Autocorrelation

The Nature of the Problem

The Nature of the Problem

Autocorrelation may be defined as "correlation between members of series of observations ordered in time (as in time series data) or space (as in cross-sectional data)

$$\text{cov}(u_i, u_j \mid x_i, x_j) = E(u_i u_j) = 0 \quad i \neq j$$

Example

- If we are dealing with quarterly time series data involving the regression of output on labor and capital inputs and if, say there is a labor strike affecting output in one quarter, there is no reason to believe that this disruption will be carried over to the next quarter
- If output is lower this quarter, there is no reason to expect it to be lower next quarter

- If we are dealing with cross-sectional data involving the regression of family consumption expenditure on family income, the effect of an increase of one family's income on its consumption expenditure is not affect the consumption expenditure of another family

Autocorrelation defines as Lag correlation of a given series with itself, lagged by a number of time units

The correlation between two time series such as $u_1, u_2, ..., u_{10}$ and $u_2, u_3, ..., u_{11}$, where the former is the latter series lagged by one time period

The correlation between time series such as $u_1, u_2, \ldots, u_{10}$ and $v_2, v_3, \ldots, v_{11}$, where u and v are two different time series, is called Serial correlation

No Autocorrelation

$$E(u_i u_j) = 0, \quad i \neq j$$

$$E(uu') = \begin{bmatrix} E(u_1^2) & E(u_1 u_2) & \dots & E(u_1 u_n) \\ E(u_2 u_1) & E(u_2^2) & \dots & E(u_2 u_n) \\ \dots & \dots & \dots & \dots \\ E(u_n u_1) & E(u_n u_2) & \dots & E(u_n^2) \end{bmatrix}$$

$$E(uu') = \begin{bmatrix} \sigma^2 & 0 & 0 & 0 \\ 0 & \sigma^2 & 0 & 0 \\ \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & \sigma^2 \end{bmatrix} = \sigma^2 \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & 1 \end{bmatrix} = \sigma^2 I$$

Why does serial correlation occur?

Inertia

A salient feature of most economic time series is sluggishness

Example

GNP, price indexes, production, employment, and unemployment exhibit business cycles

Specification Bias: excluded variables case

$$Y_t = \beta_1 + \beta_2 X_{2t} + \beta_3 X_{3t} + \beta_4 X_{4t} + u_t$$

$$Y_t = quantity\ of\ beef\ demand$$

$$X_2 = price\ of\ beef$$

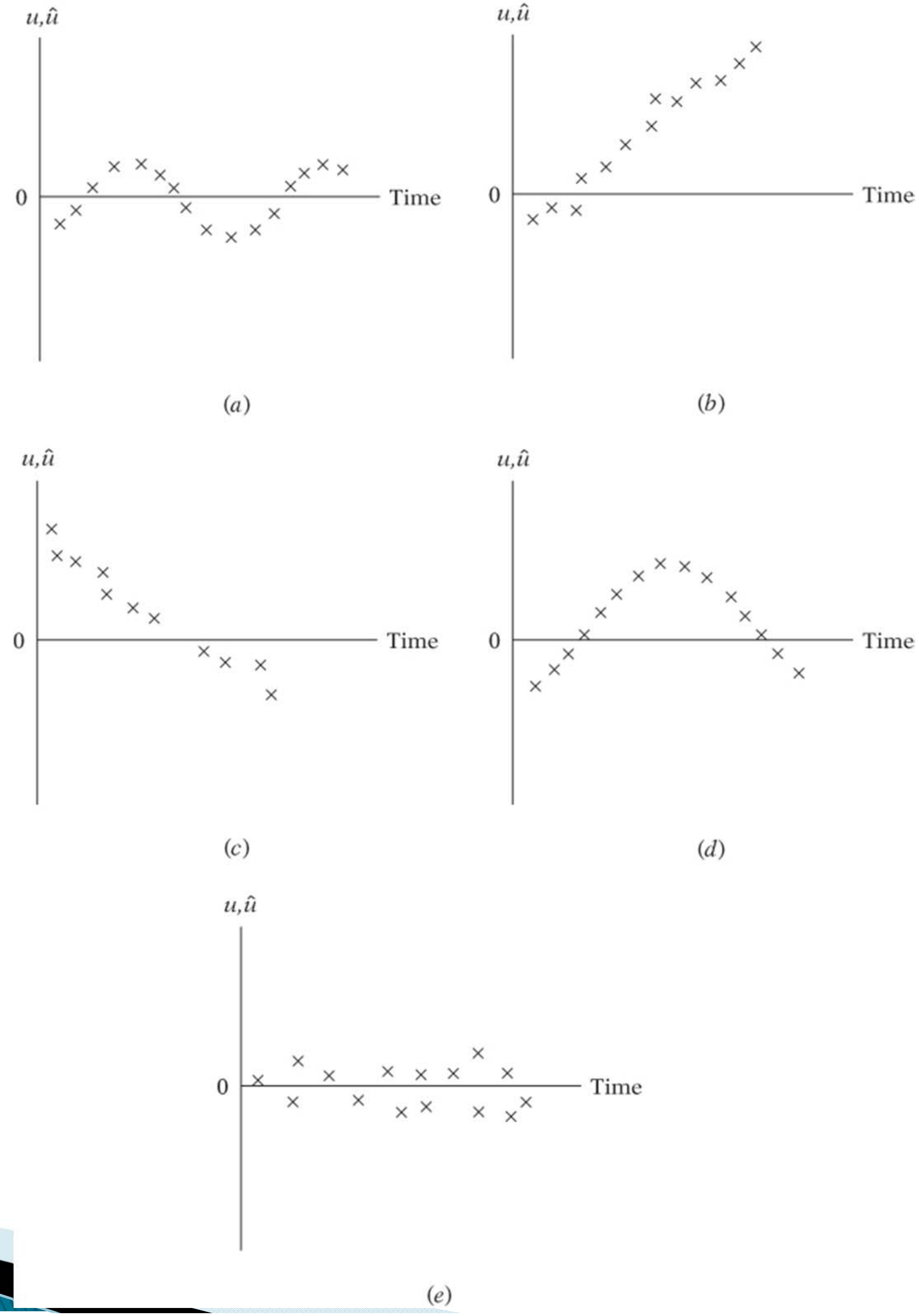
$$X_3 = consumer\ income$$

$$X_4 = price\ of\ pork$$

$$t = time$$

$$Y_t = \beta_1 + \beta_2 X_{2t} + \beta_3 X_{3t} + \nu_t$$

$$\nu_t = \beta_4 X_{4t} + u_t$$


u,û
0
Time
(a)
(b)
(c)
(d)
(e)

