United States
Department of
Agriculture

Statistical
Reporting
Service

Statistical
Research
Division

SRS Staff Report
Number AGES 840827

September 1984

Forecasting with Plant Process Models

An Introduction to a Time Series Approach

Keith N. Crank

FORECASTING WITH PLANT PROCESS MODELS: AN INTRODUCTION TO A TIME SERIES APPROACH, by Keith N. Crank, Statistical Research Division, Statistical Reporting Service, U. S. Department of Agriculture. Staff Report No. AGES840827

ABSTRACT

This paper presents a method of using data from a plant process model and nonlinear regression techniques to forecast end of season values for various plant components. The use of time series methods are also discussed. Five models are presented with a theoretical justification.

Keywords: plant process model, time series, nonlinear model, forecasting

CONTENTS

SUMMARY

This paper presents an approach and develops the statistical background for using plant process models (PPM's) as a component in yield forecasting models. Although there are other problems which must be solved before PPM's can be used in an operational program, the problem of how to use a PPM to forecast is an important one.

The method suggested in this paper uses the early season data from a plant process model to fit the nonlinear logistic model, the Gompertz model or their time series representations. One of the parameter estimates from each model will then provide the forecast of plant component values at the end of the growing season. These can then be used to obtain a forecast of yield. Analysis and summary results will be presented in a later paper.

INTRODUCTION

The purpose of this paper is to introduce a method of combining nonlinear regression techniques with data from a plant process model(PPM) to forecast plant yield components. The models obtained can be used to forecast yield directly if sufficient data is available at the time of the forecast. However, this will often not be the case. When yield data is not available, final values can be forecast for other plant components, which can in turn be used for forecasting yield. For the purposes of this paper, though, it will be assumed that the fruit weight of the plant is the component being forecast. Motivation and development of these models are given in this paper. Analysis and results of this method will be presented in a later paper.

The Statistical Reporting Service forecasts crop yields using regression parameters based on data collected in previous years. Currently data from three or five years is used to estimate the parameters. This means that the forecast for the current year is extremely dependent on the weather patterns of those previous years. If more years of data were used to estimate the parameters, it is possible that the year to year differences in weather could be incorporated into the regression parameters. Unfortunately, technological changes in crop production could invalidate regression estimates which use many years of data unless these technological changes were also modeled.

An alternative to using so many years of complex data is to develop a model which only uses data from the current year. One such model is the logistic growth function. This function has the form

$$\text{WEIGHT}(t) = \frac{\alpha}{1+\beta\rho^{t}}$$

Previous research has dealt with fitting this model to field data using the Marquardt nonlinear regression technique. ([2],[3],[4],[5],[6],[7],[8],[9],[10],[13])

Although much research went into this model, it has not been used in the operating program. Two major statistical reasons contributed to this. The first reason is the problem of unequal residual variances (heteroscedasticity). In the theory of least squares an assumption is made that the error

terms of the predicted values have a constant variance. In linear regression the parameter estimates remain unbiased even if this assumption is not valid. However, the estimates are not necessarily minimum variance estimates (that is, there may be better estimates in this case [11, pp. 144-146]). In nonlinear regression the effect of unequal residual variances is not as clear. Since there is no exact solution for the parameters of a nonlinear regression, it is not possible to directly evaluate the effect of unequal variances. Research was conducted on methods of reducing or eliminating this unequal variance problem. However, little effort was spent in determining whether or not it was a significant problem.

The second reason for not using the logistic growth model in the operating program is the problem of convergence. In linear regression convergence is not a problem, since there is a closed form solution for the parameter estimates which requires merely that two data points be provided. For the logistic growth model it is not clear how many points are needed to guarantee that exactly one model can be fit through that set of points. In some cases three points may be sufficient to identify a unique logistic model. However, in other cases, three points may determine many logistic models or possibly none.

Most of the papers dealing with the logistic growth model considered the problem of convergence. However, this consideration was usually limited to determining whether or not convergence occurred in time to make early season forecasts. Only one paper ([6]) tried to impose conditions which would make the model converge. This was done by forcing the derivative to change sign at a prespecified point. The effect was to eliminate the parameter which was hardest to estimate. This approach seemed to be an improvement but no further research was done.

As work on the logistic model decreased, research in the study of plant process models (PPM's) expanded. Plant process models are complex models from which computer programs can be written which "grow" a plant in a computer. These models require initial values for such items as variety characteristics, planting date, soil characteristics, and soil water, as well as daily values for precipitation, maximum and minimum temperature, and solar radiation. Outputs from the programs are daily values describing the size of the plant's component parts and its stage of growth. In addition, the final size of each of the components, as well as the yield, is available at the end of the season.

The purpose of this paper is to present an approach to using PPM's for forecasting. The idea behind the paper is that since data is available daily from a PPM, this data may be useful in fitting a nonlinear growth model. The next section examines some possible reasons for the statistical problems

stated previously. Then the following section describes other models and explain how these models may help to eliminate or reduce those problems. The final section describes the analysis which is to be performed and which will appear in a later paper.

