Data Analysis Project - Instructions

DD Smith

May Semester, 2023

General Instructions.

The goal of this assignment is to produce several types of analyses using R within a single report for datasets of your choice.

- Each student will use a unique version of the dataset as described in the section below "Obtaining the Data."
- An example of how the report can be formatted is available. I suggest that you use R to obtain your numerical summaries and graphics. These can be easily formatted into a word processing package such as Microsoft Word.
- The report should be typed. Do not copy-and-paste output from the Console as part of your report. This looks sloppy and will not receive any credit. Format your tables as found in the Example Solution. This can be done within a word processor such as Word, or format them using Excel and then paste them into your Word document.
- Provide a printed copy by the due date.

Obtaining your data

The packages, dataset, and variable names have been assigned for each student. The list can be found on iLearn under Content -> Empirical Data Analysis (EDA) and Simple Linear Regression (SLR) -> Data Analysis Project -> Dataset assignments. There are four parts to the project. Parts 1, 2, and 4 can use the same two variables. Part 3 requires categorical data. These variables have been assigned as well. Note: For part 4 of the assignment, you can choose which variable is the response and which is the independent variable.

Your own datasets.

If you have a dataset that you would like to use for this analysis then I am in favor of you doing so. Please check with me first so I can approve of the source and variables. Make sure that it has the minimum requirements of:

- Two continuous variables
- Two categorical variables
- Two variables that define a relationship. It does not have to be a linear relationship in order to suffice for the lab.
- 75 observations (more is even better).

Part 1 - Summary Statistics for 2 continuous variables. [20 pts]

- Create a table of statistics for the two continuous variables that are assigned or two of your choice (check with me first!). Note: If you choose your own variables, remember counts are not continuous! I cannot give you credit, for example, of constructing a `histogram` of the number of cylinders a vehicle's engine contains. Going through me first will prevent this from happening.

- Statistics to include: sample size, minimum, 1st quartile, Median, 3rd quartile, maximum, mean, and standard deviation.
- The Table description should be placed above the table (as in the example). Comment on any interesting features in the numerical summaries.
- **Rounding**: Be sure to round the mean and standard deviation two decimal places beyond the data. For example, if your data are given as integers then your mean and standard deviation should be rounded to two decimal places.

Part 2 - Histograms, density estimates, and boxplots for 2 continuous variables [20 pts]

- For the same two variables in Part 1, create a histogram and overlay a density estimate.
- Comment on modality - one, two, no mode, etc.
- Comment on symmetry or skewness. If it is skewed, then is it skewed left or skewed right?
- On boxplots, comment on any outliers. These are usually flagged with an asterisk in the boxplot.
- Comment on any gaps in the histogram.
- Although not required, you are welcome to hypothesize on what you believe is driving the shape of your graphs. Sometimes it is obvious and other times baffling.
- Be sure to label your horizontal axes. Do not include titles above your graphs.
- Include a Figure description below your graphs as in the example. You can include your discussion on modeality, skewness, etc. within the Figure comments.

Part 3 - Categorical Analysis for 2 categorical variables. [20 pts]

- Use this section to analyze the two categorical variables assgined (or, of your choice if approved). Make sure you choose variables that do not contain too many levels - say more than 7. For example, Yes and No are two levels within an opinion toward's, say, agreement with a political viewpoint. The number of flora classifications within a grid in a forest could be into the 10's or 100's - way, way too many. The variables that have been assigned are ready for analysis.
- Create a cross-tabulation of your two categorical variables. It is always helpful to include margin totals. These can aid your computations of Relative Risk and/or Odds Ratios.
- Include a mosaic plot of your two variables. Note: I have found that you may need to tweak the order of your variables in the mosaic plot so that the orientation is the same as your cross-tabulation.
- Quantify a relationship that is of interest to you using either relative risk or an odds ratio. You do not need to do both of these - you choose one or the other.

Part 4 - Simple Linear Regression Analysis [40 pts]

Graphics to include

- Scatter plot with the response on the vertical axis and independent variable on the horizontal axis. Overlay the fitted regression line. Make sure the two axes are labeled.
- Scatter plot of the Residuals Vs Fitted values. I highly recommend that you include a horizontal line beginning at zero on the vertical axis. This will help you identify lack of linearity and/or constant variance.
- qqplot of residuals. Use this to assess Normality of the residuals. Make sure you include a reference line that can be used to assess how close the residuals are matching to the Normality assumption.

- Histogram of residuals. Use this to assess Normality of the residuals. Make sure you include a density overlay as well as the Normal curve. In the example solution, the Normal curve is in red and the density estimate is in blue.

Numerical measures to include.

- Estimates of the intercept and slope parameters (these are often referred to as $\hat{\beta}_0$ and $\hat{\beta}_1$). Interpret the slope parameter **in context**.
- Compute the correlation coefficient and R-square value. Interpret R-square **in context**.
- If you choose your own variables, it is okay to use counts here. For example, in Baseball, it would acceptable to look at the number of Runs scored as a function of At-Bats. This should be correlated, right?! It is inappropriate to use the regression techniques used in this class to model 0 and 1 type responses (i.e. Survival or Presence/Absence of an attribute).

Checklist for Completeness

I hope between the instructions above and the sample projects you feel comfortable with the way your project has turned out before handing it in. I cannot pre-grade projects but for those that are meticulous and wanting more, here is a checklist.

Part 1 - Summary Statistics for Two Continuous Variables

- All required summary statistics are included.
- Table is formatted in an orderly fashion (you didn't just copy-and-paste output from the Console and call it good).
- Rounding applied where necessary. Example - No need to carry out calculations to 5 decimal places when we are dealing in thousands of dollars!
- Have a table caption beneath the table. How does the median compare to the mean? Is it close? Does the range of values surprise you?

Part 2 - Histograms, density, estimates, and boxplots.

- Did you comment on skewness? Is it correct?
- Comment on modality?
- Comment on outliers? These are usually found in the boxplots.
- Comment on gaps? These are usually found in the histograms.
- Histogram with density overlay included for both variables?
- Is the range of the horizontal axis appropriate?
- The number of bins for your histogram is subjective. There is a sweet spot though. Less than 8 is too little and more than, say 20, for a large dataset is too many.
- Boxplots included?
- Label the axes of both histograms and boxplots?

Categorical analysis

- Usually the most difficult part for some reason.
- Did you include a mosaic plot?
- Did you compute either an odds ratio or relative risk? You do not need to do both.
- Interpret the odds ratio (or, relative risk) in context?

Simple Linear Regression Analysis

- You should have four plots. See the example for the ones to include.
- Did you include your estimated regression line?
- Interpret the slope parameter in context?

- Include r-square and the correlation coefficient?
- Interpret r-square in context?
- Comment on the model assumptions? These are 1) Normality of the residuals, 2) whether a linear model is appropriate (as oppose to a curve), and 3) constant variance. There are several examples of going through these plots on iLearn.