Making Data-driven Predictions  
Using Linear Regression   
Homework Five

**Instructions:**

In this homework, you will:

* Use R for model evaluation and linear regression
* Develop a regression model
* Use the logit model on a data set
* Analyze with multicollinearity and autocorrelation
* Create a logistic regression model

Except as indicated, use this document to record all your homework work and responses to any questions. At a minimum you will need to turn in a digital copy of this document to your instructor as part of your homework completion. You may also have additional supporting documents that you will need to submit. Your instructor will provide feedback to help you work through your findings

**Note:** Though your work will only be seen by those grading the course and will not be used or shared outside the course, you should take care to obscure any information you feel might be of a sensitive or confidential nature.

*Complete each homework part as you progress through the unit. Wait to submit the homework until all parts are complete. Begin your unit homework by completing Part One below. A submit button can be found on Unit Homework assignment pages. Do not hesitate to contact your instructor if you have any questions about the homework.*

**Part One**Use R for Model Evaluation and Linear Regression  
\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

In this part of the course homework, you will use R for model evaluation and linear regression. This part of the homework requires some work in RStudio, which is located on the Use R for Model Evaluation homework page in Canvas. Use that space, along with the provided script and data file, to perform the work, then use this document to answer questions on what you discover.

1. Interpret the coefficients in the model.

|  |
| --- |
|  |

1. Compute the standard errpr for the intercept and slope in the regression equation for the problem above.

|  |
| --- |
|  |

1. What are the 90% confidence and prediction intervals for a price of 4 and for a price of 3? How does the width of the intervals change as we move from prediction to confidence interval? How about from a price of 4 to a price of 3?

|  |
| --- |
|  |

1. In R you can use the anova command to get an analysis of variance table. In the results table, you will find RSS and ESS (remember that TSS is the sum of RSS and ESS). Compute R2 from that table, and also interpret the F test.

|  |
| --- |
|  |

1. In the package MPV there is a function to compute PRESS. It is called using PRESS() where inside the () is the name of your model. After running this command and computing the PRESS statistic, interpret the value you obtain.

|  |
| --- |
|  |

1. Using the example above conduct 2-fold cross-validation. What do you notice about these results? Please include the results you have found.

|  |
| --- |
|  |

**Part Two**Develop a Regression Model  
\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

In this part of the course homework, you will develop a regression model on some election data from the 2016 US Presidential election. This part of the homework requires some work in RStudio, which is located on the Develop a Regression Model homework page in Canvas. Use that space, along with the provided script and data file, to perform the work, then use this document to answer questions on what you discover.

1. Which independent variables did you use to construct the "best" model? Evaluate the fit of this model and justify your choices. This should include which variables you have chosen to include and what the coefficients tells you about each variables impact. Address how good this model might be used in a predictive mode.

|  |
| --- |
|  |

**Part Three**Analyze with Multicollinearity and Autocorrelation  
\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

In this part of the course homework, you will use multicollinearity and autocorrelation to answer some short analytical questions.

1. Suppose you have created a regression equation with 2 independent variables. Also suppose both coefficients are statistically significant, with p-values that are less than 0.05. Now, suppose you add two new variables and now only one of the now four independent variables are statistically significant. Suppose the k-fold validation is better for this new model, too. What does the improved performance for k-fold validation mean about the second model? What might be happening?

|  |
| --- |
|  |

1. Suppose you are performing model diagnostics for autocorrelation, and one of the independent variables is time period. How might a residual versus this independent variable plot look if *negative* autocorrelation is present?

|  |
| --- |
|  |

**Part Four**Use the Logit Model  
\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

In this part of the course homework, you will use the logit model.

1. Suppose the probability that a consumer purchases unique candy offering at the movie theater concession stand is where x is the price of the candy. Graph this function and the log odds function (you can do this however you are most comfortable, such as using [wolframalpha.com](http://wolframalpha.com/)). Please show how the probabilities on the y-axis of the first graph map to the log odds on the y-axis of the second graph.

|  |
| --- |
|  |

1. Why does the log odds graph intersect the y-axis at 3?

|  |
| --- |
|  |

**Part Five**Create a Logistic Regression Model  
\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

In this part of the course homework, you will create a logistic regression model to determine if there is an interaction effect between the After variable and one of the several Price variables by creating a logistic regression model that includes the interaction term. Explain why you have come to this model and believe it is the “best” model of the underlying homeowner behavior.

This part of the homework requires some work in RStudio, which is located on the Perform Hotspot Analysis on a Dataset homework page in Canvas. Use that space, along with the provided script and data file, to perform the work, then use this document to answer questions on what you discover.

1. What Price variable did you choose? Why? What did you uncover by examining these interaction terms?

|  |
| --- |
|  |

1. Explain why you have come to this model and believe it is the “best” model of the underlying homeowner behavior. In your answer, refer to quantitative measures of fit.

|  |
| --- |
|  |

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_  
*To submit this assignment, please refer to the instructions in the course.*