AN APPLICATION OF NONLINEAR ESTIMATION

by

Dwight A. Rockwell

and

Jack Nealon

PRESENTED AT THE NINTH INTERNATIONAL BIOMETRIC CONFERENCE

August 27, 1976

STATISTICAL REPORTING SERVICE

UNITED STATES DEPARTMENT OF AGRICULTURE

GDH

TABLE OF CONTENTS

AN APPLICATION OF NONLINEAR ESTIMATION

ABSTRACT

This paper describes a nonlinear model that has application for forecasting growth phenomenon. The parameters for the model are estimated from current data rather than historic data. Two examples are presented utilizing crop characteristic data. Analyses of the residuals from the estimated model clearly indicate the violation of model assumptions. A procedure for transforming data to meet model assumptions is demonstrated. Also, a method to improve the estimating ability of the model by double sampling is shown.

INTRODUCTION

The increased dependence of world markets on United States grain production has augmented the interest in preharvest production forecasts. Farmers, consumers, grain exporters and government policy officials rely heavily upon these forecasts.

A grain production forecast consists of two components: (1) number of acres to be harvested and (2) yield per acre. Acreages for harvest are relatively stable from year to year. Also, farmers' plans to harvest a certain number of acres seldom change significantly during a growing season, except for isolated instances of total crop failure. Yield per acre, however, is considearbly more difficult to forecast, because yields vary substantially from year to year. Also, early season forecasts can be invalidated by sudden changes in growing conditions. Because of the complex nature of yield, current methodology does not always perform as well as desired. This paper presents an alternative approach to forecasting crop yield using nonlinear estimation.

TRADITIONAL APPROACH

The traditional approach to forecasting crop yields has consisted of two methods: subjective and objective. Both methods rely on what may be called "between-year" models.

The subjective method has consisted of surveying a nonprobabilistic sample of farm operators. Each farm operator is asked to report what he expects his crop will yield per acre. The results are summarized and an average yield is computed. Previous years' (historic) data are used in a linear regression model to regress estimated yield at harvest against preharvest forecasts by farmers to allow for farmers' tendency to overstate or understate potential yield.

There are several limitations to the subjective method. First, although farmers can provide a reasonable forecast of their yield, the forecast is subjective and therefore susceptible to the farmers' bias. Second, farmers' ability to judge yield potential is probably not independent of crop quality. That is, farmers' ability to forecast yield is probably not the same in good crop years as it is in bad crop years. Finally, a predetermined number of years' data are required to estimate a relationship between farmers' preharvest yield forecasts and harvested yields. Therefore, a current year forecast requires the use of historic data.

In the objective method, data are collected from randomly selected field plots. During the growing season, plant counts and fruit measurements are made for each plot. The estimated yield at harvest for each plot is regressed against the preharvest plant counts and fruit measurements. Historic data are used to estimate the parameters of a linear model for each of several maturity categories. Current year counts and measurements are then used in conjunction

with this model to forecast current year yield at harvest during the growing season.

It is necessary to determine how many years' data to use in estimating the parameters of the model. In practice, the previous three years' data are used. This means that each year, one year's historic data are excluded, the most recent historic year's data are included, and new estimates for the parameters are computed. Regardless of how many or which years are used to estimate the model's parameters, it is assumed that current year data will conform to that model. If this assumption is violated, the objective method will not provide reliable preharvest yield forecasts.

ALTERNATIVE APPROACH

Limitations of the traditional approach have prompted considerable interest in the possibility of using only current year data from randomly selected field plots as a basis for developing crop yield forecasts during the growing season. Therefore, efforts have been directed toward testing a "within-year" forecasting model for which the parameters can be estimated from current season data only.

A within-year model would have the advantage of providing yield forecasts without the dependence on historic data to estimate the parameters. Therefore, a within-year model would reflect unique characteristics of the year for which the forecast was desired.

A within-year model could be a valuable supplement to a between-year model. Supplemental information from a within-year model may improve crop forecasts for atypical years in which growing conditions differ greatly from the three previous years that were used to generate the parameter values in a between-year model.

In addition to providing supplemental information to the present yield forecasting system, a within-year model could be very useful in developing a forecasting model for crops not in the present crop yield forecasting system. A within-year model could be developed in a shorter time period, since historic data would be unnecessary.

Various within-year models have been examined to determine if biological laws relate corn or spring wheat growth, in terms of dry kernel weight, to time scales closely related to initial kernel formation. Results demonstrate that a within-year model, known as the logistic growth model, describes the process by which dry kernel weight accumulates in corn or spring wheat.[1/ 2/ 3/]

The objectives of this paper are to:

(1) Discuss the form of the logistic growth model.

(2) Describe a method for estimating the parameters of a nonlinear model, such as the logistic growth model.

(3) Illustrate the use of the logistic growth model as a forecasting methodology for corn and spring wheat.

(4) Describe two variations of this model, which allow for relaxing certain assumptions concerning the residuals.

(5) Explain the use of double sampling to refine the dependent variable.

