United States
Department of
Agriculture

Statistical
Reporting
Service

Statistical
Research
Division

SRS Staff Report
Number YRB-09

November 1985

Covariance Analysis of Soybean Objective Yield Maturity Categories 7, 8 & 9

Robert Battaglia

COVARIANCE ANALYSIS OF SOYBEAN OBJECTIVE YIELD MATURITY CATEGORIES 7, 8 & 9 By Robert J. Battaglia, Statistical Research Division, Statistical Reporting Service, U.S. Department of Agriculture, Washington, D.C. 20250. November 1985. Staff Report No. YRB-85-09.

ABSTRACT

Covariance analysis techniques were used to examine the effects of maturity category and year on forecast models in the soybean objective yield program. Models were constructed using October data from 1977 to 1983. Results showed that forecast models from maturity categories 7, 8 & 9 could not be operationally combined into one October forecast model without some loss in model fit. The year effect, which was significant in some states, shows that an unusual year can affect the slope and intercept of a forecast model.

KEYWORDS

Covariance analysis, maturity categories, forecast models, Soybean Objective Yield.

This paper was prepared for limited distribution to the research community outside the U.S. Department of Agriculture. The views expressed herein are not necessarily those of SRS or USDA.

ACKNOWLEDGMENT

The author thanks Ron Steele for his assistance with the analysis, and for the helpful comments from Tom Birkett, Ned Jones and Ron Fecso.

Washington, D.C. November 1985

CONTENTS PAGE

SUMMARY

The purpose of this research was to determine if soybean pods per plant forecast models for maturity categories 7, 8 & 9 could be combined since the three regression models use identical variables. A second objective was to measure the effect of year on the forecast models. In the operational program these models are constructed using five previous years of data which are implicitly assumed to be like the current year. Covariance models were constructed using October soybean objective yield data from 1977 to 1983. Results indicated that forecast models for the three October maturity categories could not be combined into one model without loss in model fit. This was the result of a maturity category effect which caused significant differences in slopes and/or intercepts. The analysis of year effect showed that some years contribute a significant effect to the slopes and intercepts of the forecast models in some states. Significant year effects were more prevalent in the southern soybean objective yield states.

COVARIANCE ANALYSIS OF SOYBEAN OBJECTIVE YIELD MATURITY CATEGORIES 7, 8 & 9

by Robert Battaglia [1]

INTRODUCTION

The soybean objective yield survey uses regression models to forecast number of pods with beans per plant. Models are created by state for each maturity category within a month using five previous years of data. Since all soybean fields in a state will not be at the same stage of development, the maturity categories are used to divide the sample units into homogeneous groups. Forecast models utilizing this grouped data should be more efficient. There are 10 maturity categories used in the soybean objective yield program. These categories are listed in Appendix 1.

Previous research on soybean maturity categories dealt primarily with reducing the number of plant-component counts made by field enumerators [1]. The efficiency of the categories was difficult to assess statistically because the variables used in the forecast models could change from year to year. This problem was due to the stepwise selection procedure used to determine the soybean forecast models. Research on the models which forecast number of pods with beans per plant led to the replacement of stepwise-created models with fixed-variable models in 1985 [2,3]. For each maturity category within a month, the variables used in the forecast models are the same for all states. The models used to forecast number of pods with beans per plant are listed in Appendix 2.

[1] The author is a mathematical statistician with the Statistical Reporting Service, U.S. Department of Agriculture, Washington, D.C.
[2] Numbers in brackets refer to literature cited in the references at the end of the report.

In some months, forecast models from adjacent maturity categories use identical variables. The purpose of this study is to determine statistically whether data from those maturity categories can be combined. This would reduce the number of maturity categories and increase the number of observations in the combined categories. Another objective will be to investigate the assumption that year is homogeneous in terms of the number of pods with beans per plant. Using the data from the last five years to build forecast models for the current year implies that the previous five years were no different than the current year. A test of equal year effect should give insight on the effect of unusual weather, changes in procedure, etc. on yield models.

METHODS

Soybean objective yield data from 1977 to 1983 from the 15 states was used for the analysis. A complete description of soybean objective yield methods can be found in the supervising and editing manual [6]. Outliers and leverage points were removed from the regression models using Studentized T and Cook's D statistics. Residual plots of the forecast models were also examined. The residual plots were non normally distributed and slightly negatively skewed for all states. These results would make the alpha levels for any hypothesis tests approximate but still useable since the distributions are nearly normal.

