ACE 564
Spring 2006

Lecture 10: Violations of Basic Assumptions: Autocorrelation

by
Professor Scott H. Irwin

Required Readings:

Griffiths, Hill and Judge. "An Autocorrelated Error Model," Chapter 16 in *Learning and Practicing Econometrics*

Kennedy. "Autocorrelated Disturbances," Section 8.4 in *A Guide to Econometrics*

Nature of Autocorrelation

For efficiency (most precise estimation; "best"), all systematic information needs to be incorporated into the regression model

Heteroskedasticity is one type of systematic pattern in regression errors

Autocorrelation is another systematic pattern in regression errors (auto = self)

- It may be either positive (attracting) or negative (repelling)
- When we have time-series data, where the observations follow a natural ordering through time, there is always a possibility that successive errors will be correlated with each other
- In any one period, the current error term may contain not only the effects of current shocks but also the carryover from previous shocks
- This carryover will be related to, or correlated with, the effects of the earlier shocks

Positive Autocorrelation: errors do not cross zero-line enough (attracting)

No Autocorrelation: errors cross zero-line randomly

Negative Autocorrelation: errors cross zero-line too much (repelling)

The possibility of autocorrelation should always be entertained when we are dealing with time-series data.

This pattern violates one of the fundamental assumptions of the linear regression model

The violation is captured in the assumptions about the statistical properties of the regression error term as follows,

$$E(e_t) = 0 \qquad \text{var}(e_t) = \sigma^2$$
$$\text{cov}(e_t e_s) \neq 0 \quad \forall t, s, t \neq s$$

The Problem

Need to model the area planted to sugarcane in a region of Bangladesh

Information on response of area planted to price is important for government policy and planning purposes

- Is there a need for additional milling capacity?
- The implications of a pricing policy on production?

The Economic Model

One way of modeling supply response for an agricultural crop is to relate quantity produced to price and other supply-related variables

- This was the approach used in modeling the wheat supply in a region of Australia in Lecture 9

Another way of modeling supply response is to directly relate area sown (planted) to price and other supply-related variables

- Quantity produced = (area sown) x (yield/area)
- Area sown is the part of quantity produced that is determined by farmer
- Yield will depend on weather, pests, and diseases

Concentrating on area sown has several advantages

- Eliminates yield uncertainty from the model
- Concentrates on a farmer's main decision variable
- No need to model technological change in yields

The main product that competes with sugarcane for acreage is jute

Hence, the relative price of sugarcane (P_s) to jute (P_j) is likely to be a key determinant of sugarcane acreage

A simple economic model is,

$$y_t = f(P_s, P_j)$$

Assuming a double-log functional form, the economic model is,

$$\ln y_t = \beta_1 + \beta_2 \ln x_t$$

where

- y_t is the area sown to sugarcane (1,000 hectares) in year t
- x_t is the relative price of sugarcane to jute (P_s/P_j) in year t

The Statistical Model

As before, we recognize that the economic model will not hold exactly each time period,

$$\ln y_t = \beta_1 + \beta_2 \ln x_t + e_t$$

We can estimate this regression model using LS,

$$\widehat{\ln y_t} = 6.120 + 1.0041 \ln x_t \qquad R^2 = 0.614$$
$$(0.214)\ (0.141) \qquad (s.e.)$$

Both intercept and slope estimates are significantly different from zero

Table 16.1 Data for Area Response for Sugarcane in Bangladesh

Area (1000 of hectares)	Price of Sugarcane (taka/tonne)	Price of Jute (taka/tonne)
29	73	970
71	108	940
42	94	930
90	107	970
72	110	1004
57	146	1102
44	132	931
61	171	816
42	186	988
26	174	888
88	182	805
80	183	1257
125	208	1072
232	239	884
125	237	1005
99	246	1114
250	240	630
91	297	1446
121	269	1006
162	297	1289
143	333	903
138	319	1119
230	347	963
128	343	1062
87	357	1185
124	388	1348
97	391	974
152	414	1023
197	421	1192
220	441	1075
171	448	1243
208	483	1043
237	457	1138
235	479	1223

Griffiths, William E., R. Carter Hill, George G. Judge. Learning and Practicing Econometrics. John Wiley & Sons, Inc. New York. 1993.

Sources of Autocorrelated Errors

Prolonged influence of shocks

- Influence of random shocks often occurs over several time periods (e.g. oil embargo, wars, strikes, etc.)

Inertia

- Past actions often have a strong effect on current actions

Data manipulation

- Published data often undergo revisions that smooth the true, underlying errors

Mis-specification

- Error term contains affects of excluded variables (e.g. future prices, government policies, input prices)

- If excluded variables are autocorrelated, then error term will be autocorrelated

To examine this issue for the area regression, we will analyze estimated errors for patterns

- Assume that characteristics of population errors are reflected in estimated, sample errors (we used same assumption when detecting heteroskedasticity)

