified in maturity* 6-9 *(unharvested), state i*

n_{ih} = *the number of samples* $\in J_{ih}$

n_{ik} = *the number of samples* $\in J_{ik}$

Y_{ij} = *harvested yield for sample j, state i*

$\hat{Y}_{ij}$ = *forecasted yield, sample j, state i.*

The regional estimate is

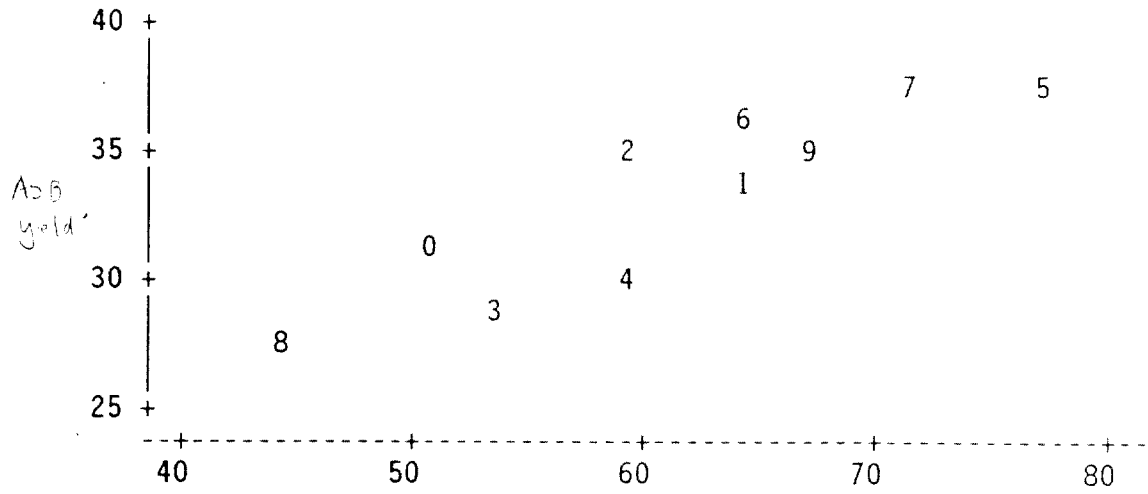
$$X = \frac{\sum_I a_i X_i}{\sum_I a_i}$$

where a_i *is the ASB acres for harvest for state i and I is the set of states in the survey.*

A plot of the ASB yields versus the values of X for 1980-1989, August-December, follows. (Note: there has been some movement of states in and out of the program during the 10-year period depicted in the plot. The X and Y variables represent whichever states were in the program in respective years).

