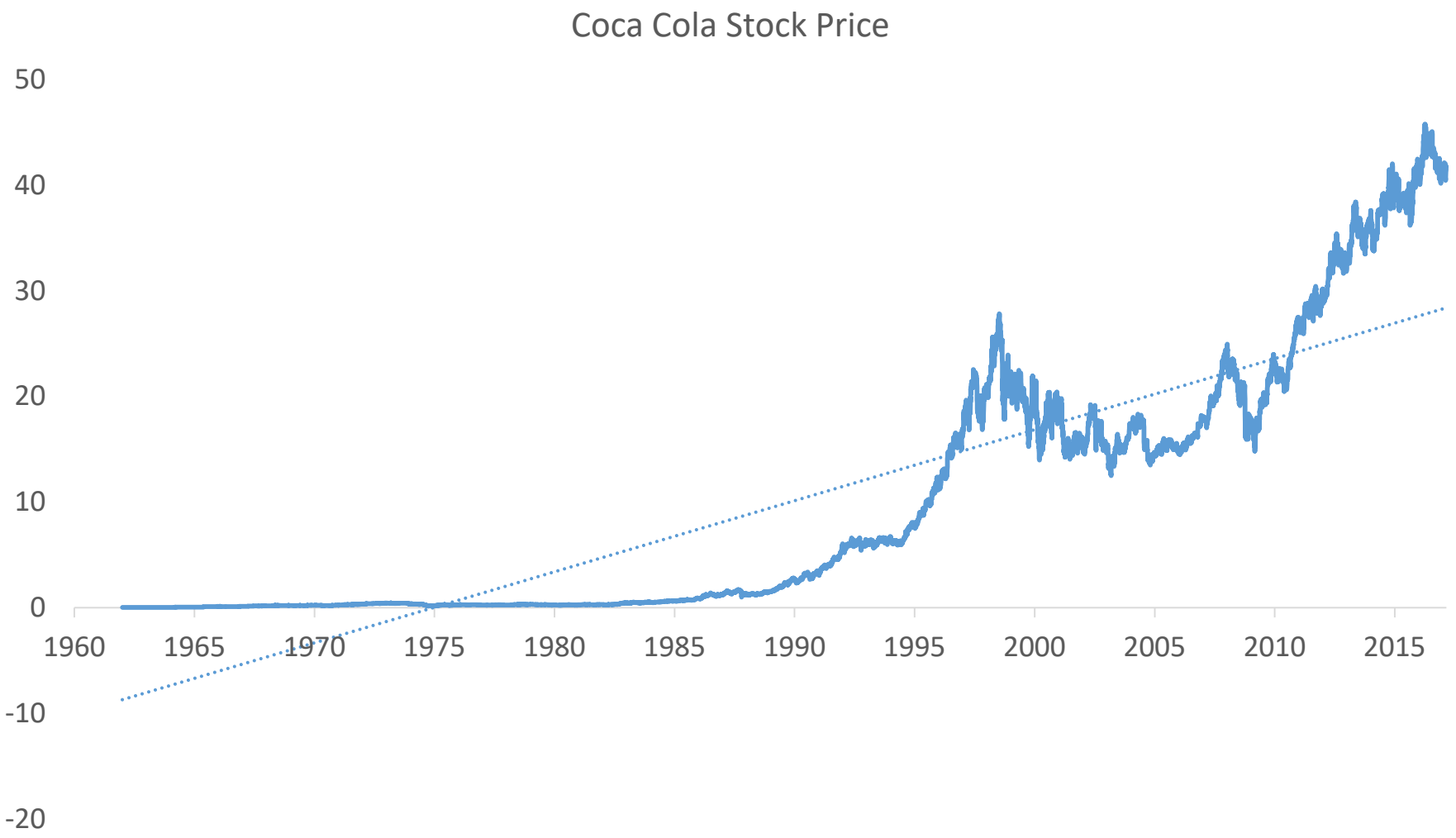
Chapter 9
Correlated Errors

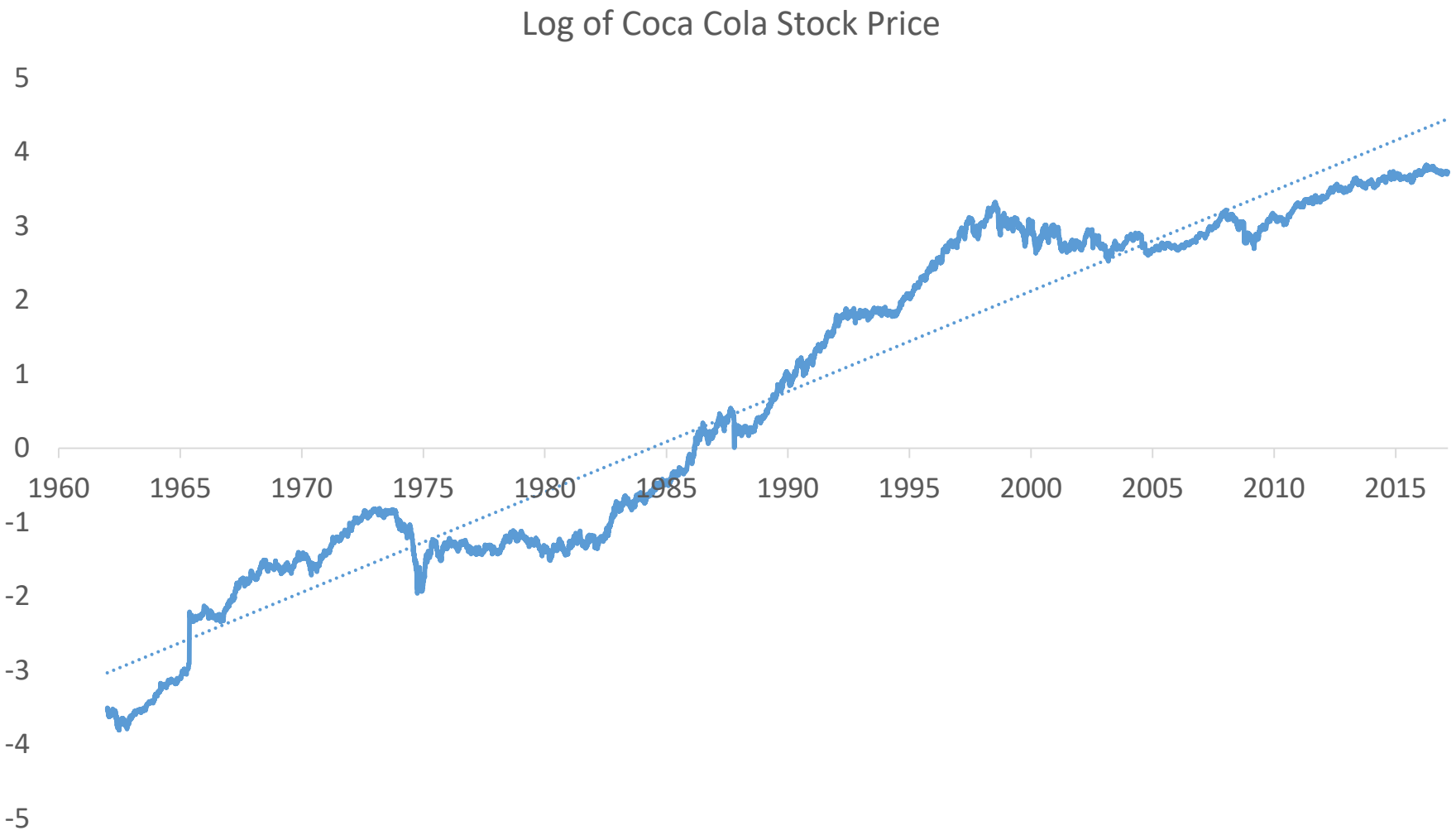
Learning Objectives

- Demonstrate the problem of correlated errors and its implications
- Conduct and interpret tests for correlated errors
- Correct for correlated errors using Newey and West's estimator (ex post) or using generalized least squares (ex ante)
- Correct for correlated errors by adding lagged variables to the model
- Show that correlated errors can arise in clustered and spatial data as well as in time-series data

