# EE325 Ch.8 Autocorrelation

Read Guajarati Ch. 12

# Outline

1. Nature of Autocorrelation
2. Consequence of Autocorrelation
3. Detection of Autocorrelation
4. Remedial Measures

# Nature of Autocorrelation

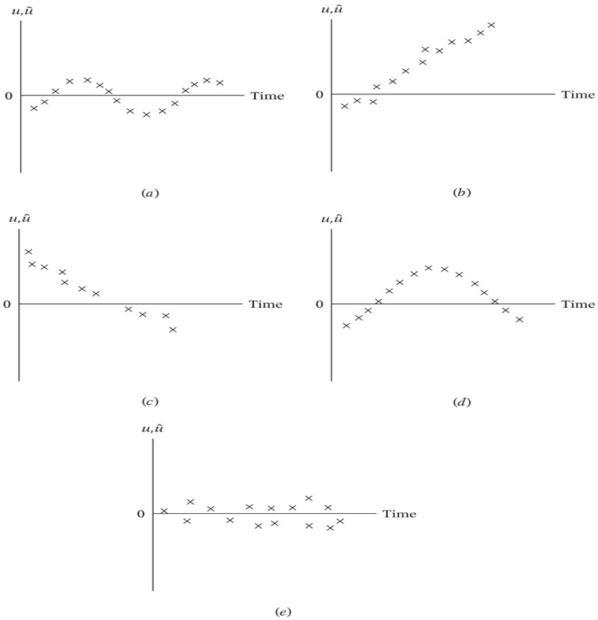
Autocorrelation may be defined as correlation between members of series of observations ordered in time (as in time series data) or space (as in cross-sectional data)

$$cov(u_i, u_j | X_i, X_j) = E(u_i, u_j) \neq 0$$

for $i \neq j$

- 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 $u_1, u_2, \ldots, u_{10}$ and $u_2, u_3, \ldots, u_{11}$, where the former is the latter series lagged by one time period

Serial correlation is the correlation between time series such as $u_1, u_2, \ldots, u_{10}$ and $\nu_2, \nu_3, \ldots, \nu_{11}$ where $u_i$ and $\nu_i$ are two different time series


u,û
0
Time
(a)
u,û
0
Time
(b)
u,û
0
Time
(c)
u,û
0
Time
(d)
u,û
0
Time
(e)

- Inertia

A salient feature of most economic time series is sluggishness

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

- Specification error – excluded variables

True model:

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

$Y_t$ = quantity of beef demand
$X_{2t}$ = price of beef
$X_{3t}$ = consumer income
$X_{4t}$ = price of pork

Instead, we model:

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

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

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

- Specification error – incorrect functional form

True model for cost-output study:

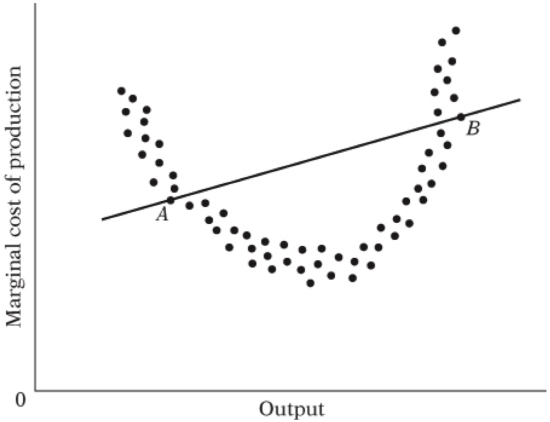
$$MC_i = \beta_1 + \beta_2 Output_{2i} + \beta_3 Output_{3i} + u_i$$

Instead, we model:

$$MC_i = \alpha_1 + \alpha_2 Output_{2i} + \nu_i$$

where $\nu_t = \beta_3 Output_{3i} + u_i$

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

$$Consumption_t = \beta_1 + \beta_2 Income_t + \beta_3 Consumption_{t-1} + u_t$$

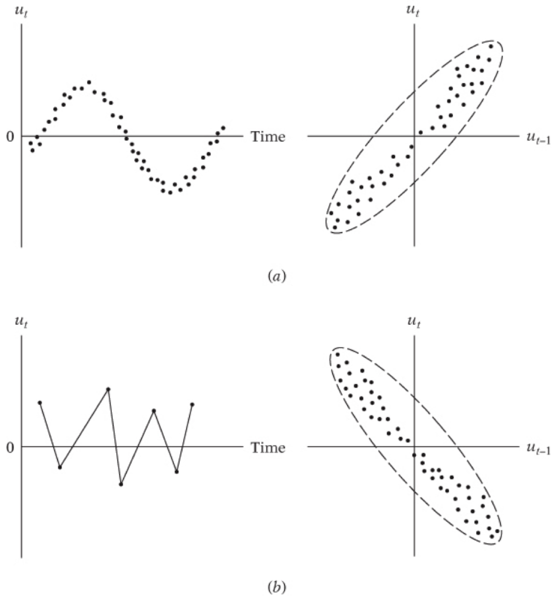
- Data Transformation

$$\begin{aligned}
\text{Level form} \quad Y_t &= \beta_1 + \beta_2 X_t + u_t \\
Y_{t-1} &= \beta_1 + \beta_2 X_{t-1} + u_{t-1} \\
\text{Difference form} \quad \Delta Y_t &= \beta_2 \Delta X_t + \Delta u_t \\
\Delta Y_t &= \beta_2 \Delta X_t + \nu_t \\
\text{where} \quad \nu_t &= \Delta u_t = (u_t - u_{t-1})
\end{aligned}$$

$$\begin{aligned}
\text{Suppose} \quad E(u_t) &= 0 \\
\text{Then} \quad E(\nu_t) &= E(u_t - u_{t-1}) = E(u_t) - E(u_{t-1})
\end{aligned}$$

$$\begin{aligned}
var(\nu_t) &= var(u_t - u_{t-1}) \\
&= var(u_t) - var(u_{t-1}) = 2\sigma^2
\end{aligned}$$

$$\begin{aligned}
cov(\nu_t, \nu_{t-1}) &= E(\nu_t - \nu_{t-1}) \\
&= E[u_t - u_{t-1}][u_{t-1} - u_{t-2}] = -\sigma^2
\end{aligned}$$


$u_t$
0
Time
$u_t$
$u_{t-1}$
(a)
$u_t$
0
Time
$u_t$
$u_{t-1}$
(b)

# Consequence of Autocorrelation

Characteristics of the Error Terms:

$$\begin{aligned} Y_i &= \beta_1 + \beta_2 X_i + u_i \\ u_t &= \rho u_{t-1} + \epsilon_t; \; -1 < \rho < 1 \end{aligned}$$

where $\rho$ is coefficient of autocovariance

Given

$$\begin{aligned} E(\epsilon_t) &= 0 \\ var(\epsilon_t) &= \sigma_\epsilon^2 \\ cov(\epsilon_t, \epsilon_{t+s}) &= 0 \end{aligned}$$

where $\epsilon$ is white-noise error term and $s \neq 0$

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

where $-1 < \rho < 1$

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

$$\begin{aligned} E(u_t) &= \rho E(u_{t-1}) + E(\epsilon_t) = 0 \\ var(u_t) &= E(u_t^2) \\ &= \rho^2 var(u_{t-1}) + var(\epsilon_t) = \frac{\sigma_\epsilon^2}{1-\rho^2} \end{aligned}$$

Note

$$\begin{aligned} var(u_t) &= var(u_{t-1}) = \sigma^2 \\ var(\epsilon_t) &= \sigma_\epsilon^2 \end{aligned}$$

Covariance between error term $s$ periods apart is
$cov(u_t, u_{t+s}) = E(u_t, u_{t+s}) = \rho^s \frac{\sigma_\epsilon^2}{1-\rho^2}$

The symmetry property of covariances follows
$cov(u_t, u_{t+s}) = cov(u_t, u_{t-s})$

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

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

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

Correlation between error terms $s$ periods apart is
$cor(u_t, u_{t+s}) = \rho^s$

The symmetry property of correlations follows
$cor(u_t, u_{t+s}) = cor(u_t, u_{t-s})$

Since $\rho$ 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 value several periods in the past

Normal OLS estimation:

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

OLS estimators are

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

and its variance is given by

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

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

$$var(\hat{\beta}_2)_{AR(1)} = \frac{\sigma^2}{\sum x_t^2}[1 + 2\rho\frac{\sum x_t x_{t-1}}{\sum x_t^2} + 2\rho^2\frac{\sum x_t x_{t-2}}{\sum x_t^2} + \cdots + 2\rho^{n-1}\frac{\sum x_t x_n}{\sum x_t^2}]$$

where

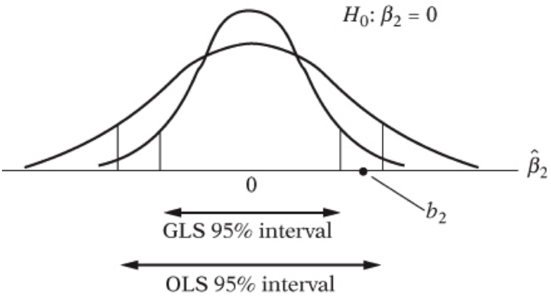
$$\begin{aligned} x_t &= (X_t - \bar{X}) \\ x_{t-1} &= (X_{t-1} - \bar{X}) \\ &\vdots \\ x_n &= (X_n - \bar{X}) \end{aligned}$$

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

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

As in the case of heteroscedasiticity 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)

$\hat{\beta}_2$ is not BLUE and even if we use $var(\hat{\beta}_2)_{AR(1)}$, the confidence intervals derived from it are likely to be wider than those based on the GLS procedure.



