STAT 213: Project 1 (Linear Regression)

Colin Reimer Dawson

Due via Upload to Blackboard by Monday 4/9 at 11:59 P.M.

Goals

The goal of this project is to use multiple linear regression to create a predictive model in a domain of your choice, using any of the tools and methods you have learned about in class.

I have provided some suggestions for datasets and topics below, but I encourage you to identify your own topic and find a relevant dataset online or in an R package. You will have a much better time doing the project if you choose a topic that you actually care about investigating (and perhaps have a bit of background knowledge about)!

You will document the process and write about the results in a reproducible format (i.e., RMarkdown). Some guidelines and particular requirements are below, but you are encouraged to use any appropriate tools from the toolbox that we have developed so far.

Your topic should have a single well identified response variable that you are trying to predict, and at least half a dozen or so potential predictors to consider (though you likely won't end up using all of them in your final model!).

Here are a few examples for topics. The datasets and codebooks referred to below are available at

`http://colindawson.net/data/<filename>.txt`

1. Model mean Life Expectancies of countries (or some other variable of interest!) using other economic, geographical, and demographic information.

   (data: `CountryData.csv`, code book: `CountryData-CodeBook.txt`)

2. Use data on liberal arts colleges to "reverse-engineer" a ranking equation — i.e., try to use various variables about schools to predict their rank (or model a different response variable!).

   (data: `colleges_2014.csv`, code book: `colleges_2014-codebook.txt`)

3. Model birth weights of babies using variables that are knowable during pregnancy, to help predict when a mother and fetus are at risk of dangerously low birth weights.

   (data: `low_birth_weight.csv`, code book: `low_birth_weight_codebook.txt`).

When choosing your dataset, start with the cases and response variable that you want to be able to model (the response must be quantitative for MLR to apply), and try to find data that contains information about that response variable for a sample of cases, as well as several other potentially useful predictor variables. You do not have to gather the data yourself, although you can if you want to. If you want to do this, you should talk to me about your plans first before carrying out any data collection.

An Outline of the Modeling Process

You can make use of any metrics, tools, code, etc. that we have used in class. There is not a set path of model-fitting and assessing that you must follow, but here is a basic set of elements to include:

1. Before fitting any models, set aside a random subset of your dataset (perhaps 20% of cases) as a "test set" that you will use only to evaluate prediction error for your handful of "finalist" models. It should not be used for any parameter fitting or hypothesis tests. (I will supply a code template for doing this)

2. Use your background knowledge together with any appropriate statistical methods you know about to narrow down the set of possible models to three or four "finalists", making sure to consider regression conditions and the problem of overfitting in some way along the way. In considering models, include some with quadratic (or higher-order) polynomial terms, interaction terms, etc., even if you do not wind up keeping them in your finalists. Don't go overboard, just consider a couple that might make sense in context. **Caution**: If using automated selection methods (such as forward, backward and stepwise selection)

to narrow down your set, be careful not to remove "lower order" terms that contribute to a polynomial or interaction in the model.

3. Once you have your three or four finalists, assess multicollinearity for each finalist, and if appropriate, do something about it.

4. For each of your (possibly revised) finalists, compute a measure of prediction error on the 20% of cases that you held out. Note that you should *not* re-estimate your coefficients on the test set; use the coefficients that were fitted on the 80% of cases in the "training set" to make predictions for the test set, comparing the predictions to reality, and reporting the cross-validation correlation and the shrinkage for each finalist.

5. Once you have settled on a final model, report and interpret the coefficients, as well as reporting and interpreting confidence and prediction intervals at a couple of representative combinations of predictors. You should *not* do hypothesis tests for the coefficients at this point: since this model was selected from among many models considered, P-values are not valid (you may, however, want to use a few carefully chosen hypothesis tests during your selection process, as you see fit).

An Outline of the Writeup

You should turn in a Markdown report documenting your data exploration, the CHOOSE, FIT, ASSESS cycle (upload both the `.Rmd` source and the compiled `.pdf` output, as usual, and if the data is in a local file as opposed to a URL, the data set as a `.csv`).

The writeup should consist of:

1. A brief introduction of why your chosen topic is of interest, in general, and to you specifically.

2. Since this is a methodology project, the technical "meat" of the writeup is the "Methods and Results" section, which walks the reader through your modeling process, describing the logic behind each choice that you make (with any appropriate supporting numerical or graphical justifications). You should as much as possible suppress "raw" R output, however, using the `echo = FALSE`, `message = FALSE`, `warning = FALSE` chunk settings, instead citing numerical results in the text itself. In the interest of reproducibility, you may want to use the `` `r ``

`variable_name` syntax in your text to include the value of `variable_name` in a sentence.

3. A "Discussion" section in which you interpret your findings (both qualitatively and quantitatively) in context, and

Throughout, discuss what you are doing in text! Where possible, interpret the components of the models you are considering (particularly the "finalists"). It may not always be possible to give an intuitive interpretation of every coefficient, but try to do this when you can.

Grading

The project is graded on the following "Content" SLOs:

1. C2: Diagnosis of regression conditions
2. C3: Application of appropriate transformations as needed
3. C4: Identifying potential excessively influential points
4. D7: Interpretation of confidence and prediction intervals
5. E5: Cross-Validation Process

In addition, the project is graded on the following "holistic" criteria (all weighted equally, on a 0-10 scale):

1. Overall technical soundness: To what extent are the tools you apply appropriate to the questions you are trying to answer, and to what extent are the technical details correctly executed in the code?
2. Chain of reasoning: How well are the decisions you make throughout your analysis motivated in terms of the research question and in terms of the results obtained so far?
3. Interpretation of results in context: How well did you take the results of your analysis and connect it back to the real world context of the problem, in such a way that a reader not trained in statistics can take something away from your analysis?
4. Clarity of communication: How easy is it to follow your writeup? This criterion comprises both the quality of the writing itself, the organization of the

report (are text sections, graphs, etc. well placed for the reader to follow what is going on?), and the aesthetic quality of the report (have you suppressed unnecessary/distracting output, are your figures well labeled and visually appealing?)