Ch10. Basic Regression Analysis with Time Series Data

Ping Yu

HKU Business School
The University of Hong Kong

The Nature of Time Series Data

The Nature of Time Series Data

- A times series is a temporal ordering of observations; it may not be arbitrarily reordered.
- Typical Features: serial correlation/nonindependence of observations.
- How should we think about the randomness in time series data?
  - The outcome of economic variables (e.g., GDP, Dow Jones) is uncertain; they should therefore be modeled as random variables.
  - Time series are sequences of r.v.'s (= stochastic process/time series process).
  - Randomness does not come from sampling from a population as the cross-sectional data.
  - "Sample" = the one realized path of the time series out of the many possible paths the stochastic process could have taken. [figure here]

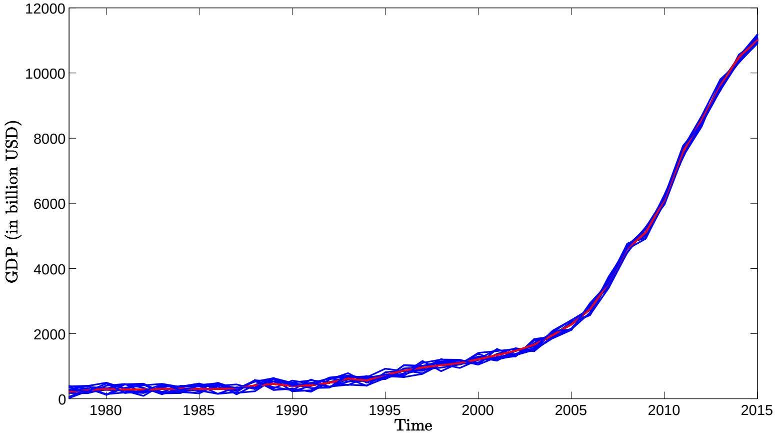


Figure: China GDP from 1978 to 2015: Red is the **Realized** GDP and Blues are **Potential** GDPs

Example: US Inflation and Unemployment Rates 1948-2003

TABLE 10.1 Partial Listing of Data on U.S. Inflation and Unemployment Rates, 1948–2003

Year	Inflation	Unemployment
1948	8.1	3.8
1949	−1.2	5.9
1950	1.3	5.3
1951	7.9	3.3
.	.	.
.	.	.
.	.	.
1998	1.6	4.5
1999	2.2	4.2
2000	3.4	4.0
2001	2.8	4.7
2002	1.6	5.8
2003	2.3	6.0

- Here, there are only two time series. There may be many more variables whose paths over time are observed simultaneously.
- Time series analysis focuses on modeling the dependency of a variable on its own past, and on the present and past values of other variables.

Autocorrelation

- The correlation of a series with its own lagged values is called autocorrelation or serial correlation.
- The first autocorrelation of y_t is $Corr\left(y_t, y_{t-1}\right)$ and the first autocovariance of y_t is $Cov\left(y_t, y_{t-1}\right)$. Thus

$$Corr\left(y_t, y_{t-1}\right) = \frac{Cov\left(y_t, y_{t-1}\right)}{\sqrt{Var\left(y_t\right) Var\left(y_{t-1}\right)}} = \rho_1.$$

- Similarly, the jth autocorrelation is

$$Corr\left(y_t, y_{t-j}\right) = \frac{Cov\left(y_t, y_{t-j}\right)}{\sqrt{Var\left(y_t\right) Var\left(y_{t-j}\right)}} = \rho_j.$$

- These are population correlations - they describe the population joint distribution of $\left(y_t, y_{t-1}\right)$ and $\left(y_t, y_{t-j}\right)$.
- The jth sample autocorrelation is an estimate of the jth population autocorrelation, i.e., the sample analog of $Corr\left(y_t, y_{t-j}\right)$.

Sample Autocorrelation

- Recall that for a sample $\{(x_i, y_i) : i = 1, \cdots, n\}$,

$$\widehat{Cov}(x, y) = \frac{1}{n} \sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y}).$$

- For a time series $\{y_t : t = 1, \cdots, T\}$, $\widehat{Cov}(y_t, y_{t-j})$ is supposedly

$$\widehat{Cov}(y_t, y_{t-j}) = \frac{1}{T} \sum_{t=1}^{T} (y_t - \overline{y_t})(y_{t-j} - \overline{y_{t-j}}).$$

- However,

$$\underbrace{y_1, \cdots, y_{j-1}, y_j, y_{j+1}, \cdots, y_{T-j-1}, y_{T-j}}_{y_{t-j}}, y_{T-j+1}, \cdots, y_T$$

and

$$y_1, \cdots, y_{j-1}, y_j, \overbrace{y_{j+1}, \cdots, y_{T-j-1}, y_{T-j}, y_{T-j+1}, \cdots, y_T}^{y_t},$$

the summation should be from $j+1$ to T; otherwise, y_{t-j} for $t \leq j$ is not defined.

