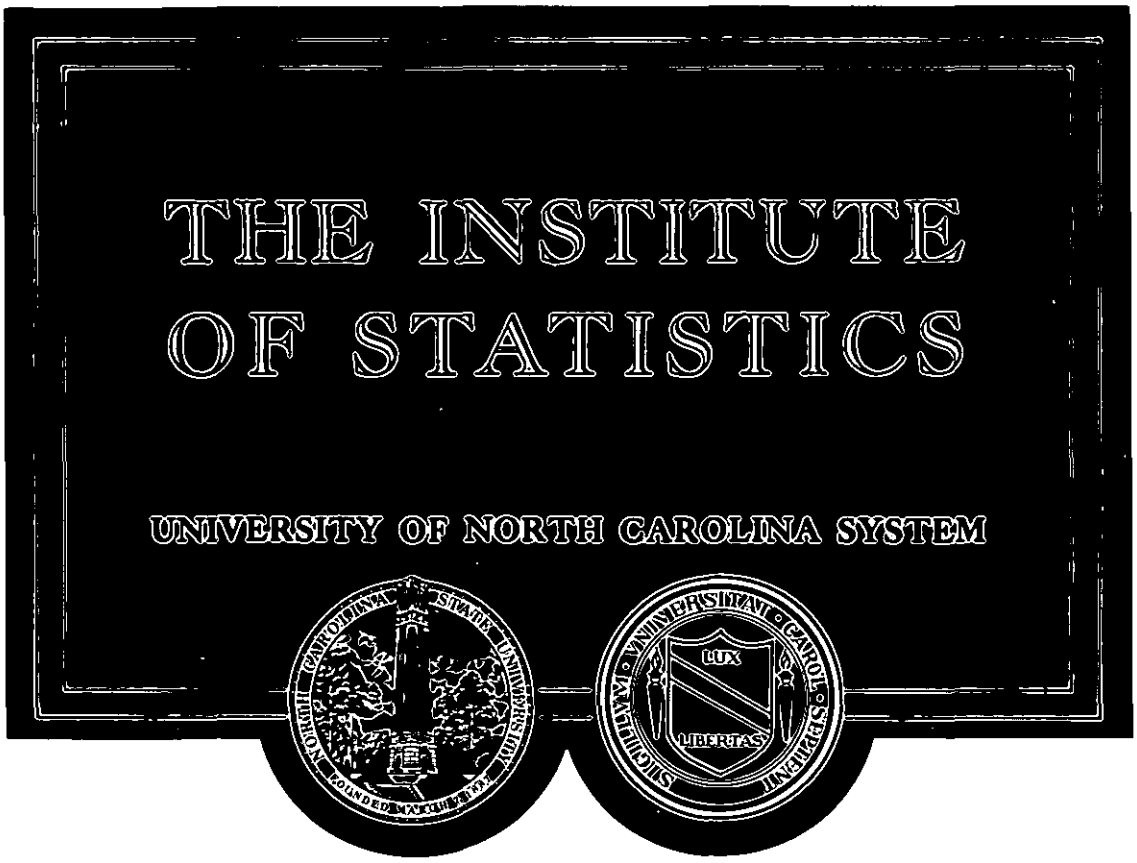

Nonlinear Statistical Models

by
A. Ronald Gallant

Chapter 2. Univariate Nonlinear Regression: Special Situations

Institute of Statistics Mimeograph Series No. 1682

December 1985

NORTH CAROLINA STATE UNIVERSITY
Raleigh, North Carolina

NONLINEAR STATISTICAL MODELS

by

A. Ronald Gallant

Chapter 2. Univariate Nonlinear Regression: Special Situations

A. Ronald Gallant
Institute of Statistics
North Carolina State University
Post Office Box 8203
Raleigh, NC 27695-8203

Phone 1-919-737-2531

NONLINEAR STATISTICAL MODELS

Table of Contents

Table of Contents (continued)

Chapter 2. Univariate Nonlinear Regression: Special Situations

In this chapter, we shall consider some special situations that arise with some frequency in the analysis of univariate nonlinear models but lie outside the scope of the standard least-squares methods that were discussed in the previous chapter.

The first situation considered is the problem of heteroskedastic errors. Two solutions are proposed Either deduce the pattern of the heteroskedasticity, transform the model, and then apply standard nonlinear methods or use least squares and substitute heteroskedastic-invariant variance estimates and test statistics. The former offers efficiency gains if a suitable transformation can be found.

The second situation is the problem of serially correlated errors. The solution is much as above. If the errors appear to be covariance stationary then a suitable transformation will reduce the model to a standard case. If the errors appear to be both serially correlated and heteroskedastic, then least squares estimators can be used with invariant variance estimates and test statistics.

The third is a testing problem involving model choice which arises quite often in applications but violates the regularity conditions needed to apply standard methods. A variant of the lack-of-fit test is proposed as a solution.

The last topic is a brief discussion of nonlinearity measures. They can be used to find transformations that will improve the performance of optimization routines and, perhaps, the accuracy of probability statements. The latter is an open question as the measures relate to sufficient conditions, not necessary conditions, and little Monte-Carlo evidence is available.

1. HETEROSKEDASTIC ERRORS

If the variance σ_t^2 of the errors in the nonlinear model

$$y_t = f(x_t,\theta^o) + e_t \qquad t = 1, 2, \ldots, n$$

is known to depend on x_t, *viz.*

$$\sigma_t^2 = \sigma^2 / \psi^2(x_t),$$

then the situation can be remedied using weighted least squares -- see Judge *et al.* (1980, Section 4.3) for various tests for heteroskedasticity. Put

$$"y_t" = \psi(x_t)y_t$$

$$"f"(x_t,\theta) = \psi(x_t)f(x_t,\theta)$$

and apply the methods of the previous chapter with $"y_t"$ and $"f"(x_t,\theta)$ replacing y_t and $f(x_t,\theta)$ throughout. The justification for this approach is straightforward. If the errors e_t are independent then the errors

$$"e_t" = "y_t" - "f"(x_t,\theta^o) \qquad t = 1, 2, \ldots, n$$

will be independent and have constant variance σ^2 as required.

If the transformation

$$\sigma_t^2 = \sigma^2 / \psi^2(x_t,\tau^o)$$

depends on an unknown parameter τ, there are a variety of approaches that one might use. If one is willing to take the trouble, the best approach is to write the model as

$$q(y_t,x_t,\theta^o,\tau^o) = \psi(x_t,\tau^o)[y_t - f(x_t,\theta)] = "e_t"$$

and estimate the parameters $\lambda = (\theta,\tau,\sigma^2)$ jointly using maximum likelihood as discussed in Section 5 of Chapter 8. If not, and the parameters τ do not depend functionally on θ -- or one is willing to forego efficiency gains if they do -- then a two-step approach can be used. It is as follows.

Let $\hat{\theta}$ denote the least squares estimator computed by minimizing

$$\Sigma_{t=1}^{n}[y_t - f(x_t,\theta)]^2.$$

Put

$$|\hat{e}_t| = |y_t - f(x_t,\hat{\theta})|$$

and estimate τ^o by $\hat{\tau}$ where $(\hat{\tau},\hat{c})$ minimizes

$$\Sigma_{t=1}^{n}[|\hat{e}_t| - c/\psi(x_t,\tau)]^2;$$

$\hat{c}$ will be a consistent estimator of $\sqrt{(2\sigma^2/\pi)}$ if the errors are normally distributed. The methods discussed in Section 4 of Chapter 1 can be used to compute this minimum. Put

$$"y_t" = \psi(x_t,\hat{\tau})y_t$$

$$"f"(x_t,\theta) = \psi(x_t,\hat{\tau})f(x_t,\theta)$$

and apply the methods of the previous chapter with $"y_t"$ and $"f"(x_t,\theta)$ replacing y_t and $f(x_t,\theta)$ throughout. Section 3 of Chapter 3 provides the theoretical justification for this approach.

If one suspects that heteroskedasticity is present but cannot deduce an acceptable form for $\psi(x_t,\tau)$, another approach is to use least squares estimators and correct the variance estimate. As above, let $\hat{\theta}$ denote the least squares estimator, the value that minimizes

$$s_n(\theta) = (1/n)\sum_{t=1}^{n}[y_t - f(x_t,\theta)]^2,$$

and let $\hat{e}_t$ denote residuals

$$\hat{e}_t = y_t - f(x_t,\hat{\theta}) \qquad t = 1, 2, \ldots, n.$$

Upon application of the results of Section 3 of Chapter 3

$$\sqrt{n}(\hat{\theta} - \theta^o) \xrightarrow{\mathcal{L}} N_p(0,V)$$

with

$$V = \mathcal{J}^{-1}\mathcal{I}\,\mathcal{J}^{-1}.$$

$\mathcal{I}$ and $\mathcal{J}$ can be estimated using

$$\hat{\mathcal{I}} = (1/n)\sum_{t=1}^{n} \hat{e}_t^2[(\partial/\partial\theta)f(x_t,\hat{\theta})][(\partial/\partial\theta)f(x_t,\hat{\theta})]'$$

and

$$\hat{\mathcal{J}} = (1/n)\sum_{t=1}^{n}[(\partial/\partial\theta)f(x_t,\hat{\theta})][(\partial/\partial\theta)f(x_t,\hat{\theta})]'.$$

For testing

$$H: h(\theta^o) = 0 \quad \text{against} \quad A: h(\theta^o) \neq 0$$

where $h: \mathbb{R}^p \to \mathbb{R}^q$ the Wald test statistic is (Theorem 11, Chapter 3)

$$W = nh'(\hat{\theta})[\hat{H}\hat{V}\hat{H}']^{-1}h(\hat{\theta})$$

where $\hat{H} = (\partial/\partial\theta')h(\hat{\theta})$ and $\hat{V} = \hat{\mathcal{J}}^{-1}\hat{\mathcal{I}}\hat{\mathcal{J}}^{-1}$. The null hypothesis $H: h(\theta^o) = 0$ is rejected in favor of the alternative hypothesis $A: h(\theta^o) \neq 0$ when the test statistic exceeds the upper $\alpha \times 100$ percentage point χ^2_α of a chi-square random variable with q degrees of freedom; $\chi^2_\alpha = (\chi^2)^{-1}(1-\alpha,q)$.