I analyzed October data since the majority of sample units are in maturity categories 7, 8 & 9 during this month. Since these maturity categories occur prior to harvest, most of the pods have filled and the beans are in the process of maturing. The number of pods with beans per plant counted in October is used to forecast final number of pods with beans for categories 7, 8 & 9. Also, the relationship between final and October pod numbers has a very good linear fit for these categories. Research on fixed-variable forecast models has shown that the model coefficients for the above October maturity categories were extremely stable over time [2,3]. Since the plants are approaching maturity and the three maturity categories use the same independent variable to forecast final number of pods with beans, it is logical to explore combining these maturity categories. An analysis of covariance was used to determine whether the maturity categories can be combined.

The covariance model can be thought of as a combination of regression and analysis of variance methods. It can be used where there is a quantitative dependent variable (final pods with beans), a quantitative independent variable (October pods with beans) and a qualitative treatment (maturity category). Covariance analysis is effective in reducing experimental errors and investigating treatment effects [5]. The covariance models assume the following properties: error terms are independent and have constant variance, the treatments have the same slope; therefore, the expected difference between treatments is the same for all values of x, the observations of October pod numbers are considered constants, and the covariate and treatment are statistically independent [4]. The last assumption may be violated since the covariate (pods with beans/plants) and the treatment (pods with beans/total fruit) share a common numerator. The consequence of this possible assumption violation makes it difficult to interpret differences in means as the result of covariate or treatment effect. However, the goal of this research is to identify differences in forecast models not the interpretation of differences.

Two covariance models were used to determine whether maturity categories 7, 8 & 9 can be combined in October. These models are described in Appendix 3. The fit of October data will be compared using sums of squared errors (SSE) from covariance models with separate and combined maturity categories 7, 8 & 9. The ratio of SSE's, defined as the relative efficiency (RE) will be computed to show the change in model fit resulting from combining maturity categories.

A third covariance model, also described in Appendix 3, was used to test the effect of year on the models. Current objective yield procedures assume that the previous five years data, from which the forecast models are constructed, are not significantly different than the present year. The covariance analysis will allow us to identify years that are significantly different with respect to the the dependent variable, the final number of pods with beans per plant.

RESULTS

Maturity Category Effect

The results of the covariance analysis are presented in this section. One of the assumptions of covariance analysis is that the regression lines for each treatment have the same slope. A test of this assumption was the first step in the analysis. Model (1) from Appendix 3 was used for this test. Results from an F test on the slopes of the forecast models for the 3 maturities showed that 4 of the 15 states had slopes that were not significantly different at alpha = .10. When the intercepts for those 4 states were tested only North Carolina had intercepts that were not significantly different. However, the F test does not give an indication which categories were different. Contrast statements were used to test for significant differences in slope between maturity category pairs. An alpha level of .01 was used to determine if a pair was different. This level of alpha is an approximation since the true significance levels of the contrast statements are not known. If the regression coefficients from a maturity category pair were significantly different, then those two categories could not be combined. Regression coefficient is defined as the slope of the linear regression model used to forecast the per plant number of pods with beans. The second column of Table 1 shows the results of this test for the 15 States using October data. This column shows pairs of maturity categories where the regression coefficients (B_j) were not significantly different. These category pairs are candidates to be combined in the next step of the analysis where the equality of intercepts is tested. In Illinois, the slopes for maturity categories 7, 8 & 9 were all significantly different. Therefore no categories were eligible for the next step of the analysis and the maturity categories cannot be combined. The data for Nebraska shows the other extreme where none of the regression coefficients were significantly different. The three categories are candidates to be combined in the next step.

Table 1: Results of Covariance Analysis on Maturity Categories 7,8 & 9
Soybean Objective Yield, October Data, 1977-83.

