Module 11: Regression for Time Series Data and Resampling Methods

Live Sessions	Office Hours		Dr. Woo
Aug 5: 9:00 - 9:50 am EDT (Blue), 10:00 - 10:50 am EDT (Orange)	Tue & Thu: 9:00 - 9:50 am EDT (Blue), 10:00 - 10:50 am EDT (Orange).		yjw4b@virginia.edu

MODULE WELCOME

In modules 1 to 9, you learned about the linear regression model. An underlying assumption in the linear regression model is that the error terms, and hence the observations, are all independent. This assumption is fine when your data are based on a random sample. However, in some instances, your data are collected in a sequence of time intervals. For example, your data could contain the prices of various stocks at the end of each day for a year. The price today is likely going to be correlated with the price yesterday and perhaps even previous days. The assumption of independent errors is no longer reasonable. In this module, you will learn how to account for this correlation between errors and how you can apply a simple transformation to the data that will allow you to still use the linear regression framework.

In the second part of this module, you will learn about cross-validation. Cross-validation is a resampling method that randomly uses part of your data set to assess how well your model performs. You have seen a simple version of cross-validation in module 7, where it was called data splitting. In this part of the module, you will learn about some variants of cross-validation, as well as the pros and cons of each variant. Cross-validation is a commonly used way to assess how well your model performs, especially when you use shrinkage methods in the next module.

In this module, you will complete and submit an evaluation of your own participation and that of your group members in the small-group discussions of the guided question sets for modules 8 through 11.

ESSENTIAL QUESTIONS

- *What are the consequences of using linear regression when the observations are not independent?*

- *How do we transform the variables to account for the correlation between observations?*

- *What are the various methods of cross-validation and what are their pros and cons?*

LEARNING OBJECTIVES

1. Detect the presence of autocorrelation in error terms.

2. Explain how the presence of autocorrelation in error terms affects the least-squares regression model.

3. Describe the setup of an autoregressive (AR) error model, and use the model to answer questions of interest.

4. Explain the pros and cons of the various methods of cross-validation.

5. Use cross-validation to decide which regression model to use in a data analysis.

ASSIGNED RESOURCES

- *Introduction to Linear Regression Analysis*, Section 14 (skip "The Maximum Likelihood Approach," pp. 485 to 488)

- Section 5 up to Section 5.1.4 in *An Introduction to Statistical Learning, with Applications in R* (after clicking this link, click on "Download the book PDF on the website")

- Module 11 R tutorial, data set: titanic.txt, data set: company.txt, and R-code: tutorial_module 11.R

- Module 11 guided question set and data set: nfl.txt

OPTIONAL

Visit the Module 11 Discussion Forum to communicate with one another about the assigned readings and resources or to discuss any other topics of interest with your instructor or fellow students.

💬 Module 11 General Discussion

MODULE OVERVIEW

- 11.1: Introduction to the Lesson
- 11.2: Autocorrelation in Errors
- 11.3: Estimating the Regression Model with Autocorrelated Errors and Using it to Make Predictions
- 11.4: Cross-Validation
- 11.5: Recap for Module 11
- 11.6: R Tutorial for Module 11
- 11.7: Module 11 Live Session
- 11.8: Optional - Module 11 Assignments Due