As a consequence of this result, a 95% confidence interval on the i^{th} element of θ^o is computed as

$$\hat{\theta}_i \pm z_{.025}\sqrt{(\hat{V}_{ii})}/\sqrt{n}$$

where $z_{.025} = -\sqrt{(\chi^2)^{-1}}(.95,1) = N^{-1}(.025|0,1)$.

Let $\tilde{\theta}$ denote the minimizer of $s_n(\theta)$, subject to the restriction that $h(\theta) = 0$. Let $\tilde{H}$, $\tilde{\mathcal{J}}$ and $\tilde{\mathcal{I}}$ denote the formulas for $\hat{H}$, $\hat{\mathcal{J}}$ and $\hat{\mathcal{I}}$ above but with $\tilde{\theta}$ replacing $\hat{\theta}$ throughout; put

$$\tilde{V} = \tilde{\mathcal{I}}^{-1}\tilde{\mathcal{J}}\,\tilde{\mathcal{I}}^{-1}.$$

The Lagrange multiplier test statistic is (Theorem 14, Chapter 3)

$$R = n[(\partial/\partial\theta)s_n(\tilde{\theta})]'\tilde{\mathcal{I}}^{-1}\tilde{H}'(\tilde{H}\tilde{V}\tilde{H}')^{-1}\tilde{H}\tilde{\mathcal{I}}^{-1}[(\partial/\partial\theta)s_n(\tilde{\theta})] .$$

Again, the null hypothesis H: $h(\theta^o) = 0$ is rejected in favor of the alternative hypothesis A: $h(\theta^o) \neq 0$ when the test statistic exceeds the upper $\alpha \times 100$ percentage point χ^2_α of a chi-square random variable with q degrees of freedom; $\chi^2_\alpha = (\chi^2)^{-1}(1-\alpha,q)$.

The likelihood ratio test cannot be used because $\mathcal{J} \neq \mathcal{I}$; see Theorem 15 of Chapter 3. Formulas for computing the power of the Wald and Lagrange multiplier tests are given in Theorems 11 and 14 of Chapter 3, respectively.

2. SERIALLY CORRELATED ERRORS

In this section we shall consider estimation and inference regarding the parameter θ^o in the univariate nonlinear model

$$y_t = f(x_t, \theta^o) + u_t, \qquad t = 1, 2, \ldots, n,$$

when the errors are serially correlated. In most applications -- methods for handling exceptions are considered at the end of the section -- an assumption that the process $\{u_t\}_{t=-\infty}^{\infty}$ generating the realized disturbances $\{u_t\}_{t=1}^{n}$ is covariance stationary is plausible. This is to say that the covariances $\text{cov}(u_t, u_{t+h})$ of the time series depend only on the gap h and not on the position t in time. In consequence, the variance-covariance matrix Γ_n of the disturbance vector

$$u = (u_1, u_2, \ldots, u_n)' \qquad (n \times 1)$$

will have a banded structure with typical element $\gamma_{ij} = \gamma(i-j)$ where $\gamma(h$ is the autocovariance function of the process, *viz.*

$$\gamma(h) = \text{cov}\,(u_t, u_{t+h}), \qquad h = 0, \pm 1, \pm 2, \ldots .$$

The appropriate estimator, were Γ_n known, would be the generalized nonlinear least squares estimator. Specifically, one would estimate θ^o by the value of θ that minimizes

$$[y - f(\theta)]'\Gamma_n^{-1}[y - f(\theta)],$$

where

$$y = (y_1, y_2, \ldots, y_n)' \qquad (n \times 1)$$

and

$$f(\Theta) = [f(x_1), f(x_2), \ldots, f(x_n)]' \qquad (n \times 1) .$$

The generalized nonlinear least squares estimator is seen to be appropriate from the following considerations. Suppose that Γ_n^{-1} can be factored as

$$\Gamma_n^{-1} = (\text{scalar}) \cdot P'P .$$

If we put

$$\text{"y"} = Py, \quad \text{"f"}(\Theta) = Pf(\Theta), \quad \text{"e"} = Pu$$

then the model

$$\text{"y"} = \text{"f"}(\Theta) + \text{"e"}$$

satisfies the assumptions -- $\mathcal{E}(\text{"e"}) = 0$, $\mathcal{C}(\text{"e"},\text{"e"}') = \sigma^2 I$ -- that justifies the use of the least squares estimator and associated inference procedures. However, the least squares estimator computed from the model

$$\text{"y"} = \text{"f"}(\Theta) + \text{"e"}$$

is the same as the generalized least squares estimator above. This justifies the approach. More importantly, it provides computational and inference procedures -- one need only transform the model using P and then apply the methods of Chapter 1 forthwith. For this approach to be practical, the matrix P must be easy to obtain, must be representable using far fewer than n^2 storage locations, and the multiplication Pw must be convenient relative to the coding requirements of standard nonlinear least squares statistical packages. As we shall see below, if an autoregressive assumption is justified, then P is

easy to obtain, can be stored using very few storage locations, and the multiplication Pw is particularly convenient.

When Γ_n is not known, as we assume here, the obvious approach is to substitute an estimator $\hat{\Gamma}_n$ in the formulas above. Section 4 of Chapter 9 furnishes the theoretical justification for this approach provided that $\hat{\Gamma}_n$ depends on a finite-length vector $\hat{\tau}_n$ of random variables with $\sqrt{n}(\hat{\tau}_n - \tau^o)$ bounded in probability for some τ^o. A proof that $\hat{\Gamma}_n$ computed as described below satisfies this restriction is given by Gallant and Goebel (1975).

An assumption frequently satisfied in applications -- at least to within errors which can reasonably be ascribed to sampling variation -- is that the disturbances can be reduced to a white noise process by using a short linear filter. Specifically, the time series $\{u_t\}_{t=-\infty}^{\infty}$ is assumed to satisfy the equations

$$u_t + a_1 u_{t-1} + a_2 u_{t-2} + \ldots + a_q u_{t-q} = e_t \qquad t = 0, \pm 1, \pm 2, \ldots$$

where $\{e_t\}_{t=-\infty}^{\infty}$ is a sequence of independently and identically distributed random variables each with mean zero and variance σ^2. In addition we assume that the roots of the characteristic polynomial

$$m^q + a_1 m^{q-1} + a_2^{q-2} + \ldots + a_q$$

are less than one in absolute value. The necessity for this assumption is discussed in Fuller (1976, Chapter 2); Pantula (1985) describes a testing strategy for determining the validity of this assumption. A time series $\{u_t\}_{t=-\infty}^{\infty}$ which satisfies this assumption is called an autoregressive process of order q.

EXAMPLE 1. (Wholesale Prices) The Wholesale Price Index for the years 1720 through 1973 provides an illustration. The data is listed in Table 1 and plotted as Figure 1. Using least squares, an exponential growth model

$$y_t = \theta_1 e^{\theta_2 \cdot t} + u_t , \qquad t = 1, 2, \ldots, n = 254$$

was fitted to the data to obtain residuals $\{\hat{u}_t\}_{t=1}^{254}$. From these residuals, the autocovariances have been estimated using

$$\hat{\gamma}(h) = (1/n)\sum_{t=1}^{n-h} \hat{u}_t \hat{u}_{t+h} \qquad h = 0, 1, \ldots, 60$$

and plotted as Plot Autocovariance in Figure 2. Using the methods discussed below, a second order autoregression

$$u_t + a_1 u_{t-1} + a_2 u_{t-2} = e_t,$$

was fitted to the residuals $\{\hat{u}_t\}_{t=1}^{254}$ to obtain

$$(\hat{a}_1, \hat{a}_2, \hat{\sigma}^2) = (-1.048, 0.1287, 34.09).$$

Estimates of the autocovariances can be calculated from these estimates using the Yule-Walker equations as discussed in Anderson (1971, p. 174). Doing so yields the estimates plotted as Plot Autoregressive in Figure 2. The two plots shown as Figure 2-- Autocovariance, unrestricted estimates requiring that 60 population quantities be estimated, and Autoregressive, requiring that only three population quantities be estimated -- are in reasonable agreement and the autoregressive assumption seems to yield an adequate approximation to the autocovariances. Indeed, it must for large enough q if the process $\{u_t\}_{t=-\infty}^{\infty}$ is, in fact, stationary (Berk, 1974).