continue

- Correspondingly,

$$\overline{y_t} = \overline{y}_{j+1,T} \equiv \frac{1}{T-j}\sum_{t=j+1}^{T} y_t,$$
$$\overline{y_{t-j}} = \overline{y}_{1,T-j} \equiv \frac{1}{T-j}\sum_{t=1}^{T-j} y_t.$$

- As a result, the jth sample autocorrelation is

$$\widehat{\rho}_j = \frac{\widehat{Cov}\left(y_t, y_{t-j}\right)}{\widehat{Var}\left(y_t\right)},$$

where

$$\widehat{Cov}\left(y_t, y_{t-j}\right) = \frac{1}{T-j}\sum_{t=j+1}^{T}\left(y_t - \overline{y}_{j+1,T}\right)\left(y_{t-j} - \overline{y}_{1,T-j}\right),$$
$$\widehat{Var}\left(y_t\right) = \frac{1}{T}\sum_{t=1}^{T}\left(y_t - \overline{y}_{1,T}\right)^2.$$

- Note: Here, we assume constant variance, i.e., $Var\left(y_t\right)$ does not depend on t.

Example: US CPI Inflation Rate

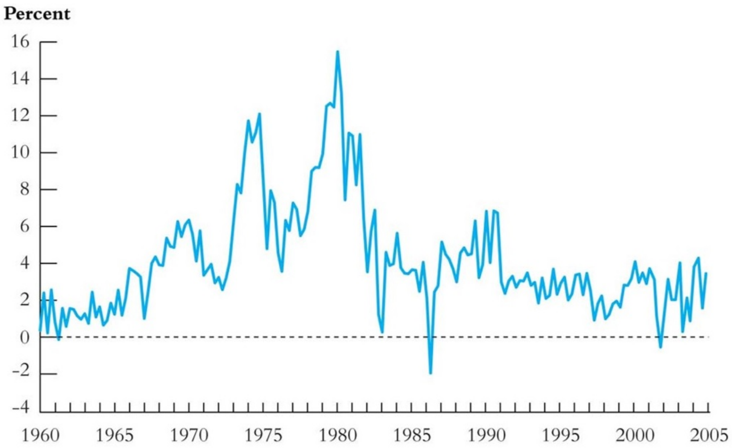


Figure: US Quarterly CPI Inflation Rate: 1960:I-2004:IV

Sample Autocorrelations of Inf_t and ΔInf_t

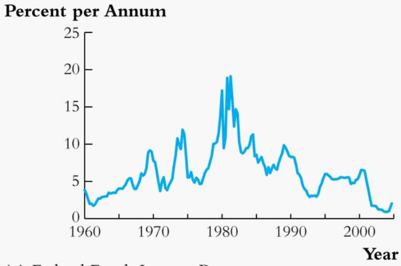
- The sample autocorrelations of the quarterly rate of U.S. inflation (Inf_t) and the quarter-to-quarter change in the quarterly rate of inflation $(\Delta Inf_t \equiv Inf_t - Inf_{t-1})$ are summarized in the following table:

Lag	Inf_t	ΔInf_t
1	0.84	−0.26
2	0.76	−0.25
3	0.76	0.29
4	0.67	−0.06

Table: First Four Sample Autocorrelations of the US Inflation Rate and Its Change

- The inflation rate is highly serially correlated ($\widehat{\rho}_1^{Inf} = 0.84$).
- Last quarter's inflation rate contains much information about this quarter's inflation rate.
- What's the intuitive meaning of $\widehat{\rho}_1^{\Delta Inf} = -0.26 < 0$?
  - the plot is also dominated by multiyear swings.

Other Economic Time Series



(a) Federal Funds Interest Rate

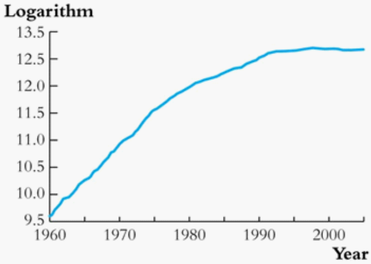
(b) U.S. Dollar/British Pound Exchange Rate

History of Cointegration

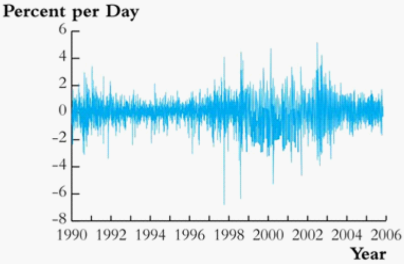
- Engle, R.F. and C.W.J. Granger, 1987, Co-integration and Error Correction: Representation, Estimation and Testing, *Econometrica*, 55, 251–276.

