6

FUNCTIONAL FORM, DIFFERENCE IN DIFFERENCES, AND STRUCTURAL CHANGE

6.1 INTRODUCTION

This chapter will examine a variety of ways that the linear regression model can be adapted for particular situations and specific features of the environment. Section 6.2 begins by using binary variables to accommodate nonlinearities and discrete shifts in the model. Sections 6.3 and 6.4 examine two specific forms of the linear model that are suited for analyzing causal impacts of policy changes, **difference in differences** models and **regression kink** and **regression discontinuity designs**. Section 6.5 broadens the class of models that are linear in the parameters. By using logarithms, quadratic terms, and **interaction terms** (products of variables), the regression model can accommodate a wide variety of functional forms in the data. Section 6.6 examines the issue of specifying and testing for discrete change in the underlying process that generates the data, under the heading of structural change. In a time-series context, this relates to abrupt changes in the economic environment, such as major events in financial markets (e.g., the world financial crisis of 2007–2008) or commodity markets (such as the several upheavals in the oil market). In a cross section, we can modify the regression model to account for discrete differences across groups such as different preference structures or market experiences of men and women.

6.2 USING BINARY VARIABLES

One of the most useful devices in regression analysis is the **binary**, or **dummy variable**. A dummy variable takes the value one for some observations to indicate the presence of an effect or membership in a group and zero for the remaining observations. Binary variables are a convenient means of building discrete shifts of the function into a regression model.

6.2.1 BINARY VARIABLES IN REGRESSION

Dummy variables are usually used in regression equations that also contain other quantitative variables,

$$y_i = \mathbf{x}_i'\boldsymbol{\beta} + \gamma d_i + \varepsilon_i, \qquad \textbf{(6-1)}$$

TABLE 6.1 Estimated Earnings Equation

$\ln earnings = \beta_1 + \beta_2 Age + \beta_3 Age^2 + \beta_4 Education + \beta_5 Kids + \varepsilon$

Sum of squared residuals:	599.4582
Standard error of the regression:	1.19044
R^2 based on 428 observations:	0.040995

Variable	*Coefficient*	*Standard Error*	*t Ratio*
Constant	3.24009	1.7674	1.833
Age	0.20056	0.08386	2.392
Age^2	−0.002315	0.000987	−2.345
Education	0.067472	0.025248	2.672
Kids	−0.35119	0.14753	−2.380

where $d_i = 1$ for some condition occurring, and 0 if not.[1] In the earnings equation in Example 5.2, we included a variable *Kids* to indicate whether there were children in the household, under the assumption that for many married women, this fact is a significant consideration in labor supply decisions. The results shown in Example 6.1 appear to be consistent with this hypothesis.

Example 6.1 Dummy Variable in an Earnings Equation

Table 6.1 reproduces the estimated earnings equation in Example 5.2. The variable *Kids* is a dummy variable that equals one if there are children under 18 in the household and zero otherwise. Because this is a semilog equation, the value of −0.35 for the coefficient is an extremely large effect, one which suggests that all other things equal, the earnings of women with children are nearly a third less than those without. This is a large difference, but one that would certainly merit closer scrutiny. Whether this effect results from different labor market effects that influence wages and not hours, or the reverse, remains to be seen. Second, having chosen a nonrandomly selected sample of those with only positive earnings to begin with, it is unclear whether the sampling mechanism has, itself, induced a bias in the estimator of this parameter.

Dummy variables are particularly useful in loglinear regressions. In a model of the form

$$\ln y = \beta_1 + \beta_2 x + \beta_3 d + \varepsilon,$$

the coefficient on the dummy variable, d, indicates a multiplicative shift of the function. The percentage change in $E[y|x, d]$ associated with the change in d is

$$\begin{aligned}
\%(\Delta E[y|x,d]/\Delta d) &= 100\%\left\{\frac{E[y|x,d=1] - E[y|x,d=0]}{E[y|x,d=0]}\right\} \\
&= 100\%\left\{\frac{\exp(\beta_1 + \beta_2 x + \beta_3)E[\exp(\varepsilon)] - \exp(\beta_1 + \beta_2 x)\,E[\exp(\varepsilon)]}{\exp(\beta_1 + \beta_2 x)E[\exp(\varepsilon)]}\right\} \\
&= 100\%[\exp(\beta_3) - 1]. \qquad \textbf{(6-2)}
\end{aligned}$$

[1] We are assuming at this point (and for the rest of this chapter) that the dummy variable in (6-1) is *exogenous*. That is, the assignment of values of the dummy variable to observations in the sample is unrelated to ε_i. This is consistent with the sort of random assignment to treatment designed in a clinical trial. The case in which d_i is endogenous would occur, for example, when individuals select the value of d_i themselves. Analyses of the effects of program participation, such as job training on wages or agricultural extensions on productivity, would be examples. The endogenous treatment effect model is examined in Section 8.5.

Example 6.2 Value of a Signature

In Example 4.10 we explored the relationship between log of sale price and surface area for 430 sales of Monet paintings. Regression results from the example are shown in Table 6.2. The results suggest a strong relationship between area and price—the coefficient is 1.33372, indicating a highly elastic relationship, and the *t* ratio of 14.70 suggests the relationship is highly significant. A variable (effect) that is clearly left out of the model is the effect of the artist's signature on the sale price. Of the 430 sales in the sample, 77 are for unsigned paintings. The results at the right of Table 6.2 include a dummy variable for whether the painting is signed or not. The results show an extremely strong effect. The regression results imply that

$$E[Price \mid Area, Aspect\ Ratio, Signature) =$$
$$\exp[-9.64 + 1.35 \ln Area - 0.0\,8\ Aspect\ Ratio + 1.23\ Signature + 0.993^2/2].$$

(See Section 4.8.2.) Computing this result for a painting of the same area and aspect ratio, we find the model predicts that the signature effect would be

$$100\% \times \frac{\Delta E[Price]}{Price} = 100\%[\exp(1.26) - 1] = 252\%.$$

The effect of a signature on an otherwise similar painting is to more than double the price. The estimated standard error for the signature coefficient is 0.1253. Using the delta method, we obtain an estimated standard error for $[\exp(b_3) - 1]$ of the square root of $[\exp(b_3)]^2 \times 0.1253^2$, which is 0.4417. For the percentage difference of 252%, we have an estimated standard error of 44.17%.

Superficially, it is possible that the size effect we observed earlier could be explained by the presence of the signature. If the artist tended on average to sign only the larger paintings, then we would have an explanation for the counterintuitive effect of size. (This would be an example of the effect of multicollinearity of a sort.) For a regression with a continuous variable and a dummy variable, we can easily confirm or refute this proposition. The average size for the 77 sales of unsigned paintings is 1,228.69 square inches. The average size of the other 353 is 940.812 square inches. There does seem to be a substantial systematic difference between signed and unsigned paintings, but it goes in the other direction. We are left with significant findings of both a size and a signature effect in the auction prices of Monet paintings. *Aspect Ratio*, however, appears still to be inconsequential.

TABLE 6.2 Estimated Equations for Log Price

$\ln price = \beta_1 + \beta_2 \ln Area + \beta_3 Aspect\ Ratio + \beta_4 Signature + \varepsilon$						
Mean of ln *Price*	0.33274					
Number of observations	430					
Sum of squared residuals	520.765			420.609		
Standard error	1.10435			1.35024		
R-squared	0.33417			0.46223		
Adjusted *R*-squared	0.33105			0.45844		
Variable	***Coefficient***	***Standard Error***	***t Ratio***	***Coefficient***	***Standard Error***	***t Ratio***
Constant	−8.34327	0.67820	−12.30	−9.65443	0.62397	−15.47
ln *Area*	1.31638	0.09205	14.30	1.34379	0.08787	16.22
Aspect ratio	−0.09623	0.15784	−0.61	−0.01966	0.14222	−0.14
Signature	—	—	—	1.26090	0.12519	10.07

Example 6.3 Gender and Time Effects in a Log Wage Equation

Cornwell and Rupert (1988) examined the returns to schooling in a panel data set of 595 heads of households observed in seven years, 1976-1982. The sample data (Appendix Table F8.1) are drawn from years 1976 to 1982 from the "Non-Survey of Economic Opportunity" from the Panel Study of Income Dynamics. A prominent result that appears in different specifications of their regression model is a persistent difference between wages of female and male heads of households. A slightly modified version of their regression model is

$$\ln Wage_{it} = \beta_1 + \beta_2 Exp_{it} + \beta_3 Exp_{it}^2 + \beta_4 Wks_{it} + \beta_5 Occ_{it} + \beta_6 Ind_{it} + \beta_7 South_{it} + \beta_8 SMSA_{it} + \beta_9 MS_{it} + \beta_{10} Union_{it} + \beta_{11} Ed_i + \beta_{12} Fem_i + \sum_{t=1977}^{1982} \gamma_t D_{it} + \varepsilon_{it}.$$

The variables in the model are listed in Example 4.6. (See Appendix Table F8.1 for the data source.)

Least squares estimates of the log wage equation appear at the left side in Table 6.3. Because these data are a panel, it is likely that observations within each group are correlated. The table reports cluster corrected standard errors, based on (4-42). The coefficient on

TABLE 6.3 Estimated Log Wage Equations

	Aggregate Effect			*Individual Fixed Effects*		
Sum of squares	391.056			81.5201		
Residual std. error	0.30708			0.15139		
R-squared	0.55908			0.90808		
Observations	4165			595 × 7		
F[17,577]	1828.50					
	Coefficient	***Clustered Std.Error***	***t Ratio***	***Coefficient***	***Clustered Std.Error***	***t Ratio***
Constant	5.08397	0.12998	39.11	Individual Fixed Effects		
EXP	0.03128	0.00419	7.47	0.10370	0.00691	15.00
EXP^2	−0.00055	0.00009	−5.86	−0.00040	0.00009	−4.43
WKS	0.00394	0.00158	2.50	0.00068	0.00095	0.72
OCC	−0.14116	0.02687	−5.25	−0.01916	0.02033	−0.94
IND	0.05661	0.02343	2.42	0.02076	0.02422	0.86
SOUTH	−0.07180	0.02632	−2.73	0.00309	0.09620	0.03
SMSA	0.15423	0.02349	6.57	−0.04188	0.03133	−1.34
MS	0.09634	0.04301	2.24	−0.02857	0.02887	−0.99
UNION	0.08052	0.02335	3.45	0.02952	0.02689	1.10
ED	0.05499	0.00556	9.88	—	—	—
FEM	−0.36502	0.04829	−7.56	—	—	—
Year(Base = 1976)						
1977	0.07461	0.00601	12.42	—	—	—
1978	0.19611	0.00989	19.82	0.04107	0.01267	3.24
1979	0.28358	0.01016	27.90	0.05170	0.01662	3.11
1980	0.36264	0.00985	36.82	0.05518	0.02132	2.59
1981	0.43695	0.01133	38.58	0.04612	0.02718	1.70
1982	0.52075	0.01211	43.00	0.04650	0.03254	1.43

FEM is −0.36502. Using (6-2), this translates to a roughly 100%[exp(−0.365) − 1] = 31% wage differential. Because the data are a panel, it is quite likely that the disturbances are correlated across the years within a household. Thus, robust standard errors are reported in Table 6.3. The effect of the adjustment is substantial. The conventional standard error for *FEM* based on $s^2(\mathbf{X}'\mathbf{X})^{-1}$ is 0.02201—less than half the reported value of 0.04829. Note the reported denominator degrees of freedom for the model *F* statistic is 595 − 18 = 577. Given that observations within a unit are not independent, it seems that 4147 would overstate the degrees of freedom. The number of groups of 595 is the natural alternative number of observations. However, if this were the case, then the statistic reported, computed as if there were 4165 observations, would not have an *F* distribution. This remains as an ambiguity in the computation of robust statistics. As we will pursue in Chapter 8, there is yet another ambiguity in this equation. It seems likely unobserved factors that influence ln *Wage* (in ε_{it}) (e.g., ability) might also be influential in the level of education. If so (i.e., if Ed_i is correlated with ε_{it}), then least squares might not be an appropriate method of estimation of the parameters in this model.

It is common for researchers to include a dummy variable in a regression to account for something that applies only to a single observation. For example, in time-series analyses, an occasional study includes a dummy variable that is one only in a single unusual year, such as the year of a major strike or a major policy event. (See, for example, the application to the German money demand function in Section 21.3.5.) It is easy to show (we consider this in the exercises) the very useful implication of this:

A dummy variable that takes the value one only for one observation has the effect of deleting that observation from computation of the least squares slopes and variance estimator (but not from R-squared).

6.2.2 SEVERAL CATEGORIES

When there are several categories, a set of binary variables is necessary. Correcting for seasonal factors in macroeconomic data is a common application. We could write a consumption function for quarterly data as

$$C_t = \beta_1 + \beta_2 x_t + \delta_1 D_{t1} + \delta_2 D_{t2} + \delta_3 D_{t3} + \varepsilon_t,$$

where x_t is disposable income. Note that only three of the four quarterly dummy variables are included in the model. If the fourth were included, then the four dummy variables would sum to one at every observation, which would replicate the constant term—a case of perfect multicollinearity. This is known as the **dummy variable trap**. To avoid the dummy variable trap, we drop the dummy variable for the fourth quarter. (Depending on the application, it might be preferable to have four separate dummy variables and drop the overall constant.[2]) Any of the four quarters (or 12 months) can be used as the base period.

The preceding is a means of *deseasonalizing* the data. Consider the alternative formulation:

$$C_t = \beta x_t + \delta_1 D_{t1} + \delta_2 D_{t2} + \delta_3 D_{t3} + \delta_4 D_{t4} + \varepsilon_t.$$

[2] See Suits (1984) and Greene and Seaks (1991).

Using the results from Section 3.3 on partitioned regression, we know that the preceding multiple regression is equivalent to first regressing C and x on the four dummy variables and then using the residuals from these regressions in the subsequent regression of deseasonalized consumption on deseasonalized income. Clearly, deseasonalizing in this fashion prior to computing the simple regression of consumption on income produces the same coefficient on income (and the same vector of residuals) as including the set of dummy variables in the regression.

Example 6.4 Genre Effects on Movie Box Office Receipts

Table 4.10 in Example 4.12 presents the results of the regression of log of box office receipts in 2009 for 62 movies on a number of variables including a set of dummy variables for four genres: *Action*, *Comedy*, *Animated*, or *Horror*. The left out category is "any of the remaining 9 genres" in the standard set of 13 that is usually used in models such as this one.[3] The four coefficients are −0.869, −0.016, −0.833, and +0.375, respectively. This suggests that, save for horror movies, these genres typically fare substantially worse at the box office than other types of movies. We note the use of b directly to estimate the percentage change for the category, as we did in Example 6.1 when we interpreted the coefficient of −0.35 on *Kids* as indicative of a 35% change in income. This is an approximation that works well when b is close to zero but deteriorates as it gets far from zero. Thus, the value of −0.869 above does not translate to an 87% difference between *Action* movies and other movies. Using (6-2), we find an estimated difference closer to 100% [exp(−0.869)−1] or about 58%. Likewise, the −0.35 result in Example 6.1 corresponds to an effect of about 29%.

6.2.3 MODELING INDIVIDUAL HETEROGENEITY

In the previous examples, a dummy variable is used to account for a specific event or feature of the observation or the environment, such as whether a painting is signed or not or the season. When the sample consists of repeated observations on a large number of entities, such as the 595 individuals in Example 6.3, a strategy often used to allow for unmeasured (and unnamed) fixed individual characteristics (effects) is to include a full set of dummy variables in the equation, one for each individual. To continue Example 6.3, the extended equation would be

$$\begin{aligned}\ln Wage_{it} = {} & \beta_1 + \sum_{i=1}^{595}\alpha_i A_{it} + \beta_2 Exp_{it} + \beta_3 Exp_{it}^2 + \beta_4 Wks_{it} + \beta_5 Occ_{it} + \\ & \beta_6 Ind_{it} + \beta_7 South_{it} + \beta_8 SMSA_{it} + \beta_9 MS_{it} + \beta_{10} Union_{it} + \\ & \beta_{11} Ed_{it} + \beta_{12} Fem_{it} + \sum_{t=1977}^{1982}\gamma_t D_{it} + \varepsilon_{it},\end{aligned}$$

where A_{it} equals one for individual i in every period and zero otherwise. The unobserved effect, α_i, in an earnings model could include factors such as ability, general skill, motivation, and fundamental experience. This model would contain the 12 variables from earlier plus the six time dummy variables for the periods, plus the 595 dummy variables for the individuals. There are some distinctive features of this model to be considered before it can be estimated.

- Because the full set of time dummy variables, D_{it}, $t = 1976, \ldots, 1982$, sums to 1 at every observation, which would replicate the constant term, one of them is dropped—1976 is

[3] Authorities differ a bit on this list. From the MPAA, we have Drama, Romance, Comedy, Action, Fantasy, Adventure, Family, Animated, Thriller, Mystery, Science Fiction, Horror, Crime.

identified as the "base year" in the results in Table 6.3. This avoids a multicollinearity problem known as the dummy variable trap.[4] The same problem will arise with the set of individual dummy variables, A_{it}, $i = 1, \ldots, 595$. The obvious remedy is to drop one of the effects, say the last one. An equivalent strategy that is usually used is to drop the overall constant term, leaving the "fixed effects" form of the model,

$$\begin{aligned}\ln Wage_{it} = {} & \sum_{i=1}^{595} \alpha_i A_{it} + \beta_2 Exp_{it} + \beta_3 Exp_{it}^2 + \beta_4 Wks_{it} + \beta_5 Occ_{it} + \beta_6 Ind_{it} + \\ & \beta_7 South_{it} + \beta_8 SMSA_{it} + \beta_9 MS_{it} + \beta_{10} Union_{it} + \beta_{11} Ed_{it} + \\ & \beta_{12} Fem_{it} + \sum_{t=1977}^{1982} \gamma_t D_{it} + \varepsilon_{it}\end{aligned}$$

(This is a application of Theorem 3.8.) Note that this does not imply that the base year time dummy variable should now be restored. If so, the dummy variable trap would reappear as

$$\sum_{i=1}^{595} A_{it} = \sum_{t=1976}^{1982} D_{it}.$$

In a model that contains a set of fixed individual effects, it is necessary either to drop the overall constant term or one of the effects.

- There is another subtle multicollinearity problem in this model. The variable Fem_{it} does not change within the block of 7 observations for individual i—it is either 1 or 0 in all 7 years for each person. Let the matrix $\mathbf{A}$ be the 4165×595 matrix in which the ith column contains $\mathbf{a}_i$, the dummy variable for individual i. Let **fem** be the 4165×1 vector that contains the variable Fem_{it}; **fem** is the column of the full data matrix that contains FEM_{it}. In the block of seven rows for individual i, the 7 elements of **fem** are all 1 or 0 corresponding to Fem_{it}. Finally, let the 595×1 vector $\mathbf{f}$ equal 1 if individual i is female and 0 if male. Then, it is easy to see that $\mathbf{fem} = \mathbf{Af}$. That is, the column of the data matrix that contains Fem_{it} is a linear combination of the individual dummy variables, again, a multicollinearity problem. This is a general result:

  In a model that contains a full set of N individual effects represented by a set of N dummy variables, any other variable in the model that takes the same value in every period for every individual can be written as a linear combination of those effects.

