# Score Tests for the Single Index Model

**Jeffrey S. Simonoff**

Leonard N. Stern School of Business
New York University
New York, NY   10012-0258
(*jsimonof@stern.nyu.edu*)

**Chih-Ling Tsai**

Graduate School of Management
University of California at Davis
Davis, CA   95616-8609
(*cltsai@ucdavis.edu*)

The single index model is a generalization of the linear regression model with $E(y|\mathbf{x}) = g(\mathbf{x}'\boldsymbol{\beta})$, where $g$ is an unknown function. The model provides a flexible alternative to the linear regression model while providing more structure than a fully nonparametric approach. Although the fitting of single index models does not require distributional assumptions on the error term, the properties of the estimates depend on such assumptions, as does practical application of the model. In this article score tests are derived for three potential misspecifications of the single index model: heteroscedasticity in the errors, autocorrelation in the errors, and the omission of an important variable in the linear index. These tests have a similar structure to corresponding tests for nonlinear regression models. Monte Carlo simulations demonstrate that the first two tests hold their nominal size well and have good power properties in identifying model violations, often outperforming other tests. Testing for the need for additional covariates can be effective, but is more difficult. The score tests are applied to three real datasets, demonstrating that the tests can identify important model violations that affect inference, and that approaches that do not take model misspecifications into account can lead to very different results.

KEY WORDS:   Autocorrelation; Heteroscedasticity; Local polynomial estimation; Model selection.

## 1.  INTRODUCTION

In recent years the classic linear regression model,

$$y_i = \mathbf{x}_i'\boldsymbol{\beta} + \varepsilon_i \qquad (1.1)$$

with $E(\varepsilon_i|\mathbf{x}_i) = 0$, has been extended and generalized in various ways. One simple, yet extremely useful, generalization is the single index model, where (1.1) is replaced by

$$y_i = g(\mathbf{x}_i'\boldsymbol{\beta}) + \varepsilon_i, \qquad (1.2)$$

where $g$ is an unknown function, often taken to be smooth. The single index model is more flexible than the linear model, because it allows for nonlinear relationships between the index variable $\mathbf{x}'\boldsymbol{\beta}$ and the target variable $y$ while avoiding many of the drawbacks of a fully nonparametric approach. Such drawbacks include decreasing precision with increasing dimension of $\mathbf{x}$ (the so-called curse of dimensionality), difficulty in representing relationships graphically, and the inability to make predictions at points $\mathbf{x}$ outside the support of the observed $\mathbf{x}_i$ values. This model and its variants have been applied successfully to data in fields including economics and medicine; see, e.g., Carroll, Fan, Gijbels, and Wand (1997) and Horowitz (1998).

Model (1.2) can be fit by using a two-step procedure. First, $\boldsymbol{\beta}$ is estimated, and the linear index variable $\mathbf{z} = X\hat{\boldsymbol{\beta}}$ is formed, where $X$ is the matrix of predictor values. Then $g$ is estimated using nonparametric regression as a smooth function based on $\{z_i, y_i\}$, where $z_i = \mathbf{x}_i'\hat{\boldsymbol{\beta}}, i = 1, \ldots, n$. We focus on two possible estimates of $\boldsymbol{\beta}$. Brillinger (1983) showed that if predictors are roughly Gaussian, the ordinary least squares (OLS) estimate based on (1.1) is a $\sqrt{n}$-consistent estimate of $\boldsymbol{\beta}$ in (1.2) up to a constant of proportionality. Note that an intercept term $\beta_0$ is not identifiable in this context, as it can be absorbed into the function $g$. (The OLS estimate is based on centered predictors and includes an intercept, but the linear index does not use the estimated intercept.) A second approach is sliced inverse regression (SIR), proposed by

Duan and Li (1991). This estimate is easy to compute and is $\sqrt{n}$-consistent (up to a proportionality constant) and asymptotically normally distributed. We do not consider here more complicated semiparametric estimators that have been proposed for this problem, such as those of Powell, Stock, and Stoker (1989) and Ichimura (1993), although the tests derived here would be equally applicable to those estimators.

Given $\mathbf{z}$, any nonparametric regression smoother can be used to estimate $g$, yielding fitted values $\hat{y}_i = \hat{g}(z_i)$. Simonoff (1996, chap. 5) discussed various approaches to this problem. We focus here on local polynomial estimation, where a $p$th-order local polynomial estimator $\hat{g}(z)$ is the constant term $\hat{\gamma}_0$ of the minimizer of

$$\sum_{i=1}^{n}\left[y_i - \gamma_0 - \cdots - \gamma_p(z_i - z)^p\right]^2 K\left(\frac{z_i - z}{h}\right), \qquad (1.3)$$

where $K$ is the kernel function, generally taken to be a symmetric probability density function with finite second derivative. In all simulations and examples treated here a local quadratic ($p = 2$) estimate is used. We also need an estimate of the derivative of $g$, $\dot{g}$, which is just the slope term $\hat{\gamma}_1$ of the minimizer of (1.3). The smoothness of $\hat{g}$ is controlled by the smoothing parameter $h$. Hurvich, Simonoff, and Tsai (1998) proposed using a corrected Akaike information criterion (AIC) to choose the smoothing parameter, where $h$ is chosen to minimize

$$\begin{aligned}
\text{AIC}_C &= \log\hat{\sigma}^2 + \frac{1 + \operatorname{tr}(H)/n}{1 - (\operatorname{tr}(H) + 2)/n}\\
&= \log\hat{\sigma}^2 + 1 + \frac{2(\operatorname{tr}(H) + 1)}{n - \operatorname{tr}(H) - 2},
\end{aligned}$$

where $\hat{\sigma}^2 = \sum \hat{e}_i^2/n$ (where $\hat{e}_i = y_i - \hat{y}_i$ is the $i$th element of the residual vector), $H$ satisfies $\hat{\mathbf{y}} = H\mathbf{y}$ and depends on $\mathbf{z}$ and depends on $\mathbf{y}$ only through $\hat{\boldsymbol{\beta}}$, and they showed that this choice is asymptotically optimal for minimizing the integrated squared error of $\hat{g}$. Note that although in theory a different smoothing parameter should be used for estimating $g$ and $\dot{g}$, in the tests constructed here we take the simple strategy of using the same smoothing parameter (the one derived to estimate $g$) for both estimates. This issue is discussed more fully in Section 3.

Although the fitting of single index models does not require distributional assumptions on $\boldsymbol{\varepsilon}$, the properties of the estimates do depend on such assumptions. Härdle, Hall, and Ichimura (1993) derived an optimal smoothing parameter for model (1.2) and showed that the efficiency of $\hat{\boldsymbol{\beta}}$ depends on the distribution of $\boldsymbol{\varepsilon}$ (focusing on an assumption of constant variance of $\varepsilon_i$). Similarly, a modified estimator is appropriate if the errors exhibit autocorrelation. In addition, the existence of violations of model (1.2) affects practical application of the model in, for example, the construction of prediction intervals.

In this article we derive score tests to test for the presence of violations of model (1.2) with $\boldsymbol{\varepsilon} \sim N(\mathbf{0}, \sigma^2 I)$. The principle here is that using the score test can be viewed as a unifying principle in identifying such violations, which can then be addressed using methods adapted to those violations. Score tests have the advantage of requiring fitting only under the null model, thus avoiding the fitting of models that incorporate those violations unless necessary. In addition, in one of the scenarios examined here, a likelihood-ratio-type alternative test is not useful, leaving the score test without apparent competition appropriate for this model.

In the next section a generalized version of (1.2) is formulated; it has as special cases heteroscedasticity, autocorrelation, and additional covariates. Score tests are then derived for each possible violation. Section 3 summarizes the results of Monte Carlo simulations investigating the usefulness of the tests, and several real datasets are analyzed in Section 4. Discussion and recommendations for further work conclude the article.