- A more objective test will be considered later

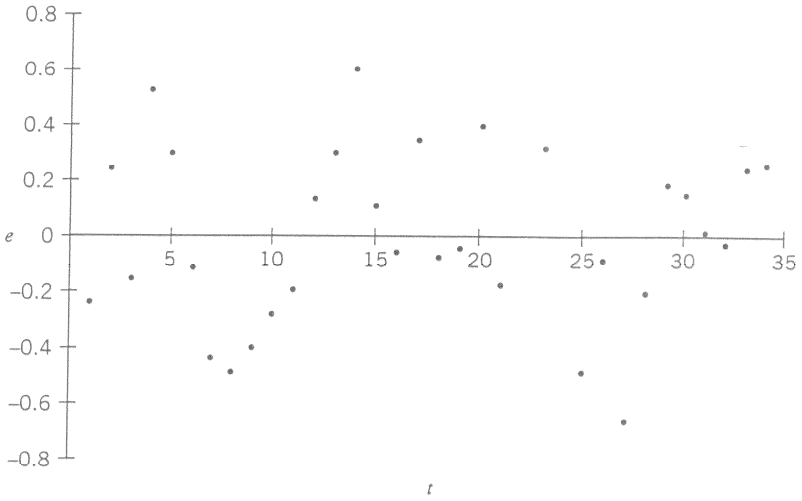


FIGURE 11.1 Least squares residuals plotted against time

Hill, C., W. Griffiths, and G. Judge. *Undergraduate Econometrics.* John Wiley & Sons, Inc., New York, NY 1997.

Order of Autoregression

Given the likely existence of autocorrelated errors, we need a specific assumption regarding the pattern in order to be of practical use

By far the most common model is the first-order autocorrelation model, or AR(1) for short,

$$e_t = \rho e_{t-1} + \upsilon_t$$

where:

- ρ is the correlation coefficient between e_t and e_{t-1}
- υ_t is a random error term with constant variance σ_υ^2

Note that υ_t has the all the properties we assumed earlier about e_t,

$$E(\upsilon_t) = 0, \quad \text{var}(\upsilon_t) = \sigma_\upsilon^2, \quad \text{cov}(\upsilon_t \upsilon_{t-1}) = 0 \quad t \neq s$$

Rationale for the model $e_t = \rho e_{t-1} + \upsilon_t$ is simple, with the random shock in time period t composed of two parts,

- ρe_{t-1} is the carryover of the random error from the previous period, where ρ determines degree of carryover
- υ_t is a "new" shock to the economic variable

The AR(1) model is the most basic form of an autoregressive model for the regression error term

We could have specified any of the following,

AR(2): $e_t = \rho_1 e_{t-1} + \rho_2 e_{t-2} + \upsilon_t$

AR(3): $e_t = \rho_1 e_{t-1} + \rho_2 e_{t-2} + \rho_3 e_{t-3} + \upsilon_t$

AR(4): $e_t = \rho_1 e_{t-1} + \rho_2 e_{t-2} + \rho_3 e_{t-3} + \rho_4 e_{t-4} + \upsilon_t$

The lag specification is only limited by degrees of freedom

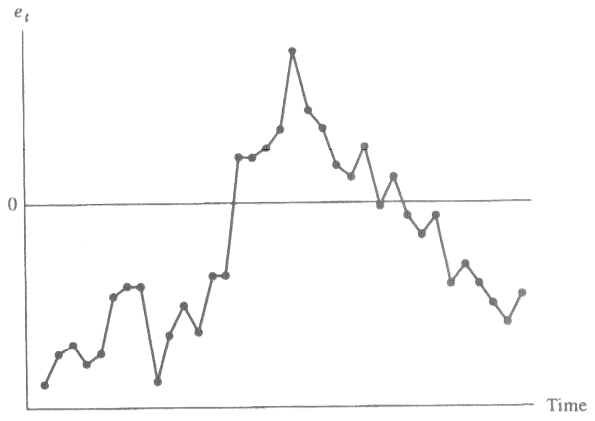


Figure 16.1 Error generated from an AR(1) process with $e_t = 0.9e_{t-1} + \varepsilon_t$.

Griffiths, William E., R. Carter Hill, George G. Judge. Learning and Practicing Econometrics. John Wiley & Sons, Inc. New York. 1993.

AR(1) is the most widely-applied model

- Simplest to estimate
- Seems to capture error correlation reasonably well for many economic applications
- Demonstrates principles without needless complication

Statistical Properties of an AR(1) Error

Must first assume that ρ is less than one in absolute value, or,

$$-1 < \rho < +1$$

Based on this assumption we can determine the mean of e_t,

$$E(e_t) = E(\rho e_{t-1} + \upsilon_t)$$

$$E(e_t) = \rho E(e_t) + 0$$

$$(1-\rho)E(e_t) = 0$$

$$E(e_t) = 0$$

Next, we can determine the variance of e_t,

$$\text{var}(e_t) = \sigma_e^2 = E(\rho e_{t-1} + \upsilon_t)^2$$

$$\sigma_e^2 = E(\rho^2 e_{t-1}^2 + \upsilon_t^2 + 2\rho e_{t-1}\upsilon_t)$$

$$\sigma_e^2 = \rho^2 E(e_{t-1}^2) + E(\upsilon_t^2) + E(2\rho e_{t-1}\upsilon_t)$$

$$\sigma_e^2 = \rho^2 \sigma_e^2 + \sigma_\nu^2$$

$$\sigma_e^2 = \frac{\sigma_\upsilon^2}{1-\rho^2}$$

It can be shown the covariances and correlations of e_t are,

$$\text{cov}(e_t e_{t-k}) = \sigma_e^2 \rho^k$$

$$corr(e_t e_{t-k}) = \rho^k$$

The last result implies that,

$$\rho^1 \geq \rho^2 \geq \rho^3 .....$$

or that, as the distance between errors increases, the correlation between them declines

Consequences of an AR(1) Error for the Least Squares Estimator

The consequences of ignoring autocorrelated errors are the same as those for heteroskedasticity,