  This means that the coefficient on Fem_{it} cannot be estimated. The natural remedy is to fix that coefficient at zero—that is, to drop that variable. In fact, the education variable, ED_{it}, has the same characteristic and must also be dropped from the model. This turns out to be a significant disadvantage of this formulation of the model for data such as these. Indeed, in this application, the gender effect was of particular interest. (We will examine the model with individual heterogeneity modeled as fixed effects in greater detail in Chapter 11.)

[4] A second time dummy variable is dropped in the model results on the right-hand side of Table 6.3. This is a result of another dummy variable trap that is specific to this application. The experience variable, *EXP*, is a simple count of the number of years of experience, starting from an individual specific value. For the first individual in the sample, $EXP_{1,t} = 3, \ldots, 9$ while for the second, it is $EXP_{2,t} = 30, \ldots, 36$. With the individual specific constants and the six time dummy variables, it is now possible to reproduce $EXP_{i,t}$ as a linear combination of these two sets of dummy variables. For example, for the first person, $EXP_{1,1} = 3 \times A_{1,1}$; $EXP_{1,2} = 3 \times A_{1,2} + D_{1,1978}$; $EXP_{1,3} = 3 \times A_{1,3} + 2D_{1,1979}$; $EXP_{1,4} = 3 \times A_{1,3} + 3D_{1,1980}$ and so on. So, each value EXP_{it} can be produced as a linear combination of A_{it} and one of the D_{it}'s. Dropping a second period dummy variable interrupts this result.

- The model with N individual effects has become very unwieldy. The wage equation now has more than 600 variables in it; later we will analyze a similar data set with more than 7,000 individuals. One might question the practicality of actually doing the computations. This particular application shows the power of the Frisch–Waugh result, Theorem 3.2—the computation of the regression is equally straightforward whether there are a few individuals or millions. To see how this works, write the log wage equation as

$$y_{it} = \alpha_i + \mathbf{x}_{it}'\boldsymbol{\beta} + \varepsilon_{it}.$$

We are not necessarily interested in the specific constants α_i, but they must appear in the equation to control for the individual unobserved effects. Assume that there are no invariant variables such as FEM_{it} in $\mathbf{x}_{it}$. The mean of the observations for individual i is

$$\bar{y}_i = \frac{1}{7}\sum_{t=1976}^{1982} y_{it} = \alpha_i + \bar{\mathbf{x}}_i'\boldsymbol{\beta} + \bar{\varepsilon}_i.$$

A strategy for estimating $\boldsymbol{\beta}$ without having to worry about α_i is to transform the data using simple deviations from group means:

$$y_{it} - \bar{y}_i = (\mathbf{x}_{it} - \bar{\mathbf{x}}_i)'\boldsymbol{\beta} + (\varepsilon_{it} - \bar{\varepsilon}_i).$$

This transformed model can be estimated by least squares. All that is necessary is to transform the data beforehand. This computation is automated in all modern software. (Details of the theoretical basis of the computation are considered in Chapter 11.)

> *To compute the least squares estimates of the coefficients in a model that contains N dummy variables for individual fixed effects, the data are transformed to deviations from individual means, then simple least squares is used based on the transformed data. (Time dummy variables are transformed as well.) Standard errors are computed in the ways considered earlier, including robust standard errors for heteroscedasticity. Correcting for clustering within the groups would be natural.*

Notice what becomes of a variable such as *FEM* when we compute $(x_{it} - \bar{x}_i)$. Because *FEM* and *ED* take the same value in every period, the group mean is that value, and the deviations from the means becomes zero at every observation. The regression cannot be computed if $\mathbf{X}$ contains any columns of zeros. Finally, for some purposes, we might be interested in the estimates of the individual effects, α_i. We can show using Theorem 3.2 that the least squares coefficients on A_{it} in the original model would be $a_i = \bar{y}_i - \bar{\mathbf{x}}_i'\mathbf{b}$.

Results of the fixed effects regression are shown at the right in Table 6.3. Accounting for individual effects in this fashion often produces quite substantial changes in the results. Notice that the fit of the model, measured by R^2, improves dramatically. The effect of *UNION* membership, which was large and significant before has essentially vanished. And, unfortunately, we have lost view of the gender and education effects.

Example 6.5 Sports Economics: Using Dummy Variables for Unobserved Heterogeneity[5]

In 2000, the Texas Rangers major league baseball team signed 24-year-old Alex Rodriguez (A-Rod), who was claimed at the time to be "the best player in baseball," to the largest contract in baseball history (up to that time). It was publicized to be some $25Million/year for

[5] This application is based on Cohen, R. and Wallace, J., "A-Rod: Signing the Best Player in Baseball," *Harvard Business School*, Case 9-203-047, Cambridge, 2003.

10 years, or roughly a quarter of a billion dollars.[6] Treated as a capital budgeting decision, the investment is complicated partly because of the difficulty of valuing the benefits of the acquisition. Benefits would consist mainly of more fans in the stadiums where the team played, more valuable broadcast rights, and increased franchise value. We (and others) consider the first of these. It was projected that A-Rod could help the team win an average of 8 more games per season and would surely be selected as an All-Star every year. How do 8 additional wins translate into a marginal value for the investors? The franchise value and broadcast rights are highly speculative. But there is a received literature on the relationship between team wins and game attendance, which we will use here.[7] The final step will then be to calculate the value of the additional attendance.

Appendix Table F6.5 contains data on attendance, salaries, games won, and several other variables for 30 teams observed from 1985 to 2001. (These are *panel data*. We will examine this subject in greater detail in Chapter 11.) We consider a **dynamic linear regression model**,

$$Attendance_{i,t} = \Sigma_i \alpha_i A_{i,t} + \gamma Attendance_{i,t-1} + \beta_1 Wins_{i,t} + \beta_2 Wins_{i,t-1} + \beta_3 All\ Stars_{i,t} + \varepsilon_{i,t},$$
$$i = 1, \ldots, 30; t = 1985, \ldots, 2001.$$

The previous year's attendance and wins are loyalty effects. The model contains a separate constant term for each team. The effect captured by α_i includes the size of the market and any other unmeasured time constant characteristics of the market.

The team specific dummy variable, $A_{i,t}$ is used to model unit specific **unobserved heterogeneity**. We will revisit this modeling aspect in Chapter 11. The setting is different here in that in the panel data context in Chapter 11, the sampling framework will be with respect to units "*i*" and statistical properties of estimators will refer generally to increases in the number of units. Here, the number of units (teams) is fixed at 30, and asymptotic results would be based on additional years of data.[8]

Table 6.4 presents the regression results for the dynamic model. Results are reported with and without the separate team effects. Standard errors for the estimated coefficients are adjusted for the clustering of the observations by team. The *F* statistic for $H_0: \alpha_i = \alpha$, $i=1, \ldots, 31$ is computed as

$$F[30,401] = \frac{(23.267 - 20.254)/30}{20.254/401} = 1.988$$

The 95% critical value for *F*[30,401] is 1.49 so the hypothesis of no separate team effects is rejected. The individual team effects appear to improve the model—note the peculiar negative loyalty effect in the model without the team effects.

In the dynamic equation, the long run equilibrium attendance would be

$$Attendance^* = (\alpha_i + \beta_1 Wins^* + \beta_2 Wins^* + \beta_3 All\ Stars^*)/(1 - \gamma).$$

(See Section 11.11.3.) The marginal value of winning one more game every year would be $(\beta_1 + \beta_2)/(1 - \gamma)$. The effect of winning 8 more games per year and having an additional All-Star on the team every year would be

$$(8(\beta_1 + \beta_2) + \beta_3)/(1 - \gamma) \times 1 \text{ million} = 268{,}270 \text{ additional fans/season.}$$

[6] Though it was widely reported to be a 10-year arrangement, the payout was actually scheduled over more than 20 years, and much of the payment was deferred until the latter years. A realistic present discounted value at the time of the signing would depend heavily on assumptions, but using the 8% standard at the time, would be roughly $160M, not $250M.

[7] See, for example, *The Journal of Sports Economics* and Lemke, Leonard, and Tlhokwane (2009).

[8] There are 30 teams in the data set, but one of the teams changed leagues. This team is treated as two observations.

TABLE 6.4 Estimated Attendance Model

Mean of Attendance	2.22048 Million					
Number of observations	437 (31 Teams)					
	No Team Effects			Team Effects		
Sum of squared residuals	23.267			20.254		
Standard error	0.23207			0.24462		
R-squared	0.74183			0.75176		
Adjusted *R*-squared	0.73076			0.71219		
Variable	***Coefficient***	***Standard Error****	***t Ratio***	***Coefficient***	***Standard Error****	***t Ratio***
$Attendance_{t-1}$	0.70233	0.03507	20.03	0.54914	0.02760	16.76
Wins	0.00992	0.00147	6.75	0.01109	0.00157	7.08
$Wins_{t-1}$	−0.00051	0.00117	−0.43	0.00220	0.00100	2.20
All stars	0.02125	0.01241	1.71	0.01459	0.01402	1.04
Constant	−1.20827	0.87499	−1.38	Individual Team Effects		

*Standard errors clustered at the team level.

In this case, the calculation of monetary value is 268,270 fans times $50 per fan (possibly somewhat high) or about $13.0 million against the cost of roughly $18 to $20 million per season.

6.2.4 SETS OF CATEGORIES

The case in which several sets of dummy variables are needed is much the same as those we have already considered, with one important exception. Consider a model of statewide per capita expenditure on education, y, as a function of statewide per capita income, x. Suppose that we have observations on all $n = 50$ states for $T = 10$ years. A regression model that allows the expected expenditure to change over time as well as across states would be

$$y_{it} = \alpha + \beta x_{it} + \delta_i + \theta_t + \varepsilon_{it}.$$

As before, it is necessary to drop one of the variables in each set of dummy variables to avoid the dummy variable trap. For our example, if a total of 50 state dummies and 10 time dummies is retained, a problem of *perfect multicollinearity* remains; the sums of the 50 state dummies and the 10 time dummies are the same, that is, 1. One of the variables in each of the sets (or the overall constant term and one of the variables in one of the sets) must be omitted.

Example 6.6 Analysis of Covariance

The data in Appendix Table F6.1 were used in a study of efficiency in production of airline services in Greene (2007a). The airline industry has been a favorite subject of study [e.g., Schmidt and Sickles (1984); Sickles, Good, and Johnson (1986)], partly because of interest in this rapidly changing market in a period of deregulation and partly because of an abundance of large, high-quality data sets collected by the (no longer existent) Civil Aeronautics Board. The original data set consisted of 25 firms observed yearly for 15 years (1970 to 1984), a "balanced panel." Several of the firms merged during this period and several others experienced strikes, which reduced the number of complete observations substantially. Omitting these and others

TABLE 6.5 *F* Tests for Firm and Year Effects

Model	*Sum of Squares*	*Restrictions on Full Model*	*F*	*Degrees of Freedom*
Full model	0.17257	0	—	
Time effects only	1.03470	5	65.94	[5, 66]
Firm effects only	0.26815	14	2.61	[14, 66]
No effects	1.27492	19	22.19	[19, 66]

because of missing data on some of the variables left a group of 10 full observations, from which we have selected 6 for the example to follow. We will fit a cost equation of the form

$$\ln C_{i,t} = \beta_1 + \beta_2 \ln Q_{i,t} + \beta_3 \ln^2 Q_{i,t} + \beta_4 \ln P_{fuel,i,t} + \beta_5 \, Load\ Factor_{i,t} + \sum_{t=1}^{14} \theta_t D_{i,t} + \sum_{i=1}^{5} \delta_i F_{i,t} + \varepsilon_{i,t}.$$

The dummy variables are $D_{i,t}$, which is the year variable, and $F_{i,t}$, which is the firm variable. We have dropped the first one in each group. The estimated model for the full specification is

$$\ln C_{i,t} = 12.89 + 0.8866 \ln Q_{i,t} + 0.01261 \ln^2 Q_{i,t} + 0.1281 \ln P_{fuel,i,t} - 0.8855 \, Load\ Factor_{i,t} + \text{time effects} + \text{firm effects} + e_{i,t}.$$

We are interested in whether the firm effects, the time effects, both, or neither are statistically significant. Table 6.5 presents the sums of squares from the four regressions. The *F* statistic for the hypothesis that there are no firm-specific effects is 65.94, which is highly significant. The statistic for the time effects is only 2.61, which is also larger than the critical value of 1.84. In the absence of the year-specific dummy variables, the year-specific effects are probably largely absorbed by the price of fuel.

6.2.5 THRESHOLD EFFECTS AND CATEGORICAL VARIABLES

In most applications, we use dummy variables to account for purely qualitative factors, such as membership in a group, or to represent a particular time period. There are cases, however, in which the dummy variable(s) represents levels of some underlying factor that might have been measured directly if this were possible. For example, education is a case in which we often observe certain thresholds rather than, say, years of education. Suppose, for example, that our interest is in a regression of the form

$$Earnings = \beta_1 + \beta_2 \, Age + Effect\ of\ Education + \varepsilon.$$

The data on education might consist of the highest level of education attained, such as less than high school (*LTHS*), high school (*HS*), college (*C*), post graduate (*PG*). An obviously unsatisfactory way to proceed is to use a variable, *E*, that is 0 for the first group, 1 for the second, 2 for the third, and 3 for the fourth. That would be $Earnings = \beta_1 + \beta_2 \, Age + \beta_3 E + \varepsilon$. The difficulty with this approach is that it assumes that the increment in income at each threshold is the same; β_3 is the difference between income with post graduate study and college and between college and high school.[9]

[9] One might argue that a regression model based on years of education instead of this sort of step function would be likewise problematic. It seems natural that in most cases, the 12th year of education (with graduation) would be far more valuable than the 11th.

This is unlikely and unduly restricts the regression. A more flexible model would use three (or four) binary variables, one for each level of education. Thus, we would write

$$Earnings = \beta_1 + \beta_2\, Age + \delta_B\, HS + \delta_M\, C + \delta_P\, PG + \varepsilon.$$

The correspondence between the coefficients and income for a given age is

$$\begin{aligned}
&\text{Less Than High School:} && E[Earnings|Age, LTHS] &&= \beta_1 + \beta_2\, Age,\\
&\text{High School:} && E[Earnings|Age, HS] &&= \beta_1 + \beta_2\, Age + \delta_{HS},\\
&\text{College:} && E[Earnings|Age, C] &&= \beta_1 + \beta_2\, Age + \delta_C,\\
&\text{Post Graduate:} && E[Earnings|Age, PG] &&= \beta_1 + \beta_2\, Age + \delta_{PG}.
\end{aligned}$$

The differences between, say, δ_{PG} and δ_C and between δ_C and δ_{HS} are of interest. Obviously, these are simple to compute. An alternative way to formulate the equation that reveals these differences directly is to redefine the dummy variables to be 1 if the individual has the level of education, rather than whether the level is the highest obtained. Thus, for someone with post graduate education, all three binary variables are 1, and so on. By defining the variables in this fashion, the regression is now

$$\begin{aligned}
&\textit{Less Than High School:} && E[Earnings|Age, LTHS] &&= \beta_1 + \beta_2\, Age,\\
&\textit{High School:} && E[Earnings|Age, HS] &&= \beta_1 + \beta_2\, Age + \delta_{HS},\\
&\textit{College:} && E[Earnings|Age, C] &&= \beta_1 + \beta_2\, Age + \delta_{HS} + \delta_C,\\
&\textit{Post Graduate:} && E[Earnings|Age, PG] &&= \beta_1 + \beta_2\, Age + \delta_{HS} + \delta_C + \delta_{PG}.
\end{aligned}$$

Instead of the difference between post graduate and the base case of less than high school, in this model δ_{PG} is the marginal value of the post graduate education, *after college*.

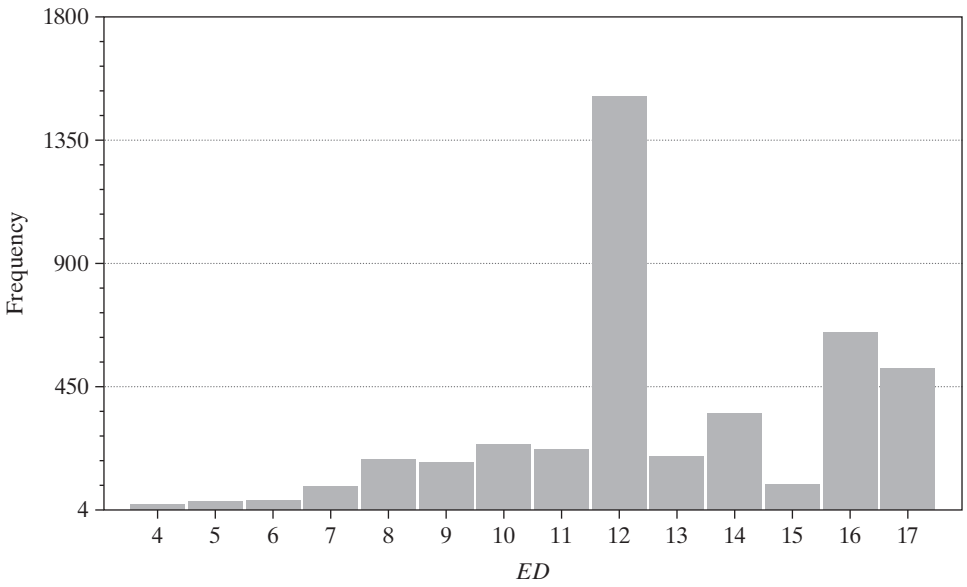
6.2.6 TRANSITION TABLES

When a group of categories appear in the model as a set of dummy variables, as in Example 6.4, each included dummy variable reports the comparison between its category and the "base case." In the movies example, the four reported values each report the comparison to the base category, the nine omitted genres. The comparison of the groups to each other is also a straightforward calculation. In Example 6.4, the reported values for *Action*, *Comedy*, *Animated*, and *Horror* are (−0.869, −0.016, −0.833, +0.375). The implication is, for example, that $E[\ln Revenue|\mathbf{x}]$ is 0.869 less for *Action* movies than the base case. Moreover, based on the same results, the expected log revenue for *Animated* movies is $-0.833 - (-0.869) = +0.036$ greater than for *Action* movies. A standard error for the difference of the two coefficients would be computed using the square root of

$$\begin{aligned}
\text{Asy.Var}[b_{Animated} - b_{Action}] = {}& \text{Asy.Var}[b_{Animated}] + \text{Asy.Var}[b_{Action}]\\
&- 2\text{Asy.Cov}[b_{Animated}, b_{Action}].
\end{aligned}$$

A similar effect could be computed for each pair of outcomes. Hodge and Shankar (2014) propose a useful framework for arranging the effects of a sequence of categories based on this principle. An application to five categories of health outcomes is shown in

Figure 6.1 Education Levels in Log Wage Data.



Contoyannis, Jones, and Rice (2004). The education thresholds example in the previous example is another natural application.

Example 6.7 Education Thresholds in a Log Wage Equation

Figure 6.1 is a histogram for the education levels reported in variable *ED* in the ln *Wage* model of Example 6.3. The model in Table 6.3 constrains the effect of education to be the same 5.5% per year for all values of *ED*. A possible improvement in the specification might be provided by treating the threshold values separately. We have recoded *ED* in these data to be

Less Than High School	$= 1$ if $ED \leq 11$	(22% *of the sample*),
High School	$= 1$ if $ED = 12$	(36% *of the sample*),
College	$= 1$ if $13 \leq ED \leq 16$	(30% *of the sample*),
Post Grad	$= 1$ if $ED = 17$	(12% *of the sample*).