	Maturity Category Pairs With:		
State 4/	Slopes not Different 1/	Intercepts not Different 2/	Conclusion
Illinois	none	3/	cannot combine
Indiana	7,8 7,9	7,8	combine 7,8
Iowa	8,9	none	cannot combine
Minnesota	7,9	7,9	combine 7,9
Missouri(1)	7,8 7,9	7,8	combine 7,8
Nebraska	7,8 7,9 8,9	7,8 7,9	combine 7,8 or 7,9
Ohio	7,8 7,9	7,8 7,9	combine 7,8 or 7,9
Alabama	none	3/	cannot combine
Arkansas	7,8 7,9 8,9	7,8 7,9	combine 7,8 or 7,9
Georgia	7,8 7,9 8,9	7,8 7,9 8,9	combine 7,8,9
Louisiana	7,8 7,9	7,8 7,9	combine 7,8 or 7,9
Mississippi	7,9	7,9	combine 7,9
Missouri(2)	7,8	7,8	combine 7,8
N. Carolina	7,8 7,9 8,9	7,8 7,9 8,9	combine 7,8,9
S. Carolina	8,9	8,9	combine 8,9
Tennessee	7,8	none	cannot combine

1/ Slopes were not significantly different at alpha =.10.
2/ Intercepts were not significantly different at alpha =.01 given the corresponding slopes were not different.
3/ Test is not necessary.
4/ Missouri soybeans are divided into northern and southern districts.

Model (2a) in Appendix 3 was used to test the equality of intercepts for maturity categories 7, 8 & 9. The intercept test was valid only when the slopes of the maturity category pairs were not different in column 2 of Table 1. The results of this test are listed in column 3 of Table 1. This column shows the maturity category pairs where the slope and intercept of the pods per plant forecast models are not significantly different. Interpretation of mean (intercept) differences as the result of the covariate or treatment effect must be done with caution since the two effects may not be independent. If the slope and intercept of the maturity category effect are not significantly different, then the covariance model (2a) in Appendix 3 is equivalent to a regression model

(2b). The results for North Carolina and Georgia imply that all of the maturity categories can be combined. One model can be used to forecast pods with beans in October since the forecast models for each of the three maturity categories are not significantly different.

Column 4 of Table 1 shows the conclusions drawn from the tests of slopes and intercepts for each of the 15 states. Notice that no maturity categories can be combined in Illinois, Iowa, Alabama or Tennessee. Where the conclusion was to combine categories 7,8 or 7,9 it would be more logical to combine 7,8 since they are adjacent categories. North Carolina and Georgia were the only states where the forecast models for the three categories could be combined into one model.

These results indicate that October maturity categories 7, 8 & 9 cannot be uniformly collapsed into one category. Operationally combining categories by state as recommended in Table 1 is not feasible. However, the covariance procedure can be used to develop forecast models for October categories 7, 8 & 9. Regression coefficients for the three maturities within a state can be created by one covariance model using a data set combined from the three categories (see example in Appendix 4).

Model Comparison

This section compares the fit of covariance model (1) in Appendix 3 which uses separate maturity categories against a model with maturity categories 7, 8 & 9 combined (model 2b). The comparison was based on a ratio of sum of squared errors (SSE) from the full model with three maturity categories to the SSE from the reduced model with combined categories. The ratio was defined as the relative efficiency (RE). An RE of less than one indicates that there is a loss in model fit associated with combining maturity categories. If the RE is close to one the three maturity categories are homogeneous.

The RE's for the 15 states are listed in Table 2 along with the number of observations in the models. The table implies that October maturity categories can be combined if we are willing to accept some loss in model fit. The loss in fit was generally greater in the northern states. In North Carolina, where maturity category effects were not significant, there was virtually no loss in fit from combining data from the three categories.

Table 2: Relative efficiencies from combining October categories 7,8,9 versus separate maturity categories, 1977-83.

State 1/	Number of observations	Relative efficiency 2/
Illinois	556	.861
Indiana	557	.881
Iowa	740	.931
Minnesota	469	.966
Missouri(1)	631	.898
Nebraska	363	.929
Ohio	541	.951
Alabama	677	.972
Arkansas	1234	.992
Georgia	695	.987
Louisiana	729	.978
Mississippi	923	.950
Missouri(2)	351	.906
N. Carolina	683	.996
S. Carolina	685	.949
Tennessee	759	.960

1/ Missouri soybeans are divided into northern and southern districts.
2/ Relative efficiency is defined as the ratio of the sum of squared errors from a covariance model with separate maturity categories vs a model with combined categories. It represents the decrease in model fit by combining October maturity categories.

Year Effect

The final part of the analysis investigated the effect of year on the number of pods per plant forecast models. Operationally these forecast models are built using data from the previous five years, under the implicit assumption that current year's data is not significantly different from the previous five years data. A significant year effect could be the result of unusual weather, changes in survey procedure, etc.

