2mif

NASA TECHNICAL NOTE

NASA TN D-7516

NASA TN D-7516

A MONTE CARLO INVESTIGATION OF EXPERIMENTAL DATA REQUIREMENTS FOR FITTING POLYNOMIAL FUNCTIONS

by George C. Canavos

Langley Research Center
Hampton, Va. 23665

NATIONAL AERONAUTICS AND SPACE ADMINISTRATION • WASHINGTON, D. C. • JUNE 1974

1. Report No. NASA TN D-7516	2. Government Accession No.	3. Recipient's Catalog No.
4. Title and Subtitle A MONTE CARLO INVESTIGATION OF EXPERIMENTAL DATA REQUIREMENTS FOR FITTING POLYNOMIAL FUNCTIONS		5. Report Date June 1974 6. Performing Organization Code
7. Author(s) George C. Canavos		8. Performing Organization Report No. L-9127
9. Performing Organization Name and Address NASA Langley Research Center Hampton, Va. 23665		10. Work Unit No. 501-06-01-08 11. Contract or Grant No.
12. Sponsoring Agency Name and Address National Aeronautics and Space Administration Washington, D.C. 20546		13. Type of Report and Period Covered Technical Note 14. Sponsoring Agency Code

15. Supplementary Notes

16. Abstract

This report examines the extent to which sample size affects the accuracy of a low-order polynomial approximation of an experimentally observed quantity and establishes a trend toward improvement in the accuracy of the approximation as a function of sample size. The task is made possible through a simulated analysis carried out by the Monte Carlo method, in which data are generated by using several transcendental or algebraic functions as models. Contaminated data of varying amounts are fitted to linear quadratic or cubic polynomials, and the behavior of the mean-squared error of the residual variance is determined as a function of sample size. Results indicate that the effect of the size of the sample is significant only for relatively small sample sizes and diminishes drastically for moderate and large amounts of experimental data.

17. Key Words (Suggested by Author(s)) Curve fitting Polynomial functions Sample size		18. Distribution Statement Unclassified – Unlimited STAR Category 19	
19. Security Classif. (of this report) Unclassified	20. Security Classif. (of this page) Unclassified	21. No. of Pages ~~21~~ 24	22. Price* $3.00

*For sale by the National Technical Information Service, Springfield, Virginia 22151

A MONTE CARLO INVESTIGATION OF EXPERIMENTAL DATA REQUIREMENTS FOR FITTING POLYNOMIAL FUNCTIONS

By George C. Canavos
Langley Research Center

SUMMARY

This report examines the extent to which sample size affects the accuracy of a low-order polynomial approximation of an experimentally observed quantity and establishes a trend toward improvement in the accuracy of the approximation as a function of sample size. The task is made possible through a simulated analysis carried out by the Monte Carlo method, in which data are generated by using several transcendental or algebraic functions as models. Contaminated data of varying amounts are fitted to linear quadratic or cubic polynomials, and the behavior of the mean-squared error of the residual variance is determined as a function of sample size. Results indicate that the effect of the size of the sample is significant only for relatively small sample sizes and diminishes drastically for moderate and large amounts of experimental data.

INTRODUCTION

The purpose of this report is to investigate by Monte Carlo simulation the effect that the number of experimental data points has on smoothing out the influence of random error in an analytic-function approximation of an experimentally observed quantity.

In an environment in which experimentation is the only source of information, it is often desired to determine and quantify the effect that some controlled variable exerts on a measured quantity which is nearly always subjected to random contamination. For many such cases, the true functional relationship is too vaguely known to be of practical use. Thus, some simple analytic function, such as a polynomial of relatively low degree, is used to approximate the behavior of the dependent variable within a prescribed range of the controlled variable. To determine the polynomial approximation, a reasonable amount of test data is needed to smooth out the effect of random error to some nominal value. However, in many instances the collection of laboratory data is becoming increasingly more difficult for reasons such as cost and complexity of test equipment. Consequently, it is advisable to plan carefully the collection of experimental data to enhance the relevancy of each data point while holding down its cost. Nevertheless, it is conceivable that economic restrictions on the sample size may compromise the accuracy of the approximation to an unacceptable level. Therefore, the purpose of this report is to shed some light on the problem of how the amount of test data affects the accuracy of a low-order polynomial approximation of a stochastic quantity and to establish, at least for typical cases, a trend toward improvement in the accuracy as a function of sample size. In addition, the report summarizes some existing techniques on how the observation points should be spaced within the range of the controlled variable to improve the predictive capability of the approximating function.

To determine the extent to which the sample size affects the accuracy of an approximating polynomial function, a simulated analysis is carried out by appealing to Monte Carlo procedures (ref. 1). In order to include a practical range of possibilities, data are generated first by using one of several polynomial or transcendental functions as models and then adding random errors generated from a Gaussian distribution. However, in all cases, the contaminated data are fitted to either linear, quadratic, or cubic polynomial functions. It is believed that these low-order polynomial functions are the most plausible to approximate the behavior of an experimentally observed quantity when compared with, say, high-order polynomials, which may fit the random error more than approximate the variable quantity.

SYMBOLS

E	expectation operator
m	degree of polynomial function
n	number of measurements
x	controlled variable
X	matrix of controlled-variable values
y	measured variable
$\underline{y}$	vector of measurements
$\underline{\beta}$	vector of unknown coefficients
$\underline{\epsilon}$	vector of random errors

σ^2	error variance

Subscripts:

$\hat{}$	estimate
T	transpose of matrix

An underlined symbol denotes a vector.

REVIEW OF FUNDAMENTAL CONCEPTS

Let the unknown functional relation between an observable quantity y and a controlled variable x be approximated by a polynomial of degree m in which the jth observation is depicted by

$$y_j = \beta_0 + \beta_1 x_j + \beta_2 x_j^2 + \ldots + \beta_m x_j^m + \epsilon_j \qquad (j = 1,2,\ldots,n) \tag{1}$$

where $\beta_0, \beta_1, \ldots, \beta_m$ are the unknown coefficients of the polynomial, ϵ_j is the random error associated with the observation y_j, and n is the number of laboratory measurements used to fit the polynomial. In vector form, this set of n equations in $m + 1$ unknown coefficients is written as

$$\underline{y} = X\underline{\beta} + \underline{\epsilon} \tag{2}$$

where

$$\underline{y} = \begin{bmatrix} y_1 \\ y_2 \\ \cdot \\ \cdot \\ \cdot \\ y_n \end{bmatrix} \qquad X = \begin{bmatrix} 1 & x_1 & x_1^2 & \ldots & x_1^m \\ 1 & x_2 & x_2^2 & \ldots & x_2^m \\ \cdot & & & & \\ \cdot & & & & \\ \cdot & & & & \\ 1 & x_n & x_n^2 & \ldots & x_n^m \end{bmatrix} \qquad \underline{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \cdot \\ \cdot \\ \cdot \\ \beta_m \end{bmatrix} \qquad \underline{\epsilon} = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \cdot \\ \cdot \\ \cdot \\ \epsilon_n \end{bmatrix}$$

If $E(\underline{\epsilon}) = \underline{0}$ and $E\left(\underline{\epsilon}\underline{\epsilon}^T\right) = \sigma^2 I$, where σ^2 is the error variance and I the appropriate identity matrix, then the least-squares estimates (refs. 2 and 3) of the components of $\underline{\beta}$ are determined by the result

$$\hat{\underline{\beta}} = \left(X^T X\right)^{-1} X^T \underline{y} \tag{3}$$

where $\hat{\underline{\beta}}$ denotes the vector of estimates.