1. The LS estimator is still a linear and unbiased estimator, but no longer the best linear unbiased estimator (no longer BLUE)

2. The standard errors usually computed for the least squares estimator are biased, and hence, confidence intervals and hypothesis tests that use these standard errors are misleading

To explore these issues, it is most straightforward to use the simple linear regression model,

$$y_t = \beta_1 + \beta_2 x_t + e_t \qquad t = 1, \ldots, T$$

where all "classical" assumptions hold except,

$$e_t = \rho e_{t-1} + \upsilon_t$$

Recall once again that the least squares estimator of the slope parameter in the simple linear regression model can be written as,

$$b_2 = \beta_2 + \sum_{t=1}^{T} w_t e_t$$

where

$$w_t = \frac{(x_t - \bar{x})}{\sum_{t=1}^{T} (x_t - \bar{x})^2}$$

The mean is derived by taking the expectation of b_2,

$$E(b_2) = E\left(\beta_2 + \sum_{t=1}^{T} w_t e_t \right)$$

$$E(b_2) = E(\beta_2) + \sum_{t=1}^{T} w_t E(e_t)$$

$$E(b_2) = \beta_2$$

This shows that β_2 is the mean of the sampling distribution of b_2, even when the variance of the error term is autocorrelated

We can derive the variance of b_2 as follows,

$$\text{var}(b_2) = \text{var}\left(\beta_2 + \sum_{t=1}^{T} w_t e_t \right) = \text{var}\left(\sum_{t=1}^{T} w_t e_t \right)$$

$$\text{var}(b_2) = \sum_{t=1}^{T} w_t^2 \, \text{var}(e_t) + \sum_{t=1}^{T} \sum_{s=1}^{T} w_t w_s \, \text{cov}(e_t, e_s) \qquad t \neq s$$

Noting that $\text{cov}(e_t e_s) = \sigma_e^2 \rho^k$, and substituting,

$$\text{var}(b_2) = \sum_{t=1}^{T} w_t^2 \, \text{var}(e_t) + \sum_{t=1}^{T} \sum_{s=1}^{T} w_t w_s \sigma_e^2 \rho^k \quad t \neq s \quad k = |t - s|$$

$$\text{var}(b_2) = \sum_{t=1}^{T} w_t^2 \, \text{var}(e_t) + \sigma_e^2 \sum_{t=1}^{T} \sum_{s=1}^{T} w_t w_s \rho^k \quad t \neq s \quad k = |t - s|$$

This can be re-written with a good bit of algebra as,

$$\text{var}(b_2) = \frac{\sigma_e^2}{\sum_{t=1}^{T} (x_t - \bar{x})^2} \left(1 + \frac{1}{\sum_{t=1}^{T} (x_t - \bar{x})^2} \sum_{t=1}^{T} \sum_{s=1}^{T} (x_t - \bar{x})(x_s - \bar{x}) \rho^k \right)$$

where $t \neq s \quad k = |t - s|$ in the double summation

Now compare,

$$\text{var}(b_2) = \frac{\sigma_e^2}{\sum_{t=1}^{T}(x_t - \bar{x})^2}\left[1 + \frac{1}{\sum_{t=1}^{T}(x_t - \bar{x})^2}\sum_{t=1}^{T}\sum_{s=1}^{T}(x_t - \bar{x})(x_s - \bar{x})\rho^k\right]$$

to the sampling variance in the non-autocorrelated case,

$$\text{var}(b_2) = \left[\frac{\sigma_e^2}{\sum_{t=1}^{T}(x_t - \bar{x})^2}\right]$$

($\sigma_e^2 = \sigma^2$ in our previous notation)

So, the sampling variance with autocorrelated errors is equal to the variance of the least squares estimator in the absence of autocorrelation times a factor that depends on the explanatory variable and ρ

In general, the variance of the least squares estimator in the absence of autocorrelation may be higher or lower than the variance of the least squares estimator in the presence of autocorrelation

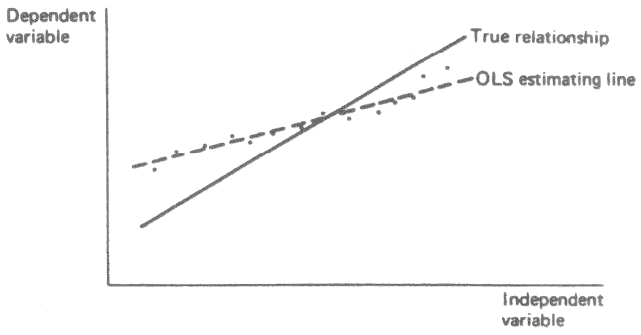


Figure 8.2 Illustrating positive autocorrelated errors

Kennedy, P. *Guide to Econometrics*, Fourth Edition. The MIT Press, Cambridge, Mass. 1998.