(Admittedly, there might be some misclassification at the margins. It also seems likely that the *Post Grad* category is "top coded"—17 years represents 17 or more.) Table 6.6 reports the respecified regression model. Note, first, the estimated gender effect is almost unchanged. But, the effects of education are rather different. According to these results, the marginal value of high school compared to less than high school is 0.13832, or 14.8%. The estimated marginal value of attending college after high school is $0.29168 - 0.13832 = 0.15336$, 16.57%—this is roughly 4% per year for four years compared to 5.5% estimated earlier. But, again, one might suggest that most of that gain would be a "sheepskin" effect attained in the fourth year by graduating. Hodge and Shankar's "transition matrix" is shown in Table 6.7. (We have omitted the redundant terms and transitions from more education to less which are the negatives of the table entries.)

TABLE 6.6 Estimated log Wage Equations with Education Thresholds

	Threshold Effects			*Education in Years*		
Sum of squared residuals	403.329			391.056		
Standard error of the regression	0.31194			0.30708		
R-squared based on 4165 observations	0.54524			0.55908		
	Coefficient	*Clustered Std.Error*	*t Ratio*	*Coefficient*	*Clustered Std.Error*	*t Ratio*
Constant	5.60883	0.10087	55.61	5.08397	0.12998	39.11
EXP	0.03129	0 .00421	7.44	0.03128	0.00419	7.47
EXP^2	−0.00056	0.00009	−5.97	−0.00055	0.00009	−5.86
WKS	0.00383	0.00157	2.44	0.00394	0.00158	2.50
OCC	−0.16410	0.02683	−6.12	−0.14116	0.02687	−5.25
IND	0.05365	0.02368	2.27	0.05661	0.02343	2.42
SOUTH	−0.07438	0.02704	−2.75	−0.07180	0.02632	−2.73
SMSA	0.16844	0.02368	7.11	0.15423	0.02349	6.57
MS	0.10756	0.04470	2.41	0.09634	0.04301	2.24
UNION	0.07736	0.02405	3.22	0.08052	0.02335	3.45
FEM	−0.35323	0.05005	−7.06	−0.36502	0.04829	−7.56
ED				0.05499	0.00556	9.88
LTHS	0.00000	—	—			
HS	0.13832	0.03351	4.13			
COLLEGE	0.29168	0.04181	6.98			
POSTGRAD	0.40651	0.04896	8.30			
Year(Base = 1976)						
1977	0.07493	0.00608	12.33	0.07461	0.00601	12.42
1978	0.19720	0.00997	19.78	0.19611	0.00989	19.82
1979	0.28472	0.01023	27.83	0.28358	0.01016	27.90
1980	0.36377	0.00997	36.47	0.36264	0.00985	36.82
1981	0.43877	0.01147	38.25	0.43695	0.01133	38.58
1982	0.52357	0.01219	42.94	0.52075	0.01211	43.00

TABLE 6.7 Education Effects in Estimated Log Wage Equation

Effects of switches between categories in education level

Initial Education	*New Education*	*Partial Effect*	*Standard Error*	*t Ratio*
LTHS	*HS*	0.13832	0.03351	4.13
LTHS	*COLLEGE*	0.29168	0.04181	6.98
LTHS	*POSTGRAD*	0.40651	0.04896	8.30
HS	*COLLEGE*	0.15336	0.03047	5.03
HS	*POSTGRAD*	0.26819	0.03875	6.92
COLLEGE	*POSTGRAD*	0.11483	0.03787	3.03

6.3 DIFFERENCE IN DIFFERENCES REGRESSION

Many recent studies have examined the causal effect of a **treatment** on some kind of **response**. Examples include the effect of attending an elite college on lifetime income [Dale and Krueger (2002, 2011)], the effect of cash transfers on child health [Gertler (2004)], the effect of participation in job training programs on income [LaLonde (1986)], the effect on employment of an increase in the minimum wage in one of two neighboring states [Card and Krueger (1994)] and pre- versus post-regime shifts in macroeconomic models [Mankiw (2006)], to name but a few.

6.3.1 TREATMENT EFFECTS

The applications can often be formulated in regression models involving a treatment dummy variable, as in

$$y_i = \mathbf{x}_i'\boldsymbol{\beta} + \delta D_i + \varepsilon_i,$$

where the shift parameter, δ (under the right assumptions), measures the causal effect of the treatment or the policy change (conditioned on $\mathbf{x}$) on the sampled individuals. For example, Table 6.6 provides a log wage equation based on a national (U.S.) panel survey. One of the variables is *UNION*, a dummy variable that indicates union membership. Measuring the effect of union membership on wages is a longstanding objective in labor economics—see, for example, Card (2001). Our estimate in Table 6.6 is roughly 0.08, or 8%. It will take a bit of additional specification analysis to conclude that the *UNION* dummy truly does measure the effect of membership in that context.[10]

In the simplest case of a comparison of one group to another, without covariates,

$$y_i = \beta_1 + \delta D_i + \varepsilon_i.$$

Least squares regression of y on D will produce

$$b_1 = (\bar{y} \,|\, D_i = 0),$$

that is, the average outcome of those who did not experience the treatment, and

$$d = (\bar{y} \,|\, D_i = 1) - (\bar{y} \,|\, D_i = 0),$$

the difference in the means of the two groups. Continuing our earlier example, if we measure the *UNION* effect in Table 6.6 without the covariates, we find

$$\ln \mathit{Wage} = 6.673\ (0.023) + 0.00834\ \mathit{UNION}\ (0.028).$$

(Standard errors are in parentheses.) Based on a simple comparison of means, there appears to be a less than 1% impact of union membership. This is in sharp contrast to the 8% reported earlier.

When the analysis is of an intervention that occurs over time to everyone in the sample, such as in Krueger's (1999) analysis of the Tennessee STAR experiment in which school performance measures were observed before and after a policy that dictated a change in class sizes, the treatment dummy variable will be a period indicator, $T_t = 0$ in period 1 and 1 in period 2. The effect in β_2 then measures the change in the outcome variable, for example, school performance, pre- to post-intervention; $b_2 = \bar{y}_1 - \bar{y}_0$.

[10] See, for example, Angrist and Pischke (2009, pp. 221–225.)

The assumption that the treatment group does not change from period 1 to period 2 (or that the treatment group and the control group look the same in all other respects) weakens this analysis. A strategy for strengthening the result is to include in the sample a group of **control observations** that do not receive the treatment. The change in the outcome for the **treatment group** can then be compared to the change for the **control group** under the presumption that the difference is due to the intervention. An intriguing application of this strategy is often used in clinical trials for health interventions to accommodate the **placebo effect**. The placebo effect is a controversial, but apparently tangible outcome in some clinical trials in which subjects "respond" to the treatment even when the treatment is a decoy intervention, such as a sugar or starch pill in a drug trial.[11] A broad template for assessment of the results of such a clinical trial is as follows: The subjects who receive the placebo are the controls. The outcome variable—level of cholesterol, for example—is measured at the baseline for both groups. The treatment group receives the drug, the control group receives the placebo, and the outcome variable is measured pre- and post-treatment. The impact is measured by the difference in differences,

$$E = [(\bar{y}_{exit}|treatment) - (\bar{y}_{baseline}|treatment)] - [(\bar{y}_{exit}|placebo) - (\bar{y}_{baseline}|placebo)].$$

The presumption is that the difference in differences measurement is robust to the placebo effect *if it exists*. If there is no placebo effect, the result is even stronger (assuming there is a result).

A common social science application of treatment effect models is in the evaluation of the effects of discrete changes in policy.[12] A pioneering application is the study of the Manpower Development and Training Act (MDTA) by Ashenfelter and Card (1985) and Card and Krueger (2000). A widely discussed application is Card and Krueger's (1994) analysis of an increase in the minimum wage in New Jersey. The simplest form of the model is one with a pre- and post-treatment observation on a group, where the outcome variable is y, with

$$y_{it} = \beta_1 + \beta_2 T_t + \beta_3 D_i + \delta(T_t \times D_i) + \varepsilon_{it},\ t = 0, 1. \qquad \textbf{(6-3)}$$

In this model, T_t is a dummy variable that is zero in the pre-treatment period and one after the treatment and D_i equals one for those individuals who received the treatment. The change in the outcome variable for the treated individuals will be

$$(y_{i2}|D_i = 1) - (y_{i1}|D_i = 1) = (\beta_1 + \beta_2 + \beta_3 + \delta) - (\beta_1 + \beta_3) = \beta_2 + \delta.$$

For the controls, this is

$$(y_{i2}|D_i = 0) - (y_{i1}|D_i = 0) = (\beta_1 + \beta_2) - \beta_1 = \beta_2.$$

The difference in differences is

$$[(y_{i2}|D_i = 1) - (y_{i1}|D_i = 1)] - [(y_{i2}|D_i = 0) - (y_{i1}|D_i = 0)] = \delta.$$

[11] See Hróbjartsson and Götzsche (2001).

[12] Surveys of literatures on treatment effects, including use of "D-i-D" estimators, are provided by Imbens and Wooldridge (2009), Millimet, Smith, and Vytlacil (2008), Angrist and Pischke (2009), and Lechner (2011).

In the multiple regression of y_{it} on a constant, T, D, and $T \times D$, the least squares estimate of δ will equal the difference in the changes in the means,

$$\begin{aligned} d &= (\bar{y}|\, D = 1, \textit{Period } 2) - (\bar{y}|\, D = 1, \textit{Period } 1) \\ &\quad - (\bar{y}|\, D = 0, \textit{Period } 2) - (\bar{y}|\, D = 0, \textit{Period } 1) \\ &= \Delta\bar{y}|\, \textit{treatment} - \Delta\bar{y}|\, \textit{control}. \end{aligned}$$

The regression is called a difference in differences estimator in reference to this result.

Example 6.8 SAT Scores

Each year, about 1.7 million American high school students take the SAT test. Students who are not satisfied with their performance have the opportunity to retake the test. Some students take an SAT prep course, such as Kaplan or Princeton Review, before the second attempt in the hope that it will help them increase their scores. An econometric investigation might consider whether these courses are effective in increasing scores. The investigation might examine a sample of students who take the SAT test twice, with scores y_{i0} and y_{i1}. The time dummy variable T_t takes value $T_0 = 0$ "before" and $T_1 = 1$ "after." The treatment dummy variable is $D_i = 1$ for those students who take the prep course and 0 for those who do not. The applicable model would be (6-3),

$$\textit{SAT Score}_{i,t} = \beta_1 + \beta_2\, \textit{2ndTest}_t + \beta_3\, \textit{PrepCourse}_i + \delta\, \textit{2ndTest}_t \times \textit{PrepCourse}_i + \varepsilon_{i,t}.$$

The estimate of δ would, in principle, be the treatment, or prep course effect.

This small example illustrates some major complications. First, and probably most important, the setting does not describe a randomized experiment such as the clinical trial suggested earlier would be. The treatment variable, *PrepCourse*, would naturally be taken by those who are persuaded that it would provide a benefit—that is, the treatment variable is not an exogenous variable. Unobserved factors that are likely to contribute to higher test scores (and are embedded in $\varepsilon_{i,t}$) would likely motivate the student to take the prep course as well. This *selection effect* is a compelling confounder of studies of treatment effects when the treatment is voluntary and self selected. Dale and Krueger's (2002, 2011) analysis of the effect of attendance at an elite college provides a detailed analysis of this issue. Second, test performance, like other performance measures, is probably subject to regression to the mean—there is a negative autocorrelation in such measures. In this regression context, an unusually high disturbance in period 0, all else equal, would likely be followed by a low value in period 1. Of course, those who achieve an unusually high test score in period 0 are less likely to return for the second attempt. Together with the selection effect, this produces a very muddled relationship between the outcome and the test preparation that is estimated by least squares. Finally, it is possible that there are other measurable factors (covariates) that might contribute to the test outcome or changes in the outcome. A more complete model might include these covariates. We do note any such variable $x_{i,t}$ would have to vary between the first and second test, else they would simply be absorbed in the constant term.

When the treatment is the result of a policy change or event that occurs completely outside the context of the study, the analysis is often termed a **natural experiment**. Card's (1990) study of a major immigration into Miami in 1979 is an application.

Example 6.9 A Natural Experiment: The Mariel Boatlift

A sharp change in policy can constitute a natural experiment. An example studied by Card (1990) is the Mariel boatlift from Cuba to Miami (May–September 1980), which increased the Miami labor force by 7%. The author examined the impact of this abrupt change in labor market conditions on wages and employment for nonimmigrants. The model compared Miami (the treatment group) to a similar city, Los Angeles (the control group). Let *i* denote an

individual and *D* denote the "treatment," which for an individual would be equivalent to "lived in the city that experienced the immigration." For an individual in either Miami or Los Angeles, the outcome variable is

$$Y_i = 1 \quad \text{if they are unemployed and 0 if they are employed.}$$

Let *c* denote the city and let *t* denote the period, before (1979) or after (1981) the immigration. Then, the unemployment rate in city *c* at time *t* is $E[y_{i,0}|c, t]$ if there is no immigration and it is $E[y_{i,1}|c, t]$ if there is the immigration. These rates are assumed to be constants. Then

$$E[y_{i,0}|c, t] = \beta_t + \gamma_c \qquad \text{without the immigration,}$$
$$E[y_{i,1}|c, t] = \beta_t + \gamma_c + \delta \quad \text{with the immigration.}$$

The effect of the immigration on the unemployment rate is measured by δ. The natural experiment is that the immigration occurs in Miami and not in Los Angeles but is not a result of any action by the people in either city. Then,

$$E[y_i|M,79] = \beta_{79} + \gamma_M \quad \textit{and} \quad E[y_i|M, 81] = \beta_{81} + \gamma_M + \delta \quad \textit{for Miami},$$
$$E[y_i|L, 79] = \beta_{79} + \gamma_L \quad \textit{and} \quad E[y_i|L, 81] = \beta_{81} + \gamma_L \qquad \textit{for Los Angeles}.$$

It is assumed that unemployment growth in the two cities would be the same if there were no immigration. If neither city experienced the immigration, the change in the unemployment rate would be

$$E[y_{i,0}|M, 81] - E[y_{i,0}|M, 79] = \beta_{81} - \beta_{79} \qquad \textit{for Miami},$$
$$E[y_{i,0}|L, 81] - E[y_{i,0}|L, 79] = \beta_{81} - \beta_{79} \qquad \textit{for Los Angeles}.$$

If both cities were exposed to migration,

$$E[y_{i,1}|M, 81] - E[y_{i,1}|M, 79] = \beta_{81} - \beta_{79} + \delta \quad \textit{for Miami},$$
$$E[y_{i,1}|L, 81] - E[y_{i,1}|L, 79] = \beta_{81} - \beta_{79} + \delta \quad \textit{for Los Angeles}.$$

Only Miami experienced the immigration (the "treatment"). The difference in differences that quantifies the result of the experiment is

$$\{E[y_{i,1}|M, 81] - E[y_{i,1}|M, 79]\} - \{E[y_{i,0}|L, 81] - E[y_{i,0}|L, 79]\} = \delta.$$

The author examined changes in employment rates and wages in the two cities over several years after the boatlift. The effects were surprisingly modest (essentially nil) given the scale of the experiment in Miami.

Example 6.10 Effect of the Minimum Wage

Card and Krueger's (1994) widely cited analysis of the impact of a change in the minimum wage is similar to Card's analysis of the Mariel Boatlift. In April 1992, New Jersey (NJ) raised its minimum wage from $4.25 to $5.05. The minimum wage in neighboring Pennsylvania (PA) was unchanged. The authors sought to assess the impact of this policy change by examining the change in employment in the two states from February to November, 1992 at fast food restaurants that tended to employ large numbers of people at the minimum wage. Conventional wisdom would suggest that, all else equal, whatever labor market trends were at work in the two states, NJ's would be affected negatively by the abrupt 19% wage increase for minimum wage workers. This certainly qualifies as a natural experiment. NJ restaurants could not opt out of the treatment. The authors were able to obtain data on employment for 331 NJ restaurants and 97 PA restaurants in the first wave. Most of the first wave restaurants provided data for the second wave, 321 and 78, respectively. One possible source of "selection" would be attrition from the sample. Though the numbers are small, the possibility that the second wave sample was substantively composed of firms that were affected by the policy change

TABLE 6.8 Full Time Employment in NJ and PA Restaurants

	PA	*NJ*
First Wave (February)	23.33	20.44
Second Wave (November)	21.17	21.03
Difference	−2.16	0.59
Difference (balanced)	−2.28	0.47

would taint the analysis (e.g., if firms were driven out of business because of the increased labor costs). The authors document at some length the data collection process for the second wave. Results for their experiment are shown in Table 6.8.

The first reported difference uses the full sample of available data. The second uses the "balanced sample" of all stores that reported data in both waves. In both cases, the difference in differences would be

$$\Delta(NJ) - \Delta(\text{PA}) = +2.75 \text{ full time employees.}$$

A superficial analysis of these results suggests that they go in the wrong direction. Employment rose in NJ compared to PA in spite of the increase in the wage. Employment would have been changing in both places due to other economic conditions. The policy effect here might have distorted that trend. But, it is also possible that the trend in the two states was different. It has been assumed throughout so far that it is the same. Card and Krueger (2000) examined this possibility in a followup study. The newer data cast some doubt on the crucial assumption that the trends were the same in the two states.

Card and Krueger (1994) considered the possibility that restaurant specific factors might have influenced their measured outcomes. The implied regression would be

$$y_{it} = \beta_2 T_t + \beta_3 D_i + \delta T_t \times D_i + (\alpha_i + \boldsymbol{\gamma}'\mathbf{x}_i) + \varepsilon_{it}, \quad t = 0, 1.$$

Note the individual specific constant term that represents the unobserved heterogeneity and the addition to the regression. In the restaurant study, $\mathbf{x}_i$ was characteristics of the store such as chain store type, ownership, and region—all features that would be the same in both waves. These would be fixed effects. In the difference in differences context, while they might indeed be influential in the outcome levels, it is clear that they will fall out of the differences:

$$\begin{aligned}\Delta E[y_{it}|D_{it} = 0, \mathbf{x}_i] &= \beta_2 \qquad + \Delta(\alpha_i + \boldsymbol{\gamma}'\mathbf{x}_i),\\ \Delta E[y_{it}|D_{it} = 1, \mathbf{x}_i] &= \beta_2 + \delta + \Delta(\alpha_i + \boldsymbol{\gamma}'\mathbf{x}_i).\end{aligned}$$

The final term in both cases is zero, which leaves, as before,

$$\Delta E[y_{it}|D_{it} = 1, \mathbf{x}_i] - \Delta E[y_{i\,t}|D_{it} = 1, \mathbf{x}_i] = \delta.$$

The useful conclusion is that in analyzing differences in differences, time invariant characteristics of the individuals will not affect the conclusions.

