7.3 MEDIAN AND QUANTILE REGRESSION

We maintain the essential assumptions of the linear regression model,

$$y = \mathbf{x}'\boldsymbol{\beta} + \varepsilon$$

where $E[\varepsilon|\mathbf{x}] = 0$ and $E[y|\mathbf{x}] = \mathbf{x}'\boldsymbol{\beta}$. If $\varepsilon|\mathbf{x}$ is normally distributed, so that the distribution of $\varepsilon|\mathbf{x}$ is also symmetric, then the median, $\text{Med}[\varepsilon|\mathbf{x}]$, is also zero and $\text{Med}[y|\mathbf{x}] = \mathbf{x}'\boldsymbol{\beta}$. Under these assumptions, least squares remains a natural choice for estimation of $\boldsymbol{\beta}$. But, as we explored in Example 4.5, **least absolute deviations** is a possible alternative that might even be preferable in a small sample. Suppose, however, that we depart from the second assumption directly. That is, the statement of the model is

+LAD

$$\text{Med}[y|\mathbf{x}] = \mathbf{x}'\boldsymbol{\beta}.$$

This result suggests a motivation for LAD in its own right, rather than as a robust (to outliers) alternative to least squares.[9] The conditional median of $y_i|\mathbf{x}_i$ might be an interesting function. More generally, other quantiles of the distribution of $y_i|\mathbf{x}_i$ might also be of interest. For example, we might be interested in examining the various quantiles of the distribution of income or spending. Quantile regression (rather than least squares) is used for this purpose. The (linear) **quantile regression model** can be defined as

FN 9

AU: KT "quantile regression model" is already KT in chap. Here also?

$$Q[y|\mathbf{x},q] = \mathbf{x}'\boldsymbol{\beta}_q \text{ such that } \text{Prob}[y \le \mathbf{x}'\boldsymbol{\beta}_q \,|\, \mathbf{x}] = q,\ 0 < q < 1. \quad (7\text{-}33)$$

The **median regression** would be defined for $q = ½$. Other focal points are the lower and upper quartiles, $q = ¼$ and $q = ¾$, respectively. We will develop the median regression in detail in Section 7.3.1, once again largely as an alternative estimator in the linear regression setting.

KT

AU: KT "median regression" not in chap. list

The quantile regression model is a richer specification than the linear model that we have studied thus far, because the coefficients in ~~(**)~~ (7-33) are indexed by q. The model is nonparametric—it requires a much less detailed specification of the distribution of $y|\mathbf{x}$. In the simplest linear model with fixed coefficient vector, $\boldsymbol{\beta}$, the quantiles of $y|\mathbf{x}$ would be defined by variation of the constant term. The implication of the model is shown in Figure 7.4. For a fixed β and conditioned on x, the value of $\alpha_q + \beta x$ such that $\text{Prob}(y < \alpha_q + \beta x)$ is shown for $q = 0.15$, 0.5 and 0.9 in Figure 7.4. There is a value of α_q for each quantile. In Section 7.3.2, we will examine the more general specification of the quantile regression model in which the entire coefficient vector plays the role of α_q in Figure 7.4.

FIG 7.4

9 In Example 4.5, we considered the possibility that in small samples with possibly thick-tailed disturbance distributions, the LAD estimator might have smaller variance than least squares.

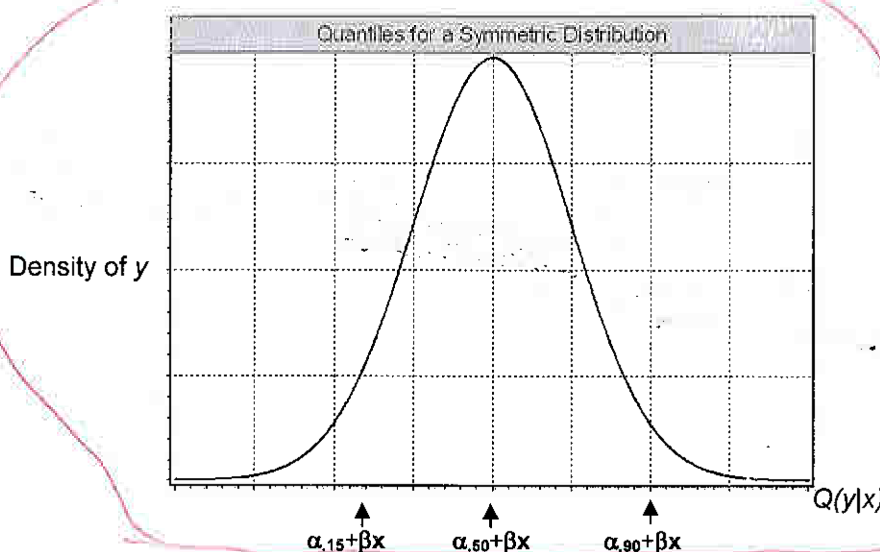


Figure 7.4 Quantile Regression Model

7-34

~~parametric estimator when the assumption of the distribution is correct. Once again, in the frontier function setting, least squares may be robust for the slopes, and it is the most efficient estimator that uses only the orthogonality of the disturbances and the regressors, but it will be inferior to the maximum likelihood estimator when the two-part normal distribution is the correct assumption.~~

~~14.3.1 GMM ESTIMATION IN ECONOMETRICS~~

~~Recent applications in economics include many that base estimation on the **method of moments.** The generalized method of moments departs from a set of model based moment equations, $E[\mathbf{m}(y_i, \mathbf{x}_i, \boldsymbol{\beta})] = \mathbf{0}$, where the set of equations specifies a relationship known to hold in the population. We used one of these in the preceding paragraph. The least squares estimator can be motivated by noting that the essential assumption is that $E[\mathbf{x}_i(y_i - \mathbf{x}_i'\boldsymbol{\beta})] = \mathbf{0}$. The estimator is obtained by seeking a parameter estimator, $\mathbf{b}$, which mimics the population result; $(1/n)\Sigma_i[\mathbf{x}_i(y_i - \mathbf{x}_i'\mathbf{b})] = \mathbf{0}$. These are, of course, the normal equations for least squares. Note that the estimator is specified without benefit of any distributional assumption. Method of moments estimation is the subject of Chapter 15, so we will defer further analysis until then.~~

7.3.1 ~~14.3.2~~ LEAST ABSOLUTE DEVIATIONS ESTIMATION

Least squares can be severely distorted by outlying observations. Recent applications in microeconomics and financial economics involving thick-tailed disturbance distributions, for example, are particularly likely to be affected by precisely these sorts of observations. (Of course, in those applications in finance involving hundreds of thousands of observations, which are becoming commonplace, this discussion is moot.) These applications have led to the proposal of "robust" estimators that are unaffected by outlying observations.[10] In this section, we will examine one of these, the least absolute deviations, or LAD estimator.

FN 10

That least squares gives such large weight to large deviations from the regression causes the results to be particularly sensitive to small numbers of atypical data points when the sample size is small or moderate. The **least absolute deviations** (LAD) estimator has been suggested as an alternative that remedies (at least to some degree) the problem. The LAD estimator is the solution to the optimization problem,

AU: KT "least absolute deviations (LAD)" already KT in chap. Here also?

$$\text{Min}_{\mathbf{b}_0} \sum_{i=1}^{n} |y_i - \mathbf{x}_i'\mathbf{b}_0|.$$

The LAD estimator's history predates least squares (which itself was proposed over 200 years ago). It has seen little use in econometrics, primarily for the same reason that Gauss's method (LS) supplanted LAD at its origination; LS is vastly easier to compute. Moreover, in a more modern vein, its statistical properties are more firmly established than LAD's and samples are usually large enough that the small sample advantage of LAD is not needed.

10 ~~7~~ For some applications, see Taylor (1974), Amemiya (1985, pp. 70–80), Andrews (1974), Koenker and Bassett (1978), and a survey written at a very accessible level by Birkes and Dodge (1993). A somewhat more rigorous treatment is given by Hardle (1990).

The LAD estimator is a special case of the quantile regression:

$$\text{Prob}[y_i \leq \mathbf{x}_i'\beta] = q. \quad \text{subscript } q$$

The LAD estimator estimates the *median regression*. That is, it is the solution to the quantile regression when $q = 0.5$. Koenker and Bassett (1978, 1982), Huber (1967), and Rogers (1993) have analyzed this regression.[3] [11] Their results suggest an estimator for the asymptotic covariance matrix of the **quantile regression** estimator,

AU: KT "quantile regression" already KT in chap. Here also?

