DIFFERENCING METHODS IN NONPARAMETRIC REGRESSION: SIMPLE TECHNIQUES FOR THE APPLIED ECONOMETRICIAN

A. Yatchew*

There has been an explosion in nonparametric regression techniques in statistics and econometrics, yet the use of these tools by *applied* economists has been much more limited. The motivation and purpose of this essay is to provide a *simple* collection of techniques for analyzing a basic class of nonparametric and semiparametric models. The unifying theme is the idea of differencing which permits one to remove the nonparametric effects from the data in order to estimate the parametric effects. The estimated parametric effects are in turn removed from the original data and the nonparametric effects are then analyzed. This device permits one to draw not only on the reservoir of parametric human capital, but also to make use of existing software. Bootstrap procedures are included for a number of the techniques.

Keywords: nonparametric regression, optimal differencing, testing equality of nonparametric regressions, partial linear model, partial parametric model, semiparametric model, specification test, monotonicity, concavity, isotonic regression, instrumental variables, two stage least squares, translog, Cobb-Douglas, CES, bootstrap, moving block bootstrap, heteroskedasticity, autocorrelation, Newey-West standard errors

JEL Classification: C14

October 20, 2000.

SurveysEssay.wpd

* Department of Economics, University of Toronto, yatchew@chass.utoronto.ca, (416) 978-7128 voice, 978-6713 (fax).

Preface

This essay is largely motivated by pedagogical interests. Nonparametric techniques are widely studied by theoretical econometricians, but much under-used by applied economists. The techniques are theoretically sophisticated and often require substantial computational and programming experience. Graduate students writing applied theses often shy away from them.

The differencing device provides a convenient means for introducing nonparametric techniques to practitioners in a way which parallels their knowledge of parametric techniques. The reason is that it allows one to (approximately) remove the nonparametric effect and to analyze the parametric portion of the model as if the nonparametric portion was not there to begin with. This permits one to draw not only on the reservoir of parametric human capital, but also to make use of existing software.

This essay draws heavily on past differencing papers: Rice (1984) who introduces the differencing estimator of the residual variance, Powell (1987) and Ahn and Powell (1993) who apply differencing to the partial linear model, Yatchew (1988) and Cox and Koh (1989) who propose nonparametric tests using first order differencing, Hall, Kay and Titterington (1990) whose important contribution is the derivation of optimal differencing coefficients, and Yatchew (1997,1998a,b,1999) who applies differencing to the partial linear model, to specification testing, to testing equality of nonparametric regressions and to panel data. In order to provide a coherent set of techniques the current paper incorporates many of the results in these papers.

The essay also contains a number of innovations to differencing techniques. These include extension of the specification test to heteroskedastic settings and to tests of monotone or concave nulls; IV estimation, testing for endogeneity and heteroskedasticity/autocorrelation consistent standard errors in the partial linear model; a simple semiparametric extension of the CES model; and, the delineation of bootstrap procedures for a number of the tests and estimators.

Table of Contents

1. INTRODUCTION

1.1 The Main Theme

Numerous works have explored estimation and inference for nonparametric regression models:

$$y = f(x) + \varepsilon \tag{1.1.1}$$

where little is assumed about the function f except that it is smooth. In the simplest models, the residuals are independently and identically distributed with mean zero and constant variance σ_ε^2 and the x's are generated by a process which ensures that eventually they are dense in the domain. Closeness of the x's combined with smoothness of f provides a basis for estimation of the regression function. By averaging or smoothing observations on y for which the corresponding x's are close to a given point, say x_o, one obtains a reasonable estimate of the regression effect $f(x_o)$.

This simple assumption -- that x's which are close will have corresponding values of the regression function which are also close -- may also be used to *remove* the regression effect. It is this removal or *differencing* which is the fundamental theme we explore in this essay. To illustrate the idea of differencing we will rely upon four applications: estimation of the residual variance σ_ε^2, a specification test on the regression function f, a test of equality of nonparametric regression functions and estimation and inference in the partial linear model $y = z\beta + f(x) + \varepsilon$.

1.2 Estimation of the Residual Variance

Suppose one has data $(y_1, x_1), \ldots, (y_n, x_n)$ on the pure nonparametric regression model (1.1.1) where x is a scalar lying in the unit interval, ε is i.i.d with $E(\varepsilon \mid x) = 0$, $Var(\varepsilon \mid x) = \sigma_\varepsilon^2$ and all that is known about f is that its first derivative is bounded. Most important, the data have been rearranged so that $x_1 \le \ldots \le x_n$. Consider the following estimator of σ_ε^2:

$$s_{diff}^2 = \frac{1}{2n}\sum_{i=2}^{n}\left(y_i - y_{i-1}\right)^2 \tag{1.2.1}$$

The estimator is consistent because as the $x's$ become close, differencing tends to remove the nonparametric effect: $y_i - y_{i-1} = f\left(x_i\right) - f\left(x_{i-1}\right) + \varepsilon_i - \varepsilon_{i-1} \cong \varepsilon_i - \varepsilon_{i-1}$ so that

$$s_{diff}^2 \cong \frac{1}{2n}\sum_{i=2}^{n}\left(\varepsilon_i - \varepsilon_{i-1}\right)^2 \cong \frac{1}{n}\sum_{i=1}^{n}\varepsilon_i^2 - \frac{1}{n}\sum_{i=2}^{n}\varepsilon_i\varepsilon_{i-1} \tag{1.2.2}$$

An obvious advantage of s_{diff}^2 is that no initial estimate of the regression function f needs to be calculated. *Indeed no consistent estimate of f is implicit in (1.2.1).* Nevertheless, the terms in s_{diff}^2 which involve f converge to zero sufficiently quickly so that the asymptotic distribution of the estimator can be derived directly from the approximation in (1.2.2). In particular,

$$n^{1/2}\left(s_{diff}^2 - \sigma_\varepsilon^2\right) \overset{D}{\sim} N\left(0, E\left(\varepsilon^4\right)\right) \tag{1.2.3}$$

1.3 Specification Test for f

Suppose one wants to test the null hypothesis that f is say a linear function. Let s_{res}^2 be the usual estimate of the residual variance obtained from a linear regression of y on x. If the linear model is correct then s_{res}^2 will be approximately equal to the average of the true squared residuals:

$$s_{res}^2 = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \hat{\gamma}_1 - \hat{\gamma}_2 x_i\right)^2 \cong \frac{1}{n}\sum_{i=1}^{n}\varepsilon_i^2 \tag{1.3.1}$$

If the linear specification is incorrect then s_{res}^2 will overestimate the residual variance while s_{diff}^2 will remain a consistent estimator, thus forming the basis of a test. Consider the test statistic:

$$V = \frac{n^{1/2}\left(s^2_{res} - s^2_{diff}\right)}{s^2_{diff}} \tag{1.3.2}$$

Using (1.2.2) and (1.3.1) the numerator of V is approximately equal to:

$$n^{1/2}\left(\frac{1}{n}\sum \varepsilon_i^2 - \frac{1}{2n}\sum\left(\varepsilon_i - \varepsilon_{i-1}\right)^2\right) \cong n^{1/2}\frac{1}{n}\sum \varepsilon_i \varepsilon_{i-1} \overset{D}{\sim} N\left(0, \sigma_\varepsilon^4\right) \tag{1.3.3}$$

Since s^2_{diff}, the denominator of V, is a consistent estimator of σ^2_ε then V is asymptotically N(0,1) under H_o.

As we shall see later, this test procedure may be used to test more general null hypotheses such as nonlinear regression models or even nonparametric regression models subject to monotonicity or concavity constraints. One simply inserts the restricted estimator of the variance in (1.3.2).

1.4 Test of Equality of Regression Functions

Suppose we are given data $\left(y_{1A}, x_{1A}\right),\ldots,\left(y_{nA}, x_{nA}\right)$ and $\left(y_{1B}, x_{1B}\right),\ldots,\left(y_{nB}, x_{nB}\right)$ from two possibly different regression models A and B. Assume x is a scalar and that *each data set has been reordered so that the x's are in increasing order*. The basic models are:

$$\begin{aligned} y_{iA} &= f_A\left(x_{iA}\right) + \varepsilon_{iA} \\ y_{iB} &= f_B\left(x_{iB}\right) + \varepsilon_{iB} \end{aligned} \tag{1.4.1}$$

where given the x's, the $\varepsilon's$ have mean 0, variance σ^2_ε and are independent within and between populations; f_A and f_B have first derivatives bounded. Using (1.2.1), define consistent 'within' differencing estimators of the variance:

$$s_A^2 = \frac{1}{2N}\sum_i^N \left(y_{iA} - y_{i-1\,A}\right)^2$$
$$s_B^2 = \frac{1}{2N}\sum_i^N \left(y_{iB} - y_{i-1\,B}\right)^2 \tag{1.4.2}$$

As we will do from time to time, we have dropped the subscript '*diff*'. Now pool *all* the data and reorder so that the pooled x's are in increasing order: $\left(y_1^*, x_1^*\right), \ldots\ldots, \left(y_{2n}^*, x_{2n}^*\right)$. (Note that the pooled data have only one subscript.) Applying the differencing estimator once again we have:

$$s_p^2 = \frac{1}{4n}\sum_j^{2n} \left(y_j^* - y_{j-1}^*\right)^2 \tag{1.4.3}$$

The basic idea behind the test procedure is to compare the pooled estimator with the average of the within estimators. If $f_A = f_B$, then the 'within' and 'pooled' estimators are consistent and should yield similar estimates. If $f_A \neq f_B$ then the 'within' estimators remain consistent while the 'pooled' estimator overestimates the residual variance as may be seen in Figure 1.4. To formalize this idea define the test statistic:

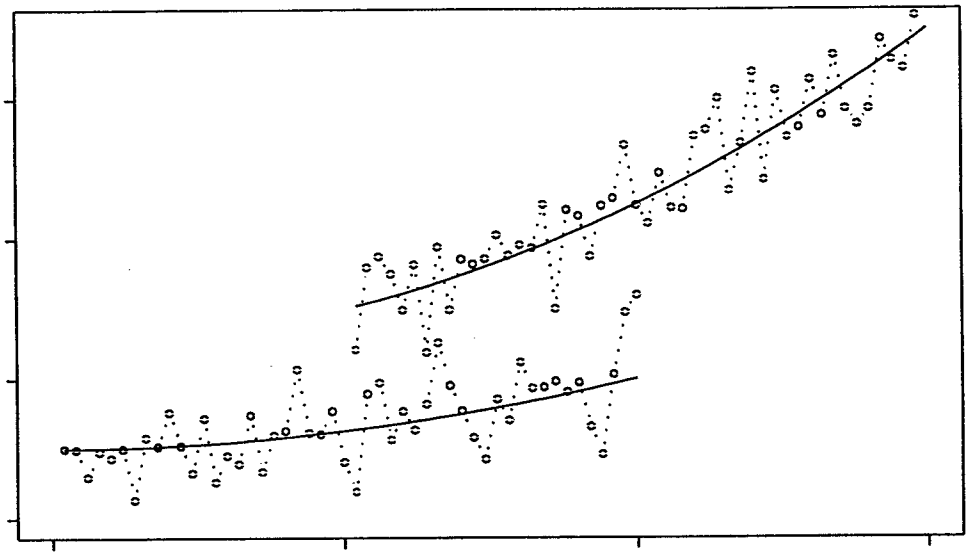
$$\Upsilon \equiv (2n)^{1/2}\left(s_p^2 - 1/2\left(s_A^2 + s_B^2\right)\right) \tag{1.4.4}$$

If $f_A = f_B$, then differencing removes the regression effect sufficiently quickly in both the 'within' and the 'pooled' estimators so that:

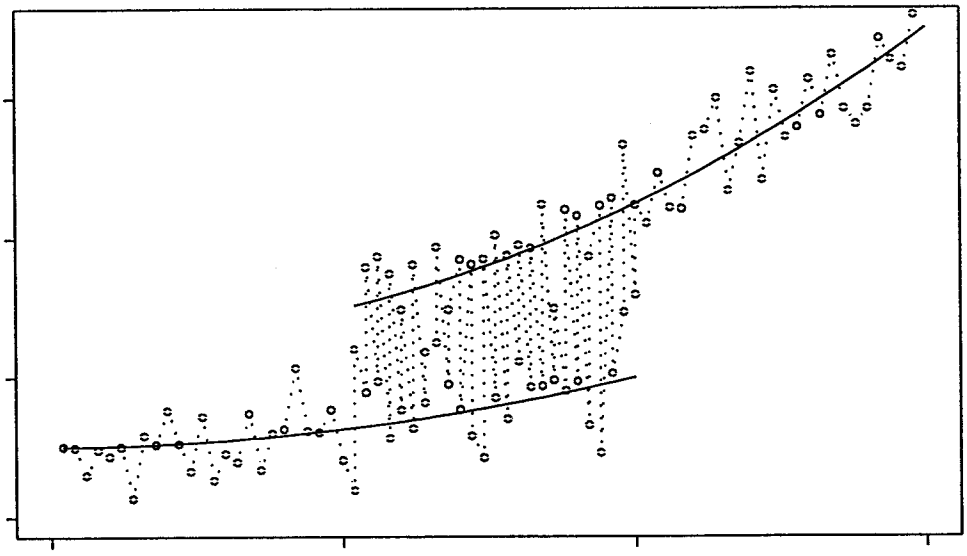
$$\begin{aligned}
\Upsilon &\equiv (2n)^{1/2}\left(s_p^2 - 1/2\left(s_A^2 + s_B^2\right)\right) \\
&\cong \frac{(2n)^{1/2}}{4n}\left(\sum_j^{2n}\left(\varepsilon_j^* - \varepsilon_{j-1}^*\right)^2 - \sum_i^n\left(\varepsilon_{iA} - \varepsilon_{i-1A}\right)^2 - \sum_i^n\left(\varepsilon_{iB} - \varepsilon_{i-1B}\right)^2\right) \\
&\cong \frac{(2n)^{1/2}}{2n}\left(\sum_j^{2n}\varepsilon_j^{*2} - \varepsilon_j^*\varepsilon_{j-1}^* - \sum_i^n \varepsilon_{iA}^2 - \varepsilon_{iA}\varepsilon_{i-1A} - \sum_i^n \varepsilon_{iB}^2 - \varepsilon_{iB}\varepsilon_{i-1B}\right) \\
&\cong \frac{1}{(2n)^{1/2}}\left(\sum_i^n \varepsilon_{iA}\varepsilon_{i-1A} + \sum_i^n \varepsilon_{iB}\varepsilon_{i-1B}\right) - \frac{1}{(2n)^{1/2}}\left(\sum_j^{2n}\varepsilon_j^*\varepsilon_{j-1}^*\right)
\end{aligned} \tag{1.4.5}$$

FIGURE 1.4: TESTING EQUALITY OF REGRESSION FUNCTIONS

Within Estimators of Residual Variance



Pooled Estimator of Residual Variance



Consider the two terms in the last line. In large samples, each is approximately $N\left(0, \sigma_\varepsilon^4\right)$. If observations which are consecutive in the individual data sets tend to be consecutive after pooling and reordering, then the *covariance* between the two terms will be large. In particular, the covariance is approximately $\sigma_\varepsilon^4(1-\pi)$ where $\pi =$ Prob [consecutive observations in the pooled reordered data set come from *different* populations]. From this it follows that under $H_o: f_A = f_B$,

$$\Upsilon \overset{D}{\sim} N\left(0, 2\pi\sigma_\varepsilon^4\right) \tag{1.4.6}$$

For example, if reordering the pooled data is equivalent to stacking data sets A and B -- because the two sets of x's, x_A and x_B do not intersect -- then $\pi \cong 0$ and indeed the statistic Υ becomes degenerate. This is not surprising, since observing nonparametric functions over different domains cannot provide a basis for testing whether they are the same. If the pooled data involve a simple interleaving of data sets A and B, then $\pi \cong 1$ and $\Upsilon \to N\left(0, 2\sigma_\varepsilon^4\right)$. If x_A and x_B are independent of each other but have the same distribution, then for the pooled reordered data, the probability that consecutive observations come from different populations is ½ and $\Upsilon \to N\left(0, \sigma_\varepsilon^4\right)$.[1] In order to implement the test, one may obtain a consistent estimate $\hat{\pi}$ by taking the proportion of observations in the pooled reordered data which are preceded by an observation from a different population.

1.5 The Partial Linear Model

Consider now the partial linear model $y = z\beta + f(x) + \varepsilon$ where for simplicity all variables are assumed to be scalars. As before, the x's have support the unit interval and have been rearranged so that $x_1 \le \ldots \le x_n$. Suppose that the conditional mean of z is a smooth function of x, say $E(z \mid x) = g(x)$ where g' is bounded and $Var(z \mid x) = \sigma_u^2$. Then we may rewrite $z = g(x) + u$. First differencing yields:

---

[1] For example, distribute n men and n women randomly along a stretch of beach facing the sunset. Then for any individual, the probability that the person to the left is of the opposite sex is ½. More generally, if x_A and x_B are independent of each other and have different distributions, then π depends on the *relative* density of observations from each of the two populations.

$$
\begin{aligned}
y_i - y_{i-1} &= \left(z_i - z_{i-1}\right)\beta + \left(f\left(x_i\right) - f\left(x_{i-1}\right)\right) + \varepsilon_i - \varepsilon_{i-1} \\
&= \left(g\left(x_i\right) - g\left(x_{i-1}\right)\right)\beta + \left(u_i - u_{i-1}\right)\beta + \left(f\left(x_i\right) - f\left(x_{i-1}\right)\right) + \varepsilon_i - \varepsilon_{i-1}
\end{aligned}
\tag{1.5.1}
$$

so that the direct effect $f(x)$ of the nonparametric variable x, and the indirect effect $g(x)$ which occurs through z, are removed through differencing. Suppose we apply the OLS estimator of β to the differenced data, i.e.,

$$
\hat{\beta}_{diff} = \frac{\sum \left(y_i - y_{i-1}\right)\left(z_i - z_{i-1}\right)}{\sum \left(z_i - z_{i-1}\right)^2} \tag{1.5.2}
$$

Then using the approximations $z_i - z_{i-1} \cong u_i - u_{i-1}$ and $y_i - y_{i-1} \cong \left(u_i - u_{i-1}\right)\beta + \varepsilon_i - \varepsilon_{i-1}$ we have

$$
n^{1/2}\left(\hat{\beta}_{diff} - \beta\right) \cong \frac{n^{1/2}\,\frac{1}{n}\sum \left(\varepsilon_i - \varepsilon_{i-1}\right)\left(u_i - u_{i-1}\right)}{\frac{1}{n}\sum \left(u_i - u_{i-1}\right)^2} \tag{1.5.3}
$$

The denominator converges to $2\sigma_u^2$ and the numerator has mean zero and variance $6\sigma_\varepsilon^2\sigma_u^2$. Thus, the ratio has mean zero and variance $6\sigma_\varepsilon^2\sigma_u^2/\left(2\sigma_u^2\right)^2 = 1.5\sigma_\varepsilon^2/\sigma_u^2$. Furthermore, using a finitely dependent central limit theorem, the ratio may be shown to be approximately normal. Thus we have:

$$
\hat{\beta}_{diff} \overset{D}{\sim} N\left(\beta, \frac{1.5}{n}\frac{\sigma_\varepsilon^2}{\sigma_u^2}\right) \tag{1.5.4}
$$

The most efficient estimator has variance $\sigma_\varepsilon^2/n\sigma_u^2$ so that the proposed estimator based on first differences has relative efficiency ⅔=1/1.5. In the upcoming chapter we will show how *higher* order differencing produces an estimator which is efficient.

Now, in order to use (1.5.4) to do inference, we will need consistent estimators of σ_ε^2 and σ_u^2. These may be obtained using:

$$s_{\varepsilon}^{2} = \frac{1}{2n}\sum_{i=2}^{n}\left(\left(y_i - y_{i-1}\right) - \left(z_i - z_{i-1}\right)\hat{\beta}_{diff}\right)^2 \cong \frac{1}{2n}\sum_{i=2}^{n}\left(\varepsilon_i - \varepsilon_{i-1}\right)^2 \overset{P}{\sim} \sigma_{\varepsilon}^{2} \tag{1.5.5}$$

and

$$s_{u}^{2} = \frac{1}{2n}\sum_{i=2}^{n}\left(z_i - z_{i-1}\right)^2 \cong \frac{1}{2n}\sum_{i=2}^{n}\left(u_i - u_{i-1}\right)^2 \overset{P}{\sim} \sigma_{u}^{2} \tag{1.5.6}$$

The above procedure generalizes straightforwardly to models where there are multiple parametric explanatory variables.

1.6 Empirical Application: Scale Economies in Electricity Distribution[2]

To illustrate these ideas, suppose we have a simple Cobb-Douglas model for the distribution of electricity:

$$\begin{aligned} tc &= f(cust) + \beta_1\, wage + \beta_2\, pcap \\ &+ \beta_3\, PUC + \beta_4\, kwh + \beta_5\, life + \beta_6\, lf + \beta_7\, kmwire + \varepsilon \end{aligned} \tag{1.6.1}$$

where *tc* is the log of total cost per customer, *cust* is the log of the number of customers, *wage* is the log wage rate, *pcap* is the log price of capital, *PUC* is a dummy variable for public utility commissions which deliver additional services and therefore may benefit from economies of scope, *life* is the log of the remaining life of distribution assets, *lf* is the log of the load factor (this measures capacity utilization relative to peak usage) and *kmwire* is the log of kilometers of distribution wire per customer. The data consist of 81 municipal distributors in Ontario, Canada during 1994. (See also Yatchew (1999).)

---

[2] Variable definitions for this and other empirical examples are summarized in Appendix C.

Since the data have been reordered so that the nonparametric variable *cust* is in increasing order, first differencing (1.6.1) tends to remove the nonparametric effect f. We also divide by $\sqrt{2}$ so that the residuals in the differenced equation (1.6.2) have the same variance as those in (1.6.1). Thus we have:

$$\begin{aligned}\left[tc - tc_{-1}\right]/\sqrt{2} \cong\ & \beta_1\left[wage - wage_{-1}\right]/\sqrt{2} + \beta_2\left[pcap - pcap_{-1}\right]/\sqrt{2} \\ & + \beta_3\left[PUC - PUC_{-1}\right]/\sqrt{2} + \beta_4\left[kwh - kwh_{-1}\right]/\sqrt{2} + \beta_5\left[life - life_{-1}\right]/\sqrt{2} \\ & + \beta_6\left[lf - lf_{-1}\right]/\sqrt{2} + \beta_7\left[kmwire - kmwire_{-1}\right]/\sqrt{2} + \left[\varepsilon - \varepsilon_{-1}\right]/\sqrt{2}\end{aligned} \tag{1.6.2}$$

Figure 1.6 summarizes our estimates of the parametric effects β using the differenced equation. It also contains estimates of a pure parametric specification where the scale effect f is modeled with a quadratic. Applying the specification test (1.3.2) where s^2_{diff} is obtained using (1.5.5) yields a value of 1.08 indicating that the quadratic model appears to be adequate.

Thus far our results suggest that by differencing, we can perform inference on β as if there were no nonparametric component f in the model to begin with. But having estimated β, we can then proceed to apply a variety of nonparametric techniques to analyze f as if β were known. Such a modular approach simplifies implementation because it permits the use of existing software designed for pure nonparametric models.

More precisely, suppose we assemble the ordered pairs $\left(y_i - z_i\hat{\beta}_{diff}, x_i\right)$ then we have :

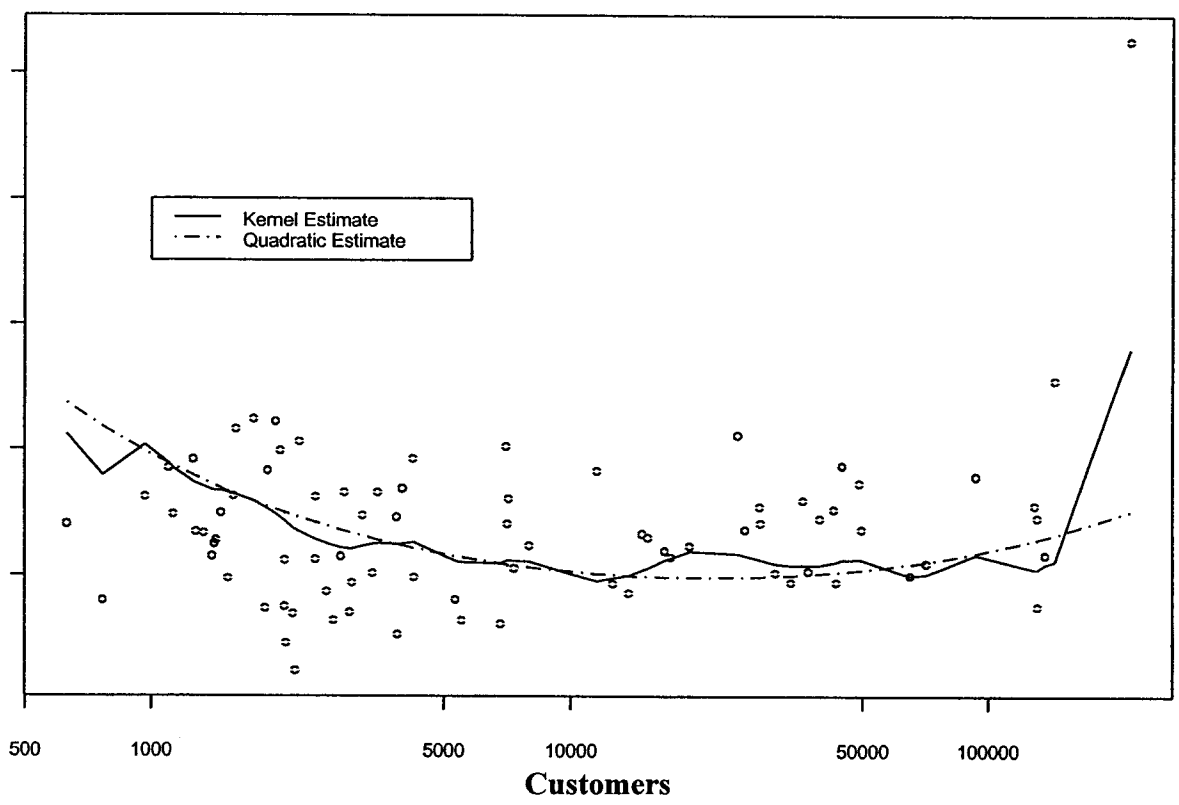
$$y_i - z_i\hat{\beta}_{diff} = z_i\left(\beta - \hat{\beta}_{diff}\right) + f\left(x_i\right) + \varepsilon_i \cong f\left(x_i\right) + \varepsilon_i \tag{1.6.3}$$

If we apply conventional kernel or spline estimation methods to the ordered pairs, then consistency, optimal rate of convergence results and the construction of confidence intervals for *f(·)* remain valid because $\hat{\beta}_{diff}$ converges sufficiently quickly to β so that the approximation in the last part of (1.6.3) leaves asymptotic arguments unaffected. (This is indeed why we could apply the specification test after removing the *estimated* parametric effect.) Thus, in Figure 1.6 we have also plotted kernel estimates of f. In subsequent sections we will elaborate this example further as well as providing additional ones.