The price of pork affects the consumption of beef, the error or disturbance term v will reflect a systematic pattern, thus creating (false) autocorrelation

Specification Bias: Incorrect Functional Form

Suppose the true or correct model in a cost-output study is as follows:

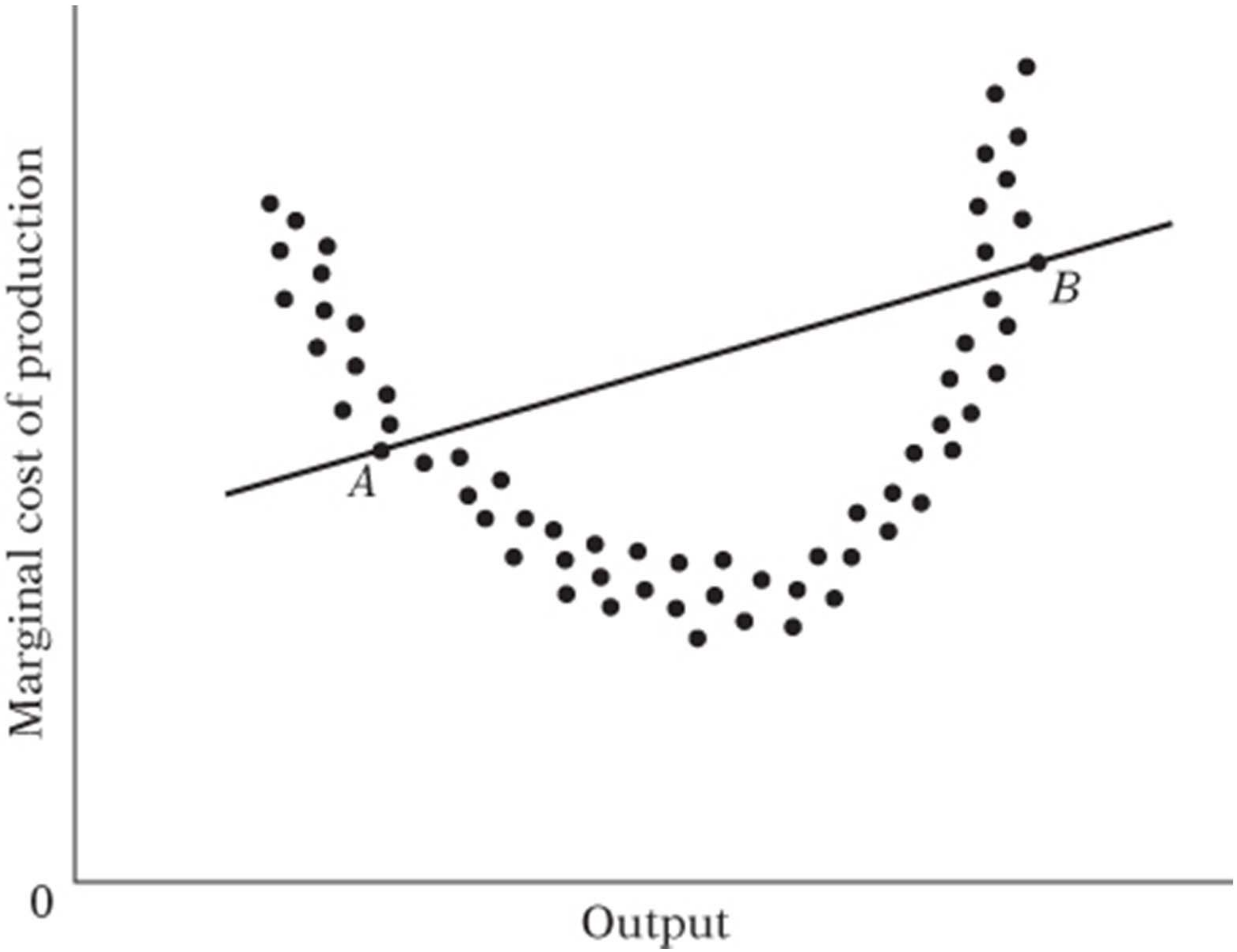
$$MARGINAL\ \ COST_i = \beta_1 + \beta_2 output_i + \beta_3 output_i^2 + u_i$$

But we fit the following model:

$$MARGINAL\ \ COST_i = \alpha_1 + \alpha_2 output_i + \nu_i$$

$$\nu_i = \beta_3 output^2 + u_i$$

The marginal cost curve corresponding to the true model is shown in the next figure along with the incorrect linear cost curve



Marginal cost of production
A
B
0
Output

Cobweb Phenomenon

The supply of many agricultural commodities reflects the so-called cobweb phenomenon, where supply reacts to price with a lag of one time period because supply decisions take time to implement.

$$Supply_t = \beta_1 + \beta_2 P_{t-1} + u_t$$

Lags

$$Consumption_t = \beta_1 + \beta_2 income + \beta_3 consumption_{t-1} + u_t$$

Data Transformation

$$Y_t = \beta_1 + \beta_2 X_t + u_t$$

$$Y = Consumption \quad Expenditure$$

$$X = income$$

$$Y_{t-1} = \beta_1 + \beta_2 X_{t-1} + u_{t-1} \longleftarrow \text{Level Form}$$