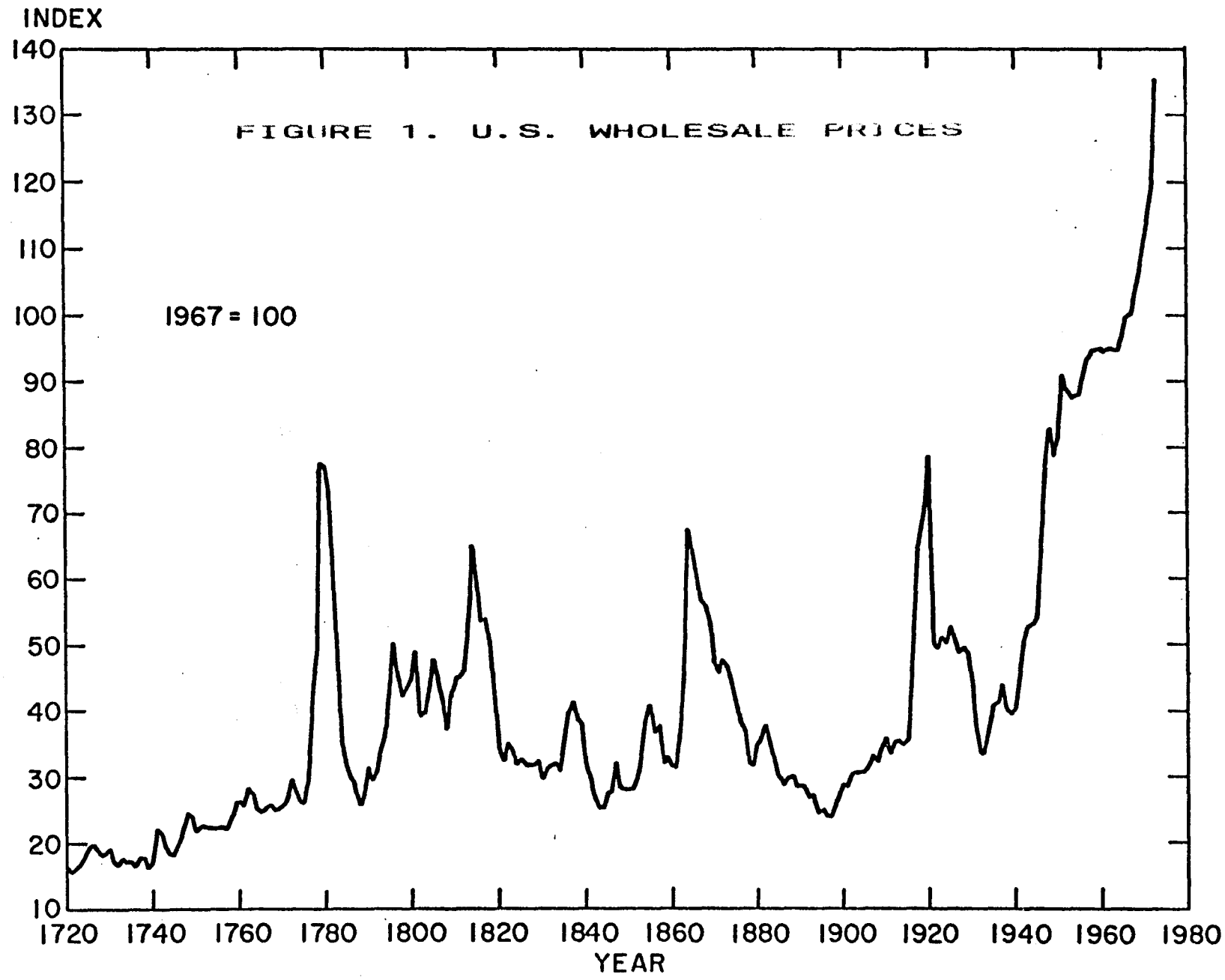

FIGURE 1. U.S. WHOLESALE PRICES

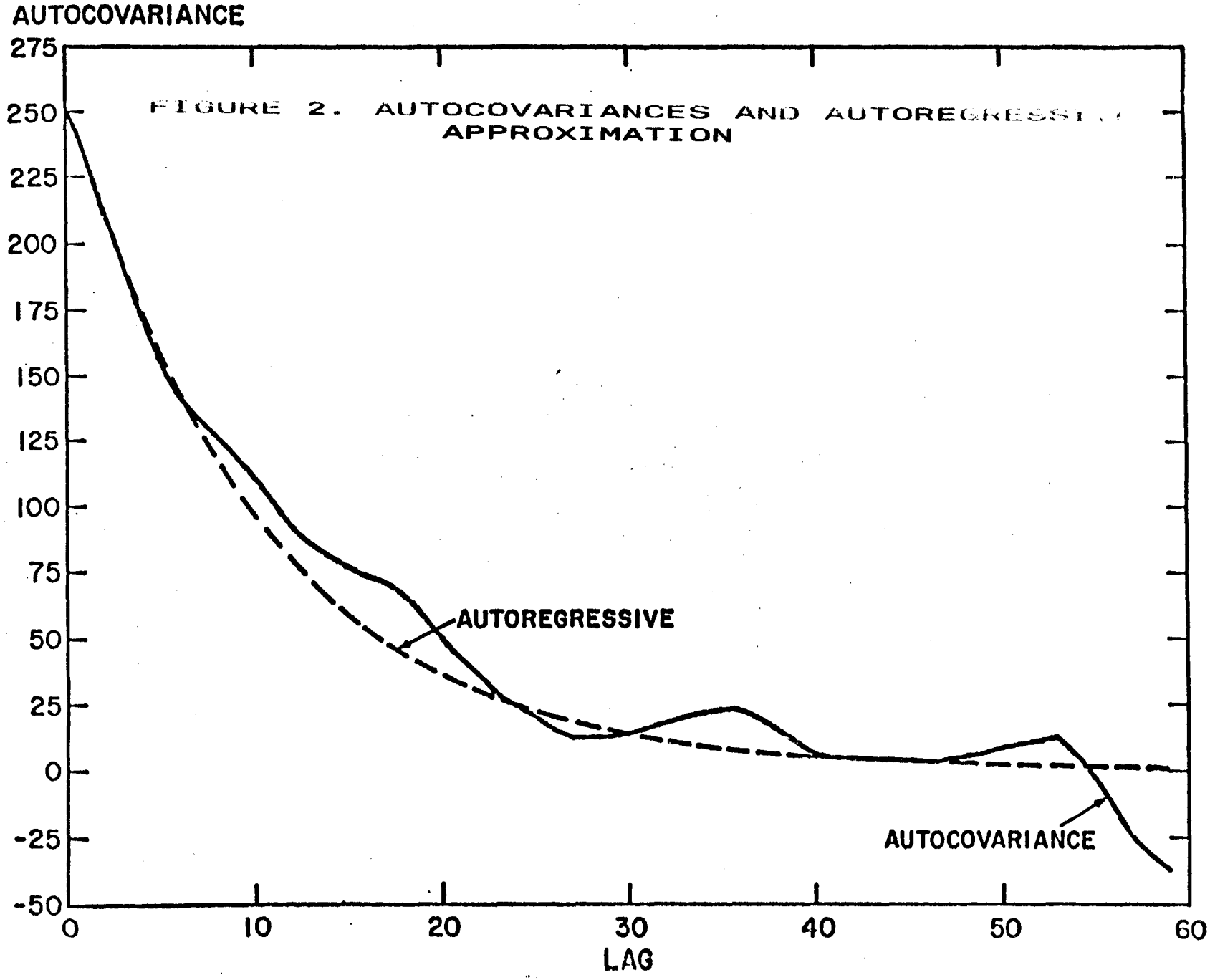

FIGURE 2. AUTOCOVARIANCES AND AUTOREGRESSIVE APPROXIMATION

Table 1. U.S. Wholesale Prices

Year	Index	Year	Index	Year	Index	Year	Index
1720	16.98	1761	25.48	1802	39.04	1843	25.06
1721	15.48	1762	28.42	1803	39.29	1844	25.20
1722	16.07	1763	27.43	1804	43.30	1845	27.47
1723	16.60	1764	24.85	1805	47.84	1846	27.62
1724	17.51	1765	24.41	1806	43.94	1847	31.99
1725	19.03	1766	25.23	1807	41.85	1848	28.02
1726	19.89	1767	25.82	1808	37.01	1849	27.79
1727	19.22	1768	24.90	1809	42.44	1850	28.17
1728	18.28	1769	25.01	1810	45.03	1851	27.77
1729	18.22	1770	25.63	1811	45.12	1852	29.37
1730	19.30	1771	26.50	1812	45.77	1853	32.71
1731	17.16	1772	29.67	1813	53.76	1854	38.34
1732	16.47	1773	28.07	1814	65.06	1855	40.81
1733	17.73	1774	26.08	1815	59.92	1856	36.38
1734	17.18	1775	25.74	1816	53.12	1857	37.77
1735	17.29	1776	29.51	1817	53.95	1858	31.56
1736	16.47	1777	42.21	1818	50.48	1859	32.97
1737	17.94	1778	48.04	1819	41.86	1860	31.48
1738	17.94	1779	77.55	1820	34.70	1861	31.11
1739	16.19	1780	77.21	1821	32.07	1862	36.35
1740	17.20	1781	74.12	1822	35.02	1863	46.49
1741	22.18	1782	57.44	1823	34.14	1864	67.47
1742	21.33	1783	44.52	1824	31.76	1865	64.67
1743	18.83	1784	34.50	1825	32.62	1866	60.82
1744	17.90	1785	31.58	1826	31.67	1867	56.63
1745	18.26	1786	29.94	1827	31.62	1868	55.23
1746	19.64	1787	28.99	1828	31.84	1869	52.78
1747	21.78	1788	25.73	1829	32.35	1870	47.19
1748	24.56	1789	27.45	1830	29.43	1871	45.44
1749	23.93	1790	31.48	1831	31.38	1872	47.54
1750	21.69	1791	29.51	1832	31.69	1873	46.49
1751	22.57	1792	30.96	1833	32.12	1874	44.04
1752	22.65	1793	34.60	1834	30.50	1875	41.25
1753	22.28	1794	36.63	1835	33.97	1876	38.45
1754	22.20	1795	44.43	1836	39.69	1877	37.05
1755	22.19	1796	50.25	1837	41.33	1878	31.81
1756	22.43	1797	45.20	1838	38.45	1879	31.46
1757	22.00	1798	42.09	1839	38.11	1880	34.96
1758	23.10	1799	43.47	1840	31.63	1881	36.00
1759	26.21	1800	44.51	1841	29.87	1882	37.75
1760	26.28	1801	48.99	1842	26.62	1883	35.31

Table 1. (Continued).

