Module 11: Time Series Regression and Cross Validation

Jeffrey Woo

MSDS, University of Virginia

Assumptions in Linear Regression

Linear regression model:

$$y_i = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + \epsilon_i.$$

Assumptions: ϵ_i are i.i.d. (independent and identically distributed) $N(0, \sigma^2)$. A by-product of the assumptions is that the observations are independent from one another.

Assumptions in Linear Regression

When data are collected in a sequence, $x_t, x_{t+1}, x_{t+2}, \cdots$, the observations are unlikely to be independent. For example:

- Earnings in successive quarters.
- Temperature in successive days.

R will always allow us to fit a regression model, regardless of assumptions being met.

Assumptions in Linear Regression

If assumptions are not met, results from hypothesis tests and confidence intervals are no longer reliable as these are calculated under the assumptions being true.

Comments on Time Series Analysis

The Orcutt-Cochrane method for regression with AR errors is a simplified version of the more general lagged regression with ARMA errors:

- In lagged regression, we regress y_t on lagged versions of the predictor series, $x_{t-1}, x_{t-2}, \cdots$.
- For ARMA(p,q) errors:
  $\epsilon_t = \phi_1 \epsilon_{t-1} + \cdots \phi_p \epsilon_{t-p} + a_t + \theta_1 a_{t-1} + \cdots + \theta_q a_{t-q}$. A good resource is: "Time Series Analysis and Its Applications, with R Examples (4th Ed)." Shumway & Stoffer.
- In financial time series, the interest is more in estimating the conditional variance, rather than the mean, of x_t. A good resource is: "An Introduction to Analysis of Financial Data with R." Tsay.

Comments on Cross Validation

- You may realize that when we briefly touched on validation in a linear regression setting, we were using the validation set approach.
- LOOCV is another approach at the other end of the spectrum. It has a fixed final value, but takes a longer time to run.
- k-fold CV is considered a compromise between the two approaches.
- LOOCV: smallest bias, highest variance. Validation set approach: largest bias, smallest variance.
- No one approach is "always better" than the other. Usually, k-fold is used most commonly in practice, with $k = 5, 10$ as fairly common values.