$$\Delta Y_t = \beta_2 \Delta X_t + \Delta u_t$$

$$\Delta Y_t = \beta_2 \Delta X_t + \nu_t$$

$$\nu_t = \Delta u_t = (u_t - u_{t-1}) \longleftarrow \text{Difference Form}$$

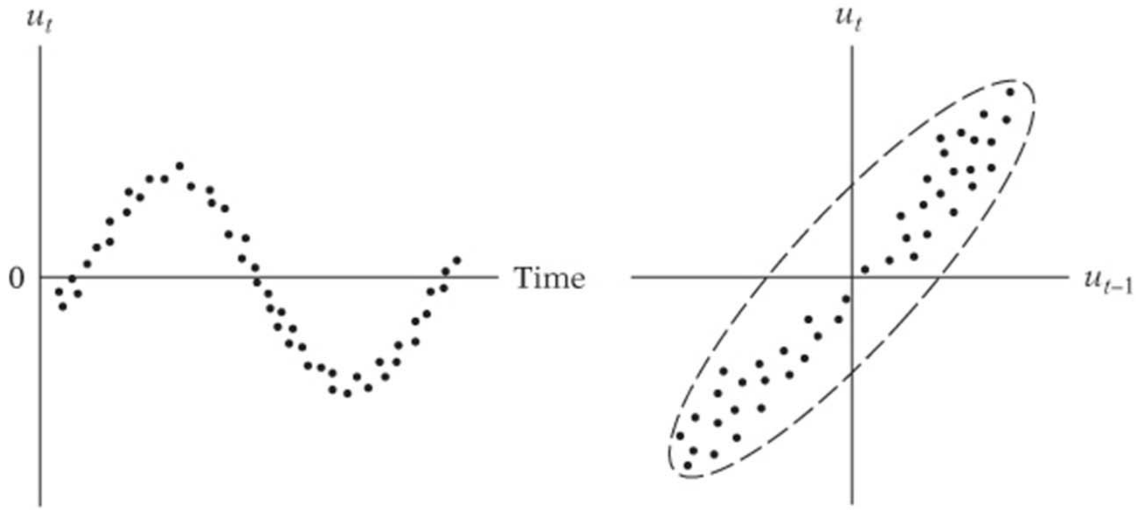
$$\nu_t = u_t - u_{t-1}$$

$$E(u) = 0$$

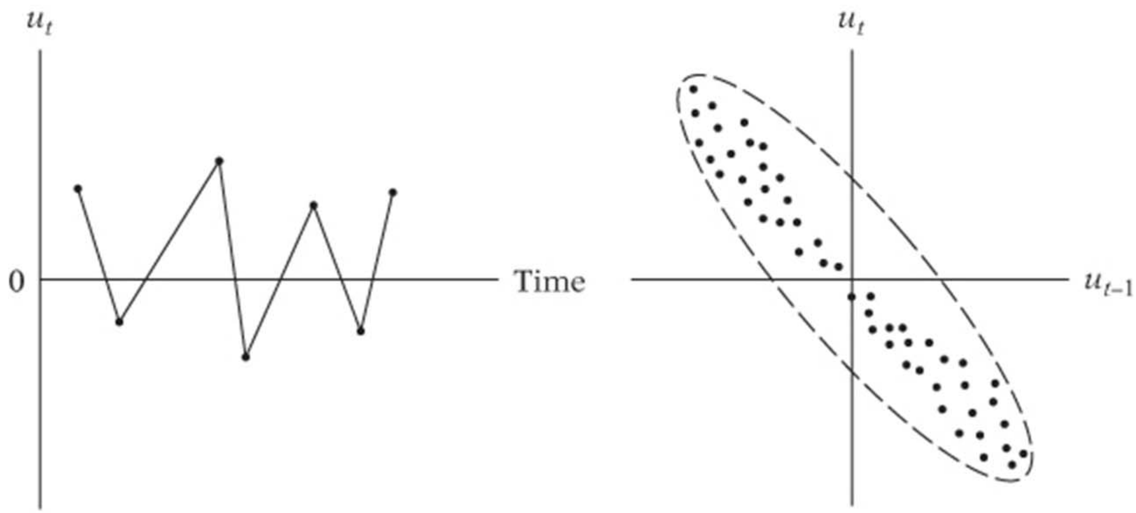
$$E(\nu_t) = E(u_t - u_{t-1}) = E(u_t) - E(u_{t-1}) = 0$$

$$\text{var}(\nu_t) = \text{var}(u_t - u_{t-1}) = \text{var}(u_t) - \text{var}(u_{t-1}) = 2\sigma^2$$

$$\text{cov}(\nu_t, \nu_{t-1}) = E(\nu_t \nu_{t-1}) = E\left[\left(u_t - u_{t-1}\right)\left(u_{t-1} - u_{t-2}\right)\right] = -\sigma^2$$



(*a*)



(*b*)

OLS Estimation in the Presence of Autocorrelation

- What happens to the OLS estimators and their variances if we introduce autocorrelation in the disturbances by assuming that

$$E(u_t u_{t+s}) \neq 0 \quad (s \neq 0)$$

but retain all the other assumptions of classical model?

$$Y_t = \beta_1 + \beta_2 X_i + u_i$$

$$u_t = \rho u_{t-1} + \varepsilon_t \qquad -1 < \rho < 1$$

ρ (coefficient of autocovariance)

$$E(\varepsilon_t) = 0$$

$$\text{var}(\varepsilon_t) = \sigma_\varepsilon^2$$

$$\text{cov}(\varepsilon_t, \varepsilon_{t+s}) = 0 \quad s \neq 0$$

ε (white noise error term)

$$u_t = \rho u_{t-1} + \varepsilon_t \qquad -1 < \rho < 1$$

This equation is known as a Markov first order autoregressive scheme or first-order autoregressive scheme AR(1)

AR (1)

$$E(u_t) = \rho E(u_{t-1}) + E(\varepsilon) = 0$$

$$\operatorname{var}(u_t) = E(u_t^2)$$

$$= \rho^2 \operatorname{var}(u_{t-1}) + \operatorname{var}(\varepsilon) = \frac{\sigma_\varepsilon^2}{1-\rho^2}$$

$$\operatorname{var}(u_t) = \operatorname{var}(u_{t-1}) = \sigma^2$$

$$\operatorname{var}(\varepsilon_t) = \sigma_\varepsilon^2$$

$$\operatorname{cov}(u_t, u_{t+s}) = E(u_t, u_{t+s}) = \rho^s \frac{\sigma_\varepsilon^2}{1-\rho^2}$$

Covariance between error terms s periods apart

The symmetry property of covariances

$$\operatorname{cov}(u_t, u_{t+s}) = \operatorname{cov}(u_t, u_{t-s})$$

- By definition, the (population) coefficient of correlation between u_t and u_{t-1} is

$$\rho = \frac{E\{[u_t - E(u_t)][u_{t-1} - E(u_{t-1})]\}}{\sqrt{\operatorname{var}(u_t)}\sqrt{\operatorname{var}(u_{t-1})}} = \frac{E(u_t u_{t-1})}{\operatorname{var}(u_{t-1})}$$

Since $E(u_t) = 0$ for each t and $\operatorname{var}(u_t) = \operatorname{var}(u_{t-1})$ because we are retaining the assumption of homoscedasticity.

$$Cor(u_t, u_{t+s}) = \rho^s$$

Correlation between error terms s period apart

The symmetry property of correlations

$$Cor(u_t, u_{t+s}) = Cor(u_t, u_{t-s})$$

Since ρ is a constant between -1 and +1, , variance equation shows that under the AR(1) scheme, the variance of u_t is still homoscedastic, but u_t is correlated not only with its immediate past value but its values several periods in the past.

