CHAPTER 9

Problems in Regression Analysis

9.1 MULTICOLLINEARITY

Multicollinearity refers to the case in which two or more explanatory variables in the regression model are highly correlated, making it difficult or impossible to isolate their individual effects on the dependent variable. With multicollinearity, the estimated OLS coefficients may be statistically insignificant (and even have the wrong sign) even though R^2 may be "high." Multicollinearity can sometimes be overcome or reduced by collecting more data, by utili zing a priori information, by transforming the functional relationship (see Prob. 9.3), or by dropping one of the highly collinear variables.

EXAMPLE 1. Table 9.1 gives the growth rate of imports Y, gross domestic product X_1, and inflation X_2 for the United States from 1985 to 1999 (the reason for using growth rates is explained in Chap. 11). It is expected that the level of imports will be greater as GDP and domestic prices increase. Regressing Y on X_1 and X_2, we get

$$\hat{Y} = 0.0015 + 1.39X_1 + 0.09X_2 \qquad R^2 = 0.42$$
$$(1.46) \quad (1.85) \qquad r_{12} = 0.38$$

Table 9.1 Growth Rate of Imports, GDP and Inflation in the United States from 1985 to 1999

Year	1985	1986	1987	1988	1989	1990	1991	1992
Y	0.0540	0.0656	0.1475	0.0686	0.0455	0.0827	−0.0157	0.0753
X_1	0.0709	0.0505	0.0780	0.0750	0.0627	0.0464	0.0399	0.0640
X_2	−0.1593	−0.2683	0.4801	0.1348	−0.0218	0.1612	−0.2511	−0.2611
Year	1993	1994	1995	1996	1997	1998	1999	
Y	0.0841	0.1540	0.0578	0.0918	0.0949	0.0555	0.1593	
X_1	0.0503	0.0621	0.0432	0.0600	0.0623	0.0585	0.0652	
X_2	0.0527	−0.1500	0.0251	−0.1119	−0.0131	−0.3613	0.2579	

Source: St. Louis Federal Reserve (Bureau of Economic Analysis).

Neither $\hat{b}_1$ nor $\hat{b}_2$ is statistically significant at the 5% level. $\hat{b}_2$ is significant at the 10% level, but the R^2 indicates that 42% of the variation in Y is explained by the model even though none of the independent variables stand out individually. The correlation is positive correlation X_1 and X_2, as indicated by r_{12}. Reestimating the regression without either X_2 or X_1, we get

$$\hat{Y} = -0.04 + \underset{(2.13)}{2.06}\ X_1 \qquad R^2 = 0.26$$

$$\hat{Y} = 0.09 + \underset{(2.48)}{0.11}\ X_2 \qquad R^2 = 0.32$$

In simple regressions, the significance of both X_1 and X_2 increases, with X_1 almost significant at the 5% level and X_2 significant at more than the 5% level, indicating that the original regression exhibited multicollinearity. However, dropping either variable from the regression leads to biased OLS estimates, because economic theory suggests that both GDP and prices should be included in the import function.

9.2 HETEROSCEDASTICITY

If the OLS assumption that the variance of the error term is constant for all observations does not hold, we face the problem of *heteroscedasticity*. This leads to unbiased but inefficient (i.e., larger than minimum variance) estimates of the coefficients, as well as biased estimates of the standard errors (and, thus, incorrect statistical tests and confidence intervals).

One test for heteroscedasticity involves arranging the data from small to large values of the independent variable X and running two regressions, one for small values of X and one for large values, omitting, say, one-fifth of the middle observations. Then, we test that the ratio of the error sum of squares (ESS) of the second regression to the first regression is significantly different from zero, using the F table with $(n - d - 2k)/2$ degrees of freedom, where n is the total number of observations, d is the number of omitted observations, and k is the number of estimated parameters.

If the error variance is proportional to X^2 (often the case), heteroscedasticity can be overcome by dividing every term of the model by X and then reestimating the regression using the transformed variables.

EXAMPLE 2. Table 9.2 gives average wages Y and the number of workers employed X by 30 firms in an industry. Regressing Y on X for the entire sample, we get

$$\hat{Y} = \underset{(40.27)}{7.5} + \underset{(16.10)}{0.009\,X} \qquad R^2 = 0.90$$

The results of regressing Y on X for the first 12 and for the last 12 observations are, respectively

$$\hat{Y} = \underset{(39.4)}{8.1} + \underset{(4.36)}{0.006X} \qquad R^2 = 0.66 \qquad \mathrm{ESS}_1 = 0.507$$

$$\hat{Y} = \underset{(4.16)}{6.1} + \underset{(3.89)}{0.013X} \qquad R^2 = 0.60 \qquad \mathrm{ESS}_2 = 3.095$$

Table 9.2 Average Wages and Number of Workers Employed

Average Wages						Workers Employed
8.40	8.40	8.60	8.70	8.90	9.00	100
8.90	9.10	9.30	9.30	9.40	9.60	200
9.50	9.80	9.90	10.30	10.30	10.50	300
10.30	10.60	10.90	11.30	11.50	11.70	400
11.60	11.80	12.10	12.50	12.70	13.10	500

Since $ESS_2/ESS_1 = 3.095/0.507 = 6.10$ exceeds $F_{10,10} = 2.97$ at the 5% level of significance (see App. 7), the hypothesis of heteroscedasticity is accepted. Reestimating the transformed model to correct for heteroscedasticity, we get

$$\frac{\hat{Y}}{X} = \underset{(14.43)}{0.008} + \underset{(76.58)}{7.8}\left(\frac{1}{X}\right) \qquad R^2 = 0.99$$

Note that the slope coefficient is now given by the intercept (i.e., 0.008), and this is smaller than before the adjustment (i.e., 0.009).

9.3 AUTOCORRELATION

When the error term in one time period is positively correlated with the error term in the previous time period, we face the problem of (positive first-order) *autocorrelation*. This is common in time-series analysis and leads to downward-biased standard errors (and, thus, to incorrect statistical tests and confidence intervals).

The presence of first-order autocorrelation is tested by utilizing the table of the Durbin-Watson statistic (App. 8) at the 5 or 1% levels of significance for n observations and k' explanatory variables. If the calculated value of d from Eq. (*9.1*) is smaller than the tabular value of d_L (lower limit), the hypothesis of positive first-order autocorrelation is accepted:

$$d = \frac{\sum_{t=2}^{n}(e_t - e_{t-1})^2}{\sum_{t=1}^{n} e_t^2} \tag{9.1}$$

The hypothesis is rejected if $d > d_U$ (upper limit), and the test is inconclusive if $d_L < d < d_U$. (For negative autocorrelation, see Prob. 9.8.)