The quality of the estimates is measured by the variance-covariance matrix of $\hat{\underline{\beta}}$ given by (ref. 2)

$$\operatorname{var}\left(\hat{\underline{\beta}}\right) = \sigma^2\left(X^T X\right)^{-1}$$

The (i,i) element of $\sigma^2\left(X^T X\right)^{-1}$ is the variance of $\hat{\beta}_i$, while the (i,j) element corresponds to the covariance between $\hat{\beta}_i$ and $\hat{\beta}_j$ for $i \neq j$. Since variance is a measure of dispersion, then the smaller the variance of a component of $\hat{\underline{\beta}}$, the better the estimate of that component. In real-world situations, however, the error variance σ^2 is not likely to be known. Therefore, to compute $\operatorname{var}\left(\hat{\underline{\beta}}\right)$, an estimate of σ^2 must be determined. This is usually the unbiased estimate (ref. 2)

$$\hat{\sigma}^2 = \frac{\underline{y}^T \underline{y} - \hat{\underline{\beta}}^T X^T \underline{y}}{n - (m + 1)} \tag{4}$$

which is nothing more than the sum of squares of the residuals divided by the number of data points less the number of estimated coefficients. Thus, the estimate $\hat{\sigma}^2$ is usually referred to as the residual variance. Whereas the error variance σ^2 measures the magnitude of the random error, the residual variance $\hat{\sigma}^2$ measures the magnitude of the degree to which the fitted equation fails to describe the change in the dependent variable. The residual variance is an unbiased estimate of σ^2 only if there is no model error; otherwise $\hat{\sigma}^2$ reflects both random variation and model error. For example, if the true model is an exponential type while the approximating function is a polynomial, then $\hat{\sigma}^2$ accounts for the error due to inherent differences between the true and fitted functions as well as for pure random error.

Of major interest in an analytic-function approximation of a variable quantity is the ability to predict that quantity without a laboratory observation. Thus, let x_p be a point of prediction. The predicted value $\hat{y}_p$ corresponding to x_p, from equation (1), is

$$\hat{y}_p = \underline{x}_p{}^T \hat{\underline{\beta}}$$

where

$$\underline{x}_p^T = \left(1, x_p, x_p^2, \ldots, x_p^m\right)$$

From matrix algebra, the variance of $\hat{y}_p$ is

$$\begin{aligned} \mathrm{var}\left(\hat{y}_p\right) &= \underline{x}_p^T\left(X^TX\right)^{-1} \underline{x}_p \hat{\sigma}^2 \\ &= \underline{x}_p^T \mathrm{var}\left(\hat{\underline{\beta}}\right)\underline{x}_p \end{aligned}$$

Therefore, the quality of $\hat{y}_p$ is directly proportional to the quality of the least-squares estimates of the polynomial coefficients. Moreover, $\mathrm{var}\left(\hat{y}_p\right)$ is a function of the residual variance $\hat{\sigma}^2$. If, for example, the residual variance is zero, the predicted and observed values will coincide and the fitted polynomial function will model the observed data without error. On the other hand, an excessively large residual variance will result in a poor prediction capability.

As stated earlier, the motivating force in an analytic-function approximation of a stochastic quantity is to predict the quantity without an actual measurement. Thus, it is imperative that the data-gathering procedure be carefully planned to control the size of the error between a laboratory measurement and the corresponding predicted value. In fact, how the observation points are spaced is related to the error of a predicted value. If, for example, some polynomial of unknown degree is to be tried as the approximating function, the optimal spacing of observation points is a uniform distribution throughout the selected range of the controlled variable (ref. 4). Such a spacing increases the likelihood of detection of an unusual behavior while holding down the size of the error. Alternatively, if the degree of the approximating polynomial is known, spacing such as that considered by De la Garza (ref. 5) is preferred. Some optimal-spacing techniques in curve fitting are summarized briefly in the appendix.

MONTE CARLO SIMULATION

In every Monte Carlo simulation, it is mandatory to specify completely the process to be simulated and to identify quantities of interest (ref. 1). It is therefore necessary to present a detailed discussion for implementing the simulation procedure.

The objective is to fit a varying number of contaminated data points to one of three models (viz., linear, quadratic, or cubic polynomial) and determine the behavior of the random error as a function of the number of data points. Establishing the effect that sample size has on smoothing out the influence of random error in a polynomial-function approximation is essentially the same as determining the stability of the residual variance as a function of sample size. A quantity which measures the stability of any estimator is the mean-squared error of the estimator. The mean-squared error depicts the average squared difference between the estimator and the quantity it is estimating. Since the error variance must be an input to the simulation, it is possible to determine the squared difference between $\hat{\sigma}^2$ and σ^2, given a model and a set of data points. By simulating and repeating such a procedure many times, the squared differences are accumulated and the mean-squared error of the residual variance is determined.

Data are generated by using polynomial and transcendental functions as models and are uniformly distributed throughout the range of x for the following 11 values of n: 5, 11, 21, 31, . . ., 101. The reason that odd sample sizes are selected is to allow for the inclusion of the midpoint and both extremes of the range of x while maintaining uniform spacing.

When polynomials are used to generate data, the range of x is restricted to the interval (-1,1) for the purpose of providing reasonable control on the magnitude of the dependent variable. Moreover, two distinct values of the error variance σ^2 are used: 1 and 225. Since the magnitude of y is not likely to be excessive within the indicated range of x, it is believed that these two values of σ^2 provide modest and significant contamination to the data, respectively. Data are generated by using a polynomial function of degree $m \leqq 3$ and are fitted to the same model after contamination. Thus, for each value of n and σ^2, this part of the simulation is carried out according to the following scheme:

(1) Values for the coefficients of the polynomial are generated from the range -100 to 100.

(2) By using the generated coefficients and the appropriate X matrix, n uncontaminated values of y are generated.

(3) Each value of y is contaminated by generating a random number from a normal distribution with a mean of zero and a variance of σ^2.

(4) The least-squares estimates of the coefficients and the residual variance are computed according to equations (3) and (4), respectively.

(5) The squared differences between $\hat{\sigma}^2$ and σ^2 are computed and stored.

(6) Steps (1) to (5) are repeated 500 times to determine the mean-squared error of the residual variance.

The results are provided in figures 1 to 3, where the behavior of the mean-squared error of $\hat{\sigma}^2$ is given as a function of n for each polynomial model and each value of σ^2.

