Math 158
Jo Hardin
due March 26, 2018
semester project (part 3)
+20 points

Multiple Linear Regression

Goal

Your task for the MLR project is to apply the tools to multiple linear regression in order to answer questions about the relationship between multiple explanatory variables and one response variable.

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 regression model you'll be fitting. Use at least 4 explanatory variables. If a factor variable has more than one level, it will need to be made into g-1 indicator variables, where g is the number of levels. Note, R will do this automatically if the variable is non-numeric. If the variable is numeric, you can use the code `as.factor()`.
- Before running the analysis, create a pairs plot on the explanatory and response variables. Comment on any interesting relationships you see. Are any of the explanatory variables highly correlated? Is there any reason to fit a quadratic term? Or do a log transformation?
- Comment on whether or not you are using interaction variables. (You can use interactions with non-factor variables, it's just that they are slightly more difficult to interpret.) If you think interaction variables are necessary, comment on why the slope of the equation would change based on the level of one of the other variables.
- Fit your model. Include quadratic, log, or interaction terms as you see fit.
- Interpret your β coefficients to the best of your ability. Are your coefficients significant? You can perform a test of significance $H_0 : \beta_i \geq 0$ or $H_0 : \beta_i \geq c$ if you think there is a reason that the slope would increase by a certain factor greater than 0 (or that the intercept would increase by a certain factor if the variable of interest is an **indicator** variable.)
- Use the F test to compare two nested models. (That means that the larger model contains all the variables in the smaller model.) Make the smaller model have at least 2 fewer variables than the larger model. Comment on the soundness of the model. Which one would you report to your boss if you could only give her one model? (You can use Occam's Razor as a guideline.)
- Report the R^2 and Adjusted-R^2 values. Comment on the fit of the model as determined by how much variability is explained. Is this a guarantee that the model will accurately describe the population? Why or why not?

- A complete analysis of the residuals and influence points. Use plots to get an idea of which points may be contributing to the fit. Consider re-fitting a model with and without certain data that have both high leverage and large residuals. Do not include every plot, but consider including plots that give the reader an idea of your analysis.
- Use a statistical method to select variables to use in the model (e.g., manual, stepwise, forward, or backward selection procedures to create the best model for your data.) Explain your method and report which criterion(a) you used. Use residual plots, significance tests, and (some) criteria (F, Cp, R^2_a, R^2, AIC, SBC,..) to justify your model. (Your final model may have a large number of explanatory variables or just a few... pick the model you think is best!)
- Try to give an interpretation of the model that makes sense. Why do you think some variables stayed significant and others dropped out? Are any of your variables highly correlated (could one have taken the place of another?)
- Give CIs for a mean predicted value and a future predicted value for at least one combination of X's (from your final linear model).
- Summarize your report.
- As an aside / follow-up. Count the number of total hypothesis tests that you ran (including all the ones you didn't include in the report). What is that number? Call the number m. If you multiplied every single p-value in this *report* by that number, would any of your conclusions / analyses have fundamentally changed? Which ones? How? [The answer to this bullet point can be very short and not integrated into the rest of the write-up.]

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 to the reader (e.g., use `echo=FALSE`).
- Do not print lists of data or all of the model building output.
- 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. item 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*...
- If you are commenting on the significance of a variable in your text, you should report the p-value. (E.g., "... price is significantly associated with sales (p = 0.047)...")
- Residuals determine model appropriateness (i.e., need for transformations etc.), not p-values or R^2.

- **Interaction is not observable from a pairs plot** (because interaction speaks to the relationship between three variables, not a pair of variables). Pairs plots could tell you, however, whether pseudo-factor variables should be treated as continuous. To see interaction visually, you need a scatterplot with two variables and a different plot symbol for the third variable.
- 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 graded projects with this project. There is no expectation that you will re-do any of the analyses in the previous sections, but you should be learning as you go. (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 6 pages (including graphics), and many people will turn in fewer pages.

Pairs

If you are working in a pair, report and interpret the coefficient of partial determination for each of the explanatory variables in the final model (the contribution of each variable, one at a time, given the other variables are included in the model). Is there a sense that one (or a few) of the variables are substantially more important than the other variable(s)?