Two variable regression model

$$Y_t = \beta_1 + \beta_2 X_t + u_t$$

OLS estimator of the slope coefficient is

$$\hat{\beta}_2 = \frac{\sum (X_t - \bar{X})(Y_t - \bar{Y})}{\sum (X_t - \bar{X})^2} = \frac{\sum x_t y_t}{\sum x_t^2}$$

and its variance is given by

$$\text{var}(\hat{\beta}_2) = \frac{\sigma^2}{\sum (X_t - \bar{X})^2} = \frac{\sigma^2}{\sum x_t^2}$$

Now under the AR(1) scheme, it can be shown that the variance of this estimator is

$$\text{var}(\hat{\beta}_2)_{AR1} = \frac{\sigma^2}{\sum x_t^2}\left[1 + 2\rho\frac{\sum x_t x_{t-1}}{\sum x_t^2} + 2\rho^2\frac{\sum x_t x_{t-2}}{\sum x_t^2} + \ldots + 2\rho^{n-1}\frac{\sum x_1 x_n}{\sum x_t^2}\right]$$

where

$$x_t = (X_t - \bar{X})$$

$$x_{t-1} = (X_{t-1} - \bar{X})$$

$$\vdots$$

$$x_n = (X_n - \bar{X})$$

The difference between the $\operatorname{var}(\hat{\beta}_2)$ and $\operatorname{var}(\hat{\beta}_2)_{AR(1)}$, assume that the regressor X also follows the first order autoregressive scheme with a coefficient of autocorrelation of r

$$\operatorname{var}(\hat{\beta}_2)_{AR(1)} = \frac{\sigma^2}{\sum x_t^2}\left(\frac{1+r\rho}{1-r\rho}\right) = \operatorname{var}(\hat{\beta}_2)_{OLS}\left(\frac{1+r\rho}{1-r\rho}\right)$$

As in the case of **heteroscedasticity** in the presence of **autocorrelation** the OLS estimators are still linear unbiased as well as consistent and asymptotically normally distributed, but they are **no longer efficient (i.e., minimum variance)**

The blue estimator in the presence of autocorrelation

Continuing with the two-variable model and assuming the AR(1) process, we can show that the BLUE estimator of β_2 is given by the following expression

$$\widehat{\beta_2{}^{GLS}} = \frac{\sum_{t=2}^{n}(x_t - px_{t-1})(y_t - \rho y_{t-1})}{\sum_{t=2}^{n}(x_t - px_{t-1})^2} + C$$

Where C is a correction factor that may be disregarded in practice. Note that the subscript t now runs from t=2 to t=n.

And its variance is given by

$$\text{Var}\,\widehat{\beta_2{}^{\text{GLS}}} = \frac{\sigma^2}{\sum_{t=2}^{n}(x_t - p x_{t-1})^2} + D$$

Where D too is a correction factor that may also be disregarded in practice

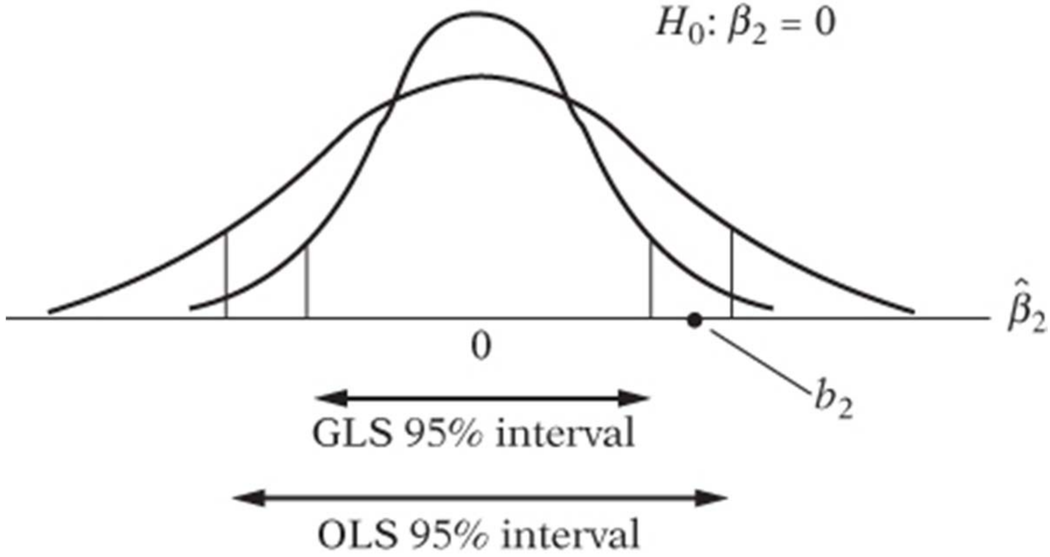
- In GLS we incorporate any additional information we have directly into the estimating procedure by transforming the variables, whereas in OLS such side information is not directly taken into consideration

- Under autocorrelation, it is the GLS estimator that is BLUE

Consequences of Using OLS in the Presence of Autocorrelation

OLS estimation allowing for autocorrelation

$\hat{\beta}_2$ is not BLUE and even if we use $\text{var}(\hat{\beta}_2)_{AR1}$, the confidence intervals derived from there are likely to be wider than those based on the GLS procedure



OLS estimation disregarding autocorrelation

The situation is potentially very serious if we not only use $\hat{\beta}_2$ but also continue to use $\operatorname{var}(\hat{\beta}_2) = \sigma^2 / \sum (X_t - \bar{X})^2$, which completely disregards the problem of autocorrelation, that is, we mistakenly believe that the usual assumptions of the classical model hold true. Errors will arise for the following reasons

1. The residual variance $\hat{\sigma}^2 = \frac{\sum \hat{u}_t^2}{(n-2)}$ is likely to underestimate the true σ^2
2. As a result, we are likely to overestimate R^2
3. Even if σ^2 is not underestimated, $\text{var}(\hat{\beta}_2)$ may underestimated $\text{var}(\hat{\beta}_2)_{AR1}$ its variance under AR(1), even though the latter is inefficient compared to $\text{var}(\hat{\beta}_2)^{GLS}$
4. Therefore, the usual t and F tests of significance are no longer valid

Example

To see how OLS is likely to underestimate σ^2 and the variance of $\hat{\beta}_2$, let us conduct the following **Monte Carlo experiment**