LOGISTIC GROWTH MODEL

The logistic growth model is a nonlinear model that uses incipient data for the independent and dependent variables to estimate values for the parameters. The model then can be used to forecast the dependent variable for a particular value of the independent variable. The form of the model is:

$$y_i = \frac{1}{\alpha + \beta(\rho)^{t_i}} + u_i; \quad i = 1, 2, \ldots, n$$

α, β and ρ = parameters

$\alpha > 0$, $\beta > 0$, $0 < \rho < 1$

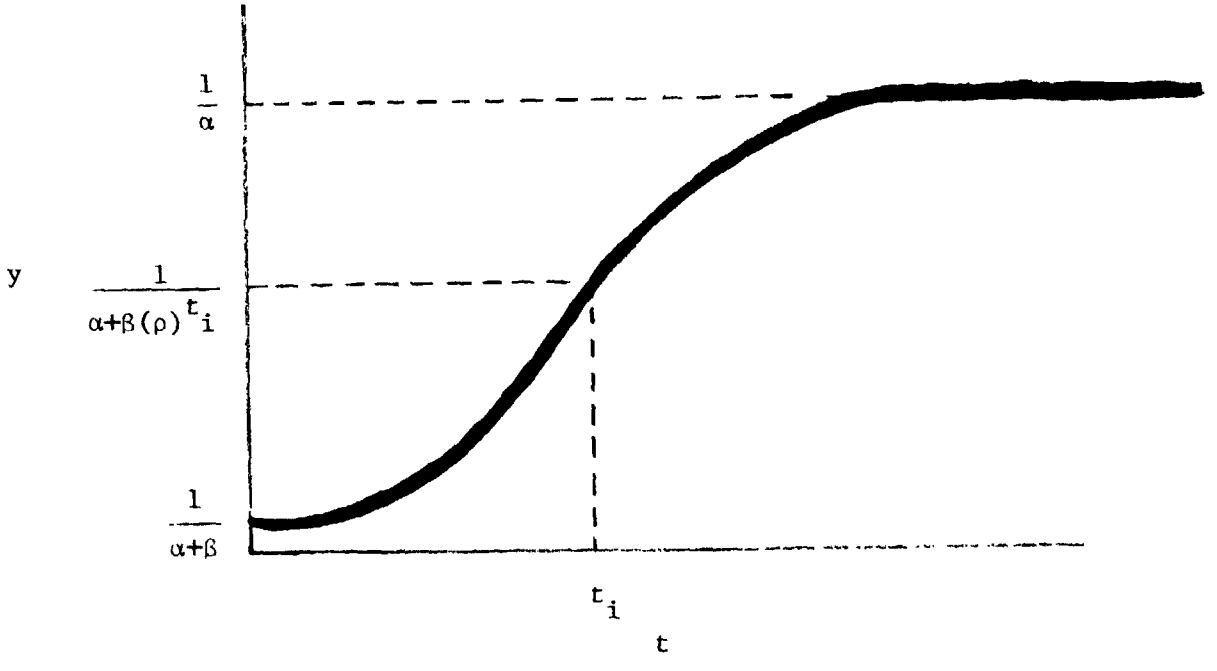
u_i = disturbance term

t_i = independent variable

y_i = dependent variable

In applying this model to dry kernel weight accumulation for an individual stalk of corn or spring wheat, the hypothesis is that accumulation begins slowly at first, increases at an increasing rate for a period of time and then increases at a decreasing rate until a maximum (asymptotic) value is attained. This asymptotic value represents the dry kernel weight per stalk at harvest. The point in the phenological development of a stalk coinciding with time equal to zero should approximate as closely as possible the initial stages of kernel development such as silk emergence in corn or flowering in wheat.

The logistic growth model is shown graphically below.



ESTIMATION

The logistic growth model is intrinsically nonlinear in the unknown parameters α, β and ρ. Therefore, the method of least squares is not directly applicable for fitting this model to sample data.

One method for estimating the parameters of a nonlinear model is the linearization (or Taylor series) method. [4/] The linearization method is a widely used method for computing nonlinear least square estimators. In general, the form of the model is:

$$y_i = f(\underline{X}_i, \underline{\Theta}) + u_i \; ; \; i = 1, \ldots, n$$

provided that y_i is the value of the dependent variable, $\underline{X}_i' = (X_{i1}, X_{i2}, \ldots, X_{ik})$ is the vector of k independent variables, u_i is the disturbance term for the i^{th} observation and $\underline{\Theta}' = (\Theta_1, \Theta_2, \ldots, \Theta_p)$ is the vector of p unknown parameters to be estimated.

Beginning with an initial estimate of the parameters, $\underline{\Theta}_0' = (\Theta_{10}, \Theta_{20}, \ldots, \Theta_{p0})$, the procedure involves carrying out a Taylor series expansion of $f(\underline{X}_i, \underline{\Theta})$ about the point $\underline{\Theta}_0$ and disregarding the terms beyond the first derivative. Then, when $\underline{\Theta}$ is close to $\underline{\Theta}_0$, y_i is approximated by:

$$y_i \doteq f(\underline{X}_i, \underline{\Theta}_0) + \sum_{j=1}^{p} \left[\frac{\partial f(\underline{X}_i, \underline{\Theta})}{\partial \Theta_j} \right]_{\underline{\Theta} = \underline{\Theta}_0} (\Theta_j - \Theta_{j0}) + u_i \; ; \; i = 1, \ldots, n.$$

All information available from theory and previous survey results concerning the population being sampled would be used in estimating initial values for the parameters.