One way to correct for autocorrelation is to first estimate ρ (Greek letter rho) from Eq. (*9.2*)

$$Y_t = b_0(1-\rho) + \rho Y_{t-1} + b_1 X_t - b_1 \rho X_{t-1} + v_t \tag{9.2}$$

and then reestimate the regression on the transformed variables:

$$(Y_t - \hat{\rho} Y_{t-1}) = b_0(1-\hat{\rho}) + b_1(X_t - \hat{\rho} X_{t-1}) + (u_t - \hat{\rho} u_{t-1}) \tag{9.3}$$

To avoid losing the first observation in the differencing process, $Y_1\sqrt{1-\hat{\rho}^2}$ and $X_1\sqrt{1-\hat{\rho}^2}$ are used for the first transformed observations of Y and X, respectively. When $\hat{\rho} \cong 1$, autocorrelation can be corrected by rerunning the regression in difference form and omitting the intercept term (see Prob. 9.12).

EXAMPLE 3. Table 9.3 gives the level of inventories Y and sales S, both in billions of dollars, in U.S. manufacturing from 1979 to 1998. Regressing Y on X, we get

$$\hat{Y}_t = \underset{(16.68)}{126.06 + 1.03X_t} \qquad R^2 = 0.94 \qquad d = 0.58$$

Table 9.3 Inventory and Sales (Both in Billions of Dollars) in U.S. Manufacturing 1979–1998

Year	1979	1980	1981	1982	1983	1984	1985	1986	1987	1988
Y	242	265	283	312	312	340	335	323	338	369
X	144	154	168	163	172	191	194	195	206	225
Year	1989	1990	1991	1992	1993	1994	1995	1996	1997	1998
Y	391	405	391	383	384	405	431	437	456	467
X	237	243	240	250	261	279	300	310	327	338

Source: *Economic Report of the President.*

Since $d = 0.58 < d_L = 1.20$ at the 5% level of significance with $n = 20$ and $k' = 1$ (from App. 8), there is evidence of autocorrelation. An estimate of ρ is given by the coefficient of Y_{t-1} in the following regression:

$$\hat{Y}_t = \underset{}{66.88} + \underset{(3.43)}{0.58}\ Y_{t-1} + \underset{(2.36)}{0.88}\ X_t - \underset{(-1.04)}{0.50}\ X_{t-1} \qquad R^2 = 0.97$$

Utilizing $\hat{\rho} = 0.58$ to transform the original variables (it is a coincidence here that $\hat{\rho} = d$), as in Eq. (9.3), and using $242\sqrt{1 - 0.58^2} = 197.14$ and $144\sqrt{1 - 0.58^2} = 117.30$ for the first transformed observations of Y and X, respectively, we rerun the regression on the transformed variables (denoted by the asterisk) and get

$$\hat{Y}_t^* = \underset{}{65.68} + \underset{(9.34)}{0.94} X_t^* \qquad R^2 = 0.83 \qquad d = 1.78$$

Since now $d = 1.78 > d_U = 1.41$ (from App. 8), there is no evidence of autocorrelation. Note that the t value of X_t^* is less than for X_t (but is still highly significant) and R^2 is also lower.

9.4 ERRORS IN VARIABLES

Errors in variables refer to the case in which the variables in the regression model include measurement errors. Measurement errors in the dependent variable are incorporated into the disturbance term and do not create any special problem. However, errors in the explanatory variables lead to biased and inconsistent parameter estimates.

One method of obtaining consistent OLS parameter estimates is to replace the explanatory variable subject to measurement errors with another variable (called an *instrumental variable*) that is highly correlated with the original explanatory variable but is independent of the error term. This is often difficult to do and somewhat arbitrary. The simplest instrumental variable is usually the lagged explanatory variable in question (see Example 4). Another method used when only X is subject to measurement errors involves regressing X on Y (inverse least squares; see Prob. 9.15).

EXAMPLE 4. Table 9.4 gives inventories Y, actual sales X, and hypothetical values of X that include measurement error X', all in billions of dollars, in U.S. retail trade from 1979 to 1998. X and Y are assumed to be error-free. Regressing Y_t on X_t, we get

$$\hat{Y}_t = \underset{(0.72)}{2.92} + \underset{(56.67)}{1.53}\ X_t \qquad R^2 = 0.99$$

Regressing Y_t on X_t' (if X_t is not available), we get

$$\hat{Y}_t = \underset{(1.70)}{6.78} + \underset{(56.23)}{1.46}\ X_t' \qquad R^2 = 0.99$$

Table 9.4 Inventories and Sales (in Billions of Dollars) in U.S. Retail Trade, 1979–1998

Year	1979	1980	1981	1982	1983	1984	1985	1986	1987	1988
Y	111	121	133	135	148	168	182	187	208	219
X	75	80	87	89	98	107	115	121	128	138
X'	76	82	89	91	100	109	118	124	132	142
Year	1989	1990	1991	1992	1993	1994	1995	1996	1997	1998
Y	237	240	243	252	269	294	310	321	330	341
X	147	154	155	163	174	188	197	209	218	229
X'	152	159	160	169	180	195	204	217	226	238

Source: *Economic Report of the President*.

Note that $\hat{b}_1' < \hat{b}_1$; furthermore, $\hat{b}_1$ falls outside the 95% confidence interval of b_1' (1.40 to 1.51). Using X_{t-1}' as an instrumental variable for X_t' (if X_t' is suspected to be correlated with u_t), we get

$$\hat{Y} = \underset{(2.48)}{13.88} + \underset{(40.19)}{1.50}\, X_{t-1}' \qquad R^2 = 0.99$$

The coefficient on X_{t-1}' is closer to the true one ($\hat{b}_1$ falls in the 95% confidence interval of 1.42 to 1.57), and is consistent.

Solved Problems

MULTICOLLINEARITY

9.1 (*a*) What is meant by *perfect multicollinearity*? What is its effect? (*b*) What is meant by *high*, but not perfect, *multicollinearity*? What problems may result? (*c*) How can multicollinearity be detected? (*d*) What can be done to overcome or reduce the problems resulting from multicollinearity?

(*a*) Two or more independent variables are *perfectly collinear* if one or more of the variables can be expressed as a linear combination of the other variable(s). For example, there is perfect multicollinearity between X_1 and X_2 if $X_1 = 2X_2$ or $X_1 = 5 - (1/3)X_2$. If two or more explanatory variables are perfectly linearly correlated, it will be impossible to calculate OLS estimates of the parameters because the system of normal equations will contain two or more equations that are not independent.

(*b*) *High*, but not perfect, *multicollinearity* refers to the case in which two or more independent variables in the regression model are highly correlated. This may make it difficult or impossible to isolate the effect that each of the highly collinear explanatory variables has on the dependent variable. However, the OLS estimated coefficients are still unbiased (if the model is properly specified). Furthermore, if the principal aim is prediction, multicollinearity is not a problem if the same multicollinearity pattern persists during the forecasted period.

(*c*) The classic case of multicollinearity occurs when none of the explanatory variables in the OLS regression is statistically significant (and some may even have the wrong sign), even though R^2 may be high (say, between 0.7 and 1.0). In the less clearcut cases, detecting multicollinearity may be more difficult. High, simple, or partial correlation coefficients among explanatory variables are sometimes used as a measure of multicollinearity. However, serious multicollinearity can be present even if simple or partial correlation coefficients are relatively low (i.e., less than 0.5).