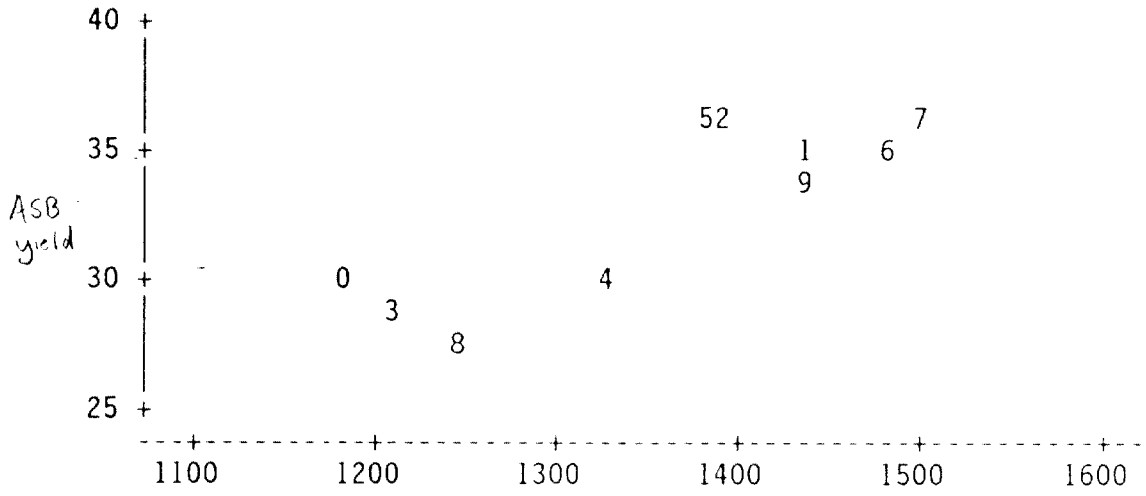
SOYBEANS

AUGUST



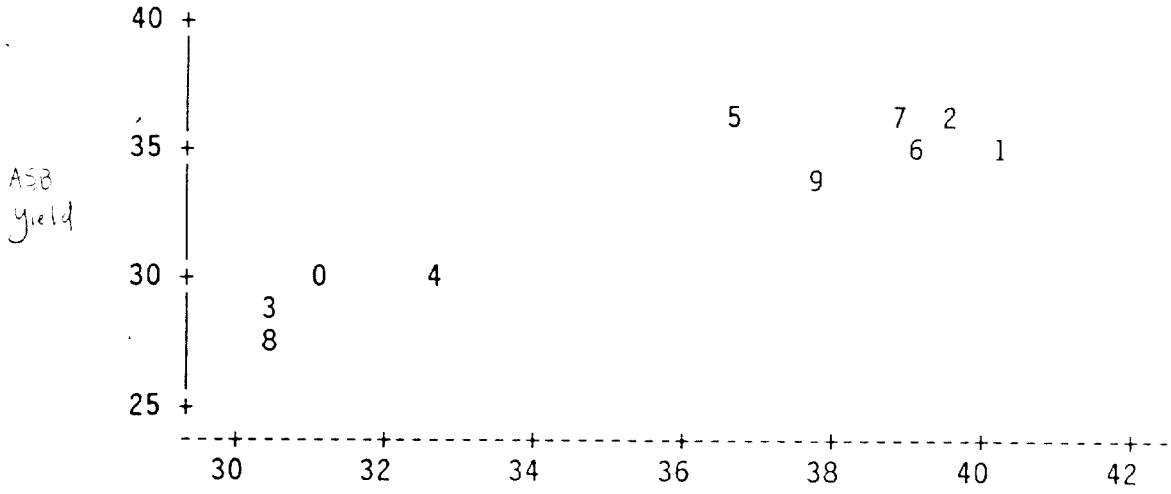
R^2 = .80 lateral branches per 18 square feet