THEORY

In applying growth models it is assumed that each plant grows according to a unique true model with certain estimable parameters. However, in sampling to estimate the model parameters more than one plant was sampled on each time visit and destructive measurements were used. Therefore, the parameters estimated were for the model of an average of plants, not an individual plant. What is desired, though, is an average of the model estimates for each individual plant. In a linear regression this is not a problem since the average of two lines is the line through the average of the points. However, for a nonlinear model this is not the case. For nonlinear models the average for a given functional form does not have the same functional form. In previous research the measurements obtained to fit the logistic model were of necessity destructive measurements. Therefore, each of the data points used to fit the model came from a different plant, and each of the points represented one of many models. However, only one model was estimated.

Both of the statistical problems encountered in previous research can be attributed in part to the way the model parameters were estimated. Since each of the points used in estimation may represent a different set of model parameters, it is not surprising that an average model fit to all of the points did not always converge. Furthermore, it is easy to see from figure 1 that heteroscedasticity is to be expected. Since all of the model functions start out near zero, there cannot be much difference in their early values. However, the later values can vary widely.

If it were possible to obtain data from a single plant without affecting the growth of the plant and use that data to fit the model, convergence might be obtained at an earlier date. However, it is not clear whether even this convergence would be early enough to provide early season forecasts. That problem will be discussed in a later paper after the analysis has been completed.

It also seems clear that the problem of unequal residual variances could be reduced by using data from a single plant. However, there may be an inherent problem with unequal residual variances in the model. It seems logical for the plant to be able to deviate more from its expected value when it is large than when it is small. Since the growth function is nonnegative and increasing, these larger deviations would occur later in plant development. Thus there may be a problem with unequal residual variances even when a single plant is

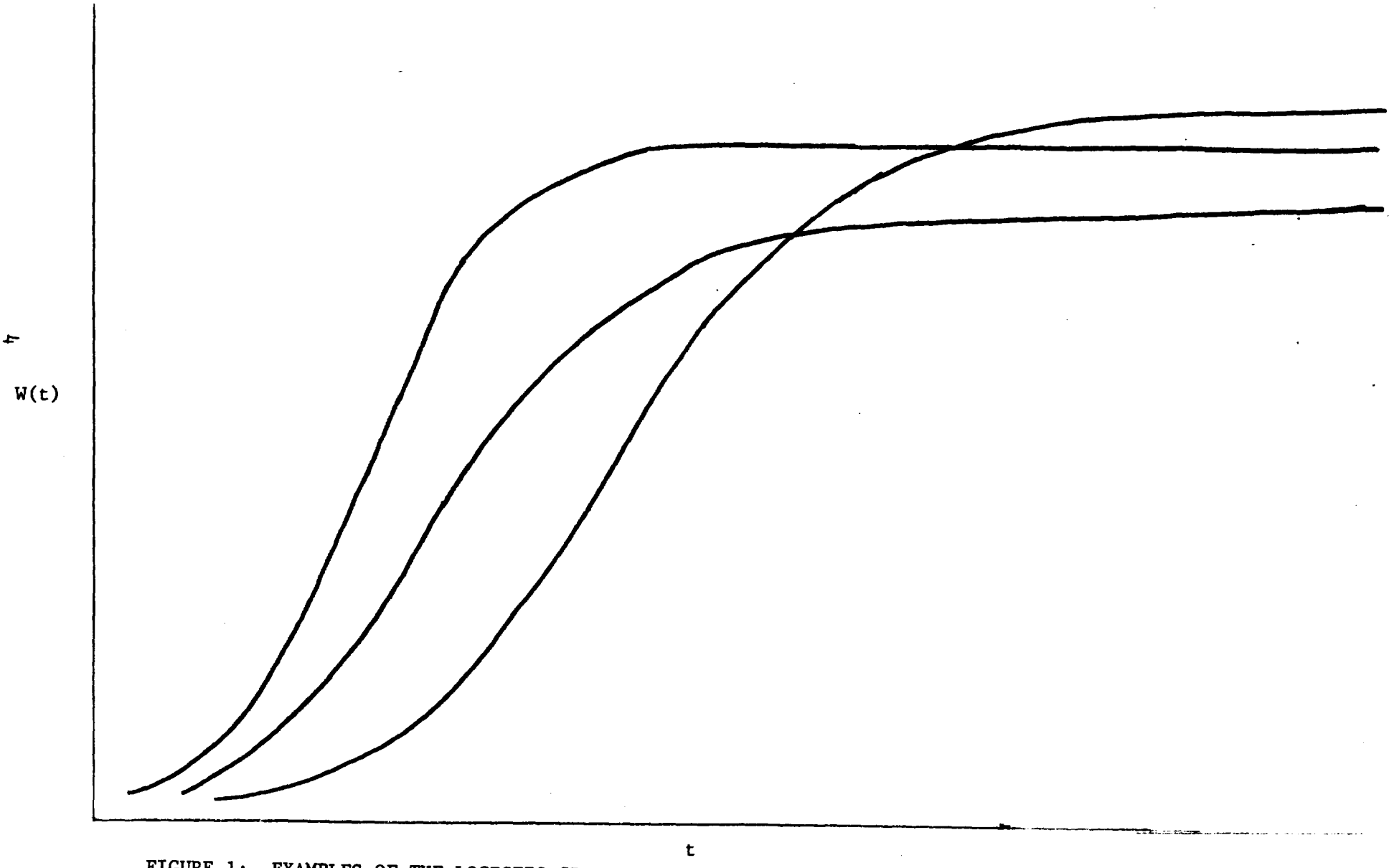

FIGURE 1: EXAMPLES OF THE LOGISTIC GROWTH FUNCTION

used. This problem should not be as large as in previous research, since between plant variability has been eliminated.