(*d*) Serious multicollinearity may sometimes be corrected by (1) extending the size of the sample data, (2) utilizing a priori information (e.g., we may know from a previous study that $b_2 = 0.25b_1$), (3) transforming the functional relationship, or (4) dropping one of the highly collinear variables (however, this may lead to specification bias or error if theory tells us that the dropped variable should be included in the model).

9.2 Table 9.5 gives the output in tons Q, the labor input in worker-hours L, and the capital input in machine-hours K, of 15 firms in an industry. (*a*) Fit a Cobb-Douglas production function of the form $Q = b_0 L^{b_1} K^{b_2} e^u$ to the data and find $\bar{R}^2$ and the simple correlation coefficient between $\ln L$ and $\ln K$. (*b*) Regress $\ln Q$ on $\ln L$ only. (*c*) Regress $\ln Q$ on $\ln K$ only. (*d*) What can be concluded from the results with regard to multicollinearity?

Table 9.5 Output, Labor, and Capital Inputs of 15 Firms in an Industry

Firm	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Q	2350	2470	2110	2560	2650	2240	2430	2530	2550	2450	2290	2160	2400	2490	2590
L	2334	2425	2230	2463	2565	2278	2380	2437	2446	2403	2301	2253	2367	2430	2470
K	1570	1850	1150	1940	2450	1340	1700	1860	1880	1790	1480	1240	1660	1850	2000

(*a*) Transforming the data into natural log form as shown in Table 9.6 and then regressing $\ln Q$ on $\ln L$ and $\ln K$, we get

$$\ln Q = 0.50 + 0.76 \ln L + 0.19 \ln K \qquad R^2 = 0.969, \quad \bar{R}^2 = 0.964, \quad r_{\ln L \ln K} = 0.992$$
$$(1.07) \qquad (1.36)$$

(*b*) $$\ln Q = -5.50 + 1.71 \ln L \qquad R^2 = 0.964$$
$$(-7.74) \qquad (18.69)$$

(*c*) $$\ln Q = 5.30 + 0.34 \ln K \qquad R^2 = 0.966$$
$$(4.78) \qquad (19.19)$$

(*d*) Since neither $\hat{b}_1$ nor $\hat{b}_2$ in part *a* is statistically significant at the 5% level (i.e., they have unduly large standard errors) while $R^2 = 0.97$, there is clear indication of serious multicollinearity. Specifically, large firms tend to use both more labor and more capital than do small firms. This is confirmed by the very high value of 0.99 for the simple correlation coefficient between $\ln L$ and $\ln K$. In parts *b* and *c*, simple regressions were reestimated with either $\ln L$ or $\ln K$ as the only explanatory variable. In these simple regressions, both $\ln L$ and $\ln K$ are statistically significant at much more than the 1% level with R^2 exceeding 0.96. However, dropping either $\ln K$ or $\ln L$ from the multiple regression leads to a biased

Table 9.6 Output, Labor, and Capital Inputs in Original and Log Form

Firm	Q	L	K	$\ln Q$	$\ln L$	$\ln K$
1	2350	2334	1570	7.76217	7.75534	7.35883
2	2470	2425	1850	7.81197	7.79359	7.52294
3	2110	2230	1150	7.65444	7.70976	7.04752
4	2560	2463	1940	7.84776	7.80914	7.57044
5	2650	2565	2450	7.88231	7.84971	7.80384
6	2240	2278	1340	7.71423	7.73105	7.20042
7	2430	2380	1700	7.79565	7.77486	7.43838
8	2530	2437	1860	7.83597	7.79852	7.52833
9	2550	2446	1880	7.84385	7.80221	7.53903
10	2450	2403	1790	7.80384	7.78447	7.48997
11	2290	2301	1480	7.73631	7.74110	7.29980
12	2160	2253	1240	7.67786	7.72002	7.12287
13	2400	2367	1660	7.78322	7.76938	7.41457
14	2490	2430	1850	7.82004	7.79565	7.52294
15	2590	2470	2000	7.85941	7.81197	7.60090

OLS slope estimate for the retained variable because economic theory postulates that both labor and capital should be included in the production function.

9.3 How can the multicollinearity difficulty faced in Prob. 9.2 be overcome if it is known that constant returns to scale (i.e., $b_1 + b_2 = 1$) prevail in this industry?

With constant returns to scale, the Cobb-Douglas production function can be rewritten as

$$Q = b_0 L^{b_1} K^{1-b_1} e^u$$

Expressing this production function in double-log form and rearranging it, we get

$$\ln Q = \ln b_0 + b_1 \ln L + (1 - b_1) \ln K + u$$
$$\ln Q - \ln K = \ln b_0 + b_1(\ln L - \ln K) + u$$

Setting $\ln Q^* = \ln Q - \ln K$ and $\ln L^* = \ln L - \ln K$ and then regressing $\ln Q^*$ on $\ln L^*$, we get

$$\ln Q^* = \underset{(9.26)}{0.07} + \underset{(39.81)}{0.83} \ln L^* \qquad R^2 = 0.992$$

Then $\hat{b}_2 = 1 - \hat{b}_1 = 1 - 0.83 = 0.17$.

HETEROSCEDASTICITY

9.4 (*a*) What is meant by *heteroscedasticity*? (*b*) Draw a figure showing homoscedastic disturbances and the various forms of heteroscedastic disturbances. (*c*) Why is heteroscedasticity a problem?

(*a*) *Heteroscedasticity* refers to the case in which the variance of the error term is not constant for all values of the independent variable; that is, $E(X_i u_i) \neq 0$, so $E(u_i)^2 \neq \sigma_u^2$. This violates the third assumption of the OLS regression model (see Prob. 6.4). It occurs primarily in cross-sectional data. For example, the error variance associated with the expenditures of low-income families is usually smaller than for high-income families because most of the expenditures of low-income families are on necessities, with little room for discretion.

(*b*) Figure 9-1*a* shows homoscedastic (i.e., constant variance) disturbances, while Fig. 9-1*b*, *c*, and *d* shows heteroscedastic disturbances. In Fig. 9-1*b*, σ_u^2 increases with X_i. In Fig. 9-1*c*, σ_u^2 decreases with X_i. In Fig. 9-1*d*, σ_u^2 first decreases and then increases as X_i increases. In economics, the heteroscedasticity shown in Fig. 9-1*b* is the most common, so the discussion that follows refers to that.