The second part of the computer simulation deals with generating data from transcendental functions, contaminating the data, and fitting them to either quadratic or cubic polynomial models. Three distinct functions are arbitrarily selected. These are

$$y = 2\left[\exp(-2x) - \exp(-4x)\right] \qquad (0 \leqq x \leqq 2) \qquad (5)$$

$$y = 2 \times 10^{4x/(100+x)} \qquad (0 \leqq x \leqq 100) \qquad (6)$$

$$y = 10x \ \exp\left(-\sqrt{x}/2\right) \qquad (0 \leqq x \leqq 200) \qquad (7)$$

where the selected range of x is indicated for each function.

Data generated by equation (5) are fitted to both quadratic and cubic models, while simulated data from equations (6) and (7) are fitted to quadratic and cubic models, respectively. The error variances used to generate Gaussian noise to contaminate the data generated by equations (5), (6), and (7) are 0.01, 81, and 9, respectively. Figures 4 to 6 are provided to show generated data before and after contamination for each one of equations (5) to (7).

The implementation of the simulation scheme for data generated by transcendental functions is analogous to that already discussed; that is, after generating the data by using one of equations (5) to (7), the simulation scheme picks up with step (3) of the outlined procedure. The results are given in figures 7 to 10, where once again the behavior of the mean-squared error of $\hat{\sigma}^2$ is depicted as a function of n. On the basis of the overall results, the following conclusions are apparent:

(1) The behavior of the mean-squared error of $\hat{\sigma}^2$ as a function of n resembles a fast-decaying exponential curve. This result is in agreement with the expected behavior of statistical estimators.

(2) In most cases, the mean-squared error is reduced dramatically as n increases to a moderate size. However, as n becomes larger, the mean-squared-error curves for nearly all cases flatten out so much that any further reduction may not be economically advantageous.

(3) Within the scope of the investigation, the substance of the results appears to be nearly invariant with the source from which the data are simulated. The same comment also applies to different values of the error variance.

From the results of this investigation, the effect that the sample size has on smoothing out the influence of random error in an analytic-function approximation of a stochastic quantity appears to be significant only for small sample sizes and diminishes considerably for larger values of n. Thus, the careful planning of only a moderate number of laboratory tests appears to be most beneficial.

CONCLUDING REMARKS

In an environment in which decisions are based on experimentation, it is often desired to determine an analytic representation of an experimentally observed quantity as a function of some controlled variable. Such a task is usually carried out by fitting laboratory measurements of the quantity to some simple analytic function, as a polynomial of relatively low degree. However, the amount of laboratory testing is becoming increasingly more restrictive mainly for economic reasons. Consequently, the purpose of this report has been to determine the effect of sample size on the accuracy of an analytic-function approximation of an experimentally observed quantity. Results obtained by using the Monte Carlo method indicate that for typical cases a moderate sample size provides an excellent trade-off between accuracy and economic restrictions.

Langley Research Center,
National Aeronautics and Space Administration,
Hampton, Va., February 12, 1974.

APPENDIX

OPTIMAL SPACING TECHNIQUES IN CURVE FITTING

When the degree of the polynomial function to be fitted is known, an optimal spacing of the independent variable has been developed by De la Garza (ref. 5). Consider the polynomial function

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \ldots + \beta_m x^m$$

of known degree m. Assume that n observations of y will be made within the range of x, which is scaled to the interval (-1,1) for convenience. De la Garza (ref. 5) showed that to minimize the maximum variance of a predicted quantity y, the optimum spacing of the n observations is accomplished by using no more than m + 1 distinct observation points within the range of x. The spacing for minimax variance is provided in the following table through the cubic polynomial function:

Model	Observation points	Number of measurements per observation point
Linear	1 -1	n/2
Quadratic	1 0 -1	n/3
Cubic	1 $\sqrt{1/5}$ $-\sqrt{1/5}$ -1	n/4

Let x_p be any arbitrary point of prediction within the interval (-1,1). If the indicated optimal spacing is used, it has also been shown (ref. 5) that the maximum variance of $\hat{y}_p$ corresponding to the prediction point x_p is

$$\max\left[\operatorname{var}\left(\hat{y}_p\right)\right] = \frac{(m + 1)\sigma^2}{n}$$

where the residual variance $\hat{\sigma}^2$ usually replaces the unknown error variance σ^2. In fact, given the polynomial model, the absolute minimum variance of $\hat{y}_p$ can also be determined. As an example, consider the cubic model. The matrix X which corresponds to the minimax variance spacing when n is a multiple of 4 is

$$X = \begin{bmatrix} 1 & -1 & 1 & -1 \\ 1 & -\sqrt{5}/5 & 1/5 & -\sqrt{5}/25 \\ 1 & \sqrt{5}/5 & 1/5 & \sqrt{5}/25 \\ 1 & 1 & 1 & 1 \\ \cdot & & & \\ \cdot & & & \\ \cdot & & & \\ 1 & -1 & 1 & -1 \\ 1 & -\sqrt{5}/5 & 1/5 & -\sqrt{5}/25 \\ 1 & \sqrt{5}/5 & 1/5 & \sqrt{5}/25 \\ 1 & 1 & 1 & 1 \end{bmatrix}$$

Thus,

$$\left(X^T X\right) = n \begin{bmatrix} 1 & 0 & 3/5 & 0 \\ 0 & 3/5 & 0 & 13/25 \\ 3/5 & 0 & 13/25 & 0 \\ 0 & 13/25 & 0 & 63/125 \end{bmatrix}$$

and

$$\left(X^TX\right)^{-1} = \frac{1}{4n}\begin{bmatrix} 13 & 0 & -15 & 0 \\ 0 & 63 & 0 & -65 \\ -15 & 0 & 25 & 0 \\ 0 & -65 & 0 & 75 \end{bmatrix}$$

Therefore, if x_p is the point of prediction,

$$\text{var}\left(\hat{y}_p\right) = \underline{x}_p{}^T\left(X^TX\right)^{-1}\underline{x}_p\,\sigma^2$$

$$= \begin{bmatrix} 1 & x_p & x_p{}^2 & x_p{}^3 \end{bmatrix}\begin{bmatrix} 13 & 0 & -15 & 0 \\ 0 & 63 & 0 & -65 \\ -15 & 0 & 25 & 0 \\ 0 & -65 & 0 & 75 \end{bmatrix}\begin{bmatrix} 1 \\ x_p \\ x_p{}^2 \\ x_p{}^3 \end{bmatrix}\frac{\sigma^2}{4n}$$

$$= \frac{13 + 33x_p{}^2 - 105x_p{}^4 + 75x_p{}^6}{4n}$$

where σ^2 is temporarily dropped for convenience. It follows that the points at which maximum or minimum variances occur are determined by solving the equation

$$\frac{d\left[\text{var}\left(\hat{y}_p\right)\right]}{dx_p} = 0$$

or

$$66x_p - 420x_p{}^3 + 450x_p{}^5 = 0$$

which, upon simplification, yields the values $x_p = 0$, $\pm\sqrt{11/15}$, $\pm\sqrt{1/5}$. By examining the second derivative, it is determined that the minimum variance for a cubic model is $2.578\sigma^2/n$ and occurs at $x_p = \pm\sqrt{11/15}$, while the maximum variance $4\sigma^2/n$ occurs at $x_p = \pm\sqrt{1/5}$.