FIGURE 1.6: PARTIAL LINEAR MODEL — LOGLINEAR COST FUNCTION
SCALE ECONOMIES IN ELECTRICITY DISTRIBUTION

VARIABLE	QUADRATIC MODEL		PARTIAL LINEAR MODEL [1]	
	Coef	SE	Coef	SE
cust	-0.8566	0.1786	-	-
cust^2	0.0860	0.0190	-	-
wage	0.6407	0.3243	0.2489	0.3969
pcap	0.5292	0.0732	0.4505	0.0815
PUC	-0.0821	0.0391	-0.0966	0.0458
kwh	0.0020	0.0884	-0.0086	0.0956
life	-0.6097	0.1201	-0.5482	0.1413
lf	0.5742	0.2146	0.4738	0.2372
kmwire	0.3989	0.0869	0.3279	0.1036
s_ε^2	.0214		.0191	
R^2	.616		.657	

ESTIMATED SCALE EFFECT



[1] Test of quadratic vs nonparametric specification of scale effect: $V = n^{1/2}\left(s^2_{res} - s^2_{diff}\right) / s^2_{diff} = 81^{1/2}(.0214 - .0191)/.0191 = 1.08$ where V is N(0,1), Section 1.3.

1.7 Why Differencing?

An important advantage of using differencing procedures is their simplicity. Consider once again the partial linear model $y = z\beta + f(x) + \varepsilon$. Conventional estimators, such as the one proposed by Robinson (1988), require one to estimate $E(y|x)$ and $E(z|x)$ using nonparametric regressions. The estimated residuals from each of these regressions (hence the term 'the double residuals method') are then used to estimate the *parametric* regression:

$$y - E(y|x) = (z - E(z|x))\beta + \varepsilon \tag{1.7.1}$$

If z is a vector, then a separate nonparametric regression is run for each component of z where the independent variable is the nonparametric variable x. In contrast, differencing eliminates these first stage regressions so that estimation of β can be performed -- regardless of its dimension -- even if nonparametric regression procedures are not available within the software being used. Similarly, tests of parametric specifications against nonparametric alternatives and tests of equality of regression functions across two or more (sub-) samples can be performed without performing a nonparametric regression.

As should be evident from the empirical example of the last section, differencing may be easily combined with other procedures. In that example, we used differencing to estimate the parametric component of a partial linear model. We then removed the estimated parametric effect and applied conventional nonparametric procedures to analyze the nonparametric component. Such modular analysis does require theoretical justification, which we will provide in Section 3.1 below.

As we have seen, the partial linear model permits a simple semiparametric generalization of the Cobb-Douglas model. Translog and other linear-in-parameters models may be generalized similarly. If we allow the parametric portion of the model to be nonlinear — so that we have a partial parametric model — then we may also obtain simple semiparametric generalizations of models such as the CES cost function. These too may be estimated straightforwardly using differencing, (Section 2.7). The key requirement is that the parametric and nonparametric portions of the model be additively separable.

Other procedures commonly used by the econometrician may be imported with relative ease into the

differencing setting. If some of the parametric variables are potentially correlated with the residuals, instrumental variable techniques can be applied, with suitable modification, as can the Hausman endogeneity test (Section 2.8). If the residuals are not independently and identically distributed then well known techniques such as White's heteroskedasticity consistent standard errors or the Newey-West standard errors can be adapted (Sections 5.1-5.4). The reader will no doubt find other procedures that can be readily transplanted.

Earlier we have pointed out that the first order differencing estimator of β in the partial linear model is inefficient when compared to the most efficient estimator (Section 1.5). The same is true for the first order differencing estimator of the residual variance (Section 1.2). This problem can be corrected using higher order differencing so long as the differencing coefficients are chosen judiciously -- as we will see in the next chapter.

The simplicity of differencing also provides a useful pedagogical device. Students can be introduced to nonparametric techniques early and with conventional econometric software. Indeed all the procedures in the example of Section 1.6 can be executed within packages such as *E-Views, SAS, Shazam, Stata* or *TSP*. Furthermore, since the partial linear model can easily accommodate multiple parametric variables, students can immediately apply these techniques to data that are of practical interest.

Simplicity and versatility, however, has a price. One of the criticisms of differencing is that it can result in greater bias in moderately sized samples than other estimators. [3] A second criticism is that differencing as proposed here works only if the dimension of the nonparametric component of the model does not exceed 3, (Section 5.5 below). Indeed, for most of this essay we will focus on the case where the nonparametric variable is a scalar. However, we would argue that even if differencing techniques were limited to one (nonparametric) dimension, they have the potential of significant 'market share'. The reason is that high dimensional nonparametric regression models, (unless they rely on additional structure such as additive separability), suffer from the 'curse of dimensionality' which severely limits one's ability to estimate the regression relationship with any degree of precision. It is not surprising therefore that the vast majority of applied papers which use nonparametric regression, limit the nonparametric component to 1 or 2 dimensions.

---

[3] Seifert, Gasser and Wolf (1993) study this issue for differencing estimators of the residual variance.

1.8 Notational Conventions

When applying the differencing procedures used in this essay, the first few observations may be treated differently or lost. For example, to calculate the differencing estimator of the residual variance $s^2_{diff} = \sum_i^n \left(y_i - y_{i-1}\right)^2 / 2n$, we begin the summation at $i=2$. For the mathematical arguments below, such effects are negligible. Thus, we will use the symbol $\doteq$ to denote 'equal except for end effects'. As must be evident by now, we will also be using the symbol $\cong$ to denote approximate equality, $\xrightarrow{P}$ for convergence in probability and $\xrightarrow{D}$ for convergence in distribution .

Since differencing is the main focus of this essay, many of estimators merit the subscript '*diff*' as in the above paragraph or in (1.2.1) or (1.5.2). For simplicity, we will sometimes suppress this annotation. However, we will frequently use subscripts to denote components of vectors or matrices, e.g., β_i, A_{ij} or $[AB]_{ij}$.

For any two matrices A, B of identical dimension we will use the notation $\left[A \odot B\right]_{ij} = A_{ij} B_{ij}$.

2. Higher Order Optimal Differencing Procedures

2.1 Differencing Matrices

Let m be the order of differencing and $d_0, d_1, \ldots, d_m$ differencing weights which satisfy the conditions:

$$\sum_{j=0}^{m} d_j = 0 \qquad \sum_{j=0}^{m} d_j^2 = 1 \tag{2.1.1}$$

Define the differencing matrix:

$$\underset{nxn}{D} = \begin{bmatrix} d_0, d_1, d_2, ..d_m, 0, \ldots\ldots\ldots\ldots, 0 \\ 0,\ d_0, d_1, d_2, ..d_m, 0, \ldots\ldots\ldots\ldots 0 \\ \vdots \quad \vdots \\ \vdots \quad \vdots \\ 0, \ldots\ldots\ldots\ldots 0,\ d_0, d_1, d_2, ..d_m, 0 \\ 0, \ldots\ldots\ldots\ldots 0, d_0, d_1, d_2, ..d_m \\ 0, \ldots\ldots\ldots\ldots\ldots\ldots 0 \\ \vdots \quad \vdots \\ 0, \ldots\ldots\ldots\ldots\ldots\ldots 0 \end{bmatrix} \tag{2.1.2}$$

It will be convenient to use lag matrices L_i. For $i>0$ define L_i to have 0's everywhere except on the i-th diagonal below the main diagonal where it has 1's. If $i<0$, L_i has 1's on the i-th diagonal above the main diagonal. L_0 is defined to be the usual identity matrix, $L_i' = L_{-i}$ and $L_i L_j \doteq L_{i+j}$. It is evident from (2.1.2) that, (except for end effects), the differencing matrix of order m is a weighted sum of lag matrices, that is,

$$D \doteq d_0 L_0 + d_1 L_1' + \ldots + d_m L_m' \tag{2.1.3}$$

We will need the matrix $D'D$ which has a symmetric band structure with ones on the main diagonal (due to the normalization restriction in (2.1.1)). Hence $tr\left(D'D\right) \doteq n$. As one moves away from the main diagonal, consecutive diagonals take values $\Sigma_{j=0}^{m-k} d_j d_{j+k}$ k=1,...,m. The remainder of the matrix is zero. Equivalently, using (2.1.3) and the properties of lag matrices one has:

$$D'D \doteq L_0 + \left(L_1 + L_1'\right)\sum_{j=0}^{m-1} d_j d_{j+1} + \left(L_2 + L_2'\right)\sum_{j=0}^{m-2} d_j d_{j+2} + \ldots\ldots + \left(L_{m-1} + L_{m-1}'\right)\sum_{j=0}^{1} d_j d_{j+m-1} + \left(L_m + L_m'\right) d_0 d_m \quad (2.1.4)$$

Since band structure is preserved by matrix multiplication, the matrix $D'DD'D$ will also have this property as well as being symmetric. The value on the main diagonal is easily determined by multiplying the expansion in (2.1.4) by itself. Note that only products of the form $L_k L_k'$ and $L_k' L_k$ yield (except for end effects) the identity L_0. Thus, the common diagonal value of $D'DD'D$ will be the sum of the coefficients of $L_0, L_1'L_1, L_1L_1', L_2'L_2, L_2L_2', \ldots L_m'L_m, L_mL_m'$, that is,

$$\left[D'DD'D\right]_{ii} \doteq 1 + 2\sum_{k=1}^{m}\left(\sum_{j=0}^{m-k} d_j d_{j+k}\right)^2 \quad (2.1.5)$$

A particularly useful quantity will be:

$$\delta = \sum_{k=1}^{m}\left(\sum_{j=0}^{m-k} d_j d_{j+k}\right)^2 \quad (2.1.6)$$

We now have $\left[D'DD'D\right]_{ii} \doteq 1 + 2\delta$ and $tr\left(D'DD'D\right) \doteq n\left(1 + 2\delta\right)$. [4]

[4] The trace of $D'DD'D$ may be obtained alternatively as follows. Since for any symmetric matrix A, tr (AA) is the sum of squares of elements of A, we may use (2.1.4) to conclude that $tr\left(D'DD'D\right) \doteq n\left(1 + 2\delta\right)$.

2.2 Variance Estimation

Let us return to the problem of estimating the residual variance in a pure nonparametric regression model $y_i = f(x_i) + \varepsilon_i$ where $\varepsilon \mid x$ is distributed with mean 0, variance σ_ε^2 and $E(\varepsilon^4 | x) = \eta$ and f has first derivative bounded. Given independently and identically distributed observations on the model $(y_1, x_1) \ldots (y_n, x_n)$ where the x's have been reordered so that they are in increasing order, define $y' = (y_1, \ldots, y_n)$ and $f(x)' = (f(x_1), \ldots, f(x_n))$. In vector notation we have

$$y = f(x) + \varepsilon \tag{2.2.1}$$

Applying the differencing matrix we have

$$Dy = Df(x) + D\varepsilon \tag{2.2.2}$$

A typical element of the vector Dy is of the form

$$d_0 y_i + \ldots + d_m y_{i+m} = d_0 f(x_i) + \ldots + d_m f(x_{i+m}) + d_0 \varepsilon_i + \ldots + d_m \varepsilon_{i+m} \tag{2.2.3}$$

so that the role of the constraints (2.1.1) is now evident. The first condition ensures that as the x's become close, the nonparametric effect is removed. The second condition ensures that the variance of the weighted sum of the residuals remains equal to σ_ε^2.

The m-th order differencing estimator of the residual variance is now defined to be

$$s_{diff}^2 = \frac{1}{n} \sum_{i=1}^{n-m} \left(d_0 y_i + d_1 y_{i+1} + \ldots + d_m y_{i+m} \right)^2 = \frac{1}{n} y' D' D y \tag{2.2.4}$$

In large samples we have

$$s_{diff}^2 \cong \frac{1}{n} \varepsilon' D' D \varepsilon \tag{2.2.5}$$

Using the mean and variance of a quadratic form, (see Appendix A, Lemma A.1), we have:

$$E\left(s^2_{diff}\right) \cong \sigma^2_\varepsilon \frac{1}{n} tr\left(D'D\right) \doteq \sigma^2_\varepsilon \tag{2.2.6}$$

and

$$Var\left(s^2_{diff}\right) \cong \frac{1}{n}\left[\left(\eta_\varepsilon - 3\sigma^4_\varepsilon\right) + \sigma^4_\varepsilon \frac{2}{n} tr\left(D'DD'D\right)\right] = \frac{1}{n}\left[Var\left(\varepsilon^2\right) + 4\,\sigma^4_\varepsilon\,\delta\right] \tag{2.2.7}$$

Thus to minimize the large sample variance of the differencing estimator one needs to make δ as small as possible. Using time series techniques, Hall, Kay and Titterington (1990) show that if the d_j are selected to minimize δ then

$$\sum_{j=0}^{m-k} d_j d_{j+k} = -\frac{1}{2m}, \quad k = 1,2,\ldots,m \qquad and \qquad \delta = \frac{1}{4m} \tag{2.2.8}$$

In this case matrix $D'D$ has (except for end effects), ones on the main diagonal, $-1/2m$ on the m adjacent diagonals and zeros elsewhere. That is,

$$D'D \doteq L_0 - \frac{1}{2m}\left(L_1 + L_1' + \ldots + L_m + L_m'\right) \tag{2.2.9}$$

so that $tr\left(D'D\right) \doteq n$. [5] Thus if optimal differencing coefficients are used then:

$$s^2_{diff} \cong \frac{\varepsilon'\varepsilon}{n} - \frac{1}{2mn}\varepsilon'\left(L_1 + L_1' + \ldots + L_m + L_m'\right)\varepsilon \tag{2.2.10a}$$

---

[5] The matrix $D'DD'D$ has a symmetric band structure with $1 + 1/2m$ on the main diagonal so that $tr\left(D'DD'D\right) \doteq n\left(1 + 1/2m\right)$. The first m diagonals adjacent to the main diagonal take the value: $-\frac{1}{m} + j\frac{1}{4m^2}$, $j = 2m-2, 2m-3, \ldots, m-1$. The next m diagonals take the value: $j\frac{1}{4m^2}$, $j = m, m-1, \ldots, 1$. The remainder of the matrix is 0.

and applying (2.2.7) and (2.2.8)

$$Var\left(s^2_{diff}\right) \cong \frac{1}{n}\left(Var\left(\varepsilon^2\right) + \frac{\sigma^4_\varepsilon}{m}\right) \tag{2.2.10b}$$

On the other hand, if the regression function were parametric, for example if it were known to be linear and we used a conventional OLS estimator then:

$$s^2_{OLS} = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \hat{\gamma}_1 - \hat{\gamma}_2 x_i\right)^2 \cong \frac{1}{n}\sum_{i=1}^{n}\varepsilon_i^2 + O_P\left(\frac{1}{n}\right) \tag{2.2.11a}$$

in which case

$$Var\left(s^2_{OLS}\right) \cong \frac{1}{n}Var\left(\varepsilon^2\right) \tag{2.2.11b}$$

Comparing the variances of the two residual estimators, (2.2.10b) and (2.2.11b) we see that as the order of differencing m increases, the variance of the differencing estimator approaches that of parametric based estimators.

Optimal differencing weights do not have analytic expressions but may be calculated easily using an optimization routine. Hall, Kay and Titterington (1990) present weights to order $m = 10$. These contain some minor errors. Figure 2.1 contains corrected weights. Appendix B discusses calculation of optimal weights. Weights up to order m=500 are available from the author upon request.

FIGURE 2.1: OPTIMAL DIFFERENCING WEIGHTS	
m	$(d_0, d_1, \ldots, d_m)$
1	(0.7071,-.0.7071)
2	(0.8090,-0.5000,-0.3090)
3	(0.8582,-0.3832,-0.2809,-0.1942)
4	(0.8873,-0.3099,-0.2464,-0.1901,-0.1409)
5	(0.9064,-0.2600,-0.2167,-0.1774,-0.1420,-0.1103)
6	(0.9200,-0.2238,-0.1925,-0.1635,-0.1369,-0.1126.-0.0906)
7	(0.9302,-0.1965,-0.1728,-0.1506,-0.1299,-0.1107.-0.0930,-0.0768)
8	(0.9380,-0.1751,-0.1565,-0.1389,-0.1224,-0.1069.-0.0925,-0.0791,-0.0666)
9	(0.9443,-0.1578,-0.1429,-0.1287,-0.1152,-0.1025.-0.0905,-0.0792.-0.0687,-0.0588)
10	(0.9494,-0.1437,-0.1314,-0.1197,-0.1085,-0.0978.-0.0877,-0.0782.-0.0691,-0.0606,-.0527)
In contrast to those in Hall et al (1990), all the above optimal weight sequences decline in absolute value towards zero.	

Proposition 2.2.1: Let $d_0, d_1, \ldots, d_m$ be optimal differencing weights then

$$n^{1/2}\left(s^2_{diff} - \sigma^2_\varepsilon\right) \sim N\left(0,\ \eta_\varepsilon - \sigma^4_\varepsilon + \frac{\sigma^4_\varepsilon}{m}\right) \qquad (2.2.12)\ \blacksquare$$

In order to make use of this results, we will need a consistent estimator of $\eta_\varepsilon = E\left(\varepsilon^4\right)$ for which we will use fourth order powers of the differenced data. To motivate the estimator it is convenient to first establish the following:

$$E\left(d_0\varepsilon_i + \ldots + d_m\varepsilon_{i+m}\right)^4 = \eta_\varepsilon\left(\sum_{i=0}^{m} d_i^4\right) + 6\sigma^4_\varepsilon\left(\sum_{i=0}^{m-1} d_i^2 \sum_{j=i+1}^{m} d_j^2\right) \qquad (2.2.13)$$

This result may be obtained by first noting that in expanding the left hand side only two types of terms will have non-zero expectations. Those which involve the fourth power of a residual, e.g., $E\varepsilon_i^4 = \eta_\varepsilon$ and those which involve products of squares of residuals, e.g., $E\varepsilon_i^2\varepsilon_{i+j}^2 = \sigma^4_\varepsilon$, $j \neq 0$. Equation (2.2.13) is then obtained by summing the coefficients of such terms. We now have:

Proposition 2.2.2: Let

$$\hat{\eta}_{\varepsilon} = \frac{\frac{1}{n}\sum_{i=m+1}^{n}\left(d_0 y_i + \ldots + d_m y_{i+m}\right)^4 - 6\left(s_{diff}^2\right)^2\left(\sum_{i=0}^{m-1} d_i^2 \sum_{j=i+1}^{m} d_j^2\right)}{\sum_{i=0}^{m} d_i^4} \tag{2.2.14}$$

then $\hat{\eta}_{\varepsilon} \overset{P}{\to} \eta_{\varepsilon}$. ■

Equations (2.2.13) and (2.2.14) are valid for arbitrary differencing weights.

These results may be used to test equality of residual variances for two possibly different regression models $y_A = f_A(x_A) + \varepsilon_A$ and $y_B = f_B(x_B) + \varepsilon_B$. Let s_A^2, $\hat{\eta}_A$, s_B^2, $\hat{\eta}_B$ be differencing estimators of the residual variances and fourth order moments using (2.2.4) and (2.2.14). Then using Propositions 2.2.1 and 2.2.2 we have under the null hypothesis:

$$\frac{s_A^2 - s_B^2}{\left(\frac{\hat{\eta}_A + (1/m-1)s_A^4}{n_A} + \frac{\hat{\eta}_B + (1/m-1)s_B^4}{n_B}\right)^{1/2}} \sim N(0,1) \tag{2.2.15}$$

2.3 Specification Test [6]

We remain with the pure nonparametric model $y = f(x) + \varepsilon$ where f has first derivative bounded. Let $h(x, \gamma)$ be a known function of x and an unknown parameter γ. We wish to test the null hypothesis that the regression function has the parametric form $h(x, \gamma)$ against the nonparametric alternative $f(x)$. Let $\hat{\gamma}_{LS}$ be obtained using for example parametric nonlinear least squares. Define the restricted estimator of the residual variance:

$$s_{res}^2 = \frac{1}{n} \sum \left(y_i - h\left(x_i, \hat{\gamma}_{LS}\right) \right)^2 \tag{2.3.1}$$

Proposition 2.3.1 Suppose $H_o : f(x) = h(x, \gamma)$ is true where h is a known function. Define s_{res}^2 as in (2.3.1). If one uses optimal differencing weights then

$$(m\,n)^{1/2} \frac{\left(s_{res}^2 - s_{diff}^2 \right)}{s_{diff}^2} \overset{D}{\sim} N(0, 1) \tag{2.3.2}$$ ■

In the denominator, s_{diff}^2 may be replaced by s_{res}^2 since under the null, both estimators of the residual variance are consistent.

A significance test on x is a special case of the above procedure. In this case, f is a constant function and so the restricted estimator of the regression function is just the sample mean of the y_i.

---

[6] A variety of results are available for testing a parametric null against a nonparametric alternative. These include Azzalini, Bowman and Hardle (1989), Eubank and Spiegelman (1990), Lee (1991), Wooldridge (1992), Gozalo (1993), Hardle and Mammen (1993), Whang and Andrews (1993) and Horowitz and Härdle (1994), Li (1994), Hong and White (1995) and Zheng (1996).

2.4 Test of Equality of Regression Functions [7]

Suppose we have samples of size n from each of T populations. Our model is given by $y_{it} = f_t(x_{it}) + \varepsilon_{it}$ where conditional on x_{it}, $E(\varepsilon_{it}) = 0$, $Var(\varepsilon_{it}) = \sigma_\varepsilon^2$ and $E(\varepsilon_{it}^4) = \eta_\varepsilon$. For each t, (y_{it}, x_{it}) are i.i.d., the x_{it} have bounded domain and f_t has a bounded first derivative.

Let $\varepsilon_t = (\varepsilon_{1t},\ldots,\varepsilon_{nt})'$ be the n-dimensional column vector of residuals from population t and $\varepsilon = (\varepsilon_1',\ldots,\varepsilon_T')'$ the nT-dimensional concatenation of these column vectors. Define $y_t = (y_{1t},\ldots,y_{nt})'$, $y = (y_1',\ldots,y_T')'$, $x_t = (x_{1t},\ldots,x_{nt})'$ and $x = (x_1',\ldots,x_T')'$ in a similar fashion. We emphasize that the data have already been ordered so that *within* each population, the x's are in increasing order, i.e., $x_{1t} \le \ldots \le x_{nt}$, $t = 1,\ldots,T$. For purposes of asymptotics, we will assume that T is fixed while $n \to \infty$.

Define the 'within' estimator of σ_ε^2:

$$s_w^2 = \frac{1}{nT} y'(I_T \otimes D')(I_T \otimes D)y = \frac{1}{nT} y'(I_T \otimes D'D)\, y \tag{2.4.1}$$

Define P_p the 'pooled' permutation matrix to be the matrix which reorders data so that the pooled x's are in increasing order. Thus if $x^* = P_p x$ then the consecutive elements of the reordered vector of length nT are in increasing order. Define the 'pooled' estimator of the variance to be:

$$s_p^2 = \frac{1}{nT} y'P_p'(I_T \otimes D')(I_T \otimes D)P_p y = \frac{1}{nT} y'P_p'(I_T \otimes D'D)\, P_p y \tag{2.4.2}$$

---

[7] A number of procedures are available for testing equality of nonparametric regression functions. These include procedures of Hall and Hart (1990), Härdle and Marron (1990), King, Hart and Wehrly (1991), Delgado (1993), Kulasekera (1995), Young and Bowman (1995), Baltagi, Hidalgo and Li (1996) and Kulasekera and Wang (1997). See also Hart (1997, p.236). Typically, these tests involve direct comparison of nonparametric estimates of regression curves or analysis of residuals from such regressions. (Certain tests also require that data be available at the same values of the x's for each regression function.) Our test circumvents such issues by avoiding the calculation of nonparametric regressions.

In (2.4.2) we may replace $I_T \otimes D'D$ by an nT-dimensional differencing matrix with structure as in (2.1.2).

PROPOSITION 2.4.1: $(nT)^{1/2}\left(s_w^2 - \sigma_\varepsilon^2\right) \overset{D}{\sim} N\left(0, \eta_\varepsilon + (1/m-1)\sigma_\varepsilon^4\right)$. If in addition all regression functions are identical then $(nT)^{1/2}\left(s_p^2 - \sigma_\varepsilon^2\right) \overset{D}{\sim} N\left(0, \eta_\varepsilon + (1/m-1)\sigma_\varepsilon^4\right)$. ■

The asymptotic variances of s_w^2, s_p^2 may be obtained using (2.2.10b), see also Appendix A, Lemma A.1 or Hall et al. (1990). Asymptotic normality follows from finitely dependent central limit theory. The following test procedure is now available.

PROPOSITION 2.4.2: Let $Q_\Upsilon = P_p'\left(I_T \otimes D'D\right)P_p - \left(I_T \otimes D'D\right)$ and $\hat{\pi}_\Upsilon = m\, tr\left(Q_\Upsilon Q_\Upsilon\right)/nT$. Suppose $\hat{\pi}_\Upsilon \overset{P}{\sim} \pi_\Upsilon > 0$. Then under the null hypothesis that all regression functions are identical:

$$\Upsilon = (mnT)^{1/2}\left(s_p^2 - s_w^2\right) = \frac{m^{1/2}}{(nT)^{1/2}} y' Q_\Upsilon y \overset{D}{\sim} N\left(0, 2\pi_\Upsilon \sigma_\varepsilon^4\right) \quad ■ \tag{2.4.3}$$

Thus, $\Upsilon / s_w^2\left(2\hat{\pi}_\Upsilon\right)^{1/2} \overset{D}{\sim} N(0,1)$ and one would reject for large positive values of the test statistic. The quantity π_Υ no longer admits the simple interpretation associated with π in the introductory Section 1.4, however it remains between 0 and 1 (see Appendix A, Comment on Proposition 2.4.2). The condition $\pi_\Upsilon > 0$ ensures that in the pooled reordered data, the proportion of observations that are 'near' observations from a *different* probability law does not go to 0. Ideally, of course, one would like the pooled x's to be intermingled so that s_p^2 contains many terms incorporating $f_s\left(x_j^*\right) - f_t\left(x_{j-1}^*\right)$, where $s \neq t$ and * denotes data reordered by P_p.

The test procedure described above is consistent. For example, suppose we use first order differencing, ($m = 1$) and x_A and x_B are independently and uniformly distributed on the unit interval. If $f_A \neq f_B$, then the within estimators in (1.4.2) remain consistent while the pooled estimator in (1.4.3) will converge as follows:

$$s_p^2 \rightarrow \sigma_\varepsilon^2 + \tfrac{1}{4}\int \left(f_A(x) - f_B(x)\right)^2 dx \tag{2.4.4}$$

and the mean of Υ in (1.4.4) diverges.[8] In general, power depends not only on the difference between the two regression functions but also on the degree to which data are generated from *both* populations at points where the difference is large. For fixed differencing order *m,* the procedure will detect local alternatives that converge to the null at a rate close to $n^{-1/4}$. The rate may be improved by permitting the order of differencing to grow with sample size. [9]

Modification to incorporate unequal sample sizes for each sub-population is also straightforward. Results along these lines are elaborated in the electricity example of Section 2.6.

