STAT 213: Project 1 (Linear Regression)

Colin Reimer Dawson

Due Electronically as RMarkdown by Friday 7/23

Summary

The goal of this project is to

1. Use multiple linear regression to create predictive models in a domain of your choice, using any of the tools and methods you have learned about in class
2. Use model assessment and selection techniques to choose a final model by striking a balance between predictive fit, interpretability, and faithfulness to regression conditions
3. Document the process and write about the results in a reproducible format (i.e., RMarkdown). Some guidelines and particular requirements are given below, but you are encouraged to use any appropriate tools that we have developed so far

Topic Selection

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)!

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.

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, **split your dataset into separate training and test sets**, with 80% of the cases in the training set and 20% in the test set. You will use the test set **only** once you have chosen a final model, to get the final coefficients and to evaluate prediction error. **It should not be used for any parameter fitting or hypothesis tests.** You can use the following code template to split the data.

```
myData_testset <- myData %>%
    slice_sample(prop = 0.20)
myData_trainset <- myData %>%
    setdiff(myData_testset)
```

(replacing the names of datasets with appropriate names for your data) You would then use `myData_trainset` to do all your model fitting and assessment.

2. Use your background knowledge together with any appropriate statistical methods you know about to **write down a few initial candidate linear models**.

3. **Assess multicollinearity** for each set of predictors. If you have high levels of multicollinearity, you may want to explore removing one or more predictors from the set before fitting any models.

4. **Assess the regression conditions** for your candidate models, possibly applying transformations to one or more variables if this seems to be called for. **Note:** If you wind up transforming your **response** variable, then you need to do this for **all** of your models, otherwise the results will not be comparable across models.

5. You may want to consider some models with quadratic (or higher-order) **polynomial terms and/or interaction terms**, either because the residual plots suggest the need for polynomials, or because you have reason to think that one predictor may moderate the effect of another. **Caution**: Be careful not to remove "lower order" terms that contribute to a polynomial or interaction in the model, and be sure you've addressed multicollinearity **before** adding in polyomial or interaction terms, because by their nature, these terms are collinear with their lower order counterparts (this is not an issue in itself, but can make it hard to tell whether there's other multicollinearity going on)

6. For each of your (possibly revised) models, **perform K-fold cross-validation** on the 80% of the data you are using to fit and calibrate your models, using $K = 10$. Report the **cross-validation correlation** and the **shrinkage** for each finalist (we will do this in lab on 7/16).

7. Once you have settled on a final model (and only then):

   (a) Find the predicted values the model gives for the test set (the held out 20% of cases)

```
myData_test <- myData_test %>%
    mutate(
        yhat = fitted(myFinalModel, newdata = myData_test))
```

(b) Compute the squared correlation between these fitted values and the actual values from the test set. This will give you a relatively "pure" measure of how well your model can be expected to do at prediction in practice.

```
cor(y ~ yhat, data = myData_test)^2
```

(c) Re-fit your final model on **all** of your data

(d) Report and interpret the coefficients from this fit

(e) Report and interpret confidence and prediction intervals for the response variable at a couple of representative combinations of predictors

Note: You should **not** do hypothesis tests for the coefficients at this point: since this model was selected from among many possibilities 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, and the CHOOSE, FIT, ASSESS, USE 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** giving some context for the investigation, and explaining 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). 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.

Markdown Tip: You should as much as possible suppress "raw" code and R output by using the `echo = FALSE`, `results = 'hide'`, `message = FALSE`, and `warning = FALSE` chunk settings, instead citing numerical results in the text itself. The goal is that **your Knitted document should look as much like a research paper – as opposed to a lab report** – as possible.

In the interest of reproducibility, you may want to refer to values of variables in the text of your writeup. Outside a code chunk, if you write `` `r someVariableName` `` within a paragraph, then when you Knit, this will be replaced by the value of `someVariableName`. (the quote marks are "backquotes", which are the ones that start and end code chunks)

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

Note: Throughout your writeup, you should discuss what you are doing in the text! Do not include pages of figures with no text breaking them up (and again, don't include raw R code or output in the Knitted document at all); instead, whenever you include a figure, include at least a sentence or two describing what it shows us.

Grading

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

1. B1: Diagnosis of regression conditions and remediation of issues
2. B2: Identifying potential outliers and/or high leverage cases
3. B3: Diagnosis and remediation of multicollinearity
4. C2: Interpret confidence and prediction intervals for respones variables
5. D3: Sensibly employ model selection tools including cross-validation

In addition, the project is graded on the following "holistic" criteria (all weighted equally, on the same 0-8 scale used for the SLOs):

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?)