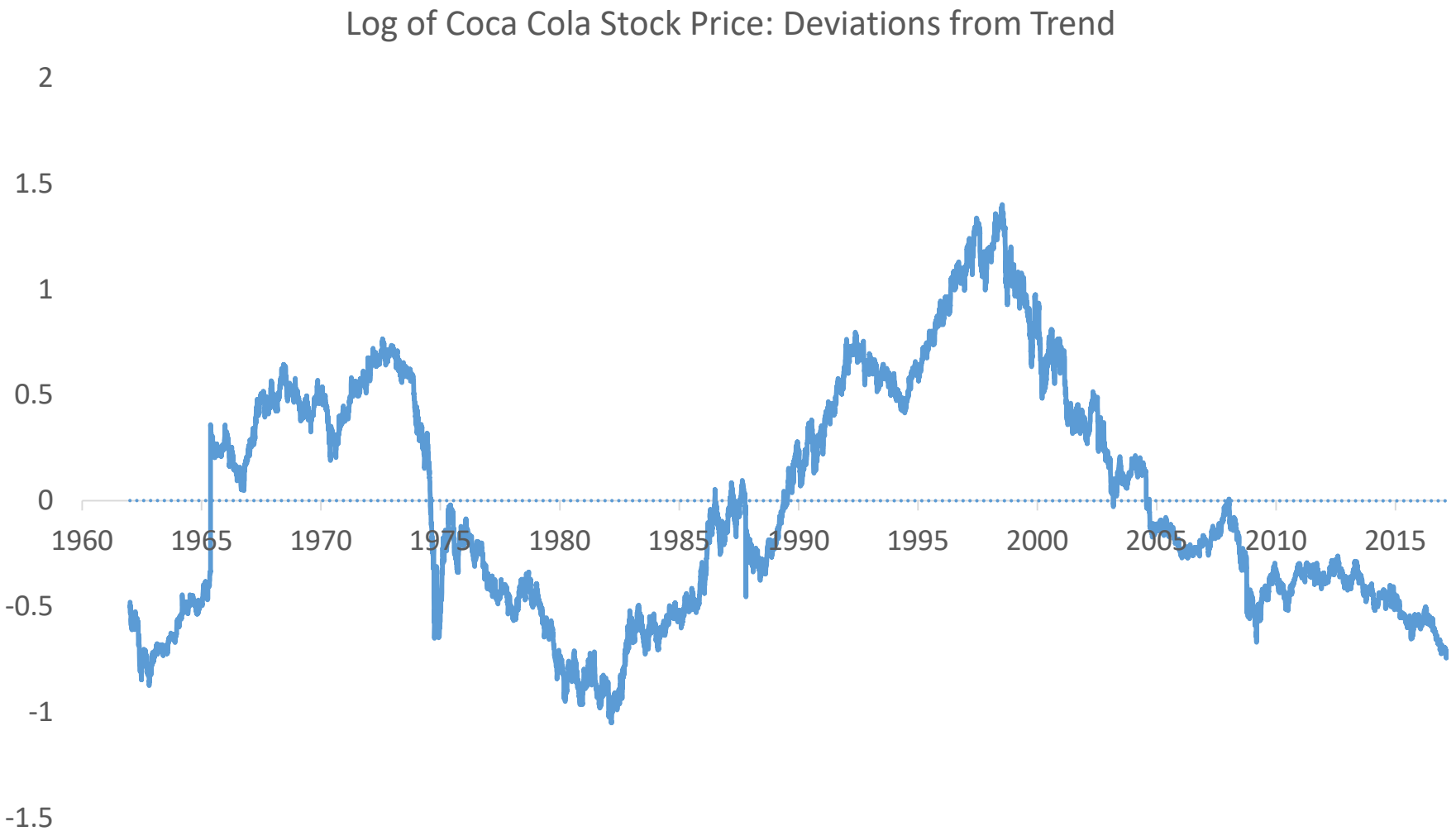
Autocorrelated Errors



Logs Make the Model Fit Better



Autocorrelated Errors



The Problem in Words

- You think you have more **information** in your data than you really do.
    - It can be the opposite – you think you have less info than you really do, but this is rare
- OLS estimates **unbiased**, but **not BLUE**
- Causes **standard errors** to be **underestimated**
- Examples
    1. Stock prices over time
    2. Consumption over time
    3. Income across space

The Problem Mathematically

$$Y_t = \beta_0 + \beta_1 X_{1t} + \beta_2 X_{2t} + \varepsilon_t$$

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

- If **ρ > 0**, then some of last period's error remains in this period's error – we have less new information each period than the standard error formula assumes.