SEPTEMBER



R^2 = .75 Pods with beans per 18 square feet

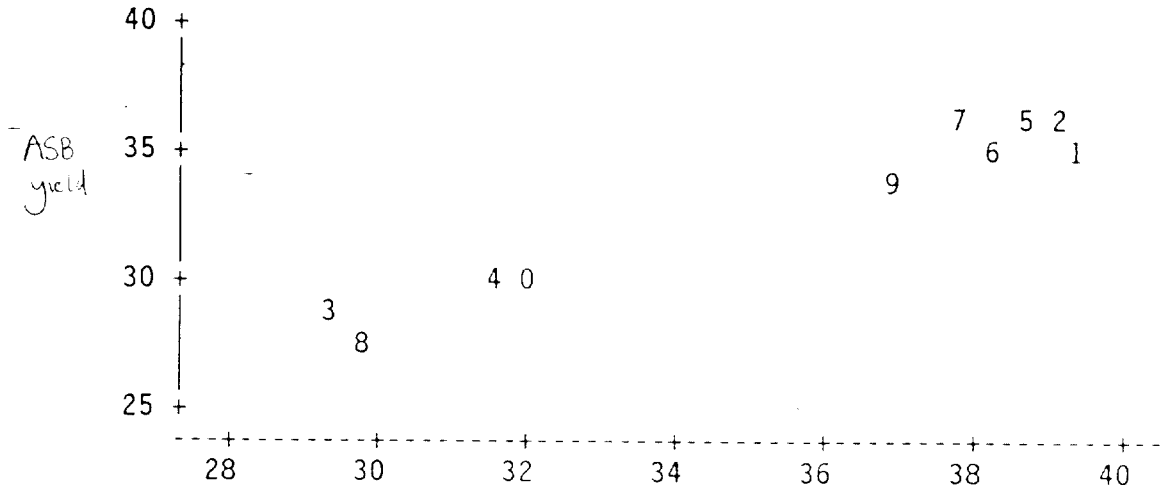
OCTOBER



R^2 = .98

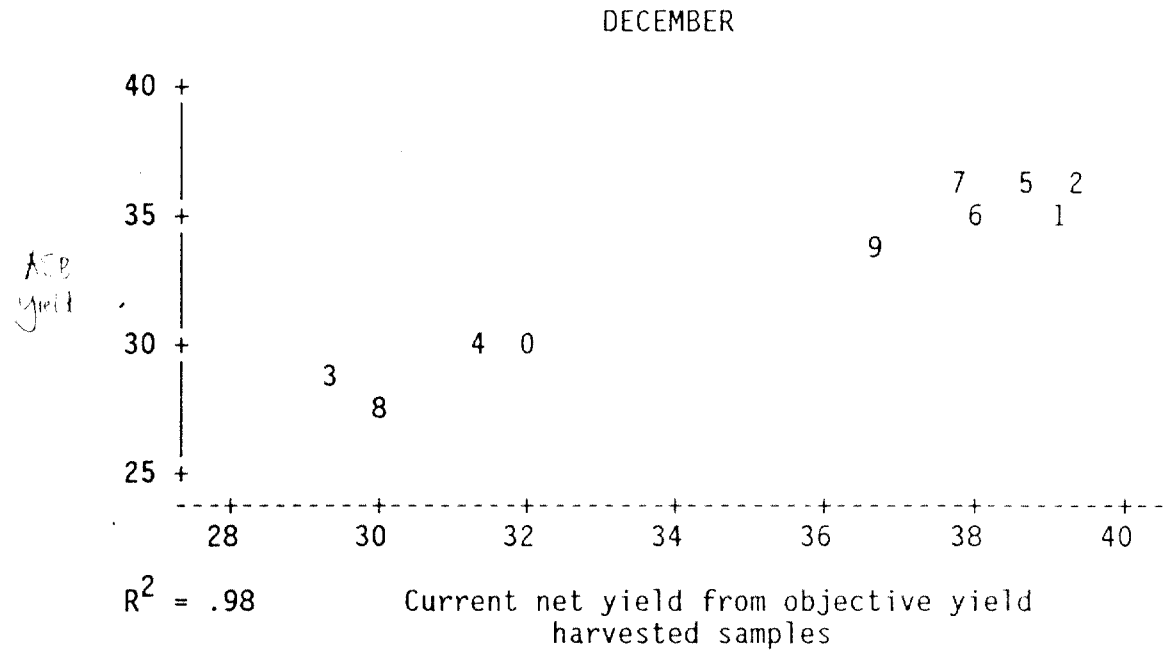
f(current net yield from objective yield harvested samples, pods with beans from unharvested samples)

NOVEMBER



R^2 = .98

Current net yield from objective yield harvested samples



Corn

For corn only samples from maturity categories 3-7 are used in the construction of X. In August and September samples falling in 3-6 are used, and from October on only 3-7's are used. The following table provides percentages of the status 1's that were included in the models for the (1980-1989) period.

% SAMPLES USED IN MODELS (1980-1989)

	August	September
Illinois	28	99
Indiana	21	97
Iowa	13	99
Michigan	2	94
Minnesota	8	97
Missouri	52	98
Nebraska	12	98
Ohio	12	94
South Dakota	4	95
Wisconsin	4	94
Region	19	97

From October on virtually all status 1 (sampled) and status 4 (harvested) samples are included in the models. Because of the small percentages in Michigan, Minnesota, South Dakota and Wisconsin, samples from those states are automatically excluded from the August variable regardless of maturity.