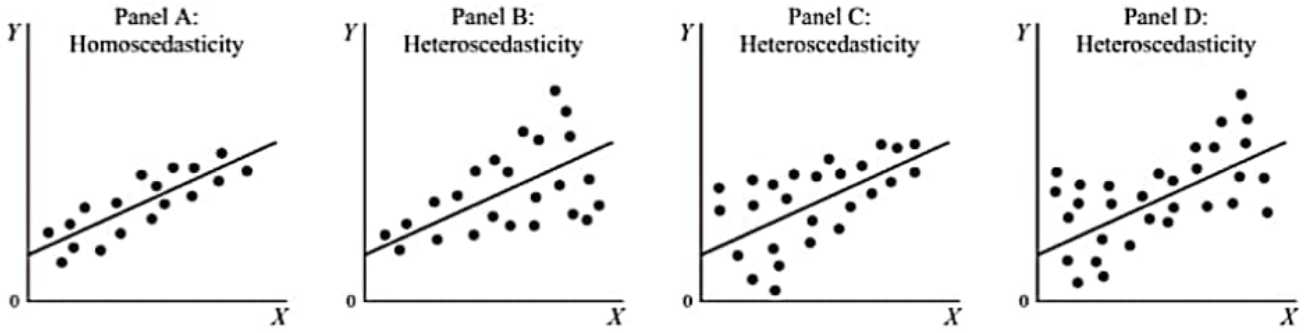

Fig. 9-1

(*c*) With heteroscedasticity, the OLS parameter estimates are still unbiased and consistent, but they are inefficient (i.e., they have larger than minimum variances). Furthermore, the estimated variances of the parameters are biased, leading to incorrect statistical tests for the parameters and biased confidence intervals.

9.5 (*a*) How is the presence of heteroscedasticity tested? (*b*) How can heteroscedasticity be corrected?

(*a*) The presence of heteroscedasticity can be tested by arranging the data from small to large values of the independent variable X_i and then running two separate regressions, one for small values of X_i and one for large values of X_i, omitting some (say, one-fifth) of the middle observations. Then the ratio of the error sum of squares of the second regression to the error sum of squares of the first regression (i.e., ESS_2/ESS_1) is tested to see if it is significantly different from zero. The F distribution is used for this test with $(n-d-2k)/2$ degrees of freedom, where n is the total number of observations, d is the number of omitted observations, and k is the number of estimated parameters. This is the *Goldfeld-Quandt test for heteroscedasticity* and is most appropriate for large samples (i.e., for $n \geq 30$). If no middle observations are omitted, the test is still correct, but it will have a reduced power to detect heteroscedasticity.

(*b*) If it is assumed (as often is the case) that $\operatorname{var} u_i = CX_i^2$, where C is a nonzero constant, we can correct for heteroscedasticity by dividing (i.e., weighting) every term of the regression by X_i and then reestimating the regression using the transformed variables. In the two-variable case, we have

$$\frac{Y_i}{X_i} = \frac{b_0}{X_i} + b_1 + \frac{u_i}{X_i} \tag{9.4}$$

The transformed error term is now homoscedastic:

$$\operatorname{var} u_i = \operatorname{var} \frac{u_i}{X_i} = \frac{1}{X_i^2} \operatorname{var} u_i = C\frac{X_i^2}{X_i^2} = C$$

Note that the original intercept has become a variable in Eq. (*9.4*), while the original slope parameter, b_1, is now the new intercept. However, care must be used to correctly interpret the results of the transformed or weighted regression. Since in Eq. (*9.4*) the errors are homoscedastic, the OLS estimates are not only unbiased and consistent, but also efficient. In the case of a multiple regression, each term of the regression is divided (i.e., weighted) by the independent variable (say, X_{2i}) that is thought to be associated with the error term, so we have

$$\frac{Y_i}{X_{2i}} = \frac{b_0}{X_{2i}} + b_1\frac{X_{1i}}{X_{2i}} + b_2 + \frac{u_i}{X_{2i}} \tag{9.5}$$

In Eq. (*9.5*), the original intercept, b_0, has become a variable, while b_2 has become the new intercept term. We can visually determine whether it is X_{2i} or X_{1i} that is related to the u_i by plotting X_{2i} and X_{1i} against the regression residuals, e_i.

9.6 Table 9.7 gives the consumption expenditures C and disposable income Y_d for 30 families. (*a*) Regress C on Y_d for the entire sample and test for heteroscedasticity. (*b*) Correct for heteroscedasticity if it is found in part *a*.

Table 9.7 Consumption and Income Data for 30 Families (in U.S. Dollars)

Consumption			Income
10,600	10,800	11,100	12,000
11,400	11,700	12,100	13,000
12,300	12,600	13,200	14,000
13,000	13,300	13,600	15,000
13,800	14,000	14,200	16,000
14,400	14,900	15,300	17,000
15,000	15,700	16,400	18,000
15,900	16,500	16,900	19,000
16,900	17,500	18,100	20,000
17,200	17,800	18,500	21,000

(*a*) Regressing C on Y_d for the entire sample of 30 observations, we get

$$\hat{C} = \underset{(3.29)}{1480.0} + \underset{(29.37)}{0.788}\, Y_d \qquad R^2 = 0.97$$

To test for heteroscedasticity, we regress C on Y_d for the first 12 and for last 12 observations, leaving the middle 6 observations out, and we get

$$\hat{C} = \underset{(0.74)}{846.7} + \underset{(9.91)}{0.837} Y_d \qquad R^2 = 0.91, \quad \text{ESS}_1 = 1{,}069{,}000$$

$$\hat{C} = \underset{(0.79)}{2{,}306.7} + \underset{(5.00)}{0.747} Y_d \qquad R^2 = 0.71, \quad \text{ESS}_2 = 3{,}344{,}000$$

Since $\text{ESS}_2/\text{ESS}_1 = 3{,}344{,}000/1{,}069{,}000 = 3.13$ exceeds $F = 2.97$ with $(30 - 6 - 4)/2 = 10$ degrees of freedom in the numerator and denominator at the 5% level of significance (see App. 7), we accept the hypothesis of heteroscedasticity.

(*b*) Assuming that the error variance is proportional to Y_d^2, and then reestimating the regression using the transformed variables of Table 9.8 to correct for heteroscedasticity, we get (in the last column of Table 9.8; 0.833333E-04 = 0.0000833333) the following:

$$\frac{\hat{C}}{Y_d} = \underset{(31.51)}{0.792} + \underset{(3.59)}{1421.3}\frac{1}{Y_d} \qquad R^2 = 0.32$$

Note that the marginal propensity to consume is now given by the intercept (i.e., 0.792) and is larger than before the adjustment (i.e., 0.788). The statistical significance of both estimated parameters is now even higher than before. The R^2 of the weighted regression (i.e., 0.32) is much lower but not directly comparable with the R^2 of 0.97 before the transformation because the dependent variables are different (Y/X as opposed to Y).

9.7 Table 9.9 gives the level of inventories I and sales S, both in millions of dollars, and borrowing rates for 35 firms in an industry. It is expected that I will be directly related to S but inversely related to R. (*a*) Regress I on S and R for the entire sample and test for heteroscedasticity. (*b*) Correct for heteroscedasticity if it is found in part *a*, assuming that the error variance is proportional to S^2.