FN 11

$$\text{Est. Asy. Var}[\mathbf{b}_q] = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{D}\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1},$$

where $\mathbf{D}$ is a diagonal matrix containing weights

$$d_i = \left[\frac{q}{f(0)}\right]^2 \text{ if } y_i - \mathbf{x}_i'\beta \text{ is positive and } \left[\frac{1-q}{f(0)}\right]^2 \text{ otherwise,}$$

FN 12

and $f(0)$ is the true density of the disturbances evaluated at 0.[4] [12] [It remains to obtain an estimate of $f(0)$.] There is a useful symmetry in this result. Suppose that the true density were normal with variance σ^2. Then the preceding would reduce to $\sigma^2(\pi/2)(\mathbf{X}'\mathbf{X})^{-1}$, 4.5 which is the result we used in Example 17.4 ~~to compare estimates of the median and the mean in a simple situation of random sampling~~. For more general cases, some other 12.4 empirical estimate of $f(0)$ is going to be required. Nonparametric methods of density estimation are available [see Section 14.4 and, e.g., Johnston and DiNardo (1997, pp. 370–375)]. But for the small sample situations in which techniques such as this are most desirable (our application below involves 25 observations), nonparametric kernel 4.5 density estimation of a single ordinate is optimistic; these are, after all, asymptotic results. But asymptotically, as suggested by Example 17.4, the results begin overwhelmingly to favor least squares. For better or worse, a convenient estimator would be a KT **kernel density estimator** as described in Section 14.4.1. Looking ahead, the computation would be

AU: Confirm Section x-ref number is correct

$$\hat{f}(0) = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{h}K\left[\frac{e_i}{h}\right] \qquad \text{ital.}$$

where h is the KT **bandwidth** (to be discussed shortly), $K[.]$ is a weighting, or kernel function and e_i, $i = 1, \ldots, n$ is the set of residuals. There are no hard and fast rules for choosing h; one popular choice is that used by Stata (2006), $h = .9s/n^{1/5}$. The kernel function is likewise discretionary, though it rarely matters much which one chooses; the logit kernel (see Table 14.2) is a common choice.

AU: Confirm table x-ref number is correct

KT The **bootstrap** method of inferring statistical properties is well suited for this application. Since the efficacy of the bootstrap has been established for this purpose, the search for a formula for standard errors of the LAD estimator is not really

11 [3]Powell (1984) has extended the LAD estimator to produce a robust estimator for the case in which data on the dependent variable are censored, that is, when negative values of y_i are recorded as zero. See Example 14.7 for discussion and Melenberg and van Soest (1996) for an application. For some related results on other semiparametric approaches to regression, see Butler et al. (1990) and McDonald and White (1993).

12 [4]Koenker suggests that for independent and identically distributed observations, one should replace d_i with the constant $a = q(1-q)/[f(F^{-1}(q))]^2 = [.50/f(0)]^2$ for the median (LAD) estimator. This reduces the expression to the true asymptotic covariance matrix, $a(\mathbf{X}'\mathbf{X})^{-1}$. The one given is a sample estimator which will behave the same in large samples. (Personal communication to the author.)

7-36

Au: Are "LAD's as Subs ok Italic?

necessary. The bootstrap estimator for the asymptotic covariance matrix can be computed as follows:

$$\text{Est. Var}[\mathbf{b}_{LAD}] = \frac{1}{R}\sum_{r=1}^{R}(\mathbf{b}_{LAD}(r) - \mathbf{b}_{LAD})(\mathbf{b}_{LAD}(r) - \mathbf{b}_{LAD})',$$

where $\mathbf{b}_{LAD}$ is the LAD estimator and $\mathbf{b}_{LAD}(r)$ is the rth LAD estimate of $\boldsymbol{\beta}$ based on a sample of n observations, drawn with replacement, from the original data set.

Example ~~14.4~~ 7.9 LAD Estimation of a Cobb–Douglas Production Function

Zellner and Revankar (1970) proposed a generalization of the Cobb–Douglas production function that allows economies of scale to vary with output. Their statewide data on Y = value added (output), K = capital, L = labor, and N = the number of establishments in the transportation industry are given in Appendix Table F14.1. For this application, estimates of the Cobb–Douglas production function, F7.2

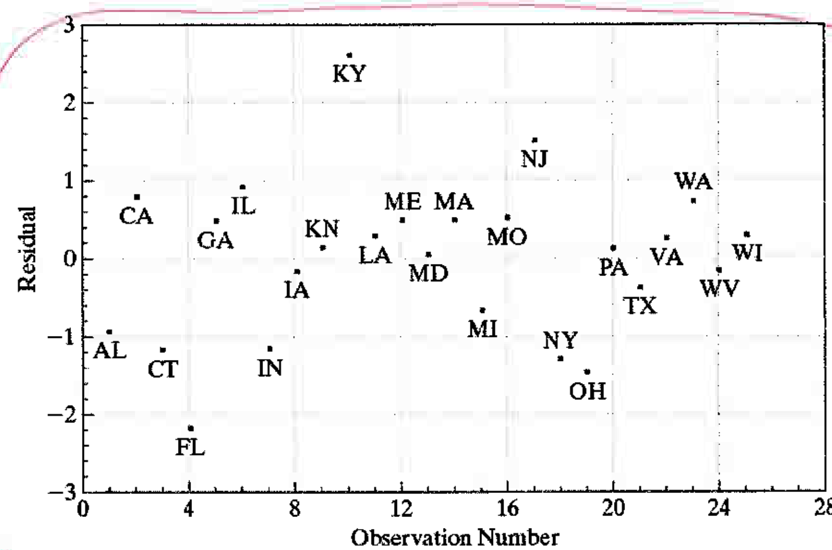
$$\ln(Y_i/N_i) = \beta_1 + \beta_2 \ln(K_i/N_i) + \beta_3 \ln(L_i/N_i) + \varepsilon_i,$$

FIG 7.5 — 7.5

TB 7.4

are obtained by least squares and LAD. The standardized least squares residuals shown in Figure 14.1 suggest that two observations (Florida and Kentucky) are outliers by the usual construction. The least squares coefficient vectors with and without these two observations are (2.293, 0.279, 0.927) and (2.205, 0.261, 0.879), respectively, which bears out the suggestion that these two points do exert considerable influence. Table 14.1 presents the LAD estimates of the same parameters, with standard errors based on 500 bootstrap replications. The LAD estimates with and without these two observations are identical, so only the former are presented. Using the simple approximation of multiplying the corresponding OLS standard error by $(\pi/2)^{1/2} = 1.2533$ produces a surprisingly close estimate of the bootstrap estimated standard errors for the two slope parameters (0.102, 0.123) compared with the bootstrap estimates of (0.124, 0.121). The second set of estimated standard errors are based on Koenker's suggested estimator, $.25/\hat{f}^2(0) = 0.25/1.5467^2 = 0.104502$. The bandwidth and kernel function are those suggested earlier. The results are surprisingly consistent given the small sample size. 7.4

FIGURE ~~14.1~~ 7.5 Standardized Residuals for Production Function.