August

In August the algebraic formula for X is algebraically more involved than it is for soybeans, because X is a function of two count variables and a size variable.

$$X = \left\{ \sum_I \frac{a_i}{\sum_I a_i} \left[\frac{1}{n_i} \sum_{J_i} \frac{(se_{ij} + ek_{ij})}{\frac{15}{4} r_{ij}} \right] \right\} \chi$$
$$\left\{ \sum_I \frac{a_i}{\sum_I a_i} \left[\frac{1}{n_i} \sum_{J_i} \frac{(se_{ij}+ek_{ij})\overline{krl}_{ij}}{\frac{15}{4} r_{ij}(se_{ij} + ek_{ij})} \right] \right\}$$

where

a_i = *current ASB acres for harvest, state i*

I = *the set of states in the survey*

J_i = *the set of samples in maturity 3-6, state i*

n_i = *the number of samples in* J_i

se_{ij} = *the number of stalks with ears, sample j, state i*

ek_{ij} = *the number of ears with kernels , sample j, state i*

r_{ij} = *four row space measurement, sample j, state i*

$\overline{krl}_{ij}$ = *average kernel row length, sample j, state i.*

- **September**

In September it is

$$X = \left\{ \sum_I \frac{a_i}{\sum_I a_i} \left[\frac{1}{n_i} \sum_{J_i} \frac{(ek_{ij})}{\frac{15}{4} r_{ij}} \right] \right\} \chi$$
$$\left\{ \sum_I \frac{a_i}{\sum_I a_i} \left[\frac{1}{n_i} \sum_{J_i} \frac{(ek_{ij})\overline{krl}_{ij}}{\frac{15}{4} r_{ij}(ek_{ij})} \right] \right\}$$

where

a_i = *current ASB acres for harvest, state i*

I = *the set of states in the survey*

J_i = *the set of samples in maturity 3-6, state i*

n_i = *the number of samples in* J_i

ek_{ij} = *the number of ears with kernels , sample j, state i*

r_{ij} = *four row space measurement, sample j, state i*

$\overline{krl}_{ij}$ = *average kernel row length, sample j, state i.*

October-December

After September X is the average of the yields from the harvested samples and the predicted yields from the unharvested samples. At this time the predicted yields are those from the current operational models. The models in categories 3-6 have ears with kernels and average kernel row length as the independent variables.

$$X = \sum_I \frac{a_i}{\sum_I a_i} \left[\frac{1}{n_{ih} + n_{ik}} \left(\sum_{J_{ih}} Y_{ij} + \sum_{J_{ik}} \hat{Y}_{ik} \right) \right]$$

where

a_i = *ASB acres for harvest, state i,*

J_{ih} *is the subset of samples classified in maturity category 7 (harvested), state i.*

J_{ik} *is the subset of samples classified in maturity 3-6 (unharvested), state i*

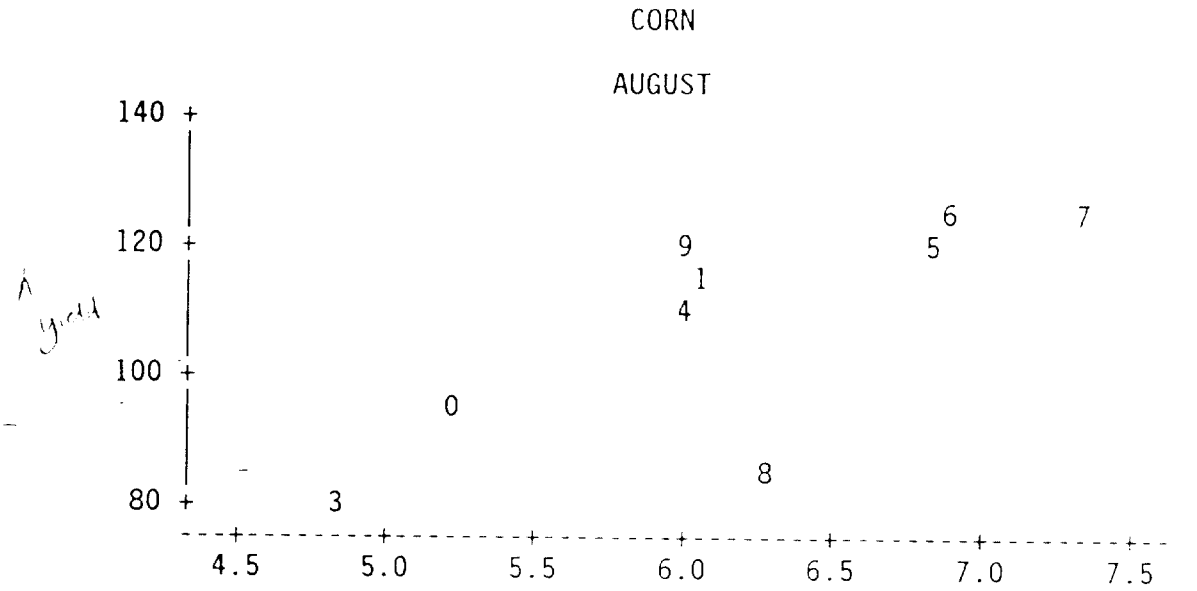
n_{ih} = *the number of samples* $\in J_{ih}$