Year	Index	Year	Index	Year	Index	Year	Index
1884	32.51	1908	32.08	1932	33.16	1956	90.70
1885	29.71	1909	34.48	1933	33.63	1957	93.30
1886	28.66	1910	35.91	1934	38.21	1958	94.60
1887	29.71	1911	33.10	1935	40.84	1959	94.80
1888	30.06	1912	35.24	1936	41.24	1960	94.90
1889	28.31	1913	35.60	1937	44.03	1961	94.50
1890	28.66	1914	34.73	1938	40.09	1962	94.80
1891	28.46	1915	35.45	1939	39.36	1963	94.50
1892	26.62	1916	43.61	1940	40.09	1964	94.70
1893	27.24	1917	59.93	1941	44.59	1965	96.60
1894	24.43	1918	66.97	1942	50.39	1966	99.80
1895	24.89	1919	70.69	1943	52.67	1967	100.00
1896	23.72	1920	78.75	1944	53.05	1968	102.50
1897	23.77	1921	49.78	1945	54.01	1969	106.50
1898	24.74	1922	49.32	1946	61.72	1970	110.40
1899	26.62	1923	51.31	1947	76.65	1971	113.90
1900	28.61	1924	50.03	1948	83.08	1972	119.10
1901	28.20	1925	52.79	1949	78.48	1973	135.50
1902	30.04	1926	51.00	1950	81.68		
1903	30.40	1927	48.66	1951	91.10		
1904	30.45	1928	49.32	1952	88.60		
1905	30.65	1929	48.60	1953	87.40		
1906	31.52	1930	44.11	1954	87.60		
1907	33.25	1931	37.23	1955	87.80		

Source: Composite derived from: Wholesale Prices for Philadelphia, 1720 to 1861, Series E82, U.S. Bureau of the Census (1960); Wholesale Prices, All Commodities, 1749 to 1890, Series E1, U.S. Bureau of the Census (1960); Wholesale Prices, All Commodities, 1890 to 1951, Series E13, U.S. Bureau of the Census (1960); Wholesale Prices, All Commodities, 1929 to 1971, Office of the President (1972); Wholesale Prices, All Commodities, 1929 to 1973, Office of the President (1974).

The transformation matrix $\hat{P}$ based on the autoregressive assumption is computed as follows. Write the model in vector form

$$y = f(\theta^{o}) + u \ ;$$

compute the least squares estimator $\hat{\theta}$, which minimizes

$$SSE(\theta) = [y - f(\theta)]'[y - f(\theta)] \ ;$$

compute the residuals

$$\hat{u} = y - f(\hat{\theta}) \ ;$$

from these, estimate the autocovariances up to lag q using

$$\hat{\gamma}(h) = (1/n)\sum_{t=1}^{n-|h|} \hat{u}_t \hat{u}_{t+|h|} \ , \qquad h = 0, 1, \ldots, q;$$

put

$$\hat{\Gamma}_q = \begin{bmatrix} \hat{\gamma}(0) & \hat{\gamma}(1) & \ldots & \hat{\gamma}(q-1) \\ \hat{\gamma}(1) & \hat{\gamma}(0) & \ldots & \hat{\gamma}(q-2) \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ \hat{\gamma}(q-1) & \hat{\gamma}(q-2) & \ldots & \hat{\gamma}(0) \end{bmatrix} \qquad (q \times q)$$

$$\hat{\gamma}_q = [\hat{\gamma}(1), \hat{\gamma}(2), \ldots, \hat{\gamma}(q)]' \qquad (q \times 1)$$

and compute $\hat{a}$ using the Yule-Walker equations

$$\hat{a} = -\hat{\Gamma}_q^{-1} \gamma_q \qquad (q \times 1)$$

$$\hat{\sigma}^2 = \gamma(0) + \hat{a}'\hat{\gamma}_q \qquad (1 \times 1) \ ;$$

factor

$$\hat{\Gamma}_q^{-1} = \hat{P}_q'\hat{P}_q$$

using, for example, Cholesky's method and set

$$\hat{P} = \left[\begin{array}{c|c} \sqrt{\sigma^2}\,\hat{P}_q & 0 \\ \hline \begin{matrix} \hat{a}_q & \hat{a}_{q-1} & \cdots & \hat{a}_1 & 1 \end{matrix} & \\ & \\ \begin{matrix} \hat{a}_q & \hat{a}_{q-1} & \cdots & \hat{a}_1 & 1 \end{matrix} & \\ \vdots & \\ \begin{matrix} \hat{a}_q & \hat{a}_{q-1} & \cdots & \hat{a}_1 & 1 \end{matrix} & \end{array}\right] \begin{matrix} \} \text{ q rows} \\ \\ \} \text{ n-q rows} \end{matrix}$$

As discussed above, $\hat{P}$ is used to transform the model $y = f(\theta) + u$ to the model "y" = "f"(θ) + "e" using

$$\text{"y"} = \hat{P}y, \quad \text{"f"}(\theta) = \hat{P}f(\theta), \quad \text{"e"} = \hat{P}u\ .$$

Thereafter the methods discussed in Chapter 1 are applied to "y" and "f"(θ). This includes computation of the estimator, methods for testing hypotheses, and for setting confidence intervals. The value $\hat{\theta}$ left over from the least squares estimation usually is an excellent starting value for the iterations to minimize

$$\text{"SSE"}(\theta) = [\text{"y"} - \text{"f"}(\theta)]'[\text{"y"} - \text{"f"}(\theta)].$$

Note that after the first q observations,

$$\text{"y"}_t = y_t + \hat{a}_1 y_{t-1} + \hat{a}_2 y_{t-2} + \ldots + \hat{a}_q y_{t-q} \qquad t = q+1,\ q+2,\ \ldots,\ n$$

$$\text{"f"}(x_t,\theta) =$$

$$f(x_t,\theta) + \hat{a}_1 f(x_{t-1},\theta) + \ldots + \hat{a}_q f(x_{t-q},\theta) \qquad t = q+1,\ q+2,\ \ldots,\ n$$

which is particularly easy to code. We illustrate with the example.

EXAMPLE 1 (continued). Code to compute $\hat{\theta}$ for the model

$$y_t = \theta_1 e^{\theta_2 \cdot t} + u_t\ , \qquad t = 1,\ 2,\ \ldots,\ n = 254$$

$$u_t + a_1 u_{t-1} + a_2 u_{t-2} = e_t, \qquad t = 0,\ \pm 1,\ \pm 2,\ \ldots$$

using the data of Table 1 is shown in Figure 3.

Most of the code has to do with the organization of data sets. First, the least squares estimator $\hat{\theta}$ is computed PROC NLIN and the residuals from the fit are stored in WORK01. The purpose of the code which follows is to arrange these residuals so that the data set WORK03 has the form

"y"	"X"	
$\hat{u}$	0	0
0	$\hat{u}$	0
0	0	$\hat{u}$

whence familiar regression formulas can be used to compute

$$\hat{\Gamma}_q = (1/n)X'X \qquad \hat{a} = -(X'X)^{-1} X'y \qquad \hat{\sigma}^2 = y'y - \hat{a}'X'X\hat{a}$$

$$= \begin{bmatrix} 252.32 & 234.35 \\ 234.35 & 252.32 \end{bmatrix} \qquad = \begin{bmatrix} 213.20 \\ 234.35 \end{bmatrix} \qquad = 34.0916$$

using PROC MATRIX. Note that the columns of "X" have been permuted to permute

the columns of $\hat{a}$ in this code. The transformation $\hat{P}$ is put in the data set WORK04 whose first two rows contain $(\hat{P}_q|0)$ and remaining rows contain $(\hat{a}_2,\hat{a}_1,1)$. The transformation is merged with the data, and lagged values of the data, and stored in the data set WORK06. The appearance of the data set WORK06 is as follows:

OBS	COL1	COL2	COL3	Y1	Y2	Y3	X1	X2	X3
1	0.991682	-0.9210	0	16.98	15.48	16.07	1	2	3
2	0.000000	0.3676	0	16.98	15.48	16.07	1	2	3
3	0.128712	-1.0483	1	16.98	15.48	16.07	1	2	3
4	0.128712	-1.0483	1	15.48	16.07	16.60	2	3	4
•									
•									
•									
254	0.128712	-1.0483	1	113.90	119.10	135.50	252	253	254

Using PROC NLIN one obtains

$$\hat{\theta} = (12.1975,\ 0.00821720)'$$

as shown in Figure 3.