Fitting a growth model using data from a single plant could provide many benefits. But it has its drawbacks as well. The biggest problem is with autocorrelation of the residuals. This means that deviations of the true value from the model value will be related for time periods which are close together. For example, if a plant's size is below what is expected on a given day, the size on the following day will likely be below its expected value also. In order to reach its expected value on the second day, the plant must not only make up the previous day's deficit, but must also account for its expected growth for that day. Thus the errors will not be uncorrelated. If the errors are not uncorrelated, then the parameter estimates may not be consistent (that is, no matter how much data is available the estimates may not converge to their true value [12]).

It would be nice if instead of comparing the plant's size to some fixed value, it could be compared in some way to its previous day's value. If the plant's expected growth for a given day were dependent on its size at the beginning of the day, then a model based on the size of the plant on the previous day would have uncorrelated errors. Such a model would be a time series model. In a time series model each value of plant size can depend on the previous day's value. However, such a model would not necessarily be linear. Thus the techniques developed in most time series courses may not be directly applicable.

The discussion of using data from a single plant assumed that such data could be obtained without influencing the future growth of the plant. In general such data is not available. In order to obtain measurements on a growing plant it is usually necessary to destroy the plant or at least parts of the plant. However, by assuming that the plant process models can reproduce the growth of a representative plant, data can be obtained for what can be considered to be a single plant. In addition this data can be obtained without affecting the future growth of the plant.

MODEL

Since plant process models (PPM's) produce daily size values for all plant components, these PPM's can be used in conjunction with within year growth models to forecast the values of those components at the end of the season. This can be done by fitting a nonlinear regression model to the daily plant data. The model with which the agency has the most experience is the logistic growth model. The form of this model is

$$W(t) = \frac{\alpha}{1+\beta\rho^{t}}$$

where W is some measure of the weight of the fruit at time t, t is time (usually measured as number of days from some observable phenological event), and α, β, and ρ are parameters to be estimated. (See Figure 2.)

Another model which is similar is the Gompertz model. This has the form

$$W(t) = \alpha\beta^{\rho^t}$$

where all of the variables are the same as in the logistic model (though the parameters may have different interpretations). (See Figure 3.)

Both of these models can be rewritten as a time series, that is, W(t) (weight at time t) is written as a function of W(t-1) (the previous day's weight) instead of as a function of t. This is done by first solving the equation for t. For the logistic model we have the following steps:

$$W(t) = \frac{\alpha}{1+\beta\rho^t}.$$

Multiplying on both sides by $1+\beta\rho^t$ we get

$$(1+\beta\rho^t)W(t) = \alpha.$$

Expanding the left hand side

$$W(t)+(\beta\rho^t)W(t) = \alpha,$$

subtracting W(t)