An Approximate Estimator for the Variance of Least Squares Standard Errors

One approach to the problem of autocorrelation is to seek "correct" standard error estimates for LS parameter estimates

One such method is based on replacing σ_e^2 with $\hat{\sigma}_e^2$ and ρ^k with $\hat{\rho}^k$ in the formula for the standard error

- The argument is that autocorrelation in the population regression model should lead to estimated residuals that are autocorrelated
- By approximating true errors with estimated LS errors, "correct" standard error estimates can be obtained
- Sometimes such standard errors are called "autocorrelation-consistent variance-covariance estimates"
- Most econometric packages have commands or options to compute autocorrelation-consistent standard errors; Excel does not

To derive the autcorrelation-consistent estimator, recall that the formula for the sampling variance of the least squares slope estimator,

$$\text{var}(b_2) = \frac{\sigma_e^2}{\sum_{t=1}^{T}(x_t - \bar{x})^2}\left(1 + \frac{1}{\sum_{t=1}^{T}(x_t - \bar{x})^2}\sum_{t=1}^{T}\sum_{s=1}^{T}(x_t - \bar{x})(x_s - \bar{x})\rho^k\right)$$

Assuming we have estimates of σ_e^2 and ρ, we simply substitute them,

$$\text{var}_{ac}(b_2) = \frac{\hat{\sigma}_e^2}{\sum_{t=1}^{T}(x_t - \bar{x})^2}\left(1 + \frac{1}{\sum_{t=1}^{T}(x_t - \bar{x})^2}\sum_{t=1}^{T}\sum_{s=1}^{T}(x_t - \bar{x})(x_s - \bar{x})\hat{\rho}^k\right)$$

We will learn later how to compute $\hat{\sigma}_e^2$ and $\hat{\rho}$

Sugar cane example:

$$\hat{\ln} y_t = 6.120 + 1.0041 \ln x_t \qquad R^2 = 0.614$$
$$(0.214)\ (0.141) \qquad (\text{"}incorrect\text{"}\ LS\ s.e.)$$
$$(0.319)\ (0.206) \qquad (\text{"}correct\text{"}\ s.e.)$$

Key points:

- Although autocorrelation can in theory lead to LS standard errors that are too big or too small, in practice, understatement of standard errors is typical!

- This means confidence intervals are too narrow and the significance of hypothesis tests is overstated

Generalized Least Squares: ρ is Known

In the previous section, we explored "correct" standard errors for LS estimators in the presence of an AR(1) error process

In parallel with our discussion of heteroskedasticity, we want to ask if it is possible to obtain an estimator for the regression parameters that is superior to LS or LS with "correct" standard errors

- Generalized least squares (GLS) will provide such an estimator

- Objective is to transform the regression model with AR(1) error so that the transformed error is uncorrelated and homoscedastic

We will once again begin with the simple model,

$$y_t = \beta_1 + \beta_2 x_t + e_t$$

where all "classical" assumptions hold except,

$$e_t = \rho e_{t-1} + \upsilon_t$$

and υ_t has "classical" *iid* properties

Objective: transform the model so that υ_t is the error, not e_t

Substitute the AR(1) model into the regression model,

$$y_t = \beta_1 + \beta_2 x_t + \rho e_{t-1} + \upsilon_t$$

We have eliminated e_t, but the model still contains e_{t-1}

To eliminate e_{t-1}, we note that the original model holds for all time periods, so we can write,

$$y_{t-1} = \beta_1 + \beta_2 x_{t-1} + e_{t-1}$$

or,

$$e_{t-1} = y_{t-1} - \beta_1 - \beta_2 x_{t-1}$$

The last result can be multiplied by ρ,

$$\rho e_{t-1} = \rho y_{t-1} - \rho\beta_1 - \rho\beta_2 x_{t-1}$$

This can be substituted into,

$$y_t = \beta_1 + \beta_2 x_t + \rho e_{t-1} + \upsilon_t$$

with the following result,

$$y_t = \beta_1 + \beta_2 x_t + \rho y_{t-1} - \rho\beta_1 - \rho\beta_2 x_{t-1} + \upsilon_t$$

Rearranging,

$$y_t - \rho y_{t-1} = \beta_1(1-\rho) + \beta_2(x_t - \rho x_{t-1}) + \upsilon_t$$

This transformed regression model has the desired uncorrelated and homoskedastic error term

We can re-state the transformed regression model as,

$$y_t^* = \beta_1 x_{1,t}^* + \beta_2 x_{2,t}^* + \upsilon_t$$

where

$$y_t^* = y_t - \rho y_{t-1} \qquad t = 2,3,...T$$

$$x_{1,t}^* = (1-\rho) \qquad t = 2,3,...T$$

$$x_{2,t}^* = x_t - \rho x_{t-1} \qquad t = 2,3,...T$$

What about the first observation?

Recovering the First Observation

Dropping the first observation and applying LS to the transformed model does not produce BLUE estimators

- Efficiency is lost because the variance of the error associated with the first observation is not equal to that of the other errors

- This is a special case of heteroskedasticity, where all errors have the same variance except the first error

  $\Rightarrow$ We must transform the first observation so that its error variance is the same as all other observations

The following transformation assures that the error variance of the first observation is equal to that of the other *T-1* observations,

$$y_1^* = \left(\sqrt{1-\rho^2}\right) y_1 \qquad x_{1,1}^* = \left(\sqrt{1-\rho^2}\right)$$

$$x_{2,1}^* = \left(\sqrt{1-\rho^2}\right) x_1$$

We can add the transformed first observation to the other transformed *T-1* observations to obtain the fully transformed model,

$$y_t^* = \beta_1 x_{1,t}^* + \beta_2 x_{2,t}^* + \upsilon_t$$

where the first observation is,

$$y_1^* = \left(\sqrt{1-\rho^2}\right) y_1$$

$$x_{1,1}^* = \left(\sqrt{1-\rho^2}\right)$$

$$x_{2,1}^* = \left(\sqrt{1-\rho^2}\right) x_1$$

and the remaining $t=2,3,...T$ observations are,

$$y_t^* = y_t - \rho y_{t-1} \qquad t = 2,3,...T$$

$$x_{1,t}^* = (1-\rho) \qquad t = 2,3,...T$$

$$x_{2,t}^* = x_t - \rho x_{t-1} \qquad t = 2,3,...T$$

Key points:

- Transformed model is linear in the parameters
- LS estimators b_1^* and b_2^* are BLUE
- b_1^* and b_2^* from the transformed model are the correct estimates of the intercept and slope of the original model
- Because an intercept is not included in the transformed model, R^2 cannot be interpreted in the usual manner

Transformation of variables should be viewed as a device for converting an autocorrelated error model into an uncorrelated error model, not as something that changes the meaning of the coefficients

Estimated Generalized Least Squares: ρ is Unknown

Once again, the normal situation is that a key parameter is unknown

In the case of heteroskedasticity, we could use some "tricks" in certain cases to get around a similar problem

There are no "tricks" in the case of autocorrelation, and therefore, we must estimate ρ

Consider again the AR(1) error model,

$$e_t = \rho e_{t-1} + \upsilon_t$$

At first glance it would appear that we could estimate this equation by LS

However, neither e_t or e_{t-1} are observable because they depend on the unknown parameters β_1 and β_2 as follows,

$$e_t = y_t - \beta_1 - \beta_2 x_t$$

As an approximation, we can use the LS residuals,

$$\hat{e}_t = y_t - b_1 - b_2 x_t$$

as substitutes,

$$\hat{e}_t = \rho \hat{e}_{t-1} + \hat{v}_t$$

The LS estimator of ρ in the above equation is,

$$\hat{\rho} = \frac{\sum_{t=2}^{T} \hat{e}_t \hat{e}_{t-1}}{\sum_{t=2}^{T} \hat{e}_{t-1}^2}$$

After obtaining this "first step" estimate , we can use it to transform the variables according to the GLS procedures

Then, as a "second step," we can apply LS to the transformed variables

Key Points on Two-Step Estimation Procedure

1. Since $\hat{\rho}$ is estimated, application of LS to transformed variables will NOT yield BLUE estimators

2. Application of LS to transformed variables will yield "consistent" estimators that are valid if the sample size is "large"

   - Implies we must be quite cautious about precision of estimates if sample is not "large"

   - Monte Carlo studies also suggest that it is better to use LS if $\rho < 0.3$

3. Two-step procedure is one application of estimated generalized least squares (EGLS)

4. Two-step procedure is also known as Cochrane-Orcutt method

Sugar Cane Regression Models

LS

$$\widehat{\ln} y_t = 6.120 + 1.0041 \ln x_t \qquad R^2 = 0.614$$
$$(0.214)\ (0.141) \qquad (\text{"}incorrect\text{"}\ LS\ s.e.)$$
$$(0.319)\ (0.206) \qquad (\text{"}correct\text{"}\ s.e.)$$

EGLS

$$\hat{\rho} = \frac{\sum_{t=2}^{T} \hat{e}_t \hat{e}_{t-1}}{\sum_{t=2}^{T} \hat{e}_{t-1}^2} = 0.4501$$

$$\widehat{\ln} y_t = 6.205 + 1.060 \ln x_t$$
$$(0.286)\ (0.181)$$

Comparison of 95% confidence intervals for β_2

AR(1): [0.691, 1.429]
Auto-Cons.: [0.584, 1.424]
LS: [0.717, 1.291]

Sugar Cane Example: EGLS Transformation of Data in Excel to Correct for Autocorrelation

y	Ps	Pj	x=Ps/Pj	ln(y)	ln(x)	ln(y)*	x1*	x2*=ln(x)*
29	73	970	0.0753	3.3673	-2.5868	3.0069	0.8930	-2.3100
71	108	940	0.1149	4.2627	-2.1637	2.7471	0.5499	-0.9994
42	94	930	0.1011	3.7377	-2.2919	1.8190	0.5499	-1.3180
90	107	970	0.1103	4.4998	-2.2045	2.8175	0.5499	-1.1729
72	110	1004	0.1096	4.2767	-2.2113	2.2513	0.5499	-1.2190
57	146	1102	0.1325	4.0431	-2.0213	2.1181	0.5499	-1.0260
44	132	931	0.1418	3.7842	-1.9535	1.9644	0.5499	-1.0437
61	171	816	0.2096	4.1109	-1.5628	2.4076	0.5499	-0.6835
42	186	988	0.1883	3.7377	-1.6699	1.8874	0.5499	-0.9665
26	174	888	0.1959	3.2581	-1.6299	1.5758	0.5499	-0.8783
88	182	805	0.2261	4.4773	-1.4868	3.0109	0.5499	-0.7532
80	183	1257	0.1456	4.3820	-1.9270	2.3668	0.5499	-1.2578
125	208	1072	0.1940	4.8283	-1.6397	2.8560	0.5499	-0.7724
232	239	884	0.2704	5.4467	-1.3080	3.2735	0.5499	-0.5699
125	237	1005	0.2358	4.8283	-1.4447	2.3767	0.5499	-0.8560
99	246	1114	0.2208	4.5951	-1.5104	2.4219	0.5499	-0.8601
250	240	630	0.3810	5.5215	-0.9651	3.4532	0.5499	-0.2853
91	297	1446	0.2054	4.5109	-1.5828	2.0256	0.5499	-1.1484
121	269	1006	0.2674	4.7958	-1.3190	2.7655	0.5499	-0.6066
162	297	1289	0.2304	5.0876	-1.4679	2.9290	0.5499	-0.8742
143	333	903	0.3688	4.9628	-0.9976	2.6729	0.5499	-0.3369
138	319	1119	0.2851	4.9273	-1.2550	2.6935	0.5499	-0.8060
230	347	963	0.3603	5.4381	-1.0207	3.2203	0.5499	-0.4559
128	343	1062	0.3230	4.8520	-1.1302	2.4044	0.5499	-0.6707
87	357	1185	0.3013	4.4659	-1.1998	2.2820	0.5499	-0.6911
124	388	1348	0.2878	4.8203	-1.2454	2.8102	0.5499	-0.7054
97	391	974	0.4014	4.5747	-0.9127	2.4051	0.5499	-0.3522
152	414	1023	0.4047	5.0239	-0.9046	2.9648	0.5499	-0.4938
197	421	1192	0.3532	5.2832	-1.0408	3.0220	0.5499	-0.6336
220	441	1075	0.4102	5.3936	-0.8910	3.0157	0.5499	-0.4226
171	448	1243	0.3604	5.1417	-1.0205	2.7140	0.5499	-0.6194
208	483	1043	0.4631	5.3375	-0.7698	3.0233	0.5499	-0.3105
237	457	1138	0.4016	5.4681	-0.9123	3.0656	0.5499	-0.5658
235	479	1223	0.3917	5.4596	-0.9374	2.9984	0.5499	-0.5267