Assuming the spacing for minimax variance with regard to linear, quadratic, and cubic polynomials, the following table provides upper and lower bounds of the variance of a predicted value $\hat{y}_p$ when the prediction point is within the range -1 to 1:

Model	Variance boundaries for a predicted value
Linear	$\frac{\sigma^2}{n} \leqq \mathrm{var}\left(\hat{y}_p\right) \leqq \frac{2\sigma^2}{n}$
Quadratic	$\frac{1.875\sigma^2}{n} \leqq \mathrm{var}\left(\hat{y}_p\right) \leqq \frac{3\sigma^2}{n}$
Cubic	$\frac{2.578\sigma^2}{n} \leqq \mathrm{var}\left(\hat{y}_p\right) \leqq \frac{4\sigma^2}{n}$

REFERENCES

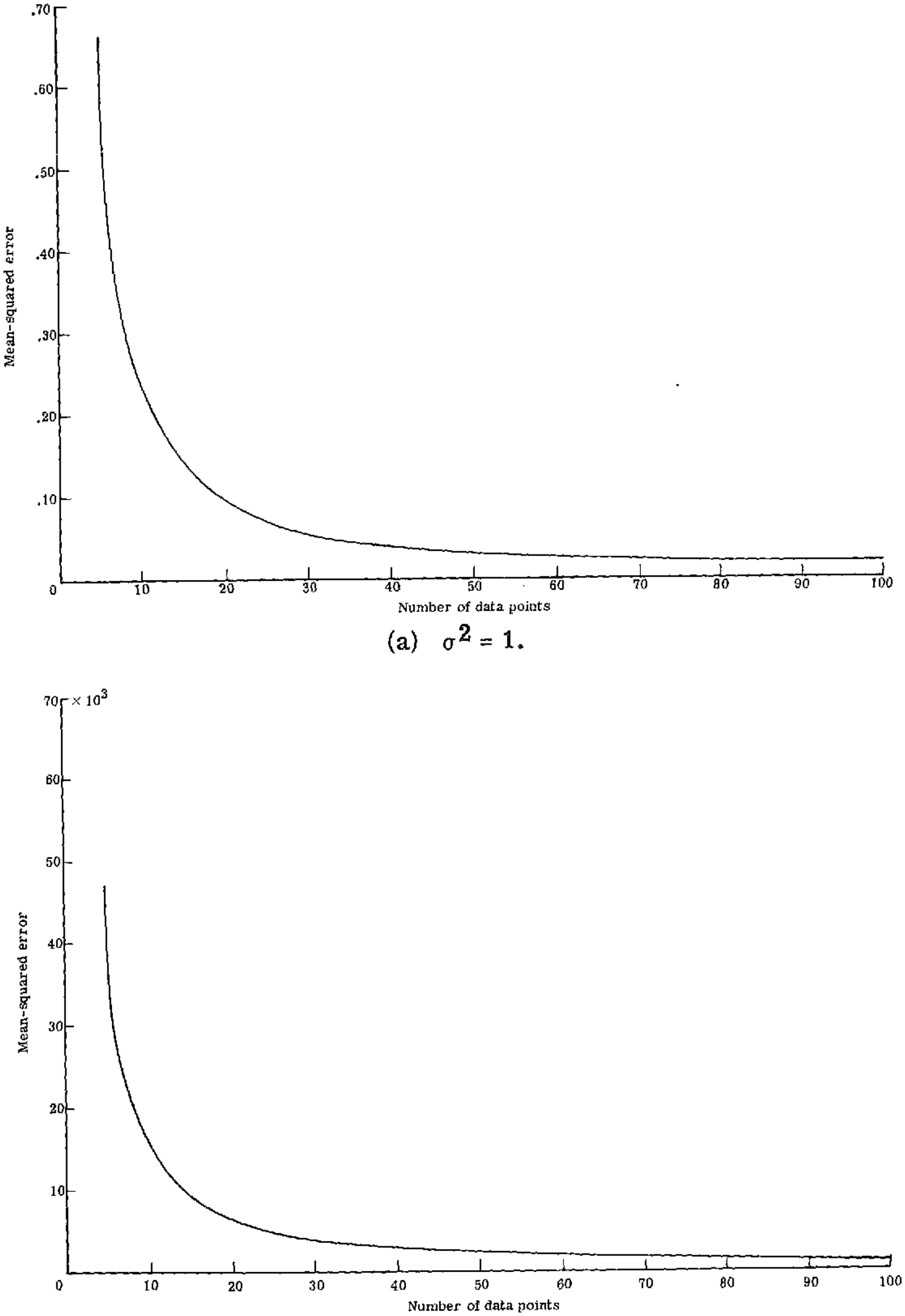
1. Naylor, Thomas H.; Balintfy, Joseph L.; Burdick, Donald S.; and Chu, Kong: Computer Simulation Techniques. John Wiley & Sons, Inc., c.1966.