## 2. DERIVATION OF SCORE TESTS

In all that follows, the notation $g(\mathbf{b})$, where $\mathbf{b}$ is a vector, is used to represent the vector with $i$th entry $g(b_i)$. Consider the generalization of (1.2)

$$\mathbf{y} = g(X_1\boldsymbol{\beta}_1 + X_2\boldsymbol{\beta}_2) + \Sigma^{1/2}\boldsymbol{\varepsilon}, \qquad \boldsymbol{\varepsilon} \sim N(\mathbf{0}, \sigma^2 I), \quad (2.1)$$

where $X_1$ and $X_2$ are an $n \times p_1$ and $n \times p_2$ matrices of predictor values, respectively, $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$ are $p_1 \times 1$ and $p_2 \times 1$ vectors of predictor coefficients, respectively, and

$$\Sigma = \frac{1}{1-\rho^2} \begin{pmatrix} w_1 & \rho & \rho^2 & \cdots & \rho^{n-1} \\ \rho & w_2 & \rho & \cdots & \rho^{n-2} \\ & & & \vdots & \\ \rho^{n-1} & \rho^{n-2} & \cdots & \rho & w_n \end{pmatrix}$$

with $|\rho| < 1$, $w_i = w(\mathbf{u}_i, \boldsymbol{\delta})$, $\mathbf{u}_i'$ the $i$th row of an $n \times k$ matrix $U$, and $\boldsymbol{\delta}$ is a $q \times 1$ vector. Define $\boldsymbol{\delta}_0$ so that $w(\mathbf{u}, \boldsymbol{\delta}_0) = 1$. As shown below, setting parameters to specified values results

in models consistent with heteroscedasticity, first-order autoregressive autocorrelation, and additional covariates.

The general form of the score test [see, for example, Cox and Hinkley (1974, p. 324)] follows a structure similar to that for nonlinear regression models (Seber and Wild 1989, pp. 228–231). Let $L(\boldsymbol{\theta})$ be the log-likelihood function of the observed value of $\mathbf{y}$, where the $m \times 1$ vector $\boldsymbol{\theta}$ is partitioned into $\boldsymbol{\theta} = (\boldsymbol{\theta}_1', \boldsymbol{\theta}_2')'$, with $\boldsymbol{\theta}_1$ being $m_1 \times 1$ and $\boldsymbol{\theta}_2$ being $m_2 \times 1$. Let

$$U_j(\boldsymbol{\theta}) = \frac{\partial L(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}_j}, \qquad j = 1, 2,$$

and

$$I_{ij}(\boldsymbol{\theta}) = E\left[-\frac{\partial^2 L(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}_i \partial \boldsymbol{\theta}_j}\right], \qquad i = 1, 2, \quad j = 1, 2.$$

Note that $I_{21}(\boldsymbol{\theta}) = I_{12}(\boldsymbol{\theta})'$. The score test statistic tests the hypotheses

$$H_0 : \boldsymbol{\theta}_2 = \boldsymbol{\theta}_{20} \qquad (2.2a)$$

versus

$$H_a : \boldsymbol{\theta}_2 \neq \boldsymbol{\theta}_{20} \qquad (2.2b)$$

and is then defined as

$$S = \left[U_2 - I_{21}I_{11}^{-1}U_1\right]'\left[I_{22} - I_{21}I_{11}^{-1}I_{12}\right]^{-1}\left[U_2 - I_{21}I_{11}^{-1}U_1\right],$$

where $U_j$ and $I_{ij}$ are evaluated at $\boldsymbol{\theta}_1 = \hat{\boldsymbol{\theta}}_1(\boldsymbol{\theta}_{20})$ (a $\sqrt{n}$-consistent estimator of $\boldsymbol{\theta}_1$ given $\boldsymbol{\theta}_2 = \boldsymbol{\theta}_{20}$) and $\boldsymbol{\theta}_2 = \boldsymbol{\theta}_{20}$.

In certain circumstances, the asymptotic distribution of the score test can be shown to be $\chi^2$ on $m_2$ degrees of freedom under the null hypothesis. For example, in the nonlinear regression situation where $g$ is known, this is the case (Seber and Wild 1989, p. 198; Tsai 1986). Similar results are sometimes available in the current context and are noted below. In any event, we appeal to this distribution for $S$ as a benchmark here, and we use Monte Carlo simulations in the next section to investigate this approximation.

### 2.1 Test for Heteroscedasticity in Errors

Consider model (2.1) with $\rho = 0$, i.e., $\Sigma = \text{diag}(w_i)$. This model incorporates possible heteroscedasticity through $w_i = w(\mathbf{u}_i, \boldsymbol{\delta})$, because $\text{var}(\varepsilon_i) = w_i\sigma^2$. The values $\mathbf{u}_i$ represent variables that are predictive for $\text{var}(\varepsilon_i)$, which can, but need not, include variables from $X$. To construct the test, a specific functional form for $w$ is required; a typical choice is an exponential function

$$w(\mathbf{u}_i, \boldsymbol{\delta}) = \exp(\mathbf{u}_i'\boldsymbol{\delta}), \qquad (2.3)$$

implying that $\boldsymbol{\delta} = \boldsymbol{\delta}_0 = 0$ corresponds to constant variance [see Simonoff and Tsai (1994) and the references therein]. The score test is based on hypotheses (2.2) with $\boldsymbol{\theta}_1 = (\boldsymbol{\beta}', \sigma^2)'$ and $\boldsymbol{\theta}_2 = \boldsymbol{\delta}$ and equals

$$S_H = \frac{1}{2}\mathbf{c}'\overline{D}(\overline{D}'\overline{D})^{-1}\overline{D}'\mathbf{c}, \qquad (2.4)$$

where $\mathbf{c}$ is the $n \times 1$ vector with $i$th entry $c_i = \hat{e}_i^2/\hat{\sigma}^2$, $\overline{D} = (I - \mathbf{1}\mathbf{1}'/n)D$, $\mathbf{1}$ is a vector of 1's, $\hat{\mathbf{e}} = \mathbf{y} - \hat{g}(X\hat{\boldsymbol{\beta}})$ is the vector of residuals, $\hat{\sigma}^2 = \hat{\mathbf{e}}'\hat{\mathbf{e}}/n$, and $D$ is the $n \times q$ matrix $\partial w(\mathbf{u}_i, \boldsymbol{\delta})/\partial \boldsymbol{\delta}$ evaluated at $\boldsymbol{\delta} = \boldsymbol{\delta}_0$. This is one-half the regression sum of

squares of the regression of $\mathbf{c}$ on $D$ and is thus easy to calculate. Koenker (1981) showed that the score test for nonconstant variance in the linear regression model can be made more robust through studentizing, where $\hat{\sigma}^2$ is replaced with $[\sum_i (\hat{e}_i^2 - \hat{\sigma}^2)^2/(2n)]^{1/2}$. Simonoff and Tsai (1994) applied this same idea to the score test for heteroscedasticity based on the modified profile likelihood.

Eubank and Thomas (1993) investigated this score test for heteroscedasticity in the context of the nonparametric regression model,

$$y_i = g(x_i) + \varepsilon_i, \qquad \varepsilon_i \sim N(0, \sigma_i^2), \qquad (2.5)$$

estimating $g$ using a cubic smoothing spline. Specifically, they proposed the test statistic

$$T = \frac{[\sum_{i=1}^n A_i(\hat{e}_i^2/\hat{\sigma}^2) - 1]^2}{2 \sum_{i=1}^n A_i^2}$$

to test for nonconstant variance of the form $\sigma_i^2 = \sigma^2 + n^{-1/2} d_i$ (putting appropriate boundedness conditions on the $\{A_i\}$ and $\{d_i\}$). They then showed that for a wide range of choices of the smoothing parameter (including ones that lead to optimal convergence rates), $T$ is asymptotically noncentral $\chi_1^2$ with noncentrality parameter depending on $\{A_i\}$ and $\{d_i\}$.