Suppose in the two-variable model

$$Y_t = 1.0 + 0.8X_t + u_t$$

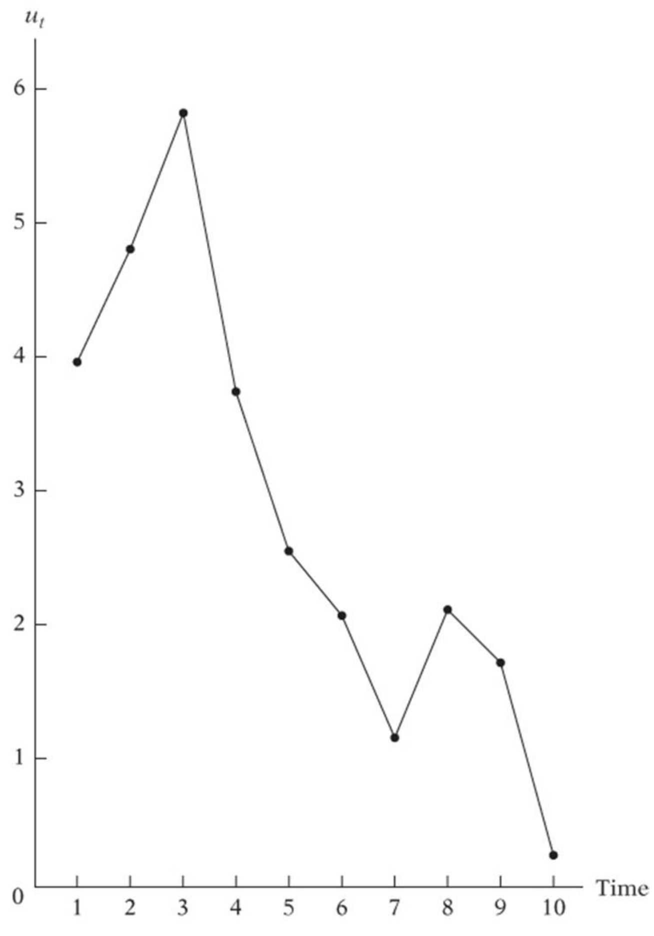
$$E(Y_t \mid X_t) = 1.0 + 0.8X_t$$

$$u_t = 0.7u_{t-1} + \varepsilon_t, \qquad \varepsilon_t \sim N(0,1)$$

TABLE 12.1
A Hypothetical Example of Positively Autocorrelated Error Terms

	ε_t	$u_t = 0.7u_{t-1} + \varepsilon_t$
0	0	$u_0 = 5$ (assumed)
1	0.464	$u_1 = 0.7(5) + 0.464 = 3.964$
2	2.026	$u_2 = 0.7(3.964) + 2.0262 = 4.8008$
3	2.455	$u_3 = 0.7(4.8010) + 2.455 = 5.8157$
4	−0.323	$u_4 = 0.7(5.8157) - 0.323 = 3.7480$
5	−0.068	$u_5 = 0.7(3.7480) - 0.068 = 2.5556$
6	0.296	$u_6 = 0.7(2.5556) + 0.296 = 2.0849$
7	−0.288	$u_7 = 0.7(2.0849) - 0.288 = 1.1714$
8	1.298	$u_8 = 0.7(1.1714) + 1.298 = 2.1180$
9	0.241	$u_9 = 0.7(2.1180) + 0.241 = 1.7236$
10	−0.957	$u_{10} = 0.7(1.7236) - 0.957 = 0.2495$

Note: ε_t data obtained from *A Million Random Digits and One Hundred Thousand Deviates,* Rand Corporation, Santa Monica, Calif., 1950.



u_t
6
5
4
3
2
1
0
1
2
3
4
5
6
7
8
9
10
Time

Now suppose the values of X are fixed at 1,2,3,…,10.