Figure 3. Example 1 Estimated Using an Autoregressive Transformation.

SAS Statements:

```
PROC NLIN DATA=EG01 METHOD=GAUSS ITER=50 CONVERGENCE=1.E-10;
OUTPUT OUT=WORK01 RESIDUAL=UHAT;
PARMS T1=1 T2=.003;
MODEL Y=T1*EXP(T2*X); DER.T1=EXP(T2*X); DER.T2=T1*X*EXP(T2*X);

DATA WORK02; SET WORK01; KEEP UHAT; OUTPUT;
IF _N_=254 THEN DO; UHAT=0; OUTPUT; OUTPUT; END;
DATA WORK03; SET WORK02;
UHAT_0=UHAT; UHAT_1=LAG1(UHAT); UHAT_2=LAG2(UHAT);
IF _N_=1 THEN DO; UHAT_1=0; UHAT_2=0; END; IF _N_=2 THEN UHAT_2=0;

PROC MATRIX;
FETCH Y DATA=WORK03(KEEP=UHAT_0); FETCH X DATA=WORK03(KEEP=UHAT_2 UHAT_1);
GG=X'*X#/254; G=X'*Y#/254; A=-INV(GG)*G; SS=Y'*Y#/254-A'*GG*A;
SPQ=HALF(SS#INV(GG));
ZERO=0/0; ONE=1; P=SPQ||ZERO; ROW=A'||ONE; DO I=1 TO 252; P=P//ROW; END;
OUTPUT P OUT=WORK04;

DATA WORK05; SET EG01;
Y3=Y; Y2=LAG1(Y); Y1=LAG2(Y); X3=X; X2=LAG1(X); X1=LAG2(X);
IF _N_=3 THEN DO; OUTPUT; OUTPUT; END; IF _N_>2 THEN OUTPUT;
DATA WORK06; MERGE WORK04 WORK05; DROP ROW Y X;

PROC NLIN DATA=WORK06 METHOD=GAUSS ITER=50 CONVERGENCE=1.E-10;
PARMS T1=1 T2=.003;
Y = COL1*Y1 + COL2*Y2 + COL3*Y3;
F = COL1*T1*EXP(T2*X1) + COL2*T1*EXP(T2*X2) +COL3*T1*EXP(T2*X3);
D1 = COL1*EXP(T2*X1) + COL2*EXP(T2*X2) +COL3*EXP(T2*X3);
D2 = COL1*T1*X1*EXP(T2*X1) + COL2*T1*X2*EXP(T2*X2) +COL3*T1*X3*EXP(T2*X3);
MODEL Y = F; DER.T1 = D1; DER.T2 = D2;
```

Output:

SAS 5

NON-LINEAR LEAST SQUARES SUMMARY STATISTICS DEPENDENT VARIABLE Y

SOURCE	DF	SUM OF SQUARES	MEAN SQUARE
REGRESSION	2	4428.02183844	2214.01091922
RESIDUAL	252	5656.56502716	22.44668662
UNCORRECTED TOTAL	254	10084.58686559	

PARAMETER	ESTIMATE	ASYMPTOTIC STD. ERROR	ASYMPTOTIC 95 % CONFIDENCE INTERVAL LOWER	UPPER
T1	12.19756397	3.45880524	5.38562777	19.00950018
T2	0.00821720	0.00133383	0.00559029	0.01084410

A word of caution. This example is intended to illustrate the computations, not to give statistical guidance. Specifically, putting $x_t = t$ violates the regularity conditions of the asymptotic theory and visual inspection of Figure 1 suggests a lack of stationarity as the variance seems to be growing with time.∎

Monte-Carlo simulations reported in Gallant and Goebel (1976) suggest that the efficiency gains, relative to least squares, using this procedure can be substantial. They also suggest that the probability statements associated to hypothesis and confidence intervals are not as accurate as one might hope but are certainly an improvement over least squares probability statements. These fact hold true whether the series $\{u_t\}_{t=-\infty}^{\infty}$ that generates the data is actually an autoregressive process of order q or some other covariance stationary process such as a moving average process that can be approximated by an autoregression

The order q of the autoregressive process which best approximates the error process $\{u_t\}$ is unknown in applications. One approach is to attempt to determine q from the least squares residuals $\{\hat{u}_t\}$.

This problem is very much analogous to the problem of determining the appropriate degree of polynomial to use in polynomial regression analysis. The correct analogy is obtained by viewing $\hat{\Gamma}_q a = -\hat{\gamma}_q$ as the normal equations with solution vector $\hat{a} = -\hat{\Gamma}_q$ and residual mean square $s^2 = [\hat{\gamma}(0) + \hat{a}'\hat{\gamma}_q]/(n-q)$. The hypotheses $H: a_i = 0$ against $A: a_i \neq 0$, $i = 1, 2, \ldots, q$, may be tested using $t_i = |\hat{a}_i|/\sqrt{(s^2\hat{\gamma}^{ii})}$ where $\hat{\gamma}^{ii}$ is the i^{th} diagonal element of $\hat{\Gamma}_q^{-1}$ by entering tables of t with n - q degrees of freedom. Standard techniques of degree determination in polynomial regression analysis may be employed of which two are: test sequentially upward, or start from a very high order and test downward (Anderson, 1971, Section 3.2.2).

Akaike's (1969) method is a variant on the familiar procedure of plotting the residual mean square against the degree of the fitted polynomial and terminating when the curve flattens. Akaike plots

$$FPE = [1 + (q + p)/n][1/(n - q - p)] \sum_{t=1}^{n}(\hat{u}_t + \sum_{j=1}^{q}\hat{a}_j\hat{u}_{t-j})^2$$

against q for all q less than an *a priori* upper bound; in this computation put $\hat{u}_0, \hat{u}_{-1}, \ldots, \hat{u}_{1-q} = 0$. That q at which the minimum obtains is selected as the order of the approximating autoregressive process.

The methods discussed above are appropriate if the error process $\{u_t\}$ is covariance stationary. If there is some evidence to the contrary, and a transformation such as discussed in the previous section will not induce stationarity, then an alternative approach is called for. The easiest is to make no attempt at efficiency gains as above but simply correct the standard errors of least squares estimators and let it go at that. The method is as follows.

As above, let $\hat{\theta}$ denote the least squares estimator, the value that minimizes

$$s_n(\theta) = (1/n)\sum_{t=1}^{n}[y_t - f(x_t,\theta)]^2,$$

and let $\hat{u}_t$ denote residuals

$$\hat{u}_t = y_t - f(x_t,\hat{\theta}) \qquad t = 1, 2, \ldots, n.$$

Upon application of the results of Section 4 of Chapter 9, approximately -- see Theorem 6 of Chapter 9 for an exact statement --

$$\sqrt{n}(\hat{\theta} - \theta^o) \sim N_p(0,V)$$

with

$$V = \mathcal{J}^{-1} \mathcal{I} \mathcal{J}^{-1}.$$

$\mathcal{I}$ and $\mathcal{J}$ can be estimated using

$$\hat{\mathcal{I}} = \sum_{\tau=-\ell(n)}^{\ell(n)} w[\tau/\ell(n)]\ \hat{\mathcal{I}}_{n\tau}$$

and

$$\hat{\mathcal{J}} = (1/n)\sum_{t=1}^{n}[(\partial/\partial\theta)f(x_t,\hat{\theta})][(\partial/\partial\theta)f(x_t,\hat{\theta})]'$$

where $\ell n(n)$ is the integer nearest $(n)^{1/5}$,

$$w(x) = \begin{cases} 1 - 6|x|^2 + 6|x|^3 & 0 \le x \le 1/2 \\ 2(1 - |x|)^3 & 1/2 \le x \le 1 \end{cases}$$

and

$$\mathcal{I}_{n\tau} = \begin{cases} (1/n)\ \sum_{t=1+\tau}^{n}[\hat{u}_t(\partial/\partial\theta)f(x_t,\hat{\theta})][\hat{u}_{t-\tau}(\partial/\partial\theta)f(x_{t-\tau},\hat{\theta})]' & \tau \ge 0 \\ (\hat{\mathcal{I}}_{n,-\tau})' & \tau < 0 \end{cases}$$

For testing

$$H: h(\theta^o) = 0 \quad \text{against} \quad A: h(\theta^o) \neq 0$$

where $h: \mathbb{R}^p \to \mathbb{R}^q$ the Wald test statistic is (Theorem 12, Chapter 9)

$$W = nh'(\hat{\theta})[\hat{H}\hat{V}\hat{H}']^{-1}h(\hat{\theta})$$

where $\hat{H} = (\partial/\partial\theta')h(\hat{\theta})$ and $\hat{V} = \hat{\mathcal{J}}^{-1}\hat{\mathcal{I}}\hat{\mathcal{J}}^{-1}$. The null hypothesis H: $h(\theta^o) = 0$ is rejected in favor of the alternative hypothesis A: $h(\theta^o) \neq 0$ when the test statistic exceeds the upper $\alpha \times 100$ percentage point χ^2_α of a chi-square random variable with q degrees of freedom; $\chi^2_\alpha = (\chi^2)^{-1}(1-\alpha,q)$.

