Time Series Analysis

Trend

Hossein Moradi Rekabdarkolaee

South Dakota State University

email: *hossein.moradirekabdarkolaee@sdstate.edu*.

Office: CAME Building, *Room: 240*.

Review

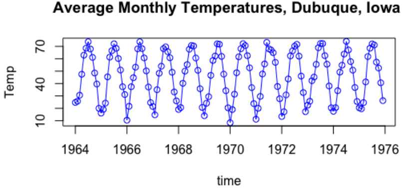
- So far, we learned that a time series has 3 major components
  - Trend,
  - Seasonality,
  - Noise.
- Trend and Seasonal variation are often referred to as *large variation*.
- In this Section we will focus on the large variation component.
- Trend is a systematic change in a time series.
- Finding the Trend could be a difficult task.

Deterministic Versus Stochastic Trends

- As an example, a simulated random walk might be considered to display a general trend.
- However, we know that the random walk process has zero mean for all time.
- Thus, this trend is just an artifact of the strong positive correlation between the series values at nearby time points and the increasing variance in the process as time goes by.
- A second and third simulation of exactly this process might show completely different "trends".
- These can be considered as **stochastic trends**.

Deterministic Versus Stochastic Trends

- How about the ”Average Monthly Temperatures” data:



- There is a clear cyclical or seasonal trend.
- We can model this time series as $Y_t = \mu_t + \epsilon_t$, where μ_t is a deterministic function that is periodic with period 12; that is μ_t, should satisfy

$$\mu_t = \mu_{t-12}$$

- We describe this model as having a **deterministic trend**.

Deterministic Trend

- A deterministic trend can be expressed as a function of time i.e. $\mu_t = g(t)$.
- This function can be
  - Constant at all time i.e. $g(t) = c$.
  - A linear function of time i.e.

    $$\mu_t = \beta_0 + \beta_1 t$$

  - A quadratic function of time

    $$\mu_t = \beta_0 + \beta_1 t + \beta_2 t^2$$

- Note that an implication of the model $Y_t = \mu_t + \epsilon_t$ with $E(\epsilon_t) = 0, \forall\ t$, is that the deterministic trend μ_t applies for all time.
- Thus, we need a good reason for this assumption.

Trend and Seasonal Variation

- Usually, we remove (subtract) these components from data to get a stationary time series.
- That is,

Data = Trend + Seasonality + Stationary TS

- Note that seasonality in time series can be deterministic or stochastic.
- Also, stochastic seasonality can be stationary or non-stationary.
- We want to get an estimate of the large variation part and remove it.

Estimation of a Constant Mean

- Consider a model with constant mean μ.
- Then the time series becomes

$$Y_t = \mu + \epsilon_t$$

- Then, the estimator for the μ for the observed time $1, 2, \ldots, n$ is

$$\overline{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i,$$

  which is the sample's mean.
- This is an unbiased estimator of μ.

Estimation of a Constant Mean

- If Y_t be a stationary time series with autocorrelation function ρ_k, then we have

$$Var(\bar{Y}) = \frac{\gamma_0}{n}\Big[\sum_{k=-n+1}^{n-1}(1-\frac{|k|}{n})\rho_k\Big]$$

$$= \frac{\gamma_0}{n}\Big[1 + 2\sum_{k=1}^{n-1}(1-\frac{k}{n})\rho_k\Big]$$

- Recall the stationary moving average time series

$$Y_t = \frac{e_t + e_{t-1}}{2}$$

where e_ts are i.i.d with mean 0 and variance σ^2.

- What is the variance of the sample mean $\bar{Y}$?

Estimation of a Constant Mean

- How about the following time series

$$Y_t = \frac{e_t - e_{t-1}}{2}$$

- What is the variance of the sample mean $\bar{Y}$?
- A positive correlation makes the estimate of the mean to be less efficient compared to the estimate obtained in the white noise case and a negative correlation improves the estimation of the mean compared with the estimation obtained in the white noise.
- For many stationary processes, the autocorrelation function decays quickly enough with increasing lags that

$$\sum_{k=0}^{\infty} |\rho_k| < \infty$$

Stationarity & the variance of the sample mean

- Given a large sample size n, and assuming $\sum_{k=0}^{\infty} |\rho_k| < \infty$, the following is a useful approximation of the sample mean-variance

$$Var(\bar{Y}) \simeq \frac{\gamma_0}{n} \left[\sum_{k=-\infty}^{\infty} \rho_k \right]$$

- As an example, suppose that $\rho_k = \phi^{|k|}$ for all k, where ϕ is a number strictly between -1 and $+1$.