$$(\beta\rho^t)W(t) = \alpha-W(t),$$

dividing by $\beta W(t)$

$$\rho^t = \frac{\alpha-W(t)}{\beta W(t)},$$

and then taking logs we get

$$t(\ln(\rho)) = \ln\left[\frac{\alpha-W(t)}{\beta W(t)}\right].$$

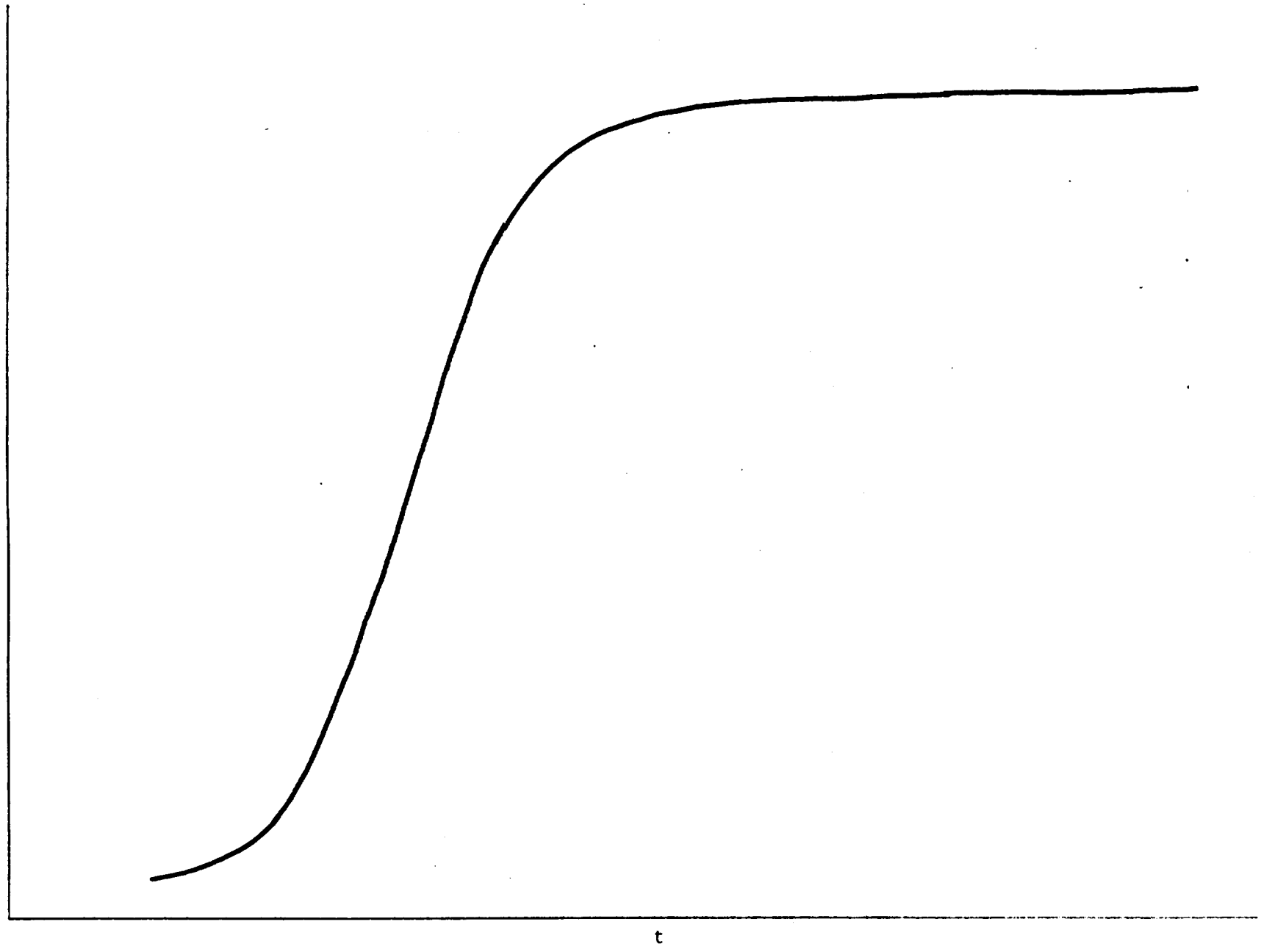

FIGURE 2: LOGISTIC FUNCTION

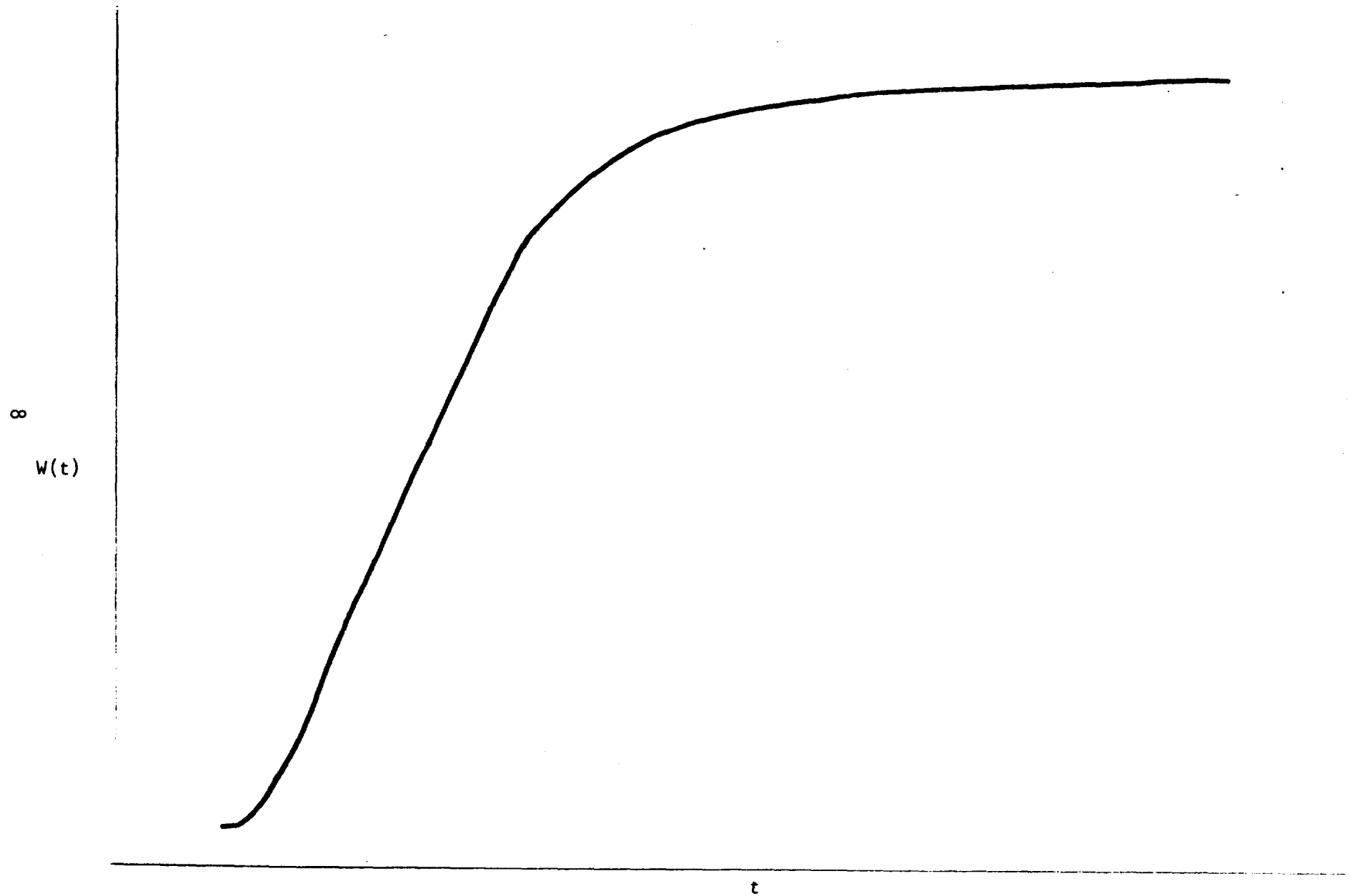

FIGURE 3: GOMPERTZ FUNCTION

We then divide by $\ln(\rho)$ to get

$$t = \frac{1}{\ln(\rho)} \ln\left[\frac{\alpha - W(t)}{\beta W(t)}\right].$$

Then we substitute (t-1) for t

$$(t-1) = \frac{1}{\ln(\rho)} \ln\left[\frac{\alpha - W(t-1)}{\beta W(t-1)}\right].$$

This is then substituted back into our original equation:

$$W(t) = \frac{\alpha}{1+\beta\rho^{t}}$$

$$= \frac{\alpha}{1+\beta\rho^{(1+t-1)}}$$

$$= \frac{\alpha}{1+\beta\rho\rho^{(t-1)}}$$

$$= \frac{\alpha}{1+\beta\rho\rho^{\left[\frac{1}{\ln(\rho)} \ln\left[\frac{\alpha - W(t-1)}{\beta W(t-1)}\right]\right]}}$$

$$= \frac{\alpha}{1+\beta\rho e^{\left[\frac{\ln(\rho)}{\ln(\rho)} \ln\left[\frac{\alpha - W(t-1)}{\beta W(t-1)}\right]\right]}}$$