This result can be adapted to the single index model when there is one variance predictor. Because $\boldsymbol{\beta}$ is $\sqrt{n}$-consistent, the linear index values $z_i$ can be treated as the predictor values $x_i$ in (2.5), implying an asymptotic $\chi_1^2$ distribution under the null hypothesis, and power to detect $\sqrt{n}$ departures from constant variance. The theorem requires the local quadratic smoothing parameter to be $o(n^{-1/12})$, which is satisfied by the smoothing parameter that minimizes the (mean) integrated squared error (which has a rate of $O(n^{-1/9})$ in the interior), although technically it does not allow for data-dependent choices of $h$.

## 2.2   Test for First-Order Autoregressive Errors

Consider model (2.1) with $\boldsymbol{\delta} = \boldsymbol{\delta}_0$. This model incorporates first-order autoregressive [AR(1)] autocorrelation into the errors if $\rho \neq 0$. The score test statistic is based on hypotheses (2.2) with $\boldsymbol{\theta}_1 = (\boldsymbol{\beta}', \sigma^2)'$ and $\theta_2 = \rho$ and equals

$$S_A = \left[\sum_{i=2}^n \hat{e}_i \hat{e}_{i-1}/\hat{\sigma}^2\right]^2 \Big/ (n-1). \qquad (2.6)$$

Defining an estimate of $\rho$ to be $\hat{\rho} = \sum_{i=2}^n \hat{e}_i \hat{e}_{i-1} / \sum_{i=1}^n \hat{e}_i^2$, the score test statistic is effectively $S_A = n\hat{\rho}^2$.

The statistic $S_A$ can be written in the form

$$S_A = \frac{\mathbf{y}'M\mathbf{y} - \text{tr}(A)}{2 \text{ tr}(A^2)},$$

where $M = (I - H)A(I - H)$ and $A$ is the $n \times n$ matrix with 1's in the main off-diagonal and 0's elsewhere. Eubank and Thomas (1993, p. 154) showed that statistics of this form are asymptotically $\chi_1^2$ under the null hypothesis in the nonparametric regression context, and this result again carries over to the single index model, justifying the use of $\chi_1^2$ critical values here.

## 2.3   Test for the Need for Additional Covariates

Consider model (2.1) with $\rho = 0$ and $\boldsymbol{\delta} = \boldsymbol{\delta}_0$. Given $\boldsymbol{\beta}_1 \neq \mathbf{0}$, this model allows for the possibility of additional useful covariates if $\boldsymbol{\beta}_2 \neq \mathbf{0}$. The score test statistic is based on hypotheses (2.2) with $\boldsymbol{\theta}_1 = \boldsymbol{\beta}_1$ and $\boldsymbol{\theta}_2 = \boldsymbol{\beta}_2$ and equals

$$S_C = \frac{\hat{\mathbf{e}}_1' \widetilde{X}_2 [\widetilde{X}_2'(I - \widetilde{H}_1)\widetilde{X}_2]^{-1} \widetilde{X}_2' \hat{\mathbf{e}}_1}{\hat{\sigma}_1^2}, \qquad (2.7)$$

where

$$\widetilde{X}_2 = \text{Diag}[\hat{\dot{g}}(X_1 \hat{\boldsymbol{\beta}}_1)]X_2^c,$$
$$\widetilde{H}_1 = \widetilde{X}_1(\widetilde{X}_1' \widetilde{X}_1)^{-1} \widetilde{X}_1,$$
$$\widetilde{X}_1 = \text{Diag}[\hat{\dot{g}}(X_1 \hat{\boldsymbol{\beta}}_1)]X_1^c,$$
$$\hat{\sigma}_1^2 = \hat{\mathbf{e}}_1' \hat{\mathbf{e}}_1/n,$$
$$\hat{\mathbf{e}}_1 = \mathbf{y} - \hat{g}(X_1 \hat{\boldsymbol{\beta}}_1),$$

$\hat{\dot{g}}(X_1 \hat{\boldsymbol{\beta}}_1)$ and $\hat{g}(X_1 \hat{\boldsymbol{\beta}}_1)$ are the estimates of $\dot{g}$ and $g$ evaluated at the observations, respectively, based on using only $X_1$ as predictors in the model, and $X_1^c$ and $X_2^c$ are mean-centered versions of $X_1$ and $X_2$, respectively. Centered versions of the predictors must be used for this test because, as noted above, the intercept $\beta_0$ is not identifiable in model (1.2). Similarly, because $\hat{\boldsymbol{\beta}}$ is determined only up to a constant, the values of $\hat{g}(X\hat{\boldsymbol{\beta}})$ must be standardized to have unit mean (as they would have if $g$ were linear). Evans and Savin (1982) proposed an adjustment to the score test for extra covariates based on degrees of freedom in the linear regression model that can also be applied here. Unfortunately, there is no asymptotic theory for $S_C$ in the single index situation, but we will continue to use $\chi^2$-based critical values as a benchmark.

## 3.   MONTE CARLO SIMULATIONS

In this section we examine the properties of the tests using Monte Carlo simulations. We examine both the size and the power properties of the tests, and how they relate to sample size, the strength of the relationship between $y$ and $\mathbf{x}$, the nature of $g$, and the type of $\hat{\boldsymbol{\beta}}$ used. All tests have the same level $\alpha = .05$ and are examined using sample sizes $n = 30, 50$. Tests for heteroscedasticity are based on the model $y = g(\beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3) + \varepsilon \times \exp(z\delta/2)$ with $\beta_1 = \beta_2 = \beta_3 = 1$, $\varepsilon \sim N(0, \sigma^2)$, $z$ standard normal, and $\delta$ taking the values $(-2, -1, -.5, 0, .5, 1, 2)$. Tests for autocorrelation are based on the model $y = g(\beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3) + \varepsilon$, where $\beta_1 = \beta_2 = \beta_3 = 1$, $\varepsilon$ follows an AR(1) process, and $\rho \in (-.9, -.5, -.1, 0, .1, .5, .9)$. Tests for an additional covariate are based on the true model $y = g(\beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3) + \varepsilon$, where $\beta_1 = \beta_2 = 1$ and $\beta_3$ takes the values $(0, 1, 5, 9)$, $\beta_3 = 0$ corresponding to the null hypothesis. For each situation $x_1$, $x_2$, and $x_3$ are generated once as uniform on $(0, 1)$, $g$ is linear, negative exponential, or a sine function, and $\sigma$ equals the range of $g(\beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3)$ or the range of $g(\beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3)$ divided by 10, respectively [that is, the signal-to-noise ratio (SNR) equals 1 or 10, respectively]. Each setting is simulated with 1,000 replications, resulting in a maximum standard error of the estimates of power of .016 (estimates of size, which are around .05, have standard errors roughly half as large).

Either OLS or SIR was used to estimate $\boldsymbol{\beta}$ for each simulation replication, followed by a local quadratic estimate of $g$ using $AIC_C$ to choose the smoothing parameter in each case. The tests based on ordinary least squares (OLS) and SIR performed similarly, with no clear advantage for one method over the other. Given that the OLS estimates are much more familiar, and simpler to calculate, all the simulation results reported here are based on using OLS.

Figures 1–3 summarize results for the tests for heteroscedasticity, autocorrelation, and an extra covariate, respectively. Each figure consists of six plots giving simulated power and size values, separated by the form of $g$ and the SNR (the solid lines refer to the score test). In each plot, the lines represent linear interpolations between the actual values obtained in the simulations, those showing higher power corresponding to sample size $n = 50$ and those with lower power corresponding to $n = 30$. The horizontal line in each plot represents the nominal size $\alpha = .05$.

Figure 1 summarizes results for testing heteroscedasticity. In addition to the score test, two other easily calculated tests were examined. Levene's test (dashed lines) is the $F$-test for significance in a regression of $|\hat{\mathbf{e}}|$ on the variance predictor(s),

and Harvey's test is the regression sum of squares of a regression of $\log[(\hat{\mathbf{e}})^2]$ on the variance predictor(s) divided by 4.9348 and is compared to a $\chi^2$ distribution. The latter test consistently exhibited much lower power than the score and Levene's tests, so its values are not given in the plots.