The analysis is more complicated if the control variables, $\mathbf{x}_{it}$, do change over time. Then,

$$y_{it} = \beta_2 T_t + \beta_3 D_i + \delta T_t \times D_i + \boldsymbol{\gamma}'\mathbf{x}_{it} + \varepsilon_{it}, \; t = 0, 1.$$

Then,

$$\Delta E[y_{it}|\mathbf{x}_{it}, D_{it} = 1] = \beta_2 + \delta + \boldsymbol{\gamma}'[\Delta\mathbf{x}_{it}|D_{it} = 1]$$
$$\Delta E[y_{it}|\mathbf{x}_{it}, D_{it} = 0] = \beta_2 + \quad \boldsymbol{\gamma}'[\Delta\mathbf{x}_{it}|D_{it} = 0]$$

$$\Delta E[y_{it}|D_{it} = 1, \mathbf{x}_i] - \Delta E[y_{it}|D_{it} = 1, \mathbf{x}_i] = \delta + \boldsymbol{\gamma}'[(\Delta\mathbf{x}_{it}|D_{it} = 1) - (\Delta\mathbf{x}_{it}|D_{it} = 0)].$$

Now, if the effect of D_{it} is measured by the simple difference of means, the result will consist of the causal effect plus an additional term explained by the difference of the changes in the control variables. If individuals have been carefully sampled so that treatment and controls look the same in both periods, then the second effect might be ignorable. If not, then the second part of the regression should become part of the analysis.

6.3.2 EXAMINING THE EFFECTS OF DISCRETE POLICY CHANGES

The differences in differences result provides a convenient methodology for studying the effects of exogenously imposed policy changes. We consider an application from a recent antitrust case.

Example 6.11 Difference in Differences Analysis of a Price Fixing Conspiracy[13]

Roughly 6.5% of all British schoolchildren, and more than 18% of those over 16, attend 2,600 independent fee-paying schools. Of these, roughly 10.5% are "boarders"—the remainder attend on a day basis. Each year from 1997 until June, 2003, a group of 50 of these schools shared information about intended fee increases for boarding and day students. The information was exchanged via a survey known as the "Sevenoaks Survey" (SS). The UK Office of Fair Trading (OFT, Davies (2012)) determined that the conspiracy, which was found to lead to higher fees, was prohibited under the antitrust law, the Competition Act of 1998. The OFT intervention consisted of a modest fine (10,000GBP) on each school, a mandate for the cartel to contribute about 3,000,000GBP to a trust, and prohibition of the Sevenoaks Survey. The OFT investigation was ended in 2006, but for the purposes of the analysis, the intervention is taken to have begun with the 2004/2005 academic year.

The authors of this study investigated the impact of the OFT intervention on the boarding and day fees of the Sevenoaks schools using a difference in differences regression. The pre-intervention period is academic years 2001/02 to 2003/04. The post-intervention period extends to 2011/2012. The sample consisted of the treatment group, the 50 Sevenoaks schools, and 178 schools that were not party to the conspiracy and therefore, not impacted by the treatment. (Not necessarily. More on that below.) The "balanced panel data set" of 12 years times 228 schools, or 2,736 observations, was reduced by missing data to 1,829 for the day fees model and 1,317 for the boarding fees model. Figure 6.2 (Figures 2 and 3 from the study) shows the behavior of the boarding and day fees for the schools for the period of the study.[14] It is difficult to see a difference in the rates of change of the fees. The difference in the levels is obvious, but not yet explained.

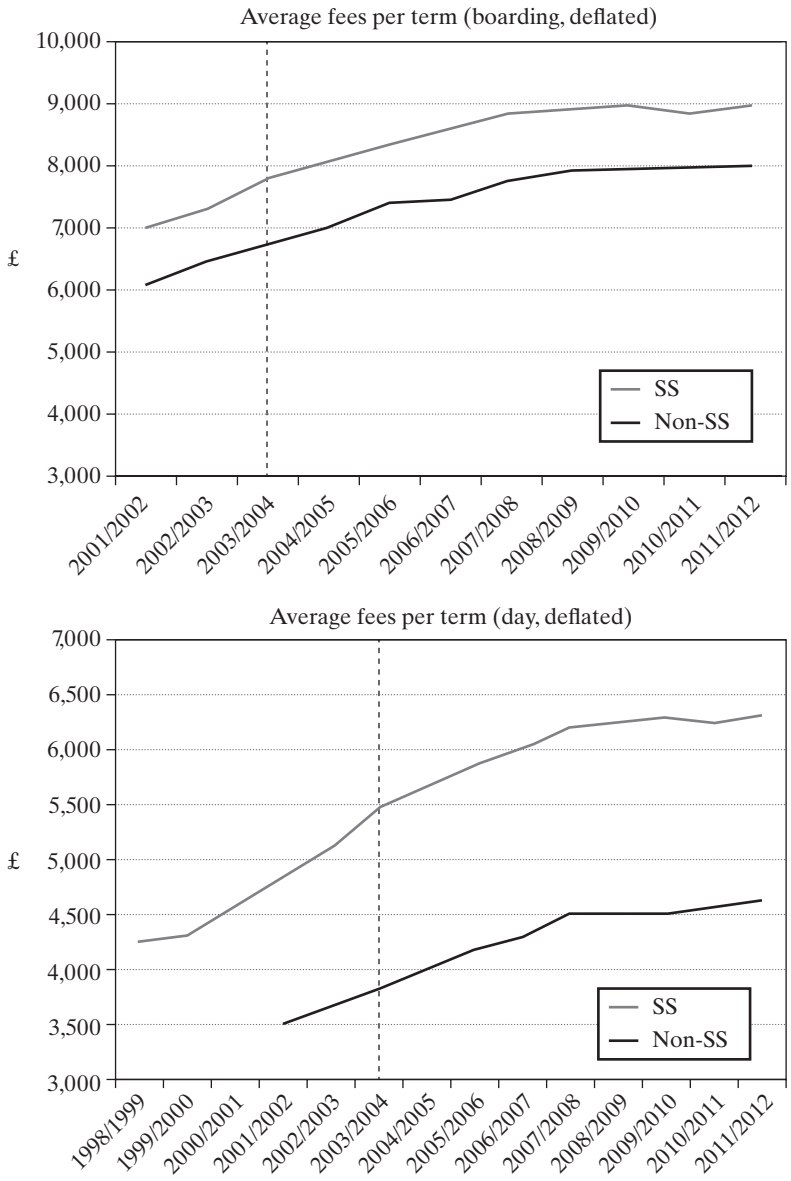
A difference in differences methodology was used to analyze the behavior of the fees. Two key assumptions are noted at the outset.

1. The schools in the control group are not affected by the intervention. This may not be the case. The non-SS schools compete with the SS schools on a price basis. If the pricing behavior of the SS schools is affected by the intervention, that of the non-SS schools may be as well.

[13] This case study is based on UK OFT (2012), Davies (2012) and Pesarisi et al. (2015).

[14] The figures are extracted from the UK OFT (2012) working paper version of the study.

Figure 6.2 Price Increases by Boarding Schools.



2. It must be assumed that the trends and influences that affect the two groups of schools outside the effect of the intervention are the same. (Recall this was an issue in Card and Krueger's analysis of the minimum wage in Example 6.10.)

The linear regression model used to study the behavior of the fees is

$$\ln Fee_{it} = \alpha_i + \beta_1 \%boarder_{it} + \beta_2 \%ranking_{it} + \beta_3 \ln pupils_{it} + \beta_4 year_t + \lambda postintervention_t + \delta SS_{it} \times postintervention_t + \varepsilon_{it}$$

Fee_{it} = inflation-adjusted day or boarding fees,
%*boarder* = percentage of the students who are boarders at school *i* in year *t*,
%*ranking* = percentile ranking of the school in *Financial Times* school rankings,
pupils = number of students in the school,
year = linear trend,
postintervention = dummy variable indicating the period after the intervention,
SS = dummy variable for Sevenoaks school,
α_i = school-specific effect, modeled using a school specific dummy variable.

The effect of interest is δ. Several assumptions underlying the data are noted to justify the interpretation of δ as the sought-after causal impact of the intervention.

a. The effect of the intervention is exerted on the fees beginning in 2004/2005.
b. In the absence of the intervention, the regime would have continued on to 2012 as it had in the past.
c. The *Financial Times* ranking variable is a suitable indicator of the quality of the ranked school.
d. As noted earlier, pricing behavior by the control schools was not affected by the intervention.

The regression results are shown in Table 6.9.

The main finding is a decline of 1.5% for day fees and 1.6% for the boarding fees. Figure 6.3 [extracted from the UK OFT (2012) version of the paper] summarizes the estimated cumulative impact of the study. The authors estimated the cumulative savings attributable to the intervention based on the results in Figure 6.3 to be roughly 85 million GBP.

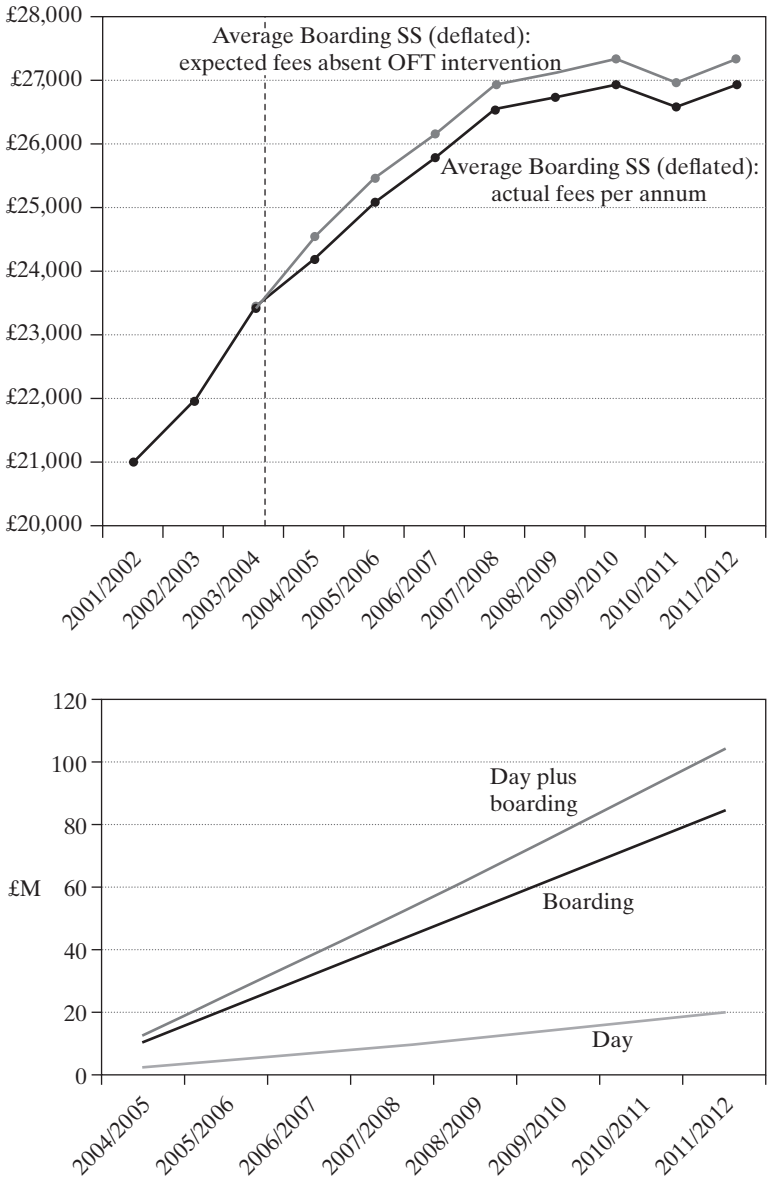
One of the central issues in policy analysis concerns measurement of treatment effects when the treatment results from an individual participation decision. In the clinical trial example given earlier, the control observations (it is assumed) do not know they they are in the control group. The treatment assignment is exogenous to the experiment. In contrast, in Krueger and Dale (1999) study, the assignment to the treatment group, attended the elite college, is completely voluntary and determined by the individual. A crucial aspect of the analysis in this case is to accommodate the almost certain outcome that the treatment dummy might be measuring the latent motivation and initiative of the participants rather than the effect of the program itself. That is the

TABLE 6.9 Estimated Models for Day and Boarding Fees*

	Day Fees	Boarding Feees
% Boarder	0.7730 (0.051)**	0.0367 (0.029)
% Ranking	−0.0147 (0.019)	0.00396 (0.015)
ln Pupils	0.0247 (0.033)	0.0291 (0.021)
Year	0.0698 (0.004)	0.0709 (0.004)
Post-intervention	0.0750 (0.027)	0.0674 (0.022)
Post-intervention and SS	−0.0149 (0.007)	−0.0162 (0.005)
N	1,825	1,311
R^2	0.949	0.957

Source: Pesaresi et al. (2015), Table 1.
* Model fit by least squares. Estimated individual fixed effects not shown.
** Robust standard errors that account for possible heteroscedasticity and autocorrelation in parentheses.

Figure 6.3 Cumulative Impact of Sevenoaks Intervention.



main appeal of the natural experiment approach—it more closely (possibly exactly) replicates the exogenous treatment assignment of a clinical trial.[15] We will examine some of these cases in Chapters 8 and 19.

[15] See Angrist and Krueger (2001) and Angrist and Pischke (2010) for discussions of this approach.

6.4 USING REGRESSION KINKS AND DISCONTINUITIES TO ANALYZE SOCIAL POLICY

The ideal situation for the analysis of a change in social policy would be a randomized assignment of a sample of individuals to treatment and control groups.[16] There are some notable examples to be found. The Tennessee STAR class size experiment was designed to study the effect of smaller class sizes in the earliest grades on short and long term student performance. [See Mosteller (1995) and Krueger (1999) and, for some criticism, Hanushek (1999, 2002).] A second prominent example is the Oregon Health Insurance Experiment.

> *The Oregon Health Insurance Experiment is a landmark study of the effect of expanding public health insurance on health care use, health outcomes, financial strain, and well-being of low-income adults. It uses an innovative randomized controlled design to evaluate the impact of Medicaid in the United States. Although randomized controlled trials are the gold standard in medical and scientific studies, they are rarely possible in social policy research. In 2008, the state of Oregon drew names by lottery for its Medicaid program for low-income, uninsured adults, generating just such an opportunity. This ongoing analysis represents a collaborative effort between researchers and the state of Oregon to learn about the costs and benefits of expanding public health insurance. (www.nber.org/oregon/)*
>
> *In 2008, a group of uninsured low-income adults in Oregon was selected by lottery to be given the chance to apply for Medicaid. This lottery provides a unique opportunity to gauge the effects of expanding access to public health insurance on the health care use, financial strain, and health of low-income adults using a randomized controlled design. In the year after random assignment, the treatment group selected by the lottery was about 25 percentage points more likely to have insurance than the control group that was not selected. We find that in this first year, the treatment group had substantively and statistically significantly higher health care utilization (including primary and preventive care as well as hospitalizations), lower out-of-pocket medical expenditures and medical debt (including fewer bills sent to collection), and better self-reported physical and mental health than the control group. [Finkelstein et al. (2011).]*

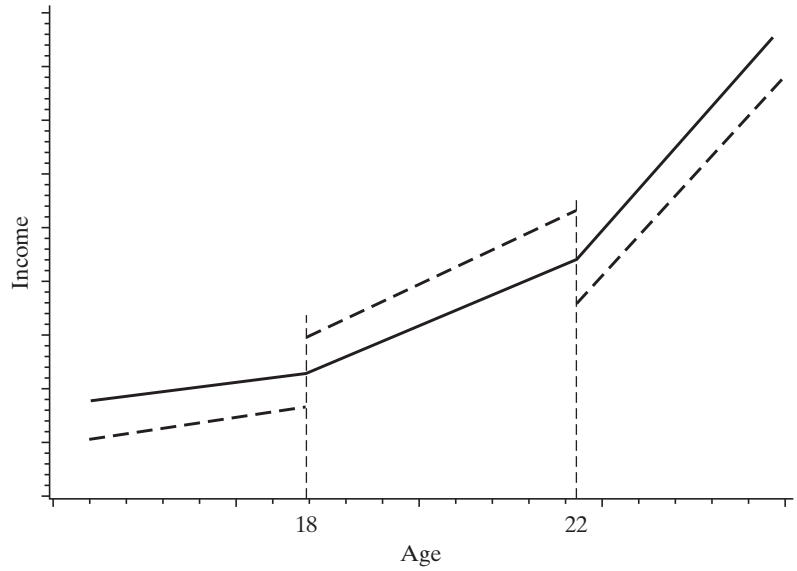
Substantive social science studies such as these, based on random assignment, are rare. The natural experiment approach, such as in Example 6.9, is an appealing alternative when it is feasible. Regression models with kinks and discontinuities have been designed to study the impact of social policy in the absence of randomized assignment.

6.4.1 REGRESSION KINKED DESIGN

A plausible description of the age profile of incomes will show incomes rising throughout but at different rates after some distinct milestones, for example, at age 18, when the typical individual graduates from high school, and at age 22, when he or she graduates from college. The profile of incomes for the typical individual in this population might appear as in Figure 6.4. We could fit such a regression model just by dividing the sample into three subsamples. However, this would neglect the continuity of the proposed function and possibly misspecify the relationship of other variables that might appear in the model. The result would appear more like the dashed figure than the continuous

[16] See Angrist and Pischke (2009).

Figure 6.4 Piecewise Linear Regression.



function we had in mind. Constrained regression can be used to achieve the desired effect. The function we wish to estimate is

$$\begin{aligned} E[income|age] = \alpha^0 + \beta^0\, age \quad & \text{if } age < 18, \\ \alpha^1 + \beta^1\, age \quad & \text{if } age \geq 18 \text{ and } age < 22, \\ \alpha^2 + \beta^2\, age \quad & \text{if } age \geq 22. \end{aligned}$$

Let

$$\begin{aligned} d_1 &= 1 \quad \text{if } age \geq t_1^*, \\ d_2 &= 1 \quad \text{if } age \geq t_2^*, \end{aligned}$$

where $t_1^* = 18$ and $t_2^* = 22$. To combine the three equations, we use

$$income = \beta_1 + \beta_2\, age + \gamma_1 d_1 + \delta_1 d_1\, age + \gamma_2 d_2 + \delta_2 d_2\, age + \varepsilon.$$

This produces the dashed function Figure 6.4. The slopes in the three segments are β_2, $\beta_2 + \delta_1$, and $\beta_2 + \delta_1 + \delta_2$. To make the function *continuous*, we require that the segments join at the thresholds—that is,

$$\begin{aligned} &\beta_1 + \beta_2 t_1^* = (\beta_1 + \gamma_1) + (\beta_2 + \delta_1) t_1^* \text{ and} \\ &(\beta_1 + \gamma_1) + (\beta_2 + \delta_1) t_2^* = (\beta_1 + \gamma_1 + \gamma_2) + (\beta_2 + \delta_1 + \delta_2) t_2^*. \end{aligned}$$

These are linear restrictions on the coefficients. The first one is

$$\gamma_1 + \delta_1 t_1^* = 0 \quad \text{or} \quad \gamma_1 = -\delta_1 t_1^*.$$

Doing likewise for the second, we obtain

$$income = \beta_1 + \beta_2\, age + \delta_1 d_1\, (age - t_1^*) + \delta_2 d_2\, (age - t_2^*) + \varepsilon.$$

Constrained least squares estimates are obtainable by multiple regression, using a constant and the variables

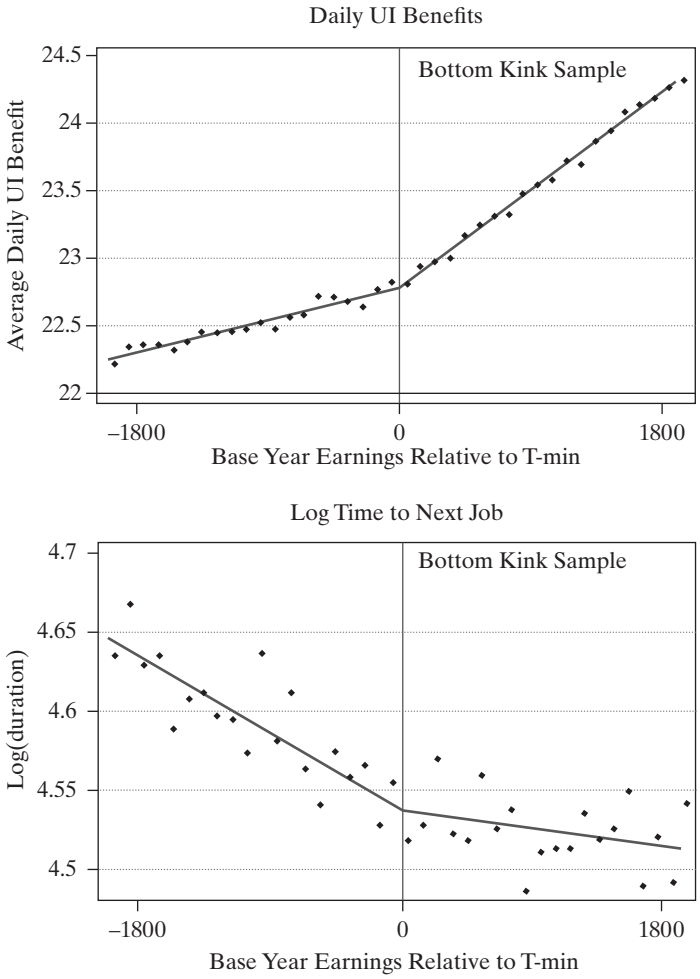
$$x_1 = age,$$
$$x_2 = age - 18 \quad \text{if } age \geq 18 \text{ and 0 othewise,}$$
$$x_3 = age - 22 \quad \text{if } age \geq 22 \text{ and 0 othewise.}$$

We can test the hypothesis that the slope of the function is constant with the joint test of the two restrictions $\delta_1 = 0$ and $\delta_2 = 0$.