- If **ρ < 0**, then we have more information than the s.e. formula assumes – this is rare!

Solutions

1. Test and fix after-the-fact (ex-post)

2. Change the model to eliminate the correlated errors
   i. Generalized least squares (ex-ante correction)
   ii. Change the model by adding lagged variables (**best approach**)

Testing for Autocorrelation

Autocorrelation Model

$$Y_t = \beta_0 + \beta_1 X_{1t} + \beta_2 X_{2t} + \varepsilon_t$$

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

Breusch-Godfrey test

1. Estimate regression

$$Y_t = b_0 + b_1 X_{1t} + b_2 X_{2t} + e_t$$

2. Auxiliary regression of residuals on lag

$$e_t = \rho e_{t-1} + \alpha_0 + \alpha_1 X_{1t} + \alpha_2 X_{2t} + u_t$$

Breusch-Godfrey Test

Auxiliary regression of residuals on lag

$$e_t = \rho e_{t-1} + \alpha_0 + \alpha_1 X_{1t} + \alpha_2 X_{2t} + u_t$$

Test statistic: $BG = (T-1)R^2$

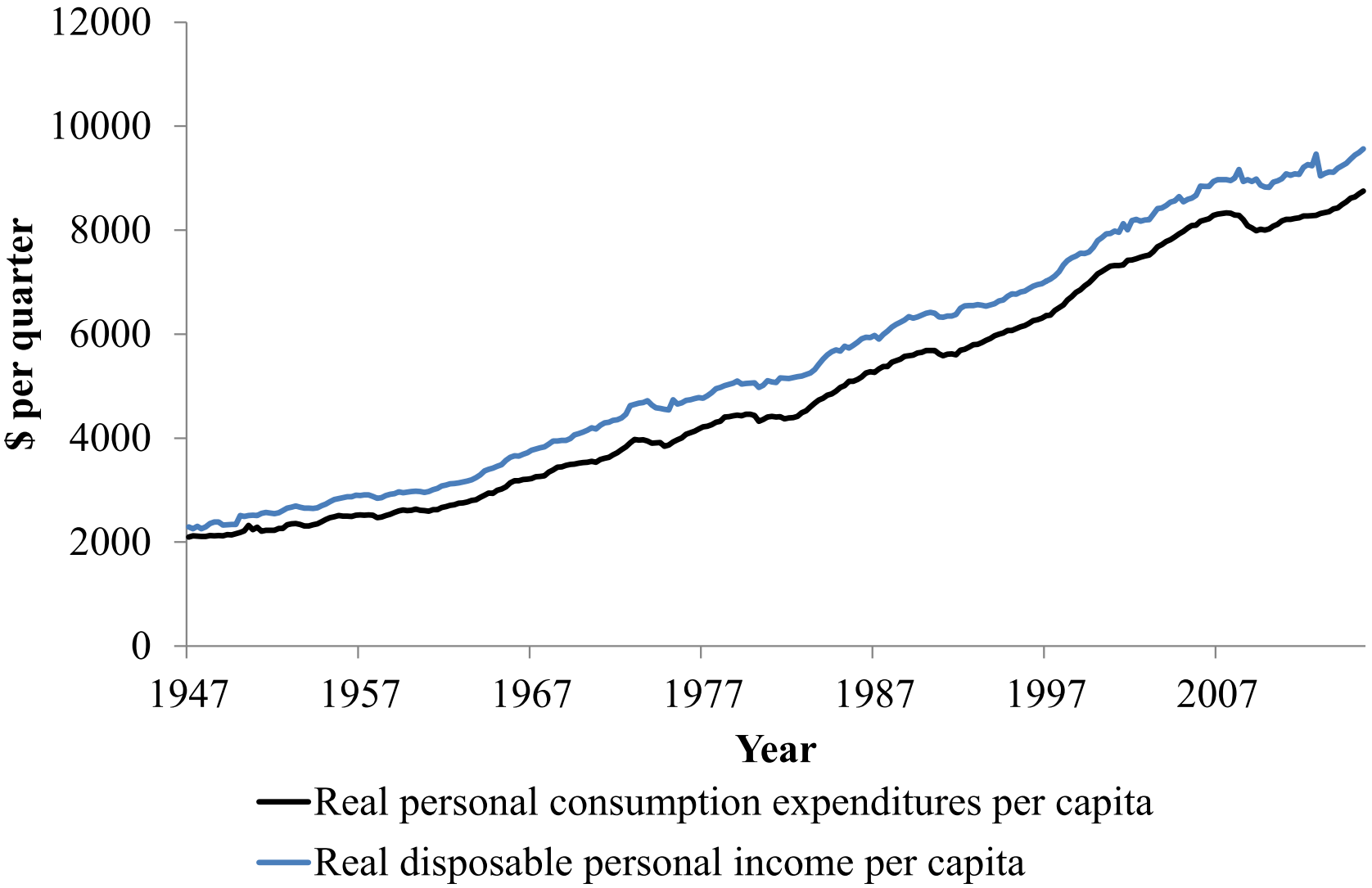
Critical value: $\chi^2_{(1)}$ (e.g., at 5% significance, c.v. = 3.84)