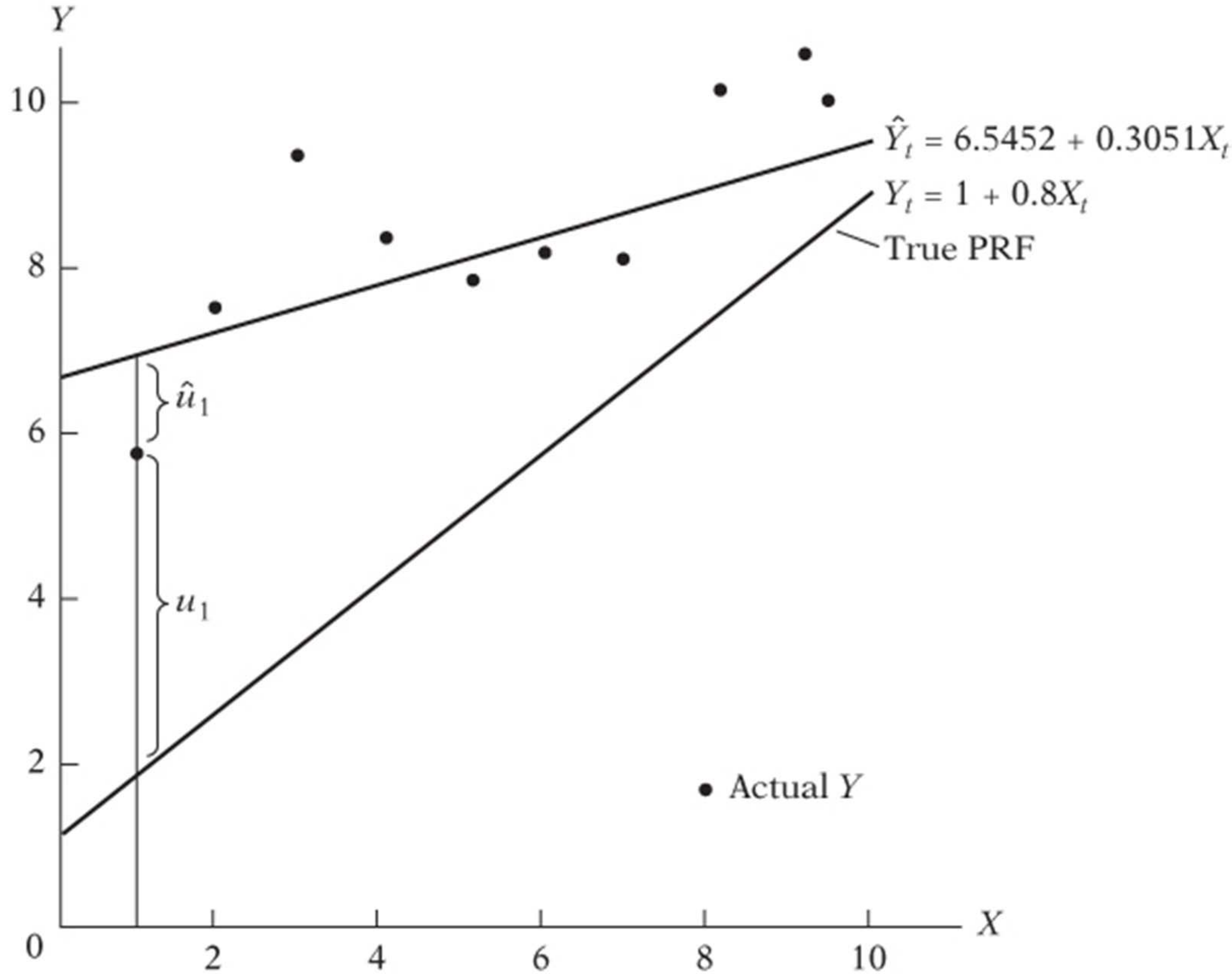
$$\hat{Y}_t = 6.5452 + 0.3051X_t$$

$$(0.6153) \quad (0.0992)$$

$$t = (10.6366) \ (3.0763)$$

$$r^2 = 0.5419$$

$$\hat{\sigma}^2 = 0.8114$$



Y
10
8
6
4
2
0
2
4
6
8
10
X
$\hat{Y}_t = 6.5452 + 0.3051X_t$
$Y_t = 1 + 0.8X_t$
True PRF
$\hat{u}_1$
u_1
Actual Y

TABLE 12.2
Generation of *Y* Sample Values

X_t	u_t	$Y_t = 1.0 + 0.8X_t + u_t$
1	3.9640	$Y_1 = 1.0 + 0.8(1) + 3.9640 = 5.7640$
2	4.8010	$Y_2 = 1.0 + 0.8(2) + 4.8008 = 7.4008$
3	5.8157	$Y_3 = 1.0 + 0.8(3) + 5.8157 = 9.2157$
4	3.7480	$Y_4 = 1.0 + 0.8(4) + 3.7480 = 7.9480$
5	2.5556	$Y_5 = 1.0 + 0.8(5) + 2.5556 = 7.5556$
6	2.0849	$Y_6 = 1.0 + 0.8(6) + 2.0849 = 7.8849$
7	1.1714	$Y_7 = 1.0 + 0.8(7) + 1.1714 = 7.7714$
8	2.1180	$Y_8 = 1.0 + 0.8(8) + 2.1180 = 9.5180$
9	1.7236	$Y_9 = 1.0 + 0.8(9) + 1.7236 = 9.9236$
10	0.2495	$Y_{10} = 1.0 + 0.8(10) + 0.2495 = 9.2495$

Note: u_t data obtained from Table 12.1.

Keeping the X_t and ε_t given in Table 2.1 and Table 2.2, let us assume $\rho = 0$, that is, no autocorrelation

TABLE 12.3
Sample of Y Values with Zero Serial Correlation

X_t	$\varepsilon_t = u_t$	$Y_t = 1.0 + 0.8X_t + \varepsilon_t$
1	0.464	2.264
2	2.026	4.626
3	2.455	5.855
4	−0.323	3.877
5	−0.068	4.932
6	0.296	6.096
7	−0.288	6.312
8	1.298	8.698
9	0.241	8.441
10	−0.957	8.043

Note: Since there is no autocorrelation, the u_t and ε_t are identical. The ε_t are from Table 12.1.

The regression based on Table 12.3 is as follows:

$$\hat{Y}_t = 2.5345 + 0.6145 X_t$$

$$(0.6796) \quad (0.1087)$$

$$t = (3.7910) \quad (5.6541)$$

$$r^2 = 0.7997$$

$$\hat{\sigma}^2 = 0.9752$$

Detecting Autocorrelation

- Durbin-Watson d Test

Durbin-Watson d Test

$$d = \frac{\sum_{t=2}^{t=n} (\hat{u}_t - \hat{u}_{t-1})^2}{\sum_{t=1}^{t=n} \hat{u}_t^2}$$

The ratio of the sum of squared differences in successive residuals to the RSS. Note that in the numerator of the d statistic the number of observations is n-1 because one observation is lost in taking successive differences

Assumptions underlying the d statistic

- The regression model includes the intercept term
- The explanatory variables, the X's are nonstochastic, or fixed in repeated sampling
- The disturbances are generated by the AR(1)

$$u_t = \rho u_{t-1} + \varepsilon_t$$

Therefore, it cannot be used to detect higher-order AR schemes

- The error term u_t is assumed to be normally distributed

- The regression model does not include the lagged value(s) of the dependent variable as one of the explanatory variables

- There are no missing observations in the data

$$d = \frac{\sum_{t=2}^{t=n} (\hat{u}_t - \hat{u}_{t-1})^2}{\sum_{t=1}^{t=n} \hat{u}_t^2}$$

$$d = \frac{\sum \hat{u}_t^2 + \sum \hat{u}_{t-1}^2 - 2\sum \hat{u}_t \hat{u}_{t-1}}{\sum \hat{u}_t^2}$$

$$d \approx 2\left(1 - \frac{\sum \hat{u}_t \hat{u}_{t-1}}{\sum \hat{u}_t^2}\right)$$

Since $\sum \hat{u}_t^2$ and $\sum \hat{u}_{t-1}^2$ differ in only one observation, they are approximately equal. Therefore, setting $\sum \hat{u}_{t-1}^2 \approx \sum \hat{u}_t^2$

$$\hat{\rho} = \frac{\sum \hat{u}_t \hat{u}_{t-1}}{\sum \hat{u}_t^2}$$

$$d \approx 2(1 - \hat{\rho})$$

But since $-1 \leq \rho \leq 1$, $0 \leq d \leq 4$

The mechanics of the Durbin-Watson test are

1. Run the OLS regression and obtain the residuals
2. Compute d
3. For the given sample size and given number of explanatory variables, find out the critical d_L and d_U values
4. Now follow the decision rules given in Table 12.6