As a consequence of this result, a 95% confidence interval on the i^{th} element of θ^o is computed as

$$\hat{\theta}_i \pm z_{.025}\sqrt{(\hat{V}_{ii})}/\sqrt{n}$$

where $z_{.025} = -\sqrt{(\chi^2)^{-1}(.95,1)} = N^{-1}(.025|0,1)$.

Let $\tilde{\theta}$ denote the minimizer of $s_n(\theta)$, subject to the restriction that $h(\theta) = 0$. Let $\tilde{H}$, $\tilde{\mathcal{J}}$ and $\tilde{\mathcal{I}}$ denote the formulas for $\hat{H}$, $\hat{\mathcal{J}}$ and $\hat{\mathcal{I}}$ above but with $\tilde{\theta}$ replacing $\hat{\theta}$ throughout; put

$$\tilde{V} = \tilde{\mathcal{J}}^{-1}\tilde{\mathcal{I}}\,\tilde{\mathcal{J}}^{-1}.$$

The Lagrange multiplier test statistic is (Theorem 16, Chapter 9)

$$R = n[(\partial/\partial\theta)s_n(\tilde{\theta})]'\tilde{\mathcal{J}}^{-1}\tilde{H}'(\tilde{H}\tilde{V}\tilde{H}')^{-1}\tilde{H}\tilde{\mathcal{J}}^{-1}[(\partial/\partial\theta)s_n(\tilde{\theta})] .$$

Again, the null hypothesis H: $h(\theta^o) = 0$ is rejected in favor of the alternative hypothesis A: $h(\theta^o) \neq 0$ when the test statistic exceeds the upper $\alpha \times 100$ percentage point χ^2_α of a chi-square random variable with q degrees of freedom; $\chi^2_\alpha = (\chi^2)^{-1}(1-\alpha,q)$.

The likelihood ratio test cannot be used because $\mathcal{J} \neq \mathcal{I}$; see Theorem 17 of Chapter 9. Formulas for computing the power of the Wald and Lagrange multiplier tests are given in Theorems 14 and 16 of Chapter 9, respectively.

3. TESTING A NONLINEAR SPECIFICATION

Often, it is helpful to be able to choose between two model specifications:

$$H: y_t = g(x_t,\psi) + e_t$$

and

$$A: y_t = g(x_t,\psi) + \tau h(x_t,\omega) + e_t \ .$$

The unknown parameters are ψ, τ, and ω of dimension u, q, and v, respectively. The functional forms, $g(x,\psi)$ and $g(x,\omega)$, are known. The errors, e_t, are normally and independently distributed with mean zero and unknown variance σ^2. Parametrically, the situation is equivalent to testing:

$$H: \tau = 0 \text{ against } A: \tau \neq 0$$

regarding ψ, ω, and σ^2 as nuisance parameters.

It would be natural to employ one of the tests discussed in Section 5 of the previous chapter. In the formal sense, the Lagrange multiplier test is undefined because ω cannot be estimated if $\tau = 0$. The likelihood ratio test is defined in the sense that the residual sum of squares from the model

$$H: y_t = g(x_t,\psi\ ) + e_t$$

can be computed and used as $SSE(\tilde{\theta})$. But one must also compute the unconstrained estimate of

$$\theta = (\psi, \omega, \tau)$$

to obtain $SSE(\hat{\theta})$ in order to compute the likelihood ratio test statistic; the Wald test statistic also requires computation of $\hat{\theta}$. When H is true, this dependence on $\hat{\theta}$ causes two difficulties.

1. It is likely that the attempt to fit the full model will fail or, at best converge very slowly as seen in Figure 6 and Table 4 of Chapter 1.

2. The regularity conditions used to obtain the asymptotic properties of the unconstrained least-squares estimator $\hat{\theta}$ -- and also of test statistics that depend on $\hat{\theta}$ -- are violated as neither the identification condition or the rank condition discussed in Section 3 of the previous chapter are satisfied.

It is useful to consider when the situation of testing H against A using data which support H is likely to arise. It is improbable that one would attempt to fit a nonlinear model which is not supported by the data if one is merely attempting to represent data parametrically without reference to a substantive problem. For example, in the cases considered in Table 4 of Chapter 1, plots of the observed response y_t against the input x_{3t} failed to reveal any visual impression of exponential growth for values of $|\theta_4|$ less than .1. Consequently, substantive rather than data analytic considerations will likely have suggested A. As we shall see, it will be helpful if these same substantive considerations also imply probable values for ω.

The lack-of-fit test has been discussed by several authors (Beale, 1960; Halperin, 1963; Hartley, 1964; Turner, Monroe, and Lucas, 1961; Williams, 1962) in the context of finding exact tests or confidence regions in nonlinear regression analysis. Here the same idea is employed, but an asymptotic theory is substituted for an exact small-sample theory. The basic idea is straightforward: If $\tau^o = 0$, then the least-squares estimator of the parameter δ in the model

$$\hat{A}: \; y_t = g(x_t,\Psi) + z_t'\delta + e_t$$

where the w-vector z_t does not depend on any unknown parameters, is estimating

the zero vector. Thus any (asymptotically) level α test of

$$H\colon\ y_t = g(x_t,\psi) + e_t \quad \text{against} \quad \hat{A}\colon\ y_t = g(x_t,\psi) + z_t'\delta + e_t$$

is a (asymptotically) level-α test of

$$H\colon\ y_t = g(x_t,\psi) + e_t \quad \text{against} \quad A\colon\ y_t = g(x_t,\psi) + \tau h(x_t,\omega) + e_t.$$

Note that since z_t does not depend on any unknown parameters, the computational problems that arise when trying to fit A by least squares will not arise when fitting $\hat{A}$.

When H is true, any of the tests considered in Section 4 of Chapter 1 are asymptotically level α. Regularity conditions such that the Wald and likelihood ratio test statistics for H against $\hat{A}$ follow the noncentral F-distribution plus an asymptotically negligible remainder term when A is true are in Gallant (1977). Simulations reported in Gallant (1977) suggest that the problem of inaccurate probability statements associated to inference based on the Wald test statistic are exacerbated in the present circumstance. The simulations support the use of the likelihood ratio test; the Lagrange multiplier text was not considered.

Let $(\hat{\psi},\hat{\delta})$ denote the least-squares estimator for the model $\hat{A}$ and define:

$$SSE(\hat{\psi},\hat{\delta}) = \sum_{t=1}^{n}[y_t - g(x_t,\hat{\psi}) + z_t'\hat{\delta}]^2,$$

$$G(\psi) = \text{the n by u matrix with typical element } (\partial/\partial\psi_j)g(x_t,\psi)\ ,$$

$$Z = \text{the n by w matrix with } t^{th}\text{row } z_t',$$

$$Q_Z = I - G(\psi^o)[G'(\psi^o)G(\psi^o)]^{-1}G'(\psi^o)$$

Let $\tilde{\psi}$ denote the least-squares estimator for the model H, and define

$$SSE(\tilde{\psi}) = \sum_{t=1}^{n}[y_t - g(x_t,\tilde{\psi})]^2.$$

The likelihood ratio test for H against $\hat{A}$ rejects when

$$L = \frac{[SSE(\tilde{\psi}) - SSE(\hat{\psi},\hat{\delta})]/w}{SSE(\hat{\psi},\hat{\delta})]/(n - u - w)}$$

exceeds F_α, the upper $\alpha \cdot 100$ percentage point an F-random variable with w numerator degrees of freedom and n - u - w denominator degrees of freedom; $F_\alpha = F^{-1}(1-\alpha;w,n-u-w)$.

The objective governing the choice of the vector z_t of additional regressors is to find those which will maximize the power of the test of H against $\hat{A}$ when A is true. The asymptotic power of the likelihood ratio test is given by the probability that a doubly noncentral F statistic exceeds F_α (Gallant,1977). The noncentrality parameters of this statistic are

$$\lambda_1 = (\tau^o)^2 h'Q_G Z(Z'Q_G'Q_G Z)^{-1}Z'Q_G'h/(2\sigma^2)$$

for the numerator, and

$$\lambda_2 = (\tau^o)^2 h'Q_G h/(2\sigma^2) - \lambda_1$$

for the denominator, where

$$h = [h(x_1,\omega^o),\ h(x_2,\omega^o),\ \ldots,\ h(x_n,\omega^o)]' .$$

Thus one should attempt to find those z_t which best approximate h in the sense of maximizing the ratio

$$h'Q_G Z(Z'Q_G'Q_G Z)^{-1}Z'Q_G'h/h'Q_G h$$

while attempting, simultaneously, to keep the number of columns of Z as small as possible. We consider, next, how this might be done in applications.

In a situation where substantive considerations or previous experimental evidence suggest a single point estimate $\hat{\omega}$ for ω^o, the natural choice is $z_t = h(x_t, \hat{\omega})$.