Can add more lags to auxiliary regression

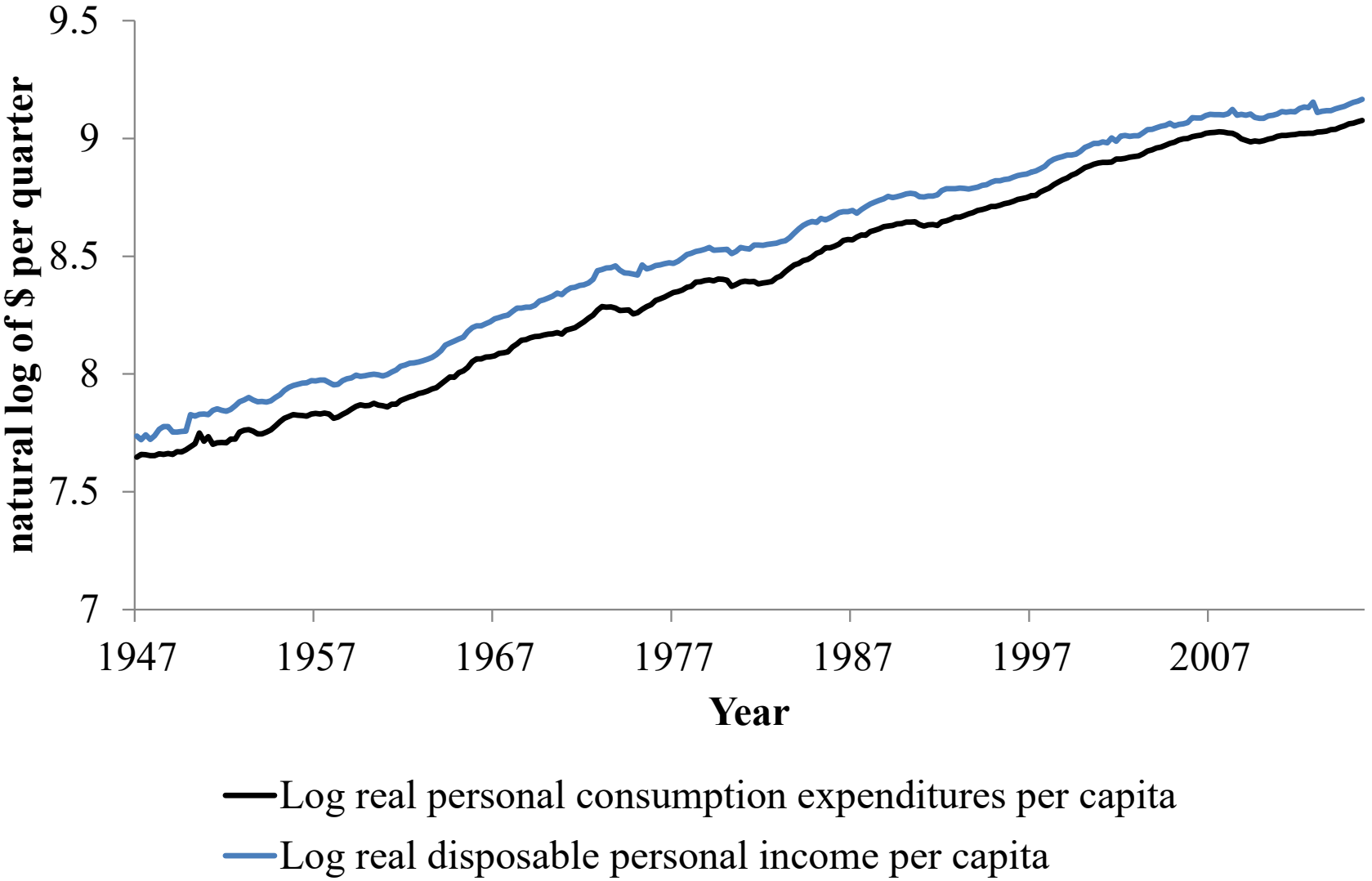
$$e_t = \rho_1 e_{t-1} + \rho_2 e_{t-2} + \ldots + \rho_m e_{t-m} + \alpha_0 + \alpha_1 X_{1t} + \alpha_2 X_{2t} + u_t$$