Non-constant variances across equations can be incorporated as follows. Suppose

$$E\left(\varepsilon\varepsilon'\right) = \begin{vmatrix} \sigma_1^2 I_n & 0 & \cdots\cdots & 0 \\ 0 & \sigma_2^2 I_n & \cdots\cdots & 0 \\ \vdots & & \vdots & \\ 0 & \cdots & 0 & \sigma_T^2 I_n \end{vmatrix} = \Omega \tag{2.4.5}$$

A differencing estimator of the variance may be applied to each equation to obtain $\hat{\Omega}$. Test statistic (2.4.3) may be used and its variance may be estimated consistently using $s_\Upsilon^2 = 2m\, tr\left(\hat{\Omega} Q_\Upsilon \hat{\Omega} Q_\Upsilon\right) / nT$. (Use Lemma A.1b in Appendix A.) For example, with two equations A and B and m=1, equation (2.4.3) becomes $\Upsilon \to N\left(0, 2\pi_\Upsilon\left(\tfrac{1}{4}\sigma_A^4 + \tfrac{1}{2}\sigma_A^2\sigma_B^2 + \tfrac{1}{4}\sigma_B^4\right)\right)$.

The test procedure may be applied to the partial linear model. Consider $y_A = z_A\beta_A + f_A\left(x_A\right) + \varepsilon_A$ and $y_B = z_B\beta_B + f_B\left(x_B\right) + \varepsilon_B$. Suppose one obtains $n^{1/2}$ consistent estimators of β_A and β_B (e.g., by using

---

[8] The factor ¼ arises for two reasons. The main bias term in s_p^2 is $(1/4n)\sum_{j=2}^{2n}\left(f_j^*(x_j^*) - f_{j-1}^*(x_{j-1}^*)\right)^2$ which has a divisor of *4n* even though there are *2n* observations. Furthermore, about ½ the consecutive observations come from the same population and hence their asymptotic contribution to bias is negligible.

[9] Suppose $f_A - f_B = kg$ for scalar k and non-zero function g. Then given fixed m, the procedure is consistent for $k = O\left(n^{-1/4+\epsilon}\right)$ where ϵ is positive and arbitrarily close to zero. However, if one permits the differencing order to depend on sample size n, then the procedure is consistent if k goes to 0 at the faster rate $k = O\left(m(n)\,n\right)^{-1/4+\epsilon}$.

the differencing estimator in the upcoming section). To test $f_A = f_B$ one can apply equations (2.4.1)-(2.4.3) to $y_A - z_A\hat{\beta}_A \cong f_A(x_A) + \varepsilon_A$ and $y_B - z_B\hat{\beta}_B \cong f_B(x_B) + \varepsilon_B$ without altering the asymptotic properties of the procedure.

2.5 Partial Linear Model

Let us return to the problem of estimating the parameters in the partial linear regression model $y_i = z_i\beta + f(x_i) + \varepsilon_i$ where x_i is a scalar, z_i is a p-dimensional row vector, $\varepsilon_i \mid x_i, z_i$ is distributed with mean 0, variance σ_ε^2 and f has first derivative bounded. We will also assume that the vector of parametric variables z has a smooth regression relationship with the nonparametric variable x. That is, we may write $z_i = g(x_i) + u_i$ where g is a vector function with first derivatives bounded, $E(u_i \mid x_i) = 0$ and $E(Var(z_i \mid x_i)) = \Sigma_{z|x}$. Assume that the data (y_1, x_1, z_1) .... (y_n, x_n, z_n) have been reordered so that the x's are in increasing order. Define $y' = (y_1, \ldots, y_n)$, $f(x)' = (f(x_1), \ldots, f(x_n))$ and $Z = (z_1', \ldots, z_n')'$. In matrix notation we have:

$$y = f(x) + Z\beta + \varepsilon \tag{2.5.1}$$

Applying the differencing matrix we have

$$Dy = Df(x) + DZ\beta + D\varepsilon \tag{2.5.2}$$

The following contains our main result.

Proposition 2.5.1: If one uses optimal differencing weights then:

$$\hat{\beta}_{diff} = \left[(DZ)'DZ\right]^{-1}(DZ)'Dy \overset{A}{\sim} N\left(\beta, \left(1 + \frac{1}{2m}\right)\frac{\sigma_\varepsilon^2}{n}\Sigma_{z|x}^{-1}\right) \tag{2.5.3}$$

$$s_{diff}^2 = \frac{1}{n}\left(Dy - DZ\hat{\beta}_{diff}\right)'\left(Dy - DZ\hat{\beta}_{diff}\right) \overset{P}{\sim} \sigma_\varepsilon^2 \tag{2.5.4}$$

$$\hat{\Sigma}_{z|x} = \frac{1}{n}(DZ)'DZ \overset{P}{\sim} \Sigma_{z|x} \qquad (2.5.5)\blacksquare$$

The covariance matrix of the differencing estimator of β may be estimated using:

$$\hat{\Sigma}_{\hat{\beta}} = \left(1 + \frac{1}{2m}\right)\frac{s^2_{diff}}{n}\hat{\Sigma}^{-1}_{z|x} \qquad (2.5.6)$$

Linear restrictions of the form $R\beta = r$ may be tested using the conventional statistic which -- if the null hypothesis is true -- has the following distribution:

$$(R\hat{\beta} - r)'\left(R\hat{\Sigma}_{\hat{\beta}}R'\right)^{-1}(R\hat{\beta} - r) \overset{D}{\sim} \chi^2_{rank(R)} \qquad (2.5.7)$$

2.6 Empirical Applications

Household Gasoline Demand in Canada

In a concurrent paper, Yatchew and No (1999) estimate a partial linear model of household demand for gasoline in Canada — a model which is very similar to those estimated by Hausman and Newey (1995) and Schmalensee and Stoker (1999). The basic model is given by:

$$dist = f(price) + \beta_1\, income + \beta_2\, drivers + \beta_3\, hhsize + \beta_4\, URBANDUM + \beta_5\, RETIREDUM + \varepsilon \quad (2.6.1)$$

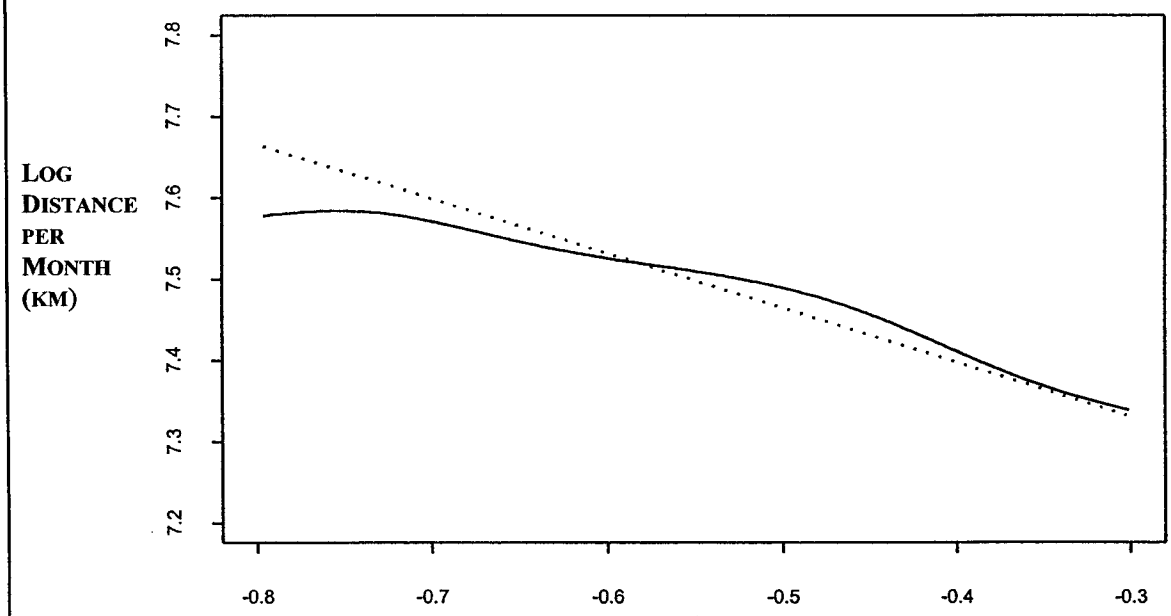
where *dist* is the log of distance traveled per month by the household, *price* is the log of price of a liter of gasoline, *drivers* is the log of the number of licensed drivers in the household, *hhsize* is the log of the size of the household, URBANDUM is a dummy for urban dwellers and RETIREDUM is a dummy for those households where the head is over the age of 65. Figure 2.6.1 summarizes the results. The 'Parametric Estimates' refer to a model where *price* enters linearly. The 'Double Residual Estimates' use Robinson (1988). That procedure requires one to estimate regression functions of the dependent variable and each of the parametric independent variables on the nonparametric variable. The residuals are then used to estimate the parametric effects. We implement Robinson's method using `ksmooth` a kernel regression estimation procedure in *S-Plus*. The 'Differencing Estimates' use third order differencing and Proposition 2.5.1 to estimate the parametric effects. The three sets of estimates are very similar except that the standard errors of the differencing estimates are 6% to 8% larger. This is consistent with equation (2.5.3) which would suggest that third order differencing should yield standard errors that are 8% larger ($\sqrt{1+1/(2*3)} = 1.0801$) than efficient estimators. If one uses 10-th order differencing, then there is little change in coefficient estimates, while standard errors are only about 2% larger which is again consistent with (2.5.3), ($\sqrt{1+1/(2*10)} = 1.0247$).

The estimated parametric effects are then removed and kernel regression is applied to obtain a nonparametric estimate of the price effect (the solid line in Figure 2.6.1). Applying the specification test in Proposition 2.3.1 yields a value of 3.1 which would suggest rejection of the loglinear specification.

FIGURE 2.6.1: HOUSEHOLD DEMAND FOR GASOLINE

VARIABLE	PARAMETRIC ESTIMATES		DOUBLE RESIDUAL ESTIMATES		DIFFERENCING ESTIMATES	
	Coef	SE	Coef	SE	Coef	SE
price	-0.6652	0.1264	-	-	-	-
income	0.2899	0.0234	0.2992	0.0234	0.2781	0.0251
drivers	0.5477	0.0365	0.5493	0.0366	0.5587	0.0389
hhsize	0.0829	0.0277	0.0768	0.0277	0.0887	0.0294
URBANDUM	-0.3002	0.0227	-0.2949	0.0227	-0.3057	0.0246
RETIREDUM	-0.2330	0.0316	-0.2289	0.0316	-0.2340	0.0338
s_ε^2	.5253		.5274		.5122	
R^2	.2278		.2247		.2471	

ESTIMATED PRICE EFFECT



LOG PRICE PER LITER ($CAD)

Variance of dependent variable is .6803. Order of differencing m=3. Number of observations is 4923. Robinson estimates of parametric effects produced using kernel procedure `ksmooth` in *S-Plus*. Solid line is kernel estimate applied to data after removal of estimated parametric effect. Dotted line is parametric estimate of price effect. Specification test of loglinear model for price effect yields value of 3.1.

Scale Economies in Electricity Distribution (continued)[10]

We now consider the example of Section 1.6 in considerably more detail. Suppose we have a slightly more general specification which is a semiparametric variant of the translog model, (variable definitions may be found in the Appendix C):

$$\begin{aligned} tc &= f(cust) + \beta_1 wage + \beta_2 pcap + \tfrac{1}{2}\beta_{11} wage^2 + \tfrac{1}{2}\beta_{22} pcap^2 + \beta_{12} wage \cdot pcap \\ &+ \beta_{31} cust \cdot wage + \beta_{32} cust \cdot pcap + \beta_4 PUC + \beta_5 kwh + \beta_6 life + \beta_7 lf + \beta_8 kmwire + \varepsilon \end{aligned} \tag{2.6.2}$$

Note that in addition to appearing nonparametrically, the scale variable *cust* interacts parametrically with wages and the price of capital. It is readily verified that if these interaction terms are zero (i.e. $\beta_{31} = \beta_{32} = 0$) then the cost function is homothetic. If in addition $\beta_{11} = \beta_{22} = \beta_{12} = 0$, then the model reduces to the loglinear specification of Section 1.6.

Differencing estimates of the parametric component of equation (2.6.2), are presented in Figure 2.6.2. (We use third order differencing (m=3)). We do not find significant statistical evidence against either the homothetic or the loglinear models. Estimates of non-price covariate effects exhibit little variation as one moves from the Full Model to the Homothetic and Loglinear Models.

We may now remove the estimated parametric effect from the dependent variable and analyse the nonparametric effect. In particular, for purposes of the tests below, the approximation $y_i - z_i\hat{\beta} = z_i(\beta - \hat{\beta}) + f(x_i) + \varepsilon_i \cong f(x_i) + \varepsilon_i$ does not alter the large sample properties of the procedures. We use the estimates of the loglinear model to remove the parametric effect.

Figure 2.6.2 displays the ordered pairs $\left(y_i - z_i\hat{\beta}_{diff}, x_i\right)$ as well as kernel and spline estimates of f. Parametric null hypotheses may be tested against nonparametric alternatives using the specification test in Section 2.3. If we insert a constant function for f then the procedure constitutes a test of significance of the scale variable x against a nonparametric alternative. The resulting statistic is 8.15 indicating a strong scale effect. Next we test a quadratic model for output. The resulting test statistic is 1.68. Since this is a one sided test, the asymptotic 5% critical value is 1.64 suggesting that the quadratic model may be inadequate.

---

[10] For a detailed treatment of these data, see Yatchew (1999).

To provide further illustrations of differencing procedures which are not contained in Yatchew (1999), we divide our data into two sub-populations: those that deliver additional services besides electricity, i.e., Public Utility Commissions, (PUCs) and those that are pure electricity distribution utilities (non-PUCs). The numbers of observations in the two sub-populations are $n_{PUC} = 37$ and $n_{nonPUC} = 44$. We denote differencing estimates of parametric effects and of residual variances as $\hat{\beta}_{PUC}$, $\hat{\beta}_{nonPUC}$, s^2_{PUC} and s^2_{nonPUC}. For each sub-population, we estimate the loglinear model using the differencing estimator and report the results in Figure 2.6.3.

To test whether PUCs and non-PUCs experience the same parametric effects we use:

$$\left(\hat{\beta}_{PUC} - \hat{\beta}_{nonPUC}\right)' \left(\hat{\Sigma}_{\hat{\beta}_{PUC}} + \hat{\Sigma}_{\hat{\beta}_{nonPUC}}\right)^{-1} \left(\hat{\beta}_{PUC} - \hat{\beta}_{nonPUC}\right)\left(R\hat{\beta} - r\right) \overset{D}{\sim} \chi^2_{dim(\beta)} \qquad (2.6.3)$$

The computed value of the χ^2_6 test statistic is 9.73, so that the null is not rejected. Next, we constrain the *parametric* effects to be equal across the two types of utilities while permitting distinct *nonparametric* effects. This is accomplished in the usual way by taking a weighted combination of the two estimates:

$$\hat{\beta}_{weighted} = \left[\hat{\Sigma}^{-1}_{\hat{\beta}_{PUC}} + \hat{\Sigma}^{-1}_{\hat{\beta}_{nonPUC}}\right]^{-1} \left[\hat{\Sigma}^{-1}_{\hat{\beta}_{PUC}} \cdot \hat{\beta}_{PUC} + \hat{\Sigma}^{-1}_{\hat{\beta}_{nonPUC}} \cdot \hat{\beta}_{nonPUC}\right] \qquad (2.6.4)$$

with estimated covariance matrix:

$$\hat{\Sigma}_{\hat{\beta}_{weighted}} = \left[\hat{\Sigma}^{-1}_{\hat{\beta}_{PUC}} + \hat{\Sigma}^{-1}_{\hat{\beta}_{nonPUC}}\right]^{-1} \qquad (2.6.5)$$

The coefficient estimates are reported in Figure 2.6.4. [11] To test whether the residual variances are

---

[11] A numerically similar estimator with the same large sample properties may be constructed by differencing the data within each sub-population, then stacking as follows:

$$\begin{bmatrix} Dy_{PUC} \\ Dy_{nonPUC} \end{bmatrix} = \begin{bmatrix} DZ_{PUC} \\ DZ_{nonPUC} \end{bmatrix} \beta + \begin{bmatrix} Df_{PUC}\left(x_{PUC}\right) \\ Df_{nonPUC}\left(x_{nonPUC}\right) \end{bmatrix} + \begin{bmatrix} Dv_{PUC} \\ Dv_{nonPUC} \end{bmatrix}$$

Let $\hat{\beta}$ be the OLS estimator applied to the above. Then the common residual variance may be estimated using::

equal in the two sub-populations, we use equation 2.2.15. The statistic which is N(0,1) under the null takes a value of -1.16.

Next we test whether the PUC and non-PUC data have the same nonparametric effects by adapting the results of Section 2.4 to the case of unequal sample sizes. Define the 'within' estimate to be the weighted average of the sub-population variance estimates keeping in mind that the estimated parametric effect has been removed using say $\hat{\beta}_{weighted}$:

$$s_w^2 = \frac{n_{PUC}}{n} s_{PUC}^2 + \frac{n_{nonPUC}}{n} s_{nonPUC}^2 \tag{2.6.6}$$

Now stack the PUC and non-PUC data to obtain the ordered pairs $\left(y_i - z_i \hat{\beta}_{weighted} , x_i \right)$ $i=1,\ldots,n$. Let P_p be the permutation matrix that reorders these data so that the nonparametric variable x is in increasing order. Note because separate equations were estimated for the two sub-populations, z does not contain the PUC dummy. Define:

$$s_p^2 = \frac{1}{n} \left(y - Z\hat{\beta}_{weighted} \right)' P_p' D' D P_p \left(y - Z\hat{\beta}_{weighted} \right) \tag{2.6.7}$$

If the null hypothesis is true, then differencing will still remove the nonparametric effect in the pooled data and s_p^2 will converge to σ_ε^2. Otherwise, it will generally converge to some larger value. In this setting with unequal sized sub-populations but with a total number of observations n, equation (2.4.3) becomes:

$$s^2 = \frac{1}{N} \left(\begin{bmatrix} Dy_{PUC} \\ Dy_{nonPUC} \end{bmatrix} - \begin{bmatrix} DZ_{PUC} \\ DZ_{nonPUC} \end{bmatrix} \hat{\beta} \right)' \left(\begin{bmatrix} Dy_{PUC} \\ Dy_{nonPUC} \end{bmatrix} - \begin{bmatrix} DZ_{PUC} \\ DZ_{nonPUC} \end{bmatrix} \hat{\beta} \right)$$

and the covariance matrix of $\hat{\beta}$ may be estimated using :

$$\hat{\Sigma}_{\hat{\beta}} = \left(1 + \frac{1}{2m} \right) \frac{s^2}{n} \left[\left(DZ_{PUC}\right)' \left(DZ_{PUC}\right) + \left(DZ_{nonPUC}\right)' \left(DZ_{nonPUC}\right) \right]^{-1}$$

$$\Upsilon = (mn)^{1/2}\left(s_p^2 - s_w^2\right) = \frac{m^{1/2}}{n^{1/2}} y' Q_\Upsilon y \overset{D}{\sim} N\left(0, 2\pi_\Upsilon \sigma_\varepsilon^4\right) \tag{2.6.8}$$

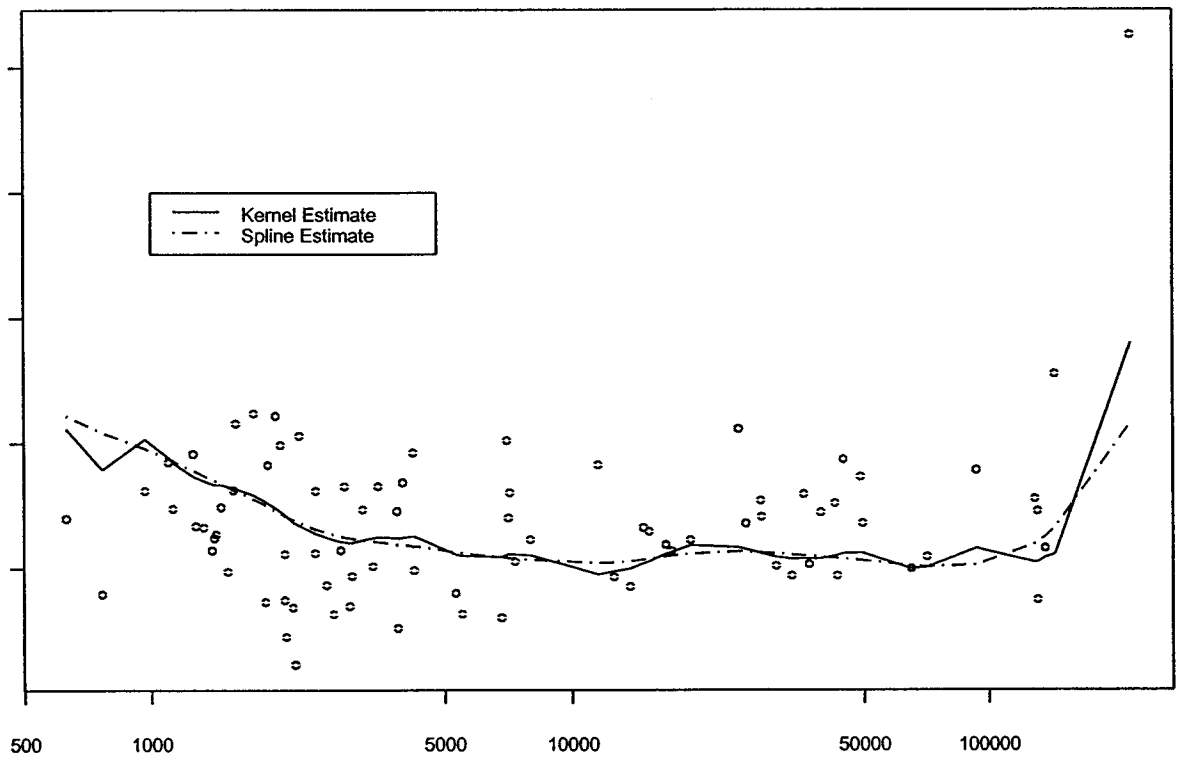
where $Q_\Upsilon = P_p' D' D P_p - D'D$ and $\hat{\pi}_\Upsilon = m\, tr\left(Q_\Upsilon Q_\Upsilon\right)/n$. Suppose $\hat{\pi}_\Upsilon \overset{P}{\sim} \pi_\Upsilon > 0$, then $\Upsilon / s_w^2 \left(2\hat{\pi}_\Upsilon\right)^{1/2} \overset{D}{\sim} N(0,1)$ and one would reject for large positive values of the test statistic.

If one wants to test whether the nonparametric regressions are *parallel*, then one should regress $DP_p\left(y - Z\hat{\beta}_{weighted}\right)$ on $DP_p z_{PUC}$ where z_{PUC} is a column of n_{PUC} ones followed by n_{nonPUC} zeros; s_p^2 would be the sum of squared residuals from this regression divided by *n*. These are in fact the test results which we report in Figure 2.6.3. The hypothesis of parallel nonparametric scale effects is not rejected, (the N(0,1) statistic is 1.49). We also plot kernel estimates of the scale effect *f(x)* for the two sub-populations: PUC's tend to have lower costs than non-PUC's.

FIGURE 2.6.2: SCALE ECONOMIES IN ELECTRICITY DISTRIBUTION

VARIABLE	FULL MODEL: SEMIPARAMETRIC TRANSLOG		HOMOTHETIC MODEL SEMIPARAMETRIC HOMOTHETIC		LOGLINEAR MODEL: SEMIPARAMETRIC COBB-DOUGLAS	
	Coef	SE	Coef	SE	Coef	SE
wage	1.6303	13.883	1.5082	12.999	0.3543	0.3170
pcap	-3.5079	2.5194	-2.0913	1.5894	0.5041	0.0671
$\frac{1}{2}wage^2$	-3.7813	4.9276	-2.2686	4.2613	-	-
$\frac{1}{2}pcap^2$	0.1667	0.1636	0.0954	0.1320	-	-
wage · pcap	0.7795	0.7067	0.4867	0.5791	-	-
cust · wage	0.1298	0.1787	-	-	-	-
cust · pcap	-0.0351	0.0479	-	-	-	-
PUC	-0.0855	0.0386	-0.0893	0.0384	-0.0870	0.0380
kwh	0.0476	0.0841	0.0476	0.0841	0.0301	0.0842
life	-0.6002	0.1168	-0.6099	0.1151	-0.6265	0.1139
lf	0.5047	0.2256	0.5789	0.2036	0.6016	0.2026
kmwire	0.3573	0.0856	0.3593	0.0860	0.3690	0.0864
s_ε^2	.0184		.0185		.0193	
R^2	.666		.664		.650	

ESTIMATED SCALE EFFECT



CUSTOMERS

Test of Homothetic Model vs Full Model, χ^2_2 under H_o: 0.54. Test of Loglinear Model vs Homothetic Model, χ^2_3 under H_o: 2.77. Order of differencing m=3. Kernel and spline estimates produced using `ksmooth` and `smooth.spline` functions in *S-Plus*

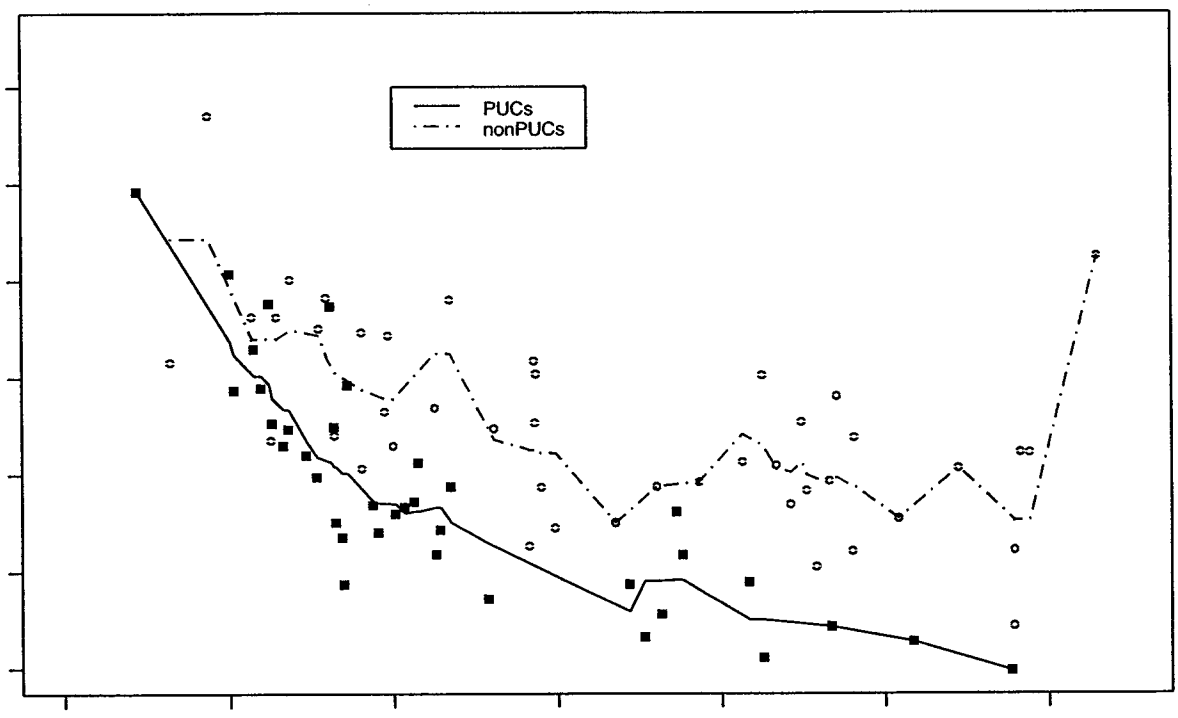
FIGURE 2.6.3: PARTIAL LINEAR MODEL — LOGLINEAR COST FUNCTION

PUC nonPUC ANALYSIS

SCALE ECONOMIES IN ELECTRICITY DISTRIBUTION

	Quadratic Model				Partial Linear Model [1]			
VARIABLE	PUC		nonPUC		PUC		nonPUC	
	Coef	SE	Coef	SE	Coef	SE	Coef	SE
cust	-0.9063	0.2201	-0.8102	0.229	-	-	-	-
cust^2	0.0852	0.0236	0.0803	0.0246	-	-	-	-
wage	0.6524	0.3074	1.3068	0.5946	0.6588	0.3249	1.2287	0.6854
pcap	0.4426	0.0842	0.5791	0.0947	0.4242	0.0872	0.6152	0.1041
kwh	0.1083	0.1055	-0.0584	0.1169	0.1077	0.1165	0.0600	0.1170
life	-0.5504	0.1227	-0.6423	0.1749	-0.4951	0.1308	-0.6602	0.1928
lf	1.8308	0.4895	0.4718	0.2326	1.9444	0.5626	0.3616	0.2377
kmwire	0.3575	0.1048	0.3760	0.1103	0.2975	0.1092	0.4254	0.1146
s_e^2	0.0133		0.0223		0.013		0.0216	

ESTIMATED SCALE EFFECT



[1] Order of differencing m=3. Test of equality of parametric effects using (2.6.3) which is χ^2_6 under H_0 yields a value of 9.73. Test of equality of residual variances using (2.2.15) which is N(0,1) under H_0 yields a value of -1.16. Test of parallel nonparametric effects, (see (2.6.8) and following paragraph) which is N(0,1) under H_0 yields a value of 1.49; (ssq_{pool} = .0193; ssq_{within} = .0176; $\hat{\pi}$ = .4707;).