Example 6.12 Policy Analysis Using Kinked Regressions

Discontinuities such as those in Figure 6.4 can be used to help identify policy effects. Card, Lee, Pei, and Weber (2012) examined the impact of unemployment insurance (UI) on the duration of joblessness in Austria using a regression kink design. The policy lever, UI, has a sharply defined benefit schedule level tied to base year earnings that can be traced through to its impact on the duration of unemployment. Figure 6.5 [from Card et al. (2012, p. 48)]

Figure 6.5 Regression Kink Design.



suggests the nature of the identification strategy. Simonsen, Skipper, and Skipper (2015) used a similar strategy to examine the effect of a subsidy on the demand for pharmaceuticals in Denmark.

6.4.2 REGRESSION DISCONTINUITY DESIGN

Van der Klaauw (2002) studied financial aid offers that were tied to SAT scores and grade point averages using a regression discontinuity design. The conditions under which the approach can be effective are when (1) the outcome, y, is a continuous variable; (2) the outcome varies smoothly with an assignment variable, A; and (3) treatment is *sharply* assigned based on the value of A, specifically $T = 1(A > A^*)$ where A^* is a fixed threshold or cutoff value. [A **fuzzy design** is based on $\text{Prob}(T = 1|A) = F(A)$. The identification problems with fuzzy design are much more complicated than with sharp design. Readers are referred to Van der Klaauw (2002) for further discussion of fuzzy design.] We assume, then, that

$$y = f(A, T) + \varepsilon.$$

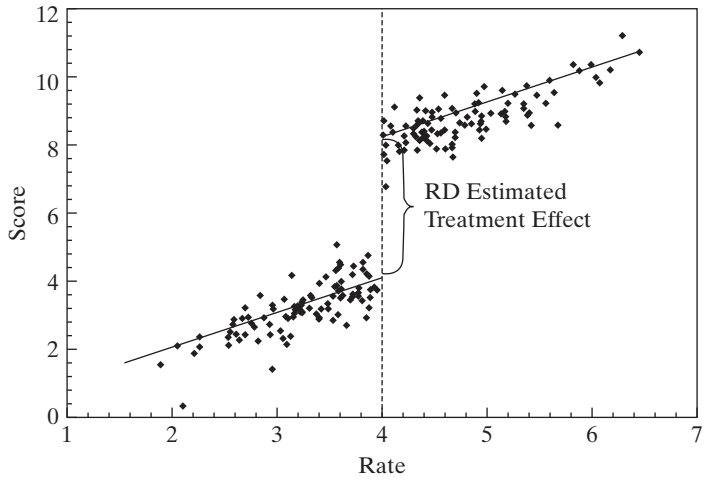
Suppose, for example, the outcome variable is a test score, and that an administrative treatment such as a special education program is funded based on the poverty rates of certain communities. The ideal conditions for a regression discontinuity design based on these assumptions are shown in Figure 6.6. The logic of the calculation is that the points near the threshold value, which have essentially the same stimulus value, constitute a nearly random sample of observations which are segmented by the treatment.

The method requires that $E[\varepsilon|A, T] = E[\varepsilon|A]$—the assignment variable—be exogenous to the experiment. The result in Figure 6.6 is consistent with

$$y = f(A) + \alpha T + \varepsilon,$$

where α will be the treatment effect to be estimated. The specification of $f(A)$ can be problematic; assuming a linear function when something more general is appropriate

Figure 6.6 Regression Discontinuity.



will bias the estimate of α. For this reason, nonparametric methods, such as the LOWESS regression (see Section 12.4), might be attractive. This is likely to enable the analyst to make fuller use of the observations that are more distant from the cutoff point.[17] Identification of the treatment effect begins with the assumption that $f(A)$ is continuous at A^*, so that

$$\lim_{A\uparrow A^*} f(A) = \lim_{A\downarrow A^*} f(A) = f(A^*).$$

Then

$$\lim_{A\downarrow A^*} E[y|A] - \lim_{A\uparrow A^*} E[y|A] = f(A^*) + \alpha + \lim_{A\downarrow A^*} E[\varepsilon|A] - f(A^*) - \lim_{A\uparrow A^*} E[\varepsilon|A]$$
$$= \alpha.$$

With this in place, the treatment effect can be estimated by the difference of the average outcomes for those individuals close to the threshold value, A^*. Details on regression discontinuity design are provided by Trochim (1984, 2000) and Van der Klaauw (2002).

Example 6.13 The Treatment Effect of Compulsory Schooling

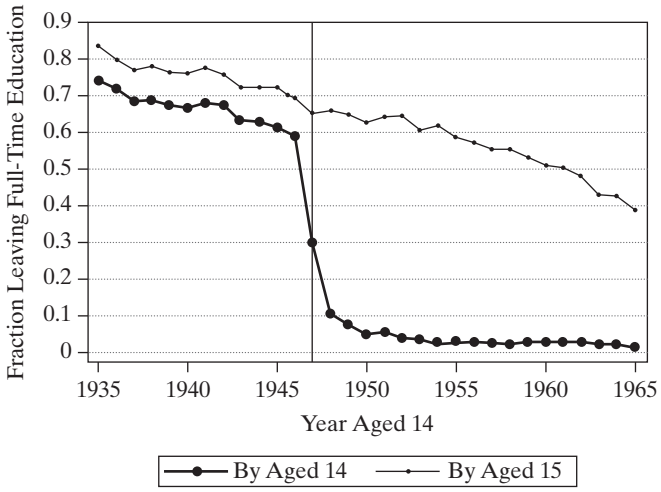
Oreopoulos (2006) examined returns to education in the UK in the context of a discrete change in the national policy on mandatory school attendance. [See, also, Ashenfelter and Krueger (2010b) for a U.S. study.] In 1947, the minimum school-leaving age in Great Britain was changed from 14 to 15 years. In this period, from 1935 to 1960, the exit rate among those old enough in the UK was more than 50%, so the policy change would affect a significant number of students. For those who turned 14 in 1947, the policy would induce a mandatory increase in years of schooling for many students who would otherwise have dropped out. Figure 6.7 (composed from Figures 1 and 6 from the article) shows the quite stark impact of the policy change. (A similar regime change occurred in Northern Ireland in 1957.) A regression of the log of annual earnings that includes a control for birth cohort reveals a distinct break for those born in 1933, that is, those who were affected by the policy change in 1947. The estimated regression produces a return to compulsory schooling of about 7.9% for Great Britain and 11.3% for Northern Ireland. (From Table 2. The figures given are based on least squares regressions. Using instrumental variables produces results of about 14% and 18%, respectively.)

Example 6.14 Interest Elasticity of Mortgage Demand

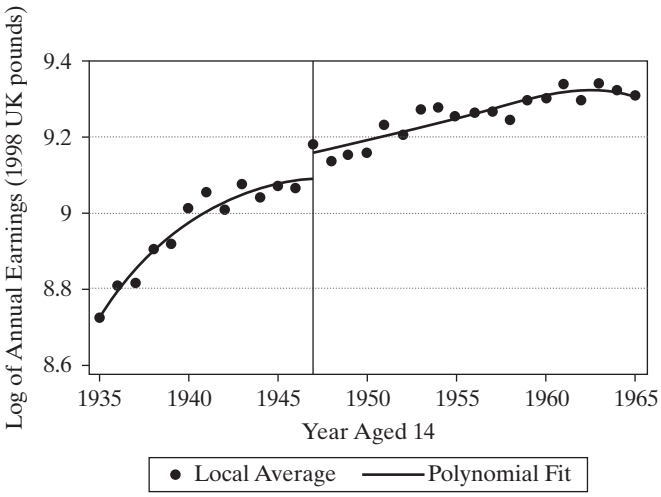
DeFusco and Paciorek (2014, 2016) studied the interest rate elasticity of the demand for mortgages. There is a natural segmentation in this market imposed by the maximum limit on loan sizes eligible for purchase by the Government Sponsored Enterprises (GSEs), Fannie Mae and Freddie Mac. The limits, set by the Federal Housing Finance Agency, vary by housing type and have been adjusted over time. The current loan limit, called the *conforming loan limit (CLL)* for single family homes has been fixed at $417,000 since 2006. A loan that is larger than the CLL is labeled a "jumbo loan." Because the GSEs are able to obtain an implicit subsidy in capital markets, there is a discrete jump in interest rates at the conforming loan limit. The relationship between the mortgage size and the interest rates is key to the specification of the denominator of the elasticity. This foregoing suggests a regression discontinuity approach to the relationship between mortgage rates and loan sizes, such as shown in the left panel

[17] See Van der Klaauw (2002).

Figure 6.7 Regression Discontinuity Design for Returns to Schooling.



Note: The lower line shows the proportion of British-born adults aged 32 to 64 from the 1983 to 1998 General Household Surveys who report leaving full-time education at or before age 14 from 1935 to 1965. The upper line shows the same, but for age 15. The minimum school leaving age in Great Britain changed in 1947 from 14 to 15.



Note: Local averages are plotted for British-born adults aged 32 to 64 from the 1983 to 1998 General Household Surveys. The curved line shows the predicted fit from regressing average log annual earnings on a birth cohort quartic polynomial and an indicator for the school leaving age faced at age 14. The school leaving age increased from 14 to 15 in 1947, indicated by the vertical line. Earnings are measured in 1998 UK pounds using the UK retail price index.

Figure 6.8 Regression Discontinuity Design for Mortgage Demand.

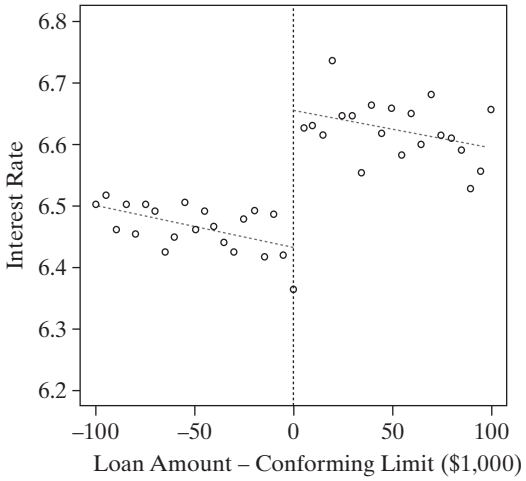


FIG. 2.—Mean Interest Rate Relative to the Conforming Limit, Fixed-Rate Mortgages Only (2006). This figure plots the mean interest rate for fixed rate mortgages originated in 2006 as a function of the loan amount relative to the conforming limit. Each dot represents the mean interest rate within a given $5,000 bin relative to the limit. The dashed lines are predicted values from a regression fit to the binned data allowing for changes in the slope and intercept at the conforming limit. Sample includes all loans in the LPS fixed-rate sample that fall within $100,000 of the conforming limit. See text for details on sample construction.

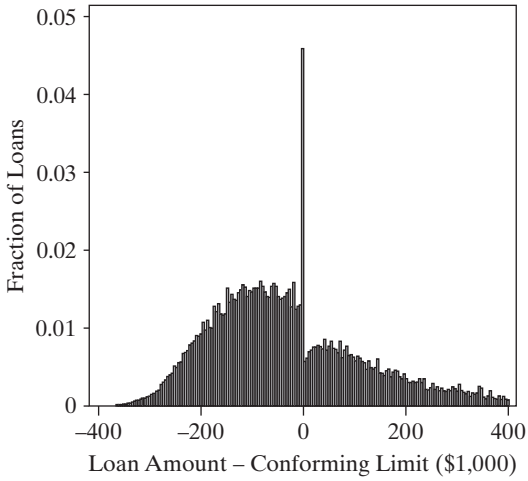


FIG. 3.—Loan Size Distribution Relative to the Conforming Limit. This figure plots the fraction of all loans that are in any given $5,000 bin relative to the conforming limit. Data are pooled across years and each loan is centered at the conforming limit in effect at the date of origination, so that a value of 0 represents a loan at exactly the conforming limit. Sample includes all transactions in the primary DataQuick sample that fall within $400,000 of the conforming limit. See text for details on sample construction.

of Figure 6.8. [Figure 2 in DeFusco and Paciorek (2014).] The semiparametric regression proposed was as follows:

$$r_{i,t} = \alpha_{z(i),t} + \beta J_{i,t} + f^{J=0}(m_{i,t}) + f^{J=1}(m_{i,t}) + s^{LTV}(LTV_{it}) + s^{DTI}(DTI_{i,t}) + s^{FICO}(FICO_{i,t}) + PMI_{i,t} + PP_{i,t} + g(TERM_{i,t}) + \varepsilon_{i,t}.$$

The variables in the specification are:

$r_{i,t}$ = interest rate on loan i originated at time t,
$\alpha_{Z(i),t}$ = fixed effect for zip code and time,
J = dummy variable for jumbo loan ($J=1$) or conforming loan ($J=0$),
$m_{i,t}$ = *size* of the mortgage,
$f^{J=0}$ = $(1-J) \times$ cubic polynomial in the mortgage size,
$f^{J=1}$ = $J \times$ cubic polynomial in the mortgage size,
$LTV_{i,t}$ = loan to value ratio,
$DTI_{i,t}$ = debt to income ratio,
$FICO_{i,t}$ = credit score of borrower,
$PMI_{i,t}$ = dummy variable for whether borrower took out private mortgage insurance,
$PP_{i,t}$ = dummy variable for whether mortgatge has a prepayment penalty,
$TERM_{i,t}$ = control for the length of the mortgage.

A coefficient of interest is β which is the estimate of the jumbo, conforming loan spread. Estimates obtained in this study were roughly 16 basis points. A complication for obtaining the numerator of the elasticity (the response of the mortgage amount) is that the crucial variable *J* is endogenous in the model. This is suggested by the bunching of observations at the CLL that can be seen in the right panel of Figure 6.8. Essentially, individuals who would otherwise take out a jumbo loan near the boundary can take advantage of the lower rate by taking out a slightly smaller mortgage. The implication is that the unobservable characteristics of many individuals who are conforming loan borrowers are those of individuals who are in principle jumbo loan borrowers. The authors consider a semiparametric approach and an instrumental variable approach suggested by Kaufman (2012) (we return to this in Chapter 8) rather than a simple RD approach. (Results are obtained using both approaches.) The instrumental variable used is an indicator related to the appraised home value; the exogeneity of the indicator is argued because home buyers cannot control the appraisal of the home. In the terms developed for IVs in Chapter 8, the instrumental variable is certainly exogenous as it is not controlled by the borrower, and is certainly relevant through the correlation between the appraisal and the size of the mortgage. The main empirical result in the study is an estimate of the interest elasticity of the loan demand, which appears to be measurable at the loan limit. A further complication of the computation is that the increase in the cost of the loan at the loan limit associated with the interest rate increase is not marginal. The increased cost associated the increased interest rate is applied to the entire mortgage, not just the amount by which it exceeds the loan limit. Accounting for that aspect of the computation, the authors obtain estimates of the semi-elasticity ranging from -0.016 to -0.052. They find, for an example, that this suggests an increase in rates from 5% to 6% (a 20% increase) attends a 2% to 3% decrease in demand.

6.5 NONLINEARITY IN THE VARIABLES

It is useful at this point to write the linear regression model in a very general form: Let $\mathbf{z} = z_1, z_2, \ldots, z_L$ be a set of L independent variables; let $f_1, f_2, \ldots, f_K$ be K linearly independent functions of $\mathbf{z}$; let $g(y)$ be an observable function of y; and retain the usual assumptions about the disturbance. The linear regression model may be written

$$\begin{aligned} g(y) &= \beta_1 f_1(\mathbf{z}) + \beta_2 f_2(\mathbf{z}) + \cdots + \beta_K f_K(\mathbf{z}) + \varepsilon \\ &= \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_K x_K + \varepsilon \\ &= \boldsymbol{x}'\boldsymbol{\beta} + \varepsilon. \end{aligned} \tag{6-4}$$

By using logarithms, exponentials, reciprocals, transcendental functions, polynomials, products, ratios, and so on, this linear model can be tailored to any number of situations.

6.5.1 FUNCTIONAL FORMS

A commonly used form of regression model is the **loglinear model**,

$$\ln y = \ln \alpha + \sum_k \beta_k \ln X_k + \varepsilon = \beta_1 + \sum_k \beta_k x_k + \varepsilon.$$

In this model, the coefficients are elasticities:

$$\left(\frac{\partial y}{\partial X_k}\right)\left(\frac{X_k}{y}\right) = \frac{\partial \ln y}{\partial \ln X_k} = \beta_k. \tag{6-5}$$

In the loglinear equation, measured changes are in proportional or percentage terms; β_k measures the percentage change in y associated with a one percent change in X_k. This removes the units of measurement of the variables from consideration in using the regression model. For example, in Example 6.2, in our analysis of auction prices of Monet paintings, we found an elasticity of price with respect to area of 1.34935. (This is an extremely large value—the value well in excess of 1.0 implies that not only do sale prices rise with area, they rise considerably faster than area.)

An alternative approach sometimes taken is to measure the variables and associated changes in standard deviation units. If the data are standardized before estimation using $x_{ik}^* = (x_{ik} - \bar{x}_k)/s_k$ and likewise for y, then the least squares regression coefficients measure changes in standard deviation units rather than natural units or percentage terms. (Note that the constant term disappears from this regression.) It is not necessary actually to transform the data to produce these results; multiplying each least squares coefficient b_k in the original regression by s_k/s_y produces the same result.

