Assignment 08

Cross-Validation

This goal of this assignment is to give you experience using cross-validation methods in regression analyses. Turn in a printed version of your responses to each of the questions on this assignment. In addition, please adhere to the following guidelines for further formatting your assignment:

- All graphics should be set to an appropriate aspect ratio and sized so that they do not take up more room than necessary. They should also have an appropriate caption.
- Any typed mathematics (equations, matrices, vectors, etc.) should be appropriately typeset within the document.
- Syntax or computer output should not be included in your assignment unless it is specifically asked for.

This assignment is worth 15 points.

Part I: Minneapolis Violent Crime

For the first part of this assignment, you will use the file *mpls-violent-crime.csv*. This file contains data collected from the Minneapolis Police Department and reported by the Star Tribune on the rate of violent crimes since 2000. The variables are:

- `year`: Year
- `crime_rate`: Violent crime rate per 100,000 people

All the analyses in Part I will examine the trend in violent crime rate over time.

Preparation

Create a variable that indicates the number of years since 2000. Use this variable in all analyses for Part I rather than the `year` variable.

Description

1. Create a scatterplot showing the violent crime rate as a function of time.
2. Based on the plot, describe the trend in violent crime rate over time.
3. If you were going to fit a polynomial model to these data, what degree polynomial would you fit to? Explain.

Use p-Value methods for Model Selection

4. Fit a series of polynomial models starting with a linear model, and then models that also include higher order polynomials that allow you to evaluate your response to Question #3. Be sure to fit models up to degree $k + 1$, where k is the degree you hypothesized in Question #3. Report the results from the ANOVA table.
5. Based on these results, which polynomial model would you adopt? Explain.

Using LOOCV for Model Selection

In this section of the assignment, you are going to use LOOCV to evaluate the MSE for the same set of polynomial models you evaluated in Question #4.

6. Write and include syntax that will carry out the LOOCV.
7. Report the cross-validated MSE for each of the models in your set of polynomial models.
8. Based on these results, which degree polynomial model should be adopted? Explain.

Part II: Course Evaluations

For the second part of this assignment, you will use the file *evaluations.csv.* This file contains data collected from student evaluations of instructors' beauty and teaching quality for several courses at the University of Texas. The teaching evaluations were conducted at the end of the semester, and the beauty judgments were made later, by six students who had not attended the classes and were not aware of the course evaluations. The variables are:

- `prof_id`: Professor ID number
- `avg_eval`: Average course rating
- `num_courses`: Number of courses for which the professor has evaluations
- `num_students`: Number of students enrolled in the professor's courses
- `perc_evaluating`: Average percentage of enrolled students who completed an evaluation
- `beauty`: Measure of the professor's beauty composed of the average score on six standardized beauty ratings
- `tenured`: Is the professor tenured? (0 = non-tenured; 1 = tenured)
- `native_english`: Is the professor a native English speaker? (0 = non-native English speaker; 1 = native English speaker)
- `age`: Professor's age (in years)
- `female`: Is the professor female? (0 = male; 1 = female)

These source of these data is: Hamermesh, D. S. & Parker, A. M. (2005). Beauty in the classroom: Instructors' pulchritude and putative pedagogical productivity. *Economics of Education Review, 24*, 369–376. The data were made available by: Gelman, A., & Hill, J. (2007). *Data analysis using regression and multilevel/hierarchical models.* New York: Cambridge University Press.

Preparation

Begin by fitting a model that predicts average course evaluation score using the following predictors: beauty, number of courses for which the professor has evaluations, whether the professor is a native English speaker, and whether the professor is female.

9. Using average course evaluation scores (y), compute the total sum of squares (SST). Show your work.
10. Using average course evaluation scores (y) and the predicted values from the model ($\hat{y}$), compute the sum of squared errors (SSE). Show your work.
11. Compute the model R^2 value using the formula: $1 - \frac{\text{SSE}}{\text{SST}}$.

Using k-Fold Cross-Validation to Estimate R^2

As mentioned in class, the estimate for R^2 is biased. We can obtain a better estimate of R^2 by using cross-validation. You will use 5-fold cross-validation to estimate the R^2 value. The algorithm for this will be:

- Randomly divide the beauty data into 5 folds.
- Hold out 1 fold as your validation data and use the remaining 4 folds as your training data.
  - Fit the model to the training data.
  - Use the estimated coefficients from those fits to compute $\hat{y}$ values using the validation data.
  - Compute the SST and SSE values for the validation data, and use those to compute R^2 based on the formula $1 - \frac{\text{SSE}}{\text{SST}}$. (Note that sometimes the R^2 may be negative when we compte it in this manner.)
- Repeat for each fold.
- Compute the cross-validated R^2 by finding the mean of the five R^2 from the cross-validations.

12. Write and include syntax that will carry out the 5-fold cross-validation. In this syntax use `set.seed(1000)` so that you and the answer key will get the same results. (This website may be useful in using the **purrr** package to obtain the y- and $\hat{y}$-values in order to compute the SST and SSE values: https://drsimonj.svbtle.com/k-fold-cross-validation-with-modelr-and-broom)

13. Report the five R^2 values from your analysis and the cross-validated R^2 value.

14. How does this value compare to the R^2 value you computed in Question #11, based on the data.

15. Explain why the cross-validated estimate of R^2 is a better estimate than the data-based R^2.