SMM1 Exam 2

John Kubale

2024-11-05

Notes

- Label each answer with the appropriate question number in your R Markdown document (e.g., 1.1, 1.2, etc.).
- Clearly demonstrate your work. Where applicable, include any R code pertinent to your answer.
- Submit a single pdf file via Canvas by the deadline (930am October 8).
- You may consult any reference materials **except tools that utilize AI (e.g., Chat GPT, Github Copilot, etc.)**
- This is not a group assignment so do not consult with your classmates. Your answers should be based on your own, individual work.
- The point value for each question is given in square brackets.

1. Why should we be cautious in using tests like the Shapiro-Wilk test of normality in assessing whether residuals are normally distributed? Be sure to describe how this is related to sample size (if applicable). [5pt]

2.1. Using the `Hitters` dataset from the `ISLR2` package, fit a linear regression model with `Salary` as a function of `Years`, `Division`, and `CHits`. Save it as mod1. Refit the same model, but this time add an interaction term between `Years` and `Hits`. Create 2 scatter plots (one for each model) showing the observations and include line plots showing the regression lines of the respective model. Your outcome should be on the y axis.[15pt]

2.2. Interpret the relationship between `CHits` and `Salary` in each model. Be sure to give your interpretation in the proper units for `CHits`, `Salary`, and `Division` (if applicable). [10pt]

2.3. Use the global F test to compare the models and state which one is preferable and why (based only on the result of the test).[5pt]

3. Use PATH_MIDTERM.csv.

- Read in the data.
- The data comes from a project called "Positive Attitudes Towards Health" in 2017 that targeted persons who inject drugs in Southeast Michigan.
- The dataset includes 408 cases and the following variables: – SAMPLEID: Id of respondents – AGE: Age in years – MALE: 1. Male; 0. Female – BLACK: 1. Race black; 0. Other than black – EDUC: 1.<=Highschool (HS) Education,2. HS Education; 3. >HS Education – LIFESAT: Life satisfaction score summarized from 5 questions; Higher scores mean higher satisfaction – AGE_DIFF: Age one feels (from a question, "How old do you feel?") minus Actual age; Positive values mean feeling older than actual age; Negative values mean feeling younger than actual age

3.1. Fit a linear model that regresses AGE_DIFF on AGE, MALE, BLACK, EDUC and LIFESAT. Interpret the results from the estimated model coefficients (including the intercept) in layman's terms. Treat each as if it were the exposure of interest. [10pt]

3.2. Plot residuals against fitted values. Comment on the zero mean error assumption and the constant error variance assumption. [5pt]

3.3. Report the mean of residuals for the entire data and for the fitted values >-5 vs. <=-5. What do you conclude about the zero mean error assumption? Make sure to incorporate your observations from #3.2 in your answer. [5pt]

3.4. Report the variance of residuals for the entire data and for the fitted values >-5 vs. <=-5. Include formal testing. What do you conclude about the constant error variance assumption? Make sure to incorporate your observations from #3.2 in your answer. [10pt]

3.5. Identify top 3 influential observations through Cook's distance. What do you conclude about these observations? What makes them stand out? [5pt]

3.6. Fit the model in #3.1 without the top 3 influential observations identified in #3.5. Based on one of the goodness of fit metrics we've discussed, state whether you would remove these observations. [10pt]

3.7. Based on the model in #3.1, predict AGE_DIFF for the mean 45 year old black female with high school education and mean life satisfaction and construct

its appropriate 95% interval. What does this interval mean in layman's terms? [15pt]

3.8. Now predict AGE_DIFF for a 45 year old black female with high school education and mean life satisfaction and construct its appropriate 95% interval. How does this interval compare to the one from 3.7? [5pt]