A hybrid of the linear and loglinear models is the **semilog equation**

$$\ln y = \beta_1 + \beta_2 x + \varepsilon. \tag{6-6}$$

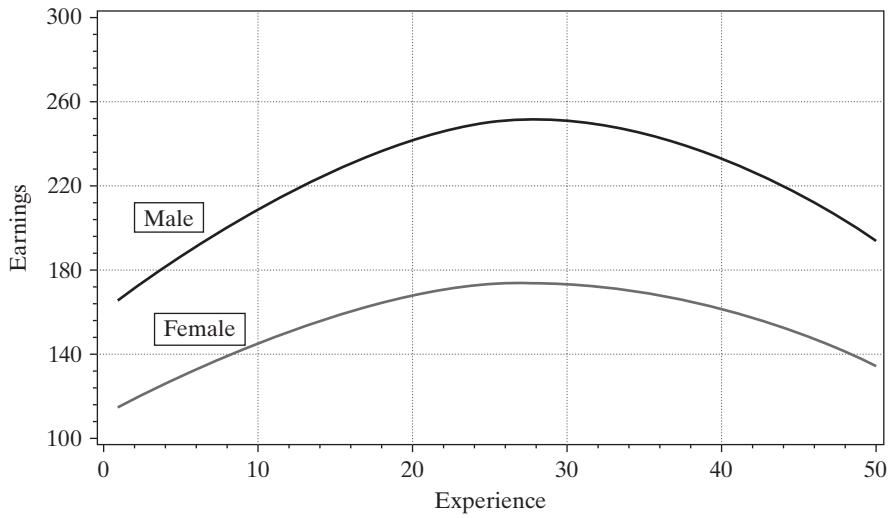
In a semilog equation with a time trend, $d \ln y/dt = \beta_2$ is the average rate of growth of y. The estimated values of 0.0750 and 0.0709 for day fees and boarding fees reported in Table 6.9 suggests that over the full estimation period, after accounting for all other factors, the average rate of growth of the fees was about 7% per year.

The coefficients in the semilog model are partial- or semi-elasticities; in (6-6), β_2 is $\partial \ln y/\partial x$. This is a natural form for models with dummy variables such as the earnings equation in Example 6.1. The coefficient on *Kids* of −0.35 suggests that all else equal, earnings are *approximately* 35% less when there are children in the household.

Example 6.15 Quadratic Regression

The quadratic earnings equation in Example 6.3 shows another use of nonlinearities in the variables. Using the results in Example 6.3, we find that the experience-wage profile appears as in Figure 6.8. This figure suggests an important question in this framework. It is tempting to conclude that Figure 6.8 shows the earnings trajectory of a person as experience accumulates. (The distinctive downturn is probably exaggerated by the use of a quadratic regression rather than a more flexible function.) But that is not what the data provide. The model is based on a cross section, and what it displays is the earnings of different people with different experience levels. How this profile relates to the expected earnings path of one individual is a different, and complicated, question.

Figure 6.9 Experience-Earnings Profile.



6.5.2 INTERACTION EFFECTS

Another useful formulation of the regression model is one with interaction terms. For example, the model for ln *Wage* in Example 6.3 might be extended to allow different partial effects of education for men and women with

$$\ln Wage = \beta_1 ED + \beta_2 FEM + \beta_3 ED \times FEM + \ \ldots \ + \varepsilon.$$

In this model,

$$\frac{\partial E[\ln Wage | ED, FEM, \ldots]}{\partial ED} = \beta_1 + \beta_3 FEM,$$

which implies that the **marginal effect** of education differs between men and women (assuming that β_3 is not zero).[18] If it is desired to form confidence intervals or test hypotheses about these marginal effects, then the necessary standard error is computed from

$$Var\left(\frac{\partial \hat{E}[\ln Wage | ED, FEM, \ldots]}{\partial ED}\right) = Var[\hat{\beta}_1] + FEM^2\, Var[\hat{\beta}_3] + 2FEM\, Cov[\hat{\beta}_1, \hat{\beta}_3].$$

(Because *FEM* is a dummy variable, $FEM^2 = FEM$.) The calculation is similar for

$$\begin{aligned}\Delta E[\ln Wage | ED, FEM, \ldots] &= E[\ln Wage | ED, FEM = 1, \ldots] - E[\ln Wage | ED, FEM = 0, \ldots] \\ &= \beta_2 + \beta_3 ED.\end{aligned}$$

[18] See Ai and Norton (2004) and Greene (2010) for further discussion of partial effects in models with interaction terms.

Example 6.16 Partial Effects in a Model with Interactions

We have extended the model in Example 6.3 by adding an interaction term between *FEM* and *ED*. The results for this part of the expanded model are

$$\ln Wage = \ldots + \underset{(0.00588)}{0.05250\, ED} - \underset{(0.15207)}{0.69799\, FEM} + \underset{(0.01055)}{0.02572\, ED \times FEM} + \ldots$$

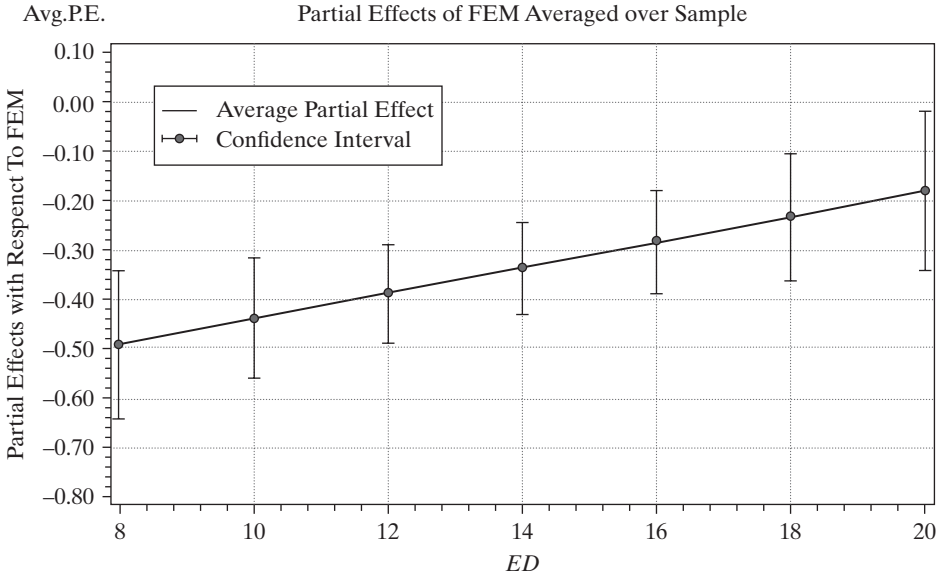
$$Est.Asy.Cov\begin{bmatrix} b_1 \\ b_2 \\ b_3 \end{bmatrix} = \begin{bmatrix} 0.0000345423 & & \\ 0.000349259 & 0.0231247 & \\ -0.0000243829 & -0.00152425 & 0.000111355 \end{bmatrix}.$$

The individual coefficients are not informative about the marginal impact of gender or education. The mean value of *ED* in the full sample is 12.8. The partial effect of a year increase in *ED* is 0.05250 (0.00588) for men and $0.05250 + 0.02572 = 0.07823$ (0.00986) for women. The gender difference in earnings is $-0.69799 + 0.02572 \times ED$. At the mean value of *ED*, this is -0.36822. The standard error would be $(0.0231247 + 12.8^2(0.000111355) - 2(12.8)(0.00152425))^{1/2} = 0.04846$. A convenient way to summarize the information is a plot of the gender difference for the different values of *ED*, as in Figure 6.10. The figure reveals a richer interpretation of the model produced by the nonlinearity—the gender difference in wages is persistent, but does diminish at higher levels of education.

6.5.3 IDENTIFYING NONLINEARITY

If the functional form is not known a priori, then there are a few approaches that may help to identify any nonlinearity and provide some information about it from the sample. For example, if the suspected nonlinearity is with respect to a single regressor in the equation, then fitting a quadratic or cubic polynomial rather than a linear function may capture some of it. The residuals from a plot of the estimated function can also help to reveal the appropriate functional form.

Figure 6.10 Partial Effects in a Nonlinear Model.



Example 6.17 Functional Form for a Nonlinear Cost Function

In a pioneering study of economies of scale in the U.S. electric power industry, Nerlove (1963) analyzed the production costs of 145 American electricity generating companies. Economies of scale are typically modeled as a characteristic of the production function. Nerlove chose a Cobb–Douglas function to model output as a function of capital, K, labor, L, and fuel, F:

$$Q = \alpha_0 K^{\alpha_K} L^{\alpha_L} F^{\alpha_F} e^{\varepsilon},$$

where Q is output and ε_i embodies the unmeasured differences across firms. The economies of scale parameter is $r = \alpha_K + \alpha_L + \alpha_F$. The value 1.0 indicates constant returns to scale. The production model is loglinear, so assuming that other conditions of the classical regression model are met, the four parameters could be estimated by least squares. But, for a firm that optimizes by choosing its factors of production, the demand for fuel would be $F^* = F^*(Q, P_K, P_L, P_F)$ and likewise for labor and capital. The three factor demands are endogenous and the assumptions of the classical model are violated.

In the regulatory framework in place at the time, state commissions set rates and firms met the demand forthcoming at the regulated prices. Thus, it was argued that output (as well as the factor prices) could be viewed as exogenous to the firm. Based on an argument by Zellner, Kmenta, and Dreze (1966), Nerlove argued that at equilibrium, the *deviation* of costs from the long-run optimum would be independent of output. The firm's objective was cost minimization subject to the constraint of the production function. This can be formulated as a Lagrangean problem,

$$\text{Min}_{K,L,F} P_K K + P_L L + P_F F + \lambda(Q - \alpha_0 K^{\alpha_K} L^{\alpha_L} F^{\alpha_F}).$$

The solution to this minimization problem is the three factor demands and the multiplier (which measures marginal cost). Inserted back into total costs, this produces a loglinear cost function,

$$P_K K + P_L L + P_F F = C(Q, P_K, P_L, P_F) = rAQ^{1/r} P_K^{\alpha_K/r} P_L^{\alpha_L/r} P_F^{\alpha_F/r} e^{\varepsilon/r},$$

or

$$\ln C = \beta_1 + \beta_q \ln Q + \beta_K \ln P_K + \beta_L \ln P_L + \beta_F \ln P_F + u, \qquad \textbf{(6-7)}$$

where $\beta_q = 1/(\alpha_K + \alpha_L + \alpha_F)$ is now the parameter of interest and $\beta_j = \alpha_j/r$, $j = K, L, F$.

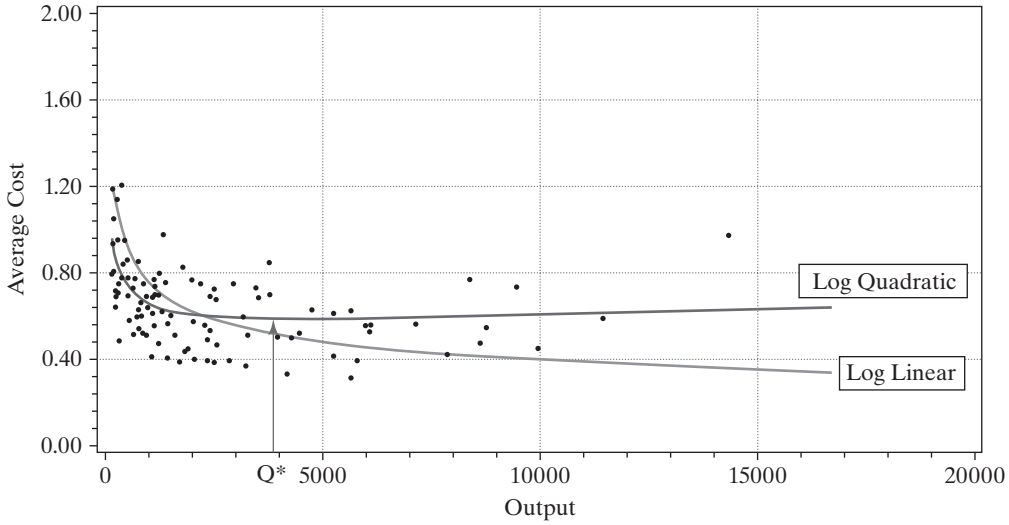
The cost parameters must sum to one; $\beta_K + \beta_L + \beta_F = 1$. This restriction can be imposed by regressing $\ln(C/P_F)$ on a constant, $\ln Q$, $\ln(P_K/P_F)$, and $\ln(P_L/P_F)$. Nerlove's results appear at the left of Table 6.10.[19] The hypothesis of constant returns to scale can be firmly rejected. The t ratio is $(0.721-1)/0.0174 = -16.03$, so we conclude that this estimate is significantly less than 1 or, by implication, r is significantly greater than 1. Note that the coefficient on the capital price is negative. In theory, this should equal α_K/r, which should be positive. Nerlove attributed this to measurement error in the capital price variable. The residuals in a plot of the average costs against the fitted loglinear cost function as in Figure 6.11 suggested that the Cobb-Douglas model was not picking up the increasing average costs at larger outputs, which would suggest diminished economies of scale. An approach used was to expand the cost function to include a quadratic term in log output. This approach corresponds to a more general model. Again, a simple t test strongly suggests that increased generality is called for; $t = 0.051/0.00054 = 9.44$. The output elasticity in this quadratic model is $\beta_q + 2\gamma_{qq} \log Q$. There are economies of scale when this value is less than 1 and constant returns to scale when it equals 1. Using the two values given in the table (0.152 and 0.0052, respectively), we

[19]Nerlove's data appear in Appendix Table F6.2. Figure 6.6 is constructed by computing the fitted log cost values using the means of the logs of the input prices. The plot then uses observations 31–145.

TABLE 6.10 Cobb–Douglas Cost Functions for log (C/P_F) based on 145 observations

	Log-linear			*Log-quadratic*		
Sum of squares	21.637			13.248		
R^2	0.932			0.958		
Variable	*Coefficient*	*Standard Error*	*t Ratio*	*Coefficient*	*Standard Error*	*t Ratio*
Constant	−4.686	0.885	−5.29	−3.764	0.702	−5.36
ln Q	0.721	0.0174	41.4	0.152	0.062	2.45
$\ln^2 Q$	0.000	0.000	—	0.051	0.0054	9.44
ln (P_L/P_F)	0.594	0.205	2.90	0.481	0.161	2.99
ln (P_K/P_F)	−0.0085	0.191	−0.045	0.074	0.150	0.49

Figure 6.11 Estimated Cost Functions.



find that this function does, indeed, produce a U-shaped average cost curve with minimum at ln $Q^* = (1 - 0.152)/(2 \times 0.051) = 8.31$, or $Q = 4079$. This is roughly in the middle of the range of outputs for Nerlove's sample of firms.

6.5.4 INTRINSICALLY LINEAR MODELS

The loglinear model illustrates a nonlinear regression model. The equation is **intrinsically linear**, however. By taking logs of $Y_i = \alpha X_i^{\beta_2} e^{\varepsilon_i}$, we obtain

$$\ln Y_i = \ln \alpha + \beta_2 \ln X_i + \varepsilon_i$$

or

$$y_i = \beta_1 + \beta_2 x_i + \varepsilon_i.$$

Although this equation is linear in most respects, something has changed in that it is no longer linear in α. But, written in terms of β_1, we obtain a fully linear model. That may not be the form of interest, but nothing is lost because β_1 is just $\ln \alpha$. If β_1 can be estimated, then the obvious estimator of α is $\hat{\alpha} = \exp(b_1)$.

This fact leads us to a useful aspect of intrinsically linear models; they have an "invariance property." Using the nonlinear least squares procedure described in the next chapter, we could estimate α and β_2 directly by minimizing the sum of squares function:

$$\text{Minimize with respect to } (\alpha, \beta_2) : S(\alpha, \beta_2) = \sum_{i=1}^{n} (\ln Y_i - \ln \alpha - \beta_2 \ln X_i)^2. \quad \textbf{(6-8)}$$

This is a complicated mathematical problem because of the appearance of the term $\ln \alpha$. However, the equivalent linear least squares problem,

$$\text{Minimize with respect to } (\beta_1, \beta_2) : S(\beta_1, \beta_2) = \sum_{i=1}^{n} ,(y_i - \beta_1 - \beta_2 x_i)^2, \quad \textbf{(6-9)}$$

is simple to solve with the least squares estimator we have used up to this point. The invariance feature that applies is that the two sets of results will be numerically identical; we will get the identical result from estimating α using (6-8) and from using $\exp(\beta_1)$ from (6-9). By exploiting this result, we can broaden the definition of linearity and include some additional cases that might otherwise be quite complex.

DEFINITION 6.1 Intrinsic Linearity
In the linear regression model, if the K parameters $\beta_1, \beta_2, \ldots, \beta_K$ can be written as K one-to-one, possibly nonlinear functions of a set of K underlying parameters $\theta_1, \theta_2, \ldots, \theta_K$, then the model is intrinsically linear in θ.

Example 6.18 Intrinsically Linear Regression

In Section 14.6.4, we will estimate by maximum likelihood the parameters of the model

$$f(y|\beta, x) = \frac{(\beta + x)^{-\rho}}{\Gamma(\rho)} y^{\rho-1} e^{-y/(\beta+x)}.$$

In this model, $E[y|x] = (\beta\rho) + \rho x$, which suggests another way that we might estimate the two parameters. This function is an intrinsically linear regression model, $E[y|x] = \beta_1 + \beta_2 x$, in which $\beta_1 = \beta\rho$ and $\beta_2 = \rho$. We can estimate the parameters by least squares and then retrieve the estimate of β using b_1/b_2. Because this value is a nonlinear function of the estimated parameters, we use the delta method to estimate the standard error. Using the data from that example,[20] the least squares estimates of β_1 and β_2 (with standard errors in parentheses) are -4.1431 (23.734) and 2.4261 (1.5915). The estimated covariance is -36.979. The estimate of β is $-4.1431/2.4261 = -1.708$. We estimate the sampling variance of $\hat{\beta}$ with

$$\begin{aligned}\text{Est. Var}[\hat{\beta}] &= \left(\frac{\partial\hat{\beta}}{\partial b_1}\right)^2 \widehat{Var}[b_1] + \left(\frac{\partial\hat{\beta}}{\partial b_2}\right)^2 \widehat{Var}[b_2] + 2\left(\frac{\partial\hat{\beta}}{\partial b_1}\right)\left(\frac{\partial\hat{\beta}}{\partial b_2}\right)\widehat{Cov}[b_1, b_2] \\ &= 8.689^2.\end{aligned}$$

[20] The data are given in Appendix Table FC.1.

TABLE 6.11 Estimates of the Regression in a Gamma Model: Least Squares versus Maximum Likelihood

	$\boldsymbol{\beta}$		ρ	
	Estimate	***Standard Error***	***Estimate***	***Standard Error***
Least squares	−1.708	8.689	2.426	1.592
Maximum likelihood	−4.719	2.345	3.151	0.794

Table 6.11 compares the least squares and maximum likelihood estimates of the parameters. The lower standard errors for the maximum likelihood estimates result from the inefficient (equal) weighting given to the observations by the least squares procedure. The gamma distribution is highly skewed. In addition, we know from our results in Appendix C that this distribution is an exponential family. We found for the gamma distribution that the sufficient statistics for this density were $\Sigma_i y_i$ and $\Sigma_i \ln y_i$. The least squares estimator does not use the second of these, whereas an efficient estimator will.