Clive Granger (1934-2009),
UCSD, 2003NP, 1959NottinghamPhD

continue



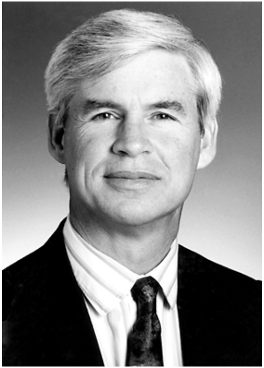
(c) Logarithm of GDP in Japan



(d) Percentage Changes in Daily Values of the NYSE Composite Stock Index

History of AutoregRessive Conditional Heteroskedasticity (ARCH)

- Engle, R.F., 1982, Autoregressive Conditional Heteroscedasticity with Estimates of Variance of United Kingdom Inflation, *Econometrica*, 50, 987-1008.



Robert Engle (1942-),
NYU, 2003NP, 1969CornellPhD

Examples of Time Series Regression Models

a: Static Models

- A static model relating y to z is

$$y_t = \beta_0 + \beta_1 z_t + u_t.$$

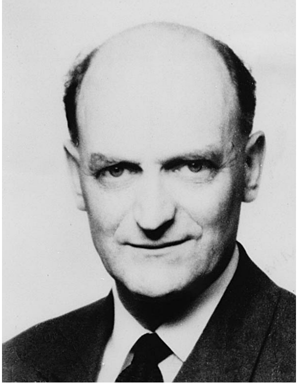
- In static time series models, the current value of one variable is modeled as the result of the current values of explanatory variables.
- Example (Phillips Curve [photo here]): There is a contemporaneous relationship between unemployment and inflation [figure here],

$$inf_t = \beta_0 + \beta_1 unem_t + u_t.$$

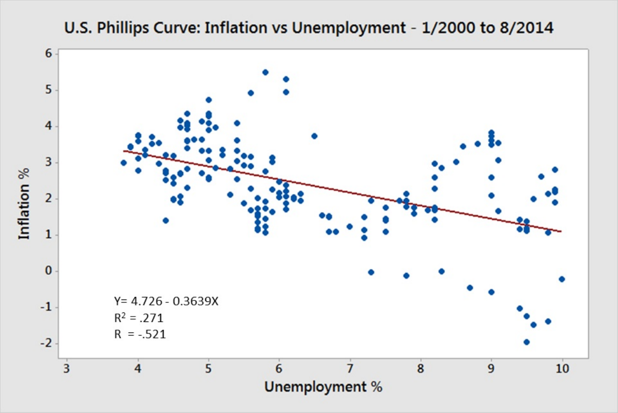
- Example: The current murder rate is determined by the current conviction rate, unemployment rate, and fraction of young males in the population,

$$mrdrte_t = \beta_0 + \beta_1 convrte_t + \beta_2 unem_t + \beta_3 yngmle_t + u_t.$$

History of The Phillips Curve



A.W. Phillips (1914-1975),
New Zealander, LSE, 1949LSESociologyPhD



Source Data: FRED Database
Inflation: CPI for All Urban Consumers

b: Finite Distributed Lag Models

- In finite distributed lag (FDL) models, the explanatory variables are allowed to influence the dependent variable with a time lag.
- Mathematically,

$$y_t = \alpha_0 + \delta_0 z_t + \delta_1 z_{t-1} + \cdots + \delta_q z_{t-q} + u_t$$

  is an FDL of order q, where q is finite.
- Example: The fertility rate may depend on the tax value of a child, but for biological and behavioral reasons, the effect may have a lag,

$$gft_t = \alpha_0 + \delta_0 pe_t + \delta_1 pe_{t-1} + \delta_2 pe_{t-2} + u_t,$$

  where
  gft_t = general fertility rate (children born per 1,000 women in year t)
  pe_t = tax exemption in year t

Effect of a Transitory Shock

- If there is a one time shock in a past period, the dependent variable will change temporarily by the amount indicated by the coefficient of the corresponding lag.
- Consider an FDL of order 2,

$$y_t = \alpha_0 + \delta_0 z_t + \delta_1 z_{t-1} + \delta_2 z_{t-2} + u_t.$$

- At time t, z increases by one unit from c to $c+1$ and then reverts to its previous level at time $t+1$:

$$\cdots, z_{t-2} = c, z_{t-1} = c, z_t = c+1, z_{t+1} = c, z_{t+2} = c, \cdots$$

- Then, by setting the errors to be zero,

$$\begin{aligned} y_{t-1} &= \alpha_0 + \delta_0 c + \delta_1 c + \delta_2 c, \\ y_t &= \alpha_0 + \delta_0 (c+1) + \delta_1 c + \delta_2 c, \\ y_{t+1} &= \alpha_0 + \delta_0 c + \delta_1 (c+1) + \delta_2 c, \\ y_{t+2} &= \alpha_0 + \delta_0 c + \delta_1 c + \delta_2 (c+1), \\ y_{t+3} &= \alpha_0 + \delta_0 c + \delta_1 c + \delta_2 c. \end{aligned}$$

continue

- So
$$y_t - y_{t-1} = \delta_0,$$
i.e., δ_0 is the immediate change in y due to the one-unit increase in z at time t, and is usually called impact propensity or impact multiplier.
- Similarly,
$$\delta_1 = y_{t+1} - y_{t-1}$$
is the change in y one period after the temporary change, and
$$\delta_2 = y_{t+2} - y_{t-1}$$
is the change in y two periods after the temporary change.
- At time $t+3$, y has reverted back to its initial level: $y_{t+3} = y_{t-1}$ because only two lags of z appears in the FDL model.
- In summary,
$$\delta_j = \frac{\partial y_{t+j}}{\partial z_t}.$$
- Lag Distribution: δ_j as a function of j, which summarizes the dynamic effect that a temporary increase in z has on y. [figure here]