The tests generally hold their size well, with estimated size less than .06 in virtually all cases, despite the small samples used here. As would be expected, the power of the tests increases as the heteroscedasticity increases ($|\delta|$ increasing), with positive and negative $\delta$ generally equally easy to identify. Whereas the score test and Levene's test are close in power, the score test is consistently the more powerful test. The strength of the regression relationship and the nature of $g$ seem to have little effect on the test, with the exception of a sine relationship and high SNR, where power is lower. Power values otherwise exceed .9 ($n = 30$) or .95 ($n = 50$) for $|\delta| = 2$.

The properties of the score test for autocorrelation (Figure 2) are broadly similar to those for heteroscedasticity. The plots also give values for an application of the Durbin–Watson statistic (dashed lines) to the observed residuals,

$$d = \frac{\sum_{i=2}^{n} (\hat{e}_i - \hat{e}_{i-1})^2}{\sum_{i=1}^{n} \hat{e}_i^2}.$$
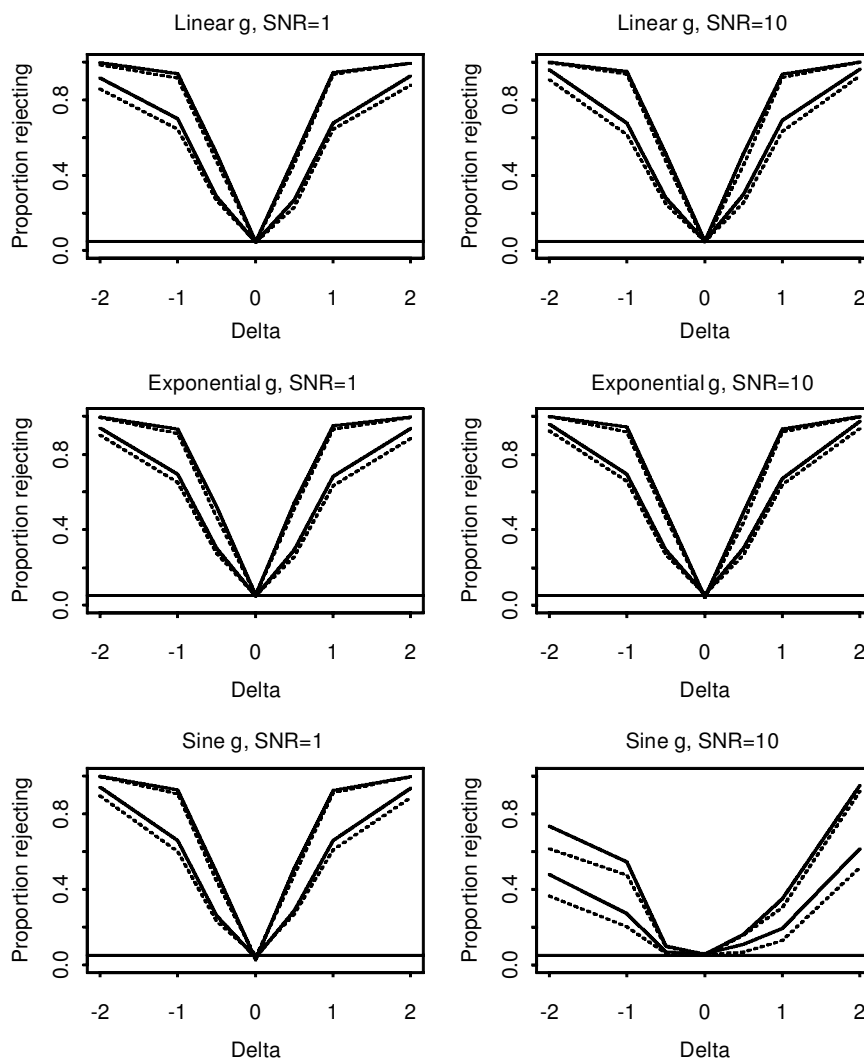


Figure 1. Results of Monte Carlo Simulations of Power and Size of Test for Heteroscedasticity as a Function of $\delta$, Separated by Type of $g$ and SNR. Solid lines refer to score test and dashed lines to Levene's test.
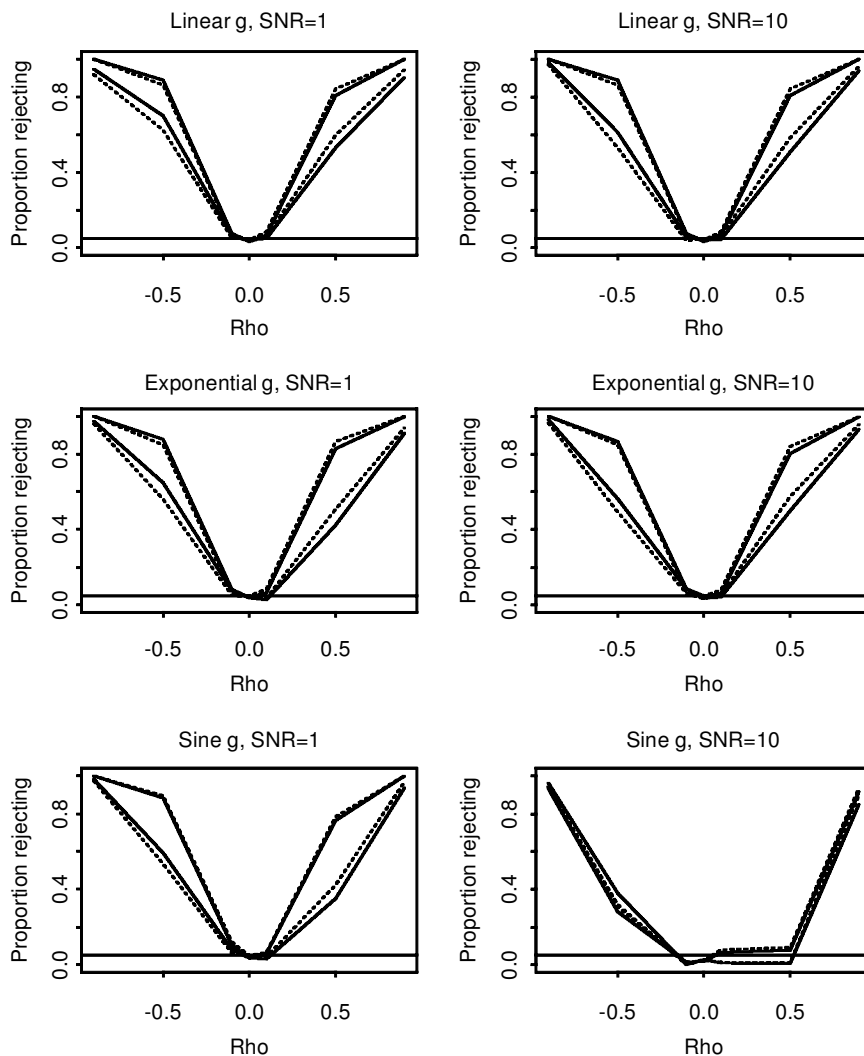
*Figure 2.   Results of Monte Carlo Simulations of Power and Size of Test for Autocorrelation as a Function of $\rho$, Separated by Type of $g$ and SNR. Solid lines refer to score test and dashed lines to Durbin–Watson test.*

Significance of the statistic is based on an asymptotic normal limit, or equivalently a $\chi_1^2$ approximation for $n(d/2-1)^2$. The tests hold the .05 size well, with all estimated sizes less than .06. The underlying form of $g$ does not have a large effect, except for the sine function with high SNR, where power is noticeably lower. This is presumed to relate to the difficulties of estimating functions with more structure (as in the heteroscedasticity test) combined with the difficulties of $\text{A IC}_C$ in differentiating between a regression function with some curvature and some autocorrelation (the actual situation) and one with even more curvature and little autocorrelation (which would erroneously lead to nonrejection of the null hypothesis). The score test and Durbin–Watson test have similar properties; the score test is more powerful for negative autocorrelation and the Durbin–Watson test is more powerful for positive autocorrelation. This pattern is easy to understand, because

$$n\left(\frac{d}{2}-1\right)^2 = n\left(\hat{\rho}+\frac{\hat{e}_1^2+\hat{e}_n^2}{\sum_{i=1}^n \hat{e}_i^2}\right)^2 \equiv n\tilde{\rho}^2.$$

Clearly $\tilde{\rho} \geq \hat{\rho}$, implying a (typically) smaller $\chi^2$ statistic (and lower power) when $\hat{\rho} < 0$ and larger $\chi^2$ statistic (and higher power) when $\hat{\rho} > 0$.