The emphasis in intrinsic linearity is on "one to one." If the conditions are met, then the model can be estimated in terms of the functions $\beta_1, \ldots, \beta_K$, and the underlying parameters derived after these are estimated. The one-to-one correspondence is an **identification condition**. If the condition is met, then the underlying parameters of the regression ($\boldsymbol{\theta}$) are said to be **exactly identified** in terms of the parameters of the linear model $\boldsymbol{\beta}$. An excellent example is provided by Kmenta (1986, p. 515).

Example 6.19 CES Production Function

The constant elasticity of substitution production function may be written

$$\ln y = \ln \gamma - \frac{\nu}{\rho} \ln [\delta K^{-\rho} + (1 - \delta) L^{-\rho}] + \varepsilon. \tag{6-10}$$

A Taylor series approximation to this function around the point $\rho = 0$ is

$$\begin{aligned} \ln y &= \ln \gamma + \nu\delta \ln K + \nu(1 - \delta) \ln L + \rho\nu\delta(1 - \delta) \left\{ -\tfrac{1}{2} [\ln K - \ln L]^2 \right\} + \varepsilon' \\ &= \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \varepsilon', \end{aligned} \tag{6-11}$$

where $x_1 = 1$, $x_2 = \ln K$, $x_3 = \ln L$, $x_4 = -\frac{1}{2} \ln^2 (K/L)$, and the transformations are

$$\begin{aligned} &\beta_1 = \ln \gamma, \quad \beta_2 = \nu\delta, \quad \beta_3 = \nu(1 - \delta), \quad \beta_4 = \rho\nu\delta(1 - \delta), \\ &\gamma = e^{\beta_1}, \quad \delta = \beta_2/(\beta_2 + \beta_3), \quad \nu = \beta_2 + \beta_3, \quad \rho = \beta_4(\beta_2 + \beta_3)/(\beta_2\beta_3). \end{aligned} \tag{6-12}$$

Estimates of β_1, β_2, β_3, and β_4 can be computed by least squares. The estimates of γ, δ, ν, and ρ obtained by the second row of (6-12) are the same as those we would obtain had we found the nonlinear least squares estimates of (6-11) directly. [As Kmenta shows, however, they are not the same as the nonlinear least squares estimates of (6-10) due to the use of the Taylor series approximation to get to (6-11).] We would use the delta method to construct the estimated asymptotic covariance matrix for the estimates of $\boldsymbol{\theta}' = [\gamma, \delta, \nu, \rho]$. The derivatives matrix is

$$\mathbf{C} = \frac{\partial \boldsymbol{\theta}}{\partial \boldsymbol{\beta}'} = \begin{bmatrix} e^{\beta_1} & 0 & 0 & 0 \\ 0 & \beta_3/(\beta_2 + \beta_3)^2 & -\beta_2/(\beta_2 + \beta_3)^2 & 0 \\ 0 & 1 & 1 & 0 \\ 0 & -\beta_3\beta_4/(\beta_2^2\beta_3) & -\beta_2\beta_4/(\beta_2\beta_3^2) & (\beta_2 + \beta_3)/(\beta_2\beta_3) \end{bmatrix}.$$

The estimated covariance matrix for $\hat{\boldsymbol{\theta}}$ is $\hat{\mathbf{C}} \{ Asy.Var[\hat{\theta}] \} \hat{\mathbf{C}}'$.

Not all models of the form

$$y_i = \beta_1(\theta)x_{i1} + \beta_2(\theta)x_{i2} + \cdots + \beta_K(\theta)x_{ik} + \varepsilon_i \quad \textbf{(6-13)}$$

are intrinsically linear. Recall that the condition that the functions be one to one (i.e., that the parameters be exactly identified) was required. For example,

$$y_i = \alpha + \beta x_{i1} + \gamma x_{i2} + \beta\gamma x_{i3} + \varepsilon_i$$

is nonlinear. The reason is that if we write it in the form of (6-13), we fail to account for the condition that β_4 equals $\beta_2\beta_3$, which is a **nonlinear restriction**. In this model, the three parameters α, β, and γ are **overidentified** in terms of the four parameters β_1, β_2, β_3, and β_4. Unrestricted least squares estimates of β_2, β_3, and β_4 can be used to obtain two estimates of each of the underlying parameters, and there is no assurance that these will be the same. Models that are not intrinsically linear are treated in Chapter 7.

6.6 STRUCTURAL BREAK AND PARAMETER VARIATION

One of the more common applications of hypothesis testing is in tests of **structural change**.[21] In specifying a regression model, we assume that its assumptions apply to all the observations in the sample. It is straightforward, however, to test the hypothesis that some or all of the regression coefficients are different in different subsets of the data. To analyze an example, we will revisit the data on the U.S. gasoline market that we examined in Examples 2.3 and 4.2. As Figure 4.2 suggests, this market behaved in predictable, unremarkable fashion prior to the oil shock of 1973 and was quite volatile thereafter. The large jumps in price in 1973 and 1980 are clearly visible, as is the much greater variability in consumption. It seems unlikely that the same regression model would apply to both periods.

6.6.1 DIFFERENT PARAMETER VECTORS

The gasoline consumption data span two very different periods. Up to 1973, fuel was plentiful and world prices for gasoline had been stable or falling for at least two decades. The embargo of 1973 marked a transition in this market, marked by shortages, rising prices, and intermittent turmoil. It is possible that the entire relationship described by the regression model changed in 1974. To test this as a hypothesis, we could proceed as follows: Denote the first 21 years of the data in $\mathbf{y}$ and $\mathbf{X}$ as $\mathbf{y}_1$ and $\mathbf{X}_1$ and the remaining years as $\mathbf{y}_2$ and $\mathbf{X}_2$. An unrestricted regression that allows the coefficients to be different in the two periods is

$$\begin{bmatrix}\mathbf{y}_1\\ \mathbf{y}_2\end{bmatrix} = \begin{bmatrix}\mathbf{X}_1 & \mathbf{0}\\ \mathbf{0} & \mathbf{X}_2\end{bmatrix}\begin{bmatrix}\boldsymbol{\beta}_1\\ \boldsymbol{\beta}_2\end{bmatrix} + \begin{bmatrix}\boldsymbol{\varepsilon}_1\\ \boldsymbol{\varepsilon}_2\end{bmatrix}. \quad \textbf{(6-14)}$$

Denoting the data matrices as $\mathbf{y}$ and $\mathbf{X}$, we find that the unrestricted least squares estimator is

$$\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} = \begin{bmatrix}\mathbf{X}_1'\mathbf{X}_1 & \mathbf{0}\\ \mathbf{0} & \mathbf{X}_2'\mathbf{X}_2\end{bmatrix}^{-1}\begin{bmatrix}\mathbf{X}_1'\mathbf{y}_1\\ \mathbf{X}_2'\mathbf{y}_2\end{bmatrix} = \begin{bmatrix}\mathbf{b}_1\\ \mathbf{b}_2\end{bmatrix}, \quad \textbf{(6-15)}$$

[21] This test is often labeled a **Chow test**, in reference to Chow (1960).

which is least squares applied to the two equations separately. Therefore, the total sum of squared residuals from this regression will be the sum of the two residual sums of squares from the two separate regressions:

$$\mathbf{e}'\mathbf{e} = \mathbf{e}_1'\mathbf{e}_1 + \mathbf{e}_2'\mathbf{e}_2.$$

The restricted coefficient vector can be obtained by imposing a constraint on least squares. Formally, the restriction $\boldsymbol{\beta}_1 = \boldsymbol{\beta}_2$ is $\mathbf{R}\boldsymbol{\beta} = \mathbf{q}$, where $\mathbf{R} = [\mathbf{I}:-\mathbf{I}]$ and $\mathbf{q} = \mathbf{0}$. The general result given earlier can be applied directly. An easy way to proceed is to build the restriction directly into the model. If the two coefficient vectors are the same, then (6-14) may be written

$$\begin{bmatrix}\mathbf{y}_1\\ \mathbf{y}_2\end{bmatrix} = \begin{bmatrix}\mathbf{X}_1\\ \mathbf{X}_2\end{bmatrix}\boldsymbol{\beta} + \begin{bmatrix}\boldsymbol{\varepsilon}_1\\ \boldsymbol{\varepsilon}_2\end{bmatrix};$$

the restricted estimator can be obtained simply by stacking the data and estimating a single regression. The residual sum of squares from this restricted regression, $\mathbf{e}_*'\mathbf{e}_*$, then forms the basis for the test.

We begin by assuming that the disturbances are homoscedastic, nonautocorrelated, and normally distributed. More general cases are considered in the next section. Under these assumptions, the test statistic is given in (5-29), where J, the number of restrictions, is the number of columns in $\mathbf{X}_2$ and the denominator degrees of freedom is $n_1 + n_2 - 2K$. For this application,

$$F[K, n_1 + n_2 - 2K] = \frac{(\mathbf{e}_*'\mathbf{e}_* - \mathbf{e}_1'\mathbf{e}_1 - \mathbf{e}_2'\mathbf{e}_2)/K}{(\mathbf{e}_1'\mathbf{e}_1 + \mathbf{e}_2'\mathbf{e}_2)/(n_1 + n_2 - 2K)}. \tag{6-16}$$

Example 6.20 Structural Break in the Gasoline Market

Figure 4.2 shows a plot of prices and quantities in the U.S. gasoline market from 1953 to 2004. The first 21 points are the layer at the bottom of the figure and suggest an orderly market. The remainder clearly reflect the subsequent turmoil in this market. We will use the Chow tests described to examine this market. The model we will examine is the one suggested in Example 2.3, with the addition of a time trend:

$$\ln(G/Pop)_t = \beta_1 + \beta_2 \ln (Income/Pop)_t + \beta_3 \ln PG_t + \beta_4 \ln PNC_t + \beta_5 \ln PUC_t + \beta_6 t + \varepsilon_t.$$

The three prices in the equation are for *G*, new cars and used cars. *Income*/*Pop* is per capita Income, and *G*/*Pop* is per capita gasoline consumption. The time trend is computed as t = Year−1952, so in the first period $t = 1$. Regression results for three functional forms are shown in Table 6.12. Using the data for the entire sample, 1953 to 2004, and for the two subperiods, 1953 to 1973 and 1974 to 2004, we obtain the three estimated regressions in the first and last two columns. Using the full set of 52 observations to fit the model, the sum of squares is $\mathbf{e}_*'\mathbf{e}_* = 0.101997$. The *F* statistic for testing the restriction that the coefficients in the two equations are the same is

$$F[6, 40] = \frac{(0.101997 - (0.00202244 + 0.007127899))/6}{(0.00202244 + 0.007127899)/(21 + 31 - 12)} = 67.645.$$

The tabled critical value is 2.336, so, consistent with our expectations, we would reject the hypothesis that the coefficient vectors are the same in the two periods.

TABLE 6.12 Gasoline Consumption Functions

Coefficients	*1953–2004*	*1953–1973*	*1974–2004*
Constant	−26.6787	−22.1647	−15.3238
ln *Income/Pop*	1.6250	0.8482	0.3739
ln *PG*	−0.05392	−0.03227	−0.1240
ln *PNC*	−0.08343	0.6988	−0.001146
ln *PUC*	−0.08467	−0.2905	−0.02167
Year	−0.01393	0.01006	0.004492
R^2	0.9649	0.9975	0.9529
Standard error	0.04709	0.01161	0.01689
Sum of squares	0.101997	0.00202244	0.007127899

6.6.2 ROBUST TESTS OF STRUCTURAL BREAK WITH UNEQUAL VARIANCES

An important assumption made in using the Chow test is that the disturbance variance is the same in both (or all) regressions. In the restricted model, if this is not true, the first n_1 elements of ε have variance σ_1^2, whereas the next n_2 have variance σ_2^2, and so on. The restricted model is, therefore, heteroscedastic, and the results for normally distributed disturbances no longer apply. In several earlier examples, we have gone beyond heteroscedasticity, and based inference on robust specifications that also accommodate clustering and correlation across observations. In both settings, the results behind the F statistic in (6-16) will no longer apply. As analyzed by Schmidt and Sickles (1977), Ohtani and Toyoda (1985), and Toyoda and Ohtani (1986), it is quite likely that the actual probability of a type I error will be larger than the significance level we have chosen. (That is, we shall regard as large an F statistic that is actually less than the *appropriate* but unknown critical value.) Precisely how severe this effect is going to be will depend on the data and the extent to which the variances differ, in ways that are not likely to be obvious.

If the sample size is reasonably large, then we have a test that is valid whether or not the disturbance variances are the same. Suppose that $\hat{\boldsymbol{\theta}}_1$ and $\hat{\boldsymbol{\theta}}_2$ are two consistent and asymptotically normally distributed estimators of a parameter based on independent samples, with asymptotic covariance matrices $\mathbf{V}_1$ and $\mathbf{V}_2$. Then, under the null hypothesis that the true parameters are the same,

$$\hat{\boldsymbol{\theta}}_1 - \hat{\boldsymbol{\theta}}_2 \text{ has mean } \mathbf{0} \text{ and asymptotic covariance matrix } \mathbf{V}_1 + \mathbf{V}_2.$$

Under the null hypothesis, the Wald statistic,

$$W = (\hat{\boldsymbol{\theta}}_1 - \hat{\boldsymbol{\theta}}_2)'(\hat{\mathbf{V}}_1 + \hat{\mathbf{V}}_2)^{-1}(\hat{\boldsymbol{\theta}}_1 - \hat{\boldsymbol{\theta}}_2), \tag{6-17}$$

has a limiting chi-squared distribution with K degrees of freedom. A test that the difference between the parameters is zero can be based on this statistic.[22] It is straightforward to apply this to our test of common parameter vectors in our regressions. Large values of the statistic lead us to reject the hypothesis.

In a small or moderately sized sample, the Wald test has the unfortunate property that the probability of a type I error is persistently larger than the critical level we

[22] See Andrews and Fair (1988). The true size of this suggested test is uncertain. It depends on the nature of the alternative. If the variances are radically different, the assumed critical values might be somewhat unreliable.

use to carry it out. (That is, we shall too frequently reject the null hypothesis that the parameters are the same in the subsamples.) We should be using a larger critical value. Ohtani and Kobayashi (1986) have devised a "bounds" test that gives a partial remedy for the problem. In general, this test attains its validity in relatively large samples.

Example 6.21 Sample Partitioning by Gender

Example 6.3 considers the labor market experiences of a panel of 595 individuals, each observed 7 times. We have observed persistent differences between men and women in the relationship of log wages to various variables. It might be the case that different models altogether would apply to the two subsamples. We have fit the model in Example 6.3 separately for men and women (omitting *FEM* from the two regressions, of course), and calculated the Wald statistic in (6-17) based on the cluster corrected asymptotic covariance matrices as used in the pooled model as well. The chi-squared statistic with 17 degrees of freedom is 27.587, so the hypothesis of equal parameter vectors is rejected. The sums of squared residuals for the pooled data set for men and for women, respectively, are 416.988, 360.773, and 24.0848; the F statistic is 20.287 with critical value 1.625. This produces the same conclusion.

Example 6.22 The World Health Report

The 2000 version of the World Health Organization's (WHO) *World Health Report* contained a major country-by-country inventory of the world's health care systems. [World Health Organization (2000). See also http://www.who.int/whr/en/.] The book documented years of research and has thousands of pages of material. Among the most controversial and most publicly debated parts of the report was a single chapter that described a comparison of the delivery of health care by 191 countries—nearly all of the world's population. [Evans et al. (2000a,b). See, e.g., Hilts (2000) for reporting in the popular press.] The study examined the efficiency of health care delivery on two measures: the standard one that is widely studied, (disability adjusted) life expectancy (DALE), and an innovative new measure created by the authors that was a composite of five outcomes (COMP) and that accounted for efficiency and fairness in delivery. The regression-style modeling, which was done in the setting of a frontier model (see Section 19.2.4), related health care attainment to two major inputs, education and (per capita) health care expenditure. The residuals were analyzed to obtain the country comparisons.

The data in Appendix Table F6.3 were used by the researchers at the WHO for the study. (They used a panel of data for the years 1993 to 1997. We have extracted the 1997 data for this example.) The WHO data have been used by many researchers in subsequent analyses.[23] The regression model used by the WHO contained DALE or COMP on the left-hand side and health care expenditure, education, and education squared on the right. Greene (2004b) added a number of additional variables such as per capita GDP, a measure of the distribution of income, and World Bank measures of government effectiveness and democratization of the political structure.

Among the controversial aspects of the study was the fact that the model aggregated countries of vastly different characteristics. A second striking aspect of the results, suggested in Hilts (2000) and documented in Greene (2004b), was that, in fact, the "efficient" countries in the study were the 30 relatively wealthy OECD members, while the rest of the world on average fared much more poorly. We will pursue that aspect here with respect to DALE. Analysis of COMP is left as an exercise. Table 6.8 presents estimates of the regression models for DALE for the pooled sample, the OECD countries, and the non-OECD countries, respectively. Superficially, there do not appear to be very large differences across the two subgroups. We first tested the joint significance of the additional variables, income distribution (GINI), per

[23] See, for example, Hollingsworth and Wildman (2002), Gravelle et al. (2002), and Greene (2004b).

TABLE 6.13 Regression Results for Life Expectancy

	All Countries		*OECD*		*Non-OECD*	
Constant	25.237	38.734	42.728	49.328	26.816	41.408
Health exp	0.00629	−0.00180	0.00268	0.00114	0.00955	−0.00178
Education	7.931	7.178	6.177	5.156	7.0433	6.499
Education^2	−0.439	−0.426	−0.385	−0.329	−0.374	−0.372
Gini coeff		−17.333		−5.762		−21.329
Tropic		−3.200		−3.298		−3.144
Pop. Dens.		−0.255e−4		0.000167		−0.425e−4
Public exp		−0.0137		−0.00993		−0.00939
PC GDP		0.000483		0.000108		0.000600
Democracy		1.629		−0.546		1.909
Govt. Eff.		0.748		1.224		0.786
R^2	0.6824	0.7299	0.6483	0.7340	0.6133	0.6651
Std. Err.	6.984	6.565	1.883	1.916	7.366	7.014
Sum of sq.	9121.795	7757.002	92.21064	69.74428	8518.750	7378.598
N	191		30		161	
GDP/Pop	6609.37		18199.07		4449.79	
F test	4.524		0.874		3.311	

capita GDP, and so on. For each group, the *F* statistic is $[(\mathbf{e}_*'\mathbf{e}_* - \mathbf{e}'\mathbf{e})/7]/[\mathbf{e}'\mathbf{e}/(n - 11)]$. These *F* statistics are shown in the last row of the table. The critical values for *F*[7,180] (all), *F*[7,19] (OECD), and *F*[7,150] (non-OECD) are 2.061, 2.543, and 2.071, respectively. We conclude that the additional explanatory variables are significant contributors to the fit for the non-OECD countries (and for all countries), but not for the OECD countries. Finally, to conduct the structural change test of OECD vs. non-OECD, we computed

$$F[11, 169] = \frac{[7757.002 - (69.74428 + 7378.598)]/11}{(69.74428 + 7378.598)/(191 - 11 - 11)} = 0.637.$$

The 95% critical value for *F*[11,169] is 1.846. So, we do not reject the hypothesis that the regression model is the same for the two groups of countries. The Wald statistic in (6-17) tells a different story. The statistic is 35.221. The 95% critical value from the chi-squared table with 11 degrees of freedom is 19.675. On this basis, we would reject the hypothesis that the two coefficient vectors are the same.