- Summing a geometric series yields

$$Var(\bar{Y}) \simeq \frac{(1+\phi)}{(1-\phi)} \frac{\gamma_0}{n}$$

Stationarity & the variance of the sample mean

- Consider the random walk process:

$$Y_t = \sum_{j=1}^{t} e_j$$

- What is the mean and its variance?
- Notice that in this special case, the variance of our estimate of the mean actually increases as the sample size n increases.
- Clearly, this is unacceptable, and we need to consider other estimation techniques for nonstationary series.

Regression Methods

- Consider the deterministic time trend to be a linear function of time expressed as
$$\mu_t = \beta_0 + \beta_1 t$$
- The **least squares (LS)** method can be used to estimate the intercept β_0 and the slope β_1 by minimizing
$$Q(\beta_0, \beta_1) = \sum_{t=1}^{n} (Y_t - \beta_0 - \beta_1 t)^2$$
- It can be shown that the LS estimators are
$$\hat{\beta}_1 = \frac{\sum_{t=1}^{n} (Y_t - \bar{Y})(t - \bar{t})}{\sum_{t=1}^{n} (t - \bar{t})^2}$$
$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{t}$$

Interpreting Regression Output

- Some of the properties of the regression output depend heavily on the usual regression assumption that
  - $\{\epsilon_t\}$ is white noise, and some depend on the further assumption that
  - $\{\epsilon_t\}$ is approximately normally distributed.
- If the $\{\epsilon_t\}$ process has a constant variance, then we can estimate the standard deviation of ϵ_t, namely $\sqrt{\gamma_0}$, by the **residual standard deviation**

$$s = \sqrt{\frac{1}{n-p}\sum_{t=1}^{n}(Y_t - \hat{\mu}_t)^2}$$

Interpreting Regression Output

- **Coefficient of determination R^2:**

$$R^2 = \frac{SSR}{SS_Y} = 1 - \frac{SSE}{SS_Y} = 1 - \frac{\sum_{t=1}^{n}(Y_t - \hat{\mu}_t)^2}{\sum_{t=1}^{n}(Y_t - \bar{Y})^2}$$

  - It is the square of the sample correlation coefficient between the observed series and the estimated trend.
  - It is also the fraction of the variation in the series that is explained by the estimated trend.

- The adjusted R-squared value is a small adjustment to R^2 that yields an approximately unbiased estimate based on the number of parameters estimated in the trend.

Residual Analysis

- The unobserved stochastic component $\{\epsilon_t\}$ can be estimated, or predicted, by the **residual**

$$\hat{\epsilon}_t = Y_t - \hat{\mu}_t$$

- If the stochastic component is white noise, then the residuals should behave roughly like independent (normal) random variables with zero mean and standard deviation *s*.

- Similar to regression, we need to run some diagnostics on residuals.

Residual Analysis

- look at the plot of the residuals over time.
- look at the standardized residuals versus the corresponding trend estimate or fitted value.
- Look at the histogram of the residuals or standardized residuals (gross non-normality can be assessed by a histogram)
- Normality can be checked more carefully by plotting the so-called normal scores or quantile-quantile (QQ) plot
- **Shapiro-Wilk** test (an excellent test of normality)

The Sample Autocorrelation Function

- Another very important diagnostic tool for examining dependence is the sample autocorrelation function.

$$r_k = \frac{\sum_{t=k+1}^{n} (Y_t - \bar{Y})(Y_{t-k} - \bar{Y})}{\sum_{t=1}^{n} (Y_t - \bar{Y})^2}, \qquad k = 1, 2, \ldots$$

- The plot of k vs r_k for $k = 1, ..., J$, where usually $10 \leq J \leq n/4$ with horizontal lines at $\pm 2/\sqrt{n}$ (two standard deviations of the sample autocorrelations) added to the plot.

- This plot is the ACF plot that you are already familiar with.