Sugar Cane Example: EGLS Regression Results in Excel with Autocorrelation Correction

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.6528
R Square	0.4262
Adjusted R Square	0.3770
Standard Error	0.3507
Observations	34

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	2	2.9228	1.4614	11.8839	0.0001
Residual	32	3.9352	0.1230		
Total	34	6.8580			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	0.0000	#N/A	#N/A	#N/A	#N/A	#N/A
X Variable 1	6.2053	0.2857	21.7223	0.0000	5.6234	6.7872
X Variable 2	1.0604	0.1813	5.8478	0.0000	0.6910	1.4297

Detecting Autocorrelation

As with heteroskedasticity, the best place to start is a graph

Plot LS residuals and examine whether they have any pattern

- Plot $\hat{e}_t$ against time
- Scatterplot of $\hat{e}_t$ vs. $\hat{e}_{t-1}$
- If the estimated errors are uncorrelated, there should be no pattern of any kind
- If a pattern is detected with visual inspection, then more formal analysis is appropriate

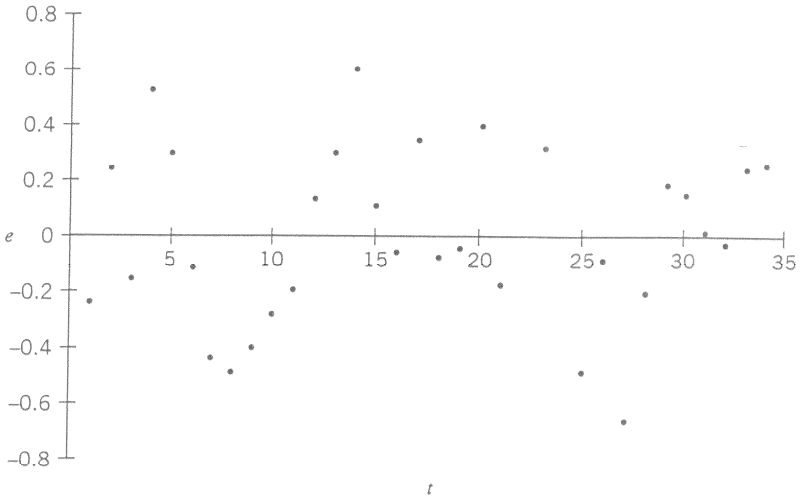
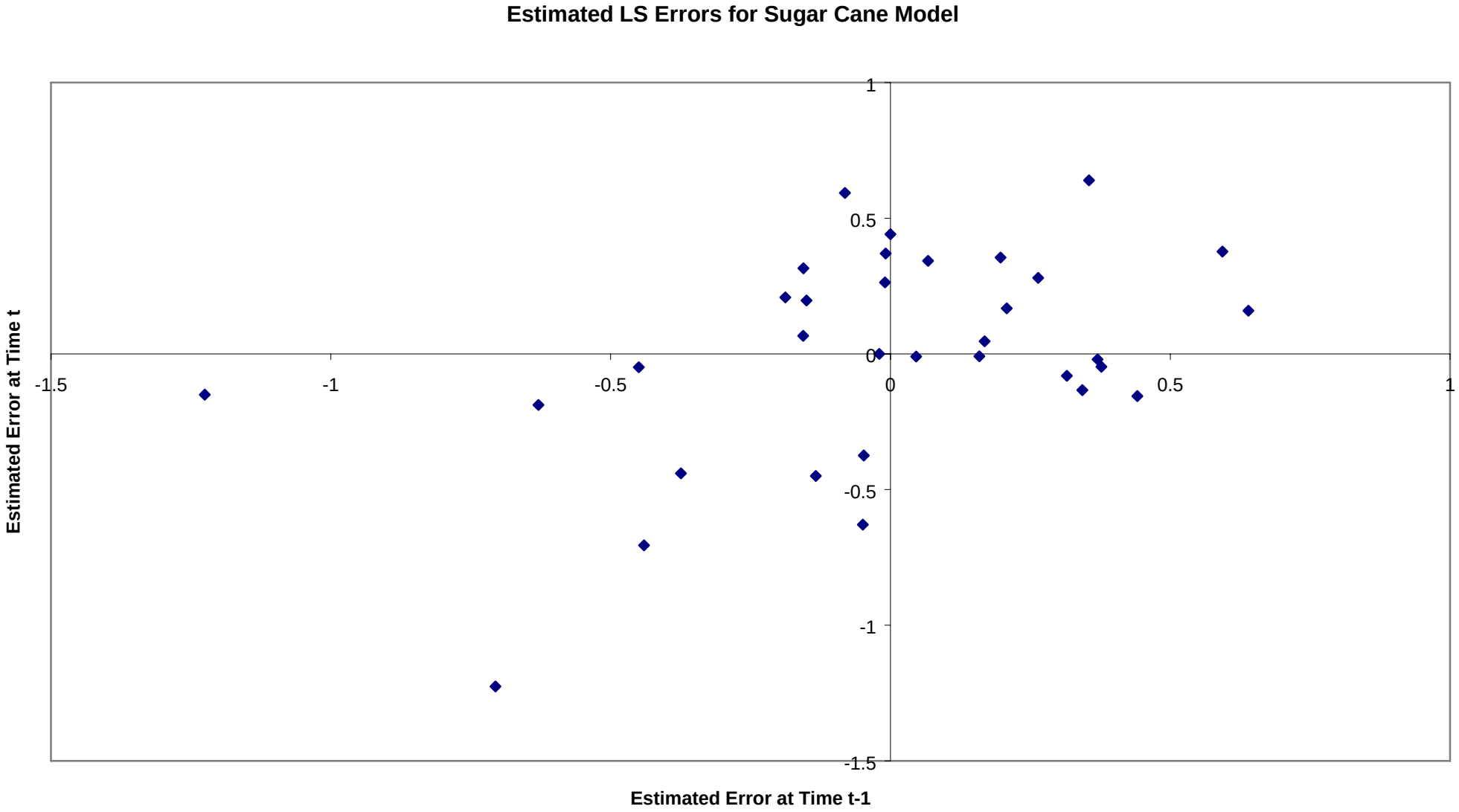