If, instead of a point estimate, ranges of plausible values for the components of ω are available then a representative selection of values of ω,

$$\{\hat{\omega}_i : i = 1, 2, \ldots, K\}$$

whose components fall within these ranges can be chosen -- either deterministically or by random sampling from a distribution defined on the plausible values -- and the vectors $h(\hat{\omega}_i)$ made the columns of Z. If, following this procedure, the number of columns of Z is unreasonably large, it may be reduced as follows. Decompose the matrix

$$B = [h(\hat{\omega}_1) | \ldots | h(\hat{\omega}_K)]$$

into its principal component vectors, and choose the first few of these to make up Z; equivalently, obtain the singular value decomposition (Bussinger and Golub, 1969) B = USV' where U'U = V'V = VV' = I, and S is diagonal with nonnegative entries , and choose the first few columns of U to make up Z. We illustrate with an example.

EXAMPLE 2 (Preschool Boys' Weight/Height Ratio). The data shown in Figure 4 are preschool boys' weight/height ratios plotted against age and were obtained from Eppright *et al.* (1972); the tabular values are shown in Table 2. The question is whether the data support the choice of a three-segment quadratic-quadratic-linear polynomial response function as opposed to a two-segment

FIGURE 4. PRESCHOOL BOY'S WEIGHT/HEIGHT RATIO

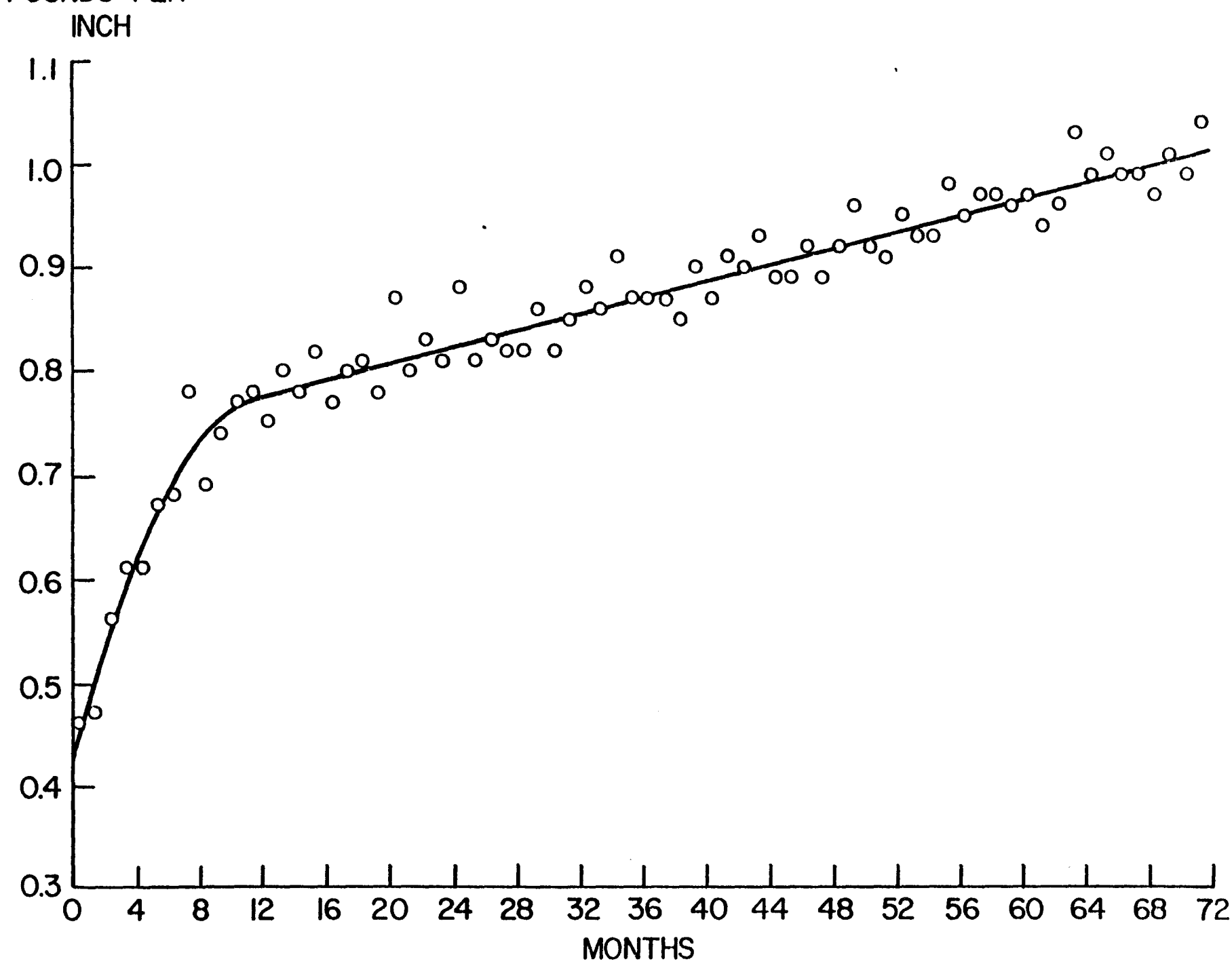

Table 2. Boys' Weight/Height vs Age

W/H	Age	W/H	Age	W/H	Age
0.46	0.5	0.88	24.5	0.92	48.5
0.47	1.5	0.81	25.5	0.96	49.5
0.56	2.5	0.83	26.5	0.92	50.5
0.61	3.5	0.82	27.5	0.91	51.5
0.61	4.5	0.82	28.5	0.95	52.5
0.67	5.5	0.86	29.5	0.93	53.5
0.68	6.5	0.82	30.5	0.93	54.5
0.78	7.5	0.85	31.5	0.98	55.5
0.69	8.5	0.88	32.5	0.95	56.5
0.74	9.5	0.86	33.5	0.97	57.5
0.77	10.5	0.91	34.5	0.97	58.5
0.78	11.5	0.87	35.5	0.96	59.5
0.75	12.5	0.87	36.5	0.97	60.5
0.80	13.5	0.87	37.5	0.94	61.5
0.78	14.5	0.85	38.5	0.96	62.5
0.82	15.5	0.90	39.5	1.03	63.5
0.77	16.5	0.87	40.5	0.99	64.5
0.80	17.5	0.91	41.5	1.01	65.5
0.81	18.5	0.90	42.5	0.99	66.5
0.78	19.5	0.93	43.5	0.99	67.5
0.87	20.5	0.89	44.5	0.97	68.5
0.80	21.5	0.89	45.5	1.01	69.5
0.83	22.5	0.92	46.5	0.99	70.5
0.81	23.5	0.89	47.5	1.04	71.5

quadratic-linear response function. In both cases, the response function is required to be continuous in x (age) and to have a continuous first derivative in x. Formally,

$$H: y_t = \theta_1 + \theta_2 x_t + \theta_3 T_2(\theta_4 - x_t) + e_t ,$$

and

$$A: y_t = \theta_1 + \theta_2 x_t + \theta_3 T_2(\theta_4 - x_t) + \theta_5 T_2(\theta_6 - x_t) + e_t,$$

where

$$T_k(z) = \begin{cases} z^k & \text{when } z \geq 0, \\ 0 & \text{when } z \leq 0; \end{cases}$$

see Gallant and Fuller (1973) for a discussion of the derivation and fitting of grafted polynomial models.

The correspondence with the notation above is:

$$\psi = (\theta_1, \theta_2, \theta_3, \theta_4)',$$

$$\tau = \theta_5,$$

$$\omega = \theta_6,$$

$$g(x,\psi) = \psi_1 + \psi_2 x + \psi_3 T_2(\psi_4 - x),$$

$$h(x,\omega) = T_2(\omega - x).$$

The parameter ω is the join point associated with the quadratic term whose omission is proposed.

Suppose plausible values for ω are $\hat{\omega}_1 = 4$, $\hat{\omega}_2 = 8$, and $\hat{\omega}_3 = 12$. The matrix B, described above, has typical row

$$B_t = [T_2(4 - x_t),\ T_2(8 - x_t),\ T_2(12 - x_t)].$$

The first principal component vector of B, with elements

$$z_t = [(2.08)T_2(4 - x_t) + (14.07)T_2(8 - x_t) + (39.9)T_2(12 - x_t)]10^{-4}$$

was chosen as the additional regressor. This choice yields

$$SSE(\tilde{\psi}) = 0.03789865 \qquad \text{(from Figure 5)},$$

$$SSE(\hat{\psi},\hat{\delta}) = 0.03769031 \qquad \text{(from Figure 5)},$$

$$L = \frac{(0.03789865 - 0.03769031)/1}{0.03769031/(72 - 4 - 1)}$$

$$= 0.370,$$

$$P[F(1,67) > 0.370] \doteq .485.$$

These data give little support to A.

Simulations reported by Gallant (1977) indicate that the best choice for z_t in this example is to take as Z the first principal component of B. For practical purposed, the power of the test is as good as if the true value ω^o were known.

Figure 5. Lack of Fit Test Illustrated Using Example 2.