n_{ik} = *the number of samples* $\in J_{ik}$

Y_{ij} = *harvested yield for sample j, state i*

$\hat{Y}_{ij}$ = *forecasted yield, sample j, state i.*

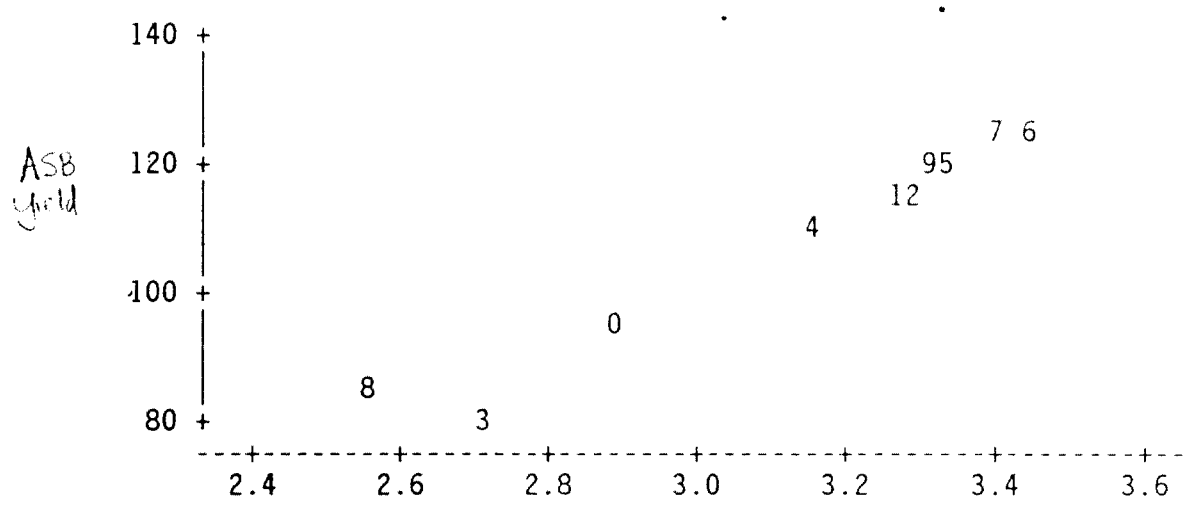
Plots of Y vs X for corn follow.