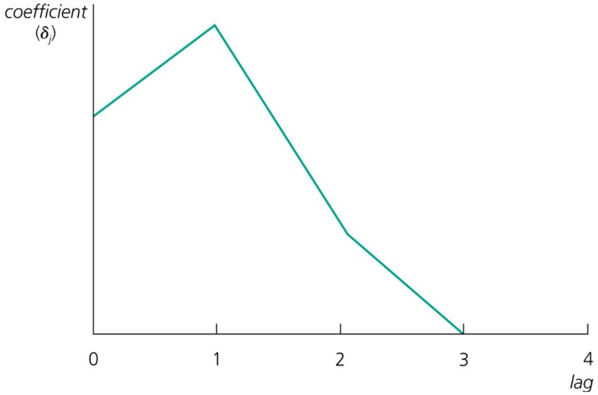


Figure: A Lag Distribution with Two Nonzero Lags

- The effect is biggest after a lag of one period. After that, the effect vanishes (if the initial shock was transitory).

Effect of Permanent Shock

- If there is a permanent shock in a past period, i.e., the explanatory variable permanently increases by one unit, the effect on the dependent variable will be the cumulated effect of all relevant lags. This is a long-run effect on the dependent variable.
- At time t, z permanently increases by one unit from c to $c+1$:

$$z_s = c, s < t \text{ and } z_s = c+1, s \geq t.$$

- Then,

$$\begin{aligned}
y_{t-1} &= \alpha_0 + \delta_0 c + \delta_1 c + \delta_2 c, \\
y_t &= \alpha_0 + \delta_0 (c+1) + \delta_1 c + \delta_2 c, \\
y_{t+1} &= \alpha_0 + \delta_0 (c+1) + \delta_1 (c+1) + \delta_2 c, \\
y_{t+2} &= \alpha_0 + \delta_0 (c+1) + \delta_1 (c+1) + \delta_2 (c+1), \\
y_{t+3} &= y_{t+2}.
\end{aligned}$$

continue

- With a permanent increase in z, after one period, y will increase by $\delta_0 + \delta_1$, and after two periods, y will increase by $\delta_0 + \delta_1 + \delta_2$ and then stay there.
- The sum of the coefficients on current and lagged z, $\delta_0 + \delta_1 + \delta_2$, is the long-run change in y given a permanent increase in z, and is called the long-run propensity (LRP) or long-run multiplier.
- In summary, in an FDL of order q,

$$LRP = \delta_0 + \delta_1 + \cdots + \delta_q = \frac{\partial y_{t+q}}{\partial z_t} + \cdots + \frac{\partial y_{t+q}}{\partial z_{t+q}}.$$

- In the figure, the long run effect of a permanent shock is the cumulated effect of all relevant lagged effects. It does not vanish (if the initial shock is a permanent one).

(**) Finite Sample Properties of OLS under Classical Assumptions

Standard Assumptions for the Time Series Model

- Assumption TS.1 (Linear in Parameters):

$$y_t = \beta_0 + \beta_1 x_{t1} + \cdots + \beta_k x_{tk} + u_t.$$

  - The time series involved obey a linear relationship.
  - The stochastic processes $y_t, x_{t1}, \ldots, x_{tk}$ are observed, the error process u_t is unobserved.
  - The definition of the explanatory variables is general, e.g., they may be lags or functions of other explanatory variables.

- Assumption TS.2 (No Perfect Collinearity): In the sample (and therefore in the underlying time series process), no independent variable is constant nor a perfect linear combination of the others.

continue

- Assumption TS.3 (Zero Conditional Mean):

$$E\left[u_t|\mathbf{X}\right] = 0,$$

where

$$\mathbf{X} = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1k} \\ \vdots & \vdots & \vdots & \vdots \\ x_{t1} & x_{t2} & \cdots & x_{tk} \\ \vdots & \vdots & \vdots & \vdots \\ x_{T1} & x_{T2} & \cdots & x_{Tk} \end{pmatrix}$$

collects all the information on the complete time paths of all explanatory variables.