SAS Statements:

```
DATA B; SET EG02;
Z1=(4-AGE>0)*(4-AGE)**2; Z2=(8-AGE>0)*(8-AGE)**2; Z3=(12-AGE>0)*(12-AGE)**2;
PROC MATRIX; FETCH B DATA=B(KEEP=Z1 Z2 Z3); SVD U Q V B; OUTPUT U OUT=WORK01;
DATA WORK02; MERGE EG02 WORK01;  KEEP AGE WH Z; Z=COL1;

PROC NLIN DATA=WORK02 METHOD=GAUSS ITER=50 CONVERGENCE=1.E-10;
PARMS T1=1 T2=.004 T3=-.002 T4=12; X=(T4-AGE>0)*(T4-AGE);
MODEL WH = T1+T2*AGE+T3*X**2;
DER.T1=1; DER.T2=AGE; DER.T3=X**2; DER.T4=2*T3*X;

PROC NLIN DATA=WORK02 METHOD=GAUSS ITER=50 CONVERGENCE=1.E-10;
PARMS T1=.73 T2=.004 T3=-5.E-5 T4=21.181 D=-.4; X=(T4-AGE>0)*(T4-AGE);
MODEL WH = T1+T2*AGE+T3*X**2+Z*D;
DER.T1=1; DER.T2=AGE; DER.T3=X**2; DER.T4=2*T3*X; DER.D=Z;
```

Output:

SAS 3

NON-LINEAR LEAST SQUARES SUMMARY STATISTICS DEPENDENT VARIABLE WH

SOURCE	DF	SUM OF SQUARES	MEAN SQUARE
REGRESSION	4	53.67750135	13.41937534
RESIDUAL	68	0.03789865	0.00055733
UNCORRECTED TOTAL	72	53.71540000	

SAS 5

NON-LINEAR LEAST SQUARES SUMMARY STATISTICS DEPENDENT VARIABLE WH

SOURCE	DF	SUM OF SQUARES	MEAN SQUARE
REGRESSION	5	53.67770969	10.73554194
RESIDUAL	67	0.03769031	0.00056254
UNCORRECTED TOTAL	72	53.71540000	

4. MEASURES OF NONLINEARITY

Consider the nonlinear model

$$y = f(\theta^{o}) + e$$

with normally distributed errors. As we have seen, the statistic

$$R_1 = \frac{[y - f(\theta^{*})]'F(\theta^{*})[F'(\theta^{*})F(\theta^{*})]^{-1}F'(\theta)[y - f(\theta^{*})]/p}{SSE(\theta^{*})/(n - p)}$$

is distributed exactly as an F with p numerator degrees of freedom and n - p denominator degrees of freedom when a null hypothesis that completely specifies the parameter,

$$H: \theta^{o} = \theta^{*},$$

is true. Beale (1960) studied the extent to which confidence contours constructed using R_1 coincide with contours constructed using the likelihood ratio test statistic

$$L = \frac{[SSE(\theta^{*}) - SSE(\hat{\theta})]/p}{SSE(\hat{\theta})/(n - p)} .$$

On the basis of this study, he constructed measures of nonlinearity that measure the extent of the coincidence and suggested corrections to critical points based on these measures to improve the accuracy of confidence statements. Coincidence is a sufficient condition for accurate probability statements, not a necessary condition. Thus a large value of Beale's nonlinearity measure does not imply inaccurate probability statements and it is possible that Beale's corrections can actually be counter productive.

Simulations reported by Gallant (1976) indicate that there are instances where such is the case.

Bates and Watts (1980) take a geometric approach in deriving their measures of nonlinearity, an approach somewhat related in spirit to Efron (1975); Ratkowsky (1983) summarizes their work and contains FORTRAN code to compute these measures. The most interesting aspect of their work is that they break their measure into two pieces, one a measure of intrinsic curvature and the other a measure of parameter-effects curvature. The latter can be reduced be reparameterization, the former cannot. In their examples, which are rather extensive, they find that the parameter-effects curvature is more important than the intrinsic in each case.

What is interesting about the work is that it sheds some intuitive light on the question of why the likelihood ratio statistic leads to more accurate probability statements regarding the size of tests and level of confidence sets than the Wald statistic. We have seen that the Wald test is not invariant to nonlinear transformation which means that there may exist a transformation that would make the total nonlinearity nearly equal to the intrinsic nonlinearity and so improve accuracy. The Bates and Watts measure provides some guidance in finding it; see Bates and Watts (1981). The likelihood ratio test is invariant to reparameterization which means that it can be regarded as a statistic where this transformation has been found automatically so that it is only the intrinsic nonlinearity that is operative. This is, of course, rather intuitive and speculative, and suffers from the same defect as was noted above: the measures, like Beale's, represent sufficient conditions, not necessary conditions; see Cook and Witmer (1985) in this regard.

Our advice, above, was to avoid the whole issue as regards inference and simply use the likelihood ratio statistic in preference to the Wald statistic. A reparameterization will probably destroy the principal advantage of the Wald statistic in that it provides ellipsoidal confidence regions on model parameters. After a reparameterization, the ellipsoid will correspond to new parameters that will not necessarily be naturally associated to the problem and one is no better off in this regard than with the likelihood ratio statistic. As regards computations, reparameterization can be very helpful; see Ross (1970).

5. REFERENCES

Anderson, Theodore W. (1971), *The Statistical Analysis of Time Series.* New York: John Wiley and Sons.

Akaike, H. (1969), "Fitting Autoregressive Models for Prediction," *Annals of the Institute of Statistical Mathematics* 21, 243-247.

Berk, K. N. (1974), "Consistent Autoregressive Spectral Estimates," *The Annals of Statistics* 2, 489-502.

Bates, Douglas M. and Donald G. Watts (1980), "Relative Curvature Measures of Nonlinearity," *Journal of the Royal Statistical Society Series B* 42, 1-25.

Bates, Douglas M. and Donald G. Watts (1981), "Parameter Transformations for Improved Approximate Confidence Regions in Nonlinear Least Squares," *The Annals of Statistics* 9, 1152-1167.

Bussinger, Peter A. and Gene H. Golub (1969), "Singular Value Decomposition of a Complex Matrix,' *Communications of the ACM* 12, 564-565.

Beale, E. M. L. (1960), "Confidence Regions in Non-Linear Estimation," *Journal of the Royal Statistical Society Series B* 22, 41-76.

Cook, R. Dennis and Jeffrey A. Witmer (1985), "A Note on Parameter-Effects Curvature," *Journal of the American Statistical Association* 80, 872-878 .

Efron, Bradley (1975), "Defining the Curvature of a Statistical Problem (with Applications to Second Order Efficiency)," *The Annals of Statistics* 3, 1189-1242.

Eppright, E. S., H. M. Fox, B. A. Fryer, G. H. Lamkin, V. M. Vivian, and E. S. Fuller (1972), "Nutrition of Infants and Preschool Children in the North Central Region of the United States of America," *World Review of Nutrition and Dietetics* 14, 269-332.

Fuller, Wayne A. (1976), *Introduction to Statistical Time Series.* New York: John Wiley and Sons.

Gallant, A. Ronald (1976), "Confidence Regions for the Parameters of a Nonlinear Regression Model," Institute of Statistics Mimeograph Series No. 1077, North Carolina State University, Raleigh, NC.

Gallant, A. Ronald (1977), "Testing a Nonlinear Regression Specification: A Nonregular Case," *Journal of the American Statistical Association* 72, 523-530.

Gallant, A. Ronald and Wayne A. Fuller (1973), "Fitting Segmented Polynomial Regression Models Whose Join Points Have to Be Estimated," *Journal of the American Statistical Association* 68, 144-147.

Gallant, A. Ronald and J. Jeffery Goebel (1975), "Nonlinear Regression with Autoregressive Errors," Institute of Statistics Mimeograph Series No. 986, North Carolina State University, Raleigh, NC.

Gallant, A. Ronald and J. Jeffery Goebel (1976), "Nonlinear Regression with Autocorrelated Errors," *Journal of the American Statistical Association* 71, 961-967.

Halperin, Max (1963), "Confidence Interval Estimation in Nonlinear Regression," *Journal of the Royal Statistical Society Series B* 25, 330-333.

Hartley, Herman O. (1964), "Exact Confidence Regions for the Parameters in Nonlinear Regression Laws," *Biometrika* 51, 347-353.

Judge, George G., William E. Griffiths, R. Carter Hill, and Tsoung-Chao Lee (1980), *The Theory and Practice of Econometrics.* New York, John Wiley and Sons.

Office of the President (1972), *Economic Report of the President, Transmitted to Congress, January 1972,* Washington, D.C.: U. S. Government Printing Office, 250.

Office of the President (1974), *Economic Report of the President, Transmitted to Congress, January 1974,* Washington, D.C.: U. S. Government Printing Office, 305.

Pantula, Sastry (1985), "Autoregressive Processes with Several Unit Roots," Institute of Statistics Mimeograph Series No. ????, North Carolina State University, Raleigh, NC.

Ratkowsky, David A. (1983), *Nonlinear Regression Modeling, A Unified Practical Approach.* New York: Marcel Dekker, Inc.

Ross, G. J. S. (1970), "The Efficient Use of Function Minimization in Non-Linear Maximum-Likelihood Estimation," *Applied Statistics* 19, 205-221.

Turner, Malcolm E., Robert J. Monroe, and Henry L. Lucas (1961), "Generalized Asymptotic Regression and Nonlinear Path Analysis," *Biometrics* 17, 120-149.

U.S. Bureau of the Census (1960), *Historical Statistics of the United States,* Washington, D. C.: U. S. Government Printing Office, 116-9.

Williams, E. J. (1962), "Exact Fiducial Limits in Nonlinear Estimation," *Journal of the Royal Statistical Society Series B* 24, 125-139.

6. INDEX