$BG = (T-m)R^2$ *critical value:* $\chi^2_{(m)}$

Example: US Consumption vs Income

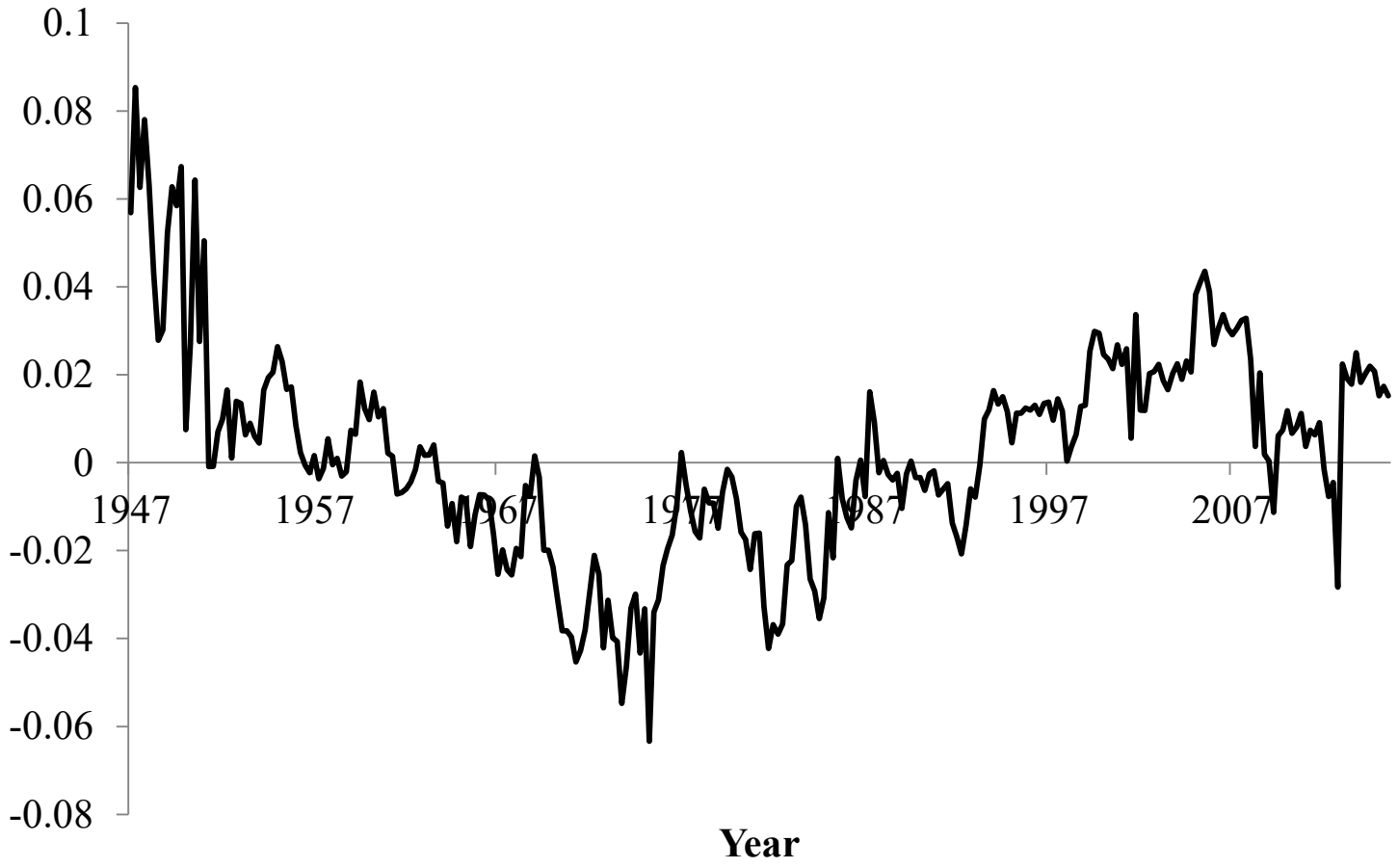


Taking Logs Straightens the Trend



Regression Residuals

$$e_t = \ln(cons_t) + 0.38 - 1.03\ln(income_t)$$



There appears to be strong autocorrelation

Breusch-Godfrey Test

Variable	Estimated Coefficient	Standard Error	t-statistic
e(t-1)	0.885	0.027	32.70
Ln(Income)	0.001	0.001	0.56
Constant	-0.007	0.012	-0.58
Sample Size	274		
R-squared	0.80		
(T-1)*R-squared	217.80		
Critical $\chi^2_{(1)}$ 1 df, 5%	3.84		

We reject the null hypothesis at the 5% level – we have **autocorrelation**

Newey-West Correction for Standard Errors

If CR2 and CR3 Hold: $$V[b_1] = \sum_{i=1}^{N} w_i^2 s^2$$

If CR2 fails (White's Method) $$V[b_1] = \sum_{i=1}^{N} w_i^2 e_i^2$$

If CR2 and CR3 fail (Newey-West)

$$V[b_1] = \sum_{t=1}^{T} w_t^2 e_t^2 \left(1 + 2 \sum_{j=1}^{L} \left(1 - \frac{j}{L} \right) r_j \right)$$

Changing the Model: GLS

Autocorrelation Model

$$Y_t = \beta_0 + \beta_1 X_t + \varepsilon_t$$

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

The error u_t satisfies CR2 and CR3

