Tutorial 4 - STAT 485/685

Trevor Thomson

Department of Statistics & Actuarial Science
Simon Fraser University, BC, Canada

October 5, 2020

Today's Plan

1. Recap of Tutorial 3
   - Trends
   - Residual Analysis
2. Examples
   - Example 1: Questions 3.5 & 3.11
   - Example 2: Questions 3.6 & 3.12
3. Random Walk Example in R

Recap of Tutorial 3

Trends

- $\{Y_t : t \in \mathcal{I}\}$ is a time series
  - ...a realization of a stochastic process, t is *time*.
- Express each term as

$$Y_t = \mu_t + X_t,$$

  where $E(Y_t) = \mu_t \Rightarrow E(X_t) = 0$.
- **Goal**: Model and estimate μ_t.
  - $\mu_t = \mu \Rightarrow$ *constant mean model*
  - $\mu_t = \beta_0 + \beta_1 t \Rightarrow$ *linear trend model*
  - $\mu_t = \mu_{t+k}$ for some $k \Rightarrow$ *seasonal mean model*
  - $\mu_t = \beta_0 + \beta_1 \cos(2\pi f t) + \beta_2 \sin(2\pi f t) \Rightarrow$ *cosine trend model*
- $\Rightarrow$ Estimate parameters in μ_t by minimizing the *least squares* objective function.

Recap of Tutorial 3

Trends

For the time series $\{Y_t : t \in \mathcal{I}\}$:

- ...textbook specifies $\mathcal{I} = \{1, 2, \cdots, n\}$.

For the `larain` dataset:

- Let Y_{t-1877} denote the value of precipitation (in inches) in Los Angeles in year t.

  We can allow for $\mathcal{I} = \{1, 2, \cdots, 115\}$ (i.e. $n = 115$).

For the `wages` and `tempdub` datasets:

- $\mathcal{I} = \{1, 2, \cdots, n\}$?

  $\Rightarrow$ depends on the time scale we want!

  - If the time scale is **months**, $\mathcal{I} = \{1, 2, \cdots, n\}$ is okay!
  - If the time scale is **years**, $\mathcal{I} = \{1, 2, \cdots, n\}$ doesn't make sense if we have monthly observations!

    Problem: What if we want to model in this years?

  You saw this issue in Question 2 of Assignment 2!

- See **Week 4 - Video 11** on specifying *indicator variables* as regressors.

Recap of Tutorial 3

Residual Analysis

- With $Y_t = \mu_t + X_t$,
  - Specify a model for μ_t.
  - Estimate the parameters
    - Obtain $\hat{\mu}_t$
  - **Now what?**

- Estimate the "error term" X_t with

$$\hat{X}_t = Y_t - \hat{\mu}_t.$$

- **Basic idea**: If we specify a "good" model for $\mu_t \Rightarrow X_t \approx 0$.
  - If we properly accounted for the autocorrelation structure present within $\{Y_t : t \in \mathcal{I}\}$, then $\{X_t : t \in \mathcal{I}\}$ should "behave" like white noise.
  - Since we don't know $\{X_t : t \in \mathcal{I}\} \Rightarrow$ we use $\{\hat{X}_t : t \in \mathcal{I}\}$ instead.

Recap of Tutorial 3

Residual Analysis

- Recall, $\{X_t : t \in \mathcal{I}\}$ is a white noise process if
  - X_t are iid random variables.
  - $E(X_t) = 0$.
  - $Var(X_t) = \sigma_e^2$.

  There are a few general checks to see if $\{X_t : t \in \mathcal{I}\}$ "behaves" like white noise, and other general goodness-of-fit procedures we can conduct:

- **Estimate** σ_e

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

where p is the number of parameters.

- **Compute R^2 - Coefficient of Determination**

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

  Here, R^2 is the proportion of the variation in Y_t that the variables in $\hat{\mu}_t$ can explain.

  - Generally, $R^2 \approx 1$ means that that the $\hat{\mu}_t$ fits the data well.

- **Compute The Adjusted R^2 - Adjusted Coefficient of Determination**

$$\bar{R}^2 = 1 - (1 - R^2)\frac{n-1}{n - p^* - 1},$$

  where p^* is the number of covariates in the model (excluding the intercept).

  - R^2 increases if we keep adding variables to the model.
  - $\bar{R}^2$ at least accounts for the number of variables included in the model.

SFU

Recap of Tutorial 3

Residual Analysis

- **Plot The Residuals Over Time**

  If $\{X_t : t \in \mathcal{I}\}$ "behaves" like white noise, we should see a random scatter around 0.

- **Histogram of Residuals**

  Empirically plot the distribution of $\{X_t : t \in \mathcal{I}\}$, to see if it follows the normal distribution.