6.6.3 POOLING REGRESSIONS

Extending the homogeneity test to multiple groups or periods should be straightforward. As usual, we begin with independent and identically normally distributed disturbances. Assume there are G groups or periods. (In Example 6.3, we are examining 7 years of observations.) The direct extension of the *F* statistic in (6-16) would be

$$F[(G - 1)K, \Sigma_{g=1}^{G}(n_g - K)] = \frac{(\mathbf{e}_*'\mathbf{e}_* - \Sigma_{g=1}^{G}\mathbf{e}_g'\mathbf{e}_g)/(G - 1)K}{(\Sigma_{g=1}^{G}\mathbf{e}_g'\mathbf{e}_g)/\Sigma_{g=1}^{G}(n_g - K)}. \quad \textbf{(6-18)}$$

To apply (6-18) to a more general case, begin with the simpler setting of possible heteroscedasticity. Then, we can consider a set of G estimators, $\mathbf{b}_g$, each with associated

asymptotic covariance matrix $\mathbf{V}_g$. A Wald test along the lines of (6-17) can be carried out by testing $H_0: \boldsymbol{\beta}_1 - \boldsymbol{\beta}_2 = \mathbf{0}, \boldsymbol{\beta}_1 - \boldsymbol{\beta}_3 = \mathbf{0}, \ldots, \boldsymbol{\beta}_1 - \boldsymbol{\beta}_G = \mathbf{0}$. This can be based on G sets of least squares results. The Wald statistic is

$$W = (\mathbf{Rb})'(\mathbf{R}(Asy.\ Var[\mathbf{b}])\mathbf{R}')^{-1}(\mathbf{Rb}), \tag{6-19}$$

where

$$\mathbf{R} = \begin{bmatrix} \mathbf{I} & -\mathbf{I} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{I} & \mathbf{0} & -\mathbf{I} & \cdots & \mathbf{0} \\ \mathbf{I} & \mathbf{0} & \mathbf{0} & \cdots & \mathbf{0} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ \mathbf{I} & \mathbf{0} & \mathbf{0} & \cdots & -\mathbf{I} \end{bmatrix}; \mathbf{b} = \begin{pmatrix} \mathbf{b}_1 \\ \mathbf{b}_2 \\ \cdots \\ \mathbf{b}_G \end{pmatrix}. \tag{6-20}$$

The results in (6-19) and (6-20) are straightforward based on G separate regressions. For example, to test equality of the coefficient vectors for three periods, (6-19) and (6-20) would produce

$$W = [(\mathbf{b}_1 - \mathbf{b}_2)' \quad (\mathbf{b}_1 - \mathbf{b}_3)'] \begin{bmatrix} (\mathbf{V}_1 + \mathbf{V}_2) & \mathbf{V}_1 \\ \mathbf{V}_1 & (\mathbf{V}_1 + \mathbf{V}_3) \end{bmatrix}^{-1} \begin{bmatrix} (\mathbf{b}_1 - \mathbf{b}_2) \\ (\mathbf{b}_1 - \mathbf{b}_3) \end{bmatrix}.$$

The computations are rather more complicated when observations are correlated, as in a panel. In Example 6.3, we are examining seven periods of data but robust calculation of the covariance matrix for the estimates results in correlation across the observations within a group. The implication for current purposes would be that we are not using independent samples for the G estimates of β_g. The following practical strategy for this computation is suggested for the particular application—extensions to other settings should be straightforward. We have seven years of data for individual i, with regression specification

$$y_{it} = \mathbf{x}_{it}'\boldsymbol{\beta} + \varepsilon_{it}.$$

For each individual, we construct

$$\widetilde{\mathbf{X}}_i = \begin{bmatrix} \mathbf{x}_{i1}' & \mathbf{0}' & \cdots & \mathbf{0}' \\ \mathbf{0}' & \mathbf{x}_{i2}' & \cdots & \mathbf{0}' \\ \cdots & \cdots & \cdots & \cdots \\ \mathbf{0}' & \mathbf{0}' & \cdots & \mathbf{x}_{i7}' \end{bmatrix} \quad \text{and} \quad \begin{pmatrix} y_{i1} \\ y_{i2} \\ \cdots \\ y_{i7} \end{pmatrix}.$$

Then, the $7K \times 1$ vector of estimated coefficient vectors is computed by least squares,

$$\mathbf{b} = \left[\sum_{i=1}^{595} \widetilde{\mathbf{X}}_i'\widetilde{\mathbf{X}}_i \right]^{-1} \sum_{i=1}^{595} \widetilde{\mathbf{X}}_i'\widetilde{y}_i$$

The estimator of the asymptotic covariance matrix of $\mathbf{b}$ is the cluster estimator from (4-41) and (4-42),

$$Est.Asy.Var[\mathbf{b}] = \left[\sum_{i=1}^{595} \widetilde{\mathbf{X}}_i'\widetilde{\mathbf{X}}_i \right]^{-1} \left\{ \sum_{i=1}^{595} (\widetilde{\mathbf{X}}_i'\mathbf{e}_i)(\mathbf{e}_i'\widetilde{\mathbf{X}}_i) \right\} \left[\sum_{i=1}^{595} \widetilde{\mathbf{X}}_i'\widetilde{\mathbf{X}}_i \right]^{-1}. \tag{6-21}$$

Example 6.23 Pooling in a Log Wage Model

Using the data and model in Example 6.3, the sums of squared residuals are as follows:

1976: 44.3242	1977: 38.7594	1978: 63.9203	1979: 61.4599
1980: 54.9996	1981: 58.6650	1982: 62.9827	Pooled: 513.767

The F statistic based on (6-18) is 14.997. The 95% critical value from the F table with 6×12 and (4165-84) degrees of freedom is 1.293. The large sample approximation for this statistic would be 72(14.997) = 1079.776 with 72 degrees of freedom. The 95% critical value for the chi-squared distribution with 72 degrees of freedom is 92.808, which is slightly less than 72(1.293). The Wald statistic based on (6-19) using (6-21) to compute the asymptotic covariance matrix is 3068.78 with 72 degrees of freedom. Finally, the Wald statistic based on (6-19) and 7 separate estimates, allowing different variances, is 1478.62. All versions of the test procedure produce the same conclusion. The homogeneity restriction is decisively rejected. We note, this conclusion gives no indication of the nature of the change from year to year.

6.7 SUMMARY AND CONCLUSIONS

This chapter has discussed the functional form of the regression model. We examined the use of dummy variables and other transformations to build nonlinearity into the model to accommodate specific features of the environment, such as the effects of discrete changes in policy. We then considered other nonlinear models in which the parameters of the nonlinear model could be recovered from estimates obtained for a linear regression. The final sections of the chapter described hypothesis tests designed to reveal whether the assumed model had changed during the sample period, or was different for different groups of observations.

Key Terms and Concepts

- Binary variable
- Chow test
- Control group
- Control observations
- Difference in differences
- Dummy variable
- Dummy variable trap
- Dynamic linear regression model
- Exactly identified
- Fuzzy design
- Identification condition
- Interaction terms
- Intrinsically linear
- Loglinear model
- Marginal effect
- Natural experiment
- Nonlinear restriction
- Overidentified
- Placebo effect
- Regression discontinuity design
- Regression kink design
- Response
- Semilog equation
- Structural change
- Treatment
- Treatment group
- Unobserved heterogeneity

Exercises

1. A regression model with $K = 16$ independent variables is fit using a panel of seven years of data. The sums of squares for the seven separate regressions and the pooled regression are shown below. The model with the pooled data allows a separate constant for each year. Test the hypothesis that the same coefficients apply in every year.

	2004	*2005*	*2006*	*2007*	*2008*	*2009*	*2010*	*All*
Observations	65	55	87	95	103	87	78	570
$\mathbf{e}'\mathbf{e}$	104	88	206	144	199	308	211	1425

2. *Reverse regression*. A method of analyzing statistical data to detect discrimination in the workplace is to fit the regression

$$y = \alpha + \mathbf{x}'\beta + \gamma d + \varepsilon, \tag{1}$$

where y is the wage rate and d is a dummy variable indicating either membership ($d = 1$) or nonmembership ($d = 0$) in the class toward which it is suggested the discrimination is directed. The regressors $\mathbf{x}$ include factors specific to the particular type of job as well as indicators of the qualifications of the individual. The hypothesis of interest is H_0:$\gamma \geq 0$ versus H_1:$\gamma < 0$. The regression seeks to answer the question, "In a given job, are individuals in the class ($d = 1$) paid less than equally qualified individuals not in the class ($d = 0$)?" Consider an alternative approach. Do individuals in the class in the same job as others, and receiving the same wage, uniformly have higher qualifications? If so, this might also be viewed as a form of discrimination. To analyze this question, Conway and Roberts (1983) suggested the following procedure:

   1. Fit (1) by ordinary least squares. Denote the estimates a, $\mathbf{b}$, and c.
   2. Compute the set of qualification indices,

$$\mathbf{q} = a\mathbf{i} + \mathbf{Xb}. \tag{2}$$

      Note the omission of $c\mathbf{d}$ from the fitted value.
   3. Regress $\mathbf{q}$ on a constant, $\mathbf{y}$ and $\mathbf{d}$. The equation is

$$\mathbf{q} = \alpha_* + \beta_*\mathbf{y} + \gamma_*\mathbf{d} + \varepsilon_*. \tag{3}$$

      The analysis suggests that if $\gamma < 0$, then $\gamma_* > 0$.

   a. Prove that the theory notwithstanding, the least squares estimates c and c^* are related by

$$c_* = \frac{(\bar{y}_1 - \bar{y})(1 - R^2)}{(1 - P)(1 - r_{yd}^2)} - c, \tag{4}$$

   where

$\bar{y}_1$ = mean of y for observations with $d = 1$,
$\bar{y}$ = mean of y for all observations,
P = mean of d,
R^2 = coefficient of determination for (1),
r_{yd}^2 = squared correlation between y and d.

   [*Hint:* The model contains a constant term]. Thus, to simplify the algebra, assume that all variables are measured as deviations from the overall sample means and use a partitioned regression to compute the coefficients in (3). Second, in (2), use the result that based on the least squares results $\mathbf{y} = a\mathbf{i} + \mathbf{Xb} + c\mathbf{d} + \mathbf{e}$, so $\mathbf{q} = \mathbf{y} - c\mathbf{d} - \mathbf{e}$. From here on, we drop the constant term. Thus, in the regression in (3) you are regressing $[\mathbf{y} - c\mathbf{d} - \mathbf{e}]$ on $\mathbf{y}$ and $\mathbf{d}$.

   b. Will the sample evidence necessarily be consistent with the theory? [*Hint:* Suppose that $c = 0$.]

   A symposium on the Conway and Roberts paper appeared in the *Journal of Business and Economic Statistics* in April 1983.

3. *Reverse regression continued*. This and the next exercise continue the analysis of Exercise 2. In Exercise 2, interest centered on a particular dummy variable in which the regressors were accurately measured. Here we consider the case in which the

crucial regressor in the model is measured with error. The paper by Kamlich and Polachek (1982) is directed toward this issue.

Consider the simple errors in the variables model,

$$y = \alpha + \beta x^* + \varepsilon, \qquad x = x^* + u,$$

where u and ε are uncorrelated and x is the erroneously measured, observed counterpart to x*.

a. Assume that x^*, u, and ε are all normally distributed with means μ^*, 0, and 0, variances σ_*^2, σ_u^2, and σ_ε^2, and zero covariances. Obtain the probability limits of the least squares estimators of α and β.
b. As an alternative, consider regressing x on a constant and y, and then computing the reciprocal of the estimate. Obtain the probability limit of this estimator.
c. Do the "direct" and "reverse" estimators bound the true coefficient?

4. *Reverse regression continued*. Suppose that the model in Exercise 3 is extended to $y = \beta x^* + \gamma d + \varepsilon, x = x^* + u$. For convenience, we drop the constant term. Assume that x^*, ε, and u are independent normally distributed with zero means. Suppose that d is a random variable that takes the values one and zero with probabilities π and $1 - \pi$ in the population and is independent of all other variables in the model. To put this formulation in context, the preceding model (and variants of it) have appeared in the literature on discrimination. We view y as a "wage" variable, x^* as "qualifications," and x as some imperfect measure such as education. The dummy variable, d, is membership ($d = 1$) or nonmembership ($d = 0$) in some protected class. The hypothesis of discrimination turns on $\gamma < 0$ versus $\gamma \geq 0$.
   a. What is the probability limit of c, the least squares estimator of γ, in the least squares regression of y on x and d? [*Hints:* The independence of x^* and d is important. Also, plim $\mathbf{d}'\mathbf{d}/n = \text{Var}[d] + E^2[d] = \pi(1 - \pi) + \pi^2 = \pi$. This minor modification does not affect the model substantively, but it greatly simplifies the algebra.] Now suppose that x^* and d are not independent. In particular, suppose that $E[x^* | d = 1] = \mu^1$ and $E[x^* | d = 0] = \mu^0$. Repeat the derivation with this assumption.
   b. Consider, instead, a regression of x on y and d. What is the probability limit of the coefficient on d in this regression? Assume that x^* and d are independent.
   c. Suppose that x^* and d are not independent, but γ is, in fact, less than zero. Assuming that both preceding equations still hold, what is estimated by $(\bar{y} | d = 1) - (\bar{y} | d = 0)$? What does this quantity estimate if γ does equal zero?

5. *Dummy variable for one observation*. Suppose the data set consists of n observations, $(\mathbf{y}_n, \mathbf{X}_n)$ and an additional observation, $(\text{y}_s, \mathbf{x}_s')$. The full data set contains a dummy variable, $\mathbf{d}$, that equals zero save for one (the last) observation. Then, the full data set is

$$(\mathbf{X}_{n,s}, \mathbf{d}_{n,s}) = \begin{bmatrix} \mathbf{X}_n & \mathbf{0} \\ \mathbf{x}_s' & \mathbf{1} \end{bmatrix} \text{ and } \mathbf{y}_{n,s} = \begin{bmatrix} \mathbf{y}_n \\ y_s \end{bmatrix}.$$

It is claimed in the text that in the *full* regression of $\mathbf{y}_{n,s}$ on $(\mathbf{X}_{n,s}, \mathbf{d}_{n,s})$ using all $n+1$ observations, the slopes on $\mathbf{X}_{n,s}$, $\mathbf{b}_{n,s}$, and their estimated standard errors will be the same as those on $\mathbf{X}_n$, $\mathbf{b}_n$ in the *short* regression of $\mathbf{y}_n$ on $\mathbf{X}_n$, and the sum of squared residuals in the full regression will be the same as the sum of squared residuals in

the short regression. That is, the last observation will be ignored. However, the R^2 in the full regression will not be the same as the R^2 in the short regression. Prove these results.

Applications

1. In Application 1 in Chapter 3 and Application 1 in Chapter 5, we examined Koop and Tobias's data on wages, education, ability, and so on. We continue the analysis here. (The source, location and configuration of the data are given in the earlier application.) We consider the model

$$\begin{aligned}\ln Wage = {} & \beta_1 + \beta_2 Educ + \beta_3 Ability + \beta_4 Experience \\ & + \beta_5 Mother's\ education + \beta_6 Father's\ education + \beta_7 Broken\ home \\ & + \beta_8 Siblings + \varepsilon.\end{aligned}$$

   a. Compute the full regression by least squares and report your results. Based on your results, what is the estimate of the marginal value, in $/hour, of an additional year of education, for someone who has 12 years of education when all other variables are at their means and *Broken home* = 0?
   b. We are interested in possible nonlinearities in the effect of education on ln *Wage*. (Koop and Tobias focused on experience. As before, we are not attempting to replicate their results.) A histogram of the education variable shows values from 9 to 20, a spike at 12 years (high school graduation), and a second at 15. Consider aggregating the education variable into a set of dummy variables:

$$\begin{aligned} HS &= 1 \text{ if } Educ \leq 12, 0 \text{ otherwise} \\ Col &= 1 \text{ if } Educ > 12 \text{ and } Educ \leq 16, 0 \text{ otherwise} \\ Grad &= 1 \text{ if } Educ > 16, 0 \text{ otherwise.}\end{aligned}$$

   Replace *Educ* in the model with (*Col, Grad*), making high school (*HS*) the base category, and recompute the model. Report all results. How do the results change? Based on your results, what is the marginal value of a college degree? What is the marginal impact on ln *Wage* of a graduate degree?
   c. The aggregation in part b actually loses quite a bit of information. Another way to introduce nonlinearity in education is through the function itself. Add $Educ^2$ to the equation in part a and recompute the model. Again, report all results. What changes are suggested? Test the hypothesis that the quadratic term in the equation is not needed—that is, that its coefficient is zero. Based on your results, sketch a profile of log wages as a function of education.
   d. One might suspect that the value of education is enhanced by greater ability. We could examine this effect by introducing an interaction of the two variables in the equation. Add the variable

$$Educ\_Ability = Educ \times Ability$$

   to the base model in part a. Now, what is the marginal value of an additional year of education? The sample mean value of ability is 0.052374. Compute a confidence interval for the marginal impact on ln *Wage* of an additional year of education for a person of average ability.

e. Combine the models in c and d. Add both $Educ^2$ and *Educ_Ability* to the base model in part a and reestimate. As before, report all results and describe your findings. If we define *low ability* as less than the mean and *high ability* as greater than the mean, the sample averages are −0.798563 for the 7,864 low-ability individuals in the sample and +0.717891 for the 10,055 high-ability individuals in the sample. Using the formulation in part c, with this new functional form, sketch, describe, and compare the log wage profiles for low- and high-ability individuals.

2. (An extension of Application 1.) Here we consider whether different models as specified in Application 1 would apply for individuals who reside in "Broken homes." Using the results in Section 6.6, test the hypothesis that the same model (not including the *Broken home* dummy variable) applies to both groups of individuals, those with *Broken home* = 0 and with *Broken home* = 1.
3. In Solow's classic (1957) study of technical change in the U.S. economy, he suggests the following aggregate production function: $q(t) = A(t)f[k(t)]$, where $q(t)$ is aggregate output per work hour, $k(t)$ is the aggregate capital labor ratio, and $A(t)$ is the technology index. Solow considered four static models, $q/A = \alpha + \beta \ln k$, $q/A = \alpha - \beta/k$, $\ln(q/A) = \alpha + \beta \ln k$, and $\ln(q/A) = \alpha + \beta/k$. Solow's data for the years 1909 to 1949 are listed in Appendix Table F6.4.
   a. Use these data to estimate the α and β of the four functions listed above. (*Note:* Your results will not quite match Solow's. See the next exercise for resolution of the discrepancy.)
   b. In the aforementioned study, Solow states:

      > A scatter of q / A against k is shown in Chart 4. Considering the amount of a priori doctoring which the raw figures have undergone, the fit is remarkably tight. Except, that is, for the layer of points which are obviously too high. These maverick observations relate to the seven last years of the period, 1943–1949. From the way they lie almost exactly parallel to the main scatter, one is tempted to conclude that in 1943 the aggregate production function simply shifted.

      Compute a scatter diagram of q / A against k and verify the result he notes above.
   c. Estimate the four models you estimated in the previous problem including a dummy variable for the years 1943 to 1949. How do your results change? (*Note:* These results match those reported by Solow, although he did not report the coefficient on the dummy variable.)
   d. Solow went on to surmise that, in fact, the data were fundamentally different in the years before 1943 than during and after. Use a Chow test to examine the difference in the two subperiods using your four functional forms. Note that with the dummy variable, you can do the test by introducing an interaction term between the dummy and whichever function of k appears in the regression. Use an F test to test the hypothesis.