The situation is potentially very serious if we not only use $\hat{\beta}_2$ but also continue to use $var(\hat{\beta}_2)$, 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:

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

To see how OLS is likely to underestimate $\sigma^2$ and the variance of $\hat{\beta}_2$, let's conduct the following Monte Carlo Experiment

Suppose in the two-variable model

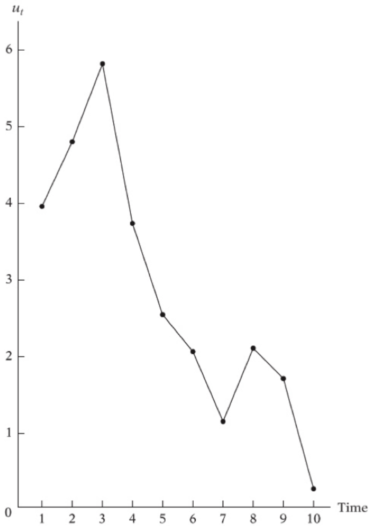
$$\begin{aligned} Y_t &= 1.0 + 0.8X_t + u_t \\ E(Y_t|X_t) &= 1.0 + 0.8X_t \end{aligned}$$

$$\begin{aligned} u_t &= 0.7u_{t-1} + \epsilon_t \\ \epsilon &\sim N(0,1) \end{aligned}$$

**TABLE 12.1**
**A Hypothetical Example of Positively Autocorrelated Error Terms**

| | $\varepsilon_t$ | $u_t = 0.7u_{t-1} + \varepsilon_t$ |
|---|---|---|
| 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:* $\varepsilon_t$ 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 and generate estimated Y

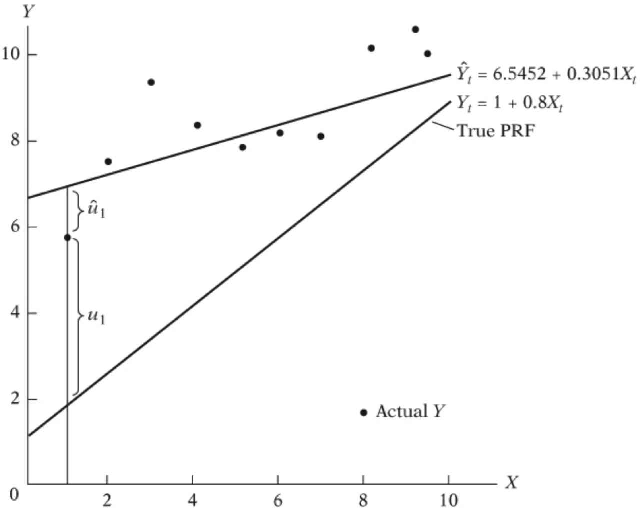
**TABLE 12.2**
**Generation of *Y* Sample Values**

| $X_t$ | $u_t$ | $Y_t = 1.0 + 0.8X_t + u_t$ |
|---|---|---|
| 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.

Then we run the regression:

$$
\begin{array}{rcccc}
\hat{Y}_t & = & 6.5452 & + & 0.3051X_t \\
se & & (0.6153) & & (0.0992) \\
t & & (10.6366) & & (3.0763) \\
R^2 & = & 0.5419 & & \\
\hat{\sigma}^2 & = & 0.814 & &
\end{array}
$$


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

Keeping the $X_t$ and $\epsilon_t$ given in Table 12.1 and 12.2, let's 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 $\varepsilon_t$ are identical. The $\varepsilon_t$ are from Table 12.1.