This equation expressed in matrix notation is:

$$(\underline{Y} - \underline{f}_0) \doteq \underline{Z}_0 \, \underline{\gamma}_1 + \underline{U}$$

provided that

$$(\underline{Y} - \underline{f}_0) = \begin{pmatrix} y_1 - f(\underline{X}_1, \underline{\theta}_0) \\ y_2 - f(\underline{X}_2, \underline{\theta}_0) \\ \vdots \\ y_n - f(\underline{X}_n, \underline{\theta}_0) \end{pmatrix}, \quad \underline{\gamma}_1 = \begin{pmatrix} \theta_1 - \theta_{10} \\ \theta_2 - \theta_{20} \\ \vdots \\ \theta_p - \theta_{p0} \end{pmatrix}, \quad \underline{U} = \begin{pmatrix} u_1 \\ u_2 \\ \vdots \\ u_n \end{pmatrix}$$

and

$$\underline{Z}_0 = \begin{pmatrix} \left.\frac{\partial f(\underline{X}_1, \underline{\theta})}{\partial \theta_1}\right|_{\underline{\theta} = \underline{\theta}_0} & \left.\frac{\partial f(\underline{X}_1, \underline{\theta})}{\partial \theta_2}\right|_{\underline{\theta} = \underline{\theta}_0} & \cdots & \left.\frac{\partial f(\underline{X}_1, \underline{\theta})}{\partial \theta_p}\right|_{\underline{\theta} = \underline{\theta}_0} \\ \left.\frac{\partial f(\underline{X}_2, \underline{\theta})}{\partial \theta_1}\right|_{\underline{\theta} = \underline{\theta}_0} & & & \\ \vdots & & \ddots & \vdots \\ \left.\frac{\partial f(\underline{X}_n, \underline{\theta})}{\partial \theta_1}\right|_{\underline{\theta} = \underline{\theta}_0} & \cdots & & \left.\frac{\partial f(\underline{X}_n, \underline{\theta})}{\partial \theta_p}\right|_{\underline{\theta} = \underline{\theta}_0} \end{pmatrix}$$

The parameter vector, $\underline{\gamma}_1$, can then be estimated by applying ordinary least squares to obtain

$$\hat{\underline{\gamma}}_1 = (\underline{Z}_0' \, \underline{Z}_0)^{-1} \, \underline{Z}_0' \, (\underline{Y} - \underline{f}_0)$$

provided that

$$\hat{\gamma}_1 = \begin{pmatrix} \theta_{11} - \theta_{10} \\ \theta_{21} - \theta_{20} \\ \vdots \\ \theta_{p1} - \theta_{p0} \end{pmatrix}.$$

The vector, $\hat{\gamma}_1$, will minimize the error sum of squares,

$$SS(\hat{\gamma}_1) = \sum_{i=1}^{n} \left\{ \left[y_i - f(\underline{X}_i, \underline{\theta}_0) \right] - \sum_{j=1}^{p} \left[\frac{\partial f(\underline{X}_i, \underline{\theta})}{\partial \theta_j} \right]_{\underline{\theta} - \underline{\theta}_0} (\theta_j - \theta_{j0}) \right\}^2,$$

with respect to the $(\theta_j - \theta_{j0})$; $j = 1, \ldots, p$.

Using $\underline{\theta}_1' = (\theta_{11}, \theta_{21}, \ldots, \theta_{p1})$ as a revised estimate of the unknown parameter vector $\underline{\theta}' = (\theta_1, \theta_2, \ldots, \theta_p)$, the θ_{j1} ; $j = 1, 2, \ldots, p$ can be placed in the same role as the θ_{j0} ; $j = 1, \ldots, p$ in the previous equations. The process of deriving the least squares solution is repeated, and another revised estimate $\underline{\theta}_2' = (\theta_{12}, \theta_{22}, \ldots, \theta_{p2})$ is obtained.

This iterative process is continued until the solution converges. The criterion for convergence might be

$$\left| \frac{\theta_{j(k+1)} - \theta_{jk}}{\theta_{jk}} \right| < \delta_1 \; ; \; j = 1, \ldots, p.$$

or, alternatively,

$$\left| \frac{SS(\hat{\gamma}_{(k+1)}) - SS(\hat{\gamma}_k)}{SS(\hat{\gamma}_k)} \right| < \delta_2$$

in successive iterations k and (k+1), where δ_1 or δ_2 would be predetermined tolerance values.

Note that with the terminating $(k+1)^{th}$ iteration, the $SS(\hat{\gamma}_{k+1})$ will be the minimum attainable error sum of squares to the accuracy level imposed by the termination criterion chosen. One should be aware of the effects of this

limitation. For example, even though the error term, u, of the nonlinear model is assumed to be normally distributed, $\hat{\underline{\theta}}$, is not normally distributed, $\hat{\sigma}^2 = SS(\hat{\gamma}_{k+1})/(n-p)$ is not an unbiased estimate of σ^2, and confidence intervals constructed for population parameters are only approximate. Of course, the more closely the sample data fit the hypothesized model and the smaller the termination criterion, the better the approximation will be. Preliminary simulation using data collected for corn was analyzed to determine the distribution of each parameter in the logistic growth model. A test designed by Wilk and Shapiro (the W test)[5/] failed to reject at the .05 signficance level the null hypothesis that estimates of each parameter were normally distributed. Therefore, although $\hat{\underline{\theta}}$ is not normally distributed, simulation has exemplified that it may be very close to being normally distributed.

The results to be described now and the experience of other people[6/] indicate that in practice the linearization method provides reasonable estimates.

APPLICATION

As previously mentioned, the logistic growth model has been tested as a forecasting methodology for corn and spring wheat. Data were collected from plots in 10 Iowa corn fields and 3 North Dakota spring wheat fields.

For corn, the independent time variable was the number of days from silk emergence to the time the corn plant was sampled. The dependent variable was the mean dry kernel weight (grams) of all ears per plant for all plants with the same associated value of time and drawn from the same sample field. The period of time since silk emergence for a plant was based on the time of the primary ear. It was assumed the residuals in this model are independently distributed with mean zero and a constant variance, σ_u^2. That is,

$$E(\underline{U}) = 0 \qquad \text{and}$$

$$E(\underline{U}\,\underline{U}') = \sigma_u^2 I_n$$

The model, $y_i = \dfrac{1}{\alpha + \beta(\rho)^{t_i}} + u_i$, $i = 1, \ldots, n$, (1)

was fitted to sample data available through August 15, September 1, September 15, October 1, October 15, and the end of the growing season. This incrementing of data was done to indicate how early in the growing season the parameters of the model could be estimated and how these estimates changed as additional information became available. To evaluate the estimated model, the asymptotic value for each of the six calendar date cutoffs was compared with an estimate of the mean dry grain weight per plant at harvest for the 10 sample fields combined.