FIGURE 2.6.4: MIXED ESTIMATION OF PUC/NONPUC EFFECTS SCALE ECONOMIES IN ELECTRICITY DISTRIBUITON		
VARIABLE	COEF	SE
wage	0.7443	0.3043
pcap	0.5099	0.0667
kwh	0.0851	0.0833
life	-0.5509	0.1063
lf	0.6198	0.2106
kmwire	0.3653	0.0782

Estimates of parametric effects are obtained separately for PUC and nonPUC sub-populations. Hence no PUC effect is estimated. The above estimates are obtained using (2.6.4) and (2.6.5).

2.7 Partial Parametric Model: CES Cost Function

A natural generalization of the partial linear model replaces the linear portion with a nonlinear parametric specification $y_i = f(x_i) + r(z_i, \beta) + \varepsilon_i$ where the regression function r is known and β is a p-dimensional vector. Suppose that the data (y_1, x_1, z_1) .... (y_n, x_n, z_n) have been reordered so that the x's are in increasing order. Let $x' = (x_1, ..., x_n)$, $y' = (y_1, ..., y_n)$ and $Z = (z_1', ..., z_n')'$.

Define $f(x)' = (f(x_1), ..., f(x_n))$ to be the column vector of nonparametric effects and $r(Z,\beta) = (r(z_1,\beta), ..., r(z_n,\beta))'$ to be the column vector of parametric effects. Let $\partial r(z,\beta)/\partial\beta$ be the $px1$ column vector of partial derivatives of r with respect to β for some vector z and $\partial r(Z,\beta)/\partial\beta$ the pxn matrix of partials of $r(z_1,\beta), ..., r(z_n,\beta)$ with respect to β. In matrix notation we may write the model as:

$$y = f(x) + r(Z,\beta) + \varepsilon \tag{2.7.1}$$

Applying the differencing matrix we have

$$Dy = Df(x) + Dr(Z,\beta) + D\varepsilon \tag{2.7.2}$$

Proposition 2.7.1: Suppose one uses optimal differencing weights. Let $\hat{\beta}_{diffnls}$ satisfy

$$\min_{\beta} \frac{1}{n}\left(Dy - Dr(Z,\beta)\right)'\left(Dy - Dr(Z,\beta)\right) \tag{2.7.3}$$

then

$$\hat{\beta}_{diffnls} \overset{A}{\sim} N\left(\beta, \left(1 + \frac{1}{2m}\right)\frac{\sigma_\varepsilon^2}{n}\Sigma^{-1}_{\frac{\partial r}{\partial\beta}|x}\right) \tag{2.7.4}$$

where

$$\Sigma_{\frac{\partial r}{\partial\beta}|x} = E\left(\frac{\partial r}{\partial\beta}\frac{\partial r}{\partial\beta'} \mid x\right) \tag{2.7.5}$$

Furthermore

$$s^2_{diffnls} = \frac{1}{n}\left(Dy - Dr(Z,\hat{\beta}_{nls})\right)'\left(Dy - Dr(Z,\hat{\beta}_{nls})\right) \overset{P}{\sim} \sigma^2_\varepsilon \tag{2.7.6}$$

and

$$\hat{\Sigma}_{\frac{\partial r}{\partial \beta}|x} = \frac{1}{n}\frac{\partial r(Z,\hat{\beta})}{\partial \beta}D'D\frac{\partial r(Z,\hat{\beta})}{\partial \beta'} \overset{P}{\sim} \Sigma_{\frac{\partial r}{\partial \beta}|x} \tag{2.7.7}$$ ■

As will be illustrated below, nonlinear least squares procedures (e.g., in *E-Views* or *TSP*) may be applied to equation (2.7.2) to obtain estimates of β. However, the covariance matrix produced by such programs needs to be multiplied by $1+\frac{1}{2m}$ as indicated by equation (2.7.4). (See also footnote to Figure 2.7.)

Example: Partial Parametric CES Cost Function

We continue with our example on electricity distribution costs. Consider a conventional CES cost function (see e.g., Varian (1992,p.56):

$$tc = \log(TC/CUST) = \beta_0 + \frac{1}{\rho}\log\left(\beta_1 WAGE^\rho + (1-\beta_1)PCAP^\rho\right) \tag{2.7.8}$$

where as before the number of customers served (*CUST*) measures the level of output. We are interested in assessing whether cost per customer is affected by the scale of operation, i.e., the number of customers. We therefore introduce a nonparametric scale effect (as well as several covariates):

$$\begin{aligned} tc &= f(cust) + \frac{1}{\rho}\log\left(\beta_1 WAGE^\rho + (1-\beta_1)PCAP^\rho\right) \\ &+ \beta_2 PUC + \beta_3 kwh + \beta_4 life + \beta_5 lf + \beta_6 kmwire + \varepsilon \end{aligned} \tag{2.7.9}$$

where lower case italicized variables correspond to transformed variables in log form (again, variable definitions are in Appendix C). First differencing and dividing by $\sqrt{2}$ so that the variance of the residual remains the same, yields:

$$\begin{aligned}\left[tc-tc_{-1}\right]/\sqrt{2} \cong \frac{1}{\rho}\left[\log\left(\beta_1 WAGE^{\rho}+(1-\beta_1)PCAP^{\rho}\right)-\log\left(\beta_1 WAGE^{\rho}_{-1}+(1-\beta_1)PCAP^{\rho}_{-1}\right)\right]/\sqrt{2} \\ +\beta_2\left[PUC-PUC_{-1}\right]/\sqrt{2}+\beta_3\left[kwh-kwh_{-1}\right]/\sqrt{2}+\beta_4\left[life-life_{-1}\right]/\sqrt{2} \\ +\beta_5\left[lf-lf_{-1}\right]/\sqrt{2}+\beta_6\left[kmwire-kmwire_{-1}\right]/\sqrt{2}+\left[\varepsilon-\varepsilon_{-1}\right]/\sqrt{2}\end{aligned} \tag{2.7.10}$$

Our parametric null consists of a quadratic specification for the scale effect. That is $f(cust) = \gamma_0 + \gamma_1 cust + \gamma_2 cust^2$ in equation (2.7.9). Results for this parametric specification and for equation (2.7.10) are presented in Figure 2.7. The model was estimated using nonlinear least squares in *EViews.* Applying the differencing specification test yields an asymptotically N(0,1) statistic of 1.48 suggesting that the quadratic null model may be adequate. The effects of covariates (PUC, *kwh, life, lf* and *kmwire*) remain fairly similar across the various parametric and semiparametric specifications contained in Figures 1.6, 2.6.2 and 2.7. Variants of the Leontief model may be implemented by imposing the restriction $\rho = 1$, which is not rejected in either the parametric or semiparametric specifications of Figure 2.7.

FIGURE 2.7: PARTIAL PARAMETRIC MODEL — CES COST FUNCTION
SCALE ECONOMIES IN ELECTRICITY DISTRIBUITON

VARIABLE	PARAMETRIC MODEL		PARTIAL PARAMETRIC MODEL[1]	
	Coef	SE	Coef	SE
cust	-0.7914	0.1802	-	-
$cust^2$	0.0398	0.0098	-	-
wage	0.6141	0.2113	0.7709	0.1777
pcap	1-.6141	-	1-.7709	-
ρ	0.3517	0.5614	0.7221	0.6199
PUC	-0.0856	0.0376	-0.0841	0.0430
kwh	0.0130	0.0866	-0.0095	0.0946
life	-0.5966	0.1215	-0.5219	0.1400
lf	0.5820	0.2141	0.4937	0.2357
kmwire	0.4004	0.0863	0.3669	0.0977
s_ε^2	.0213		.0188	
R^2	.618		.663	

[1] Order of differencing m=1. Model estimated using nonlinear least squares in *EViews*. Standard errors produced by *EViews* multiplied by $\sqrt{1.5}$ as per equation (2.7.4). Test of quadratic vs nonparametric specification of scale effect N(0,1): $V = (mn)^{1/2}\left(s_{res}^2 - s_{diff}^2\right) / s_{diff}^2 = 81^{1/2}(.0213 - .0188)/.0188 = 1.48$.

2.8 Instrumental Variable and Two Stage Least Squares Estimation in the Partial Linear Model

We return to the framework of Section 2.5 the partial linear model. Suppose one or more of the p parametric variables in Z are correlated with the residual and suppose W is an nxq matrix of observations on instruments for Z. We will assume that there are at least as many instruments as parametric variables, i.e., $q \geq p$ and that each instrument has a smooth regression function on x the nonparametric variable. Let $\widehat{DZ}$ be the predicted values of DZ:

$$\widehat{DZ} = DW\left((DW)'DW\right)^{-1}(DW)'DZ \tag{2.8.1}$$

Analogously to the conventional linear model, instrumental variable estimation may be motivated by multiplying equation (2.5.2) by $(\widehat{DZ})'$:

$$(\widehat{DZ})'Dy = (\widehat{DZ})'Df(x) + (\widehat{DZ})'DZ\beta + (\widehat{DZ})'D\varepsilon \tag{2.8.2}$$

Since differencing removes the nonparametric effect in large samples, this suggests the two stage least squares estimator $((\widehat{DZ})'DZ)^{-1}(\widehat{DZ})'Dy$. Define the conditional moment matrices: $\Sigma_{w|x} = E_x Var(w|x)$ and $\Sigma_{zw|x} = E_x Cov(z,w|x)$ where $Cov(z,w|x)$ is the pxq matrix of covariances between the z and w variables conditional on x.

Proposition 2.8.1: If one uses optimal differencing weights, then:

$$\hat{\beta}_{diff2sls} = \left[(\widehat{DZ})'DZ\right]^{-1}(\widehat{DZ})'Dy \overset{A}{\sim} N\left(\beta, \left(1+\frac{1}{2m}\right)\frac{\sigma_\varepsilon^2}{n}\left[\Sigma_{zw|x}\Sigma_{w|x}^{-1}\Sigma_{zw|x}'\right]^{-1}\right) \tag{2.8.3}$$

$$s^2_{diff2sls} = \frac{1}{n}\left(Dy - DZ\hat{\beta}_{diff2sls}\right)'\left(Dy - DZ\hat{\beta}_{diff2sls}\right) \overset{P}{\sim} \sigma_\varepsilon^2 \tag{2.8.4}$$

$$\hat{\Sigma}_{w|x} = \frac{1}{n}(DW)'DW \overset{P}{\sim} \Sigma_{w|x} \tag{2.8.5}$$

$$\hat{\Sigma}_{zw|x} = \frac{1}{n}(DZ)'DW \overset{P}{\sim} \Sigma_{zw|x} \qquad (2.8.6)\blacksquare$$

We can now produce a Hausman (1978) type test of endogeneity.

Proposition 2.8.2: let $\Sigma_{\hat{\beta}_{diff}}$ be the large sample covariance matrix of $\hat{\beta}_{diff}$ defined in equation (2.5.6) and $\Sigma_{\hat{\beta}_{diff2sls}}$ the corresponding covariance matrix for $\hat{\beta}_{diff2sls}$ in equation (2.8.3), then under the null hypothesis that z is uncorrelated with ε,

$$\left(\hat{\beta}_{diff} - \hat{\beta}_{diff2sls}\right)'\left[\Sigma_{\hat{\beta}_{diff2sls}} - \Sigma_{\hat{\beta}_{diff}}\right]^{-1}\left(\hat{\beta}_{diff} - \hat{\beta}_{diff2sls}\right) \overset{D}{\sim} \chi^2_p \qquad (2.8.7)$$

where p is the dimension of β. ■

The covariance matrices of each of the two estimators may be replaced by consistent estimates.

2.9 Increasing the Order of Differencing With Sample Size

With a scalar nonparametric variable, the order of differencing m may be increased with sample size so long as $m/n \to 0$. For example, the large sample behavior of the differencing estimators of the residual variance and the parameters of the partial linear model remain as stated above. [12] We consider the specification test of Section 2.3 in some detail in order to illustrate the impact of increasing the order of differencing on the power of the test. Suppose that $m = n^{1-\epsilon}$, $0<\epsilon<1$. If the null hypothesis is true and one uses optimal differencing weights then the standardized test statistic (2.3.2) continues to be N(0,1).

Consider a sequence of local alternatives to a linear null hypothesis of the form:

$$f_n(x) = \gamma_0 + \gamma_1 x + n^{-1/2+\zeta} g(x) \to \gamma_0 + \gamma_1 x \tag{2.9.1}$$

where $g(x)$ is a fixed function and $0<\zeta<1/2$. For ease of exposition, we assume that g is orthogonal to $\gamma_0 + \gamma_1 x$. In particular, $\int g(x)\,dP(x) = 0$ and $\int x g(x)\,dP(x) = 0$ so that $\int (\gamma_0 + \gamma_1 x) g(x)\,dP(x) = 0$.

Note that $s^2_{res} = \frac{1}{n}\sum \varepsilon_i^2 + n^{-1+2\zeta}\int g^2(x) + O_P(n)^{-1}$ and $s^2_{diff} = \frac{1}{n}\sum \varepsilon_i^2 + O_P(mn)^{-1/2}$. The latter may be obtained using Appendix Lemma A.3 and equation (2.2.10a). Then,

$$(mn)^{1/2}\left(s^2_{res} - s^2_{diff}\right) = (mn)^{1/2} n^{-1+2\zeta}\int g^2 + O_P(1) = n^{-1/2\epsilon+2\zeta}\int g^2 + O_P(1) \tag{2.9.2}$$

which diverges so long as $\epsilon < 4\zeta$. Thus, by allowing m to grow at a rate close to the sample size n, the test can detect alternative hypotheses that converge to the null at a rate close to $n^{-1/2}$.[13] On the other hand, if m remains fixed, then the right hand expression in (2.9.2) becomes $n^{-1/2+2\zeta}\int g^2 + O_P(1)$ which diverges if $1/4 < \zeta$. Substituting into (2.9.1) we see that under such circumstances, the test can only detect alternative hypotheses that converge to the null at a rate close to $n^{-1/4}$.

---

[12] To prove that asymptotic normality is preserved, one may use for example, McLeish ((1974).

[13] Li (1994) and Zheng (1996) achieve the same result using a conditional moment specification test.

2.10 Alternative Differencing Coefficients

Our focus has been on *optimal* differencing weights. For *general* differencing weights $d_0, d_1, \ldots, d_m$ satisfying constraints (2.1.1), define δ using (2.1.6). Then various results which we have provided may be generalized.

In particular, let s^2_{diff} be the differencing estimator in (2.2.4). Then (2.2.12) becomes (see Hall et al (1990)),

$$n^{1/2}\left(s^2_{diff} - \sigma^2_\varepsilon\right) \sim N\left(0,\ \eta_\varepsilon - \sigma^4_\varepsilon + 4\sigma^4_\varepsilon\delta\right) = N\left(0,\ \eta_\varepsilon - 3\sigma^4_\varepsilon + 2\sigma^4_\varepsilon(1+2\delta)\right) \qquad (2.10.1)$$

The differencing estimator for the partial linear model, section 2.5, now behaves as:

$$\hat{\beta}_{diff} = \left[(DZ)'DZ\right]^{-1}(DZ)'Dy \overset{A}{\sim} N\left(\beta,\ (1+2\delta)\frac{\sigma^2_\varepsilon}{n}\Sigma^{-1}_{z|x}\right) \qquad (2.10.2)$$

for the partial parametric model, section 2.7, we have:

$$\hat{\beta}_{diffnls} \overset{A}{\sim} N\left(\beta,\ (1+2\delta)\frac{\sigma^2_\varepsilon}{n}\Sigma^{-1}_{\frac{\partial r}{\partial\beta}|x}\right) \qquad (2.10.3)$$

and for the differencing instrumental variable estimator, section 2.8, we have:

$$\hat{\beta}_{diff2sls} = \left[(\widehat{DZ})'DZ\right]^{-1}(\widehat{DZ})'Dy \overset{A}{\sim} N\left(\beta,\ (1+2\delta)\frac{\sigma^2_\varepsilon}{n}\left[\Sigma_{zw|x}\Sigma^{-1}_{w|x}\Sigma'_{zw|x}\right]^{-1}\right) \qquad (2.10.4)$$

If one sets $\delta = 0$, then one obtains the large sample variance of the asymptotically efficient estimator. Thus, the relative efficiency of these estimators depends on the quantity $1+2\delta$. [14]

---

[14]The specification test in equation (2.3.2) becomes: $n^{1/2}\left(s^2_{res} - s^2_{diff}\right)/s^2_{diff} \overset{D}{\sim} N(0,\ 4\delta)$.

Hall, Kay and Titterington (1990, p. 515, Table 2) compare the relative efficiency of alternative differencing estimators of the residual variance.[15] They find that for small m, optimal weights perform substantially better than a 'spike' sequence where the differencing weight near the middle of the sequence is close to unity while others are equal and close to zero. For large m, both types of weights have similar properties. In particular, 'spike' weights are asymptotically efficient.

They also compare optimal weights to the usual weights used for numerical differentiation. (These are equivalent to m-th order divided differences for equally spaced data.) They find that these weights become progressively less efficient relative to optimal weights as m increases.

Seifert, Gasser and Wolf (1993) study the mean squared error of various differencing type estimators of the residual variance. They find that the bias resulting from the use of HKT optimal weights can be substantial in some cases, particularly if sample size is small and the signal to noise ratio is high. The mean squared error of differencing estimators of the partial linear model has apparently not been studied.

Since HKT differencing weights put maximum weight at the extreme of a sequence, one would expect that in some cases bias would be exacerbated. On the other hand, weights that are symmetric about a mid-point and decline as one moves away might have better bias properties. In particular, for even m we solve the optimization problem given by:

$$\min_{d_0,\ldots,d_m} \quad \delta = \sum_{k=1}^{m}\left(\sum_{j=0}^{m-k} d_j d_{j+k}\right)^2$$

$$s.t. \quad \sum_{j=0}^{m} d_j = 0 \qquad \sum_{j=0}^{m} d_j^2 = 1 \tag{2.10.5}$$

$$d_0 = d_m \quad d_1 = d_{m-1} \quad d_2 = d_{m-2} \;\ldots\ldots\; d_{m/2-1} = d_{m/2+1}$$

$$d_{m/2+1} \le d_{m/2+2} \;\ldots\; \le d_m$$

[15] If the residuals are normal, the variance of the differencing estimator in equation (2.10.1) becomes $2\sigma_\varepsilon^4(1+2\delta)$ and the variance of the efficient (e.g., parametric) estimator becomes $2\sigma_\varepsilon^4$ (see (2.2.11b)).

With m even, the number of weights is odd. Optimal values are presented below.

FIGURE 2.10.1: SYMMETRIC OPTIMAL DIFFERENCING WEIGHTS	
m	$(d_0, d_1, \ldots, d_m)$
2	(-0.4082, 0.8165, -0.4082)
4	(-0.1872,-0.2588, 0.8921,-0.2588,-0.1872)
6	(-0.1191,-0.1561,-0.1867,0.9237,-0.1867,-0.1561,-0.1191)
8	(-0.0868,-0.1091,-0.1292,-0.1454,0.9410,-0.1454,-0.1292,-0.1091,-0.0868)
10	(-0.0681,-0.0830,-0.0969,-0.1091,-0.1189,0.9519,-0.1189,-0.1091,-0.0969, -0.0830,-0.0681)

The optimization problems were solved using GAMS (see Brooke, A., D. Kendrick and A. Meeraus (1992)). Figure 2.10.2 compares the efficiency of optimal weights to symmetric optimal weights.

Figure 2.10.2
Relative Efficiency of Alternative Differencing Sequences
$(1+2\delta)$

m	Optimal	Symmetric Optimal
2	1.250	1.940
4	1.125	1.430
6	1.083	1.276
8	1.063	1.204
10	1.050	1.161
20	1.025	1.079
100	1.005	1.015
200	1.003	1.008
500	1.001	1.003

It is not surprising that symmetric optimal weights are substantially less efficient (since we are free to choose only about half as many coefficients).

3. Combining Differencing and Other Procedures

3.1 Modular Approach to Analysis of the Partial Linear Model

Our applications of the partial linear model $y = z\beta + f(x) + \varepsilon$ leave some untidy loose ends. Typically our analysis is divided into two components: first we obtain a differencing estimate of β and undertake inference procedures on β as if f were not present in the model. Then we analyze f by performing nonparametric estimation and inference on the newly constructed data $\left(y_i - z_i\hat{\beta}_{diff}, x_i\right)$ as if β were known. Is such a modular approach valid? Separate analysis of the *parametric* portion *is* justified by virtue of results like Proposition 2.5.1. However, a little more justification is necessary with respect to the appropriateness of our analysis of the *nonparametric* part.[16]

In the following sections we will provide justification for various modular procedures that we have already implemented -- whether they involve combining differencing procedures in sequence or combining differencing procedures with other nonparametric methods. We will also introduce some additional procedures such as tests of monotonicity and concavity.

3.2 Combining Differencing Procedures in Sequence

Recall the estimator of the residual variance in the partial linear model as defined in (2.5.4):

$$s^2_{diff} = \frac{1}{n}\left(Dy - DZ\hat{\beta}_{diff}\right)'\left(Dy - DZ\hat{\beta}_{diff}\right) = \frac{1}{n}\left(y - Z\hat{\beta}_{diff}\right)'D'D\left(y - Z\hat{\beta}_{diff}\right) \qquad (3.2.1)$$

It is easy to show that:

$$n^{1/2}\left(\frac{1}{n}\left(y - Z\hat{\beta}_{diff}\right)'D'D\left(y - Z\hat{\beta}_{diff}\right) - \frac{1}{n}\varepsilon'D'D\varepsilon\right) \xrightarrow{P} 0 \qquad (3.2.2)$$

This in turn implies that inference on the residual variance (Propositions 2.2.1, 2.2.2 and equation

---

[16] For example, we have applied tests of specification, tests of equality of nonparametric regression functions and conventional kernel and spline estimation procedures *after* removing estimated parametric effects. See e.g., Sections 2.6 and 2.7.

(2.2.15)), specification testing (Proposition 2.3.1) and tests of equality of regression functions (Propositions 2.4.1, 2.4.2) may be applied to the data with the estimated parametric effect removed. In each case, differencing is used first to estimate the parametric effect, then to perform a specific inference procedure.

3.3 Combining Differencing With Other Nonparametric Procedures

Suppose we perform a kernel regression of $y_i - z_i \hat{\beta}_{diff}$ on x_i. For simplicity assume the x's to be uniformly distributed on the unit interval and that the rectangular kernel is used. Define λ as the bandwidth and $N(x_o) = \{x_i \mid x_i \in x_o \pm \lambda/2\}$ to be the neighborhood of x_o over which smoothing is being performed. Using a Taylor approximation we have:

$$\begin{aligned}
\hat{f}(x_o) &\cong \frac{1}{\lambda n} \sum_{N(x_o)} y_i - z_i \hat{\beta}_{diff} \\
&= \frac{1}{\lambda n} \sum_{N(x_o)} f(x_i) + \frac{1}{\lambda n} \sum_{N(x_o)} \varepsilon_i + \left(\beta - \hat{\beta}_{diff}\right) \frac{1}{\lambda n} \sum_{N(x_o)} z_i \qquad (3.3.1) \\
&\cong f(x_o) + \tfrac{1}{2} f''(x_o) \frac{1}{\lambda n} \sum_{N(x_o)} (x_i - x_o)^2 + \frac{1}{\lambda n} \sum_{N(x_o)} \varepsilon_i + \left(\beta - \hat{\beta}_{diff}\right) \frac{1}{\lambda n} \sum_{N(x_o)} z_i
\end{aligned}$$

Each summation will have close to λn terms so that in each case we are calculating a simple average. Consider the term involving the second derivative which corresponds to the bias: $\sum_{N(x_o)} (x_i - x_o)^2 / \lambda n$ is like the variance of a uniform variable on an interval of width λ in which case it is $O(\lambda^2)$. The next corresponds to the variance term: it has mean 0 and variance $\sigma_\varepsilon^2 / \lambda n$ so that it is $O((\lambda n)^{1/2})$. The last term arises out of the removal of the estimated parametric effect. Thus,

$$\hat{f}(x_o) - f(x_o) = O(\lambda^2) + O_P((\lambda n)^{-1/2}) + O_p(n^{-1/2}) O_P(1) \qquad (3.3.2)$$

So long as $\lambda \to 0$ and $\lambda n \to \infty$, consistency of the kernel estimator is unaffected since all three terms converge to zero. Furthermore, $\lambda = O(n^{-1/5})$ still minimizes the rate at which the (sum of the) three terms converge to zero:

$$\hat{f}(x_o) - f(x_o) = O(n^{-2/5}) + O_P(n^{-2/5}) + O_P(n^{-1/2}) O_P(1) \qquad (3.3.2)$$

so that the optimal rate of convergence is unaffected. The order of each of the first two terms is $O_P\left(n^{-2/5}\right)$, while the third term converges to zero more quickly and independently of λ. Confidence intervals may also be constructed in the usual way since:

$$\begin{aligned}(\lambda n)^{1/2}\left(\hat{f}(x_o)-f(x_o)\right) &= O\left((\lambda n)^{1/2}\lambda^2\right) + O_P(1) + O_P\left(\lambda^{1/2}\right) \\ &= O(1) \quad + O_P(1) + O_P\left(n^{-1/10}\right) \quad \text{if } \lambda = O\left(n^{-1/5}\right)\end{aligned} \tag{3.3.3}$$

and the third term goes to zero, albeit slowly. If the optimal bandwidth $\lambda = O\left(T^{-1/5}\right)$ is selected, then confidence intervals must correct for a bias term.