The regression based on Table 12.3 is as following:

$$
\begin{array}{rcccc}
\hat{Y}_t & = & 2.5345 & + & 0.6145X_t \\
se & & (0.6796) & & (0.1087) \\
t & & (3.7910) & & (5.6541) \\
R^2 & = & 0.7997 & & \\
\hat{\sigma}^2 & = & 0.9752 & &
\end{array}
$$

# Detection of Autocorrelation

- Informal method
  - graphical method
- Formal method
  - 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 statistics the number of observations is n-1 because one observation is lost in taking successive difference

- The regression model includes the intercept term
- The explanatory variables, X's, are non stochastic, or fixed in the repeated sampling
- The disturbances are generated by the AR(1):
  $u_t = \rho u_{t-1} + \epsilon_t$
  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

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

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

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

$$\hat{\rho} = \frac{\sum \hat{u}_t \hat{u}_{t-1}}{\sum \hat{u}_t^2}$$
$$\therefore d \approx 2(1 - \hat{\rho})$$

Since $-1 \leq \rho \leq 1$, thus $-4 \leq d \leq 4$

The procedure of the Durbin-Watson test are

- Run the OLS regression and obtain the residuals
- Compute d
- For the given sample size and given number of explanatory variables, find out the critical $d_L$ and $d_U$ values
- Follow the decision rules as following

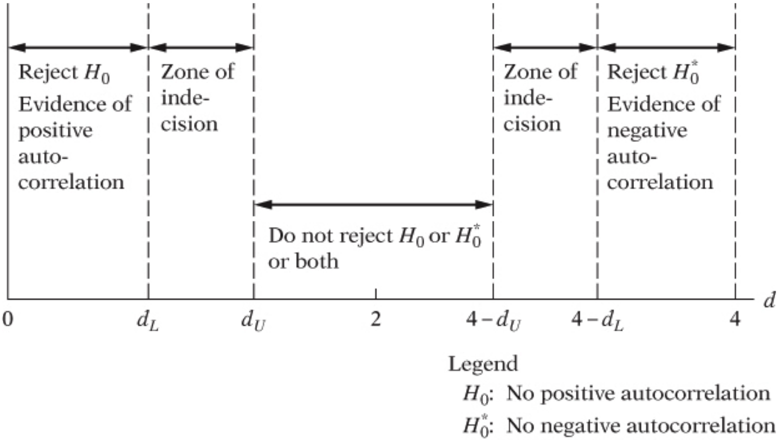



**TABLE 12.6**
**Durbin–Watson *d* Test: Decision Rules**

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

Example:
U.S. Consumption expenditure between 1947 - 2000

$$\ln Consumption_t = \beta_1 + \beta_2 \ln Income_t + \beta_3 \ln Wealth_t + \beta_4 \ln Interest_t + u_t$$

| 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 |

$$\begin{aligned} \ln Consumption_t &= -0.4677 + 0.8049 \ln Income_t + 0.2013 \ln Wealth_t \\ &\quad -0.0027 \ln Interest_t \end{aligned}$$

Hypothesis Testing

$$\begin{aligned} H_o &: \text{No positive autocorrelation} \\ H_a &: \text{Otherwise} \end{aligned}$$

$$\begin{aligned} H_o &: \text{No negative autocorrelation} \\ H_a &: \text{Otherwise} \end{aligned}$$

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

. estat dwatson

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

$H_o$ : No positive autocorrelation
$H_a$ : Otherwise

With n = 54, k = 4 at 0.05 level of significance;
$d_L = 1.452$
$d_U = 1.681$
Reject null hypothesis, evidence of positive correlation

# Remedial Measures

1. The method of generalized least squares (GLS)
2. The Newey West method - to obtain standard errors of PLS estimators that are corrected for autocorrelation

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,

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

where $1 < \epsilon < 1$

Now we consider two cases:

- $\rho$ is known
- $\rho$ is not known but has to be estimated

When $\rho$ is known

$$\begin{aligned} Y_{t-1} &= \beta_1 + \beta_2 X_{t-1} + u_{t-1} \\ \rho Y_{t-1} &= \rho\beta_1 + \rho\beta_2 X_{t-1} + \rho u_{t-1} \end{aligned}$$

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

where $\epsilon_t = u_t - \rho u_{t-1}$

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

Since the error term ($\epsilon_t$) satisfiers the usual OLS assumptions, we can apply OLS to the transformed variables $Y^*$ and $X^*$ and obtain estimators with all the optimum properties, namely, BLUE

When $\rho$ is not known

- First difference method
- based on Durbin-Watson d statistics

$$\begin{aligned} Y_t - Y_{t-1} &= \beta_2(X_t - X_{t-1}) + (u_t - u_{t-1}) \\ \Delta Y &= \beta_2 \Delta X_t + \epsilon_t \end{aligned}$$

Appropriate when d is quite low or $d < R^2$

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

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

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

where $\epsilon_t = u_t - \rho u_{t-1}$

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

```
. 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 |

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