The nonlinear least squares option of the Biomedical Computer Programs (BMD) package[7/] was used to estimate the parameters of the model for each of the six cutoffs. This computer program uses a variation of the linearization method. Table 1 shows the estimated value for each parameter, the estimated

Table 1

cut off	n	$\hat{\alpha}$	$\hat{\beta}$	$\hat{\rho}$	$\hat{\sigma}_{\hat{\alpha}/\hat{\alpha}}$	$\hat{\sigma}_{\hat{\beta}/\hat{\beta}}$	$\hat{\sigma}_{\hat{\rho}/\hat{\rho}}$	$\lim_{t_i \to \infty} y_i$	% of est'ed hv. wt.
all obs.	278	.0061541	.15655	.91866	3.59	34.36	0.98	162.49	106.3
10/15	256	.0058769	.15263	.92037	3.85	31.85	0.90	170.16	111.4
10/1	197	.0053225	.16184	.91977	5.66	29.67	0.90	187.88	123.0
9/15	128	.0063958	.40740	.88626	6.39	38.86	1.34	156.35	102.3
9/1	70	.0063116	.69776	.86809	15.59	49.90	1.91	158.44	103.7
8/15	19	.016119	14.127	.74074	18.90	134.18	6.96	62.04	40.6

relative standard error for each estimated parameter, the estimated asymptotic value and the estimated asymptotic value as a percent of the estimated grain weight per plant at harvest for each of the six cutoffs. These results show considerable variability in the asymptotic value among the six cutoffs. Also, these data indicate little success may be expected in estimating a reliable model based only on data collected up through mid-August. The plot on page 12 (Figure 1) shows the data being fitted and the estimated model for the October 1 forecast.

The logistic growth model was also fitted to data collected for spring wheat to determine if dry kernel weight accumulation follows the growth phenomenon described by this model. The dependent variable, y_i, was defined as the mean dry kernel weight in grams for stalks from the same sample field with the same value of the independent variable. The independent variable was the period of time in days since a phenological event occurred until the stalk was sampled. Phenological events observed were flowering, head emergence and head swelling. Table 2 shows the number of observations, the estimated value and relative standard error of each parameter and the estimated value of the dependent variable at harvest for each phenological event using all the data.

Table 2

Phenological Event	n	$\hat{\alpha}$	$\hat{\beta}$	$\hat{\rho}$	$\hat{\sigma}_{\hat{\alpha}/\alpha}$ %	$\hat{\sigma}_{\hat{\beta}/\beta}$ %	$\hat{\sigma}_{\hat{\rho}/\rho}$ %	$\lim_{t_i \to \infty} y_i$
Flowering	40	1.7016	35.125	.80761	7.14	57.68	4.52	.588
Head Emergence	63	1.7253	184.66	.78174	6.01	80.75	4.56	.580
Head Swelling	66	1.6593	165.25	.8051	6.88	79.36	4.09	.603

The data fit for the phenological event, flowering, is shown in Figure 2 on page 13. These results indicate that spring wheat does follow the growth