FIGURE 11.1 Least squares residuals plotted against time

Hill, C., W. Griffiths, and G. Judge. *Undergraduate Econometrics.* John Wiley & Sons, Inc., New York, NY 1997.


Estimated LS Errors for Sugar Cane Model
Estimated Error at Time t
Estimated Error at Time t-1
1
0.5
0
-0.5
-1
-1.5
-1.5
-1
-0.5
0
0.5
1

The Durbin-Watson Test

The Durbin-Watson (DW) test is by far the most widely used statistical test for the presence of autocorrelation

To develop the DW test, consider once again the two-variable regression model,

$$y_t = \beta_1 + \beta_2 x_t + e_t$$

where all "classical" assumptions hold except,

$$e_t = \rho e_{t-1} + \upsilon_t$$

and υ_t is an *iid* random error

Let's now consider the following hypotheses,

$$H_0 : \rho \leq 0$$
$$H_1 : \rho > 0$$

We use this setup because positive autocorrelation is the most likely alternative with time-series economic data

It would seem natural at this point to compute $\hat{\rho}$ and test whether it is significantly greater than zero

- However, derivation of the exact probability distribution of $\hat{\rho}$ turns out to be quite difficult

- Consequently, Durbin and Watson choose to work with a different, but closely related, statistic that has a more easily derived probability distribution

DW d statistic,

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

To see the relationship between d and $\hat{\rho}$, we can re-write d as,

$$d = \frac{\sum_{t=2}^{T} \hat{e}_t^2 + \sum_{t=2}^{T} \hat{e}_{t-1}^2 - 2\sum_{t=2}^{T} \hat{e}_t \hat{e}_{t-1}}{\sum_{t=1}^{T} \hat{e}_t^2}$$

$$d = \frac{\sum_{t=2}^{T} \hat{e}_t^2}{\sum_{t=1}^{T} \hat{e}_t^2} + \frac{\sum_{t=2}^{T} \hat{e}_{t-1}^2}{\sum_{t=1}^{T} \hat{e}_t^2} - \frac{2\sum_{t=2}^{T} \hat{e}_t \hat{e}_{t-1}}{\sum_{t=1}^{T} \hat{e}_t^2}$$

$$d \approx \quad 1 \quad + \quad 1 \quad - 2\hat{\rho}$$

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

Finally,

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

Consequently, we can (approximately) place some bounds on the relationship between d and $\hat{\rho}$,

Value of $\hat{\rho}$	**Value of d**
-1	4
0	2
1	0

Many econometric packages automatically print out the DW test statistic

Excel does not print out d automatically, so it must be computed "by hand"

Normally, we would set the alpha value of the test, determine the probability distribution of the test statistic, and look up the critical value from a table

The situation for the DW *d*-statistic is more complicated, as the probability distribution of *d*, *f(d)*, depends on the values of the independent variables

- It is impossible to tabulate critical values that can be used for every possible problem.
- Need a different table for each different data set!

