STAT 215: Mini-Project 2 (Logistic Regression)

Colin Reimer Dawson

Due Tuesday, December 5th by class time

This project is exactly analogous to MP1, except you will use logistic regression in place of linear regression. That is, your goal is to create a predictive model in a domain of your choice, where the prediction is a probability that a binary outcome occurs, using any of the tools and methods you have learned about in class.

You can choose your own topic, or use one of the suggestions below. As with previous projects, you will document the process and write about the results in an RMarkdown document.

Here are a few possible directions you could go (though again, you can choose any dataset and response variable you want, provided there are at least a half dozen or so possible predictors to consider):

1. Modeling the 2016 presidential election on a county-level (i.e., try to predict the probability of a Trump plurality vote using other county-level variables)
2. Adapting the Low Birth Weight problem from MP1 by dichotomizing the response variable at a medically motivated threshold
3. Build a model to predict whether a tumor is malignant using various measurements taken using imaging. A training and test set are available here:

   Training Set: `http://colindawson.net/data/cancer_train.csv`
   Test Set: `http://colindawson.net/data/cancer_test.csv`

4. Something else!

In the interest of clean presentation of results, before submitting your final document, add the following snippet in your first code chunk. This will suppress code, unwanted

messages, and text output from the final Knitted document. Note that this means that you cannot rely on raw R output to convey your results: you will need to refer to key results in the text. (If there is particular output you want to include, you can put it in an Appendix section, and set `results = 'markdown'` for those chunks.)

```
opts_chunk$set(
    echo = FALSE,
    message = FALSE,
    results = 'hide'
)
```

What to Turn In

You should turn in a Markdown report documenting your data exploration, the CHOOSE, FIT, ASSESS cycle. 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 are some things that would be good to include:

1. A brief introduction of why your chosen topic is of interest.
2. Before fitting any models, set aside a random subset of your dataset (perhaps 20% of cases) 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.
3. Consider some quadratic (or higher-order) polynomial terms, interaction terms, etc., and interpreting these, even if you do not wind up keeping them in your final model. Don't go overboard considering all sorts of polynomials and interactions: just consider a couple that might make sense in context.
4. Report any relevant fit measures that you used to decide among models.
5. Some assessment of multicollinearity in your (nearly?) final model, and if appropriate, doing something about it
6. Once you have settled on a final model, give two regression equations, one in probability form and one in logit form:

$$\hat{\pi} = \frac{exp(\beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k)}{1 + exp(\beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k)}$$

$$\text{logit}(\hat{\pi}) = \beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k$$

7. Report and interpret confidence intervals at a couple of representative values of the predictors.

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.