$$\begin{aligned} Y_t &= \beta_0 + \beta_1 X_t + \rho \varepsilon_{t-1} + u_t \\ &= \beta_0 + \beta_1 X_t + \rho \left(Y_{t-1} - \beta_0 - \beta_1 X_{t-1} \right) + u_t \end{aligned}$$

$$Y_t - \rho Y_{t-1} = \beta_0 (1 - \rho) + \beta_1 \left(X_t - \rho X_{t-1} \right) + u_t$$

We have a new model

$$Y_t^* = \beta_0^* + \beta_1 X_t^* + u_t$$

Changing the Model: GLS

We have a new model

$$Y_t^* = \beta_0^* + \beta_1 X_t^* + u_t$$

where $Y_t^* = Y_t - \rho Y_{t-1}$ and $X_t^* = X_t - \rho X_{t-1}$

- But we don't know ρ. Solution: **"Feasible GLS"**

1. OLS regression $Y_t = b_0 + b_1 X_t + e_t$
2. Error autocorrelation $r = \sum_{t=2}^{T} e_t e_{t-1} \Big/ \sum_{t=2}^{T} e_{t-1}^2$
3. Transform variables $Y_t^* = Y_t - rY_{t-1}, \quad X_t^* = X_t - rX_{t-1}$
4. OLS regression $Y_t^* = b_0^{GLS} + b_1^{GLS} X_t^* + e_t^{GLS}$

OLS with Newey-West vs FGLS

Variable	OLS (Newey West std. error with *L*=40)		FGLS	
	Estimated Coefficient	Standard Error	Estimated Coefficient	Standard Error
Income	1.030	0.015	1.012	0.013
Constant	-0.379	0.132	-0.026	0.012
Sample Size	275		274	
r			0.88	