2. Draper, N. R.; and Smith, H.: Applied Regression Analysis. John Wiley & Sons, Inc., c.1966.

3. Graybill, Franklin A.: An Introduction to Linear Statistical Models. Vol. I. McGraw-Hill Book Co., Inc., 1961.

4. Guest, P. G.: Numerical Methods of Curve Fitting. Cambridge Univ. Press, 1961.

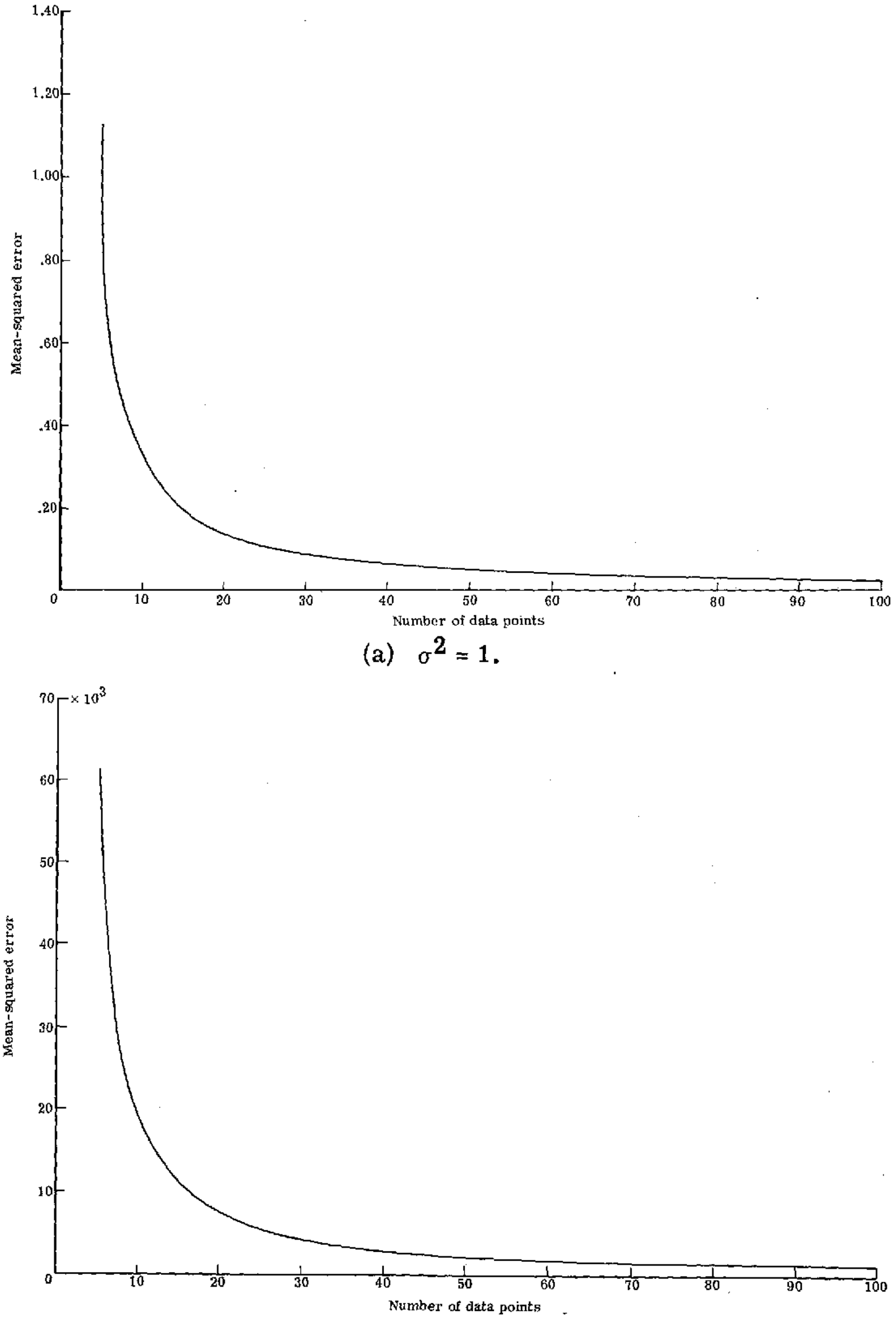
5. De la Garza, A.: Spacing of Information in Polynomial Regression. Ann. Math. Statist., vol. 25, no. 1, Mar. 1954, pp. 123-130.



(a) $\sigma^2 = 1.$

(b) $\sigma^2 = 225.$

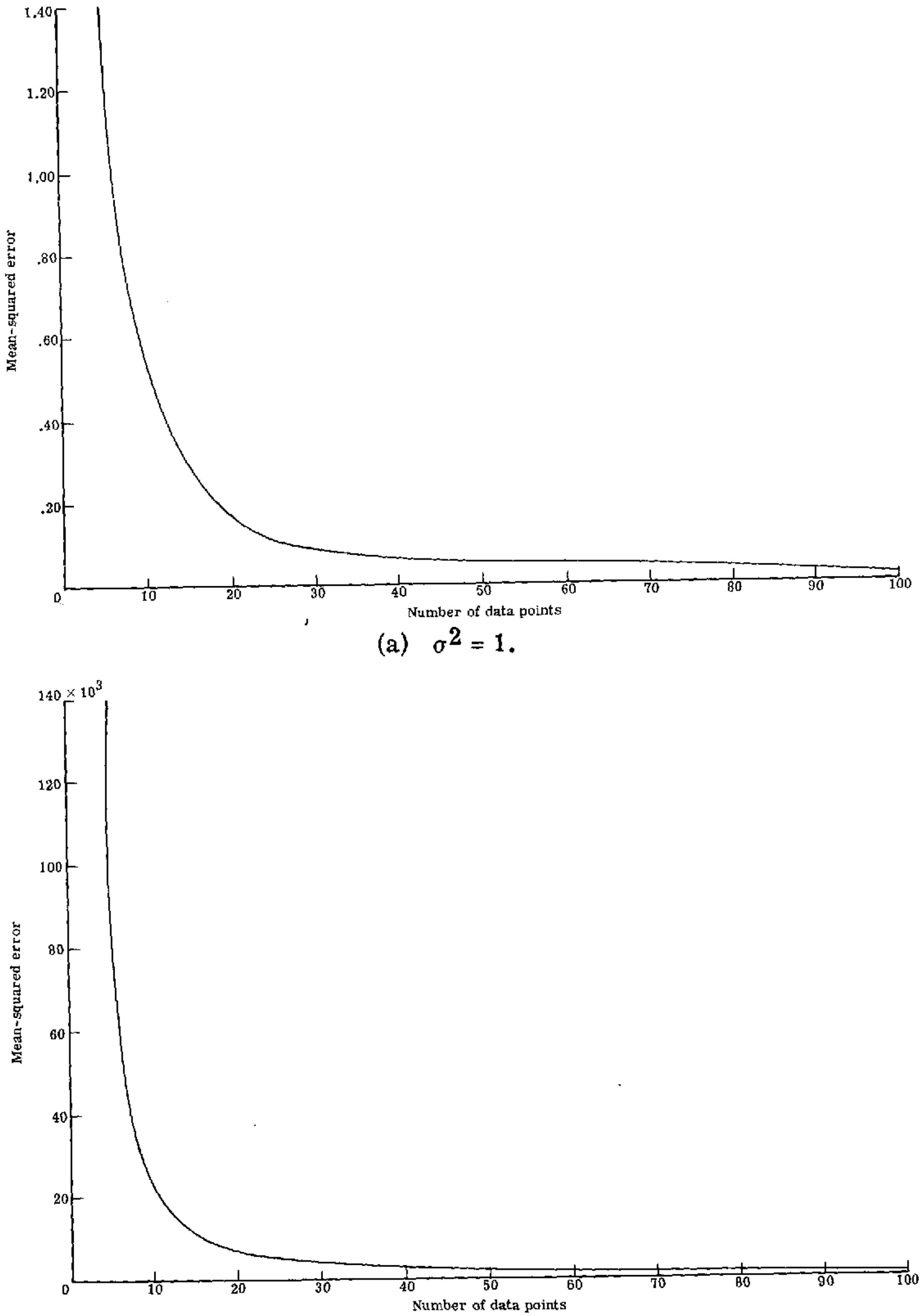
Figure 1.- Mean-squared error of residual variance for a linear model.



(a) $\sigma^2 = 1.$

(b) $\sigma^2 = 225.$

Figure 2.- Mean-squared error of residual variance for a quadratic model.



(a) $\sigma^2 = 1.$

(b) $\sigma^2 = 225.$

Figure 3.- Mean-squared error of residual variance for a cubic model.