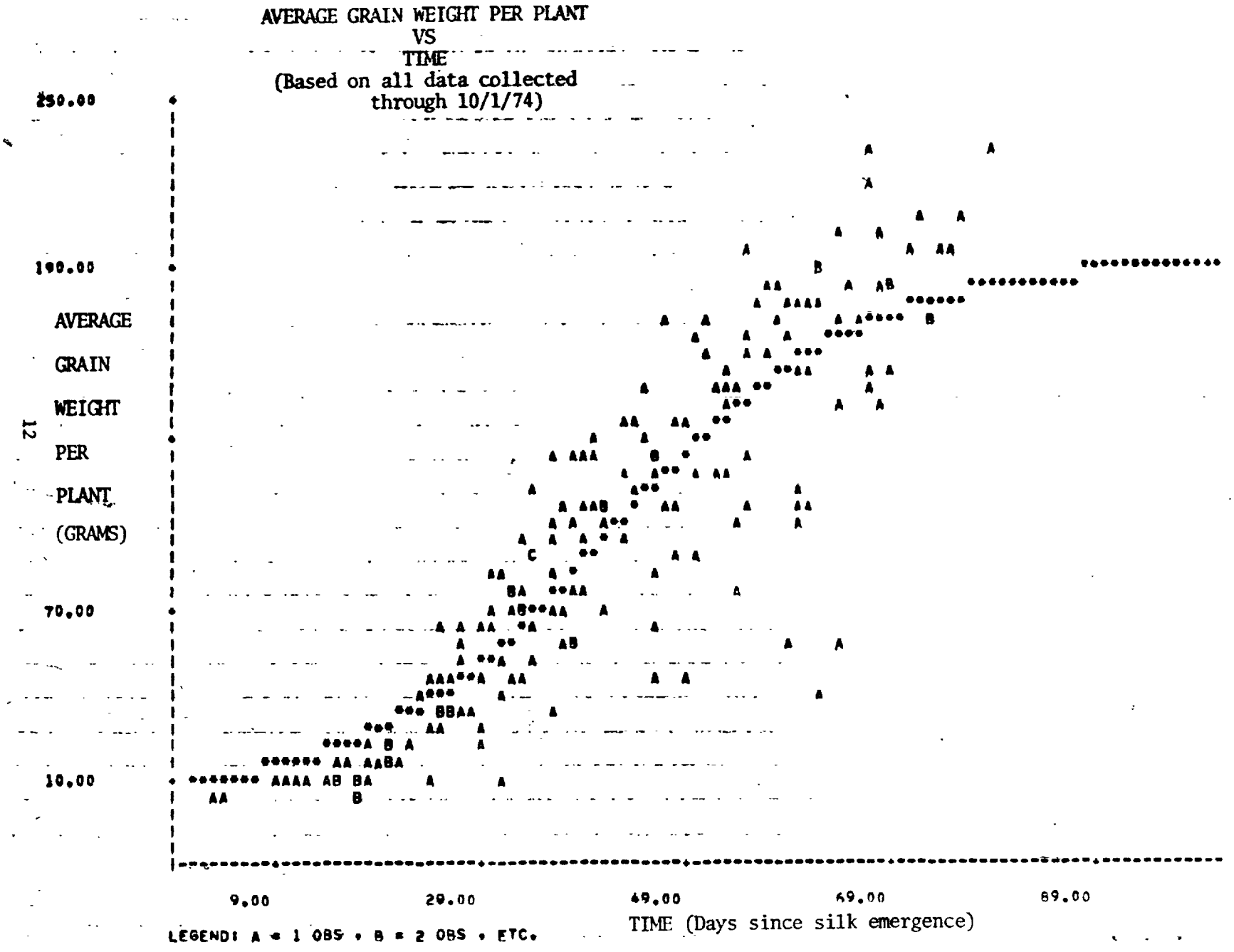

AVERAGE GRAIN WEIGHT PER PLANT
VS
TIME
(Based on all data collected
through 10/1/74)
250.00
190.00
70.00
10.00
AVERAGE
GRAIN
WEIGHT
PER
PLANT
(GRAMS)
9.00
29.00
49.00
69.00
89.00
TIME (Days since silk emergence)
LEGEND: A = 1 OBS, B = 2 OBS, ETC.

Figure 1

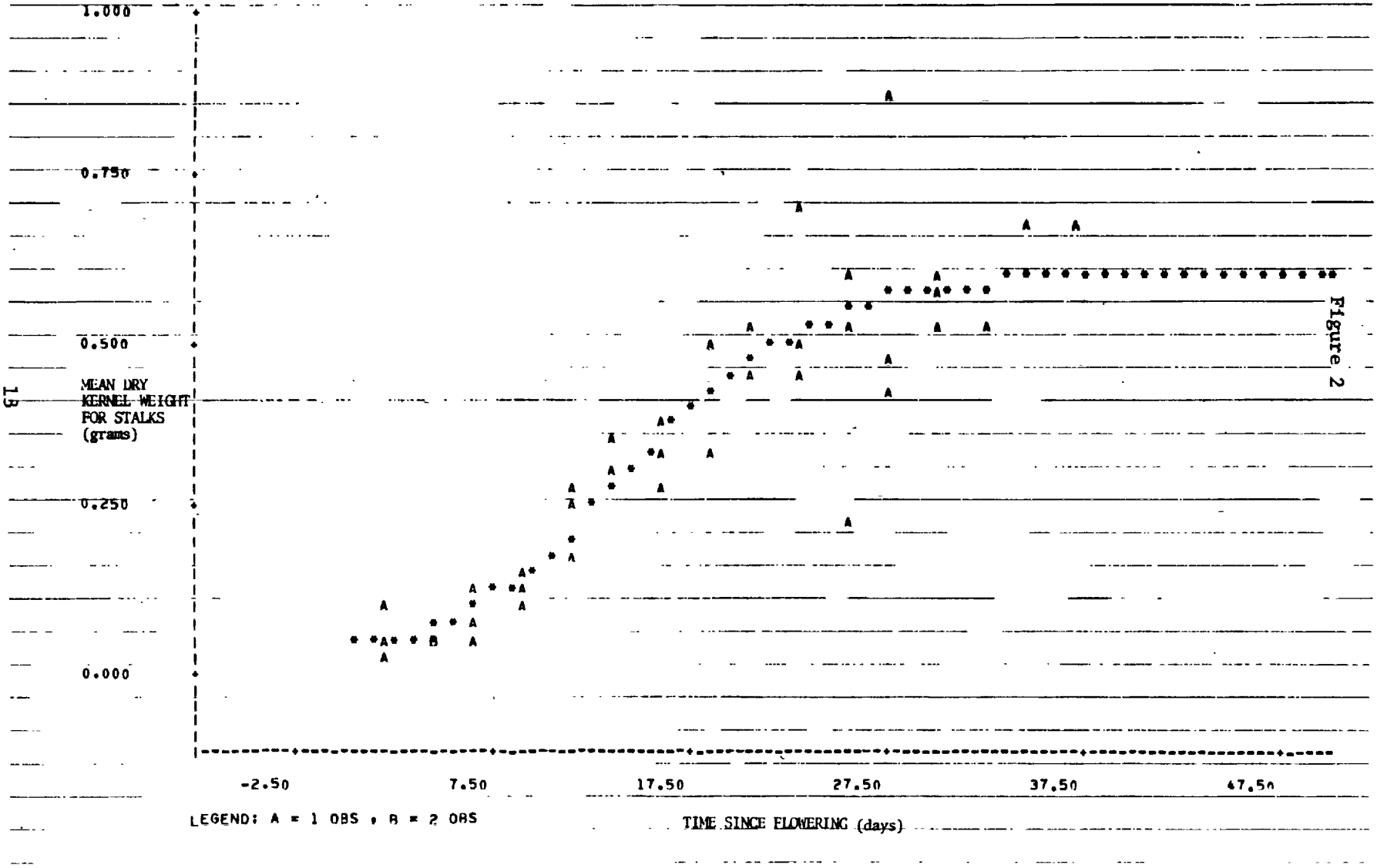

1.000
0.750
0.500
0.250
0.000
MEAN DRY KERNEL WEIGHT FOR STALKS (grams)
-2.50
7.50
17.50
27.50
37.50
47.50
LEGEND: A = 1 OBS , B = 2 OBS
TIME SINCE FLOWERING (days)

Figure 2

behavior described by the logistic growth model. Future research will investigate the use of this model as a forecasting technique for spring wheat.