(since $\rho = e^{\ln(\rho)}$)

$$= \frac{\alpha}{1+\beta\rho e^{\ln\left[\frac{\alpha - W(t-1)}{\beta W(t-1)}\right]}}$$

$$= \frac{\alpha}{1+\beta\rho\left[\frac{\alpha - W(t-1)}{\beta W(t-1)}\right]}$$

$$= \frac{\alpha}{1+\left[\frac{\rho(\alpha - W(t-1))}{W(t-1)}\right]}$$

$$= \frac{\alpha W(t-1)}{W(t-1)+\rho\alpha-\rho W(t-1)}$$

$$= \frac{\alpha W(t-1)}{(1-\rho)W(t-1)+\rho\alpha}$$

(See Figure 4.)

The same procedure is used to derive a time series for the Gompertz model:

$$W(t) = \alpha\beta^{\rho^t}$$

$$\beta^{\rho^t} = \frac{W(t)}{\alpha}$$

$$\rho^t \ln(\beta) = \ln\left[\frac{W(t)}{\alpha}\right]$$

$$\rho^t = \frac{1}{\ln(\beta)}\ln\left[\frac{W(t)}{\alpha}\right]$$

$$t(\ln(\rho)) = \ln\left[\frac{1}{\ln(\beta)}\ln\left[\frac{W(t)}{\alpha}\right]\right]$$

$$t = \frac{1}{\ln(\rho)}\ln\left[\frac{1}{\ln(\beta)}\ln\left[\frac{W(t)}{\alpha}\right]\right]$$

Then

$$(t-1) = \frac{1}{\ln(\rho)}\ln\left[\frac{1}{\ln(\beta)}\ln\left[\frac{W(t-1)}{\alpha}\right]\right],$$