7-37

7.4

TABLE ~~14.1~~ LS and LAD Estimates of a Production Function

	Least Squares			LAD				
					Bootstrap		Kernel Density	
Coefficient	Estimate	Standard Error	t Ratio	Estimate	Std. Error	t Ratio	Std. Error	t Ratio
Constant	2.293	0.107	21.396	2.275	0.202	11.246	0.183	12.374
β_k	0.279	0.081	3.458	0.261	0.124	2.099	0.138	1.881
β_l	0.927	0.098	9.431	0.927	0.121	7.637	0.169	5.498
Σe^2	0.7814			0.7984				
$\Sigma\|e\|$	3.3652			3.2541				

14.3.3 PARTIALLY LINEAR REGRESSION

The proper functional form in the linear regression is an important specification issue. We examined this in detail in Chapter 6. Some approaches, including the use of dummy variables, logs, quadratics, and so on, were considered as means of capturing nonlinearity. The translog model in particular (Example 2.4) is a well-known approach to approximating an unknown nonlinear function. Even with these approaches, the researcher might still be interested in relaxing the assumption of functional form in the model. The **partially linear model** [analyzed in detail by Yatchew (1998, 2000)] is another approach. Consider a regression model in which one variable, x, is of particular interest, and the functional form with respect to x is problematic. Write the model as

$$y_i = f(x_i) + \mathbf{z}_i'\boldsymbol{\beta} + \varepsilon_i,$$

where the data are assumed to be well behaved and, save for the functional form, the assumptions of the classical model are met. The function $f(x_i)$ remains unspecified. As stated, estimation by least squares is not feasible until $f(x_i)$ is specified. Suppose the data were such that they consisted of pairs of observations (y_{j1}, y_{j2}), $j = 1, \ldots, n/2$, in which $x_{j1} = x_{j2}$ within every pair. If so, then estimation of $\boldsymbol{\beta}$ could be based on the simple transformed model

$$y_{j2} - y_{j1} = (\mathbf{z}_{j2} - \mathbf{z}_{j1})'\boldsymbol{\beta} + (\varepsilon_{j2} - \varepsilon_{j1}), \quad j = 1, \ldots, n/2.$$

As long as observations are independent, the constructed disturbances, v_i still have zero mean, variance now $2\sigma^2$, and remain uncorrelated across pairs, so a classical model applies and least squares is actually optimal. Indeed, with the estimate of $\boldsymbol{\beta}$, say, $\hat{\boldsymbol{\beta}}_d$ in hand, a noisy estimate of $f(x_i)$ could be estimated with $y_i - \mathbf{z}_i'\hat{\boldsymbol{\beta}}_d$ (the estimate contains the estimation error as well as ε_i).[5]

The problem, of course, is that the enabling assumption is heroic. Data would not behave in that fashion unless they were generated experimentally. The logic of the partially linear regression estimator is based on this observation nonetheless. Suppose that the observations are sorted so that $x_1 < x_2 < \cdots < x_n$. Suppose, as well, that this variable is well behaved in the sense that as the sample size increases, this sorted data vector more tightly and uniformly fills the space within which x_i is assumed to vary. Then, intuitively, the difference is "almost" right, and becomes better as the sample size

[5] See Estes and Honoré (1995) who suggest this approach ~~(with simple~~ differencing of the data).

7-38

7.3.2 QUANTILE REGRESSION MODELS

The quantile regression model is

$$Q[y|\mathbf{x},q] = \mathbf{x}'\boldsymbol{\beta}_q \text{ such that } \text{Prob}[y \le \mathbf{x}'\boldsymbol{\beta}_q \,|\, \mathbf{x}] = q,\ 0 < q < 1.$$

This is essentially a nonparametric specification. No assumption is made about the distribution of $y|\mathbf{x}$ or about its conditional variance. The fact that q can vary continuously (strictly) between zero and one means that there are an infinite number of possible "parameter vectors." It seems reasonable to view the coefficients, which we might write $\boldsymbol{\beta}(q)$ less as fixed "parameters," as we do in the linear regression model, than loosely as *features* of the distribution of $y|\mathbf{x}$. For example, it is not likely to be meaningful to view $\boldsymbol{\beta}(.49)$ to be discretely different from $\boldsymbol{\beta}(.50)$ or to compute precisely a particular difference such as $\boldsymbol{\beta}(.5) - \boldsymbol{\beta}(.3)$. On the other hand, the qualitative difference, or possibly the lack of a difference, between $\boldsymbol{\beta}(.3)$ and $\boldsymbol{\beta}(.5)$ as displayed in our following example ~~below~~, may well be an interesting characteristic of the sample.

note double quotes

minus

The estimator, $\mathbf{b}_q$ of $\boldsymbol{\beta}_q$ for a specific quantile is computed by minimizing the function

$$F_n(\boldsymbol{\beta}_q \,|\, \mathbf{y},\mathbf{X}) = \sum_{i:y_i \ge \mathbf{x}_i'\boldsymbol{\beta}_q}^{n} q\,|\,y_i - \mathbf{x}_i'\boldsymbol{\beta}_q\,| + \sum_{i:y_i < \mathbf{x}_i'\boldsymbol{\beta}_q}^{n} (1-q)\,|\,y_i - \mathbf{x}_i'\boldsymbol{\beta}_q\,|$$

$$= \sum_{i=1}^{n} g\left(y_i - \mathbf{x}_i'\boldsymbol{\beta}_q \,|\, q\right)$$

where

$$g(e_{i,q} \,|\, q) = \begin{cases} q e_{i,q} & \text{if } e_{i,q} \ge 0 \\ (1-q)e_{i,q} & \text{if } e_{i,q} < 0 \end{cases},\quad e_{i,q} = y_i - \mathbf{x}_i'\boldsymbol{\beta}_q.$$

When $q = 0.5$, the estimator is the least absolute deviations estimator we examined in Example 4.5 and Section 7.3.1. Solving the minimization problem requires an iterative estimator. It can be set up as a linear programming problem.* [See Keonker and D'Oray (1987).]

FN 13 13

We cannot use the methods of Chapter 4 to determine the asymptotic covariance matrix of the estimator. But, the fact that the estimator is obtained by minimizing a sum does lead to a set of results similar to those we obtained in Section 4.4 for least squares. [See Buchinsky (1998).] Assuming that the regressors are "well behaved," the quantile regression estimator of $\boldsymbol{\beta}_q$ is consistent and asymptotically normally distributed with asymptotic covariance matrix

Ao: Is "Asy.Var." ok italics?

$$Asy.Var.[b_q] = \frac{1}{n}\mathbf{H}^{-1}\mathbf{G}\mathbf{H}^{-1}$$

where

$$\mathbf{H} = \text{plim}\frac{1}{n}\sum_{i=1}^{n} f_q(0\,|\,\mathbf{x}_i)\mathbf{x}_i\mathbf{x}_i'$$

and

$$\mathbf{G} = \text{plim}\frac{q(1-q)}{n}\sum_{i=1}^{n} \mathbf{x}_i\mathbf{x}_i'.$$

13 *Quantile regression is supported as a built in procedure in contemporary software such as Statas, SAS and NLOGIT.

This is the result we had earlier for the LAD estimator, now with quantile q instead of 0.5. As before, computation is complicated by the need to compute the density of ε_q at zero. This will require either an approximation of uncertain quality or a specification of the particular density, which we have hoped to avoid. The usual approach, as before, is to use bootstrapping.

Example 7.~~12~~ 10 Income Elasticity of Credit Card Expenditure

Greene (1992, 2007) analyzed the default behavior and monthly expenditure behavior of a large sample (13,444 observations) of credit card users. Among the results of interest in the study was an estimate of the income elasticity of the monthly expenditure. A conventional regression approach might be based on

$$Q[\ln \mathit{Spending}|\mathbf{x},q] \;=\; \beta_{1,q} + \beta_{2,q} \ln \mathit{Income} + \beta_{3,q}\, \mathit{Age} + \beta_{4,q}\, \mathit{Dependents}$$

The data in Appendix Table F7.~~4~~ 3 contain these and numerous other covariates that might explain spending; we have chosen these three for this example only. The 13,444 observations in the data set are based on credit card applications. Of the full sample, 10,499 applications were approved and the next 12 months of spending and default behavior were observed.[14] Spending is the average monthly expenditure in the twelve months after the account was initiated. Average monthly income and number of household dependents are among the demographic data in the application. Table 7.5 presents least squares estimates of the coefficients of the conditional mean function as well as full results for several quantiles.[15] Standard errors are shown for the least squares and median (1 = 0.5) results. The results for the other quantiles are essentially the same. The least squares estimate of 1.08344 is slightly and significantly greater than one – the estimated standard error is 0.03212 so the t statistic is (1.08344-1)/.03212 = 2.60. This suggests an aspect of consumption behavior that might not be surprising. However, the very large amount of variation over the range of quantiles might not have been expected. We might guess that at the highest levels of spending for any income level, there is (comparably so) some saturation in the response of spending to changes in income.

FN 14 / TB 7.5 / FN 15 / minus

AU: Confirm Fig 7.7 is correct.

Figure 7.~~8~~ 6 displays the estimates of the income elasticity of expenditure for the range of quantiles from 0.1 to 0.9, with the least squares estimate which would correspond to the fixed value at all quantiles shown in the center of the figure. Confidence limits shown in the figure are based on the asymptotic normality of the estimator. They are computed as the estimated income elasticity plus and minus 1.96 times the estimated standard error. Figure 7.6 7 shows the implied quantile regressions for q = .1, .3, .5, .7 and .9. The relatively large increase from the .1 quantile to the .3 suggests some skewness in the spending distribution. In broad terms, the results do seem to be largely consistent with our earlier result of the quantiles largely being differentiated by shifts in the constant term, in spite of the seemingly large change in the coefficient on ln Income in the results.

