Math 327 - Data Analysis Project 1 Final Report Checklist

Title

___ Does the title give an accurate preview of what the report is about? (i.e. Is it informative, specific and precise?)

Abstract

___ One paragraph stating the data, problem, and/or questions that are being addressed. Are the main points of the paper/poster described clearly and succinctly? If a friend asked you about your project and you had just a couple of minutes to tell them between classes, what would you say? What are the high points of the data and your results?

Introduction

The introduction should be a more detailed description of the data (compared to the abstract) and will not have any results.

___ Does the Introduction have a logical organization? *Does it move from the general to the specific?*

___ Has sufficient background been provided to understand the topic?

___ Is the final paragraph a brief description of the hypothesis, questions, and/or goals of the report?

Data Collection (as needed) and Data Characteristics

___ Data characteristics adequately described

___ Sufficient assessment of data distributions and the need, or not, to transform certain variables

1. Check the distribution of each variable
   a. If obviously right-skewed, try log transformation – add a small offset if some original values are zero
   b. Or try a square root transformation
   c. The goal is to make the distribution more symmetric, though not necessarily exactly symmetric
   d. If still obviously skewed in the same direction after transformation consider making 3-4 categories
2. Scatterplot matrix and correlations
   a. Check for simple linear associations between each pair of variables
3. Which of your predictor variables are, or should be, categorical?

Results

NOTE: You may remove the simple linear regression and the manual backwards elimination sections that were in your first draft report.

1. Fit a first-order model with all predictor variables, quantitative and categorical
    a. Which predictors are significant?
    b. Which predictors are highly correlated with each other?
    c. Any evidence of curvature in the residuals?
    d. Check for constant residual variance
    e. Do a Box-Cox analysis to see if the response variable should be transformed. If so, apply that transformation and re-run the first-order model with residual analysis
    f. Provide qualitative (directional) interpretations of the parameter estimates

2. Apply stepwise regression to the first order model
    a. Use the step function with the AIC criterion and direction=”both”; if you think that leaves too many predictors in the model, try again with the BIC criterion.
    b. You do not need to do a residual analysis of this model

3. Create centered interaction effects
    a. For the quantitative predictors in the “best” first-order model (from step 2), create new centered predictor variables by subtracting the mean for each variable
        i. Categorical predictors do not need to be centered (nor does is make sense to)
    b. Fit a model with all predictors retained by the stepwise regression in step 2 and all of their interaction effects
        i. Use this syntax to get all of the two-way interaction effects:
            1. lm (Y ~ (X1.c + X2.c + X3.c + … + Xk.c)^2)

4. Apply stepwise regression to the model with centered interaction effects from step 3
    a. Use the option direction = “both”
    b. Use the AIC criteria first, then BIC if you want fewer predictors
    c. What are the results?

5. Using the final model from Step 4
    a. Do Residual diagnostics
        i. Apply the plot() function to the fitted model object to get these 4 plots and interpret them:
            1. Residuals vs fitted values
            2. QQ plot
            3. Square root of absolute standardized residuals vs fitted values
            4. Standardized residuals vs Leverage, with Cook’s distance
        ii. Box plot of residuals
        iii. Residuals vs. time or sequence of data collection, if appropriate
        iv. Visual checks for model departures, using the residual plots above
            1. Non-constant variance

2. Curvature effects

3. Are there obvious outliers?

4. Do the residuals follow a normal distribution?

v. Optional: Added variable plots, as defined in section 10.1 – use avPlots function (in the car package) and compare to the parameter estimates

b. Plot response variable vs fitted values, add a line with intercept 0 and slope 1

c. Report and interpret the Variance Inflation Factors

d. Regarding the fourth residual plot, Standardized residuals vs. Leverage

i. Calculate the high leverage cutoff, per the book and note where that cutoff is on the Leverage axis

1. Are there any obviously high leverage values?

2. What proportion of the leverage values are above the cutoff (expect 5%)?

ii. Comment on any points with high Cook's Distance

iii. Create a scatterplot matrix using just the variables in the final model and highlight the points with high leverage and/or high Cook's distance. Comment on where these points are with respect to the rest of the data and discuss how much influence they might have on the overall fit.

e. Do you need to take any remedial measures needed based the residual and/or influence analysis?

f. Make interaction plots for all significant interaction effects

i. Or at most 3, if you have more than 3 significant interactions

g. Interpretation

i. Meaning/interpretation of regression parameters

1. Qualitative, directional statements are sufficient

2. Note: Parameter estimates of individual predictor variables that are involved in an interaction effect or a quadratic effect are not interpretable – Plot and summarize the interaction or quadradic effect from the plot, instead

ii. Make some example response predictions with confidence intervals and interpret those results.

1. Note: With many predictors, it can be difficult to specify a set of predictor values at which to predict. The syntax, predict(fit, interval='<type>'), where <type> is either confidence or prediction, will produce predicted values and interval limits for all observations in the data set. You can save that result and then subset it to get a few rows, for example, two rows each where the response is small, medium, and large

___ Overall assessment of Graphs and Figures

- Are the figures appropriate for the data being discussed?
- Are the figure legends and titles clear and concise?
- Are axis labels legible (e.g., large enough to read)?

Conclusion

___ Describe the overall conclusions of your analysis

___ Does your analysis raise any questions that can’t be answered from the current data set? If so, what are they?

Writing Quality

___ Is the paper well organized? (Paragraphs are organized in a logical manner)

___ Is each paragraph well written? (Clear topic sentence, single major point)

___ Is the paper generally well written? (Good use of language, sentence structure)