and

$$W(t) = \alpha\beta^{\rho^t}$$

$$= \alpha\beta^{\rho\rho^{(t-1)}}$$

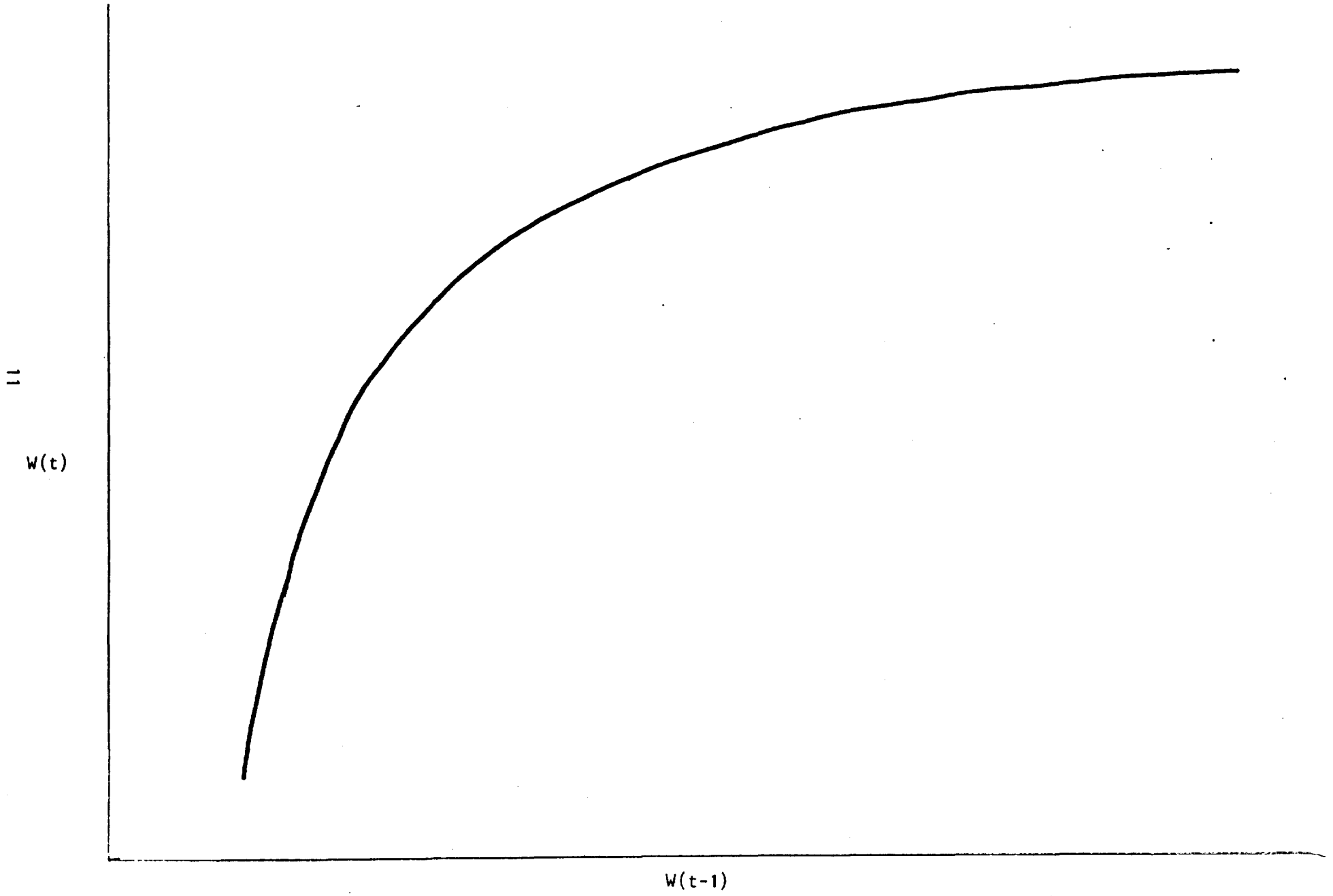

FIGURE 4: LOGISTIC TIME SERIES

$$= \alpha\beta^{\rho\rho^{\left[\frac{1}{\ln(\rho)}\ln\left[\frac{1}{\ln(\beta)}\ln\left[\frac{W(t-1)}{\alpha}\right]\right]\right]}}$$

$$= \alpha\beta^{\rho e^{\ln\left[\frac{1}{\ln(\beta)}\ln\left[\frac{W(t-1)}{\alpha}\right]\right]}}$$

$$= \alpha\beta^{\rho\left[\frac{1}{\ln(\beta)}\ln\left[\frac{W(t-1)}{\alpha}\right]\right]}$$

$$= \alpha e^{\left[\frac{\rho\ln(\beta)}{\ln(\beta)}\ln\left[\frac{W(t-1)}{\alpha}\right]\right]}$$

$$= \alpha e^{\left[\rho\ln\left[\frac{W(t-1)}{\alpha}\right]\right]}$$

$$= \alpha\left[e^{\ln\left[\frac{W(t-1)}{\alpha}\right]}\right]^{\rho}$$

$$= \alpha\left[\frac{W(t-1)}{\alpha}\right]^{\rho}$$

(See Figure 5.)

Since the theory of linear regression is much better developed than the theory of nonlinear regression, it is often useful to make a transformation which will result in a linear model. Many times this is not possible and in such cases the model is called intrinsically nonlinear. When such a transformation is possible, the equations are called intrinsically linear.([1]) Of the four models mentioned, the first three are all intrinsically nonlinear. However, the last one (the Gompertz time series) is intrinsically linear. Taking logrithms of both sides we obtain