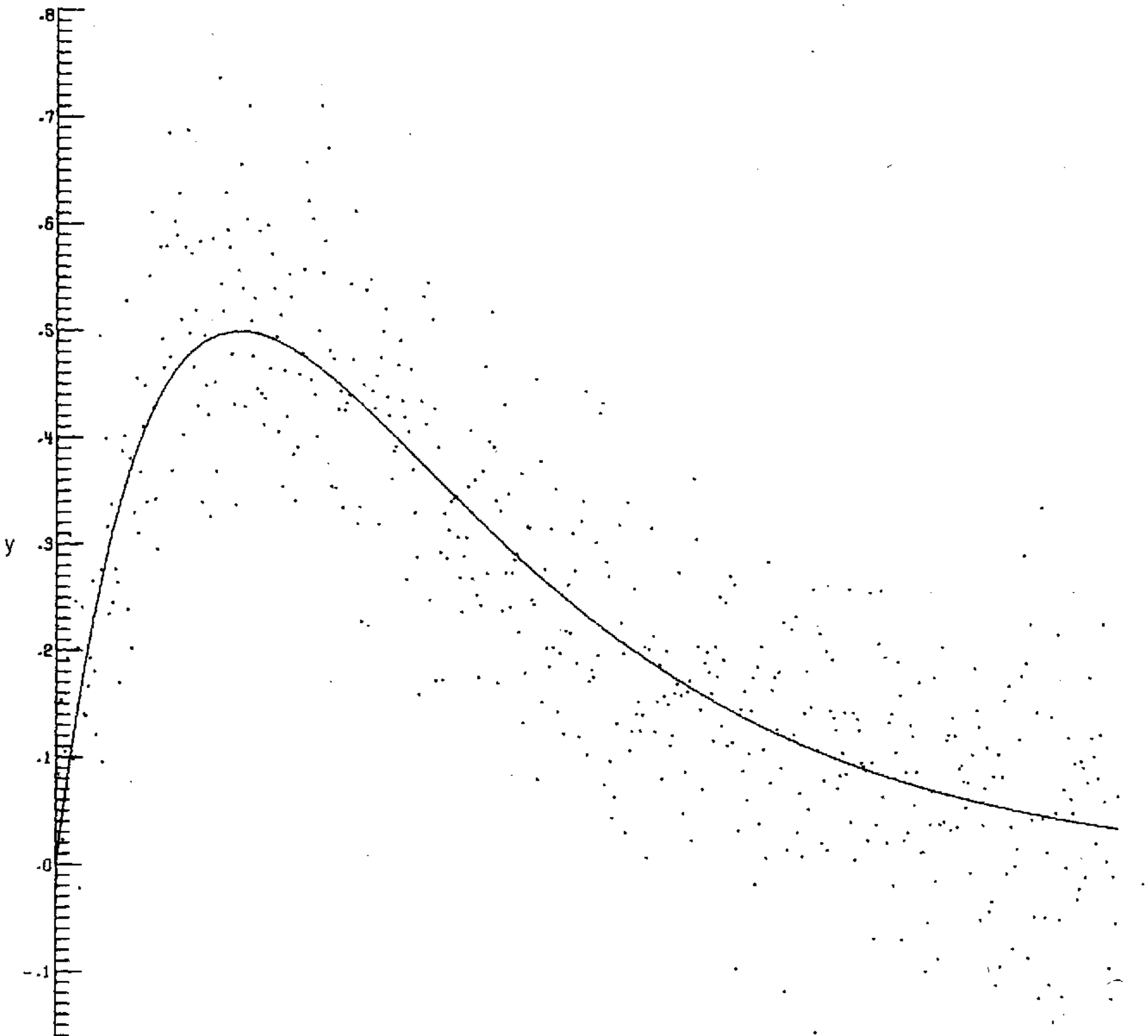


Figure 4.- Pure and contaminated data generated by equation (5).

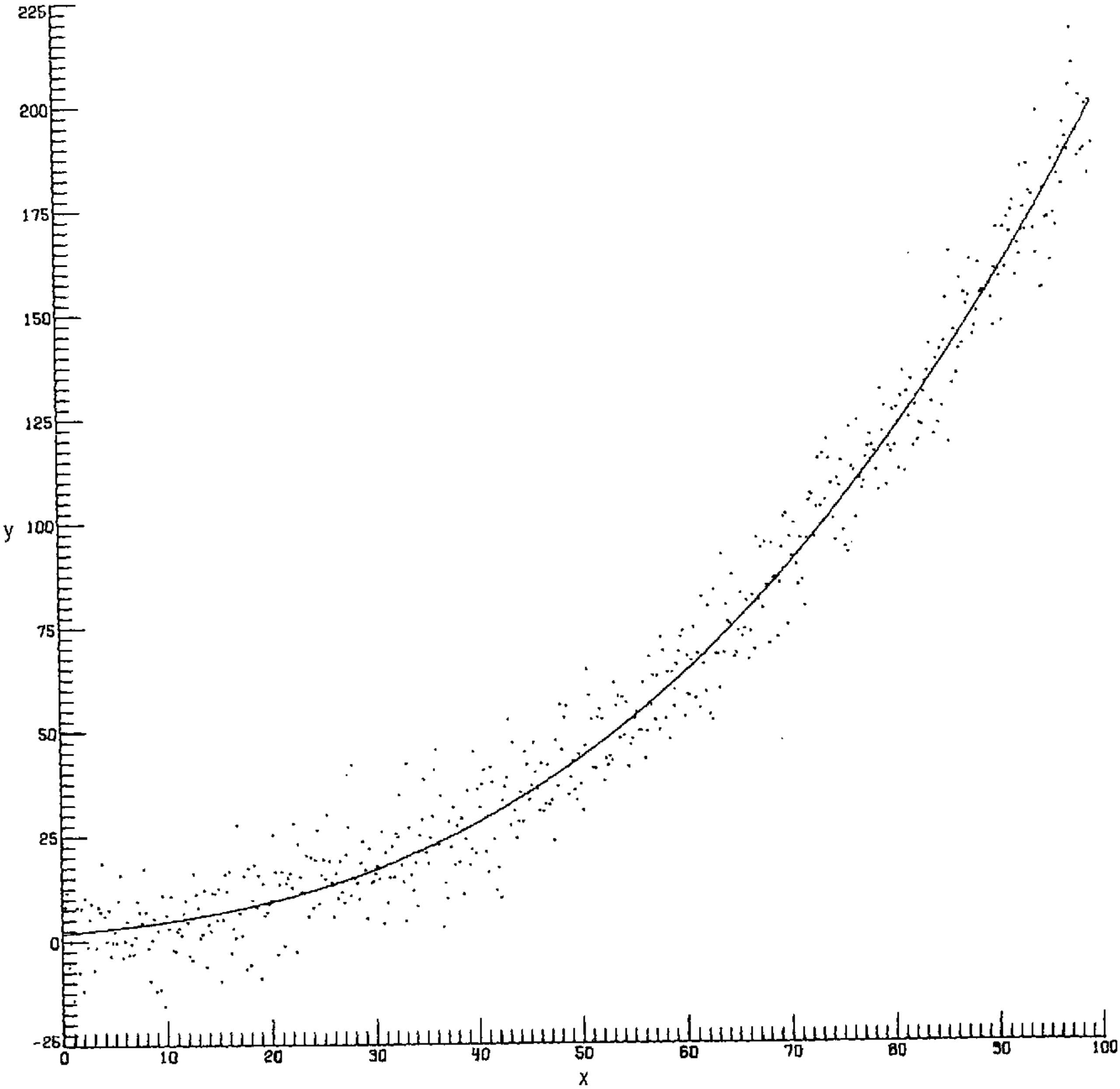


Figure 5.- Pure and contaminated data generated by equation (6).