(*a*) Regressing I on S and R for the entire sample of 35 firms, we get

$$\hat{I} = \underset{}{-6.17} + \underset{(12.39)}{0.20\, S} - \underset{(-2.67)}{0.25\, R} \qquad R^2 = 0.98$$

To test for heteroscedasticity, we regress I on S and R for the first 14 and for the last 14 observations, leaving the middle 7 observations out, and we get

$$\hat{I} = \underset{}{-2.23} + \underset{(1.90)}{0.16S} - \underset{(-0.81)}{0.22R} \qquad R^2 = 0.94, \quad \text{ESS}_1 = 0.908$$

$$= \underset{}{16.10} + \underset{(3.36)}{0.11S} - \underset{(-3.35)}{1.40R} \qquad R^2 = 0.96, \quad \text{ESS}_2 = 5.114$$

Since $\text{ESS}_2/\text{ESS}_1 = 5.114/0.908 = 5.63$ exceeds $F_{11,11} = 2.82$ at the 5% level of significance (see App. 7), we accept the hypothesis of heteroscedasticity.

(*b*) Assuming that the error variance is proportional to S^2 and reestimating the regression using the transformed variable to correct for heteroscedasticity, we get

$$\frac{\hat{I}}{S} = \underset{(12.34)}{0.21} - \underset{}{8.45(1/S)} - \underset{(-2.98)}{0.18(R/S)} \qquad R^2 = 0.93$$

Table 9.8 Consumption C and Disposable Income (Y_d) in Original and Transformed Form

Family	C, $	Y_d, $	C/Y_d, %	$1/Y_d$, %
1	10,600	12,000	0.883333	0.833333E-04
2	10,800	12,000	0.900000	0.833333E-04
3	11,100	12,000	0.925000	0.833333E-04
4	11,400	13,000	0.876923	0.769231E-04
5	11,700	13,000	0.900000	0.769231E-04
6	12,100	13,000	0.930769	0.769231E-04
7	12,300	14,000	0.878571	0.714286E-04
8	12,600	14,000	0.900000	0.714286E-04
9	13,200	14,000	0.942857	0.714286E-04
10	13,000	15,000	0.866667	0.666667E-04
11	13,300	15,000	0.886667	0.666667E-04
12	13,600	15,000	0.906667	0.666667E-04
13	13,800	16,000	0.862500	0.625000E-04
14	14,000	16,000	0.875000	0.625000E-04
15	14,200	16,000	0.887500	0.625000E-04
16	14,400	17,000	0.847059	0.588235E-04
17	14,900	17,000	0.876471	0.588235E-04
18	15,300	17,000	0.900000	0.588235E-04
19	15,000	18,000	0.833333	0.555556E-04
20	15,700	18,000	0.872222	0.555556E-04
21	16,400	18,000	0.911111	0.555556E-04
22	15,900	19,000	0.836842	0.526316E-04
23	16,500	19,000	0.868421	0.526316E-04
24	16,900	19,000	0.889474	0.526316E-04
25	16,900	20,000	0.845000	0.500000E-04
26	17,500	20,000	0.875000	0.500000E-04
27	18,100	20,000	0.905000	0.500000E-04
28	17,200	21,000	0.819048	0.476190E-04
29	17,800	21,000	0.847619	0.476190E-04
30	18,500	21,000	0.880952	0.476190E-04

$b_0 = 0.21$ is now the slope coefficient associated with the variable S (instead of 0.16 before the transformation), while $b_2 = -0.18$ is the slope coefficient associated with the variable R (instead of -0.25 before the transformation). Both these slope coefficients remain highly significant before and after the transformation, as does R^2. The new constant is -8.45 instead of -6.17.

AUTOCORRELATION

9.8 (*a*) What is meant by *autocorrelation*? (*b*) Draw a figure showing positive and negative first-order autocorrelation. (*c*) Why is autocorrelation a problem?

(*a*) *Autocorrelation* or *serial correlation* refers to the case in which the error term in one time period is correlated with the error term in any other time period. If the error term in one time period is correlated with the error term in the *previous* time period, there is *first-order* autocorrelation. Most of the applications in econometrics involve first rather than second- or higher-order autocorrelation. Even though *negative* autocorrelation is possible, most economic time series exhibit *positive*

Table 9.9 Inventories, Sales, and Borrowing Rates for 35 Firms

Firm	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
I	10	10	10	11	11	11	12	12	12	12	12	13	13	13	14	14	14	15
S	100	101	103	105	106	106	108	109	111	111	112	113	114	114	116	117	118	120
R	17	17	17	16	16	16	15	15	14	14	14	14	13	13	12	12	12	11
Firm	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	
I	15	15	15	16	16	16	17	17	17	17	18	18	19	19	19	20	20	
S	122	123	125	128	128	131	133	134	135	136	139	143	147	151	157	163	171	
R	11	11	11	10	10	10	10	9	9	9	8	8	8	8	8	7	7	

autocorrelation. Positive, first-order serial or autocorrelation means that $E_{u_t u_{t-1}} > 0$, thus violating the fourth OLS assumption (see Prob. 6.4). This is common in time-series analysis.

(*b*) Figure 9-2*a* shows positive and Fig. 9-2*b* shows negative first-order autocorrelation. Whenever several consecutive residuals have the same sign as in Fig. 9-2*a*, there is positive first-order autocorrelation. However, whenever consecutive residuals change sign frequently, as in Fig. 9-2*b*, there is negative first-order autocorrelation.

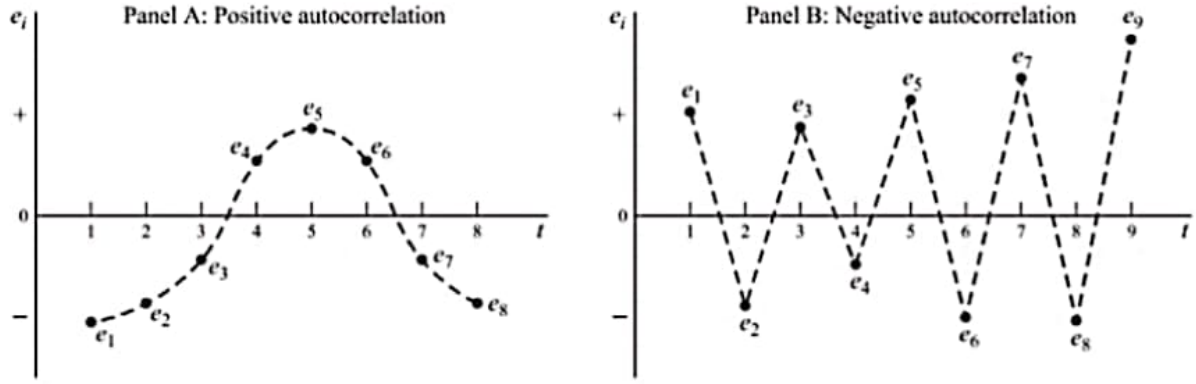

Fig. 9-2

(*c*) With autocorrelation, the OLS parameter estimates are still unbiased and consistent, but the standard errors of the estimated regression parameters are biased, leading to incorrect statistical tests and biased confidence intervals. With positive first-order autocorrelation, the standard errors of the estimated regression parameters are biased downward, thus exaggerating the precision and statistical significance of the estimated regression parameters.