$$\ln(W(t)) = \ln(\alpha)+\rho(\ln(W(t-1))-\ln(\alpha))$$

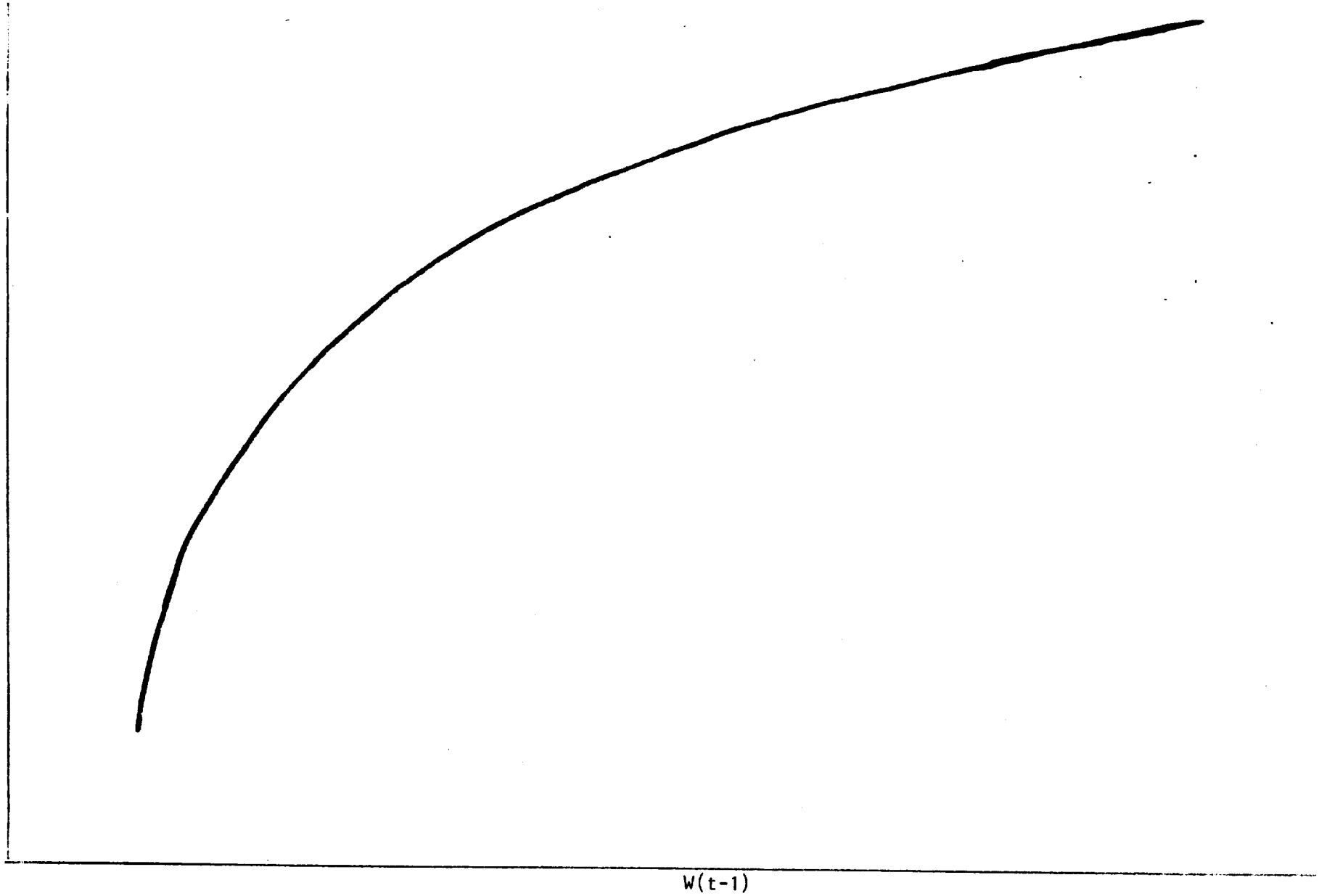

FIGURE 5: GOMPERTZ TIME SERIES

Letting $L(t)=\ln(W(t))$, $d=\ln(\alpha)-\rho\ln(\alpha)$, and $c=\rho$, we get the linear equation

$$L(t) = d+cL(t-1).$$

Of the parameters to be estimated, the parameter α is of special importance (See the Appendix). This parameter is the upper asymptote of the curve and the estimate of fruit weight at the end of the growing season. Hence, if the model can be fit to the data before the season is over, the estimated value of α can be used as a forecast of the final value of fruit weight. This is true for both the logistic model and the Gompertz model. In the nonlinear regressions, the parameter α is estimated directly. However, in the linear regression $d=(1-\rho)\ln(\alpha)$ is the estimated parameter. We can obtain an estimate of α as $\alpha=\exp(\frac{d}{1-c})$, but if c is close to one, this may not be a good estimator. In addition this estimator is not unbiased, and it is difficult to estimate its variance.

We now have five models which can be compared. As described in the Theory section, items of importance in this comparison are

1) How early in the season can we obtain good parameter estimates (convergence)?

2) Is the assumption of equal residual variances satisfied?

3) Are the errors uncorrelated?

These questions will be studied and the results presented in a later paper. In addition the models will be compared in terms of their ability to accurately forecast end of season values for items of interest.

ANALYSIS FOR SECOND PAPER

Twenty-two datasets are available for analysis. Variables of interest from these datasets will be fit to the five models described in the previous section for various forecasting dates as well as for an entire growing season. The models will be compared to find which ones are capable of providing accurate early season forecasts of final PPM results.

The steps involved in the analysis are as follows:

1) For selected variables in each dataset, the parameters for each of the five models will be estimated using data for the entire growing season.

2) For each model the residuals will be tested for autocorrelation.

3) If autocorrelation exists, a modification will be made which will reduce or eliminate the autocorrelation, while still allowing the parameters to be used in forecasting. Parameters will be estimated for this new model. (The method of modification will be described in the next paper).

4) The models will be compared. Steps 1-3 will be performed again using data that would have been available at a forecasting date. All of the models will compete unless one proves significantly inferior to the rest. The comparisons to be made include looking at the mean squared errors of the models and the amount of autocorrelation in the residuals of the models. Since the purpose of these models is for forecasting end of season values, no comparisons will be made as to how well the models fit the data early in the season.

5) Step 4 will be repeated at an earlier forecasting date if there is some hope that convergence is possible.

A later paper will present the results of the analysis and recommendations as to whether any further work should be done in this area.

REFERENCES

1. Draper, Norman, and Harry Smith, Applied Regression Analysis, New York: John Wiley & Sons, Inc., 1966.

2. House, Carol C., Forecasting Corn Yields: A Comparison Study Using 1977 Missouri Data, U. S. Department of Agriculture, Economics, Statistics and Cooperatives Service, June 1979.

3. Larsen, Greg A., Alternative Methods of Adjusting for Heteroscedasticity in Wheat Growth Data, U. S. Department of Agriculture, Economics, Statistics and Cooperatives Service, February 1978.