Similar arguments apply to other nonparametric estimators. If one uses nonparametric least squares or spline estimators in a regression of $y_i - z_i\hat{\beta}_{diff}$ on x_i then the estimator $\hat{f}$ remains consistent and its rate of convergence is unchanged.

3.4 Monotonicity and Concavity Restrictions

Consider the pure nonparametric regression model $y = f(x) + \varepsilon$ and suppose one wants to test whether the function f is monotone increasing (or more precisely, nondecreasing). One may then use isotonic regression to obtain the restricted estimator of the residual variance:

$$s^2_{mon} = \min_{\hat{f}(x_1)\ldots\hat{f}(x_n)} \frac{1}{n}\sum\left(y_i - f(x_i)\right)^2 \quad s.t.\ \hat{f}(x_1) \le \ldots \le \hat{f}(x_n) \tag{3.4.1}$$

Then $n^{-1/2}\sum\left(f(x_i) - \hat{f}_{mon}(x_i)\right)^2$ and $n^{-1/2}\sum \varepsilon_i\left(f(x_i) - \hat{f}_{mon}(x_i)\right)$ converge to 0 sufficiently quickly, so that we may apply the specification test of Proposition 2.3.1.[17] In particular:

$$(m\,n)^{1/2}\frac{\left(s^2_{mon} - s^2_{diff}\right)}{s^2_{diff}} \overset{D}{\sim} N(0,1) \tag{3.4.2}$$

---

[17] Van de Geer (1990) demonstrates that if one imposes monotonicity only, as in (3.4.1), then $\int(\hat{f}-f)^2 = O_P\left(T^{-2/3}(\log T)^{2/3}\right)$.

Monotonic regression may be implemented using the function `monreg` in XploRe, (see Hardle, Klinke and Turlach (1995)) or using GAMS (see Brooke, A., D. Kendrick and A. Meeraus (1992)). Tests of concavity and convexity as well as of other restrictions may be implemented in a similar fashion. For example, convexity constraints may be imposed as follows:

$$s^2_{con} = \min_{\hat{f}(x_1)\ldots\hat{f}(x_n)} \frac{1}{n}\sum\left(y_i - \hat{f}(x_i)\right)^2$$

$$s.t.\quad \hat{f}(x_{i+1}) \le \frac{x_{i+2}-x_{i+1}}{x_{i+2}-x_i}\hat{f}(x_i) + \frac{x_{i+1}-x_i}{x_{i+2}-x_i}\hat{f}(x_{i+2}) \quad \forall i$$

In general, validity of the test procedure requires first demonstrating that the restricted estimator of the regression function converges sufficiently quickly. If the model is partially linear, then the estimated parametric effect may be first removed using the differencing estimator without altering the asymptotic validity of the specification test or the consistency of the monotone or concave/convex regression estimator.

3.5 Conclusion

The practical point of this chapter is that for the partial linear model $y = z\beta + f(x) + \varepsilon$, (or more generally the partial parametric model) we can separate the analysis of the parametric portion from the analysis of the nonparametric portion. Given a differencing estimate of β (or for that matter, any $n^{1/2}$-consistent estimate), we may construct the new dependent variable, $y_i^* = y_i - z_i\hat{\beta}_{diff}$, set aside the original y_i and analyze the data $\left(y_i^*, x_i\right)$ as if they came from the pure nonparametric model $y_i^* = f\left(x_i\right) + \varepsilon_i$. None of the large sample properties that we have discussed will be affected. This holds true regardless of the dimension of the parametric variable z .

4. Bootstrapping Differencing Procedures

4.1 Introduction

In the following sections we will be proposing bootstrap methods for certain differencing procedures. We will rely on 'residual' based resampling rather than 'joint' resampling, principally because the former appears to perform better in simulations. Validity of the proposed bootstrap methods is ensured by extant results in the literature, to which we will refer as necessary. As a matter of convention, the superscript 'B' will denote a bootstrap sample, estimator, test statistic....

The models in which we are interested are nested within the partial linear model $y = z\beta + f(x) + \varepsilon$. If the main objective is to perform inference on β or on properties of the residual ε, then differencing may be used to first remove the nonparametric effect f. (Indeed, it does not matter if the data arrive in differenced form $Dy = Df(x) + Dz\beta + D\varepsilon$.) But in this case, the residuals follow a moving average process. We will use the moving block bootstrap to accommodate the time series structure that has been introduced into the residuals. Alternative procedures, such as the stationary bootstrap, may also be used.

On the other hand, one may be interested in testing hypotheses about the nonparametric effect f. If under the null, f (and β) is being estimated,[18] then it becomes convenient to use independent sampling from the estimated residuals $\hat{\varepsilon} = y - z\hat{\beta}_{diff} - \hat{f}(x)$, after they have been recentered. Such specification test procedures are relatively simpler so we will begin with them.

Simulations on the various procedures below have been performed, the results of which are available from the author.

[18] e.g., if under the null, f is a pure parametric model or if f is monotone or concave and isotonic regression is being applied;

4.2 Specification Testing

Consider again the specification test where under the null hypothesis, the regression function lies in a parametric family $h(x,\gamma)$ and h is known. If the null is true then:

$$V = (mn)^{1/2} \frac{\left(s^2_{res} - s^2_{diff}\right)}{s^2_{diff}} \overset{D}{\sim} N(0,\ 1) \tag{4.1.1}$$

(See sections 1.3 and 2.3 above.) Residual resampling to obtain bootstrap critical values may be conducted as follows:

FIGURE 4.2: SPECIFICATION TEST: RESIDUAL RESAMPLING BOOTSTRAP

1. Suppose the data (y_1,x_1) .... (y_n,x_n) have been reordered so that the x's are in increasing order. Estimate the model imposing the null hypothesis, e.g., $y_i = \hat{\gamma}_1 + \hat{\gamma}_2 x_i + \hat{\varepsilon}_i$. Retrieve and recenter the estimated residuals and let $\hat{F}_\varepsilon$ be their empirical distribution function. [19]

2.a) Obtain n independent observations $\hat{\varepsilon}^B_1, \ldots, \hat{\varepsilon}^B_n$ from $\hat{F}_\varepsilon$ and construct the bootstrap data set $\left(y^B_1, x_1\right)$ .... $\left(y^B_n, x_n\right)$ where $y^B_i = \hat{\gamma}_1 + \hat{\gamma}_2 x_i + \hat{\varepsilon}^B_i$.

b) Using the bootstrap data set estimate the restricted model and calculate the usual parametric estimator of the residual variance $s^{2^B}_{res}$. Calculate $s^{2^B}_{diff}$ and $V^B = (mn)^{1/2}\left(s^{2^B}_{res} - s^{2^B}_{diff}\right) / s^{2^B}_{diff}$.

c) Repeat 2a and 2b to achieve the desired degree of accuracy and calculate the empirical distribution of V^B. Calculate the bootstrap critical value V^B_{crit} from the right hand tail of this distribution using the desired significance level α.

3. Calculate V from the original data and compare to the bootstrap critical value.

[19] In the linear model with a constant, the estimated residuals already have mean zero. In the absence of a constant term or if the parametric null hypothesis is nonlinear, the estimated residuals may not have a mean of zero and the residuals need to be recentered in order to ensure validity of the bootstrap. For further details, see Freedman (1981).

Independent sampling from $\hat{F}_{\varepsilon}$ is of course equivalent to sampling with replacement from the recentered estimated residuals with equal probability assigned to each residual. The validity of the bootstrap test procedure may be established under fairly weak conditions.[20] In particular, for significance level α, $\lim Prob\left[V \geq V^{B}_{crit}\right] = \alpha$ if the null hypothesis is true, and $\lim Prob\left[V \geq V^{B}_{crit}\right] = 1$ if the null hypothesis is false. If the residuals are possibly heteroskedastic, then the wild bootstrap Wu (1986), may be used for resampling.

The above procedure may be applied to tests of monotonicity and concavity where such constraints are imposed on the estimator under the null hypothesis (see Section 3.4 above). It may also be used for testing parametric specifications for f within the partial linear model $y = z\beta + f(x) + \varepsilon$.

4.3 Bootstrap Inference on the Residual Variance

For purposes of inference on the residual variance in the pure nonparametric regression model we have been using:

$$\frac{n^{1/2}\left(s^{2}_{diff} - \sigma^{2}_{\varepsilon}\right)}{\left(\hat{\eta} + \left(1/m - 1\right)s^{4}_{diff}\right)^{1/2}} \overset{D}{\sim} N(0,1) \tag{4.3.1}$$

(See Section 2.2, Propositions 2.2.1 and 2.2.2.) We will use the moving block bootstrap to simulate the distribution of this statistic. [21] The basic idea is as follows. Suppose one has a vector of length n with a stationary dependent structure. Rather than randomly sampling individual elements, suppose one randomly samples sub-sequences or blocks, then strings them together. This bootstrapped vector thus retains some of the original dependence structure. If n is large and the dependence dies down fairly quickly, then a judiciously chosen block-length L will cause the

[20] Beran and Ducharme (1991, Proposition 4.3) lists a simple set of sufficient conditions for consistency of a bootstrap test. In our case, bounded support for x, a bounded first derivative on f and bounded fourth order moments on ε are sufficient.

[21] For details of the moving block bootstrap, see Kunsch (1989) and Liu and Singh (1992). The latter paper specifically deals with the case of m-dependent data which is precisely the dependence structure of our differenced residuals.

resampled vectors to closely mimic the dependence structure of the original data. (The usual condition is that $L \to \infty$ and $L/n \to 0$.)

To obtain the differencing estimator of the residual variance we have differenced the data to obtain $Dy = Df(x) + D\varepsilon$ whence $D\varepsilon$ is m-dependent (m is as usual the order of differencing). To obtain a bootstrap confidence interval for σ_ε^2 using the moving block bootstrap, select the block-length L. We will be sampling sufficient blocks to fill a vector whose length is the same as $D\varepsilon$. (The last sampled block may not be used in its entirety since $dim(D\varepsilon)/L$ may not be an integer.) Then proceed as in Figure 4.3.1. Note that the bootstrap statistic is centered around $E\left(s_{diff}^{2^B}\right)$ rather than around the initial estimate s_{diff}^2. Lahiri (1992) shows that this modification ensures that the bootstrap approximation is second order correct.

Next, suppose we have data on two nonparametric regression models $y_A = f_A(x_A) + \varepsilon_A$ and $y_B = f_B(x_B) + \varepsilon_B$. We want to test equality of the residual variances while allowing possibly different regression functions. Figure 4.3.2 provides a bootstrap procedure for testing this property using the moving block bootstrap. Both of the above bootstrap procedures may be applied to the partial linear model.

FIGURE 4.3.1: CONFIDENCE INTERVAL FOR RESIDUAL VARIANCE: MOVING BLOCK BOOTSTRAP

1. Suppose the data (y_1, x_1) .... (y_n, x_n) have been reordered so that the x's are in increasing order. Let $y' = (y_1, \ldots, y_n)$ and difference to obtain Dy.

2.a) From the vector Dy, randomly sample sufficient blocks to fill a vector of length n. Concatenate them to obtain Dy^B.

b) Calculate and store the following quantities

$$s_{diff}^{2^B} = \frac{1}{N}\left(Dy^B\right)' Dy^B$$

and

$$\hat{\eta}^B = \frac{\frac{1}{N}\sum_i^N \left[Dy^B\right]_i^4 - 6\left(s_{diff}^{2^B}\right)^2 \left(\sum_{i=0}^{m-1} d_i^2 \sum_{j=i+1}^{m} d_j^2\right)}{\sum_{i=0}^{m} d_i^4}$$

c) Repeat 2a and 2b to achieve the desired degree of accuracy and calculate $E\left(s_{diff}^{2^B}\right)$ as the average of the $s_{diff}^{2^B}$. Using the stored values construct the empirical distribution of the statistic:

$$\frac{n^{1/2}\left(s_{diff}^{2^B} - E\left(s_{diff}^{2^B}\right)\right)}{\left(\hat{\eta}^B + \left(1/m - 1\right)\left(s_{diff}^{2^B}\right)^2\right)^{1/2}}$$

3. To obtain a two-sided α-confidence interval for σ_ε^2, let L and U be the lower and upper percentiles of the above empirical distribution which leave $(1-\alpha)/2$ probability in each tail. Then isolate σ_ε^2 in the middle of the probability statement:

$$Prob\left[L \le \frac{n^{1/2}\left(s_{diff}^2 - \sigma_\varepsilon^2\right)}{\left(\hat{\eta} + \left(1/m - 1\right)s_{diff}^4\right)^{1/2}} \le U\right] \cong \alpha$$

Reference: Propositions 2.2.1 and 2.2.2.

FIGURE 4.3.2: TESTING EQUALITY OF VARIANCES: MOVING BLOCK BOOTSTRAP

1. Suppose that each data set (y_{A1}, x_{A1}) .... (y_{An}, x_{An}) and (y_{B1}, x_{B1}) .... (y_{Bn}, x_{Bn}) has been reordered so that the x's are in increasing order. Let $y_A' = (y_{A1}, \ldots, y_{An})$, $y_B' = (y_{B1}, \ldots, y_{Bn})$ and difference to obtain Dy_A and Dy_B. Concatenate these two vectors to obtain Dy.

2.a) From the vector Dy, randomly sample sufficient blocks to fill a vector of length n. Concatenate these to obtain Dy_A^B. Calculate and store the quantities $s_A^{2^B}$, $\hat{\eta}_A^B$.

b) From the vector Dy, randomly sample sufficient blocks to fill a vector of length n. Concatenate these to obtain Dy_B^B. Calculate and store the quantities $s_B^{2^B}$, $\hat{\eta}_B^B$.

c) Repeat 2a and 2b to achieve the desired degree of accuracy and calculate $E\left(s_A^{2^B} - s_B^{2^B}\right)$ as the average of the $s_A^{2^B} - s_B^{2^B}$. Using the stored values construct the empirical distribution of the statistic:

$$\frac{s_A^{2^B} - s_B^{2^B} - E\left(s_A^{2^B} - s_B^{2^B}\right)}{\left(\dfrac{\hat{\eta}_A^B + (1/m-1)\left(s_A^{2^B}\right)^2}{n_A} + \dfrac{\hat{\eta}_B + (1/m-1)\left(s_B^{2^B}\right)^2}{n_B}\right)^{1/2}}$$

and obtain critical values.

3. Compute the value of the test statistic:

$$\frac{s_A^2 - s_B^2}{\left(\dfrac{\hat{\eta}_A + (1/m-1)s_A^4}{n_A} + \dfrac{\hat{\eta}_B + (1/m-1)s_B^4}{n_B}\right)^{1/2}}$$

and compare to the bootstrap critical values.

Reference: equation (2.2.15).

4.4 Bootstrap Inference on the Partial Linear Model

We return to the differencing estimator of the partial linear model, (Proposition 2.5.1). Using Lahiri (1996, Th.2.2) it can be shown that if one applies the moving block bootstrap, then

$$n^{1/2}\frac{\hat{\Sigma}_{z|x}^{1/2}\left(\hat{\beta}_{diff}-\beta\right)}{\left(1+\frac{1}{2m}\right)^{1/2}s_{diff}} \tag{4.4.1a}$$

is well approximated by,

$$n^{1/2}\frac{\hat{\Sigma}_{z|x}^{1/2}\left(\hat{\beta}_{diff}^{B}-E\hat{\beta}_{diff}^{B}\right)}{\left(1+\frac{1}{2m}\right)^{1/2}s_{diff}^{B}} \tag{4.4.1b}$$

This in turn may be used to construct confidence intervals for components of β, details of which are contained in Figure 4.4.1.

Suppose we want to test the general linear restriction $R\beta = r$ (see section 2.5 above). So long as we impose the restrictions on the model and use the resulting residuals for bootstrap resampling then the sampling distribution of

$$\frac{n\left(R\hat{\beta}_{diff}-r\right)'\left(R\hat{\Sigma}_{z|x}^{-1}R'\right)^{-1}\left(R\hat{\beta}_{diff}-r\right)}{\left(1+\frac{1}{2m}\right)s_{diff}^{2}} \tag{4.4.2a}$$

which under the null hypothesis is $\chi^2_{rank(R)}$, is well approximated by:

$$\frac{n\left(R\left(\hat{\beta}^B_{diff} - E\hat{\beta}^B_{diff}\right)\right)'\left(R\hat{\Sigma}^{-1}_{z|x}R'\right)^{-1}\left(R\left(\hat{\beta}^B_{diff} - E\hat{\beta}^B_{diff}\right)\right)}{\left(1+\frac{1}{2m}\right)s^{2^B}_{diff}} \tag{4.4.2b}$$

A bootstrap procedure for testing linear restrictions is provided in Figure 4.4.2.

FIGURE 4.4.1: CONFIDENCE INTERVALS FOR β IN THE PARTIAL LINEAR MODEL: MOVING BLOCK BOOTSTRAP

1. Suppose the data $(y_1, x_1, z_1) \ldots (y_n, x_n, z_n)$ have been reordered so that the x's are in increasing order. Let $y' = (y_1, \ldots, y_n)$ and $Z = \begin{pmatrix} z_1', \ldots, z_n' \end{pmatrix}$. Difference to obtain Dy and DZ. Regress Dy on DZ to obtain $\hat{\beta}_{diff}$ and the differencing estimate of the residual variance s^2_{diff}. Calculate $\hat{\Sigma}_{z|x} = \frac{1}{n}(DZ)'DZ$. Calculate $Dy - DZ\hat{\beta}_{diff}$ and re-center this vector to obtain $\widehat{D\varepsilon}$.

2.a) From the vector $\widehat{D\varepsilon}$, randomly sample sufficient blocks to fill a vector of length n. Concatenate them to obtain $D\varepsilon^B$. Define $Dy^B = DZ\hat{\beta}_{diff} + D\varepsilon^B$.

b) Regress Dy^B on DZ to obtain $\hat{\beta}^B_{diff}$ and the differencing estimate of the residual variance $s^{2^B}_{diff}$. Store these values.

c) Repeat 2a and 2b to achieve the desired degree of accuracy and calculate $E\hat{\beta}^B_{diff}$.

3. To obtain a confidence interval for the i-th component of β use the stored values to construct the empirical distribution of:

$$\frac{n^{1/2}\left(\hat{\beta}^B_{diff_i} - E\hat{\beta}^B_{diff_i}\right)}{\left(1+\frac{1}{2m}\right)^{1/2} s^B_{diff}\left(\left[\hat{\Sigma}^{-1}_{z|x}\right]_{ii}\right)^{1/2}}$$

For a two-sided α-confidence interval, let L and U be the lower and upper percentiles of the above empirical distribution which leave $(1-\alpha)/2$ probability in each tail. Then isolate σ^2_ε in the middle of the probability statement:

$$Prob\left[L \le \frac{n^{1/2}\left(\hat{\beta}_{diff_i} - \beta_i\right)}{\left(1+\frac{1}{2m}\right)^{1/2} s_{diff}\left(\left[\hat{\Sigma}^{-1}_{z|x}\right]_{ii}\right)^{1/2}} \le U\right] \cong \alpha$$

FIGURE 4.4.2: TESTING LINEAR RESTRICTIONS $R\beta = r$ IN THE PARTIAL LINEAR MODEL: MOVING BLOCK BOOTSTRAP

1. Suppose the data $(y_1, x_1, z_1) \ldots (y_n, x_n, z_n)$ have been reordered so that the x's are in increasing order. Let $y' = (y_1, \ldots, y_n)$ and $Z = (z_1, \ldots, z_n)$ difference to obtain Dy and DZ. Regress Dy on DZ **subject to** $R\beta = r$ to obtain $\hat{\beta}_{res}$ and the differencing estimate of the residual variance $s^2_{res} = \frac{1}{n}\left(Dy - DZ\hat{\beta}_{res}\right)'\left(Dy - DZ\hat{\beta}_{res}\right)$. Calculate $\hat{\Sigma}_{z|x} = \frac{1}{n}(DZ)'DZ$. Calculate $Dy - DZ\hat{\beta}_{res}$ and re-center this vector to obtain $\widehat{D\varepsilon}$.

2.a) From the vector $\widehat{D\varepsilon}$, randomly sample sufficient blocks to fill a vector of length n. Concatenate them to obtain $D\varepsilon^B$. Define $Dy^B = DZ\hat{\beta}_{res} + D\varepsilon^B$.

b) Perform the **unrestricted** regression of Dy^B on DZ to obtain $\hat{\beta}^B$ and the differencing estimate of the residual variance $s^{2^B}_{diff} = \frac{1}{n}\left(Dy^B - DZ\hat{\beta}^B_{diff}\right)'\left(Dy^B - DZ\hat{\beta}^B_{diff}\right)$. Store these values.

c) Repeat 2a and 2b to achieve the desired degree of accuracy and calculate $E\hat{\beta}^B_{diff}$. Use the stored values to construct the empirical distribution of the statistic in (4.4.2b):

$$\frac{n\left(R\left(\hat{\beta}^B_{diff} - E\hat{\beta}^B_{diff}\right)\right)'\left(R\hat{\Sigma}^{-1}_{z|x}R'\right)^{-1}\left(R\left(\hat{\beta}^B_{diff} - E\hat{\beta}^B_{diff}\right)\right)}{\left(1 + \frac{1}{2m}\right)s^{2^B}_{diff}}$$

3.Calculate the bootstrap critical value from the right hand tail of this distribution using the desired significance level α. Calculate (4.4.2a) from the original data using the **unrestricted** differencing estimator of β and compare to the bootstrap critical value.

We illustrate two of the above bootstrap procedures by applying them to our data on electricity distribution costs (see Sections 1.6 and 2.6 above). With wages and capital prices entering in a Cobb-Douglas format, the specification is given by:

$$tc = f(cust) + \beta_1 wage + \beta_2 pcap + \beta_3 PUC + \beta_4 kwh + \beta_5 life + \beta_6 lf + \beta_7 kmwire + \varepsilon \qquad (4.5.1)$$

We re-estimate this model using differencing and apply the moving block bootstrap as described in Figure 4.4.1 to obtain confidence intervals for the components of β. (We use 10000 bootstrap replications.) These are compared to the asymptotic confidence intervals obtained by applying Proposition 2.5.1. (See also Figure 2.6.2.) The results are summarized in Figure 4.5. The bootstrap confidence intervals are about 15% narrower than the asymptotic intervals.

Next we test a quadratic specification for f. Since this is a partial linear model, step 2b of Figure 4.2 is modified. In particular, s^2_{diff} is obtained using equation 2.5.4. Again using 10000 bootstrap replications, the bootstrap critical value at a 5% significance level is 1.643, very close to the normal value, (keeping in mind that this is a one-sided test). The value of the specification test statistic is 1.681 suggesting that there is some evidence against the quadratic model.

FIGURE 4.5: ASYMPTOTIC VS BOOTSTRAP CONFIDENCE INTERVALS
SCALE ECONOMIES IN ELECTRICITY DISTRIBUTION

VARIABLE	PARTIAL LINEAR MODEL [1]			
	COEF	SE	95% CONFIDENCE INTERVALS Asymptotic *Bootstrap*	
wage	0.3543	0.3170	-0.2670 *-0.1734*	0.9757 *0.8801*
pcap	0.5041	0.0671	0.3725 *0.3911*	0.6357 *0.6152*
PUC	-0.0870	0.0380	-0.1615 *-0.1503*	-0.0125 *-0.0251*
kwh	0.0301	0.0842	-0.1350 *-0.1082*	0.1952 *0.1672*
life	-0.6265	0.1139	-0.8498 *-0.8170*	-0.4032 *-0.4371*
lf	0.6016	0.2026	0.2044 *0.2589*	0.9987 *0.9487*
kmwire	0.3690	0.0864	0.1996 *0.2280*	0.5383 *0.5116*
s^2 R^2	.0193 .650			

[1] Order of differencing m=3. Number of observations n=*81*. Resampling using moving block bootstrap as in Figure 4.4.1. Block length L = 3. Number of bootstrap replications = 10000.

5. EXTENSIONS

5.1 Heteroskedasticity Consistent Covariance Matrix Estimation

Consider the basic linear regression model expressed in matrix notation $y = Z\beta + \varepsilon$ where coefficients are estimated using ordinary least squares $\hat{\beta}_{ols} = \left(Z'Z\right)^{-1} Z'y$. If the residuals are independent but heteroskedastic with unknown covariance matrix Ω then:

$$Var\left(\hat{\beta}_{ols}\right) = \frac{1}{n}\left(\frac{Z'Z}{n}\right)^{-1} \frac{Z'\Omega Z}{n} \left(\frac{Z'Z}{n}\right)^{-1} \tag{5.1.1}$$

White (1980) demonstrates that in order to estimate this covariance matrix, one need only obtain a consistent estimator of $plim\left(Z'\Omega Z/n\right)$ and not Ω itself. He proposes $Z'\hat{\Omega}Z/n$ where the diagonal elements of $\hat{\Omega}$ are the squares of the estimated OLS residuals $\hat{\varepsilon}_i^2$, (off-diagonal elements are zero). He then shows that $Z'\Omega Z/n - Z'\hat{\Omega}Z/n \overset{P}{\to} 0$. Substitution into (5.1.1) yields a heteroskedasticity consistent covariance matrix estimator for $\hat{\beta}_{ols}$. As will be evident shortly, it is convenient to think of $\hat{\Omega}$ as the matrix $\hat{\varepsilon}\hat{\varepsilon}'$ with all terms whose expectation is known to be zero, set to zero (which in this case means that all off-diagonal terms are set to zero).

Let us now return to our differencing estimator $\hat{\beta}_{diff} = \left[(DZ)'DZ\right]^{-1}(DZ)'Dy$ of the partial linear model $y = Z\beta + f(x) + \varepsilon$. If $Var\left(\varepsilon\varepsilon'\right) = \Omega$ then:

$$Var\left(\hat{\beta}_{diff}\right) = \frac{1}{n}\left(\frac{Z'D'DZ}{n}\right)^{-1} \frac{Z'D'D\,\Omega\,D'DZ}{n} \left(\frac{Z'D'DZ}{n}\right)^{-1} \tag{5.1.2}$$

Consider the structure of $D\Omega D'$. Since Ω is diagonal, the nonzero elements of $D\Omega D'$ consist of the main diagonal and the m adjacent diagonals (m is the order of differencing). This is because differencing introduces a moving average process of order m into the residuals.[22] White's (1985) generalizations may then be applied to our differenced model $Dy \cong DZ\beta + D\varepsilon$ where as usual, differencing has (approximately) removed the nonparametric effect. Define $\widehat{D\varepsilon} = Dy - DZ\hat{\beta}_{diff}$.

---

[22] Alternatively, note that $D\Omega D' = (d_0 L_0 + \ldots + d_m L_m)\,\Omega\,(d_0 L_0' + \ldots + d_m L_m')$. The lag matrices L_i, L_i' shift the main diagonal of Ω to the i-th off-diagonals.