- **Normal Q-Q Plot**

  Another empirical check to see if $\{X_t : t \in \mathcal{I}\}$ is normally distributed.

- **Perform The Runs Test**:

$$\begin{aligned} H_0 &: \text{Elements of } \{X_t : t \in \mathcal{I}\} \text{ are mutually independent} \\ H_a &: \text{Not } H_0. \end{aligned}$$

If we fail to reject H_0, X_t are independent random variables.

- **Compute the Sample Autocorelation Function (ACF)**

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

This is a very useful quantity, **we will see this again later in the course!**

We can perform the following hypothesis test for each k with $\rho_k = Corr(Y_t, Y_{t-k})$:

$$H_0 : \rho_k = 0$$
$$H_a : \rho_k \neq 0.$$

If we fail to reject H_0 for each k, Y_t are independent random variables.

$\Rightarrow$ Useful for both $\{Y_t : t \in \mathcal{I}\}$ and $\{X_t : t \in \mathcal{I}\}$.

- `acf()` in R plots r_k vs. k (a *correlogram*) and 95% confidence intervals under H_0.

SFU

Examples

- See the R file `Tutorial3.R` where solutions are provided for Questions 3.5, 3.6, 3.11, and 3.12.

Random Walk Example in R

- Recall the *random walk* from Tutorial 1:

  Let $Y_t = Y_{t-1} + e_t$, where $Y_0 = 0$, where $\{e_t : t \in \mathbb{N}\}$ is a white noise process.

$$
\begin{aligned}
Y_1 &= e_1 \\
Y_2 &= Y_1 + e_2 = e_1 + e_2 \\
Y_3 &= Y_2 + e_3 = e_1 + e_2 + e_3 \\
&\vdots \\
Y_t &= \sum_{u=1}^{t} e_u.
\end{aligned}
$$

Random Walk Example in R

- Recall the *random walk* from Tutorial 1:

  Let $Y_t = Y_{t-1} + e_t$, where $Y_0 = 0$, where $\{e_t : t \in \mathbb{N}\}$ is a white noise process.

$$\begin{aligned} Y_1 &= e_1 \\ Y_2 &= Y_1 + e_2 = e_1 + e_2 \\ Y_3 &= Y_2 + e_3 = e_1 + e_2 + e_3 \\ &\vdots \\ Y_t &= \sum_{u=1}^{t} e_u. \end{aligned}$$

  Recall the

  - mean function: $\mu_t = E(Y_t)$
  - autocovariance function: $\gamma_{t,t-k} = Cov(Y_t, Y_{t-k})$
  - autorcorrelation function: $\rho_{t,t-k} = Corr(Y_t, Y_{t-k})$

Random Walk Example in R

For $k \geq 0$:

μ_t:

$$\mu_t = E(Y_t) = E\left(\sum_{u=1}^{t} e_u\right) = \sum_{u=1}^{t} E(e_u) = \sum_{u=1}^{t} 0 = 0$$

$\gamma_{t,t-k}$:

$$\gamma_{t,t} = Var(Y_t) = Var\left(\sum_{u=1}^{t} e_u\right) = \sum_{u=1}^{t} Var(e_u) = \sum_{u=1}^{t} \sigma_e^2 = t\sigma_e^2$$

$$\gamma_{t,t-k} = Cov(Y_t, Y_{t-k}) = Cov\left(\sum_{u=1}^{t} e_u\ ,\ \sum_{v=1}^{t-k} e_v\right) = (t-k)\sigma_e^2$$

$\rho_{t,s}$:

$$\rho_{t,t-k} = \frac{\gamma_{t,t-k}}{\sqrt{\gamma_{t,t}} \times \sqrt{\gamma_{t-k,t-k}}} = \frac{(t-k)\sigma_e^2}{\sqrt{t\sigma_e^2} \times \sqrt{(t-k)\sigma_e^2}} = \sqrt{\frac{t-k}{t}}.$$

Random Walk Example in R

With the random walk, consider the following two questions:

- **Question**: Suppose we propose the following regression model

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

$\Rightarrow E(Y_t) = \beta_0 + \beta_1 t$. But we know that $E(Y_t) = 0$.

$\Rightarrow \beta_0 = \beta_1 = 0$.

If we use `lm()` in R, can we use

(1) the reported estimates $\hat{\beta}_0$ and $\hat{\beta}_1$?

(2) the reported standard errors of $\hat{\beta}_0$ and $\hat{\beta}_1$?

$\Rightarrow$ Let's conduct a *simulation study* in R to answer (1) and (2)!

- See `Tutorial4.R`

SFU

Office hour Tomorrow:

Tuesday, October 6, 7:00-8:00 PM (PT)

See Canvas for the Zoom link

Good luck on your midterm!