HETEROSCEDASITY

After fitting the logistic growth model to the six subsets of corn data and all data for spring wheat, an attempt was made to evaluate how well the underlying assumptions concerning the residuals had been met. An examination of the plots showing the fitted model and data points for each cutoff date for corn and all data for spring wheat indicated that there may be a statistically significant relationship between the variation in the residuals and the independent time variable. Plots showed that the estimated residuals become larger for larger values of time. In other words, the data may be violating the assumption

$$E\ (u_i^2) = \sigma_u^2$$

for all i. This condition is commonly referred to as heteroscedasity.

To pursue this possibility, a method suggested by Glejser was used.[8/] It is assumed that each residual, u_i, can be expressed as

$$u_i = v_i\ f(t_i);\ i = 1, \ldots, n$$

provided that v_i is a random variable with

$$E\ (\underline{V}) = 0 \text{ and}$$

$$E\ (\underline{V}\ \underline{V}') = \sigma_v^2\ I_n.$$

Also, it is assumed that the form of the function, f, is known, but at least one of its parameters is unknown. It then follows that instead of the original assumption that

$$E\ (\underline{U}\ \underline{U}') = \sigma_u^2\ I_n,$$

now it is assumed that

$$E\ (\underline{U}\ \underline{U}') = \sigma_v^2\ \Omega$$

provided that

$$\Omega = \begin{pmatrix} [f(t_1)]^2 & 0 & \cdots & 0 \\ 0 & [f(t_2)]^2 & & 0 \\ \vdots & & \ddots & \vdots \\ 0 & 0 & \cdots & [f(t_n)]^2 \end{pmatrix}.$$

It can be shown that if the assumption of heteroscedasity holds true, using model (1) will give less efficient estimates of the parameters. That is, the estimated standard errors of the parameters will be unnecessarily large. Applying the method of generalized least squares at each iteration of the linearization procedure, an estimte of $\underset{\sim}{\gamma}_{k+1}$ would be given by

$$\hat{\underset{\sim}{\gamma}}_{k+1} = (\underset{\sim}{Z}_k' \, \underset{\sim}{\Omega}^{-1} \, \underset{\sim}{Z}_k)^{-1} \, \underset{\sim}{Z}_k' \, \underset{\sim}{\Omega}^{-1} \, (\underset{\sim}{Y} - \underset{\sim}{f}_k).$$

Alternatively, the same estimate of $\underset{\sim}{\gamma}_{k+1}$ would be obtained if the model

$$\frac{1}{f(t_i)} \, y_i = \frac{1}{f(t_i)} \, \frac{1}{\alpha + \beta(\rho)^{t_i}} + \frac{1}{f(t_i)} \, u_i; \quad i = 1, \ldots, n \qquad (2)$$

were fitted to the sample data using ordinary least squares at each iteration. Either procedure can be used with BMD. Note that the residuals of model (2),

$$\frac{u_i}{f(t_i)} = v_i \; ; \; i = 1, \ldots, n$$

do have the desired characteristic of being independently distributed with mean zero and a constant variance, σ_v^2.

Since neither the function, f, nor its parameters are known, they must be estimated. Corn data will be used to illustrate the fit to the logistic growth model based upon the heteroscedastic-error adjustment. Following the procedure outlined by Glejser, the absolute value of the estimated residuals obtained from fitting model (1), $|\hat{u}_i|$; $i = 1, \ldots, n$ were regressed on an

estimated function of time. An examination of plots of the absolute value of the residuals against time suggested the function

$$|\hat{u}_i| = f(t_i) = \tau_0 + \tau_1 t_i + e_i \; ; \; i = 1, \ldots, n.$$

Since the estimated value of τ_0 was not significantly different from zero for any of the six cutoff dates, the function

$$|\hat{u}_i| = f(t_i) = \tau t_i + e_i \; ; \; i = 1, \ldots, n$$

was used. Estimates of τ were significant for all cutoffs. The results of fitting model (2) using the estimates for f (t) are shown in Table 3.

Table 3

cut off	n	$\hat{\alpha}$	$\hat{\beta}$	$\hat{\rho}$	$\hat{\sigma}_{\hat{\alpha}}/\hat{\alpha}$ (%)	$\hat{\sigma}_{\hat{\beta}}/\hat{\beta}$ (%)	$\hat{\sigma}_{\hat{\rho}}/\hat{\rho}$ (%)	$\lim_{t_i \to \infty} y_i$	% of est'ed hv. wt.
all obs.	278	.0068499	.50744	.88328	2.99	23.22	0.79	145.99	95.5
10/15	256	.0066370	.47529	.88618	3.22	22.26	0.76	150.67	98.6
10/1	197	.0063319	.50529	.88455	3.87	22.03	0.76	157.93	103.4
9/15	128	.0070626	.76550	.86517	5.32	25.63	1.00	141.59	92.7
9/1	70	.0069929	.95989	.85613	12.07	29.70	1.33	143.00	93.6
8/15	19	.017680	28.315	.71327	16.90	128.11	6.89	56.56	37.0

A comparison of Tables 1 and 3 shows the estimated relative error of the estimated parameters is now smaller, as expected. The asymptotic values of the estimated heteroscedastic-error model are at a somewhat lower level than those of model (1). Perhaps the most attractive aspect of the results from model (2) is that the asymptotic values are less variable among the six cutoffs. Note the October 1 value is substantially more in line with the other cutoffs than before. The August 15 value is still far from being realistic.