- Assumption TS.3 assumes that the mean value of the unobserved factors is unrelated to the values of the explanatory variables in all periods:

$$Cov\left(u_t, x_{sj}\right) = 0 \text{ for all } j \text{ and all } t \text{ and } s.$$

- Assumption TS.3 is the strict exogeneity assumption for the regressors,

Discussion of Assumption TS.3

- Contemporaneous Exogeneity:

$$E[u_t|\mathbf{x}_t] = 0,$$

  where $\mathbf{x}_t = (x_{t1}, \ldots, x_{tk})$ is the values of all explanatory variables in period t.
  - It implies $Cov(u_t, x_{tj}) = 0$ for all j.
- So strict exogeneity is stronger than contemporaneous exogeneity. In the cross-sectional case, they are equivalent due to random sampling.
- TS.3 rules out feedback from the dependent variable on future values of the explanatory variables; this is often questionable especially if explanatory variables "adjust" to past changes in the dependent variable, e.g., $x_t = y_{t-1}$, then $Cov(u_{t-1}, x_t) = Cov(u_{t-1}, y_{t-1}) \neq 0$.
- Example: In a simple static model to explain a city's murder rate in terms of police officers per capita,

$$mrdrte_t = \beta_0 + \beta_1 polpc_t + u_t,$$

  if $Cov(polpc_{t+1}, u_t) \neq 0$ then TS.3 fails even if $Cov(u_t, polpc_{t-j}) = 0$ for $j = 0, 1, 2, \cdots$
- Example: In an agricultural production function, the rainfall is strictly exogenous, while the labor input is not.
- If the error term is related to past values of the explanatory variables, one should include these values as contemporaneous regressors, i.e., use the FDL model.

a: Unbiasedness of OLS

- Theorem 10.1: Under assumptions TS.1-TS.3,

$$E\left[\widehat{\beta}_j \middle| \mathbf{X}\right] = \beta_j, j = 0, 1, \cdots, k,$$

  for any values of β_j.
- The proof is similar to that of Theorem 3.1.
- The analysis of omitted variables bias is also similar.
- As in Chapter 3, we condition on the regressors **X**. Conditional unbiasedness implies unconditional unbiasedness:

$$E\left[\widehat{\beta}_j\right] = E\left[E\left[\widehat{\beta}_j \middle| \mathbf{X}\right]\right] = E\left[\beta_j\right] = \beta_j.$$

- (*) Law of Iterated Expectations (LIE): For two r.v.'s, $E[E[Y|X]] = E[Y]$.
  - Intuition: $E[Y|X] = \mu(X)$ averages over Y for each group of individuals with a specific X value, and then $E[E[Y|X]] = E[\mu(X)]$ averages over X by the probability of each X group, which results in the unconditional mean of Y. [figure here]

  - $E[u|x] = 0$ implies the unconditional mean of u, $E[u] \overset{LIE}{=} E[E[u|x]] = E[0] = 0$. We can also show, by applying the LIE, that $Var(u|x) = \sigma^2$ implies $Var(u) = \sigma^2$, and $E[u|x] = 0$ implies $Cov(x, u) = 0$.

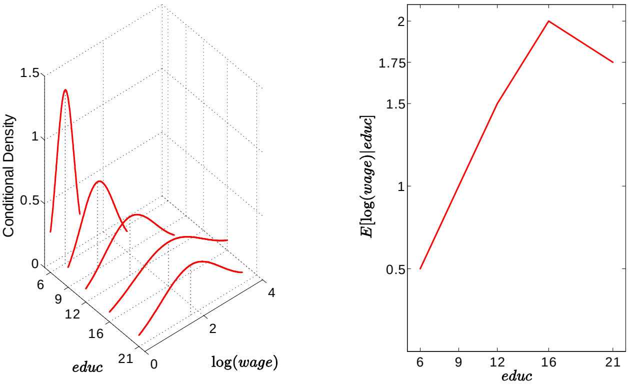


Figure: Illustration of the LIE: the distribution of log$(wage)$ given *educ* mimics the real data in Chapter 8

- $E\left[\log(wage)\right] = E\left[E\left[\log(wage|educ)\right]\right] = E\left[m(educ)\right] = \sum_{j=1}^{J} m(educ_j)p_j.$

Standard Assumptions for the Time Series Model (continue)

- Assumption TS.4 (Homoskedasticity): $Var\left(u_t|\mathbf{X}\right) = Var\left(u_t\right) = \sigma^2$.
  - The volatility of the errors must not be related to the explanatory variables in any of the periods.
- A sufficient condition is that u_t is independent of **X** and that $Var\left(u_t\right)$ is constant over time.
- In the time series context, homoskedasticity may also be easily violated, e.g., if the volatility of the dependent variable depends on regime changes (recall the Chow-test, see also the example below).
- Assumption TS.5 (No Serial Correlation): $Cov\left(u_t, u_s|\mathbf{X}\right) = 0$, $t \neq s$.
  - Conditional on the explanatory variables, the unobserved factors must not be correlated over time.
- This assumption is specific to time series. In the cross-sectional case, it automatically holds due to random sampling.
- The assumption may also serve as substitute for the random sampling assumption if sampling a cross-section is not done completely randomly.
- In this case, given the values of the explanatory variables, errors have to be uncorrelated across cross-sectional units (e.g., states).