Model (3) in Appendix 3 was used to test the effect of year on the pods per plant forecast models. This model has a second treatment for year effect. The maturity category term was left in model (3) to analyze year effect. This was done since the operational models are developed by maturity category and this research suggests that the categories should not be combined. Results from this model are listed in Table 3. This table shows the conclusions of an F test on the year treatment. The first step in this analysis was to test the effect of year on the slope of the forecast models. A slope adjustment for each of the 7 years used to build the model was computed. The F statistic was used to determine whether any of the year effects were significantly different. The second column in Table 3 shows that the year treatment did not affect the slope of the forecast models in the northern states. However, year effect did significantly affect slope in 5 of the 9 southern soybean objective states. The next step of this analysis was to test the effect of the year treatment on the intercept of the forecast models. This test was only applicable when the year effect on slope was not significant. The results in Table 3 show that the year treatment was significant on the intercepts in 3 states.

Overall the year effect was significant in 8 of the 15 soybean objective yield states. This indicates the data from some years are "different" and will affect the forecast model coefficients. The cause of the significant year effect is difficult to identify but may be due to combinations of weather, changes in survey procedures, changes in work force, etc.

Table 3: Results of F test on year effect from soybean objective yield data, October, 1977-83.

State	Slopes different 1/	Intercepts different 2/
Illinois	no	no
Indiana	no	no
Iowa	no	no
Minnesota	no	no
Missouri(1)	no	no
Nebraska	no	yes
Ohio	no	yes
Alabama	yes	3/
Arkansas	yes	3/
Georgia	yes	3/
Louisiana	no	no
Mississippi	no	no
Missouri(2)	no	yes
N. Carolina	yes	3/
S. Carolina	yes	3/
Tennessee	no	no

1/ Test that the slopes of forecast models from each of the 7 years are equal at alpha = .10.

2/ Test that the intercepts of forecast models from each year are equal at alpha = .10 given that the slopes are equal.

3/ If slopes are not equal the intercept test is invalid.

RECOMMENDATIONS

The results of the covariance analysis show that October soybean forecast models for maturity categories 7, 8 & 9 cannot be combined into one model without some loss in model fit. This was due to a maturity category effect which resulted in significant differences in slopes and/or intercepts between the forecast models. An analysis of year effect indicated that some years contributed a significant change to the slopes and intercepts of the pods per plant forecast models in some states. Significant year effects were more prevalent in southern than northern objective yield states. The year effect on model development is under study in a cooperative agreement with Iowa State University.

Based on these findings, we recommend the following:

1. Do not combine forecast models to predict the number of pods with beans per plant for October maturity categories 7, 8 & 9.

2. Investigate the possibility of combining soybean categories 3,4 & 5 since the forecast models for these maturity categories use the same independent variables.

3. Methods staff should consider the use of covariance procedures to develop forecast models where the variables used in the models do not change across adjacent maturity categories.

REFERENCES

1. Battaglia, Robert and Bovard, Gary, "Soybean Objective Yield Research: Assessment of the Revised Forecast Procedure," Statistical Reporting Service, U.S. Department of Agriculture, 1984.

2. Battaglia, Robert and Klugh, Benjamin, "Assessment of Fixed Variable vs Stepwise Forecast Models to Predict Number of Soybean Pods with Beans Per Plant," Statistical Reporting Service, U.S Department of Agriculture, 1984.

3. Battaglia, Robert and Klugh, Benjamin, "Fixed vs Stepwise Forecast Models to Predict Number of Pods with Beans per Soybean Plant in Southern States," Statistical Reporting Service, U.S. Department of Agriculture, 1985.

4. Neter, John and Wasserman, William, "Applied Linear Statistical Models," Illinois, Richard D. Irwin, Inc., 1974.

5. Freund, Rudolf and Littell, Ramon, "SAS for Linear Models," North Carolina, SAS Institute Inc, 1981.

6. U.S. Department of Agriculture, Statistical Reporting Service, "Objective Yield Supervising and Editing Manual," 1984.

APPENDIX 1

Soybean objective yield maturity category definitions.