There are two ways to overcome this problem

Exact p-value

The first way is to use software that computes the exact *p*-value for the explanatory variables in the model under consideration

Instead of comparing the calculated *d* value with some tabulated values of d_c, econometric software calculates the *p*-value of the test

If this *p*-value is less than the specified significance level, the null hypothesis is rejected and we conclude that autocorrelation does exist

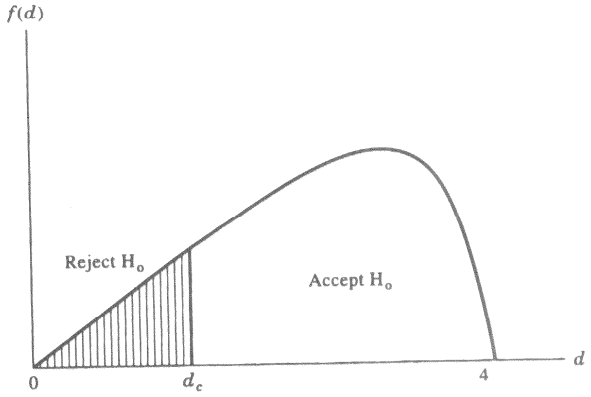


Figure 16.2 Testing for positive autocorrelation.

Griffiths, William E., R. Carter Hill, George G. Judge. Learning and Practicing Econometrics. John Wiley & Sons, Inc. New York. 1993.

Testing for Autocorrelation in the Sugar Cane Regression Model using Exact p-value

1. Hypotheses

$$H_0 : \rho \leq 0$$
$$H_1 : \rho > 0$$

2. Test statistic

We apply LS to the sugarcane data to compute the test statistic d,

$$d = \frac{\sum_{t=2}^{T} (\hat{e}_t - \hat{e}_{t-1})^2}{\sum_{t=1}^{T} \hat{e}_t^2} = 1.0934$$

3. Rejection Region

Reject the null hypothesis if the p-value for d is less than α

4. Decision

Use econometric software (SHAZAM, SAS, etc.) to compute the *p*-value for *d*

$$p-value = P(d < 1.0934) = 0.00132$$

For $\alpha = 0.05$, the *p*-value is much smaller

Reject the null hypothesis and accept the alternative that errors are positively autocorrelated

DW Bounds

In the absence of software that computes a p-value, a test known as the bounds test can be used to partially overcome the problem of not having general critical values

Durbin and Watson considered two other statistics d_L and d_U whose probability distributions do not depend on the explanatory variables and which have the property that

$$d_L < d < d_U$$

That is, irrespective of the explanatory variables in the model under consideration, d will be bounded by an upper bound d_U and a lower bound d_L

Decision rules:

- If $d < d_{Lc}$, reject $H_0 : \rho \leq 0$ and accept $H_1 : \rho > 0$
- If $d > d_{Uc}$, do not reject $H_0 : \rho \leq 0$
- If $d_{Lc} < d < d_{Uc}$, the test is inconclusive

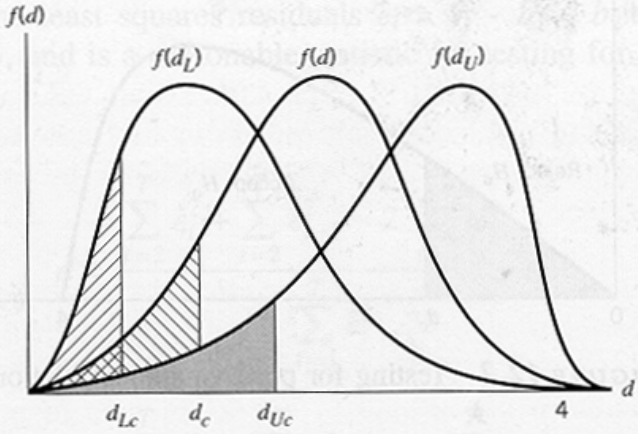


FIGURE 12.3 Upper and lower critical value bounds for the Durbin–Watson test.

Hill, R.C., W.E. Griffiths, G.G. Judge. *Undergraduate Econometrics*, Second Edition, John Wiley and Sons: Nbew York, 2001

Testing for Autocorrelation in the Sugar Cane Regression Model Using DW Bounds

1. Hypotheses

$$H_0 : \rho \leq 0$$
$$H_1 : \rho > 0$$

2. Test statistic

We apply LS to the sugarcane data to compute the test statistic d,

$$d = \frac{\sum_{t=2}^{T} (\hat{e}_t - \hat{e}_{t-1})^2}{\sum_{t=1}^{T} \hat{e}_t^2} = 1.0934$$

3. Rejection Region

To find the critical bounds for the sugar cane example we consult Table 5 at the end of LPE for $T = 34$ and $K = 2$

The values are: $d_{Lc} = 1.393$ $d_{Uc} = 1.514$

- If $d < d_{Lc} = 1.393$ reject $H_0 : \rho \leq 0$ and accept $H_1 : \rho > 0$
- If $d > d_{Uc} = 1.514$ do not reject $H_0 : \rho \leq 0$
- If $d_{Lc} = 1.393 < d < d_{Uc} = 1.514$ the test is inconclusive

4. Decision

- Since $d = 1.0934 < d_{Lc}$, we conclude that $d < d_c$, and hence reject H_0
- There is evidence to suggest that positive autocorrelation exists

Final Thoughts

It is now routine to report the DW *d*-statistic for regression results

"Full" report of standard regression output for LS regression using the sugar cane data,

$$\hat{\ln} y_t = 6.120 + 1.0041 \ln x_t \qquad R^2 = 0.614$$
$$(0.214)\ (0.141) \qquad (s.e.)$$
$$F = 50.921 \qquad d = 1.0934$$

Since the reporting of *d* is routine, it is important to understand the cases where it is not appropriate to use *d*,

- The regression model does not include an intercept
- The errors are correlated at higher orders than lag one
- Lagged values of the dependent variable are included in the regression
- The data include missing observations