Figure 3 presents results for the score test for an additional covariate, along with ones based on the ordinary $t$ test (dashed lines). The degrees-of-freedom adjustment of Evans and Savin (1982) has little effect in the situations studied here, so its results are not presented. The tests are slightly, but consistently, anticonservative, with sizes typically between .06 and .07. The power properties of the score test are considerably more complex than those of the heteroscedasticity and autocorrelation score tests, for several reasons. Recall that the score test is derived assuming that the null hypothesis is true, which means that the model is misspecified under the alternative hypothesis. This is only a second-order effect in the presence of heteroscedasticity or autocorrelation (in the sense that the mean function is still correctly specified), which means that $\hat{g}(\mathbf{x}_i/\boldsymbol{\beta})$ is still a consistent estimator of $E(y_i)$. This is not the case when testing for the need for an extra covariate, because then the estimated linear index is not a consistent estimate of the true linear index. Thus, $\hat{g}$ is not a consistent estimate of the true $g$, and $\hat{\dot{g}}$ is not a consistent estimate of $\dot{g}$.

This accounts for several of the patterns in Figure 3. The score test is best behaved for linear $g$, but the power decreases with increasing $\beta_3$ when SNR equals 10 (the same is true for
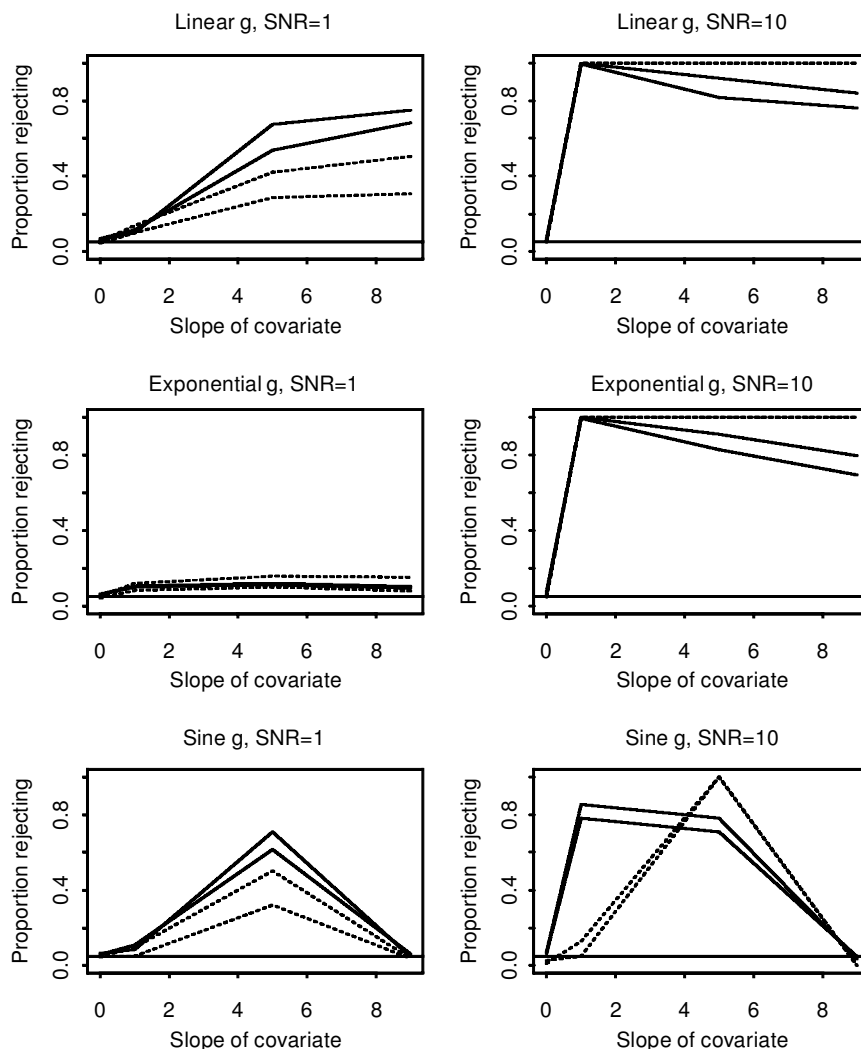
*Figure 3. Results of Monte Carlo Simulations of Power and Size of Test for an Additional Covariate as a Function of $\beta_3$, Separated by Type of g and SNR. Solid lines refer to score test and dashed lines to t test.*

the exponential function). The problem here is in estimation of $\dot{g}$. Recall that the same smoothing parameter is used to estimate both $g$ and $\dot{g}$, although theory shows that they should be different. This leads to degradation of the test. If the true values of $\dot{g}$ are used in the simulations (this can be done because $\dot{g}$ is known), the power functions for linear $g$ monotonically increase to 1 with increasing $\beta_3$ as would be expected, which confirms that the problem is in estimation of $\dot{g}$.

The poor performance of the test for exponential $g$ and SNR $= 1$ arises because $E(y_i)$ is close to zero and with the added noise exhibits little structure. In this case including the predictor $x_3$ adds very little to the fit. The structure of $x_3$ used here results in a similar pattern for the sine function when $\beta_3 = 9$.

On the other hand, the flexibility of the single index model means that the estimate $\hat{g}$ based on the misspecified (null) model will often fit better than the true $g$ based on the null model, which can translate into improved ability to identify structure related to $x_3$. The score test based on the true $g$ (results not presented here) is sometimes more powerful than that based on the single index model (reflecting the benefit of

not having to estimate $g$), and sometimes it is less powerful (reflecting better fit of the nonparametrically estimated $g$).

The $t$ test performs reasonably well in many of the situations, because the partial relationships between $y$ and $x_3$ given $x_1$ and $x_2$ for these functions have enough of a linear component to be identified. It is not at all difficult, however, to construct situations in which the score test is much more powerful than the $t$ test. For example, if the entries in $x_1$ and $x_2$ have mean zero, and the $x_3$ values are symmetric around zero, the $t$ test has no power to identify nonzero $\beta_3$ in the quadratic function $(\beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3)^2$, whereas the score test can have power greater than .5 when $n = 30$. The superiority of the score test over the $t$ test when $g$ is linear and the SNR is low seems paradoxical. This occurs because $\hat{\sigma}^2$ is biased downward, reflecting the flexibility of the single index model noted earlier ($\hat{g}$ is closer to the observed $y$ values than is $g$).

