Basic Longitudinal Data Vocabulary

- A process is a series of actions or operations that lead to a particular end.
  - A *stochastic process* is a collection of random variables that quantify a process of interest.
- *Longitudinal data* - numerical realizations of a process that evolves over time.
  - Ordering is the key, not time. Ordering could also be spatial (oil exploration).
  - A single measurement of a process yields a variable over time, denoted by $y_1, \ldots, y_T$ and referred to as a *time series*.
  - Another type of longitudinal data where we examine a cross-section of entities, such as firms, and examine their evolution over time. This type of data is also known as *panel data*.
- *Cross-sectional data* - observations for which there is no natural ordering such as space or time

Time Series versus Causal Models

- A *causal model* can take the form

$$y_t = \beta_0 + \beta_1 x_t + \epsilon_t$$

  - Requires additional theory (for example, economics)
  - In contrast, statistical models can only validate empirical relationships ("correlation, not causation").
- In this spurious regression example,
  - Both variables evolve over time
  - Time series patterns in the explanatory variables may mask or induce a significant relationship with the dependent variable.
- In contrast, regression modeling can be readily applied when explanatory variables are simply functions of time.

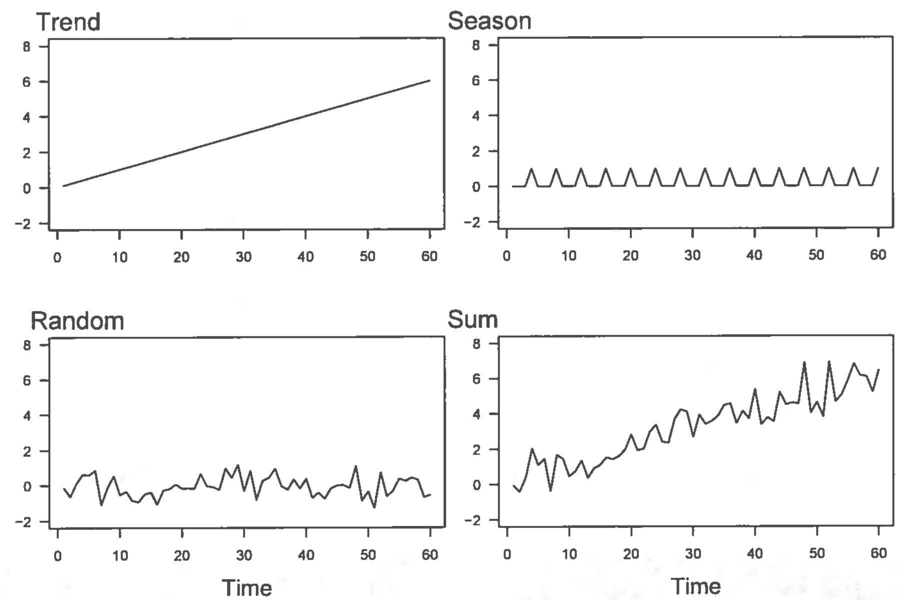
Decomposing a Time Series

- Think of a time series as being composed of
  - trends in time (T_t),
  - seasonal patterns (S_t), and
  - random, or irregular, patterns (ε_t).
- Combine these patterns in an additive fashion,

$$y_t = T_t + S_t + \varepsilon_t,$$

or in a multiplicative fashion,

$$y_t = T_t \times S_t + \varepsilon_t.$$



Fitting Trends in Time

- We can readily fit regression models to time series data if the explanatory variables are pure functions of time - no potential feedback - feedforward problems!
- The simplest model is no trend

$$y_t = \beta_0 + \varepsilon_t.$$

- Another easy model is the linear trend in time model

$$y_t = \beta_0 + \beta_1 t + \varepsilon_t.$$

- Could also a quadratic trend in time model,

$$y_t = \beta_0 + \beta_1 t + \beta_2 t^2 + \varepsilon_t,$$

or a higher-order polynomial.