AUTOCORRELATION

Further examination of the plots from model (1) for the Iowa corn data indicated the residuals may not be independently distributed. Specifically, for small values of time most of the data points lie below the function, particularly for the last three cutoffs. This led to hypothesizing a third set of assumptions concerning the residuals.

Assume the residuals in model (1), u_i, can be expressed as

$$u_i = v_i \, f(t_i); \; i = 1, \ldots, n$$

same as for the heteroscedastic-error model, but further assume the v_i follow a first-order autoregressive scheme:[9/]

$$v_i = \lambda \, v_{i-1} + \varepsilon_i$$

provided that $|\lambda| < 1$ and the ε_i satisfy the following assumptions:

$$E(\varepsilon_i) = 0$$

$$E(\varepsilon_i \, \varepsilon_{i+s}) = \sigma_\varepsilon^2 \; ; \; s = 0$$

$$= 0 \; ; \; s \neq 0$$

for all i. It then follows that

$$E(\underline{U} \, \underline{U}') = \sigma_v^2 \, \underline{\Omega}$$

provided that

$$\underline{\Omega} = \begin{pmatrix} [f(t_1)]^2 & f(t_1) f(t_2) \lambda & f(t_1) f(t_3) \lambda^2 \cdots & f(t_1) f(t_n) \lambda^{n-1} \\ f(t_1) f(t_2) \lambda & [f(t_2)]^2 & f(t_2) f(t_3) \lambda & f(t_2) f(t_n) \lambda^{n-2} \\ f(t_1) f(t_3) \lambda^2 & f(t_2) f(t_3) \lambda & [f(t_3)]^2 & f(t_3) f(t_n) \lambda^{n-3} \\ \vdots & & & \\ f(t_1) f(t_n) \lambda^{n-1} & f(t_2) f(t_n) \lambda^{n-2} & f(t_3) f(t_n)^{n-3} \cdots & [f(t_n)]^2 \end{pmatrix}.$$

It can be shown that if the assumption of autocorrelation holds true, using model (2) will underestimate the true sampling variance of the estimated parameters. As in the case of the heteroscedastic-error model, the method of generalized least squares can be applied at each iteration of the linearization procedure and an estimate of $\underline{\gamma}_{k+1}$ obtained by

$$\hat{\underline{\gamma}}_{k+1} = (\underline{Z}_k' \, \underline{\Omega}^{-1} \, \underline{Z}_k)^{-1} \, \underline{Z}_k' \, \underline{\Omega}^{-1} \, (\underline{Y} - \underline{f}_k) \quad .$$

However, this approach cannot be used with BMD when $\underline{\Omega}^{-1}$ is not a diagonal matrix. Therefore, a transformation matrix, T, must be utilized such that a new model will be formulated that can be fitted by ordinary least squares and that will have a scalar dispersion matrix. That is,

$$E \, (\underline{T} \, \underline{U} \, \underline{U}' \, \underline{T}') = \sigma^2 \, I_n \quad .$$

It can be verified by multiplying out that if $\underline{T}_1$ is defined as

$$\underline{T}_1 = \begin{pmatrix} \frac{\sqrt{1-\lambda^2}}{f(t_1)} & 0 & 0 \cdots\cdots 0 & 0 & 0 \\ \frac{-\lambda}{f(t_1)} & \frac{1}{f(t_2)} & 0 \cdots\cdots 0 & 0 & 0 \\ 0 & \frac{-\lambda}{f(t_2)} & \frac{1}{f(t_3)} \cdots\cdots 0 & 0 & 0 \\ \vdots & & & & \vdots \\ 0 & 0 & 0 \cdots\cdots \frac{-\lambda}{f(t_{n-2})} & \frac{1}{f(t_{n-1})} & 0 \\ 0 & 0 & 0 \cdots\cdots 0 & \frac{-\lambda}{f(t_{n-1})} & \frac{1}{f(t_n)} \end{pmatrix} ,$$

then

$$E \, (\underline{T}_1 \, \underline{U} \, \underline{U}' \, \underline{T}_1') = \sigma_\varepsilon^2 \, I_n \quad .$$

The result of applying this transformation to the original model is:

$$\frac{\sqrt{1-\lambda^2}}{f(t_1)}\, y_1 = \frac{\sqrt{1-\lambda^2}}{f(t_1)}\, \frac{1}{\alpha + \beta\, \rho^{t_1}} + \frac{\sqrt{1-\lambda^2}}{f(t_1)}\, u_1$$

and

$$\frac{y_i}{f(t_i)} - \frac{\lambda}{f(t_{i-1})}\, y_{i-1} = \frac{1}{f(t_i)}\, \frac{1}{\alpha + \beta\, \rho^{t_i}} - \frac{\lambda}{f(t_{i-1})}\, \frac{1}{\alpha + \beta\, \rho^{t_i}} + \frac{u_i}{f(t_i)} - \frac{\lambda}{f(t_{i-1})}\, u_{i-1}\,; \quad i = 2, \ldots, n. \quad (3)$$

Having formulated an autocorrelated-error model that retains the assumption of heteroscedasity, the next step was to test the data for each of the six cutoff dates. The test used was the von Neumann ratio[10]. Table 4 shows the results of this model for the cutoff dates that displayed autocorrelation. The remaining three cutoff dates did not show autocorrelated residuals.