R^2 = .98 f(stalks with ears, ears with kernels, kernel row length)

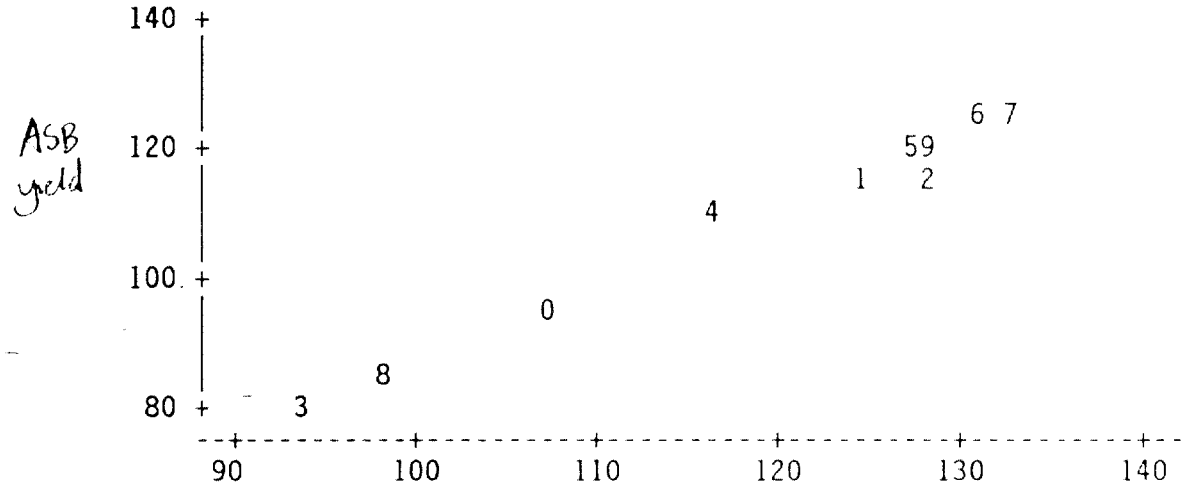
NOTE: 1 obs hidden, and 1988 is an outlier not included in the August model. This model has as independent variables X and X^2.

SEPTEMBER



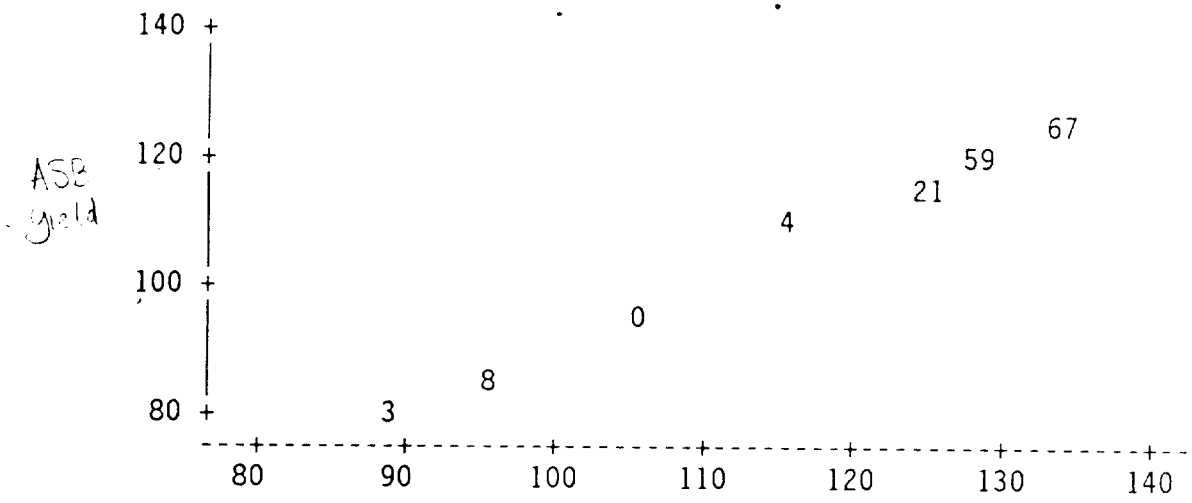
R^2 = .97 f(ears with kernels, kernel row length)