9.9 (*a*) How is the presence of positive or negative first-order autocorrelation tested? (*b*) How can autocorrelation be corrected?

(*a*) The presence of autocorrelation can be tested by calculating the Durbin-Watson statistic d given by Eq. (*9.1*). This is routinely given by most computer programs such as SAS:

$$d = \frac{\sum_{t=2}^{n} (e_t - e_{t-1})^2}{\sum_{t=1}^{n} e_t^2} \tag{9.1}$$

The calculated value of d ranges between 0 and 4, with no autocorrelation when d is in the neighborhood of 2. The values of d indicating the presence or absence of positive or negative first-order autocorrelation, and for which the test is inconclusive, are summarized in Fig. 9-3. When the lagged dependent appears as an explanatory variable in the regression, d is biased toward 2 and its power to detect autocorrelation is hampered.

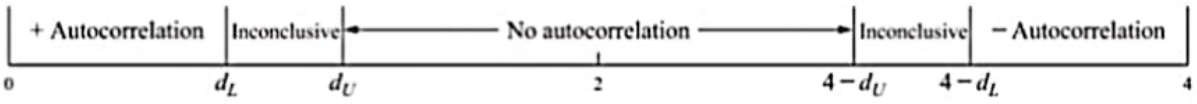

Fig. 9-3

(*b*) One method to correct positive first-order autocorrelation (the usual type) involves first regressing Y on its value lagged one period, the explanatory variable of the model, and the explanatory variable lagged one period:

$$Y_t = b_0(1-\rho) + \rho Y_{t-1} + b_1 X_t - b_1 \rho X_{t-1} + v_t \tag{9.2}$$

(The preceding equation is derived by multiplying each term of the original OLS model lagged one period by ρ, subtracting the resulting expression from the original OLS model, transposing the term ρY_{t-1} from the left to the right side of the equation, and defining $v_t = u_t - \rho u_{t-1}$.) The second step involves using the value of ρ found in Eq. (*9.2*) to transform all the variables of the original OLS model, as indicated in Eq. (*9.3*), and then estimating Eq. (*9.3*):

$$Y_t - \hat{\rho} Y_{t-1} = b_0(1-\hat{\rho}) + b_1(X_t - \hat{\rho} X_{t-1}) + \varepsilon_t \tag{9.3}$$

The error term, ε_t, in Eq. (*9.3*) is now free of autocorrelation. This procedure, known as the *Durbin two-stage method*, is an example of generalized least squares. To avoid losing the first observation in the differencing process, $Y_1\sqrt{1-\hat{\rho}^2}$ and $X_1\sqrt{1-\hat{\rho}^2}$ are used for the first transformed observation of Y and X, respectively. If the autocorrelation is due to the omission of an important variable, wrong functional form, or improper model specification, these problems should be removed first, before applying the preceding correction procedure for autocorrelation.

9.10 Table 9.10 gives the level of U.S. imports M and GDP (both seasonally adjusted in billions of dollars) from 1980 to 1999. (*a*) Regress M on GDP and test for autocorrelation at the 5% level of significance. (*b*) Correct for autocorrelation if it is found in part *a*.

(*a*)
$$\hat{M}_t = -201.80 + 0.14\ \text{GDP}_t \qquad R^2 = 0.98$$
$$(-6.48)\quad (29.44) \qquad\qquad d = 0.54$$

Since $d = 0.54 < d_L = 1.20$ at the 5% level of significance with $n = 20$ and $k' = 1$ (from App. 8), there is evidence of positive first-order autocorrelation.

Table 9.10 Seasonally Adjusted U.S. Imports and GDP (Both in Billions of Dollars) from 1980 to 1999

Year	1980	1981	1982	1983	1984	1985	1986	1987	1988	1989
M	299.2	319.4	294.9	358.0	416.4	438.9	467.7	536.7	573.5	599.6
GDP	2918.8	3203.1	3315.6	3688.8	4033.5	4319.3	4537.5	4891.6	5258.3	5588.0
Year	1990	1991	1992	1993	1994	1995	1996	1997	1998	1999
M	649.2	639.0	687.1	744.9	859.6	909.3	992.8	1087.0	1147.3	1330.1
GDP	5847.3	6080.7	6469.8	6795.5	7217.7	7529.3	7981.4	8478.6	8974.9	9559.7

Source: St. Louis Federal Reserve (Bureau of Economic Analysis).

(*b*) To correct for autocorrelation, first the following regression is run:

$$\hat{M}_t = -103.21 + \underset{(4.72)}{0.82}\, M_{t-1} + \underset{(4.68)}{0.36}\, \text{GDP}_t - \underset{(-4.23)}{0.33}\, \text{GDP}_{t-1} \qquad R^2 = 0.98$$

Then, using $\hat{\rho} = 0.82$ (the coefficient on M_{t-1} in the preceding regression), we transform the original variables as indicated in Eq. (*9.3*). The original variables (M and GDP) and the transformed variables (M^* and GDP*) are given in Table 9.11.

$$M^*_{1980} = 299.2\sqrt{1 - 0.82^2} = 171.251 \qquad \text{and} \qquad \text{GDP}^*_{1980} = 2918.8\sqrt{1 - 0.82^2} = 1670.615$$

Table 9.11 U.S. Imports and GDP in Original and Transformed Form

Year	M	GDP	M^*	GDP*
1980	299.2	2918.8	171.250	1670.610
1981	319.4	3203.1	74.056	809.684
1982	294.9	3315.6	32.992	689.058
1983	358.0	3688.8	116.182	970.008
1984	416.4	4033.5	122.840	1008.684
1985	438.9	4319.3	97.452	1011.830
1986	467.7	4537.5	107.802	995.674
1987	536.7	4891.6	153.186	1170.850
1988	573.5	5258.3	133.406	1247.188
1989	599.6	5588.0	129.330	1276.194
1990	649.2	5847.3	157.528	1265.140
1991	639.0	6080.7	106.656	1285.914
1992	687.1	6469.8	163.120	1483.626
1993	744.9	6795.5	181.478	1490.264
1994	859.6	7217.7	248.782	1645.390
1995	909.3	7529.3	204.428	1610.786
1996	992.8	7981.4	247.174	1807.374
1997	1087.0	8478.6	272.904	1933.852
1998	1147.3	8974.9	255.960	2022.448
1999	1330.1	9559.7	389.314	2200.282

Regressing M^* on GDP*, we get

$$\hat{M}^*_t = \underset{(7.79)}{579.53} + \underset{(11.91)}{4.75}\, \text{GDP}^*_t \qquad R^2 = 0.88 \qquad d = 1.69$$

Since now $d = 1.69 > d_U = 1.41$ at the 5% level of significance with $n = 20$ and $k' = 1$ (from App. 8), there is no evidence of autocorrelation. Note that though GDP$^*_t$ remains highly significant, its t value is lower than the t value of GDP$_t$. In addition, $R^2 = 0.88$ now, as opposed to $R^2 = 0.98$ before the correction for autocorrelation.