Maturity Category	Description
0	No plants were present in either row of the two 6-inch row sections.
1	No pods with beans are present and the ratio of total fruit to mainstem nodes is less than .20.
2	No pods with beans are present and the ratio of total fruit to mainstem nodes is between .20 and 1.75.
3	No pods with beans are present and the ratio of total fruit to mainstem nodes is greater than 1.75.
4	Pods with beans are present and the ratio of pods with beans to total fruit is less than .05
5	The ratio of pods with beans to total fruit is between .05 and .2.
6	The ratio of pods with beans to total fruit is between .20 and .65.
7	The ratio of pods with beans to total fruit is between .65 and .85.
8	Pods filled, leaves turning yellow or the ratio of pods with beans to total fruit is greater than .85.
9	Pods turning brown, leaves shedding.
10	Pods brown, almost mature or pods mature.

APPENDIX 2

Northern states fixed-variable forecast models to forecast number of pods with beans

Month	Maturity	Forecast variables
Aug	1	Plants, Mainstem nodes
Aug	2	Lateral branches with pods, Plants
Aug	3	Lateral branches with pods, Total blooms & pods
Aug	4	Lateral branches with pods, Plants
Aug	5	Lateral branches with pods, Total blooms & pods
Aug	6	Lateral branches with pods, Total blooms & pods
Sept	5	Total blooms & pods
Sept	6	Total blooms & pods, Pods with beans
Sept	7	Pods with beans
Sept	8	Pods with beans
Oct	7	Pods with beans
Oct	8	Pods with beans
Oct	9	Pods with beans

Southern states fixed-variable forecast models to forecast number of pods with beans

Month	Maturity	Forecast variables
Sept	2	Plants, Lateral branches with pods
Sept	3	Lateral branches with pods, Total blooms & pods
Sept	4	Lateral branches with pods, Total blooms & pods
Sept	5	Lateral branches with pods, Total blooms & pods
Sept	6	Lateral branches with pods, Pods with beans
Sept	7	Pods with beans
Sept	8	Pods with beans
Oct	6	Lateral branches with pods, Pods with beans
Oct	7	Pods with beans
Oct	8	Pods with beans
Oct	9	Pods with beans
Nov	9	Pods with beans

APPENDIX 3 Covariance models used in the analysis.

Model to test equality of slopes

$$Y_{ij} = U + A_j + (B_j)(Z_{ij}) + E_{ij} \quad (1)$$

where: Y_{ij} = Final number of pods with beans i^{th} observation in j^{th} maturity
U = Overall mean
A_j = Treatment effect of jth maturity
B_j = Regression coefficient
Z_{ij} = October number of pods with beans
E_{ij} = Error term

Model to test equality of intercepts

$$Y_{ij} = U + A_j + (B)(Z_{ij}) + E_{ij} \quad (2a)$$

where B is constant for all maturities

If the Aj are not different the covariance model becomes a regression model.

$$Y_{ij} = (U + A) + BZ_{ij} + E_{ij} \quad (2b)$$

where A is constant for all ınaturities

Model to test equality of years

$$Y_{ijk} = U + A_j + C_k + B_jZ_{ijk} + C_kZ_{ijk} + E_{ijk} \quad (3)$$

where C_k = treatment effect of k^{th} year

APPENDIX 4

October forecast models for maturity categories 7,8 & 9 can be developed using a covariance model. In this example Illinois data from 1977-83 was used to construct the forecast models. SAS procedures GLM or REG can be used but REG offers more diagnostics. The form of the model used in REG is:

FP = INT + MC7 + MC8 + OP + OP*MC7 + OP*MC8

where: FP = Final number of pods with beans.
OP = October number of pods with beans.
INT = Intercept
MC7 = Dummy variable for maturity category 7.
MC8 = Dummy variable for maturity category 8.
OP*MC7 = Slope adjustment for maturity category 7.
OP*MC8 = slope adjustment for maturity category 8.
N = Number of observations used to construct model = 555

ESTIMATED PARAMETERS	SE	T	PROB > |T|
INT = 0.9023	.2423	0.4	.7049
MC7 = 2.6188	.9453	2.8	.0058
MC8 = -0.1290	.3488	-0.4	.7116
OP = 0.9818	.0077	127.9	.0001
OP*MC7 = -0.1866	.0397	-4.7	.0001
OP*MC8 = -0.0431	.0111	-3.9	.0001

Forecast model for category 9 = .9023 + .9818(OP)

Forecast model for category 8 = .7733 + .9387(OP)

Forecast model for category 7 = 3.5211 + .7951(OP)