In order to mimic the structure of $D\Omega D'$ we define $D\widehat{\Omega}D'$ to be the matrix $\widehat{D\varepsilon}\,\widehat{D\varepsilon}'$ with all terms which are more than m diagonals away from the main diagonal set to zero. Heteroskedasticity consistent standard errors for $\hat{\beta}_{diff}$ may be obtained by using:

$$Var\left(\hat{\beta}_{diff}\right) = \frac{1}{n}\left(\frac{Z'D'DZ}{n}\right)^{-1}\frac{Z'D'D\widehat{\Omega}D'DZ}{n}\left(\frac{Z'D'DZ}{n}\right)^{-1} \tag{5.1.3}$$

Figure 5.1 contains White standard errors generated by equation (5.1.3) for our data on electricity distribution (Sections 1.6 and 2.6 above).

FIGURE 5.1: WHITE HETEROSKEDASTICITY CONSISTENT STANDARD ERRORS SCALE ECONOMIES IN ELECTRICITY DISTRIBUTION

VARIABLE	LOGLINEAR MODEL		
	Coef	SE	White SE's
wage	0.3543	0.3170	0.3480
pcap	0.5041	0.0671	0.1059
PUC	-0.0870	0.0380	0.0298
kwh	0.0301	0.0842	0.0879
life	-0.6265	0.1139	0.1066
lf	0.6016	0.2026	0.1992
kmwire	0.3690	0.0864	0.1039

Order of differencing m=3.

5.2 Specification Test in the Presence of Heteroskedasticity

The specification test we have proposed can also be readily modified to accommodate heteroskedasticity. Suppose that under the alternative we have a pure nonparametric model $y = f(x) + \varepsilon$. We assume the residuals are independent but heteroskedastic with unknown covariance matrix Ω. Let $h(x, \gamma)$ be a known function of x and an unknown parameter γ. We wish to test the null hypothesis that the regression function has the parametric form $h(x, \gamma)$. Let $\hat{\gamma}_{LS}$ be obtained using for example parametric nonlinear least squares and define:

$$s_{res}^2 = \frac{1}{n}\sum \hat{\varepsilon}_i^2 = \frac{1}{n}\sum \left(y_i - h\left(x_i, \hat{\gamma}_{LS}\right)\right)^2 \tag{5.2.1}$$

As usual let:

$$s_{diff}^2 = \frac{1}{n} y' D' D y \tag{5.2.2}$$

where D is an optimal differencing matrix of order m. The specification test in Proposition 2.3.1 may be extended as follows. If the null hypothesis is true then:

$$(m\,n)^{1/2} \frac{\left(s_{res}^2 - s_{diff}^2\right)}{\hat{\xi}^{1/2}} \overset{D}{\sim} N(0, 1) \tag{5.2.3a}$$

where

$$\hat{\xi} = \frac{1}{m}\left(\frac{1}{n}\sum \hat{\varepsilon}_i^2 \hat{\varepsilon}_{i-1}^2 + \ldots + \frac{1}{n}\sum \hat{\varepsilon}_i^2 \hat{\varepsilon}_{i-m}^2\right) \tag{5.2.3b}$$

Thus far in this essay we have assumed that the original observations are independent of each other. As a result, we have been able to presume that the data have arrived reordered so that the nonparametric variable is in increasing order. Suppose now we have time series data with correlation in the original residuals. Reordering has implications for the structure of the residual covariance matrix. To keep track of this we will use permutation matrices. Suppose then we have data (y_1, x_1, z_1) .... (y_T, x_T, z_T) on the model $y = z\beta + f(x) + \varepsilon$. (To reflect the time series nature of the data, observations are subscripted by '*t*' rather than '*i*'.) Let P be the permutation matrix which reorders the data so that x is in increasing order. Our reordered and differenced model may now be written in matrix notation as:

$$DPy = DPf(x) + DPZ\beta + DP\varepsilon \tag{5.3.1}$$

where $Var(\varepsilon\varepsilon') = \Omega$ is not necessarily diagonal, and the differencing estimator is given by:

$$\hat{\beta}_{diff} = \left[(DPZ)'DPZ\right]^{-1}(DPZ)'DPy \tag{5.3.2}$$

We have then:

$$Var\left(\hat{\beta}_{diff}\right) = \frac{1}{n}\left(\frac{Z'P'D'DPZ}{n}\right)^{-1}\frac{Z'P'D'DP\Omega P'D'DPZ}{n}\left(\frac{Z'P'D'DPZ}{n}\right)^{-1} \tag{5.3.3}$$

Under general conditions, $Z'P'D'DPZ/n \overset{P}{\sim} \Sigma_{z|x}$. We need a consistent estimate of the interior matrix in (5.3.3). For simplicity, suppose that the correlation in the residuals declines sufficiently quickly so that we may assume that any element of Ω which is more than $\mathcal{L}$ diagonals away from the main diagonal is zero. Consider the matrix $DP\Omega P'D'$ and define the indicator matrix H:

$$\begin{aligned} H_{ij} &= 1 \quad \text{if } \left[DP\Omega P'D'\right]_{ij} \neq 0 \\ &= 0 \quad \text{otherwise} \end{aligned} \tag{5.3.4}$$

We will again use the estimated residuals from the differenced regression (5.3.1). In particular, define $\widehat{DP\varepsilon} = DPy - DPZ\hat{\beta}_{diff}$. In order to mimic the structure of $DP\Omega P'D'$ let:[23]

$$DP\widehat{\Omega P}'D' = \left(\widehat{DP\varepsilon}\,\widehat{DP\varepsilon}'\right) \odot H \tag{5.3.5}$$

and substitute (5.3.5) into (5.3.3) to obtain a heteroskedasticy and autocorrelation consistent covariance matrix estimator.

With modest additional effort we may construct Newey-West (1987) standard errors. Once again let $\mathcal{L}$ be the maximum lag which exhibits non-zero autocorrelation in the matrix Ω. Define matrices H^{ℓ}, $\ell = 0,\ldots,\mathcal{L}$ as follows. Let H^0 be the identity matrix. For $\ell = 1,\ldots,\mathcal{L}$ let

$$\begin{aligned} H^{\ell}_{ij} &= 1 \quad \textit{if} \left[DP\left(L_{\ell} + L_{\ell}'\right)P'D'\right]_{ij} \neq 0 \\ &= 0 \quad \textit{otherwise} \end{aligned} \tag{5.3.6}$$

Define

$$DP\widehat{\Omega P}'D' = \left(\widehat{DP\varepsilon}\,\widehat{DP\varepsilon}'\right) \odot \left(\sum_{\ell=0}^{\mathcal{L}} \frac{\ell}{\mathcal{L}+1} H^{\ell}\right) \tag{5.3.7}$$

and substitute (5.3.7) into (5.3.3) to obtain Newey-West standard errors in this setting. The order of dependence $\mathcal{L}$ may be permitted to increase with sample size.

---

[23] For any two matrices A, B of identical dimension define $[A \odot B]_{ij} = A_{ij}B_{ij}$.

5.4 Empirical Application: Weather and Electricity Demand

In a classic paper Engle, Granger, Rice and Weiss (1986), used the partial linear model to study the impact of weather and other variables on electricity demand. We estimate a similar model where weather enters nonparametrically and other variables enter parametrically. Our data consist of 288 quarterly observations in Ontario for the period 1971 to 1994. Our specification is:

$$elec_t = f\left(TEMP_t\right) + \beta_1\, relprice_t + \beta_2\, gdp_t + \varepsilon \tag{5.4.1}$$

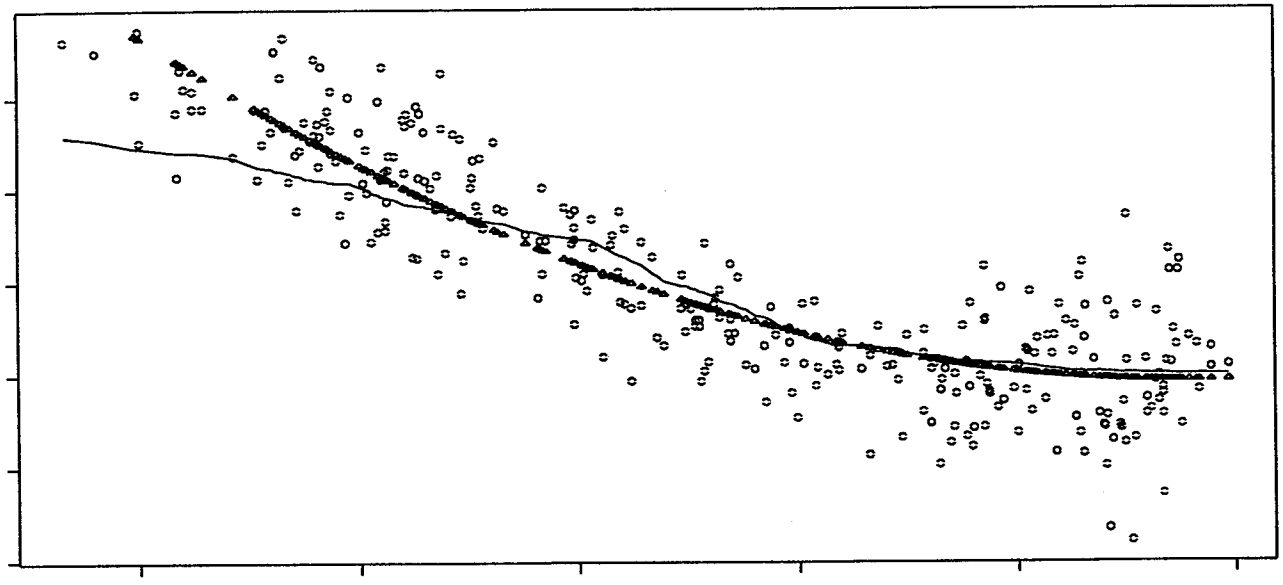
where *elec* is the log of electricity sales, TEMP is heating/cooling degree days measured relative to 68° Fahrenheit, *relprice* is the log of the ratio of the price of electricity to the price of natural gas, and *gdp* is the log of gross provincial product for Ontario. We begin by testing whether electricity sales and *gdp* are cointegrated under the assumption that *relprice* and TEMP are stationary, (though there is a growing body of evidence that global temperatures are indeed not stationary but trending upwards). The Johansen test indicates a strong cointegrating relationship. We therefore re-estimate the model in the form:

$$elec_t - gdp_t = f\left(TEMP_t\right) + \beta_1\, relprice_t + \varepsilon \tag{5.4.2}$$

Figure 5.4 contains estimates of a pure parametric specification where the temperature effect is modeled using a quadratic as well as estimates of the partial linear model (5.4.2). The price of electricity relative to natural gas is negative and strongly significant. In the partial linear model, the ratio of the coefficient estimate to the Newey-West standard error obtained using equation (5.3.7) is -3.23.

FIGURE 5.4: HETEROSKEDASTICITY/AUTOCORRELATION CONSISTENT STANDARD ERRORS
WEATHER AND ELECTRICITY DEMAND

VARIABLE	QUADRATIC MODEL		PARTIAL LINEAR MODEL[1]	
	Coef	Newey-West SE	Coef	Newey-West SE
c	-1.7072	0.0288	-	-
TEMP	-1.29×10^{-4}	3.82×10^{-5}	-	-
$TEMP^2$	4.07×10^{-7}	5.11×10^{-8}	-	-
relprice	-0.0695	0.0257	-.1103	.0341
s^2	.00317		.	
R^2	.788		.	



[1] Order of differencing m=3. Model estimated using *S-Plus*. The smooth curve represents the quadratic estimate of the temperature effect. The irregular curve is a kernel estimate of this effect obtained using the `ksmooth` function in *S-plus*.

5.5 Nonparametric Functions of Two of Three Variables

Consider again the pure nonparametric model:

$$y = f(x) + \varepsilon \tag{5.5.1}$$

but now permit x to be a vector of dimension k. Suppose again we are interested in estimating the residual variance. We would like to ensure that the data (y_1, x_1) .... (y_n, x_n) have been reordered so that the x's are 'close'.

For illustrative purposes, suppose the x's lie on a uniform grid on the unit square. Each point may be thought of as 'occupying' an area of $1/n$ and the distance between adjacent observations is therefore $1/n^{1/2}$.[24] If we use the first order differencing estimator $s^2_{diff} = \sum_{i=2}^{n} (y_i - y_{i-1})^2 / 2n$ then:

$$\begin{aligned}\left(s^2_{diff} - \sigma^2_\varepsilon \right) \\ = \left(\frac{1}{2n} \sum_{i=2}^{n} \left(\varepsilon_i - \varepsilon_{i-1}\right)^2 - \sigma^2_\varepsilon \right) + \frac{1}{2n} \sum_{i=2}^{n} \left(f(x_i) - f(x_{i-1})\right)^2 + \frac{2}{2n} \sum_{i=2}^{n} \left(\varepsilon_i - \varepsilon_{i-1}\right)\left(f(x_i) - f(x_{i-1})\right)\end{aligned} \tag{5.5.2}$$

The first term is $O_p\left(n^{-1/2}\right)$. Use $\|\cdot\|$ to denote the Euclidean distance between points and assume that f satisfies a Lipshitz constraint $\left|f(x_a) - f(x_b)\right| \le L \left\|x_a - x_b\right\|$. Then for the second term we have:

$$\frac{1}{2n} \sum_{i=2}^{n} \left(f(x_i) - f(x_{i-1})\right)^2 \le \frac{1}{2n} \sum_{i=2}^{n} L^2 \left\|x_i - x_{i-1}\right\|^2 = O\left(\frac{1}{n}\right) \tag{5.5.3}$$

---

[24] If x a scalar and the data are equally spaced on the unit interval, the distance between adjacent observations is $1/n$, much closer.

Consider now:

$$Var\left(\frac{2}{2n}\sum_{i=2}^{n}\varepsilon_i\left(f(x_i)-f(x_{i-1})\right)\right) = \frac{\sigma_\varepsilon^2}{n^2}\sum_{i=2}^{n}\left(f(x_i)-f(x_{i-1})\right)^2 \le \frac{\sigma_\varepsilon^2}{n^2}\sum_{i=2}^{n}L^2\left\|x_i - x_{i-1}\right\|^2 = O\left(\frac{1}{n^2}\right) \quad (5.5.4)$$

from which we can conclude that the third term of (5.5.2) is $O_P(1/n)$.

The point of this exercise is that if x is two-dimensional, differencing removes the nonparametric effect sufficiently quickly so that:

$$\left(s_{diff}^2 - \sigma_\varepsilon^2\right) = \left(\frac{1}{2n}\sum_{i=2}^{n}\left(\varepsilon_i - \varepsilon_{i-1}\right)^2 - \sigma_\varepsilon^2\right) + O\left(\frac{1}{n}\right) + O_P\left(\frac{1}{n}\right) \quad (5.5.5)$$

Compare this to the one-dimensional case where the second and third terms are $O\left(1/n^2\right)$ and $O_P\left(1/n^{3/2}\right)$ respectively.

With n points distributed on the uniform grid in the unit cube in $\mathbb{R}^k$, each point continues to 'occupy' a volume $1/n$ but the distance between adjacent observations is $1/n^{1/k}$ and $\sum\left(f(x_i)-f(x_{i-1})\right)^2 = O\left(n^{1-2/k}\right)$. Thus (5.5.3) becomes $O\left(1/n^{2/k}\right)$ and the *variance* of the third term, becomes $O\left(1/n^{1+2/k}\right)$. Hence with the x's lying in a k-dimensional cube, (5.5.5) becomes:

$$\begin{aligned}\left(s_{diff}^2 - \sigma_\varepsilon^2\right) &= \left(\frac{1}{2n}\sum_{i=2}^{n}\left(\varepsilon_i - \varepsilon_{i-1}\right)^2 - \sigma_\varepsilon^2\right) + O\left(\frac{1}{n^{2/k}}\right) + O_P\left(\frac{1}{n^{1/2+1/k}}\right) \\ &= O_P\left(\frac{1}{n^{1/2}}\right) + O\left(\frac{1}{n^{2/k}}\right) + O_P\left(\frac{1}{n^{1/2+1/k}}\right)\end{aligned} \quad (5.5.6)$$

How does the dimensionality of x affect the differencing estimator? First, s_{diff}^2 remains consistent regardless of the dimension of x. However, the bias term (the second term of (5.5.2) or (5.5.6)) converges to 0 more slowly as k increases. Second, s_{diff}^2 is $n^{1/2}$-consistent and

$$n^{1/2}\left(s_{diff}^2 - \sigma_\varepsilon^2\right) - n^{1/2}\left(\frac{1}{2n}\sum_{i=2}^{n}\left(\varepsilon_i - \varepsilon_{i-1}\right)^2 - \sigma_\varepsilon^2\right) \xrightarrow{P} 0 \quad (5.5.7)$$

only if k does not exceed 3. [25] This is important because throughout this essay, whenever we have derived the large sample distribution of an estimator or test statistic, we have used the property that differencing removes the nonparametric effect sufficiently quickly so that it can be ignored. Essentially, this has required that a condition like (5.5.7) holds.

With random x 's similar results hold so long as reasonable ordering rules are used. If x is a scalar, the obvious ordering rule is $x_1 \leq \ldots \leq x_n$. If x is of higher dimension, we propose the following ordering rule based on the nearest neighbor algorithm because it is simple to compute. (Other ordering rules for which the conclusion of Proposition 5.5.1 holds can easily be devised.)

Proposition 5.5.1: suppose x has support the unit cube in $\mathbb{R}^k$ with density bounded away from 0. Select ϵ positive and arbitrarily close to 0. Cover the unit cube with sub-cubes of volume $1/n^{1-\epsilon}$ each with sides $1/n^{(1-\epsilon)/k}$. Within each sub-cube construct a path using the nearest neighbor algorithm. Following this, knit the paths together by joining endpoints in contiguous sub-cubes to obtain a reordering of all the data. Then for any $\epsilon>0$, $1/n\sum \|x_i - x_{i-1}\|^2 = O_P\left(n^{-2(1-\epsilon)/k}\right)$.■

Since ϵ may be chosen arbitrarily close to 0 we write $1/n\sum \|x_i - x_{i-1}\|^2 \cong O_P\left(n^{-2/k}\right)$. *We may now assert that the key results of previous sections continues to hold so long as k does not exceed 3 and the above ordering rule is employed.*

Thus, Propositions 2.2.1,2.2.2 may be used to perform inference on the residual variance. Proposition 2.3.1 may be used to perform specification tests. Propositions 2.4.1,2.4.2 may be used for testing equality of regression functions. And, Propositions 2.5.1, 2.7.1, 2.8.1, 2.8.2 may be used to analyze data on the partial linear or partial parametric models. Combining differencing procedures with others as in Chapter 3 retains its validity, as do the bootstrap procedures in Chapter 4 and the adaptations to heteroskedastic and autocorrelated models.[26] [27]

---

[25] We note that Robinson's (1988) estimator of the partial linear model requires special treatment for dimensions exceeding 3.

[26] Additional testing procedures on nonparametric functions of up to dimension 3 may be found in Yatchew (1988). Tests of significance, symmetry and homogeneity are proposed. While that paper uses sample splitting to obtain the distribution of the test statistic, the device is unnecessary and the full data-set can be used to

Suppose one wants to test a null hypothesis which involves a *nonparametric* component against the semiparametric alternative $H_1 : y = z\beta + f(x_1, x_2) + \varepsilon$. For example, suppose one wants to test the significance of x_2, i.e., $H_0 : y = z\beta + f(x_1) + \varepsilon$. In such cases the test procedure in Proposition 2.3.1 needs modification.

Let P_{x_1} be the permutation matrix that reorders the data $(y_1, x_{11}, x_{21}, z_1), \ldots, (y_n, x_{1n}, x_{2n}, z_n)$ so that the values of x_1 are close and let $P_{x_1x_2}$ be the permutation matrix which reorders the data so that the vectors (x_{1i}, x_{2i}) are close. Define the unrestricted estimator of the residual variance

$$s^2_{x_1x_2} = \frac{1}{n}\left(DP_{x_1x_2}y - DP_{x_1x_2}Z\hat{\beta}_{diffx_1x_2}\right)'\left(DP_{x_1x_2}y - DP_{x_1x_2}Z\hat{\beta}_{diffx_1x_2}\right) \tag{5.5.8}$$

where $\hat{\beta}_{diffx_1x_2} = \left[\left(DP_{x_1x_2}Z\right)'DP_{x_1x_2}Z\right]^{-1}\left(DP_{x_1x_2}Z\right)'DP_{x_1x_2}y$. Next, define the restricted estimator as:

$$s^2_{x_1} = \frac{1}{n}\left(DP_{x_1}y - DP_{x_1}Z\hat{\beta}_{diffx_1}\right)'\left(DP_{x_1}y - DP_{x_1}Z\hat{\beta}_{diffx_1}\right) \tag{5.5.9}$$

which reorders according to just the variable x_1 and $\hat{\beta}_{diffx_1} = \left[\left(DP_{x_1}Z\right)'DP_{x_1}Z\right]^{-1}\left(DP_{x_1}Z\right)'DP_{x_1}y$.

Proposition 5.5.2: Suppose one wants to test $H_0 : y = z\beta + f(x_1) + \varepsilon$ against the alternative hypothesis $H_1 : y = z\beta + f(x_1, x_2) + \varepsilon$. If the null is is true, then

$$\left(\frac{mn}{2}\right)^{1/2}\frac{\left(s^2_{x_1} - s^2_{x_1x_2}\right)}{s^2_{x_1x_2}} \overset{D}{\sim} N(0, 1) \tag{5.5.10}$$ ■

---

calculate the restricted and unrestricted estimators of the residual variance. A test of homotheticity may also be devised.

[27] Hall, Kay and Titterington (1991) propose various optimal differencing coefficients for configurations of points on the grid in $\mathbb{R}^2$. The application is noise reduction in image processing using pixel data. Since economic data rarely comes in grid form, we have not described these procedures here.

Note the subtle change in the statistic from equation (2.3.2) which does not have $2^{1/2}$ in the denominator. If one wants to test the specification H_0: $y = f(x_1) + x_2\gamma + z\beta + \varepsilon$ against the alternative $H_1 : y = z\beta + f(x_1, x_2) + \varepsilon$ then the 'restricted' estimator of the residual variance $s^2_{x_1}$ needs to be redefined as

$$s^2_{x_1} = \frac{1}{n}\left(DP_{x_1} y - DP_{x_1} x_2 \hat{\gamma}_{diffx_1} - DP_{x_1} Z \hat{\beta}_{diffx_1} \right)' \left(DP_{x_1} y - DP_{x_1} x_2 \hat{\gamma}_{diffx_1} - DP_{x_1} Z \hat{\beta}_{diffx_1} \right)'$$

where $\hat{\gamma}_{diffx_1}, \hat{\beta}_{diffx_1}$ are obtained from an OLS regression of $DP_{x_1} y$ on $DP_{x_1} x_2$ and $DP_{x_1} Z$.

5.6 Empirical Applications

Hedonic Pricing of Housing Attributes

Housing prices are very much affected by location, an effect that has no natural parametric specification. Depending on the market, the price surface may be uni-modal, multi-modal or it may even have ridges (for example, prices along subway lines are often higher). We thus include a two-dimensional nonparametric location effect $f(x_1, x_2)$ where x_1, x_2 are location coordinates. The partial linear model below was estimated by Michael Ho (1995) using semiparametric least squares. The data consist of 92 detached homes in the Ottawa area which sold during 1987. The dependent variable y is SALE PRICE; z variables include lot size (LOTAREA), square footage of housing (USESPC), number of bedrooms (NRBED), average neighbourhood income (AVGINC), distance to highway (DHWY), presence of garage (GRGE), fireplace (FRPLC) or luxury appointments (LUX). Figure 5.6.1 contains estimates of a pure parametric model where the location effect is modeled using a linear specification. It also contains estimates of the partial linear model. Having estimated the parametric effects using differencing, smoothing is applied to the constructed data $\left(y_i - z_i \hat{\beta}_{diff}, x_{1i}, x_{2i} \right)$ to estimate the nonparametric effect.

We now respecify the model in (2.6.1) allowing both PRICE and AGE to appear nonparametrically:

$$dist = f(price, age) + \beta_1 \, income + \beta_2 \, drivers + \beta_3 \, hhsize + \beta_4 \, URBANDUM + \varepsilon \qquad (5.6.1)$$

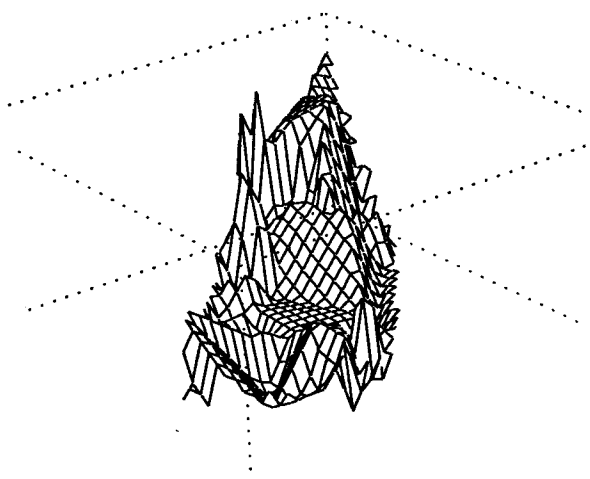
The upper panel of Figure 5.6.2 illustrates the scatter of data on PRICE and AGE and the path which we take through the data points for purposes of differencing. Estimates of the parametric effects are provided using third order differencing. These do not differ substantially from those where only PRICE is modeled nonparametrically (Figure 2.6.1 above). A test of the joint significance of the nonparametric variables using Proposition 2.3.1 yields a value of 8.3 which is strongly significant. A test of a fully parametric specification where PRICE and AGE enter log-linearly yields a statistic of .102. For much more extensive analysis of these data including the application of other testing procedures such as those of Aït-Sahalia, Bickel and Stoker (1998), see Yatchew and No (1999).

FIGURE 5.6.1: HEDONIC PRICES OF HOUSING ATTRIBUTES

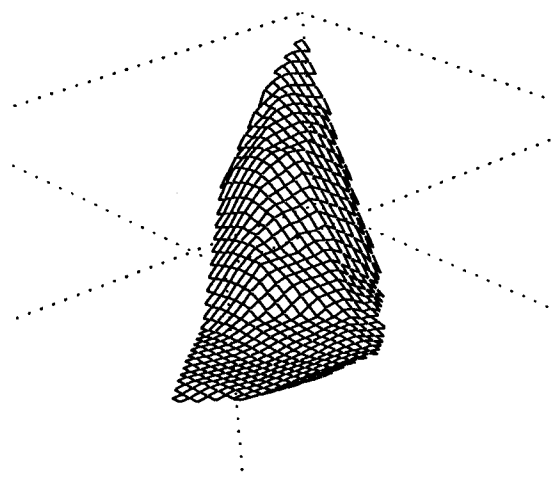
ESTIMATED MODELS

$y = \alpha + z\beta + \gamma_1 x_1 + \gamma_2 x_2 + \varepsilon$			$y = z\beta + f(x_1, x_2) + \varepsilon$		
OLS			Second Order Optimal Differencing [2]		
	Coeff	SE		Coeff	SE
α	74.0	18.0			
FRPLC	11.7	6.2	FRPLC	8.4	6.0
GRGE	11.8	5.1	GRGE	7.3	5.1
LUX	60.7	10.5	LUX	52.6	10.3
AVGINC	.478	.22	AVGINC	.10	.29
DHWY	-15.3	6.7	DHWY	-1.4	11.2
LOTAREA	3.2	2.3	LOTAREA	5.6	2.2
NRBED	6.6	4.9	NRBED	2.5	4.5
USESPC	21.1	11.0	USESPC	36.5	11.1
γ_1	7.5	2.2			
γ_2	-3.2	2.5	R^2	.71	
R^2	.62		s^2_{diff}	324.8	
s^2_{res}	424.3[1]				

DATA WITH PARAMETRIC EFFECT REMOVED



ESTIMATED LOCATION EFFECTS [3]



[1] Under the null that location has no effect, (f is constant), $s^2_{res} = 507.4$. [2] Data reordered using nearest neighbour algorithm for (x_{1i}, x_{2i}), $i = 1, \ldots, n$ (n=92). For the partial linear model we calculate $R^2 = 1 - s^2_{diff}/s^2_y$.
[3] Smoothed estimate obtained using `loess` function in *S-Plus*. Dependent variable is $y_i - z_i \hat{\beta}_{diff}$.