More on Assumption TS.5

- This assumption may easily be violated if, conditional on knowing the values of the independent variables, omitted factors are correlated over time.
- Strictly, suppose the true model is

$$y_t = \beta_0 + \beta_1 x_t + \beta_2 z_t + u_t,$$

while we fit

$$y_t = \beta_0 + \beta_1 x_t + v_t.$$

- If $Cov(z_t, z_s) \neq 0$, $t \neq s$, then

$$Cov(v_t, v_s) = Cov(\beta_2 z_t + u_t, \beta_2 z_s + u_s) = \beta_2^2 Cov(z_t, z_s)$$

even if

$$Cov(z_t, u_s) = 0 \text{ for all } t, s \text{ and } Cov(u_t, u_s) = 0 \text{ for } t \neq s.$$

b: Three Parallel Theorems

- Theorem 10.2 (OLS Sampling Variances): Under assumptions TS.1-TS.5,

$$Var\left(\widehat{\beta}_j \middle| \mathbf{X}\right) = \frac{\sigma^2}{SST_j\left(1-R_j^2\right)}, j = 1, \cdots, k.$$

- Theorem 10.3 (Unbiased Estimation of σ^2): Under assumptions TS.1-TS.5,

$$E\left[\widehat{\sigma}^2\right] = \sigma^2,$$

  where $\widehat{\sigma}^2 = \frac{SSR}{T-k-1}$.
- Theorem 10.4 (The Gauss-Markov Theorem): Under assumptions TS.1-TS.5, the OLS estimators are BLUEs conditional (or unconditional) on **X**.
- Assumptions TS.1-TS.5 are the appropriate Gauss-Markov assumptions for time series applications.

c: Inference under the Classical Linear Model Assumptions

- Assumption TS.6 (Normality): u_t is independent of $\mathbf{X}$, and $u_t \overset{iid}{\sim} N\left(0, \sigma^2\right)$.
  - This assumption implies TS.3-TS.5.
- Theorem 10.5 (Normal Sampling Distributions): Under assumptions TS.1-TS.6, the OLS estimators have the usual normal distribution (conditional on $\mathbf{X}$). The usual F- and t-tests are valid. The usual construction of CIs is also valid.
- Example (Static Phillips Curve): The fitted regression line is

$$\begin{aligned} \widehat{inf}_t &= 1.42 + .468 unem_t \\ & \quad (1.72)(.289) \\ n &= 49, R^2 = .053, \overline{R}^2 = .033 \end{aligned}$$

- Contrary to theory, the estimated Phillips Curve does not suggest a tradeoff between inflation and unemployment.

Discussion of CLM Assumptions

- TS.1: $inf_t = \beta_0 + \beta_1 unem_t + u_t$. The error term contains factors such as monetary shocks, income/demand shocks, oil price shocks, supply shocks, or exchange rate shocks.
- TS.2: A linear relationship might be restrictive, but it should be a good approximation. Perfect collinearity is not a problem as long as unemployment varies over time.
- TS.3: $E\left[u_t | unem_1, \cdots, unem_T\right] = 0$ is easily violated. $unem_{t-1} \uparrow \Longrightarrow u_t \downarrow$: past unemployment shocks may lead to future demand shocks which may dampen inflation. $u_{t-1} \uparrow \Longrightarrow unem_t \uparrow$: an oil price shock means more inflation and may lead to future increases in unemployment.
- TS.4: $Var\left(u_t | unem_1, \cdots, unem_T\right) = \sigma^2$ is violated if monetary policy is more "nervous" in times of high unemployment.
- TS.5: $Corr\left(u_t, u_s | unem_1, \cdots, unem_T\right) = 0$ is violated if exchange rate influences persist over time (they cannot be explained by unemployment).
- TS.6: If TS.3-5 is questionable, TS.6 is also questionable. Also, normality is questionable.

Functional Form and Dummy Variables

Functional Form and Dummy Variables

- Logarithmic transformation has the usual elasticity interpretation. For example, in the FDL model, we can define short-run elasticity (corresponding to impact propensity) and long-run elasticity (corresponding to LRP) by using logarithmic functional forms.
- Dummy variables are often used to isolate certain periods that may be systematically different from other periods.
- Example (Effects of Personal Exemption on Fertility Rates): The fitted regression line is

$$\begin{aligned}\widehat{gfr}_t &= \underset{(3.68)}{98.68} + \underset{(.030)}{.083} pe_t - \underset{(7.46)}{24.24} ww2_t - \underset{(4.08)}{31.59} pill_t \\ n &= 72, R^2 = .473, \overline{R}^2 = .450\end{aligned}$$

  where
  $ww2$ = dummy for World War II years (1941-45)
  $pill$ = dummy for availability of contraceptive pill (1963-present)
- During World War II, the fertility rate was temporarily (much) lower (gfr ranges from 65 to 127 during 1913-1984).
- It has been permanently lower since the introduction of the pill in 1963.
- The effect of tax exemption is significant both statistically and economically ($12 tax exemption$\Longrightarrow$one more baby per 1,000 women).