Example. Hong Kong Exchange Rate

- We have $T = 502$ daily observations for the period April 1, 2005 through May 31, 2007 that were obtained from the Federal Reserve (H10 report).
  - The fitted regression equation turns out to be:

$\widehat{INDEX}_t =$	7.797	$-3.68 \times 10^{-4} t$	$+8.269 \times 10^{-7} t^2$
t-statistics	(8,531.9)	(-44.0)	(51.2)

  - Looks good. $R^2 = 92.9\%$ and typical error has dropped from $s_y = 0.0183$ down to $s = 0.0068$.
- Suppose that we wanted to predict the exchange rate for April 1, 2007, or $t = 503$. Our prediction is

$$\widehat{INDEX}_{503} = 7.797 - 3.68 \times 10^{-4}(503) + 8.269 \times 10^{-7}(503)^2 = 7.8208.$$

Stationarity

- A process that is stable over time is called *stationary*.
  - Stationarity is the formal mathematical concept corresponding to the "stability" of a time series of data.
  - A series is said to be (weakly) stationary if
    - the mean $\mathrm{E}\, y_t$ does not depend on t and
    - the covariance between y_s and y_t depends only on the difference between time units, $|t - s|$.
  - For example, under weak stationarity
    - $\mathrm{E}\, y_4 = \mathrm{E}\, y_8$ because the means do not depend on time (thus equal).
    - $\mathrm{Cov}(y_4, y_6) = \mathrm{Cov}(y_6, y_8)$ (both are two time units apart)
    - Further, $\sigma^2 = \mathrm{Cov}(y_t, y_t) = \mathrm{Cov}(y_s, y_s) = \sigma^2$. (Both are zero time units apart)
    - A weakly stationary series has a constant mean as well as a constant variance (homoscedastic).
  - The idea is that successive samples of modest size should have approximately the same distribution
    - We are particularly concerned with the mean level and variation.
    - If the process is stationary, we may define a distribution.

White Noise

- *White noise* - a process that is i.i.d.
  - It is a stationary process.
  - It displays no apparent patterns through time
- Forecasting
  - Suppose that $y_1, \ldots, y_T$ is a white noise process
  - We wish to forecast y_{T+l}, for "l" lead units in the future
  - Let $\overline{y}$ and s_y denote the sample average and standard deviation.
  - A forecast of y_{T+l} is $\overline{y}$.
  - Further, a 95% forecast interval is

$$\overline{y} \pm (t - value) s_y \sqrt{1 + \frac{1}{T}}.$$

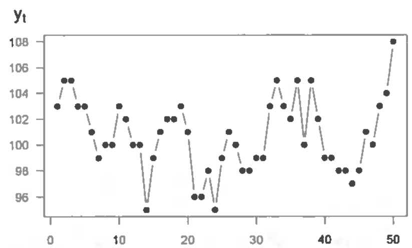
  - This interval does *not* depend on the choice of l, the number of lead units that we forecast into the future.
- White noise is both the least and the most important model.
  - Least important - most series of interest are unlikely to be i.i.d.
  - Most important - our modeling efforts are directed towards reducing a series to a white noise process.
- Procedure for reducing a series to white noise is called a *filter*.

Example: Creating a Random Walk

- Roll 2 dice - sum = c_t^*, Payoff = $c_t^* - 7$.
- Start with initial capital $y_0 = 100$.
- Update recursively $y_t = y_{t-1} + c_t$.
- The first 5 throws

t	1	2	3	4	5
c_t^*	10	9	7	5	7
c_t	3	2	0	-2	0
y_t	103	105	105	103	103

- The figure gives the first 50



Random Walk Properties

- A random walk can be expressed recursively as

$$y_t = y_{t-1} + c_t,$$

where c_t is white noise (i.i.d.)
- By repeated substitution, we have

$$y_t = c_t + y_{t-1} = c_t + (c_{t-1} + y_{t-2}) = \cdots = y_0 + c_1 + \cdots + c_t$$

where y_0 is an initial level. The random walk is the *partial sum* of a white noise process.
- The random walk is not a stationary process
  - The mean is $\mathrm{E}\, y_t = y_0 + t\mu_c$, where $\mathrm{E}\, c_t = \mu_c$
  - The variance is $\mathrm{Var}\, y_t = t\sigma_c^2$, where $\mathrm{Var}\, c_t = \sigma_c^2$.
- The random walk process is nonstationary in the variance.
- If $\mu_c \neq 0$, then the random walk process is nonstationary in the mean.

Random Walk Forecasting

- Suppose that $y_1, \ldots, y_T$ is a realization of a random walk model. We wish to forecast y_{T+l}
  - Let $c_t = y_t - y_{t-1}$ represent the differences in the series, so that

$$\begin{aligned} y_{T+l} &= y_{T+l-1} + c_{T+l} = (y_{T+l-2} + c_{T+l-1}) + c_{T+l} = \cdots \\ &= y_T + c_{T+1} + \cdots + c_{T+l}. \end{aligned}$$

  - We interpret y_{T+l} to be the current value of the series, y_T, plus the partial sum of future differences.
  - The forecast of c_{T+k} is $\overline{c}$ for $k = 1, 2, \ldots, l$.
  - The forecast of y_{T+l} is $y_T + l \times \overline{c}$.
- An approximate 95% prediction interval for y_{T+l} is

$$y_T + l\overline{c} \pm 2 s_c \sqrt{l}$$

where s_c is the standard deviation computed using the changes $c_2, c_3, \ldots, c_T$.

