Math 158, Spring 2018
Jo Hardin
due Feb 12, 2018
semester project (part 2)
+10 points

Simple Linear Regression

Goal

Your task for the SLR project is to apply the tools to simple linear regression in order to answer questions about the relationship between two continuous (quantitative) variables.

The report should include:

- Introduction (briefly refresh the reader's mind as to the variables of interest). Remember that you should include a reference for the original data source, and the reader should know to what population you are inferring your results.
- The hypotheses that you'll be addressing. It will probably be that the two variables are linearly related. (Positively? Negatively? Remember, R gives a two-sided p-value, but you can just as easily test that $\beta_1 > 0$ or $\beta_1 < 0$.)
- Check the assumptions for linear regression. Look at plots of explanatory vs. response and residual vs. predicted (include only what is interesting in your report). Comment on whether you think the data are linear with constant variability. If not, try transforming the data. Remember, transforming X gives a different relationship between X & Y (might make the relationship more linear); however, transforming Y changes the variability around the line (might make the standard deviation more constant *and* the relationship different.)
- Compute the test of β_1 (or other test from above) or find a CI for β_1. Remember that if you have transformed data, you should be careful about your interpretations. Your test or CI should include an interpretation in the words of your variables.
- Plot your (transformed?) variables, try to think of one as explanatory and the other as response. Give the reader a CI for both the mean and individual response at some interesting value of the explanatory variable. (That is, at some x-value that is interesting to you.) Interpret these intervals.
- Asses the fit of your model. Discuss the R^2 value and the residual plot(s). Remember that residual plots (not R^2) determine whether a linear model is appropriate.
- A Conclusion (Summarize your results. Comment on anything of interest that occurred in doing the project. Were the data approximately what you expected or did some of the results surprise you? What other questions would you like to ask about the data?)

Format

- The assignment should be turned in on paper by printing a pdf file that came from either R Markdown (Rmd) or R Sweave (Rnw).
- Do not print any warning or error messages. Only print code that is interesting and relevant to the reader (e.g., use `echo=FALSE`).
- Do not print lists of data.
- Be careful of overplotting. Use boxplots instead of scatterplots when appropriate. Use `alpha=0.1` for transparent plotting symbols.
- Remember a few things we've learned: e.g., provide the reader with residual plots which are most informative.
- Be very careful with the difference between individual prediction intervals and mean (average) intervals.
- A p-value is a probability of the *data*... the relationships you are testing are *linear*...
- Summarize any output from R; do not include any technical calculations. Use complete sentences.
- There are a series of tasks above, make sure the sections flow nicely into one another. This is a report on the data not a homework assignment. (Try to tell a good story.) You do not need to answer the questions above in any order, and certainly not with bullet points or enumeration.
- Do not be tempted to turn in everything you do. Only turn in the interesting parts of the analysis. One of the hardest parts of being a consultant is figuring out what to tell the researcher.
- Turn in the previous project with this project. (But let this assignment stand alone, that is, don't expect me to remember your variables.)
- Remember to label all graphs, email me if you are having trouble in R with labels (or really, any troubles in R!)
- The completed file should be no more than 4 pages (including graphics), and many people will turn in fewer pages.

Pairs

If you are working in pairs, the additional task is to do simultaneous inference.

- Use the three methods covered in the notes to find mean and prediction intervals for all n observations in your dataset.
- Plot 3 bands (preferably using `ggplot`) for mean intervals (i.e., the line) for the n points. (No adjustment, Bonferroni, and Working-Hotelling.)

- Plot 3 bands (preferably using `ggplot`) for prediction intervals (i.e., the points) for the n points. (No adjustment, Bonferroni, and Scheffé.)

- Write a few sentences on why it is important to adjust for multiple comparisons. Also, give an indication of which of the 3 lines (for each method) is most useful for communicating results.