4. _____, Forecasting 1977 Winter Wheat Growth, U. S. Department of Agriculture, Economics, Statistics and Cooperatives Service, August 1978.

5. _____, 1978 Kansas Winter Wheat Yield Estimation and Modeling, U. S. Department of Agriculture, Economics, Statistics and Cooperatives Service, September 1979.

6. _____, Forecasting Final Corn Grain Yield Per Plant with a Constrained Logistic Growth Model, U. S. Department of Agriculture, Economics, Statistics and Cooperatives Service, January 1980.

7. Nealon, Jack, Within-Year Spring Wheat Growth Models, U. S. Department of Agriculture, Statistical Reporting Service, January 1976.

8. _____, The Development of Within-Year Forecasting Models for Winter Wheat, U. S. Department of Agriculture, Statistical Reporting Service, October 1976.

9. _____, The Development of Within-Year Forecasting Models for Spring Wheat, U. S. Department of Agriculture, Statistical Reporting Service, November 1976.

10. Rockwell, Dwight A., Nonlinear Estimation, U. S. Department of Agriculture, Statistical Reporting Service, April 1975.

11. Seber, G. A. F., Linear Regression Analysis, New York, John Wiley & Sons, Inc., 1977.

12. Tsay, Ruey S.,"Regression Models with Time Series Errors," JASA, 1984, vol. 79, pp. 118-124.

13. Wilson, Wendell W., Preliminary Report on the Use of Time Related Growth Models in Forecasting Components of Corn Yield, U. S. Department of Agriculture, Statistical Reporting Service, May 1974.

APPENDIX

The purpose of fitting a nonlinear model to the data from a plant process model is to provide forecasts of yield components. Thus it is necessary to have some way of producing a forecast from the parameters estimated in the regression. This Appendix explains how this can be done and provides a mathematical justification.

For the Logistic model we have the following form

$$W(t) = \frac{\alpha}{1+\beta\rho^t} + e_t$$

Under the usual assumption that the errors have mean zero, we obtain the following

$$E(W(t)) = E\left[\frac{\alpha}{1+\beta\rho^t} + e_t\right] = \frac{\alpha}{1+\beta\rho^t}$$

If we let t increase to ∞, we find that the limiting value of $E(W(t))$ is α (provided $0<\rho<1$).

Similarly, the form for the Gompertz model is

$$W(t) = \alpha\beta^{\rho^t} + e_t$$

Taking expected values, we obtain

$$E(W(t)) = E\left[\alpha\beta^{\rho^t} + e_t\right] = \alpha\beta^{\rho^t}$$

Once again the limiting value of $E(W(t))$ as $t \to \infty$ is α provided $0<\rho<1$.

For the time series models the problem is a little more complicated. First, let's consider the Logistic Time Series

$$W(t) = \frac{\alpha W(t-1)}{(1-\rho)W(t-1)+\rho\alpha} + e_t$$

If we take expected values of both sides we obtain

$$E(W(t)) = E\left[\frac{\alpha W(t-1)}{(1-\rho)W(t-1)+\rho\alpha} + e_t\right] = E\left[\frac{\alpha W(t-1)}{(1-\rho)W(t-1)+\rho\alpha}\right]$$

Unfortunately, there is no simple way to evaluate the right hand side, since there is a random variable in both the numerator and the denominator of the expression. However, we will avoid this problem by making assumptions which seem realistic. These assumptions will allow us to look at the limit of the process itself instead of at the limit of its expected value. The first assumption is that the limit of W(t) exists. This is a reasonable assumption when W(t) represents the weight of the grain from a PPM and this reaches some fixed value at the end of the season. Thus, we have

$$\lim_{t \to \infty} W(t) = \lim_{t \to \infty} \left[\frac{\alpha W(t-1)}{(1-\rho)W(t-1)+\rho\alpha} + e_t \right]$$

The second assumption we make is on the errors:

$$\lim_{t \to \infty} e_t = 0.$$

Then (letting $\lim W(t) = \lim W(t-1) = x$)

$$x = \frac{\alpha x}{(1-\rho)x+\rho\alpha}$$

or

$$(1-\rho)x = (1-\rho)\alpha.$$

Therefore,

$$\lim_{t \to \infty} W(t) = x = \alpha$$

(since $\rho \neq 1$).

The same argument works for the Gompertz Time Series Model. The last model is just a simple transformation of the Gompertz Time Series Model. Thus the parameter of interest in all cases is α. (It should be noted that the assumptions made for the time series models are more realistic than the assumption of constant variance made for the first two models. However, in what is to follow it should be clear that the same procedure will be used for estimation in either case.)

When estimating parameters in a regression problem, it is common to use the method of least squares. Generally this requires the assumption that the errors have mean zero and constant variance. Unfortunately, for three of our models we are actually assuming that the variance becomes zero for large t. (This was a necessary assumption to have the parameter α represent the final value of the plant component.) However, since we are going to be estimating the parameters before maturity, we will assume that the variances over the range of our data are approximately constant. (Actually we will be assuming that over the range of our data, the unequal variances will not produce a bias in our estimate of α).

Based on the preceding discussion, we will be estimating the parameters of the regressions using the method of least squares. The parameter which will be used for forecasting is α in the nonlinear models, and we will assume that our estimates of it are unbiased.