FIG 7.6 / FIG 7.7

14 The expenditure data are taken from the credit card records while the income and demographic data are taken from the applications. While it might be tempting to use, e.g., Powell's (1986a,b) censored quantile regression estimator to accommodate this large cluster of zeros for the dependent variable, this approach would misspecify the model – the "zeros" represent nonexistent observations, not missing ones. A more detailed approach – the one used in the 1992 study – would model separately the presence or absence of the observation on spending, and then model spending conditionally on acceptance of the application. We will revisit this issue in Chapter 17 in the context of the sample selection model. The income data are censored at 100,000 and 220 of the observations have expenditures that are filled with $1 or less. We have not "cleaned" the data set for these aspects. The full 10,499 observations have been used as they are in the original data set.

for example

AU: OK to spell out "e.g." in text?

15 We would note, if (7-~~XXX~~33) is the statement of the model, then it does not follow that that the conditional mean function is a linear regression. That would be an additional assumption.

Figure 7.6 Estimates of Income Elasticity of Expenditure

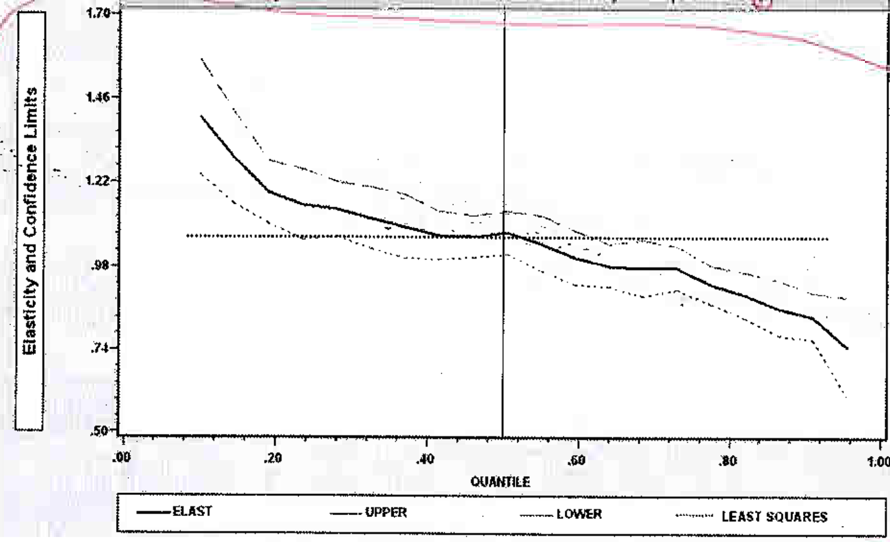
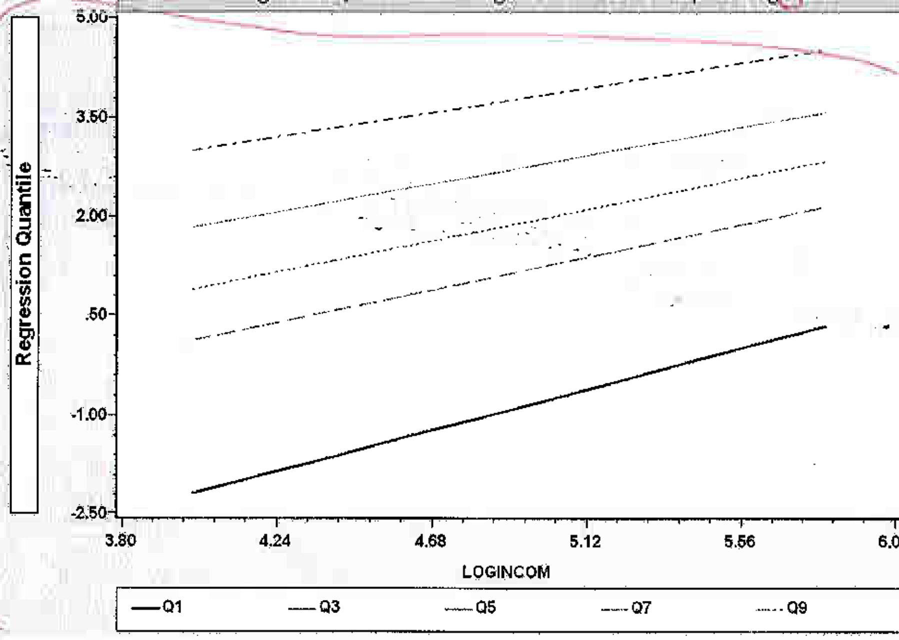


Table 7.5 Estimated Quantile Regression Models

Quantile	Constant	lnIncome	Age	Dependents
	Estimated Parameters			
0.1	-6.73560	1.40306	-.03081	-.04297
0.2	-4.31504	1.16919	-.02460	-.04630
0.3	-3.62455	1.12240	-.02133	-.04788
0.4	-2.98830	1.07109	-.01859	-.04731
(Median) 0.5	-2.80376	1.07493	-.01699	-.04995
Std.Error	(.24564)	(.03223)	(.00157)	(.01080)
t	-11.41	33.35	-10.79	-4.63
Least Squares	-3.05581	1.08344	-.01736	-.04461
Std.Error	(.23970)	(.03212)	(.00135)	(.01092)
t	-12.75	33.73	-12.88	-4.08
0.6	-2.05467	1.00302	-.01478	-.04609
0.7	-1.63875	.97101	-.01190	-.03803
0.8	-.94031	.91377	-.01126	-.02245
0.9	-.05218	.83936	-.00891	-.02009

Figure 7.6 Quantile Regressions for Ln Spending



7-42

AU: KT "partially linear model" already KT in chap. Here also?

TABLE 14.1 LS and LAD Estimates of a Production Function

	Least Squares			LAD				
					Bootstrap		Kernel Density	
Coefficient	Estimate	Standard Error	t Ratio	Estimate	Std. Error	t Ratio	Std. Error	t Ratio
Constant	2.293	0.107	21.396	2.275	0.202	11.246	0.183	12.374
β_k	0.279	0.081	3.458	0.261	0.124	2.099	0.138	1.881
β_l	0.927	0.098	9.431	0.927	0.121	7.637	0.169	5.498
Σe^2	0.7814			0.7984				
$\Sigma\|e\|$	3.3652			3.2541				

7.4 ~~14.3.3~~ PARTIALLY LINEAR REGRESSION

cumlant

The proper functional form in the linear regression is an important specification issue. We examined this in detail in Chapter 6. Some approaches, including the use of dummy variables, logs, quadratics, and so on, were considered as means of capturing nonlinearity. The translog model in particular (Example 2.4) is a well-known approach to approximating an unknown nonlinear function. Even with these approaches, the researcher might still be interested in relaxing the assumption of functional form in the model. The **partially linear model** [analyzed in detail by Yatchew (1998, 2000) and Härdle, Liang, and Gao (2000)] is another approach. Consider a regression model in which one variable, x, is of particular interest, and the functional form with respect to x is problematic. Write the model as

$$y_i = f(x_i) + \mathbf{z}_i'\boldsymbol{\beta} + \varepsilon_i,$$

where the data are assumed to be well behaved and, save for the functional form, the assumptions of the classical model are met. The function $f(x_i)$ remains unspecified. As stated, estimation by least squares is not feasible until $f(x_i)$ is specified. Suppose the data were such that they consisted of pairs of observations (y_{j1}, y_{j2}), $j = 1, \ldots, n/2$, in which $x_{j1} = x_{j2}$ within every pair. If so, then estimation of $\boldsymbol{\beta}$ could be based on the simple transformed model

$$y_{j2} - y_{j1} = (\mathbf{z}_{j2} - \mathbf{z}_{j1})'\boldsymbol{\beta} + (\varepsilon_{j2} - \varepsilon_{j1}), \quad j = 1, \ldots, n/2.$$

As long as observations are independent, the constructed disturbances, v_i still have zero mean, variance now $2\sigma^2$, and remain uncorrelated across pairs, so a classical model applies and least squares is actually optimal. Indeed, with the estimate of $\boldsymbol{\beta}$, say, $\hat{\boldsymbol{\beta}}_d$ in hand, a noisy estimate of $f(x_i)$ could be estimated with $y_i - \mathbf{z}_i'\hat{\boldsymbol{\beta}}_d$ (the estimate contains the estimation error as well as ε_i).[16]

FN 16

The problem, of course, is that the enabling assumption is heroic. Data would not behave in that fashion unless they were generated experimentally. The logic of the partially linear regression estimator is based on this observation nonetheless. Suppose that the observations are sorted so that $x_1 < x_2 < \cdots < x_n$. Suppose, as well, that this variable is well behaved in the sense that as the sample size increases, this sorted data vector more tightly and uniformly fills the space within which x_i is assumed to vary. Then, intuitively, the difference is "almost" right, and becomes better as the sample size

acute

16 ~~7~~ See Estes and Honoré (1995) who suggest this approach (with simple differencing of the data).