Table 4

cut off	n	$\hat{\alpha}$	$\hat{\beta}$	$\hat{\rho}$	$\hat{\sigma}_{\hat{\alpha}}/\hat{\alpha}$ (%)	$\hat{\sigma}_{\hat{\beta}}/\hat{\beta}$ (%)	$\hat{\sigma}_{\hat{\rho}}/\hat{\rho}$ (%)	$\lim_{t \to \infty} y_i$	% of est'ed hv. wt.
all obs.	277	.0068072	.46516	.88591	3.71	29.25	1.00	146.90	96.1
10/15	255	.0065846	.43236	.88903	3.94	27.50	0.93	151.87	99.4
10/1	196	.0062570	.45860	.88753	5.41	31.02	1.07	159.82	104.6

A comparison of Tables 3 and 4 shows the estimated relative error of the estimated parameters for model (3) are larger than for model (2), as expected. However, the asymptotic values changed only slightly.

DOUBLE SAMPLING

Obtaining data for the dependent variable, dry kernel weight, involves mailing samples from the field to the laboratory and oven drying the kernels

from the samples. This is a tedious and somewhat expensive process. If a plant characteristic, which can be cheaply obtained in the field, is highly correlated with the dry kernel weight of the plant, a double sampling scheme can be designed to increase the accuracy of the dependent variable. This scheme involves obtaining data on the plant characteristic and dry kernel weight for the sample being sent to the laboratory and, in addition, collecting data cheaply on the plant characteristic in the field on a larger sample. The correlation between the plant characteristic and the dry kernel weight for the samples sent to the laboratory are used with data on the plant characteristic from the larger sample in the field to refine by means of a linear regression estimator the mean dry kernel weight per plant so that a greater precentage of the field is represented.

The form of the double sampling linear regression estimate is:

$$\bar{Y}_L = \bar{Y}_s + \hat{\beta}_1 (\bar{X}_L - \bar{X}_s)$$

provided that

$\bar{Y}_s$ = mean value for the dry kernel weight for the sample sent to the laboratory

$\bar{X}_s$ = mean value for the plant characteristic (auxiliary variable) for the sample sent to the laboratory

$\bar{X}_L$ = mean value for the plant characteristic (auxiliary variable) for samples sent to the laboratory and samples observed in the field

$\hat{\beta}_1$ = linear regression coefficient

$\bar{Y}_L$ = mean value for the dry kernel weight provided by the linear regression estimate to represent a larger sample.

The auxiliary variables observed in the field for the corn project were ear circumference, length and weight. The length of the wheat head and fertile spikelet count were obtained in the spring wheat project. Results for the spring wheat study are presented.

The data were subsetted into distinct time intervals to strengthen the correlation. The dependent variable was refined by a linear regression estimator for each time interval. Refinements were made for each auxiliary variable. These refined values and the associated time since flowering in days were fitted to the logistic growth model. Table 5 displays the results for each auxiliary variable used to more accurately provide the dependent variable.

Table 5

Auxiliary Variable	n	$\hat{\alpha}$	$\hat{\beta}$	$\hat{\rho}$	$\hat{\sigma}_{\hat{\alpha}}/\hat{\alpha}$ %	$\hat{\sigma}_{\hat{\beta}}/\hat{\beta}$ %	$\hat{\sigma}_{\hat{\rho}}/\hat{\rho}$ %	$\lim_{t_i \to \infty} y_i$
Fertile Spikelet Count	40	1.6988	31.07	.81607	7.73	56.17	4.42	.589
Head Length	40	1.7317	34.46	.80722	7.17	58.12	4.61	.577

Comparison of the relative standard errors in Tables 2 and 5 for the phenological event, flowering, indicates that the use of auxiliary variables to refine the dependent variable did not improve the performance of the model. This may be due to the fact that by subsetting the data into time intervals the sample size in each linear regression estimate is small, and therefore, large biases may exist in the estimates. In future research, the sample sizes will be much larger. Therefore, the ratio of the bias to the standard error should be reduced.

CONCLUSION

Results based on corn and spring wheat data indicate that a nonlinear model, such as the logistic growth model, may be beneficial as a forecasting methodology for these crops. Plans are to continue this research for corn and spring wheat and to test this forecasting methodology on other crops, such as winter wheat, soybeans and cotton.

FOOTNOTES

1/ Wendell W. Wilson, *Preliminary Report on the Use of Time Related Growth Models in Forecasting Components of Corn Yield*, Statistical Reporting Service, United States Department of Agriculture, Washington, D.C., 1974.

2/ Dwight A. Rockwell, *Nonlinear Estimation*, Statistical Reporting Service, United States Department of Agriculture, Washington, D.C., 1975.

3/ Jack Nealon, *Within-Year Spring Wheat Growth Models*, Statistical Reporting Service, United States Department of Agriculture, Washington, D.C., 1976.

4/ N. R. Draper and H. Smith, *Applied Regression Analysis*, New York: John Wiley & Sons, Inc., 1966, Chapter 10.

5/ S. S. Shapiro and M. B. Wilk, *Biometrika*, 52:591, 1965.

6/ A. R. Gallant, Nonlinear Regression, *The American Statistician*, Volume 29, Number 2, 1975, pages 73 - 81.

7/ *Biomedical Computer Programs*, Department of Biomathematics, School of Medicine, University of California Press, 1973.

8/ H. Glejser, A New Test for Heteroscedasity, *Journal of the American Statistical Association*, Volume 64, 1969, pages 316 - 323.

9/ J. Johnston, *Econometric Methods*, New York: McGraw-Hill, 1963, pages 244 - 246.

10/ Ibid.