Durbin-Watson d statistic

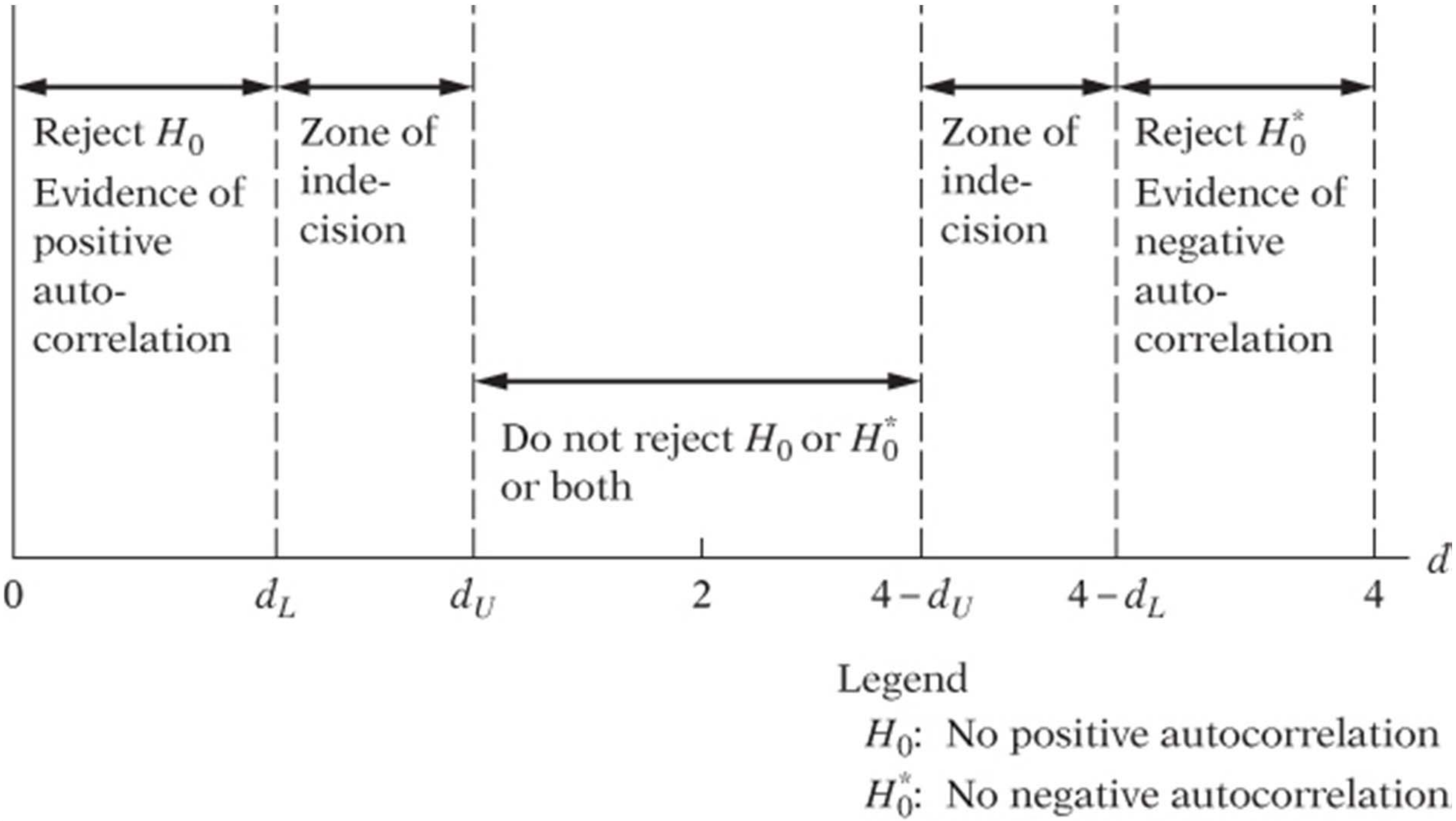


TABLE 12.6
Durbin–Watson *d* Test: Decision Rules

Null Hypothesis	Decision	If
No positive autocorrelation	Reject	$0 < d < d_L$
No positive autocorrelation	No decision	$d_L \leq d \leq d_U$
No negative correlation	Reject	$4 - d_L < d < 4$
No negative correlation	No decision	$4 - d_U \leq d \leq 4 - d_L$
No autocorrelation, positive or negative	Do not reject	$d_U < d < 4 - d_U$

Example

- U.S. Consumption Expenditure for the period 1947-2000

$$\ln Consumption = \beta_1 + \beta_2 \ln income + \beta_3 \ln wealth + \beta_4 Interest$$

$$\ln Consumption = -0.4677 + 0.8049 \ln income + 0.2013 \ln wealth + 0.0027 Interest$$

Source	SS	df	MS
Model	16.1637474	3	5.3879158
Residual	.007120721	50	.000142414
Total	16.1708681	53	.305110719

Number of obs	= 54
F(3, 50)	=37832.66
Prob > F	= 0.0000
R-squared	= 0.9996
Adj R-squared	= 0.9995
Root MSE	= .01193

lnc	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
lnincome	.8048728	.0174978	46.00	0.000	.7697273	.8400182
lnwealth	.2012702	.0175926	11.44	0.000	.1659345	.236606
i	-.0026891	.0007619	-3.53	0.001	-.0042194	-.0011587
_cons	-.467712	.042778	-10.93	0.000	-.5536342	-.3817899