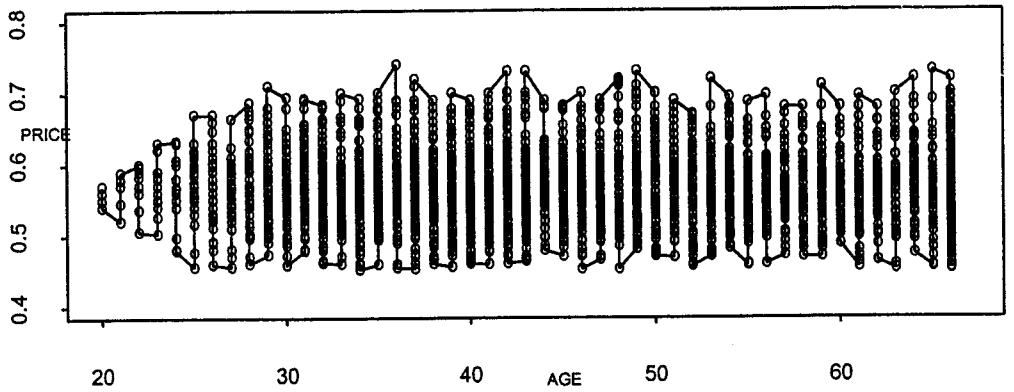
FIGURE 5.6.2: HOUSEHOLD GASOLINE DEMAND IN CANADA

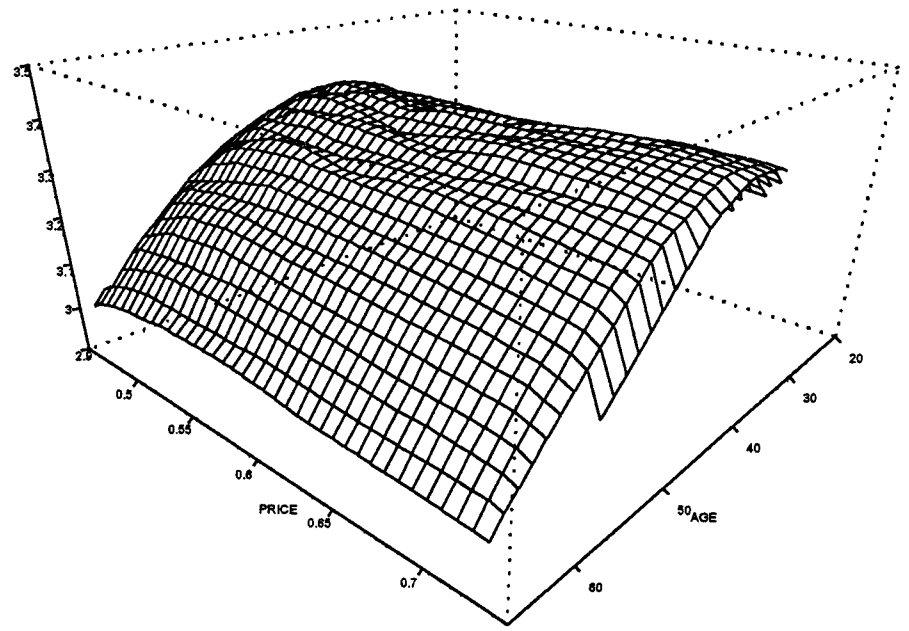
	coef	se
log income	0.2858	0.0258
log drivers	0.5501	0.0419
log hhsize	0.0945	0.0333
Urban dummy	-0.3018	0.0245
R^2	.2286	
s^2	.5248	

[1] Dependent variable is log(DIST) Var(dependent variable): .6803. Order of differencing m=3.

Upper panel right illustrated ordering of data: data are reordered first by age; then for even ages, data are reordered so that price is decreasing and for odd ages, data are reordered so that price is increasing.

Lower panel right illustrates nonparametric estimate of AGE and PRICE effects after removal of estimated parametric effects.





6. CONCLUDING REMARKS

In this essay, we have outlined a variety of techniques which use the differencing device. Our principal focus has been the partial linear model which we see as the entry level model for economists seeking to incorporate some variables nonparametrically. The differencing procedures described above include estimation of the partial linear model, specification testing, correction for heteroskedastic and autocorrelated residuals, testing equality of nonparametric regression functions, instrumental variable estimation and extensions to models where the parametric portion is nonlinear. Bootstrap procedures are also outlined.

Other familiar techniques may readily be adapted to this setting. For example, endogeneity of the nonparametric effects may be incorporated and tested using the specification described by Blundell and Duncan (1998), multi-equation seemingly unrelated regression estimation may be performed and differencing may be applied in a panel data setting, (see e.g., Yatchew (1999) and references therein). Various other data sets are also suitable for the application of the partial linear model — prominent among these are Engel curve models.

The differencing techniques described in this paper are applicable so long as the dimension of the nonparametric effect does not exceed 3. [28] (The parametric effect may be of arbitrary dimension.) While this may appear to be a limitation, it is our view that even if differencing techniques were limited to one (nonparametric) dimension, they should have the potential of significant 'market share'. The reason of course is that high dimensional nonparametric regression models, (unless they rely on additional structure such as separability), suffer from the curse of dimensionality. It is not surprising therefore that the vast majority of applied papers using nonparametric regression limit the nonparametric component to 1 or 2 dimensions.

[28] Indeed, from equation (5.5.6) it is clear that they work best if k=1.

REFERENCES

Ahn H. and J. Powell (1993): "Semiparametric Estimation of Censored Selection Models With a Nonparametric Selection Mechanism", *Journal of Econometrics*, 58, 2-29.

Aït-Sahalia, Y., P. Bickel, T. Stoker (1998): "Goodness-of-Fit Tests for Regression Using Kernel Methods", manuscript.

Anderson, T.W. (1971): *The Statistical Analysis of Time Series*, New York, Wiley.

Azzalini A., Adrian Bowman and Wolfgang Härdle (1989), "On the Use of Nonparametric Regression for Model Checking", *Biometrika*, 76, 1-11.

Baltagi, B., J. Hidalgo and Qi Li (1996): "A Nonparametric Test for Poolability Using Panel Data", *Journal of Econometrics*, 75, 345-367.

Barbeau E. (1995): *Polynomials*, Springer-Verlag, New York.

Beran Rudy and Gilles Ducharme (1991): *Asymptotic Theory for Bootstrap Methods in Statistics*, Centre de Recherches Mathématiques, Univerisité de Montréal.

Blundell, R. and A. Duncan (1998): "Kernel Regression in Empirical Microeconomics", *Journal of Human Resources,*33, 62-87.

Brooke, A., D. Kendrick and A. Meeraus (1992): *GAMS*, Scientific Press, Redwood City, California.

Cox D. and E. Koh (1989): "A Smoothing Spline Based Test of Model Adequacy in Polynomial Regression", *Annals of the Institute of Statistical Mathematics*, 41, 383-400.*

Delgado, M.A. (1993): "Testing the Equality of Nonparametric Regression Curves", *Statist. Probab. Lett.*, 17, 199-204.

Engle, Robert, Clive W.J. Granger, John Rice and Andrew Weiss (1986), "Semiparametric Estimates of the Relation Between Weather and Electricity Sales", *Journal of the American Statistical Association*, 81, 310-20.

Eubank Randall and Clifford H. Spiegelman (1990): "Testing the Goodness of Fit of a Linear Model Via Nonparametric Regression Techniques", *Journal of the American Statistical Association*, 85, 387-392.

Freedman, D.A. (1981): "Bootstrapping Regression Models", *Annals of Statistics*, 9,1218-1228.

Gozalo, Pedro L., (1993): "A Consistent Specification Test for Nonparametric Estimation of Regression Function Models", *Econometric Theory*, 9, 451-77.

Hall, P. and J.D. Hart (1990): "Bootstrap Test for Difference Between Means in Nonparametric Regression", *Journal of the American Statistical Association*, 85, 1039-1049.

Hall, P., J.W. Kay and D.M. Titterington (1990): "Asymptotically Optimal Difference-Based Estimation of Variance in Nonparametric Regression", *Biometrika*, 77, 521-8.

Hall, P., J.W. Kay and D.M. Titterington (1991): "On Estimation of Noise Variance in Two-Dimensional Signal Processing", *Advances in Applied Probability*, 23, 476-95.

Härdle Wolfgang, Sigbert Klinke and Berwin Turlach, (1995): *XploRe: An Interactive Statistical Computing Environment*, Springer-Verlag, New York.

Härdle Wolfgang and Enno Mammen, (1993), "Comparing Nonparametric vs Parametric Regression Fits", *Annals of Statistics*, 21, 1926-47.

Härdle W. and J.S. Marron (1990): "Semiparametric Comparison of Regression Curves", *Annals of Statistics*, 18, 63-89.

Hart, J. (1997): *Nonparametric Smoothing and Lack-of-Fit Tests*, Springer, New York.

Hausman, J. (1978): "Specification Tests in Econometrics", *Econometrica*, 46, 1251-71.

Hauman, J. and W. Newey (1995): "Nonparametric Estimation of Exact Consumer Surplus and Deadweight Loss", *Econometrica*. 63, 1445-76.

Ho, Michael (1995) *Essays on the Housing Market*, unpublished Ph.D. dissertation, University of Toronto.

Hong, Yongmiao and Halbert White, (1995): "Consistent Specification Testing via Nonparametric Series Regression", *Econometrica*, 63, 1133-1160.

Horowitz, Joel and Wolfgang Härdle, (1994): "Testing a Parametric Model Against a Semiparametric Alternative", *Econometric Theory*, 10, 821-48.

King, E., J.D. Hart and T.E. Wehrly (1991): "Testing the Equality of Regression Curves Using Linear Smoothers", *Statist. Probab. Lett.*, 12, 239-247.

Kulasekera, K.B. (1995): "Comparison of Regression Curves Using Quasi-Residuals", *Journal of the American Statistical Association*, 90, 1085-93.

Kulasekera, K.B. and J. Wang (1997): "Smoothing Parameter Selection for Power Optimality in Testing of Regression Curves", *Journal of the American Statistical Association*, 92, 500-11.

Kunsch, H.R. (1989): "The Jackknife and the Bootstrap for General Stationary Observations", *Annals of Statistics*, 17, 1217-1241.

Lahiri, S.N. (1992): "Edgeworth Correction by Moving Block Bootstrap for Stationary and Nonstationary Data", in *Exploring the Limits of the Bootstrap*, ed. R. LePage and L. Billard, 183-214, Wiley, New York.

Lahiri, S.N. (1996): "On Edgeworth Expansion and Moving Block Bootstrap for Studentized *M*-Estimators in Multiple Regression Models", *Journal of Multivariate Analysis*, 56, 42-59.

Lee, B.J. (1991): "A Model Specification Test Against the Nonparametric Alternative", Department of Economics, University of Colorado, manuscript.

Li, Qi (1994): "Some Simple Consistent Tests for a Parametric Regression Function versus Semiparametric or Nonparametric Alternatives", Department of Economics, University of Guelph, manuscript.

Liu, R.Y. and K. Singh (1992): "Moving Blocks Jackknife and Bootstrap Capture Weak Dependence", in *Exploring the Limits of the Bootstrap*, ed. R. LePage and L. Billard, 225-248, Wiley, New York.

McLeish, D.L.: (1974): "Dependent Central Limit Theorems and Invariance Principles", *Annals of Probability*, 2, 620-8.

Newey W. and K. West (1987): "A Simple Positive Semi-definite Heteroskedasticity and Autocorrelation Consistent Covariance Matrix", *Econometrica*, 55, 703-8.

Powell, J. (1987): "Semiparametric Estimation of Bivariate Latent Variable Models", Working Paper 8704, Social Systems Research Institute of Wisconsin, University of Wisconsin, Madison.

Rice J. (1984): "Bandwidth Choice for Nonparametric Regression", *Annals of Statistics*, 12, 1215-30.

Robinson, P.M. (1988): "Root-*N*-Consistent Semiparametric Regression", *Econometrica*, 56, 931-954.

Schmalensee R. and T. Stoker (1999): "Household Gasoline Demand in the United States, *Econometrica*, 67,645-662.

Schott, J.R. (1997): *Matrix Analysis for Statistics*, Wiley, New York.

Seifert, B., T. Gasser, A. Wolf (1993): "Nonparametric Estimation of Residual Variance Revisited", *Biometrika*, 80, 373-83.

Varian, H. (1992): *Microeconomic Analysis*, Third Edition, W. W. Norton and Co., New York.

Van de Geer, Sara, (1990): "Estimating a Regression Function", *Annals of Statistics*, 18, 907-924.

Whang Yoon-Jae and Donald W.K. Andrews (1993): "Tests of Specification for Parametric and Semiparametric Models", *Journal of Econometrics*, 57, 277-318.

White, H. (1980):"A Heteroskedasticity Consistent Covariance Matrix Estimator and a Direct Test for Heteroskedasticity", *Econometrica*, 48, 817-838.

White, H. (1985): *Asymptotic Theory for Econometricians,* Academic Press, New York.

Wooldridge, Jeffrey M. (1992): "A Test for Functional Form Against Nonparametric Alternatives", *Econometric Theory*, 8, 452-75.

Wu, Chien-Fu Jeff (1986): "Jackknife, Bootstrap and Other Resampling Methods in Regression Analysis" (with discussion), *Annals of Statistics,* 14, 1261-1350.

Yatchew, A.J. (1988): "Some Tests of Nonparametric Regressions Models", *Dynamic Econometric Modelling*, Proceedings of the Third International Symposium in Economic Theory and Econometrics, W. Barnett, E. Berndt and H. White (eds.), Cambridge University Press, 121-135.

Yatchew, A.J. (1997): "An Elementary Estimator of the Partial Linear Model", *Economics Letters,* 57, 135-143, additional examples contained in *Economics Letters* 1998, 59, 403-405.

Yatchew, A.J. (1998a): "Nonparametric Regression Techniques in Economics", *Journal of Economic Literature,* vol. XXXVI, 669-721.

Yatchew, A.J. (1998b): "An Elementary Nonparametric Differencing Test of Equality of Regression Functions", forthcoming, *Economics Letters.*

Yatchew,A.J. (1999):"Scale Economies in Electricity Distribution", forthcoming *Journal of Applied Econometrics.*

Yatchew,A.J. and Angela No (1999): "Household Gasoline Demand in Canada", manuscript.

Young S.G. and A.W. Bowman (1995): "Nonparametric Analysis of Covariance", *Biometrics*, 51, 920-931.

Zheng, John Xu (1996): "A Consistent Test of Functional Form Via Nonparametric Estimation Techniques", *Journal of Econometrics*, 75, 263-289.

APPENDIX A — PROOFS

NOTATION: if A, B are matrices of identical dimension define $\left[A \odot B\right]_{ij} = A_{ij} B_{ij}$.

LEMMA A.1: a) Suppose the components of $\vartheta = \left(\vartheta_1, \ldots, \vartheta_\xi\right)'$ are i.i.d. with $E\vartheta_i = 0$, $Var\left(\vartheta_i\right) = \sigma_\vartheta^2$ $E\vartheta_i^4 = \eta_\vartheta$ and covariance matrix $\sigma_\vartheta^2 I_\xi$. If A is a symmetric matrix, then $E\left(\vartheta' A \vartheta\right) = \sigma_\vartheta^2 \, trA$ and $Var\left(\vartheta' A \vartheta\right) = \left(\eta_\vartheta - 3\sigma_\vartheta^4\right) trA \odot A + \sigma_\vartheta^4 2\, trAA$. b) Consider the heteroskedastic case where $Var\left(\vartheta_i\right) = \sigma_i^2$, $E\vartheta_i^4 = \eta_i$, ϑ has the diagonal covariance matrix Ω and η is the diagonal matrix with entries η_i. Then $E\left(\vartheta' A \vartheta\right) = trA\Omega$ and $Var\left(\vartheta' A \vartheta\right) = tr\left(\eta \odot A \odot A - 3\Omega^2 \odot A \odot A\right) + 2\, tr\left(\Omega A \Omega A\right)$. For results of this type see e.g., Schott (1997, p.391, Theorem 9.18) or they may be proved directly.■

LEMMA A.2: suppose x has support the unit interval with density bounded away from 0. Given n observations on x, reorder them so that they are in increasing order: $x_1 \le \ldots \le x_n$. Then for any ε positive and arbitrarily close to 0, $1/n \sum \left(x_i - x_{i-1}\right)^2 = O_P\left(n^{-2(1-\varepsilon)}\right)$ ■

Proof: partition the unit interval into $n^{1-\varepsilon}$ sub-intervals and note that the probability of an empty sub-interval goes to zero as n increases. The maximum distance between observations in adjacent sub-intervals is $2/n^{1-\varepsilon}$ and the maximum distance between observations within a sub-interval is $1/n^{1-\varepsilon}$ from which the result follows immediately. ■

COMMENT ON LEMMA A.2: since ε may be chosen arbitrarily close to zero, we write $1/n \sum \left(x_i - x_{i-1}\right)^2 \cong O_P\left(n^{-2}\right)$. Note also that for fixed j, $1/n \sum \left(x_i - x_{i-j}\right)^2 \cong O_P\left(n^{-2}\right)$. For an *arbitrary* collection of points in the unit interval, the maximum value that $1/n \sum \left(x_i - x_{i-1}\right)^2$ can take is $1/n$ which occurs when all observations are at one of the two endpoints of the interval. ■

LEMMA A.3: suppose $\left(x_i, \varepsilon_i\right)$, $i=1,\ldots,n$ are i.i.d.. The x_i have density bounded away from zero on the unit interval and $\varepsilon_i \mid x_i \sim \left(0, \sigma_\varepsilon^2\right)$. Assume data have been reordered so that $x_1 \le \ldots \le x_n$. Define $f(x) = \left(f(x_1), \ldots, f(x_n)\right)'$ where the function f has a bounded first derivative. Let D be a differencing matrix of say order m. Then $f(x)' D' D f(x) = O_P\left(n^{-1+\varepsilon}\right)$ and $Var\left(f(x)' D' D \varepsilon\right) = O_P\left(n^{-1+\varepsilon}\right)$ where ε is positive and arbitrarily close to 0. ■

Proof: the result follows immediately from Yatchew (1997, Appendix, equations (A.2) and (A.3)).■

LEMMA A.4: suppose we are given n i.i.d. observations on (ψ, x) with probability law $f_{\psi x}(\psi, x) = f_{\psi|x}(\psi|x) f_x(x)$ where ψ is a vector and x is a scalar. Suppose we reorder the data to obtain $(\psi_1, x_1) \ldots (\psi_n, x_n)$ so that $x_1 \le \ldots \le x_n$. Then the ψ_i are conditionally independent of each other. That is, $f(\psi_1 \ldots \psi_n | x_1 \ldots x_n) = f_{\psi|x}(\psi_1|x_1) \ldots f_{\psi|x}(\psi_n|x_n)$. ■

Proof: note that the conditional probability law for the ψ_i is unaffected whether they are drawn before or after the x_i are reordered. ■

LEMMA A.5: suppose (y_i, x_i, z_i, w_i), $i=1,\ldots,n$ are i.i.d. where y and x are scalars and z and w are p and q dimensional vectors respectively. Suppose the data have been reordered so that $x_1 \le \ldots \le x_n$. Let Z be the nxp matrix of observations on z and W the nxq matrix of observations on w. Suppose $E(z|x)$ and $E(w|x)$ are smooth vector functions of x having first derivatives bounded. Let $\Sigma_{z|x} = E_x Var(z|x)$, $\Sigma_{w|x} = E_x Var(w|x)$ and $\Sigma_{zw|x} = E_x Cov(z,w|x)$ where $Cov(z,w|x)$ is the pxq matrix of covariances between the z and w variables conditional on x. Let $d_0, d_1, \ldots, d_m$ be differencing weights satisfying constraints (2.1.1), define δ using (2.1.6) and let D be the corresponding differencing matrix as in (2.1.2). Then:

$$\frac{Z'D'DZ}{n} \xrightarrow{P} \Sigma_{z|x} \qquad \frac{Z'D'DD'DZ}{n} \xrightarrow{P} (1+2\delta)\Sigma_{z|x}$$

$$\frac{W'D'DW}{n} \xrightarrow{P} \Sigma_{w|x} \qquad \frac{W'D'DD'DW}{n} \xrightarrow{P} (1+2\delta)\Sigma_{w|x}$$

$$\frac{Z'D'DW}{n} \xrightarrow{P} \Sigma_{zw|x} \qquad \frac{Z'D'DD'DW}{n} \xrightarrow{P} (1+2\delta)\Sigma_{zw|x}$$

Proof: since z has a smooth regression function on x, write $z_i = g(x_i) + u_i$ where g is a vector function with first derivatives bounded, $E(u_i | x_i) = 0$ and $E(Var(z_i | x_i)) = \Sigma_{z|x}$. Define $g(x) = (g(x_1), \ldots, g(x_n))'$, $Z = (z_1, \ldots, z_n)'$ and $U = (u_1, \ldots, u_n)'$. Then

$$\begin{aligned}\frac{Z'D'DZ}{n} &= \frac{U'D'DU}{n} + \frac{g(x)'D'Dg(x)}{n} + \frac{2g(x)'D'DU}{n} \\ &\cong \frac{U'D'DU}{n} + O_P\left(\frac{1}{n^{3/2}}\right)\end{aligned}$$

The second line uses Lemma A.3 above. Using (2.1.4) write

$$\frac{U'D'DU}{n} \doteq \frac{U'L_0U}{n} + \sum_{j=0}^{m-1} d_j d_{j+1} \frac{U'(L_1+L_1')U}{n} + \ldots\ldots + d_0 d_m \frac{U'(L_m+L_m')U}{n}$$

and note that all terms but the first on the right hand side converge to zero matrices. Thus, $U'DD'U/n \overset{P}{\to} \Sigma_{z|x}$ and $Z'DD'Z/n \overset{P}{\to} \Sigma_{z|x}$. Since the diagonal entries of $D'DD'D$ are $1+2\delta$, we may use similar arguments to show that $U'D'DD'DU/n \overset{P}{\to} (1+2\delta)\Sigma_{z|x}$ and that $Z'D'DD'DZ/n \overset{P}{\to} (1+2\delta)\Sigma_{z|x}$. Convergence of other quantities in the statement of the lemma may be proved by analogous reasoning. ■

COMMENTS ON LEMMA A.5:

More generally, suppose $(y_i, x_i, h(z_i)')$, $i=1,\ldots,n$ are i.i.d. where h is a p-dimensional vector function such that $E(h(z)|x)$ has first derivative bounded. Define $\sum_{h(z)|x}$ to be the pxp conditional covariance matrix of $h(z)$ given x. Let $h(Z)$ be the nxp matrix whose i-th row is $h(z_i)'$. Then:

$$\frac{h(Z)'D'Dh(Z)}{n} \overset{P}{\to} \Sigma_{h(z)|x} \qquad \frac{h(Z)'D'DD'Dh(Z)}{n} \overset{P}{\to} (1+2\delta)\Sigma_{h(z)|x} \quad ■$$

PROOF OF PROPOSITION 2.2.1: for the mean and variance use (2.2.6) and (2.2.10b). From (2.2.10a) note that s^2_{diff} has a band structure so that a finitely dependent CLT may be applied.■

PROOF OF PROPOSITION 2.2.2: use (2.2.13) to conclude that

$$\frac{1}{n}\sum_{i=m+1}^{n} (d_0 y_i + \ldots + d_m y_{i+m})^4 \overset{P}{\to} \eta_\varepsilon \left(\sum_{i=0}^{m} d_i^4 \right) + 6\sigma_\varepsilon^4 \left(\sum_{i=0}^{m-1} d_i^2 \sum_{j=i+1}^{m} d_j^2 \right)$$

from which the result follows immediately. ■

PROOF OF PROPOSITION 2.3.1: in large samples,

$$
\begin{aligned}
& n^{1/2}\left(s_{res}^2 - s_{diff}^2\right) \\
& \cong n^{1/2}\left(\frac{1}{n}\sum \varepsilon_i^2 - \frac{1}{n}\sum \left(d_0\varepsilon_i + d_1\,\varepsilon_{i+1} + \ldots + d_m\,\varepsilon_{i+m}\right)^2 \right) \\
& = -n^{1/2}\left(\left(\sum_{j=0}^{m-1} d_j d_{j+1}\right)\frac{2}{n}\sum \varepsilon_i\,\varepsilon_{i+1} + \left(\sum_{j=0}^{m-2} d_j d_{j+2}\right)\frac{2}{n}\sum \varepsilon_i\,\varepsilon_{i+2} + \ldots + d_0 d_m \frac{2}{n}\sum \varepsilon_i\,\varepsilon_{i+m} \right) \\
& = +\frac{n^{1/2}}{m}\left(\frac{1}{n}\sum \varepsilon_i\,\varepsilon_{i+1} + \frac{1}{n}\sum \varepsilon_i\,\varepsilon_{i+2} + \ldots + \frac{1}{n}\sum \varepsilon_i\,\varepsilon_{i+m} \right)
\end{aligned}
$$

which is asymptotically $N\left(0, \sigma_\varepsilon^4/m\right)$. To obtain the third line we use the condition $\sum_0^m d_j^2 = 1$. To obtain the fourth, we again use $\sum_j d_j d_{j+k} = -1/2m$, $k = 1,\ldots,m$. ■

PROOF OF PROPOSITION 2.4.1: using Lemma A.3 we have $(nT)^{1/2}\left(s_w^2 - \varepsilon'\left(I_T \otimes D'D\right)\varepsilon/nT\right) \overset{P}{\sim} 0$. Set $Q = I_T \otimes D'D$ and using the properties of $D'D$ and $D'DD'D$ note that $tr\,Q \doteq tr\,Q \odot Q \doteq nT$ and $tr\,QQ \doteq nT(1 + 1/2m)$. Use Lemma A.1 to conclude that $(nT)^{1/2}\left[\varepsilon'\left(I_T \otimes D'D\right)\varepsilon/nT - \sigma_\varepsilon^2\right]$ has asymptotic mean 0 and variance $\eta_\varepsilon + (1/m - 1)\sigma_\varepsilon^4$. To obtain asymptotic normality, note that $I_T \otimes D'D$ has a band diagonal structure and thus we may apply a finitely dependent central limit theorem. Hence $(nT)^{1/2}\left(s_w^2 - \sigma_\varepsilon^2\right) \overset{D}{\sim} N\left(0, \eta_\varepsilon + (1/m - 1)\sigma_\varepsilon^4\right)$. To prove the same result for s_p^2 (assuming all regression functions are identical) replicate the above argument substituting ε^* for ε where $\varepsilon^* = P_p\varepsilon$. ■