Changing the Model: Distributed Lags

Autocorrelation Model

$$Y_t = \beta_0 + \beta_1 X_t + \varepsilon_t$$

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

The error u_t satisfies CR2 and CR3

$$\begin{aligned} Y_t &= \beta_0 + \beta_1 X_t + \rho \varepsilon_{t-1} + u_t \\ &= \beta_0 + \beta_1 X_t + \rho \left(Y_{t-1} - \beta_0 - \beta_1 X_{t-1} \right) + u_t \\ &= \beta_0 (1-\rho) + \rho Y_{t-1} + \beta_1 X_t - \rho \beta_1 X_{t-1} + u_t \end{aligned}$$

We have a new model

$$Y_t = \beta_0^* + \beta_1^* Y_{t-1} + \beta_2^* X_t + \beta_3^* X_{t-1} + u_t$$

Distributed Lag Model

$$Y_t = \beta_0^* + \beta_1^* Y_{t-1} + \beta_2^* X_t + \beta_3^* X_{t-1} + u_t$$

Interpreting the Model: **Two Tricks**

1. Drop the time subscripts to get "long-run"

$$Y = \beta_0^* + \beta_1^* Y + \beta_2^* X + \beta_3^* X$$

$$Y = \frac{\beta_0^*}{1 - \beta_1 *} + \frac{\beta_2^* + \beta_3^*}{1 - \beta_1 *} X$$

Distributed Lag Model

$$Y_t = \beta_0^* + \beta_1^* Y_{t-1} + \beta_2^* X_t + \beta_3^* X_{t-1} + u_t$$

Interpreting the Model: **Two Tricks**

1. Drop the time subscripts to get "long-run"

$$Y = \frac{\beta_0^*}{1-\beta_1 *} + \frac{\beta_2^* + \beta_3^*}{1-\beta_1 *} X$$

2. Derive the "error correction model"

$$Y_t - Y_{t-1} = \beta_0^* + \beta_1^* Y_{t-1} + \beta_2^* X_t + \beta_3^* X_{t-1} + u_t - Y_{t-1} + \beta_3^* X_t - \beta_3^* X_t$$

$$= -(1-\beta_1^*)\left(Y_{t-1} - \frac{\beta_0^*}{1-\beta_1^*} - \frac{\beta_2^* + \beta_3^*}{1-\beta_1^*} X_t \right) - \beta_3^* \left(X_t - X_{t-1} \right) + u_t$$

Distributed Lag Estimates

Variable	Estimated Coefficient	Standard Error	t-statistic
Consumption(t-1)	0.927	0.020	47.20
Income	0.297	0.046	6.44
Income(t-1)	-0.222	0.048	-4.66
Constant	-0.021	0.012	-1.75
Sample Size	274		
R-squared	0.9997		

Distributed Lag Model

$$Y_t = -0.021 + 0.927 Y_{t-1} + 0.297 X_t - 0.222 X_{t-1} + u_t$$

Interpreting the Model: **Two Tricks**

1. Drop the time subscripts to get "long-run"

$$Y = -0.29 + 1.03 X$$

2. Derive the "error correction model"

$$Y_t - Y_{t-1} = -0.073\left(Y_{t-1} + 0.29 - 1.03 X_t\right) + 0.222\left(X_t - X_{t-1}\right) + u_t$$

Correlated Errors Across Space

- Famous study by Brent Moulton in 1990
- Data: (i) Wages for 18,946 workers in the US
  (ii) 14 garbage variables
- Moulton regressed wages on the 14 garbage variables plus education and work experience
  - **6 of the 14 garbage variables were significant**

- **WHY?** **Spatial correlation in errors made standard errors 3-5 times too small**

“Clustering” Standard Errors

Correlation over time: Newey-West

$$V[b_1] = \sum_{t=1}^{T} w_t^2 e_t^2 \left(1 + 2 \sum_{j=1}^{L} \left(1 - \frac{j}{L} \right) r_j \right)$$

Correlation over space: Clustering

$$V[b_1] = \sum_{i=1}^{N} \sum_{j=1}^{N} 1(i, j \text{ in same cluster}) w_i e_i w_j e_j$$

What We Learned

- Correlated errors cause OLS to lose its “best” property and the estimated standard errors to be biased.
    - Same as heteroskedasticity

- As long as the autocorrelation is not too strong, the standard error bias can be corrected with Newey and West’s heteroskedasticity and autocorrelation consistent estimator.

- Getting the model right by adding lagged variables to the model is usually the best approach to deal with autocorrelation in time-series data.