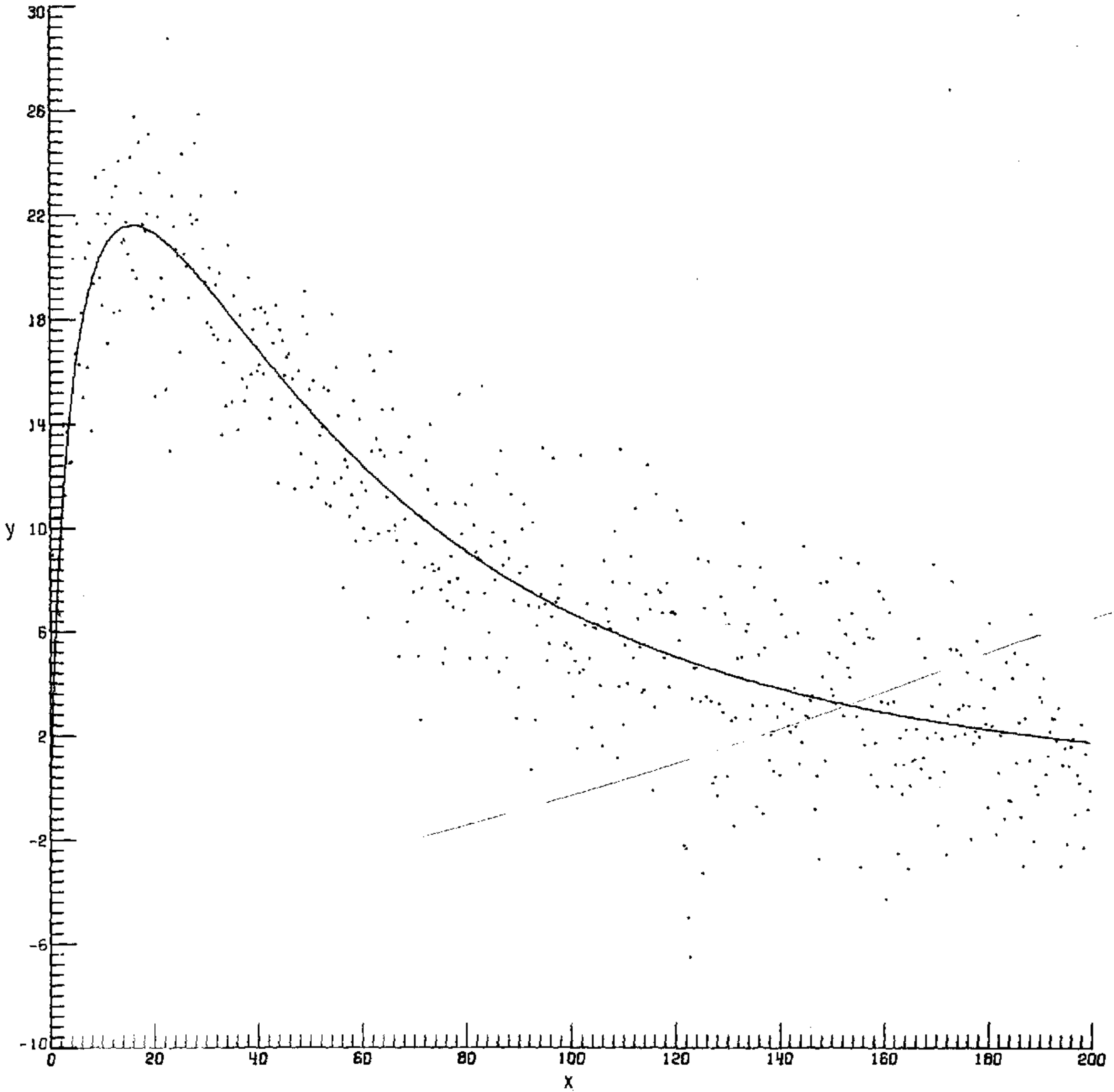


Figure 6.- Pure and contaminated data generated by equation (7).

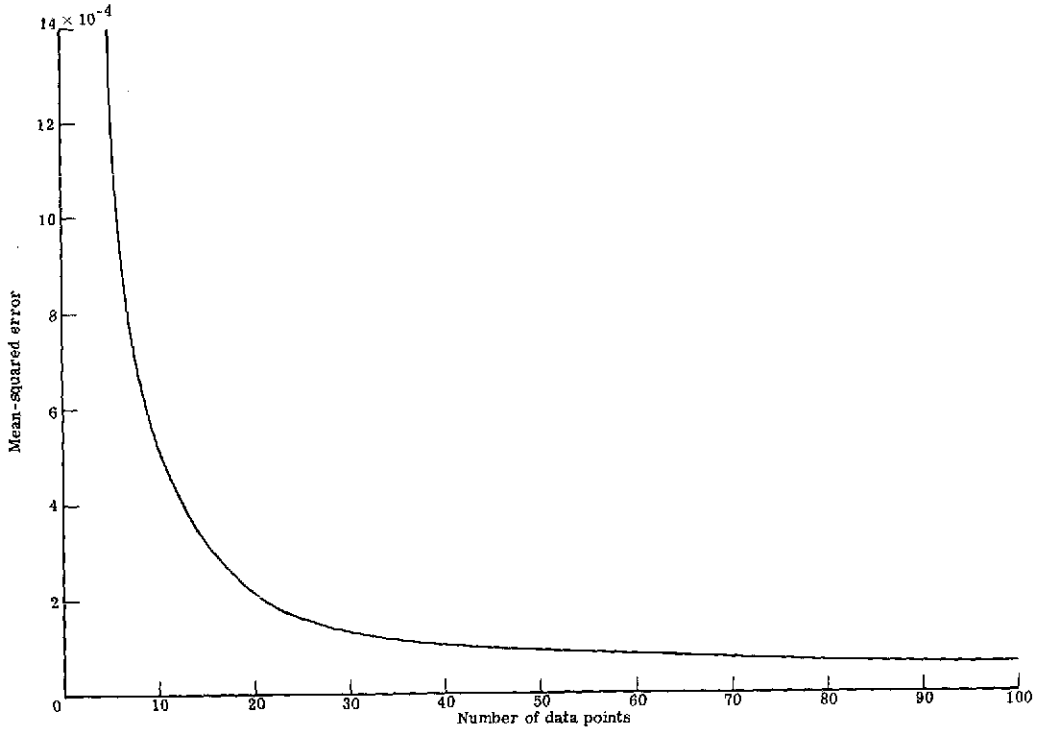


Figure 7.- Mean-squared error of residual variance for data generated by equation (5) and fitted to a quadratic polynomial.