7-43

grows. [Yatchew (1997, 1998) goes more deeply into the underlying theory.] A theory is also developed for a better differencing of groups of two or more observations. The transformed observation is $y_{d,i} = \sum_{m=0}^{M} d_m y_{i-m}$ where $\sum_{m=0}^{M} d_m = 0$ and $\sum_{m=0}^{M} d_m^2 = 1$. (The data are not separated into nonoverlapping groups for this transformation—we merely used that device to motivate the technique.) The pair of weights for $M = 1$ is obviously $\pm\sqrt{0.5}$—this is just a scaling of the simple difference, 1, −1. Yatchew [1998, p. 697)] tabulates "optimal" differencing weights for $M = 1, \ldots, 10$. The values for $M = 2$ are (0.8090, −0.500, −0.3090) and for $M = 3$ are (0.8582, −0.3832, −0.2809, −0.1942). This estimator is shown to be consistent, asymptotically normally distributed, and have asymptotic covariance matrix[17]

FN 17

$$\text{Asy. Var}[\hat{\boldsymbol\beta}_d] = \left(1 + \frac{1}{2M}\right)\frac{\sigma_v^2}{n} E_x[\text{Var}[\mathbf{z} \mid x]].$$

The matrix can be estimated using the sums of squares and cross products of the differenced data. The residual variance is likewise computed with

$$\hat{\sigma}_v^2 = \frac{\sum_{i=M+1}^{n}(y_{d,i} - \mathbf{z}_{d,i}'\hat{\boldsymbol\beta}_d)^2}{n - M}.$$

Yatchew suggests that the partial residuals, $y_{d,i} - \mathbf{z}_{d,i}'\hat{\boldsymbol\beta}_d$ be smoothed with a kernel density estimator to provide an improved estimator of $f(x_i)$.

Example ~~14.5~~ 7.11 Partially Linear Translog Cost Function

Yatchew (1998, 2000) applied this technique to an analysis of scale effects in the costs of electricity supply. The cost function, following Nerlove (1963) and Christensen and Greene (1976) was specified to be a translog model (see Example 2.4 and Section ~~10.4.2~~ 10.5.2) involving labor and capital input prices, other characteristics of the utility and the variable of interest, the number of customers in the system, C. We will carry out a similar analysis using Christensen and Greene's 1970 electricity supply data. The data are given in Appendix Table ~~F4.3~~ F4.4. (See Section ~~10.4.1~~ 10.5.1 for description of the data.) There are 158 observations in the data set, but the last 35 are holding companies which are comprised of combinations of the others. In addition, there are several extremely small New England utilities whose costs are clearly unrepresentative of the best practice in the industry. We have done the analysis using firms 6–123 in the data set. Variables in the data set include Q = output, C = total cost, and PK, PL, and PF = unit cost measures for capital, labor, and fuel, respectively. The parametric model specified is a restricted version of the Christensen and Greene model,

$$\ln c = \beta_1 k + \beta_2 l + \beta_3 q + \beta_4(q^2/2) + \beta_5 + \varepsilon.$$

where $c = \ln[C/(Q \times PF)]$, $k = \ln(PK/PF)$, $l = \ln(PL/PF)$, and $q = \ln Q$. The partially linear model substitutes $f(Q)$ for the last three terms. The division by PF ensures that average cost is homogeneous of degree one in the prices, a theoretical necessity. The estimated equations, with estimated standard errors, are shown here.

-7.42   061   260   567   057

(parametric) $c = -\cancel{6.83} + 0.\cancel{168}k + 0.\cancel{140}l - 0.\cancel{590}q + 0.\cancel{061}q^2/2 + \varepsilon,$

$(0.\cancel{353})$ $(0.\cancel{042})$ $(0.\cancel{048})$ $(0.\cancel{075})$ $(0.\cancel{010})$ $s = 0.\cancel{10383}$

341   065   070   040   005   13826

(partially linear) $c_d = \cancel{0.170}k_d + \cancel{0.127}l_d + f(q) + v$

$\cancel{(0.049)}$ $\cancel{(0.057)}$ $s = \cancel{0.14044}$

0.16478

[17] Yatchew (2000, p. 191) denotes this covariance matrix $E[\text{Cov}[\mathbf{z} \mid x]]$.

0.108 0.170

(0.080) (0.083)

Example 7.11 Partially Linear Translog Cost Function

Yatchew (1998, 2000) applied this technique to an analysis of scale effects in the costs of electricity supply. The cost function, following Nerlove (1963) and Christensen and Greene (1976), was specified to be a translog model (see Example 2.4 and Section 10.5.2) involving labor and capital input prices, other characteristics of the utility, and the variable of interest, the number of customers in the system, *C*. We will carry out a similar analysis using Christensen and Greene's 1970 electricity supply data. The data are given in Appendix Table F4.4. (See Section 10.5.1 for description of the data.) There are 158 observations in the data set, but the last 35 are holding companies which are comprised of combinations of the others. In addition, there are several extremely small New England utilities whose costs are clearly unrepresentative of the best practice in the industry. We have done the analysis using firms 6–123 in the data set. Variables in the data set include *Q*=output, *C* =total cost, and *PK*, *PL*, and *PF* =unit cost measures for capital, labor, and fuel, respectively. The parametric model specified is a restricted version of the Christensen and Greene model,

$$\ln c = \beta_1 k + \beta_2 l + \beta_3 q + \beta_4 (q^2/2) + \beta_5 + \varepsilon,$$

where $c = \ln[C/(Q \times PF)]$, $k = \ln(PK/PF)$, $l = \ln(PL/PF)$, and $q = \ln Q$. The partially linear model substitutes $f(Q)$ for the last three terms. The division by *PF* ensures that average cost is homogeneous of degree one in the prices, a theoretical necessity. The estimated equations, with estimated standard errors, are shown here.

(parametric) $c = -7.32 + 0.069k + 0.241l - 0.569q + 0.057q^2/2 + \varepsilon,$
$\qquad (0.333) \quad (0.065) \quad (0.069) \quad (0.042) \quad (0.006) \qquad s = 0.13949$

(partially linear) $c_d = 0.108k_d + 0.163l_d + f(q) + v$
$\qquad (0.076) \quad (0.081) \qquad s = 0.16529$

7-45

TABLE ~~14.2~~ Kernels for Density Estimation

Kernel	*Formula K[z]*
Epanechnikov	$0.75(1 - 0.2z^2)/2.236$ if $\lvert z\rvert \le 5$, 0 else
Normal	$\phi(z)$ (normal density),
Logit	$\Lambda(z)[1 - \Lambda(z)]$ (logistic density)
Uniform	0.5 if $\lvert z\rvert \le 1$, 0 else
Beta	$0.75(1 - z)(1 + z)$ if $\lvert z\rvert \le 1$, 0 else
Cosine	$1 + \cos(2\pi z)$ if $\lvert z\rvert \le 0.5$, 0 else
Triangle	$1 - \lvert z\rvert$, if $\lvert z\rvert \le 1$, 0 else
Parzen	$4/3 - 8z^2 + 8\lvert z\rvert^3$ if $\lvert z\rvert \le 0.5$, $8(1 - \lvert z\rvert)^3/3$ if $0.5 < \lvert z\rvert \le 1$, 0 else.

candidates have been suggested, including the (long) list in Table 14.2. Each of these is smooth, continuous, symmetric, and equally attractive. The logit and normal kernels are defined so that the weight only asymptotically falls to zero whereas the others fall to zero at specific points. It has been observed that in constructing a density estimator, the choice of kernel function is rarely crucial, and is usually minor in importance compared to the more difficult problem of choosing the bandwidth. (The logit and normal kernels appear to be the default choice in many applications.)