PROOF OF PROPOSITION 2.4.2: Apply Proposition 2.4.1 to conclude that Y is asymptotically normal with mean 0. Recall $Q_Y = P_p'\left(I_T \otimes D'D\right)'P_p - \left(I_T \otimes D'D\right)$. Note that (except for 'end effects'), $I_T \otimes D'D$ has ones on the main diagonal and that pre- and post-multiplying by a permutation matrix reorders but otherwise does not change these elements. Hence, (except for 'end effects') Q_Y has zeros on the main diagonal and $m\,tr\,Q_Y \odot Q_Y/nT \to 0$. Now apply Lemma A.1 to conclude that $Var(Y) \to 2\pi_Y\sigma_\varepsilon^4$. ■

COMMENT ON PROPOSITION 2.4.2: to see that $0 \le \hat{\pi}_Y = m\,tr\left(Q_Y Q_Y\right)/nT \le 1$, first note that since Q_Y is symmetric, $tr\left(Q_Y Q_Y\right)$ is the sum of squared entries of Q_Y thus $\hat{\pi}_Y \ge 0$. Furthermore,

$$
\begin{aligned}
tr(Q_Y Q_Y) &= \sum_{ij} \left[P_p'(I_T \otimes D'D) P_p - (I_T \otimes D'D) \right]_{ij}^2 \\
&\doteq \sum_{ij} \left[P_p' \left(I_T \otimes \left(L_0 - \frac{1}{2m}(L_1 + L_1' \ldots + L_m + L_m') \right) \right) P_p - \left(I_T \otimes \left(L_0 - \frac{1}{2m}(L_1 + L_1' \ldots + L_m + L_m') \right) \right) \right]_{ij}^2 \\
&\doteq \frac{1}{4m^2} \sum_{ij} \left[\left(I_T \otimes (L_1 + L_1' \ldots + L_m + L_m') \right) - P_p' \left(I_T \otimes (L_1 + L_1' \ldots + L_m + L_m') \right) P_p \right]_{ij}^2 \\
&\leq \frac{1}{4m^2} 4\, m\, nT = \frac{nT}{m}
\end{aligned}
$$

Hence, $\hat{\pi}_\Upsilon \leq 1$. To obtain the third line, recall that L_0 is an identity matrix of order *n*. To obtain the inequality in the last line, note that $I_T \otimes (L_1 + L_1' \ldots + L_m + L_m')$ and $P_p'(I_T \otimes (L_1 + L_1' \ldots + L_m + L_m')) P_p$ each have (except for end effects) *2mnT* ones and the remainder zeros. Hence, in the second last line we are taking the sum of squared elements of a matrix which has at most *4mnT* non-zero entries each of which equals ±1.■

PROOF OF PROPOSITION 2.5.1:

Define $g(x)$ and U as in the proof of Lemma A.5 above. Using Lemma A.3 note that differencing removes both $f(x)$ the direct effect of *x*, and $g(x)$ the indirect effect of *x* sufficiently quickly so that we have the following approximation:

$$
n^{1/2}(\hat{\beta} - \beta) \cong \left(\frac{Z'D'DZ}{n} \right)^{-1} \frac{Z'D'D\varepsilon}{n^{1/2}} \cong \left(\frac{U'D'DU}{n} \right)^{-1} \left(\frac{U'D'D\varepsilon}{n^{1/2}} \right) \tag{A2.5.1}
$$

Using Lemma A.5 and $\delta = 1/4m$ (see (2.2.8) and Appendix B) note that,

$$
Var\left(\frac{U'D'D\varepsilon}{n^{1/2}} \right) = \sigma_\varepsilon^2 E\left[\frac{U'D'DD'DU}{n} \right] \overset{P}{\to} \sigma_\varepsilon^2 \left(1 + \frac{1}{2m} \right) \Sigma_{z|x}
$$

and $(U'DD'U/n)^{-1} \overset{P}{\to} \Sigma_{z|x}^{-1}$. Thus,

$$
Var\left(n^{1/2}(\hat{\beta} - \beta) \right) \overset{P}{\to} \sigma_\varepsilon^2 \left(1 + \frac{1}{2m} \right) \Sigma_{z|x}^{-1}
$$

Finally, use (2.2.9) to write

$$\frac{U'D'D\varepsilon}{n^{1/2}} \doteq \frac{U'L_0\varepsilon}{n^{1/2}} - \frac{1}{2m}\left(\frac{U'\left(L_1+L_1'\right)\varepsilon}{n^{1/2}} + \ldots\ldots + \frac{U'\left(L_m+L_m'\right)\varepsilon}{n^{1/2}} \right)$$

and conclude asymptotic normality. ■

PROOF OF PROPOSITION 2.7.1: Take the first order conditions for β

$$\frac{1}{n}\frac{\partial r(Z,\hat{\beta})}{\partial\beta}D'D'\left(y - r(Z,\hat{\beta})\right) = 0$$

then expand in a first order Taylor series to obtain

$$\frac{1}{n}\frac{\partial r(Z,\beta)}{\partial\beta}D'D(y-r(Z,\beta)) + \left[\frac{1}{n}\sum_i \frac{\partial^2 r(Z,\beta^*)}{\partial\beta\partial\beta'}\left[D'D(y-r(Z,\beta^*))\right]_i - \frac{1}{n}\frac{\partial r(Z,\beta^*)}{\partial\beta}D'D\frac{\partial r(Z,\beta^*)}{\partial\beta'}\right]\left(\hat{\beta}_{nls}-\beta\right) = 0$$

where β^* lies between $\hat{\beta}_{diffnls}$ and β. Consistency and conditional independence of the ε_i implies that

$$\frac{1}{n}\sum_i \frac{\partial^2 r(Z,\beta^*)}{\partial\beta\partial\beta'}\left[D'D(y-r(Z,\beta^*))\right]_i \overset{P}{\to} 0$$

Refer to **Comments on Lemma A.5**, setting $h(z) = \partial r(z,\beta)/\partial\beta$ to conclude that

$$\frac{1}{n}\frac{\partial r(Z,\beta)}{\partial\beta}D'D\frac{\partial r(Z,\beta)}{\partial\beta'} \overset{P}{\to} \sum_{\frac{\partial r}{\partial\beta}|x}$$

$$\frac{1}{n}\frac{\partial r(Z,\beta)}{\partial\beta}D'DD'D\frac{\partial r(Z,\beta)}{\partial\beta'} \overset{P}{\to} \left(1+\frac{1}{2m}\right)\sum_{\frac{\partial r}{\partial\beta}|x}$$

The convergence is retained if we replace β with $\hat{\beta}_{diffnls}$ or β^*. Thus we may write

$$n^{1/2}\left(\hat{\beta}_{diffnls}-\beta\right) \cong \sum_{\frac{\partial r}{\partial\beta}|x}^{-1} \cdot \frac{\partial r(Z,\beta)}{\partial\beta}\frac{D'D\varepsilon}{n^{1/2}}$$

Next, note that

$$Var\left(\frac{\partial r(Z,\beta)}{\partial\beta}\frac{D'D\varepsilon}{n^{1/2}}\right) = \sigma_\varepsilon^2 E\left[\frac{1}{n}\frac{\partial r(Z,\beta)}{\partial\beta}D'DD'D\frac{\partial r(Z,\beta)}{\partial\beta'}\right] \overset{P}{\to} \left(1+\frac{1}{2m}\right)\sum_{\frac{\partial r}{\partial\beta}|x}$$

so that

$$Var\left(n^{1/2}\left(\hat{\beta}_{nls}-\beta\right)\right) \overset{P}{\sim} \left(1+\frac{1}{2m}\right)\Sigma^{-1}_{\frac{\partial r}{\partial \beta}|x}$$

Asymptotic normality follows straightforwardly. ■

PROOF OF THEOREM 2.8.1:

Define $g(x)$ and U as in the proof of Lemma A.5 above. Since w, the vector of instruments, has a smooth regression function on x, write $w_i = h(x_i) + v_i$ where h is a vector function with first derivatives bounded, $E(v_i \mid x_i) = 0$ and $E\left(Var(w_i \mid x_i)\right) = \Sigma_{w|x}$. Define $h(x)' = (h(x_1),\ldots,h(x_n))$, $W = (w_1,\ldots,w_n)'$ and $V = (v_1,\ldots,v_n)'$. Using Lemma A.3, note that differencing removes $f(x)$, $g(x)$ and $h(x)$ sufficiently quickly so that we have the following approximation:

$$\begin{aligned} n^{1/2}\left(\hat{\beta}_{diff2sls}-\beta\right) &\cong \left(\frac{Z'D'DW}{n}\left(\frac{W'D'DW}{n}\right)^{-1}\frac{W'D'DZ}{n}\right)^{-1}\frac{Z'D'DW}{n}\left(\frac{W'D'DW}{n}\right)^{-1}\frac{W'D'D\varepsilon}{n^{1/2}} \\ &\cong \left(\frac{U'D'DV}{n}\left(\frac{V'D'DV}{n}\right)^{-1}\frac{V'D'DU}{n}\right)^{-1}\frac{U'D'DV}{n}\left(\frac{V'D'DV}{n}\right)^{-1}\frac{V'D'D\varepsilon}{n^{1/2}} \end{aligned} \tag{A2.8.1}$$

Using Lemma A.5 and $\delta = 1/4m$ (see (2.2.8) and Appendix B) note that

$$Var\left(\frac{V'D'D\varepsilon}{n^{1/2}}\right) = \sigma^2_\varepsilon E\left[\frac{V'D'DD'DV}{n}\right] \overset{P}{\sim} \sigma^2_\varepsilon\left(1+\frac{1}{2m}\right)\Sigma_{w|x}$$

$(W'DD'W/n)^{-1} \overset{P}{\sim} \Sigma^{-1}_{w|x}$ and $(Z'DD'W/n)^{-1} \overset{P}{\sim} \Sigma^{-1}_{zw|x}$. Thus, after simplification

$$Var\left(n^{1/2}\left(\hat{\beta}_{diff2sls}-\beta\right)\right) \overset{P}{\sim} \sigma^2_\varepsilon\left(1+\frac{1}{2m}\right)\left[\Sigma_{zw|x}\Sigma^{-1}_{w|x}\Sigma'_{zw|x}\right]^{-1}$$

Finally, use (2.2.9) to write

$$\frac{V'D'D\varepsilon}{n^{1/2}} \doteq \frac{V'L_0\varepsilon}{n^{1/2}} - \frac{1}{2m}\left(\frac{V'(L_1+L_1')\varepsilon}{n^{1/2}}+\ldots\ldots+\frac{V'(L_m+L_m')\varepsilon}{n^{1/2}}\right)$$

and conclude asymptotic normality. ■

PROOF OF PROPOSITION 2.8.2:

Note that

$$Var\left(n^{1/2}\left(\hat{\beta}_{diff} - \hat{\beta}_{diff2sls} \right) \right) = nVar\left(\hat{\beta}_{diff} \right) + nVar\left(\hat{\beta}_{diff2sls} \right) - 2nCov\left(\hat{\beta}_{diff}, \hat{\beta}_{diff2sls} \right)$$

Using Propositions 2.5.1 and 2.8.1 we know that the first term on the right hand side converges to $\sigma_\varepsilon^2(1+1/2m)\Sigma_{z|x}^{-1}$ and the second to $\sigma_\varepsilon^2(1+1/2m)\left[\Sigma_{zw|x}\Sigma_{w|x}^{-1}\Sigma_{zw|x}'\right]^{-1}$. We need to establish the limit of the third term. Using appendix equations (A2.5.1) and (A2.8.1) and Lemma A.5 we have

$$\begin{aligned}
nCov\left(\hat{\beta}_{diff}, \hat{\beta}_{diff2sls} \right) &\cong E\left[n^{1/2}\left(\hat{\beta}_{diff} - \beta \right) n^{1/2}\left(\hat{\beta}_{diff2sls} - \beta \right)' \right] \\
&\cong E\left[\left(\frac{Z'D'DZ}{n} \right)^{-1} \frac{Z'D'D\varepsilon}{n^{1/2}} \frac{\varepsilon'D'DW}{n^{1/2}} \left(\frac{W'D'DW}{n} \right)^{-1} \frac{W'D'DZ}{n} \left(\frac{Z'D'DW}{n} \left(\frac{W'D'DW}{n} \right)^{-1} \frac{W'D'DZ}{n} \right)^{-1} \right] \\
&= \sigma_\varepsilon^2 E\left[\left(\frac{Z'D'DZ}{n} \right)^{-1} \frac{Z'D'D'D'DW}{n} \left(\frac{W'D'DW}{n} \right)^{-1} \frac{W'D'DZ}{n} \left(\frac{Z'D'DW}{n} \left(\frac{W'D'DW}{n} \right)^{-1} \frac{W'D'DZ}{n} \right)^{-1} \right] \\
&\overset{P}{\to} \sigma_\varepsilon^2\left(1 + \frac{1}{2m} \right) \Sigma_{z|x}^{-1}\left[\Sigma_{zw|x}\Sigma_{w|x}^{-1}\Sigma_{zw|x}' \right]\left[\Sigma_{zw|x}\Sigma_{w|x}^{-1}\Sigma_{zw|x}' \right]^{-1} \\
&= \sigma_\varepsilon^2\left(1 + \frac{1}{2m} \right) \Sigma_{z|x}^{-1}
\end{aligned}$$

Hence

$$Var\left(n^{1/2}\left(\hat{\beta}_{diff} - \hat{\beta}_{diff2sls} \right) \right) \overset{P}{\to} \sigma_\varepsilon^2\left(1 + \frac{1}{2m} \right)\left(\left[\Sigma_{zw|x}\Sigma_{w|x}^{-1}\Sigma_{zw|x}' \right]^{-1} - \Sigma_{z|x}^{-1} \right)$$

and the result follows immediately. ■

PROOF OF PROPOSITION 5.5.1: There are $n^{1-\varepsilon}$ sub-cubes and note that the probability of an empty sub-cube goes to zero as n increases. The maximum segment within each sub-cube is proportional to $1/n^{(1-\varepsilon)/k}$ as is the maximum segment between points in contiguous sub-cubes from which the result follows immediately. ■

PROOF OF PROPOSITION 5.5.2: if the null hypothesis H_0: $y = z\beta + f(x_1) + \varepsilon$ is true, then reordering according to either P_{x_1} or $P_{x_1x_2}$ will remove the nonparametric effect f in large samples. Furthermore, both $\hat{\beta}_{diffx_1}$ and $\hat{\beta}_{diffx_1x_2}$ will be $n^{1/2}$-consistent estimators of β. For purposes of the arguments below, the following approximations are valid:

$$s^2_{x_1x_2} \cong \frac{1}{n}\varepsilon' P'_{x_1x_2} D' D P_{x_1x_2}\varepsilon$$

and

$$s^2_{x_1} \cong \frac{1}{n}\varepsilon' P'_{x_1} D' D P_{x_1}\varepsilon$$

Hence

$$\begin{aligned} s^2_{x_1} - s^2_{x_1x_2} &\cong \frac{1}{n}\varepsilon' P'_{x_1} D' D P_{x_1}\varepsilon - \frac{1}{n}\varepsilon' P'_{x_1x_2} D' D P_{x_1x_2}\varepsilon \\ &\cong \frac{1}{2mn}\varepsilon'\left[P'_{x_1x_2}\left(L_1 + L_1' + \ldots + L_m + L_m'\right)P_{x_1x_2} - P'_{x_1}\left(L_1 + L_1' + \ldots + L_m + L_m'\right)P_{x_1}\right]\varepsilon \end{aligned}$$

Since the diagonal elements of the interposing matrix in square brackets are zero, we have $E\left(s^2_{x_1} - s^2_{x_1x_2}\right) \cong 0$ and using Lemma A.1 above:

$$\begin{aligned} &Var\left(s^2_{x_1} - s^2_{x_1x_2}\right) \\ &\cong \frac{2\sigma^4}{4m^2n^2}\left[2\,tr\left(L_1 + L_1' + \ldots + L_m + L_m'\right)^2 - 2\,trP'_{x_1x_2}\left(L_1 + L_1' + \ldots + L_m + L_m'\right)P_{x_1x_2}P'_{x_1}\left(L_1 + L_1' + \ldots + L_m + L_m'\right)P_{x_1}\right] \\ &\cong \frac{2\sigma^4}{mn} - \frac{\sigma^4}{m^2n}\frac{trQ}{n} \end{aligned}$$

where $Q = P'_{x_1x_2}\left(L_1 + L_1' + \ldots + L_m + L_m'\right)P_{x_1x_2}P'_{x_1}\left(L_1 + L_1' + \ldots + L_m + L_m'\right)P_{x_1}$. The quantity trQ/n measures, roughly speaking, the degree to which observations which are close when reordering according to x_1

remain close when reordering according to (x_1, x_2).[1] Hence we have

$$\frac{\left(s^2_{x_1x_2} - s^2_{x_1}\right)}{\left(\frac{2s^4_{x_1x_2}}{mn} - \frac{s^4_{x_1x_2}}{m^2 n}\frac{tr\,Q}{n}\right)^{1/2}} \overset{D}{\sim} N(0,1)$$

which yields an asymptotically valid test. However, since (x_1, x_2) has density bounded away from 0 on the unit square, *tr Q/n* converges to zero which yields equation (5.5.10). ■

[1]Storage and multiplication of permutation matrices may be performed efficiently using vectors of length *n* where the *i*th entry of the vector records in which row the '1' is located. This in turn dramatically reduces the burden of computing *tr Q* since *Q* may be rewritten as the sum of $4m^2$ permutation matrices where *m* is the order of differencing.

APPENDIX B: OPTIMAL DIFFERENCING WEIGHTS[2]

PROPOSITION B1: Consider the optimization problem

$$\min_{d_0,d_1,\ldots,d_m} \delta = \sum_{k=1}^{m}\left(\sum_{j=0}^{m-k} d_j d_{j+k}\right)^2 \qquad s.t. \qquad \sum_{j=0}^{m} d_j = 0 \quad \sum_{j=0}^{m} d_j^2 = 1 \tag{B1.1}$$

then

$$\sum_{j=0}^{m-k} d_j d_{j+k} = -\frac{1}{2m} \qquad k = 1,..,m \tag{B1.2}$$

in which case $\delta = 1/4m$. Furthermore, $d_0, d_1, \ldots, d_m$ may be chosen so that the roots of

$$d_0 z^m + d_1 z^{m-1} + \ldots + d_{m-1} z^1 + d_m = 0 \tag{B1.3}$$

lie on or outside the unit circle. ■

PROOF OF PROPOSITION B1: For purposes of interpretation it will be convenient to think of the differenced residuals as a moving average process. Define

$$\varepsilon_i^* = d_0 \varepsilon_i + \ldots + d_m \varepsilon_{i+m}$$

and

$$\rho_k = \sum_{j=0}^{m-k} d_j d_{j+k} \qquad k = 0,..,m$$

Note that $\rho_k = corr\left(\varepsilon_i^*, \varepsilon_{i+k}^*\right)$. For k =0, we have $\rho_0 = \sum_{j=0}^{m} d_j^2 = 1$. Next we have

$$0 = \left(\sum_{j=0}^{m} d_j\right)^2 = \sum_{j=0}^{m} d_j^2 + 2\sum_{k=1}^{m}\left(\sum_{j=0}^{m-k} d_j d_{j+k}\right) = 1 + 2\sum_{k=1}^{m} \rho_k$$

---

[2] Reference: Hall et al (1990).

which implies $\sum_{k=1}^{m} \rho_k = -½$. Thus (B1.1) may be written as

$$\min_{\rho_1,\ldots,\rho_m} \sum_{k=1}^{m} \rho_k^2 \qquad s.t. \qquad \sum_{k=1}^{m} \rho_k = -½$$

which is minimized when the ρ_k are equal to each other in which case $\rho_k = -1/2m,\ k = 1,\ldots,m$ so we have proved (B1.2).

As Hall et al (1990) point out, the objective is to select moving average weights which reproduce the covariance structure:

$$\begin{aligned} cov\left(d_0\varepsilon_t+\ldots+d_m\varepsilon_{t+m},\, d_0\varepsilon_{t+k}+\ldots+d_m\varepsilon_{t+m+k}\right) &= -\frac{1}{2m}\sigma_\varepsilon^2 && k = 1,\ldots,m \\ &= 0 && k > m \qquad \text{(B1.4)} \\ var\left(d_0\varepsilon_t+\ldots+d_m\varepsilon_{t+m}\right) &= \sigma_\varepsilon^2 \end{aligned}$$

This can be achieved by solving:

$$R(z) = -\frac{1}{2m}\left[z^{2m}+z^{2m-1}+\ldots+z^{m+1}-2mz^m+z^{m-1}+\ldots+z+1\right] = 0$$

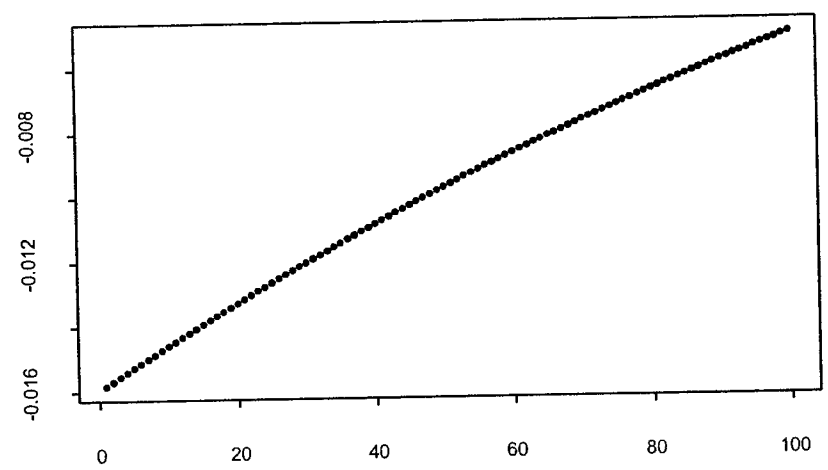
It is easy to show that '1' is a root of $R(z)$ with multiplicity 2. Furthermore, the polynomial is 'self-reciprocal' (see e.g. Anderson (1971, p.224), Barbeau (1995, p.22-23,152)), so that if $r = (a+b\iota)$ is a root then so is $1/r = (a-b\iota)/(a^2+b^2)$. (As usual, ι denotes $\sqrt{-1}$.) Thus, the set of all roots is $\Re = \{1, r_2,\ldots,r_m, 1, 1/r_2,\ldots,1/r_m\}$ where $|r_j|>1$, $j=2,\ldots,m$. A self-reciprocal polynomial may be rewritten in the form $R(z) = z^m M(z)M(1/z)$ where M is a polynomial with real coefficients. There are, however, multiple ways to construct M. In particular, obtain any partition of the roots $\Re = S \cup S^c$ satisfying the following conditions: if $s \varepsilon S$ then $1/s \varepsilon S^c$; if in addition s is complex and $s \varepsilon S$ then $\bar{s} \varepsilon S$. Compose M using the roots in S and normalize the coefficients of M so that their sum of squares equals 1. Then, by construction, the coefficients reproduce the covariance structure in (B1.4) and are therefore optimal differencing weights. Valid partitioning requires only that reciprocal pairs be separated (so that $z^m M(z)M(1/z) = R(z)$) and that conjugate pairs be kept together (to ensure that M has real coefficients). Of course, there is only one partition which separates the two unit roots and those that are respectively inside and outside the unit circle. ■

COMMENT: For example, if $m=4$, then the roots of $R(z)$ are given by:

$$\Re = \{1,\ r_2=-.2137-1.7976\iota,\ r_3=-.2137+1.7976\iota,\ r_4=-1.9219,\ 1,\ 1/r_2,\ 1/r_3,\ 1/r_4\}$$

Note that r_2, r_3, r_4 lie outside the unit circle. Taking $S = \{1, 1/r_2, 1/r_3, r_4\}$ yields differencing weights $(0.2708, -0.0142, 0.6909, -0.4858, -0.4617)$ which are those obtained by Hall et al (1990, p. 523, Table 1). If one takes $S = \{1, r_2, r_3, r_4\}$ then the differencing weights become $(0.8873, -0.3099, -0.2464, -0.1901, -0.1409)$. Figure 2.1 in the text tabulates differencing weights up to order 10 where S consists of the root '1' and all roots outside the unit circle.[3] Note that the 'spike' occurs at d_0 whether m is even or odd. The remaining weights $d_1,..,d_m$ are negative and monotonically increasing to 0. (Order or sign reversal preserves optimality of a sequence.) The pattern continues to be present as m increases. Weights for m=100, (excluding d_0 which equals 0.99454) are plotted in the figure below. ■

OPTIMAL DIFFERENCE SEQUENCE : m=100



$(d_1, \ldots, d_{100})$

Results obtained by minimizing $\sum_{k=1}^{m}\left(\sum_j d_j d_{j+k}\right)^2$ subject to $\sum_{j=0}^{m} d_j = 0$, $\sum_{j=0}^{m} d_j^2 = 1$ and $d_0 \geq .9$. Software used was GAMS, see Brooke, Kendrick and Meeraus (1992); $d_0 = 0.99454$ and is not shown.

[3] Higher order optimal differencing weights are available from the author.

APPENDIX C: VARIABLE DEFINITIONS

Here we summarize variable definitions for the various examples which have been included in the text. We follow the convention that upper case variables are in levels, while lower case italicized variables correspond to transformed variables in log form.

Scale Economies in Electricity Distribution

TC	total costs	*tc*	log(TC/CUST)
CUST	number of customers	*cust*	log(CUST)
WAGE	wage of lineman	*wage*	log(WAGE)
TOTPLANT	accum. gross investment	*pcap*	log(TOTPLANT/KMWIRE)
PUC	public utility commission dummy		
KWH	kilo-watt hour sales	*kwh*	log(KWH/CUST)
LIFE	fixed assets remaining lifetime	*life*	log(LIFE)
LF	load factor	*lf*	log(LF)
KMWIRE	kilometres of distribution wire	*kmwire*	log(KMWIRE/CUST)

Household Demand for Gasoline in Canada

dist	log of distance traveled per month
price	log of price of liter of gasoline
income	log of annual household income
drivers	log of number of licenced drivers in household
hhsize	log of number of members of household
URBANDUM	dummy for urban dwellers
RETIREDUM	dummy for households where head is over 65

Weather and Electricity Demand

elec	log of monthly electricity sales
relprice	log of ratio of price of electricity to the price of natural gas
gdp	log of Ontario gross GDP
TEMP	heating/cooling degree days relative to 68° Fahrenheit

Hedonic Pricing of Housing Attributes

SALEPRICE	sale price of house
FRPLC	dummy for fireplace(s)
GRGE	dummy for garage
LUX	dummy for luxury appointments
AVGINC	income
DWHY	distance to highway
LTAREA	area of lot
NRBED	number of bedrooms
USESPC	useable space