OCTOBER



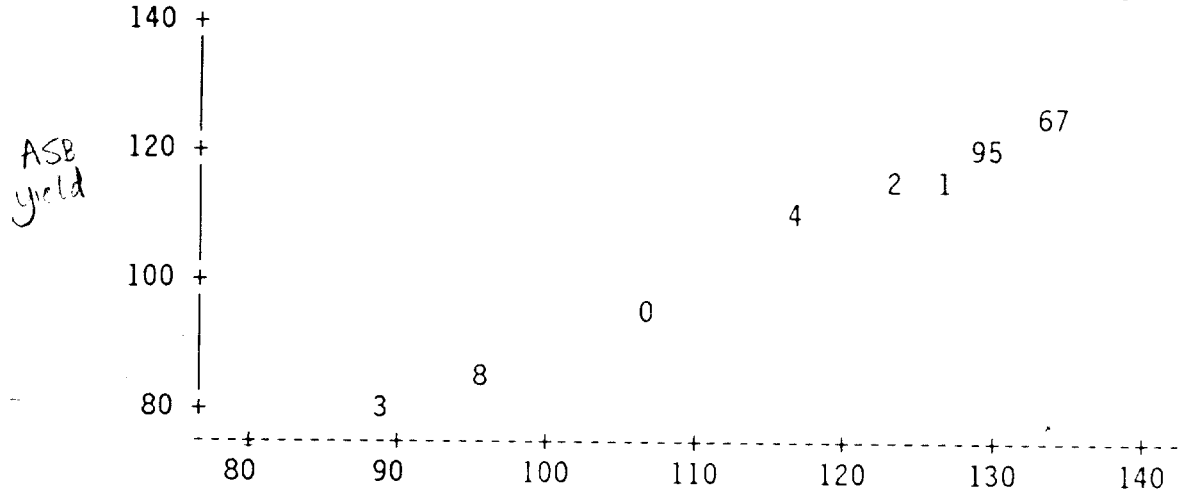
R^2 = .98 f(current net yield from objective yield harvested samples, ears with kernels and avg kernel row length from unharvested samples)

NOVEMBER



R^2 = .98 Current net yield from objective yield harvested samples

DECEMBER



R^2 = .98 Current net yield from objective yield harvested samples

II. LINEAR MODEL FRAMEWORK FOR THE NEW MODELS

a. State Models

Once the regional estimate is made, state forecasts are developed that are constrained to weight (by ASB acres) to the forecast of the regional model. To accomplish this a global model that predicts all states simultaneously is developed. The parameters of this model are estimated subject to the linear restriction that the state forecasts weight to the regional forecast.

The theory behind the state models does not require that the state models use the same independent variables as the regional models, only that they weight to the regional forecast. The key is to use independent variables that do the best job of making good state forecasts. It turns out that in August the objective yield variables are too weak at the state level to generate reliable estimates, although there is enough information to generate regional estimates. August is the only month where the state independent variables are not the same as the regional independent variables. For both corn and soybeans the August state independent variable is a function of crop condition percent as of the last week in July, obtained through the crop weather. X_i in August is

$$X_i = \frac{(vp) + 2(p) + 3(f) + 4(g) + 5(ex)}{15}$$

where

vp = *percent very poor*
p = *percent poor*
f = *percent fair*
g = *percent good*
ex = *percent excellent*

as of the last week in July.

Of course each crop has a separate set of conditions.

b. Details of the Constrained State Linear Models

Definitions

The individual state models can be written as a global state model using matrix notation. First we will define the various matrices that go into the model for a given month. For a specific example we will use the soybean objective yield

survey.

Let

$\mathbf{X}_i$ = the data matrix for the linear model for state i, i=5,17,18,19,20,22,27,28,29,31,39, the states in the survey.

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

$$\boldsymbol{X}_i = [1 \; \boldsymbol{x}_i]$$

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

Recall that for soybeans $\underline{x}_i$ is crop condition percent (as of the last week in July) for August, pods with beans in September, and a combination of current net yield from harvested samples and pods with beans from unharvested samples for October-December.

In September 1990 $\mathbf{X}_{20}$ will be 1 x 1 and will contain the 1980 observed value for pods with beans per 18 square foot. This will force $\underline{\beta}_{20}$ to be 1 x 1 also. (1980 is the only history available for Kansas).

For corn the state variables ($\underline{x}_i$'s) are crop condition percent (as of the last week in July) for August, a function of ears with kernels and kernel row length in September, and a combination of current net yield from harvested samples, and ears with kernels and average

kernel row length for unharvested samples, for October-December.

