Assignment 05

Information Criterion

This assignment is intended to have you build your understanding of using information criteria for model selection. *Do not include any R syntax or output unless it is specifically required in the question.* Please submit your responses to each of the questions below. Please submit your responses to each of the questions below in a printed document. All graphics should be resized so that they do not take up more room than necessary and should have an appropriate caption. All tables should also have an appropriate caption.

This assignment is worth 20 points. Each question is worth 1 point unless otherwise noted.

For this assignment, you will use the file *wine.csv*. This file contains data on 200 different wines. These data are a subset of a larger database ($n = 6,613$) from Wine.com, one of the biggest wine e-commerce retailers in the U.S. It allows customers to buy wine according to any price range, grape variety, country of origin, etc. The data were made available at http://insightmine.com/.The variables are:

- `wine`: Wine name
- `vintage`: Year the wine was produced (centered so that 0 = 2008, 1 = 2009, etc.)
- `region`: Region of the world where the wine was produced
- `varietal`: Grape varietal (e.g., Cabernet Sauvignon)
- `rating`: Wine rating on a 100 pt. scale (these are from sources such as *Wine Spectator*, the *Wine Advocate*, and the *Wine Enthusiast*)
- `price`: Price in U.S. dollars

You will be using these data to examine several different predictors of wine rating (a measure of the wine's quality). The literature has suggested that price of wine is quite predictive of a wine's quality. You will be carrying out a replication study (using a different data set) of a study published by Snipes and Taylor (2014).

Preparation

Read the article Model selection and Akaike Information Criteria: An example from wine ratings and prices. Fit the same nine models that Snipes and Taylor fitted in their analysis.

Table of Model Evidence

1. Create a table that includes the following information for each of the candidate models. **(2pts.)**

- Model name
- K
- AICc
- Delta
- Akaike Weight
- Logarithm (base-10) of the Evidence Ratio (see the Snipes and Taylor paper)
- Model Probability

Model Selection

2. Using information from the table you created, which candidate model(s) you will adopt? Explain by referring to the evidence ratio of the second-best model. (Hint: Back-transform the log of the evidence ratio.) **(2pts.)**

3. Interpret the model probability for your adopted model.

4. Present the model coefficients, standard errors for your adopted model in a table. Also include the R-squared value for this model. **(2pts.)**

Assumptions and Transformations

5. Examine whether the regression assumptions (linearity, independence, normality, homoscedasticity) are satisfied for the adopted model. Provide any evidence (graphical or numerical) you use in this endeavor. **(2pts.)**

6. Re-fit all of the models using the natural logarithm of rating as the outcome. Also log-transform price in all of the models. Create the same table you did in Question 1, but for the re-fitted models. **(2pts.)**

7. Use the AICc value to select the best log-transformed model. Examine whether the regression assumptions are satisfied for this model. Based on the assumptions of the regression model, which set of candidate models should be used: the un-transformed, or log-transformed models. Explain. Provide any evidence (graphical or numerical) you use in this endeavor. **(2pts.)**

8. Using the evidence from the new table you created, does the empirical evidence support more than one candidate model? Explain. **(2pts.)**

Discussion of Results

10. Based on previous literature, Snipes and Taylor hypothesized that price was an important predictor of wine quality. Using evidence from the table of candidate models you identified in Question 7, is price an important predictor of wine quality? Explain by referring to evidence about Model 0.

11. Did you select the same best model as Snipes and Taylor? Explain. Also comment on whether the uncertainty in this selection was the same as that in the Snipes and Taylor paper. **(2pts.)**

12. The model with the highest R^2 value and the highest log-likelihood value is Model 8. Explain conceptually why this model was not the one selected by referring to how AICc is computed. **(2pts.)**