Although the score test $S_C$ is useful, it should be used with caution in practice. However, because its difficulties are much more related to power than to size, a statistically significant test can indicate that the additional covariates are important to the model fitting. It might seem that the solution to the

difficulties in using the score test is to use a likelihood-ratio-type test,

$$LR = n[\log(\hat{\sigma}_1^2) - \log(\hat{\sigma}_{1,2}^2)],$$

where $\hat{\sigma}_{1,2}^2$ is the variance estimate $\sum \hat{e}_i^2 / n$ based on the full dataset $(X_1, X_2)$. (Recall that $\hat{\sigma}_1^2$ is the corresponding estimate based on using only $X_1$ as predictors.) Unfortunately, this is not the case, because the statistic does not follow a standard distribution under the null hypothesis. The unique characteristics of the single index model make it impossible to appeal to standard inferential arguments. Unlike in linear regression models, it is not necessarily the case that the unrestricted model (including the extra covariate) fits better than the restricted model (omitting it). Because the form of $g$ is different under the null and alternative hypotheses, it can happen that $\hat{\sigma}_1^2 < \hat{\sigma}_{1,2}^2$, implying $LR < 0$. For the cases examined here, this happened 15–25% of the time for high SNR situations and more than 30% of the time for low SNRs. Further, the flexibility of estimation of $g$ in the single index model implies that even under the null hypothesis the model using the additional covariate improves estimation accuracy more than would usually be expected. Using a $\chi^2$ approximation, the tests are seriously anticonservative, with null sizes for both $n = 30$ and $n = 50$ around $.15 - .2$ for low signal-to-noise and $.08 - .1$ for high SNR.

Small-scale simulations based on $t_3$-distributed errors showed that although the score tests for autocorrelation and an additional covariate were relatively insensitive to the nonnormal errors, for heteroscedasticity it was not, becoming strongly anticonservative. Studentizing the score test brings its size back to the proper level, but at the price of reduced power, making that approach problematic. For example, the studentized score test is less powerful than the ordinary score test for normal errors and is less powerful than Levene's test for $t_3$ errors.

## 4.   ANALYSIS OF REAL DATA

In this section three real datasets are analyzed using the single index model score tests, and it is shown that not taking model misspecifications into account can lead to very different inferences. In all cases the linear index was calculated using OLS on centered predictors, and $g$ was estimated using a local quadratic estimator with smoothing parameter chosen using $AIC_C$. For all three examples SIR-based linear indices led to substantially similar results.

### 4.1   Automobile Prices

The data come from the 1999 Car Buyers issue of *Consumer Reports* (1999) and represent a sample of 41 autos. The price (in thousands of dollars) of the auto is modeled as a function of engine power (in horsepower), weight (in thousands of pounds), passenger space (in cubic feet), and reliability rating (from 1 to 5, with 1 being worst and 5 being best), as these are all characteristics that could be viewed as important by consumers. Figure 4(a) gives a scatter plot of price versus the linear index, with the estimated $g$ superimposed, indicating nonlinearity in the relationship. Figure 4(b) is a scatter plot of the residuals from this fit versus the linear index, which strongly suggests heteroscedasticity. The score test $S_H$ given in (2.4),
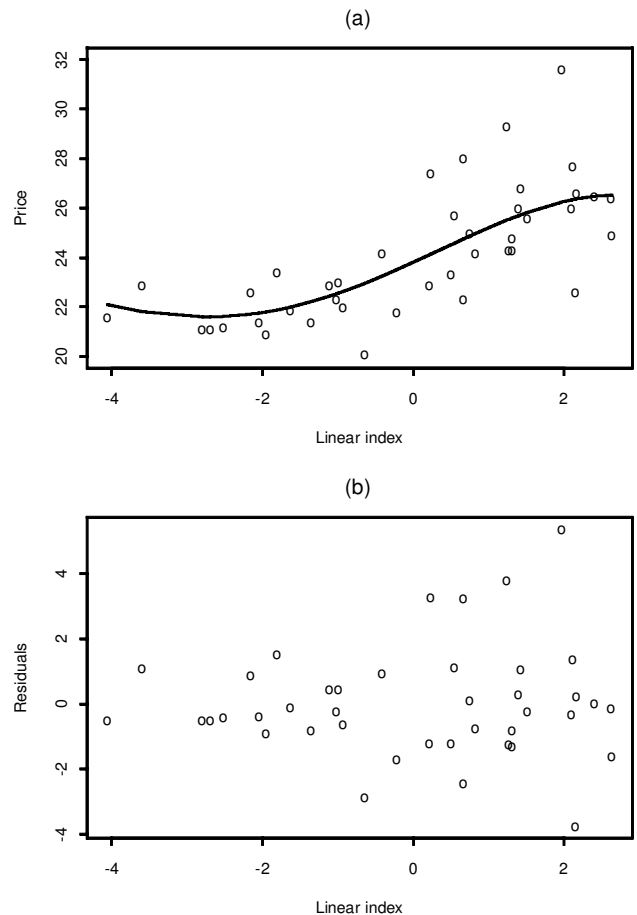


*Figure 4.   Plots. (a) Automobile price (in thousands of dollars) versus linear index, with $\hat{g}$ superimposed; (b) residuals from the single index model versus linear index.*

assuming an exponential model $V(\varepsilon_i | z_i) = \sigma_i^2 = \sigma^2 \exp(\delta z_i)$, equals 6.13 on one degree of freedom ($p = .013$), strongly rejecting constant variance. This inference is insensitive to the choice of smoothing parameter. Figure 5 gives a significance trace for this test (Bowman and Azzalini 1997, p. 89), which gives the tail probability of the test based on different choices
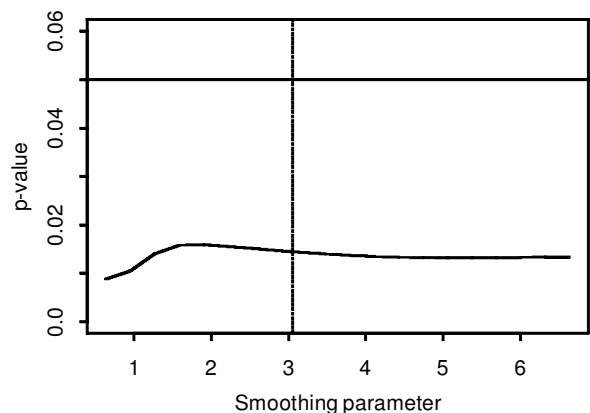


*Figure 5.   Signi'cance Trace for Test of Heteroscedasticity for Automobile Price Data. The dashed vertical line corresponds to the $AIC_C$ smoothing parameter choice. A horizontal line is given at the $\alpha = .05$ level.*

of the smoothing parameter (the $AIC_C$ choice is represented by the vertical dashed line). It is clear that the $p$-value of the test is virtually constant over a wide range of smoothing parameters.

Härdle, Hall, and Ichimura (1993) showed that a weighted version of $\hat{\beta}$ has higher efficiency than an unweighted version when there is nonconstant variance, and they proposed a multistage estimator. We do not pursue this because their heteroscedasticity model is different from the one used here. Still, even if $\hat{\beta}$ is not changed, heteroscedasticity needs to be taken into account in making predictions, for example. A rough pointwise 95% prediction interval at $z = z_0$ is

$$\hat{g}(z_0) \pm 2\sqrt{[\widehat{V}(\hat{g}(z_0)) + \widehat{V}(\varepsilon|z_0)]}. \qquad (4.1)$$

This is only roughly a prediction interval, because it ignores the bias in $\hat{g}(z_0)$, but it is useful as a guide for the effect of nonconstant variance on predictions. Once heteroscedasticity is suspected, a weighted version of $\hat{g}$ is appropriate, with weights $\hat{\sigma}_i^{-2}$. Asymptotic forms for $V(\hat{g}(z_0))$ are well known; see, for example, Fan and Gijbels (1996, p. 62).

Estimating the weights requires estimating the variance coefficients $\boldsymbol{\delta}$. Because $E(\varepsilon_i^2) = \sigma_i^2$ is an exponential function of $z_i$, a Poisson regression of $\varepsilon^2$ on $\mathbf{z}$ using the canonical logarithmic link function

$$\log[E(\varepsilon_i^2)] = \delta_0 + \delta_1 z_i$$

provides consistent estimates for $\boldsymbol{\delta}$, and hence $\sigma_i^2$ (and thus $V(\varepsilon|z_0)$). The errors $\boldsymbol{\varepsilon}$ are not available, of course, so a Poisson regression of $\hat{\mathbf{e}}^2$ on $\mathbf{z}$ provides a practical implementation of this idea.