For both crops the crop condition variable goes back to 1985, while the objective yield variables extend back to 1980.

For the dependent variable let

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

The rows of the $\mathbf{X}$'s and $\underline{Y}$'s represent each of the years 80-89 (although some states are missing some years, corresponding to years when the survey was not done in those states).

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

$$\begin{bmatrix} \underline{Y}_5 \\ \cdot \\ \cdot \\ \cdot \\ \underline{Y}_{39} \end{bmatrix} = \begin{bmatrix} \boldsymbol{X}_5 & & & & \\ & \boldsymbol{X}_{17} & & & \\ & & \cdot & & \\ & & & \cdot & \\ & & & & \boldsymbol{X}_{39} \end{bmatrix} \begin{bmatrix} \underline{\beta}_5 \\ \cdot \\ \cdot \\ \cdot \\ \underline{\beta}_{39} \end{bmatrix} + \begin{bmatrix} \underline{\varepsilon}_5 \\ \cdot \\ \cdot \\ \cdot \\ \underline{\varepsilon}_{39} \end{bmatrix}$$

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

To make inferences we will assume

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

Definitions for Predicting the Current Year

In the current year we will observe the values of the independent variables and predict the unknown value of the future Y_{fi}. To this end 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, for example, in September x_{fi} is the value of pods with beans per 18 square feet for the just completed survey for state i.

Another variable we will need is a_i, the current ASB 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 Constraint 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,

$$k' = \frac{\left[a_5 x'_{f5} \quad a_{17} x'_{f17} \quad \cdots\cdots \quad a_{39} x'_{f39}\right]}{A}$$

Parameter Estimation Under 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

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

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

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

Estimation of σ^2 under the Restricted Model

σ^2 is estimated under the restricted model as

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

where $\underline{r}_{rs}$ is the vector of residuals under the restricted model.

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

In addition,

N_s = the number of rows in X_s and

p = the number of columns in 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

$$\hat{Y}_{fs} = \begin{bmatrix} X'_{f5} \\ \cdot \\ \cdot \\ X'_{f39} \end{bmatrix} b_{rs}$$

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

$$\left\{ \begin{bmatrix} X'_{f5} \\ \cdot \\ \cdot \\ X'_{f39} \end{bmatrix} (X'_S X_S)^{-1} \begin{bmatrix} X_{f5} \\ \cdot \\ \cdot \\ X_{f39} \end{bmatrix} + I \right\} \hat{\sigma}^2$$

c. Forecasting Accuracy

To assess the forecasting accuracy of the new models compared to the old models, the estimated standard error for the 1989 forecast of each is listed in the following table (in bushels per acre). (For the standard error of the old models see Birkett (1990)).

REGIONAL MODELS

MONTH	CORN		SOYBEANS	
	Old	New	Old	New
Aug	16.0	2.9	5.1	1.9
Sept	12.0	3.1	3.8	2.2
Oct	9.2	2.7	3.1	1.2
Nov	9.2	2.6	2.6	0.6
Dec	9.6	2.5	2.6	0.6

While the standard error of September is slightly larger than that of August for both crops, this is considered to be a statistical anomaly that will be reversed with time. Also the August and September soybean models represent different regions.

For the state forecasts the average standard errors under the restricted model for predicting 1989 are listed in the following table.

STATE MODELS

MONTH	CORN	SOYBEANS
Aug	6	3.3
Sept	6	3.5
Oct	5	2.7
Nov	4.5	2
Dec	4.3	1.7

References

Birkett, T.R. (1987), "A Production Forecasting Model for Corn", National Agricultural Statistics Service, SRB-87-02, Washington, DC

Birkett, T.R. (1990), "The General Linear Model Framework for the Objective Yield Models with Suggested Improvements", National Agricultural Statistics Service, SMB-90-01, Washington, DC

Objective Yield Survey S & E Manual, (1990), National Agricultural Statistics Service, Washington, DC, Sections 8.2 and 8.6

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