Assignment 07

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 20 points.

Part I: Minneapolis Violent Crime

For the firt part of this assignment, you will use the data provided in the file *mpls-violent-crime.csv* (see the data codebook) to build a model that examines 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? 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. Analyze each of the polynomial terms (including the linear term) by using a series of nested F-tests. Report these results in an ANOVA table. (Note: If you need a refresher on fitting polynomial models and carrying out a nested F-test, see the Polynomial Regression notes from EPsy 8252.)

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 data provided in the file *evaluations.csv* (see the data codebook) to build a model that predicts variation in course evaluation scores.

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.

Description

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.

Part III: Credit Balance

For the third part of this assignment, you will again use the file the data provided in the file *credit.csv* (see the data codebook) to build a model that predicts customers' credit card balance.

16. Use the `lm.ridge()` function to fit the same sequence of *d* values you used in the FOR loop from Assignment 6, Question 7. Running `select()` on this output, provides *d* values based on different criteria. Report the *d* value associated with the generalized cross-validation (GCV) metric.
17. Re-run the FOR loop from Assignment 6, Question 7. Except this time compute the AICc and select the *d* value based on using the AICc. How does this compare to the *d* value you found using the GCV metric from the previous question? (Show your syntax.)
18. Use 10-fold cross-validation to find the modified *d* value that has the lowest CV-MSE based on fitting an elastic net. (Prior to carrying out this analysis, set the seed for the random number generation to 100.) Then use this modified *d* value to re-fit the elastic net. Report the modified *d* value and the fitted equation from the elastic net.
19. Compare the coefficients from the elastic net (from Question 18) to those from the ridge regression analyses fitted using the *d* value you found in Question 16. How are they different? Explain based on the penalty terms in the two models.
20. Use the elastic net to compute a multiple R^2 value. Remember that multiple R^2 is the squared correlation between the observed and predicted values. Report this value. (Show your work.)