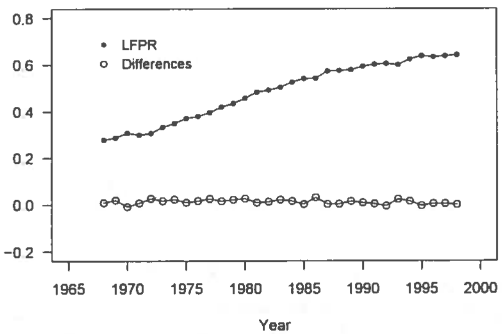
Random Walk Forecasting

- Two Dice example
  - At time 50, it turned out that our sum of money available was $y_{50} = \$93$.
  - Starting with $y_0 = \$100$, the average change was $\overline{c} = -7/50 = \$-0.14$, with standard deviation $s_c = \$2.703$.
  - Suppose that we would like to forecast y_{60}.
  - Thus, the forecast at time 60 is $93 + 10(-.14) = 91.6$.
  - The corresponding 95% prediction interval is

$$91.6 \pm 2\,(2.703)\,\sqrt{10} = 91.6 \pm 17.1 = (74.5, 108.7)\,.$$

Example. Labor Force Participation Rates

- Labor force participation rate (*LFPR*) forecasts, coupled with population forecasts, provide a picture of the future workforce.
$$LFPR = \frac{\text{civilian labor force}}{\text{civilian noninstitutional population}}.$$
- Consider data 1968-1998 for females, aged 20-44, living in a household with a spouse present and at least one child under six years of age.
- Figure shows a rapid increase in *LFPR* for this group over $T = 31$ years.

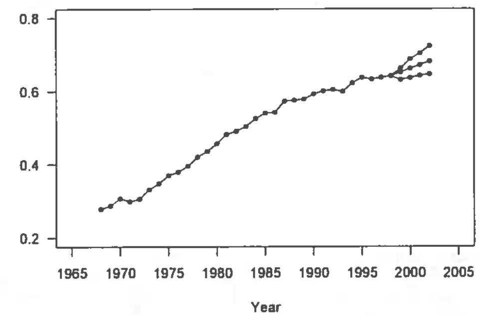


Example. Labor Force Participation Rates

- Assume that LFPR can be modeled as a random walk
- The most recent observation is $LFPR_{31} = 0.6407$.
- The average change is $\bar{c} = 0.0121$ with std dev $s_c = 0.0101$.
- An approximate 95% prediction interval for the l-step forecast is

$$0.6407 + 0.0121 l \pm 0.0202\sqrt{l}.$$

Figure illustrates prediction intervals for 1999 through 2002, inclusive. The upper and lower series represent the upper and lower 95% forecast intervals. Data for 1968-1998 represent actual values.



Identifying Random Walks

- How do we identify a series as a realization from a random walk?
- Start by deciding whether or not the series is stationary
  - For this, use a "control chart," (the basic idea is to superimpose reference lines called *control limits* on a time series plot of the data)
  - For a stationary series, successive samples of modest size should have approximately the same distribution. Look for trends in the mean or variability.
- If non-stationary, try taking differences of the series. If the differenced series is white noise, then the original series is a random walk

Random Walk versus Linear Trend in Time Models

- The LFPR example appears as if both the random walk and the linear trend in time are suitable models. How do they differ? How are they similar?
- Recall that the linear trend in time model

$$y_t = \beta_0 + \beta_1 t + \varepsilon_t,$$

  where $\{\varepsilon_t\}$ is a random error process.
- Now suppose that $\{y_t\}$ is a random walk
  - Write it as $y_{T+l} = y_T + c_{T+1} + \cdots + c_{T+l}$
  - Use $c_t = \mu_c + \varepsilon_t$.
  - Combining these two ideas, we see that

$$y_t = y_0 + \mu_c t + u_t$$

  where $u_t = \sum_{j=1}^{t} \varepsilon_j$.
- The linear trend and random walk have the same mean portion,
- The variability about the line differs dramatically.