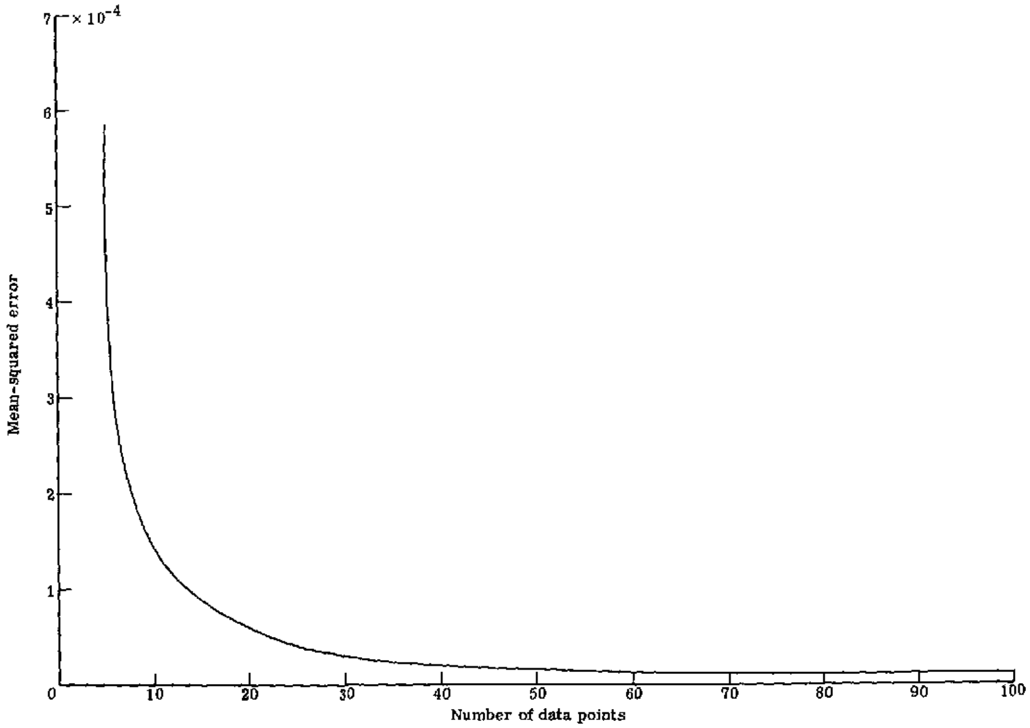


Figure 8.- Mean-squared error of residual variance for data generated by equation (5) and fitted to a cubic polynomial.

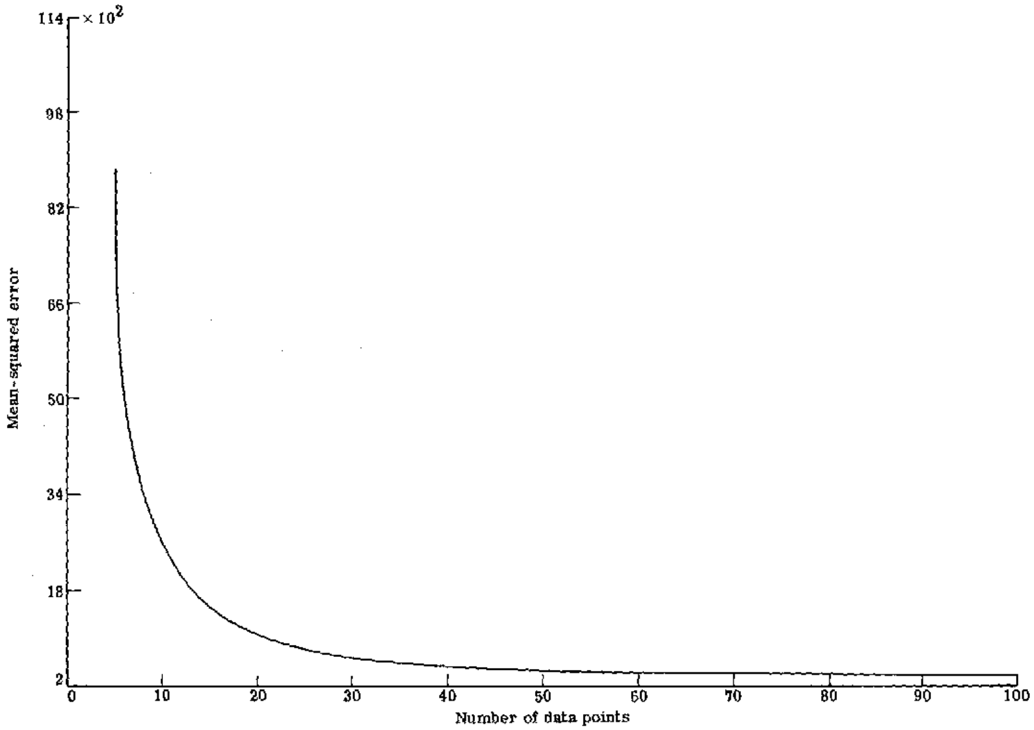


Figure 9.- Mean-squared error of residual variance for data generated by equation (6) and fitted to a quadratic polynomial.

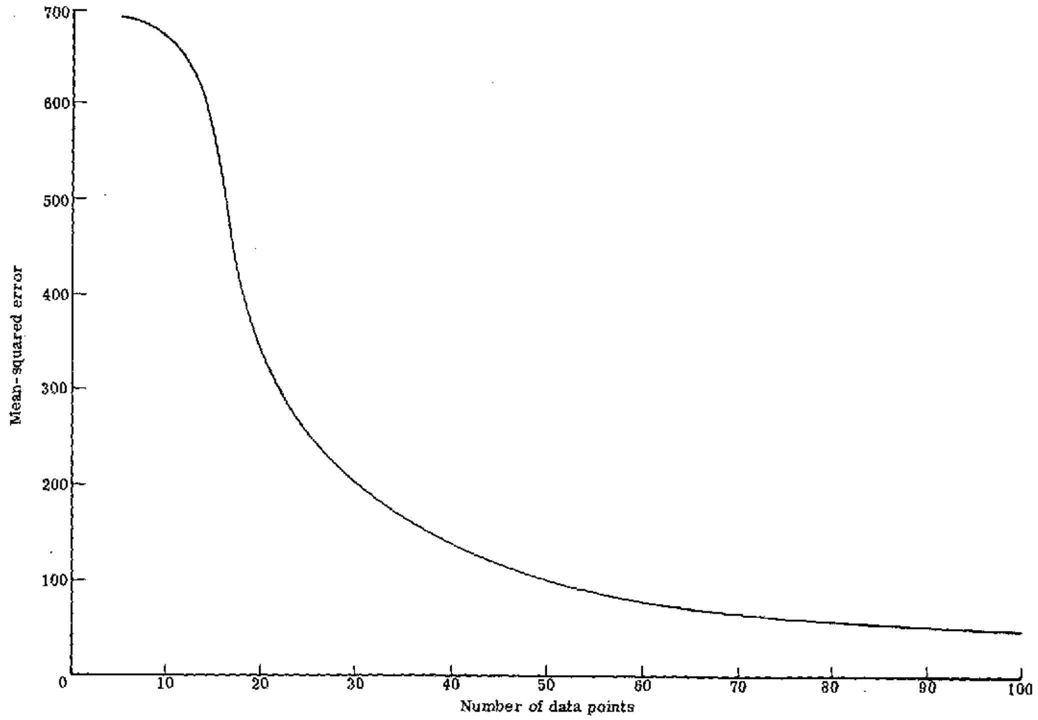


Figure 10.- Mean-squared error of residual variance for data generated by equation (7) and fitted to a cubic polynomial.