The kernel density function is an estimator. For any specific x, $\hat{f}(x)$ is a sample statistic,

$$\hat{f}(z) = \frac{1}{n}\sum_{i=1}^{n} g(x_i \mid z, h).$$

Because $g(x_i \mid z, h)$ is nonlinear, we should expect a bias in a finite sample. It is tempting to apply our usual results for sample moments, but the analysis is more complicated because the bandwidth is a function of n. Pagan and Ullah (1999) have examined the properties of kernel estimators in detail and found that under certain assumptions, the estimator is consistent and asymptotically normally distributed but biased in finite samples. The bias is a function of the bandwidth but for an appropriate choice of h, the bias does vanish asymptotically. As intuition might suggest, the larger is the bandwidth, the greater is the bias, but at the same time, the smaller is the variance. This might suggest a search for an optimal bandwidth. After a lengthy analysis of the subject, however, the authors' conclusion provides little guidance for finding one. One consideration does seem useful. For the proportion of observations captured in the bin to converge to the corresponding area under the density, the width itself must shrink more slowly than $1/n$. Common applications typically use a **bandwidth** equal to some multiple of $n^{-1/5}$ for this reason. Thus, the one we used earlier is $h = 0.9 \times s/n^{1/5}$. To conclude the illustration begun earlier, Figure 14.5 is a logit-based kernel density estimator for the distribution of slope estimates for the model estimated earlier. The resemblance to the histogram in Figure 14.4 is to be expected.

7.5 ~~14.4.2~~ NONPARAMETRIC REGRESSION

The regression function of a variable y on a single variable x is specified as

$$y = \mu(x) + \varepsilon.$$

No assumptions about distribution, homoscedasticity, serial correlation or, most importantly, functional form are made at the outset; $\mu(x)$ may be quite nonlinear. Because this is the conditional mean, the only substantive restriction would be that deviations

7-46

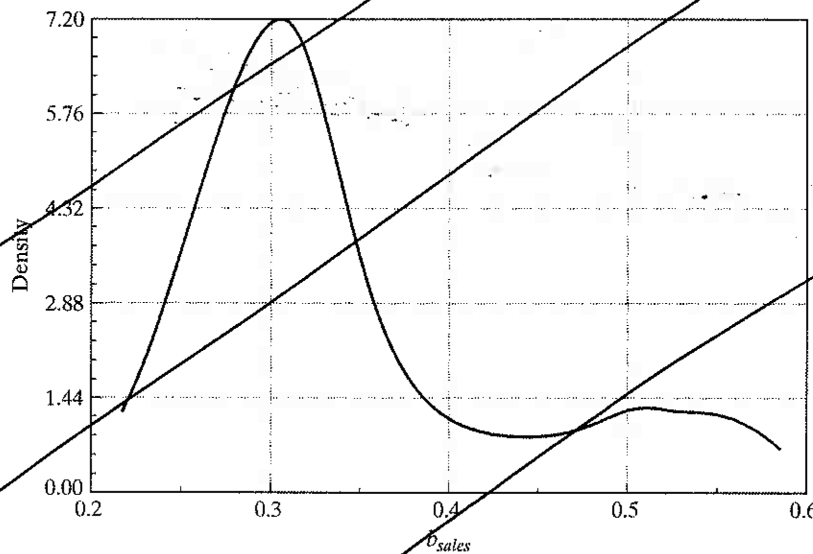


FIGURE 14.5 Kernel Density for b_{sales} Coefficients.

from the conditional mean function are not a function of (correlated with) x. We have already considered several possible strategies for allowing the conditional mean to be nonlinear, including spline functions, polynomials, logs, dummy variables, and so on. But, each of these is a "global" specification. The functional form is still the same for all values of x. Here, we are interested in methods that do not assume any particular functional form.

The simplest case to analyze would be one in which several (different) observations on y_i were made with each specific value of x_i. Then, the conditional mean function could be estimated naturally using the simple group means. The approach has two shortcomings, however. Simply connecting the points of means, $(x_i, \bar{y} \mid x_i)$ does not produce a smooth function. The method would still be assuming something specific about the function between the points, which we seek to avoid. Second, this sort of data arrangement is unlikely to arise except in an experimental situation. Given that data are not likely to be grouped, another possibility is a piecewise regression in which we define "neighborhoods" of points around each x of interest and ~~fit a separate~~ linear or quadratic regression in each neighborhood. This ~~returns~~ us to the problem of continuity that we noted earlier, but the method of splines is actually designed specifically for this purpose. Still, unless the number of neighborhoods is quite large, such a function is still likely to be crude.

discussed in Section 6.3.1

Smoothing techniques are designed to allow construction of an estimator of the conditional mean function without making strong assumptions about the behavior of the function between the points. They retain the usefulness of the **nearest neighbor** concept, but use more elaborate schemes to produce smooth, well behaved functions. The general class may be defined by a conditional mean estimating function

KT

$$\hat{\mu}(x^*) = \sum_{i=1}^{n} w_i(x^* \mid x_1, x_2, \ldots, x_n) y_i = \sum_{i=1}^{n} w_i(x^* \mid \mathbf{x}) y_i,$$

where the weights sum to 1. The linear least squares regression line is such an estimator. The predictor is

$$\hat{\mu}(x^*) = a + bx^*.$$

where a and b are the least squares constant and slope. For this function, you can show that

$$w_i(x^* \mid \mathbf{x}) = \frac{1}{n} + \frac{x^*(x_i - \bar{x})}{\Sigma_{i=1}^{n}(x_i - \bar{x})^2}.$$

The problem with this particular weighting function, which we seek to avoid here, is that it allows every x_i to be in the neighborhood of x^*, but it does not reduce the weight of any x_i when it is far from x^*. A number of **smoothing functions** have been suggested that are designed to produce a better behaved regression function. [See Cleveland (1979) and Schimek (2000).] We will consider two.

The locally weighted smoothed regression estimator ("loess" or "lowess" depending on your source) is based on explicitly defining a neighborhood of points that is close to x^*. This requires the choice of a **bandwidth**, h. The **neighborhood** is the set of points for which $|x^* - x_i|$ is small. For example, the set of points that are within the range $x^* \pm h/2$ might constitute the neighborhood. The choice of bandwith is crucial, as we will explore in the example below, and is also a challenge. There is no single best choice. A common choice is **Silverman's** (1986) **rule of thumb**,

$$h_{Silverman} = \frac{.9[\min(s, IQR)]}{1.349\ n^{0.2}}$$

where s is the sample standard deviation and IQR is the interquartile range (.75 quantile minus .25 quantile). A suitable weight is then required. Cleveland (1979) recommends the tricube weight,

$$T_i(x^* \mid \mathbf{x}, h) = \left[1 - \left(\frac{|x_i - x^*|}{h}\right)^3\right]^3.$$

Combining terms, then the weight for the loess smoother is

$$w_i(x^*|\mathbf{x},h) = 1(x_i \text{ in the neighborhood}) \times T_i(x^*|\mathbf{x},h).$$

The bandwidth is is essential in the results. A wider neighborhood will produce a smoother function. But the wider neighborhood will track the data less closely than a narrower one. A second possibility, similar to the least squares approach, is to allow the neighborhood to be all points, but make the weighting function decline smoothly with the distance between x^* and any x_i. A variety of **kernel functions** are ussed for this purpose. Two common choices are the **logistic kernel**,

$$K(x^*|x_i,h) \;=\; \Lambda(v_i)[1\text{-}\Lambda(v_i)] \text{ where } \Lambda(v_i) = \exp(v_i)/[1+\exp(v_i)],\ v_i = (x_i - x^*)/h.$$

and the Epanechnikov kernel,

$$K(x^*|x_i,h) \;=\; 0.75\ (1 - 0.2\ v_i^2)/\sqrt{5} \text{ if } |v_i| \le 5 \text{ and } 0 \text{ otherwise.}$$

KT

following

KT

minus

AU: KT "bandwidth" already KT in chap. Here also?

AU: KTs "neighborhood" & "Silverman's rule of thumb" not in chap. list.

AU: Is subscript "Silverman" ok italics?

AU: These 2 KTs not in chap. list.