9.11 Table 9.12 gives gross private domestic investment (GPDI) and GDP, both in seasonally adjusted billions of 1996 dollars, and the GDP deflator price index P for the United States from 1980 to 1999. (*a*) Regress GPDI on GDP and P and test for autocorrelation at the 5% level of significance. (*b*) Correct for autocorrelation if it is found in part *a*.

Table 9.12 U.S. GPDI, GDP (Both in Seasonally Adjusted Billions of 1996 Dollars), and GDP Deflator Price Index, 1982–1999

Year	1982	1983	1984	1985	1986	1987	1988	1989	1990
GPDI	571.1	762.2	876.9	887.8	838.2	929.3	916.7	922.9	849.6
GDP	4915.6	5286.8	5583.1	5806.0	5969.5	6234.4	6465.2	6633.5	6664.2
P	67.44	69.75	72.24	74.40	76.05	78.46	81.36	84.24	87.76
Year	1991	1992	1993	1994	1995	1996	1997	1998	1999
GPDI	864.2	941.6	1015.6	1150.5	1152.4	1283.7	1438.5	1609.9	1751.6
GDP	6720.9	6990.6	7168.7	7461.1	7621.9	7931.3	8272.9	8654.5	9084.1
P	90.47	92.56	94.79	96.74	98.79	100.63	102.49	103.69	105.31

Source: St. Louis Federal Reserve (Bureau of Economic Analysis).

(*a*)
$$\widehat{\text{GPDI}}_t = -199.71 + 0.56\ \text{GDP}_t - 29.70\ P_t \qquad R^2 = 0.97$$
$$(10.61) \quad (-6.07) \qquad d = 0.56$$

Since $d = 0.56 < d_L = 1.05$ at the 5% level of significance with $n = 18$ and $k' = 2$ (from App. 8), there is evidence of autocorrelation.

(*b*) To correct for autocorrelation, first, the following regression is run:

$$\widehat{\text{GPDI}}_t = -291.79 + \underset{(2.99)}{0.74\ \text{GPDI}_{t-1}} + \underset{(7.12)}{0.76\ \text{GDP}_t} - \underset{(-4.28)}{0.73\ \text{GDP}_{t-1}} + \underset{(0.06)}{1.91\ P_t} + \underset{(0.06)}{1.40\ P_{t-1}}$$
$$R^2 = 0.99$$

Then, using $\hat{\rho} = 0.74$ (the coefficient on GPDI_{t-1} in the preceding regression), we transform the original variables as indicated in Eq. (*9.3*). The original and the transformed variables (the latter indicated by an asterisk) are given in Table 9.13.

$$\text{GPDI}^*_{1982} = 571.1\sqrt{1 - 0.74^2} = 384.126$$
$$\text{GDP}^*_{1982} = 4915.6\sqrt{1 - 0.74^2} = 3306.266$$
$$P^*_{1982} = 67.44\sqrt{1 - 0.74^2} = 45.361$$

Regressing GPDI^*_t on GDP^*_t and P^*_t, we get

$$\widehat{\text{GPDI}}^*_t = 31.05 + 0.52\ \text{GDP}^*_t - 30.02 P^*_t \qquad R^2 = 0.88$$
$$(9.81) \quad (-6.54) \qquad d = 1.77$$

Since $d = 1.77 > d_U = 1.53$ at the 5% level of significance with $n = 18$ and $k' = 2$ (from App. 8), there is no evidence of autocorrelation. Both variables remain highly significant, and R^2 falls.

9.12 Table 9.14 gives personal consumption expenditures C and disposable personal income Y, both in billions of dollars, for the United States from 1982 to 1999. (*a*) Regress C_t on Y_t and test for autocorrelation. (*b*) Correct for autocorrelation if it is found in part *a*.

(*a*)
$$\hat{C}_t = -293.46 + 0.97 Y_t \qquad R^2 = 0.99$$
$$(-6.58) \quad (99.65) \qquad d = 0.58$$

Since $d = 0.58$, there is evidence of autocorrelation at both the 5 and 1% levels of significance.

Table 9.13 GPDI, GDP, and P in Original and Transformed Form

Year	GPDI	GDP	P	GPDI*	GDP*	P^*
1980	662.2	4936.6	59.16	384.126	3306.266	45.3610
1981	708.8	4997.1	64.10	218.772	1344.016	20.3216
1982	571.1	4915.6	67.44	46.588	1217.746	20.0060
1983	762.2	5286.8	69.75	339.586	1649.256	19.8444
1984	876.9	5583.1	72.24	312.872	1670.868	20.6250
1985	887.8	5806.0	74.40	238.894	1674.506	20.9424
1986	838.2	5969.5	76.05	181.228	1673.060	20.9940
1987	929.3	6234.4	78.46	309.032	1816.970	22.1830
1988	916.7	6465.2	81.36	229.018	1851.744	23.2996
1989	922.9	6633.5	84.24	244.542	1849.252	24.0336
1990	849.6	6664.2	87.76	166.654	1755.410	25.4224
1991	864.2	6720.9	90.47	235.496	1789.392	25.5276
1992	941.6	6990.6	92.56	302.092	2017.134	25.6122
1993	1015.6	7168.7	94.79	318.816	1995.656	26.2956
1994	1150.5	7461.1	96.74	398.956	2156.262	26.5954
1995	1152.4	7621.9	98.79	301.030	2100.686	27.2024
1996	1283.7	7931.3	100.63	430.924	2291.094	27.5254
1997	1438.5	8272.9	102.49	488.562	2403.738	28.0238
1998	1609.9	8654.5	103.69	545.410	2532.554	27.8474
1999	1751.6	9084.1	105.31	560.274	2679.770	28.5794

Table 9.14 U.S. Consumption Expenditures and Disposable Income (in Billions of Dollars), 1982-1999

Year	1982	1983	1984	1985	1986	1987	1988	1989	1990
C	2079.3	2286.4	2498.4	2712.6	2895.2	3105.3	3356.6	3596.7	3831.5
Y	2406.8	2586.0	2887.6	3086.5	3262.5	3459.5	3752.4	4016.3	4293.6
Year	1991	1992	1993	1994	1995	1996	1997	1998	1999
C	3971.2	4209.7	4454.7	4716.4	4969.0	5237.5	5524.4	5848.6	6254.9
Y	4474.8	4754.6	4935.3	5165.4	5422.6	5677.7	5982.8	6286.2	6639.2

Source: *Economic Report of the President*.