Trends and Seasonality

a: Characterizing Trending Time Series

- Modelling a Linear Time Trend:

$$y_t = \alpha_0 + \alpha_1 t + e_t.$$

- $E[y_t] = \alpha_0 + \alpha_1 t$, the expected value of the dependent variable is a linear function of time. [figure here]
- This is equivalent to say that

$$E[\Delta y_t] = E[y_t - y_{t-1}] = \alpha_0 + \alpha_1 t - (\alpha_0 + \alpha_1 (t-1)) = \alpha_1.$$

- Modelling an Exponential Time Trend:

$$\log(y_t) = \alpha_0 + \alpha_1 t + e_t.$$

- $\frac{\partial \log(y_t)}{\partial t} = \frac{\partial y_t / y_t}{\partial t} = \alpha_1$, so the growth rate is constant over time. [figure here]
- This is equivalent to say that

$$E[\Delta \log(y_t)] = E[\log(y_t) - \log(y_{t-1})] = \alpha_0 + \alpha_1 t - (\alpha_0 + \alpha_1 (t-1)) = \alpha_1.$$

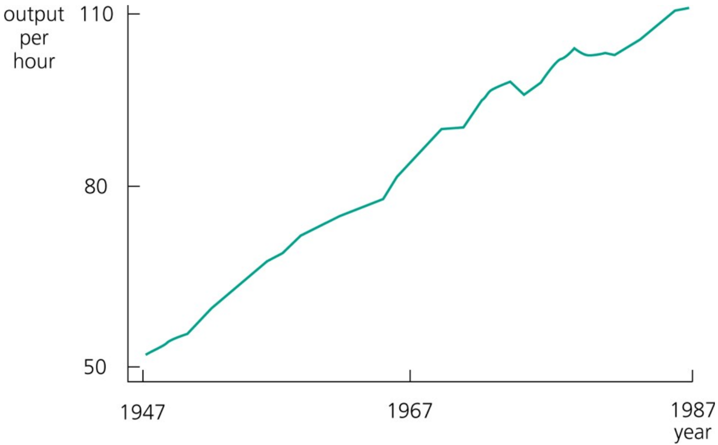


Figure: Example for a Time Series with a Linear Upward Trend

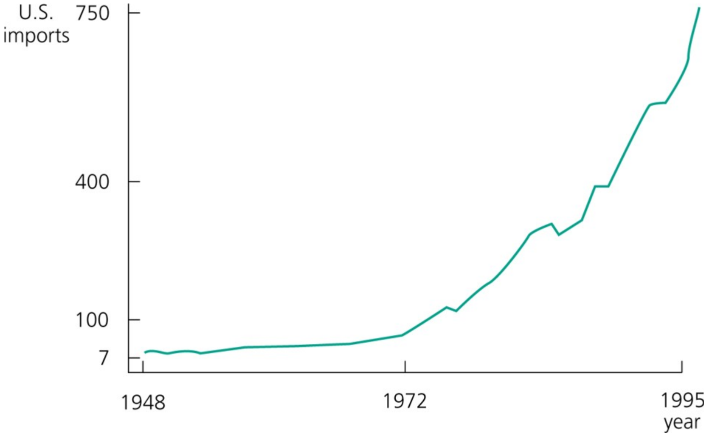


Figure: Example for a Time Series with an Exponential Trend

b: Using Trending Variables in Regression Analysis

- If trending variables are regressed on each other, a spurious relationship may arise if the variables are driven by a common trend. This is an example of spurious regression problem.
- In this case, it is important to include a trend in the regression:

$$y_t = \beta_0 + \beta_1 x_t + \beta_3 t + u_t.$$

- If x_t also includes a trend, then x_t is correlated with t. The regression

$$y_t = \beta_0 + \beta_1 x_t + u_t \quad (1)$$

  would have an omitted variable bias. This is why the spurious regression problem appears.
- If x_t has a trend but y_t does not, then $\widehat{\beta}_1$ in (1) tends to be insignificant. This is because the trending in x_t might be too dominating and obscure any partial effect it might have on y_t.
- Either y_t or some of the regressors has a time trend, add t in.