Figure 6 gives a plot of the pointwise prediction intervals for the unweighted (dashed) and weighted (solid) models based on (4.1). It is apparent that the intervals without accounting for heteroscedasticity are not appropriate, being too wide for smaller $z$ and too narrow for larger $z$. In contrast, the intervals accounting for heteroscedasticity follow the observed relationship closely, and they could be used in predictions for other automobiles not in the sample.

## 4.2 Carbon Monoxide Emissions Data

These data track carbon monoxide emissions in the United States from 1960 through 1989. The target variable is per capita emissions (in metric tonnes per person), with predictors per capita constant dollar gasoline consumption, millions of acres destroyed by forest fires, and a linear time trend, thereby taking into account two major causes of carbon monoxide (internal combustion engines and forest fires), while still allowing for a trend in emissions given these causes. The data come from the web sites of the Census Bureau, the Environmental Protection Agency, the Department of Agriculture, and *Statistical Review of World Energy*, published by British Petroleum. The coefficient for time in the OLS-based linear index is negative, indicating a declining trend in emissions given these causes. The linear model indicates a strong relationship between the predictors and emissions ($R^2 = 91.3\%$) but also shows strong evidence of autocorrelation in the residuals (for example, the Durbin–Watson statistic equals .70, which is highly statistically significant).

Figure 7 gives a plot of emissions versus the linear index, with estimated $g$ superimposed. The plot suggests some nonlinearity in the relationship. This nonlinearity of $\hat{g}$ completely changes inferences about the existence of autocorrelation, because the score test for autocorrelation given in (2.6) is $S_A = .34$, which does not give any sign of autocorrelation ($p = .56$). Figure 8 shows why the nonlinearity of $g$ is so important. That figure plots the residuals versus year from the linear model (dashed line) and single index model (solid line). It is apparent that the Durbin–Watson test is interpreting the pattern in Figure 7 as autocorrelation in the errors, rather than nonlinearity in $g$. This is also apparent in the significance trace for this test given in Figure 9. Although the test is not statistically significant for a wide range of smoothing parameters, as the smoothing parameter becomes larger the estimated curve becomes straighter, and eventually the test becomes statistically significant. It is impossible to tell which interpretation (autocorrelation in the errors or nonlinearity of $g$) is correct for these real data, given the existence of other factors related to air pollution that are not in the model, but nonlinearity seems reasonable. For example, the emissions observations at the far right of Figure 7 correspond to pre-1973 data, before emissions began to steadily fall. This drop
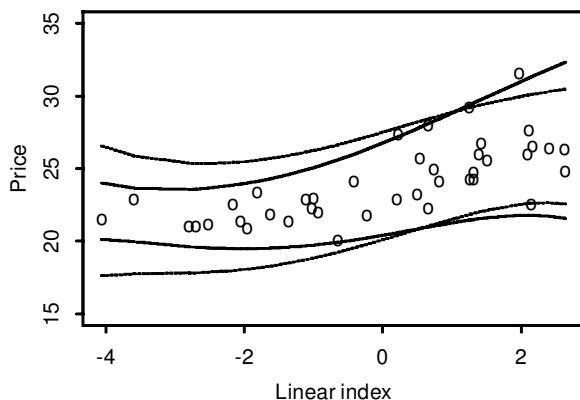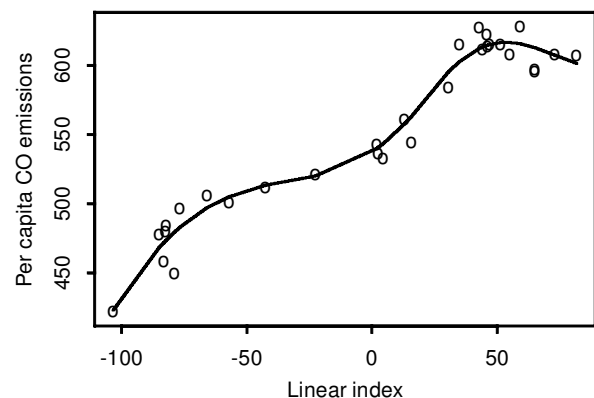
*Figure 6. Plot of Approximate Pointwise 95% Prediction Intervals for Automobile Price. Not accounting for heteroscedasticity (dashed lines) and accounting for heteroscedasticity (solid lines).*

*Figure 7. Plot of Per Capita Carbon Monoxide Emissions (in Metric Tonnes Per Person) Versus Linear Index, With $\hat{g}$ Superimposed.*
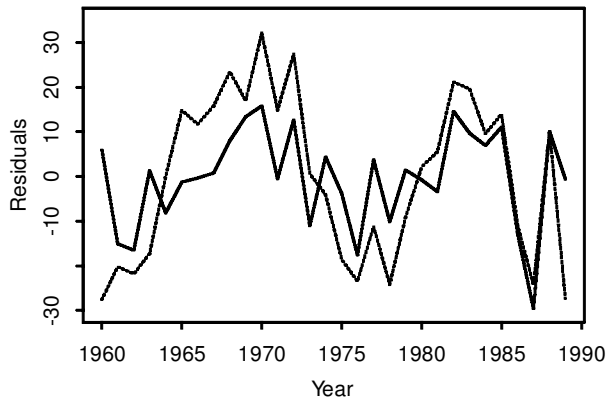
*Figure 8. Plot of Residuals Versus Year for Linear Model. Carbon monoxide emissions (dashed line) and single index model (solid line).*
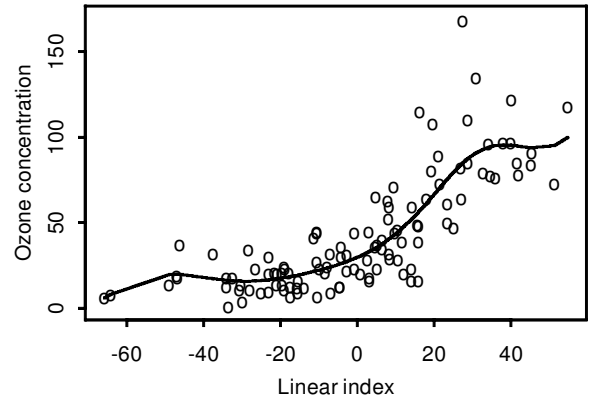


*Figure 10. Plot of Ozone Concentration Versus Linear Index, With $\hat{g}$ Superimposed.*

is attributable in large part to the implementation of the Clear Air Act of 1970, which put sharp limitations on sources of air pollution.

## 4.3 Ozone Concentration Data

The data come from Cook and Weisberg (1994, p. 127) and refer to air quality readings taken on 111 days in the New York City area in 1973. The target variable is the ozone concentration in parts per billion, and the potential predictors are the temperature (in degrees Fahrenheit), wind speed (in miles per hour), and solar radiation (in Langleys). The $t$ statistics for the three variables from an OLS fit are 6.52, $-5.10$, and 2.58, respectively. Thus, the evidence for the need for the solar radiation variable given the other two variables is barely statistically insignificant at a .01 level ($p = .011$). This is dependent, of course, on an appropriate linear model.

Figure 10 is a plot of ozone concentration versus the linear index based on only temperature and wind speed, with estimated $g$ superimposed; noticeable curvature in $g$ is again evident. Given the use of these two predictors, the score test (2.7) for the need for solar radiation level in the single index model is $S_C = 11.96$, which is highly statistically significant ($p = .0005$). The significance trace for this test given in Figure 11 shows that this inference is very insensitive to the choice

of smoothing parameter. Thus, when the curvature of $g$ is considered, the usefulness of the solar radiation variable is considerably greater than would have been thought otherwise.

These data form a time series, of course, so issues of autocorrelation are also relevant here. According to the score test for autocorrelation, this is not a problem, as it gives no indication of any autocorrelation.

## 5. CONCLUSIONS

In this article we used the general structure of the score test to derive tests for three possible misspecifications of the single index model. Monte Carlo simulations show that the tests hold their size well and can be reasonably powerful, even for small samples.