This produces the kernel weighted regression estimator,

$$\hat{\mu}(x^* \mid \mathbf{x}, h) = \frac{\sum_{i=1}^{n} \frac{1}{k} K\left[\frac{x_i - x^*}{h}\right] y_i}{\sum_{i=1}^{n} \frac{1}{k} K\left[\frac{x_i - x^*}{h}\right]},$$

which has become a standard tool in nonparametric analysis.

Example 7.12 A Nonparametric Average Cost Function

FIGS 7.8 7.9

In Example 7.11, we fit a partially linear regression for the relationship between average cost and output for electricity supply. Figures 7.8 and 7.9 show the less ambitious nonparametric regressions of average cost on output. The overall picture is the same as in the earlier example. The kernel function is the logistic density in both cases. The function in Figure 7.8 uses a bandwidth of 2,000. Because this is a fairly large proportion of the range of variation of output, the function is quite smooth. The regression in Figure 7.0 uses a bandwidth of only 200. The function tracks the data better, but at an obvious cost. The example demonstrates what we and others have noted often; the choice of bandwidth in this exercise is crucial.

7.9

7-49

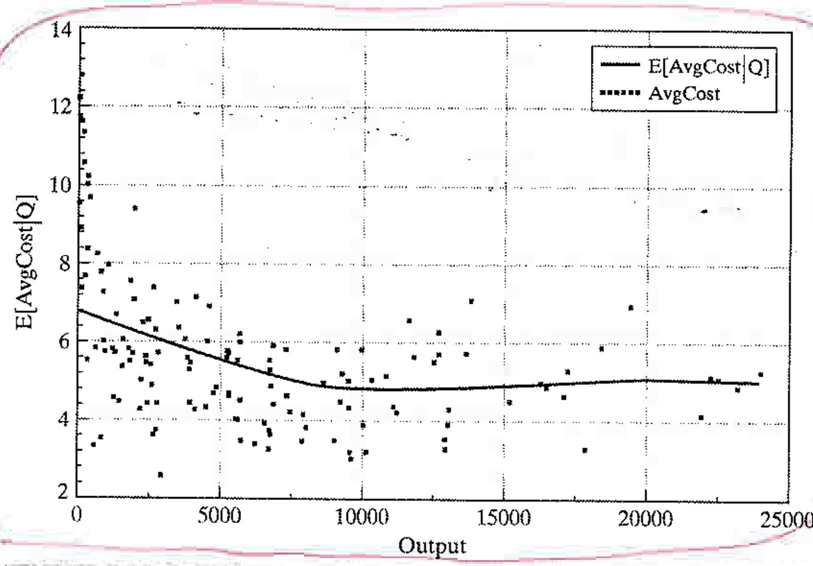
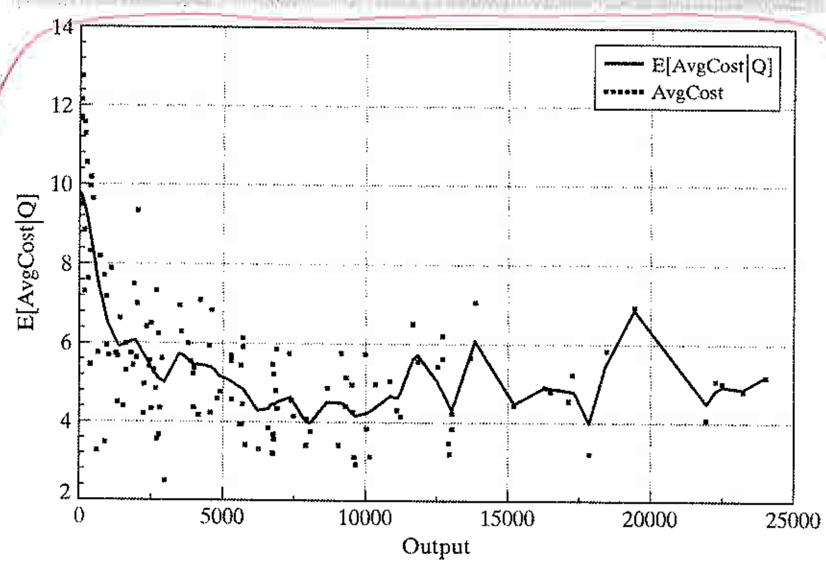


FIGURE ~~14.6~~ 7.8 Nonparametric Cost Function.

FIGURE ~~14.7~~ 7.9 Nonparametric Cost Function.



7-50

Data smoothing is essentially data driven. As with most nonparametric techniques, inference is not part of the analysis—this body of results is largely descriptive. As can be seen in the example, nonparametric regression can reveal interesting characteristics of the data set. For the econometrician, however, there are a few drawbacks. There is no danger of misspecifying the conditional mean function, for example. But, the great generality of the approach limits the ability to test one's specification or the underlying theory. [See, for example, Blundell, Browning, and Crawford's (2003) extensive study of British expenditure patterns.] Most relationships are more complicated than a simple conditional mean of one variable. In the example just given, some of the variation in average cost relates to differences in factor prices (particularly fuel) and in load factors. Extensions of the fully nonparametric regression to more than one variable is feasible, but very cumbersome. [See Härdle (1990) and Li and Racine (2007).] A promising approach is the partially linear model considered earlier.

umlaut

7.125

14.5 PROPERTIES OF ESTIMATORS

The preceding has been concerned with methods of estimation. We have surveyed a variety of techniques that have appeared in the applied literature. We have not yet examined the statistical properties of these estimators. Although, as noted earlier, we will leave extensive analysis of the asymptotic theory for more advanced treatments, it is appropriate to spend at least some time on the fundamental theoretical platform which underlies these techniques.

14.5.1 STATISTICAL PROPERTIES OF ESTIMATORS

Properties that we have considered are as follows:

- **Unbiasedness:** This is a finite sample property that can be established in only a very small number of cases. Strict unbiasedness is rarely of central importance outside the linear regression model. However, "asymptotic unbiasedness" (whereby the expectation of an estimator converges to the true parameter as the sample size grows), might be of interest. [See, e.g., Pagan and Ullah (1999, Section 2.5.1 on the subject of the kernel density estimator).] In most cases, however, discussions of asymptotic unbiasedness are actually directed toward consistency, which is a more desirable property.
- **Consistency:** This is a much more important property. Econometricians are rarely willing to place much credence in an estimator for which consistency cannot be established.
- **Asymptotic normality:** This property forms the platform for most of the statistical inference that is done with common estimators. When asymptotic normality cannot be established, it sometimes becomes difficult to find a method of progressing beyond simple presentation of the numerical values of estimates (with caveats). However, most of the contemporary literature in macroeconomics and time-series analysis is strongly focused on estimators that are decidedly not asymptotically normally distributed. The implication is that this property takes its importance only in context, not as an absolute virtue.
- **Asymptotic efficiency:** Efficiency can rarely be established in absolute terms. Efficiency within a class often can, however. Thus, for example, a great deal can

7-51

effects) that we observe in the linear model only carries over to a very few nonlinear models and, unfortunately, this is not one of them.

An approach that can be used, albeit at the cost of an additional assumption, is the simulation based estimator in Section 9.8.2. If we assume that u_i is normally distributed with mean zero and variance σ_u^2, then an analog to (9-57) for least squares would be

$$S_n^S = \frac{1}{2}\sum_{i=1}^{n}\frac{1}{R}\sum_{r=1}^{R}\sum_{t=1}^{T_i}[y_{it} - h(\mathbf{x}_{it}, \boldsymbol{\beta}, \sigma_u v_{ir})]^2 \tag{11-37}$$

The approach from this point would be the same as in Section 9.8.2. [If it is further assumed that ε_{it} is normally distributed, then after incorporating σ_ε^2 in the criterion function, (11-37) would actually be the extension of (9-57) to a nonlinear regression function. The random parameter vector there is specialized here to a nonrandom constant term.]