Example: Housing Investment and Prices (spurious regression)

- The fitted regression line is

$$\begin{aligned}\widehat{\log(invpc)} &= -.550 + 1.241 \log(price) \\ &\quad\ (.043) \quad\ (.382) \\ n &= 42, R^2 = .208, \overline{R}^2 = .189\end{aligned}$$

  where
  invpc = per capita housing investment (in thousands of dollars)
  price = housing price index (equal to 1 in 1982)

- It looks as if investment and prices are positively related.
- If adding the time trend in,

$$\begin{aligned}\widehat{\log(invpc)} &= -.913 - .381 \log(price) + .0098t \\ &\quad\ (.136) \quad (.679) \qquad\qquad\quad (.0035) \\ n &= 42, R^2 = .341\,(> .208)\,, \overline{R}^2 = .307\end{aligned}$$

- There is no significant relationship between price and investment anymore.

c: A Detrending Interpretation of Regressions with a Time Trend

- It turns out that the OLS coefficients in a regression including a trend are the same as the coefficients in a regression without a trend but where all the variables have been **detrend**ed before the regression.
- This follows from the FWL theorem.
- Specifically, suppose $y \sim 1, x, t$, and we are interested in $\widehat{\beta}_1$.
  1. $x_t \sim 1, t \Longrightarrow \ddot{x}_t$
  2. $y_t \sim 1, t \Longrightarrow \ddot{y}_t$
  3. $\ddot{y}_t \sim \ddot{x}_t \Longrightarrow \widehat{\beta}_1$

d: Computing R-Squared when the Dependent Variable Is Trending

- Due to the trend, the variance of the dependent variable will be overstated.
- It is better to first detrend the dependent variable and then run the regression on all the independent variables (plus a trend if they are trending as well).
- Specifically,
  1. $y_t \sim 1, t \Longrightarrow \ddot{y}_t$
  2. $\ddot{y}_t \sim 1, x_t, t$
- The R-squared of the second-step regression is a more adequate measure of fit:

$$R^2 = 1 - \frac{SSR}{\sum_{t=1}^T \ddot{y}_t^2},$$

where $\frac{1}{T}\sum_{t=1}^T \ddot{y}_t = 0$, and

$$\overline{R}^2 = 1 - \frac{SSR/(T-3)}{\left(\sum_{t=1}^T \ddot{y}_t^2\right)\Big/(T-2)}.$$

continue

- Recall that

$$\overline{R}^2 = 1 - \frac{\widehat{\sigma}_u^2}{\widehat{\sigma}_y^2}.$$

- Although $\widehat{\sigma}_u^2 = \frac{1}{T-k-1}\sum_{t=1}^T \widehat{u}_t^2$ is unbiased to σ_u^2, $\widehat{\sigma}_y^2 = \frac{1}{T-1}\sum_{t=1}^T (y_t - \overline{y})^2$ is **not** an unbiased estimator of σ_y^2 if y_t has a trend [figure here], where $\widehat{u}_t$ is from

$$y_t \sim 1, x_t, t,$$

which is the same as the residual from $\ddot{y}_t \sim 1, x_t, t$.

- Note that

$$\left(\sum_{t=1}^T \ddot{y}_t^2\right) \leq \sum_{t=1}^T (y_t - \overline{y})^2,$$

so

$$1 - \frac{SSR}{\sum_{t=1}^T \ddot{y}_t^2} \leq 1 - \frac{SSR}{\sum_{t=1}^T (y_t - \overline{y})^2},$$

i.e., the R^2 tends to be larger if y_t has not been detrended before calculating R^2.

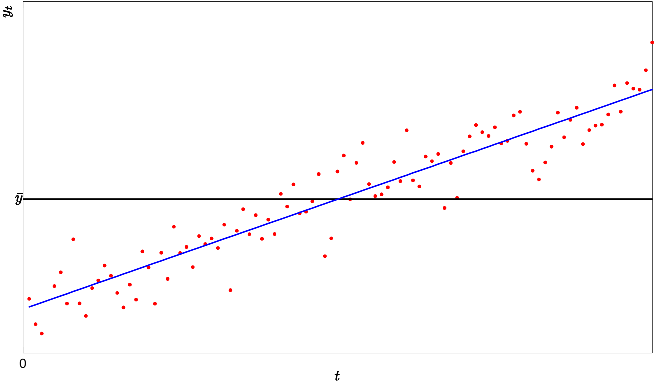


Figure: $\widehat{\sigma}_y^2 = \frac{1}{T-1}\sum_{t=1}^{T}(y_t - \overline{y})^2$ Is NOT an Unbiased Estimator of σ_y^2 If y_t Has a Trend

e: Modelling Seasonality in Time Series

- Seasonality can be used to model monthly housing starts (higher in June than in January), retails sales (higher in fourth quarter than in the previous three quarters because of Christmas), etc..
- A simple method is to include a set of seasonal dummies:

$$\begin{aligned} y_t &= \beta_0 + \delta_1 feb_t + \delta_2 mar_t + \delta_2 apr_t + \cdots + \delta_{11} dec_t \\ &\quad + \beta_1 x_{t1} + \cdots + \beta_k x_{tk} + u_t, \end{aligned}$$

  where

$$dec_t = \begin{cases} = 1, & \text{if the observartion is from december,} \\ = 0, & \text{otherwise,} \end{cases}$$

  and other dummies are similarly defined.
- Similar remarks apply as in the case of deterministic time trends:
  - The regression coefficients on the explanatory variables can be seen as the result of first deseasonalizing the dependent and the explanatory variables.
  - An *R*-squared that is based on first **deseasonalizing** the dependent variable may better reflect the explanatory power of the explanatory variables.