Score tests also can be derived for other models of interest. For example, if autocorrelation (if present) is assumed to be AR($p$), rather than AR(1), the score test generalizes to the intuitive form

$$S = n\left(\hat{\rho}_1^2 + \cdots + \hat{\rho}_p^2\right),$$

where $\hat{\rho}_j = \sum_{i=j+1}^{n} \hat{e}_i \hat{e}_{i-j} / \sum_{i=1}^{n} \hat{e}_i^2$, referenced to a $\chi_p^2$ distribution. Obviously the Durbin–Watson test is potentially useless for models of this type, if $\rho_1$ is small but higher-order autocorrelations are large.
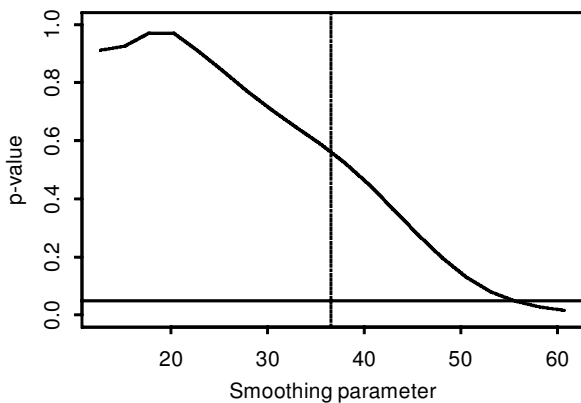


*Figure 9. Signi' cance Trace for Test of Autocorrelation for Carbon Monoxide Emissions Data. The dashed vertical line corresponds to the $AIC_C$ smoothing parameter choice. A horizontal line is given at the $\alpha = .05$ level.*
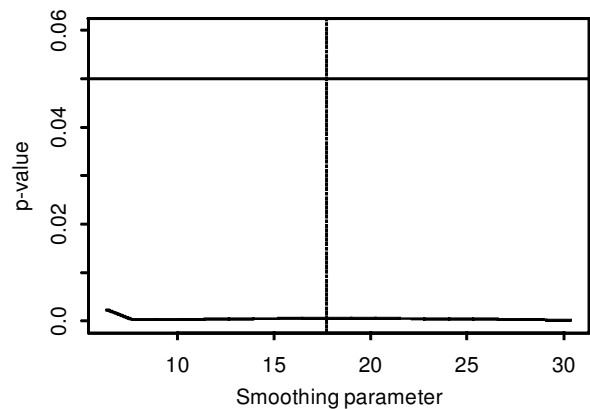


*Figure 11. Signi' cance Trace for Test of an Additional Covariate for Ozone Concentration Data. The dashed vertical line corresponds to the $AIC_C$ smoothing parameter choice. A horizontal line is given at the $\alpha = .05$ level.*

The reduced power of the test for autocorrelation when the underlying function was a sine function shows that smoothing parameter selection for this test could be improved. The test is derived assuming the null is true, so the $AIC_C$ choice is not inappropriate, but it can lead to autocorrelation being mistaken for signal. Hart (1994) and Hart and Yi (1998) proposed time series cross-validation for smoothing parameter selection under autocorrelation, and it is reasonable to think that using a smoothing parameter based on this method in the autocorrelation test could improve power.

Problems with smoothing parameter choice also affect the properties of the test for an additional covariate, in that they affect estimation of $\dot{g}$. Using a different smoothing parameter for estimating $g$ and $\dot{g}$ is appropriate in theory and would definitely help for $g$ functions close to linear. There has not been a great deal of work done on smoothing parameter choice for estimation for regression derivatives, although the plug-in method of Fan and Gijbels (1995) is one possible approach. The deterioration of power for other $g$ functions as the coefficient of the additional covariate gets larger is clearly a concern, but it seems to be an inherent part of the score test construction. In addition, the flexibility of the single index model means that the estimated $g$ under the null can be far from the "true" $g$, potentially reducing the power of the test, but the failure of the likelihood ratio to provide a useful test makes it clear that further research is needed in this area.

S-Plus code to implement these tests is available from the authors.

## ACKNOWLEDGMENTS

*[Received August 2000. Revised September 2001.]*

## REFERENCES

Bowman, A. W., and Azzalini, A. (1997), *Applied Smoothing Techniques for Data Analysis*, Oxford: Clarendon Press.

Brillinger, D. R. (1983), "A Generalized Linear Model with 'Gaussian' Regression Variables," in *A Festschrift for Erich L. Lehmann in Honor of His Sixty-Fifth Birthday*, eds. P. J. Bickel, K. A. Doksum, and J. L. Hodges, Belmont, CA: Wadsworth, pp. 97–114.

Carroll, R. J., Fan, J., Gijbels, I., and Wand, M. P. (1997), "Generalized Partially Linear Single-Index Models," *Journal of the American Statistical Association*, 92, 477–489.

Consumer Reports (1999), "The '99s," *Consumer Reports*, Vol. 64, No. 4, Yonkers, NY: Consumers Union.

Cook, R. D., and Weisberg, S. (1994), *An Introduction to Regression Graphics*, New York: Wiley.

Cox, D. R., and Hinkley, D. V. (1974), *Theoretical Statistics*, London: Chapman and Hall.

Duan, N., and Li, K. C. (1991), "Slicing Regression: A Link Free Regression Method," *Annals of Statistics*, 19, 505–530.

Eubank, R. L., and Thomas, W. (1993), "Detecting Heteroscedasticity in Nonparametric Regression," *Journal of the Royal Statistical Society*, Ser. B, 55, 145–155.

Evans, G. B. A., and Savin, N. E. (1982), "Conflict Among the Criteria Revisited; The *W*, *LR* and *LM* Tests," *Econometrica*, 50, 737–748.

Fan, J., and Gijbels, I. (1995), "Data-Driven Bandwidth Selection in Local Polynomial Fitting: Variable Bandwidth and Spatial Adaptation," *Journal of the Royal Statistical Society*, Ser. B, 57, 371–394.

——— (1996), *Local Polynomial Modelling and its Applications*, London: Chapman and Hall.

Härdle, W., Hall, P., and Ichimura, H. (1993), "Optimal Smoothing in Single-Index Models," *Annals of Statistics*, 21, 157–178.

Hart, J. D. (1994), "Automated Kernel Smoothing of Dependent Data by Using Time Series Cross-Validation," *Journal of the Royal Statistical Society*, Ser. B, 56, 529–542.

Hart, J. D., and Yi, S. (1998), "One-Sided Cross-Validation," *Journal of the American Statistical Association*, 93, 620–631.

Horowitz, J. L. (1998), *Semiparametric Methods in Econometrics*, New York: Springer-Verlag.

Hurvich, C. M., Simonoff, J. S., and Tsai, C.-L. (1998), "Smoothing Parameter Selection in Nonparametric Regression Using an Improved Akaike Information Criterion," *Journal of the Royal Statistical Society*, Ser. B, 60, 271–293.

Ichimura, H. (1993), "Semiparametric Least Squares (SLS) and Weighted SLS Estimation of Single Index Models," *Journal of Econometrics*, 58, 71–120.

Koenker, R. (1981), "A Note on Studentizing a Test for Heteroscedasticity," *Journal of Econometrics*, 17, 107–112.

Powell, J. L., Stock, J. H., and Stoker, T. M. (1989), "Semiparametric Estimation of Index Coefficients," *Econometrica*, 61, 123–138.

Seber, G. A. F., and Wild, C. J. (1989), *Nonlinear Regression*, New York: Wiley.

Simonoff, J. S. (1996), *Smoothing Methods in Statistics*, New York: Springer-Verlag.

Simonoff, J. S., and Tsai, C.-L. (1994), "Use of Modified Profile Likelihood for Improved Tests of Constancy of Variance in Regression," *Applied Statistics*, 43, 357–370.

Tsai, C.-L. (1986), "Score Test for the First-Order Autoregressive Model With Heteroscedasticity," *Biometrika*, 73, 455–460.