Hypothesis testing

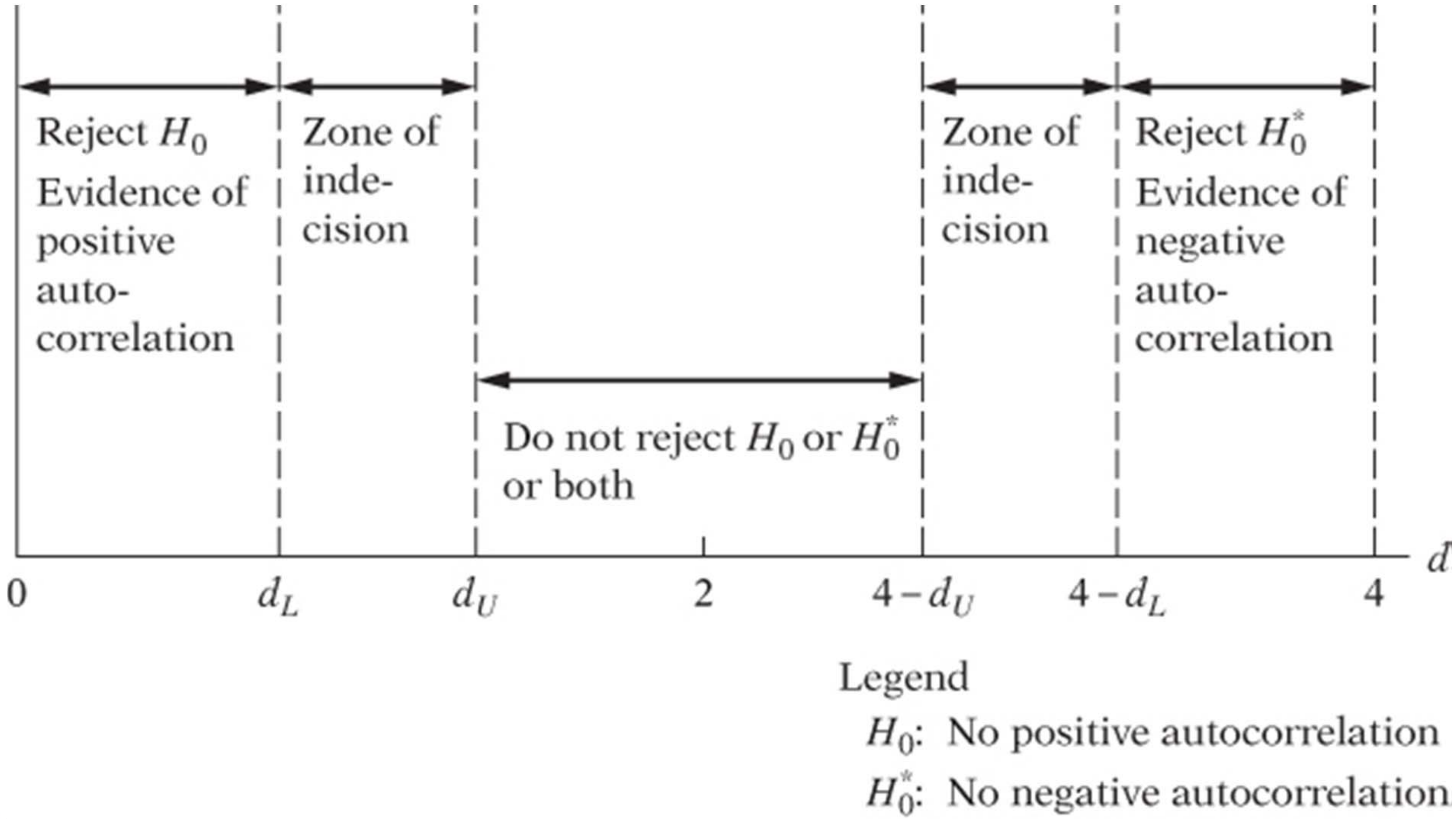
$$H_0 : No \ \ positive \ autocorrelation$$

$$H_0^* : No \ negative \ autocorrelation$$

$$H_1 : otherwise$$

$$H_1^* : otherwise$$

Durbin-Watson d statistic



Durbin-Watson test

```
. tsset year
        time variable:  year, 1947 to 2000
                delta:  1 unit

. estat dwatson

Durbin-Watson d-statistic(  4,    54) =  1.289232
```

$$(n = 54, k = 4)\ \ At\ \ 5\%\ \ Significance\ level$$

$$d_L = 1.414$$

$$d_U = 1.724$$

Reject null hypothesis

Evidence of positive autocorrelation

Correcting for (Pure) Autocorrelation

Correcting for (Pure) Autocorrelation

1. **The method of Generalized Least Squares (GLS) method**

2. The Newey West method – to obtain standard errors of OLS estimators that are corrected for autocorrelation

The method of Generalized Least Squares (GLS)

The two-variable regression model:

$$Y_t = \beta_1 + \beta_2 X_t + u_t$$

and assume that the error term follows the AR(1) scheme, namely,

$$u_t = \rho u_{t-1} + \varepsilon_t, \qquad -1 < \rho < 1$$

Now we consider two cases:

(1) ρ is known

(2) ρ is not known but has to be estimated

When ρ is known

$$Y_{t-1} = \beta_1 + \beta_2 X_{t-1} + u_{t-1}$$

$$\rho Y_{t-1} = \rho\beta_1 + \rho\beta_2 X_{t-1} + \rho u_{t-1}$$

$$(Y_t - \rho Y_{t-1}) = \beta_1(1-\rho) + \beta_2(X_t - \rho X_{t-1}) + (u_t - \rho u_{t-1})$$

$$(Y_t - \rho Y_{t-1}) = \beta_1(1-\rho) + \beta_2(X_t - \rho X_{t-1}) + \varepsilon_t$$

$$where\ \varepsilon_t = (u_t - \rho u_{t-1})$$

$$Y_t^* = \beta_1^* + \beta_2^* X_t^* + \varepsilon_t$$

Since the error term in

$$Y_t^* = \beta_1^* + \beta_2^* X_t^* + \varepsilon_t$$

satisfies the usual OLS assumptions, we can apply OLS to the transformed variables X^* and Y^* and obtain estimators with all the optimum properties, namely, BLUE.

When ρ is not known

- based on Durbin-Watson d Statistic

ρ based on Durbin-Watson d Statistic

$$\hat{\rho} \approx 1 - \frac{d}{2}$$

In reasonably large sample one can obtain ρ and use it to transform the data as shown in the generalized difference equation

$$(Y_t - \rho Y_{t-1}) = \beta_1(1-\rho) + \beta_2(X_t - \rho X_{t-1}) + \varepsilon_t$$

$$where\ \varepsilon_t = (u_t - \rho u_{t-1})$$

The Newey West method

The Newey West method of correcting the OLS standard errors

The corrected standard errors are known as **HAC (Heteroscedasticity and autocorrelation-consistent) standard errors** or simply **Newey West standard errors**

The Newey West method

```
. newey  lnc lnincome lnwealth i, lag(3)

Regression with Newey-West standard errors          Number of obs  =        54
maximum lag: 3                                      F(  3,     50)  =  22341.20
                                                    Prob > F       =    0.0000
```

lnc	Coef.	Newey-West Std. Err.	t	P>|t|	[95% Conf.	Interval]
lnincome	.8048728	.0171172	47.02	0.000	.7704919	.8392536
lnwealth	.2012702	.0154469	13.03	0.000	.1702441	.2322963
i	-.0026891	.0008798	-3.06	0.004	-.0044563	-.0009219
_cons	-.467712	.0439367	-10.65	0.000	-.5559616	-.3794625

Source

Gujarati, D.N. (2009) Basic Econometrics. 5th ed. Singapore, McGraw-Hill.