(*b*) To correct for autocorrelation, first the following regression is run:

$$\hat{C}_t = \underset{}{93.90} + \underset{(5.18)}{1.23}\, C_{t-1} + \underset{(1.79)}{0.40}\, Y_t - \underset{(-3.08)}{0.60}\, Y_{t-1} \qquad R^2 = 0.99$$

Since $\hat{\rho} \cong 1$ (the coefficient on C_{t-1} in the preceding regression), we rerun the regression on the first differences of the original variables (i.e., ΔC_t and ΔY_t), omitting the intercept, and get

$$\Delta\hat{C}_t = \underset{(25.88)}{0.97}\Delta Y_t \qquad R^2 = 0.98 \qquad d = 1.75$$

The new value of d indicates no evidence of autocorrelation at either the 1 or at the 5% level of significance. (*Note*: R^2 is not well defined in regression with no intercept and therefore is not comparable with the previous regressions. For a more in-depth study of procedure when $\rho = 1$, see Sec. 11.3.)

ERRORS IN VARIABLES

9.13 (*a*) What is meant by *errors in variables*? (*b*) What problems do errors in variables create? (*c*) Is there any test to detect the presence of errors in variables? (*d*) How can the problems created by the existence of errors in variables be corrected?

(*a*) *Errors in variables* refer to the case in which the variables in the regression model include measurement errors. These are probably very common in view of the way most data are collected and elaborated.

(*b*) Measurement errors in the dependent variable are incorporated into the disturbance term leaving unbiased and consistent (although inefficient or larger than minimum variance) OLS parameter estimates. However, with measurement errors in the explanatory variables, the fifth of the OLS assumption of independence of the explanatory variables and error term is violated (see Prob. 6.4), leading to biased and inconsistent OLS parameter estimates. In a simple regression, $\hat{b}_1$ is biased downward, while $\hat{b}_0$ is biased upward.

(*c*) There is no formal test to detect the presence of errors in variables. Only economic theory and knowledge of how the data were gathered can sometimes give some indication of the seriousness of the problem.

(*d*) One method of obtaining consistent (but still biased and inefficient) OLS parameter estimates is to replace the explanatory variable subject to measurement errors with another variable that is highly correlated with the explanatory variable in question but which is independent of the error term. In the real world, it might be difficult to find such an instrumental variable, and one could never be sure that it would be independent of the error term. The most popular instrumental variable is the lagged value of the explanatory variable in question. Measurement errors in the explanatory variable only also can be corrected by inverse least squares. This involves regressing X on Y. Then, $\hat{b}_0 = -\hat{b}'_0/\hat{b}'_1$ and $\hat{b}_1 = 1/b'_1$, where $\hat{b}_0$ and $\hat{b}_1$ are consistent estimates of the intercept and slope parameter of the regression of Y_t on X_t.

9.14 Table 9.15 gives inventories Y, actual sales X, and hypothetical values of X that include measurement errors, X', all in billions of dollars, in U.S. manufacturing from 1983 to 1998. Y and X are assumed to be free of measurement errors. (*a*) Regress Y_t on X_t. (*b*) Regress Y_t on X'_t (on the assumption that X is not available). What type of bias results in the estimates in using X' instead of X? (*c*) Use instrumental variables to obtain consistent parameter estimates, on the assumption that X_t is correlated with u_t. How do these parameter estimates compare with those obtained in part *b*?

(*a*)
$$\hat{Y}_t = 169.69 + 0.90\ X_t \qquad R^2 = 0.95$$
$$(11.66)\quad (16.46)$$

Table 9.15 Inventory and Sales (Both in Billions of Dollars) in U.S. Manufacturing, 1983–1998

Year	1983	1984	1985	1986	1987	1988	1989	1990
Y	312	340	335	323	338	369	391	405
X	172	191	194	195	206	225	237	243
X'	176	195	199	200	212	232	245	252
Year	1991	1992	1993	1994	1995	1996	1997	1998
Y	391	383	384	405	431	437	456	467
X	240	250	261	279	300	310	327	338
X'	251	263	276	296	320	333	352	366

Source: *Economic Report of the President.*

(*b*) Regressing Y_t on X_t' (if X_t is not available), we get

$$\hat{Y}_t = 182.50 + 0.78\ X_t' \qquad R^2 = 0.94$$
$$(13.38)\quad (15.23)$$

Note that $\hat{b}_1' < \hat{b}_1$; furthermore, b_1 falls outside the 95% confidence interval of b_1' (0.67 to 0.89).

(*c*) Using X_{t-1}' as an instrumental variable for X_t' (if X_t' is believed to be correlated with u_t), we get

$$\hat{Y}_t = 187.90 + 0.80\ X_{t-1}' \qquad R^2 = 0.92$$
$$(11.44)\quad (12.57)$$

The coefficient on X_{t-1}' is closer to the true one ($\hat{b}_1$ falls in the 95% confidence interval of 0.66 to 0.94), and is consistent. Of course, in the real world it is rarely known what error of measurement might be present (otherwise, the errors could be corrected before running the regression). It is also difficult or impossible to establish whether X_t' is correlated with u_t.

9.15 Using the data in Table 9.15, (*a*) regress X_t' on Y_t in order to overcome errors in measuring X_t. (*b*) How do these results compare with those in Prob. 9.14(*c*)?

(*a*) Since only X_t (i.e., the explanatory variable) is subject to measurement errors, inverse least squares is another method for obtaining consistent parameter estimates. Regressing X_t' on Y_t, we get

$$\hat{X}_t' = -206.10 + 1.21\ Y_t \qquad R^2 = 0.94$$
$$(-6.68)\quad (15.23)$$

$$\hat{b}_0 = -\frac{\hat{b}_0'}{\hat{b}_1'} = -\frac{(-206.10)}{1.21} = 170.33 \qquad \text{and} \qquad \hat{b}_1 = \frac{1}{\hat{b}_1'} = \frac{1}{1.21} = 0.83$$

where $\hat{b}_0$ and $\hat{b}_1$ are consistent (but still biased) estimates of the intercept and slope parameters of the regression of Y_t on X_t.

(*b*) Using inverse least squares gives better results in this case compared to the instrumental-variable method [see Prob. 9.14(*c*)]. With instrumental variables, both the estimated intercept and slope parameter are farther from the true values. However, the results may very well differ in other cases. In any event, in the real world we seldom know what types of errors are present, what type of adjustment is appropriate, and how close the adjusted parameters are to the true parameter values.

Supplementary Problems

MULTICOLLINEARITY

9.16 Why can the following consumption function not be estimated?

$$C_t = b_0 + b_1 Y_{dt} + b_2 Y_{dt-1} + b_3 \Delta Y_{dt} + u_t$$

where $\Delta Y_{dt} = Y_{dt} - Y_{dt-1}$.

Ans. Because there is a perfect multicollinearity between ΔY_{dt} on one hand and Y_{dt} and Y_{dt-1} on the other. As a result, there are only three independent normal equations and four coefficients to estimate, and so no unique solution is possible.

9.17 Table 9.16 gives hypothetical data on consumption expenditures C, disposable income Y_d, and wealth W, all in thousands of dollars, for a sample of 15 families. (*a*) Regress C on Y_d and W and find $\bar{R}^2$ and r_{Y_dW}. (*b*) Regress C on Y_d only. (*c*) Regress C on W only. (*d*) What can you conclude from the preceding with regard to multicollinearity?