7.6 ~~11.8~~ SUMMARY AND CONCLUSIONS

In this chapter, we extended the regression model to a form that allows nonlinearity in the parameters in the regression function. The results for interpretation, estimation, and hypothesis testing are quite similar to those for the linear model. The two crucial differences between the two models are, first, the more involved estimation procedures needed for the nonlinear model and, second, the ambiguity of the interpretation of the coefficients in the nonlinear model (because the derivatives of the regression are often nonconstant, in contrast to those in the linear model). ~~Finally, we added an additional level of generality to the model. Two-step nonlinear least squares is suggested as a method of allowing a model to be fit while including functions of previously estimated parameters.~~

Key Terms and Concepts

Insert terms on msp 7-52 where indicated

(1)(2)
- Box–Cox transformation
- Delta method
- GMM estimator
- Identification condition
- Incidental parameters problem
- Index function model
- Indirect utility function
- Iteration
- Jacobian
- (3) Linearized regression model
- Lagrange multiplier test
- Logit model
- (5) Loglinear model
- Nonlinear regression model
- Normalization
- (4) Nonlinear least squares
- Orthogonality condition
- (6) Overidentifying restrictions
- Pseudoregressors
- (7) Roy's identity
- Semiparametric
- (8) Starting values
- Two-step estimation
- Wald test

Exercises

1. Describe how to obtain nonlinear least squares estimates of the parameters of the model $y = \alpha x^\beta + \varepsilon$.
2. Verify the following differential equation, which applies to the Box–Cox transformation:

$$\frac{d^i x^{(\lambda)}}{d\lambda^i} = \left(\frac{1}{\lambda}\right)\left[x^\lambda (\ln x)^i - \frac{i d^{i-1}x^{(\lambda)}}{d\lambda^{i-1}}\right]. \tag{7-34 ~~11-38~~}$$

Insert terms on ms p. 7-51 where indicated. Do not set numbers

7-52

(1) • Bandwidth
(2) • Bootstrap
(3) • Kernel density estimator
(4) • Least absolute deviations (LAD)
(5) • Nearest neighbor
• Nonparametric estimators
(6) • Partially linear model
(7) • Quantile regression
• Semiparametric estimation
(8) • Smoothing function
• ~~Lagrange multiplier test~~

AU: The following terms were not bold KTs in text:
~~Incidental parameters problem~~
~~Index function model~~
~~Logit model~~
~~Normalization~~
~~Two-step estimation~~
~~Wald test~~
~~Nonparametric estimators~~
~~Semiparametric estimators~~

7-53

Show that the limiting sequence for $\lambda = 0$ is

$$\lim_{\lambda \to 0} \frac{d^i x^{(\lambda)}}{d\lambda^i} = \frac{(\ln x)^{i+1}}{i+1}. \qquad \text{7-35 } (\cancel{11\text{-}39})$$

These results can be used to great advantage in deriving the actual second derivatives of the log-likelihood function for the Box–Cox model.

Applications

1. The data in Appendix table F5.~~1~~ 3 present 27 statewide observations on value added (output), labor input (labor), and capital stock (capital) for SIC 33 (primary metals). We are interested in determining whether a linear or loglinear production model is more appropriate for these data. Use MacKinnon, White, and Davidson's (1983) P_E test to determine whether a linear or log-linear production model is preferred.
2. Using the Box–Cox transformation, we may specify an alternative to the Cobb–Douglas model as

$$\ln Y = \alpha + \beta_k \frac{(K^\lambda - 1)}{\lambda} + \beta_l \frac{(L^\lambda - 1)}{\lambda} + \varepsilon.$$

   Using Zellner and Revankar's data in Appendix Table F14.1, F7.5 estimate α, β_k, β_l, and λ by using the scanning method suggested in Section 11.3.2. (Do not forget to scale Y, K, and L by the number of establishments.) Use (11-16), (11-12), and (11-13) to compute the appropriate asymptotic standard errors for your estimates. Compute the two output elasticities, $\partial \ln Y/\partial \ln K$ and $\partial \ln Y/\partial \ln L$, at the sample means of K and L. (Hint: $\partial \ln Y/\partial \ln K = K\partial \ln Y/\partial K$.)

   AU: Provide correct x-ref. equation numbers – 5x
3. For the model in Application 2, test the hypothesis that $\lambda = 0$ using a Wald test and a Lagrange multiplier test. Note that the restricted model is the Cobb–Douglas log-linear model. The LM test statistic is shown in (11-22). To carry out the test, you will need to compute the elements of the fourth column of $\mathbf{X}^0$, the pseudoregressor corresponding to λ is $\partial E[y \mid x]/\partial\lambda \mid \lambda = 0$. Result (11-39) will be useful.
4. The National Institute of Standards and Technology (NIST) has created a Web site that contains a variety of estimation problems, with data sets, designed to test the accuracy of computer programs. (The URL is http://www.itl.nist.gov/div898/strd/.) One of the five suites of test problems is a set of 27 nonlinear least squares problems, divided into three groups: easy, moderate, and difficult. We have chosen one of them for this application. You might wish to try the others (perhaps to see if the software you are using can solve the problems). This is the Misralc problem (http://www.itl.nist.gov/div898/strd/nls/data/misralc.shtml). The nonlinear regression model is

   URL URL

$$\begin{aligned} y_i &= h(x, \beta) + \varepsilon \\ &= \beta_1 \left(1 - \frac{1}{\sqrt{1 + 2\beta_2 x_i}}\right) + \varepsilon_i \end{aligned}$$

7-54

The data are as follows:

Y	X
10.07	77.6
14.73	114.9
17.94	141.1
23.93	190.8
29.61	239.9
35.18	289.0
40.02	332.8
44.82	378.4
50.76	434.8
55.05	477.3
61.01	536.8
66.40	593.1
75.47	689.1
81.78	760.0

For each problem posed, NIST also provides the "certified solution," (i.e., the right answer). For the Misralc problem, the solutions are as follows:

	Estimate	*Estimated Standard Error*
β_1	6.3642725809E + 02	4.6638326572E + 00
β_2	2.0813627256E − 04	1.7728423155E − 06
$\mathbf{e}'\mathbf{e}$	4.0966836971E − 02	
$s^2 = \mathbf{e}'\mathbf{e}/(n - K)$	5.8428615257E − 02	

Finally, NIST provides two sets of starting values for the iterations, generally one set that is "far" from the solution and a second that is "close" from the solution. For this problem, the starting values provided are $\boldsymbol{\beta}^1 = (500, 0.0001)$ and $\boldsymbol{\beta}^2 = (600, 0.0002)$. The exercise here is to reproduce the NIST results with your software. [For a detailed analysis of the NIST nonlinear least squares benchmarks with several well-known computer programs, see McCullough (1999).]

7-55/End 7

5. In Example 7.1, the CES function is suggested as a model for production;

$$\ln y = \ln \gamma - \frac{\nu}{\rho}\ln\left[\delta K^{-\rho} + (1-\delta)L^{-\rho}\right] + \varepsilon. \qquad \textbf{(7-36)}$$

Example 6.8 suggested an indirect method of estimating the parameters of this model. The function is linearized around $\rho = 0$, which produces an intrinsically linear approximation to the function,

$$\ln y = \beta_1 + \beta_2 \ln K + \beta_3 \ln L + \beta_4 [1/2 (\ln K - \mathrm{Ln} L)^2] + \varepsilon$$

minus minus

where $\beta_1 = \ln\gamma$, $\beta_2 = \nu\delta$. $\beta_3 = \nu(1 - \delta)$ and $\beta_4 = \rho\nu\delta(1 - \delta)$. The approximation can be estimated by linear least squares. Estimates of the structural parameters are found by inverting the four equations above. An estimator of the asymptotic covariance matrix is suggested using the delta method. The parameters of (7-36) can also be estimated directly using nonlinear least squares and the results given earlier in this Chapter.

preceding/

Christensen and Greene's (1976) data on U.S. electricity generation are given in Appendix Table F4.4. The data file contains 158 observations. Using the first 123, fit the CES production function, using capital and fuel as the two factors of production rather than capital and labor. Compare the results obtained by the two approaches, and comment on why the differences (which are substantial) arise.

The following exercises require specialized software. The relevant techniques are available in several packages that might be in use, such as SAS, Stata, or LIMDEP. The exercises are suggested as departure points for explorations using a few of the many estimation techniques listed in this chapter.

6. Using the gasoline market data in Appendix Table F2.2, use the partially linear regression method in Section 7.4 to fit an equation of the form

$$\ln(G/Pop) = \beta_1 \ln(Income) + \beta_2 \ln P_{new\ cars} + \beta_3 \ln P_{used\ cars} + g(\ln P_{gasoline}) + \varepsilon$$

AU: Are Subscripts OK italics?

7. To continue the analysis in Question 5, consider a nonparametric regression of G/Pop on the price. Using the nonparametric estimation method in Section 7.5, fit the nonparametric estimator using a range of bandwidth values to explore the effect of bandwidth.