Problem set 5

S681

Upload your typed answers to Canvas as a PDF by 11:59 pm, Sunday 7th April. Include R code where applicable.

1. (10 points.) The data frame `salmonella` in `library(faraway)` was collected in a salmonella reverse mutagenicity assay. The predictor is the `dose` level of quinoline and the response is the numbers of revertant `colonies` of TA98 salmonella observed on each of three replicate plates.

   Build a model to predict `colonies` from `dose`. Justify your modeling choices (e.g. type of model, probability distribution, link function if any, corrections for overdispersion or lack thereof.) Plot your model, including standard errors if possible.

2. (10 points.) The dataset `wbca-training.txt` (subsetted from `wbca` from `library(faraway)`) comes from a study of breast cancer in Wisconsin. There are 481 cases of potentially cancerous tumors, which may be malignant or benign. Determining whether a tumor is really malignant is traditionally determined by an invasive surgical procedure. The purpose of this study was to determine whether a new procedure called fine needle aspiration, which draws only a small sample of tissue, could be effective in determining tumor status. The response, `Class`, is 0 if the tumor is malignant and 1 if the tumor is benign. The other variables are predictors on a 1 (normal) to 10 (abnormal) scale, as rated by a doctor. (See the help for `wbca` for more details.)

   Build **two** models to predict if a tumor is malignant or benign:

   (a) A logistic regression that includes all variables as predictors;

   (b) A more advanced model that uses variable selection, shrinkage, or both.

   Use your model to make predictions for the 200 additional observations in `wbca-test.txt`, where a tumor is predicted to be benign if the model probability is at least 0.5, and malignant otherwise. For each model, what proportion of the observations are predicted correctly?

3. (30 points.) To what extent do attitudes toward immigration explain the switching of votes of 2012 Obama supporters who became 2016 Trump supporters?

   Polls have shown that people from certain demographic groups were more likely to switch their votes than others. But what might explain why some people within a group switched, while others didn't? One theory is that attitudes toward immigration became especially salient during the 2016 campaign. Most attempts I've seen to assess this have involved fitting big, complicated logistic regression models. This is useful but not really sufficient to study this problem, so YOU will explore the data and then fit a big and complicated logistic regression model. We won't attempt to assess cause-and-effect, but we can get a sense of whether attitudes toward immigration had explanatory power over and above demographic shifts.

We'll use the 2016 Cooperative Congressional Election Study, a very large survey of a nationally representative sample of 64,600 adults. The investigators asked questions to the sample both before and after the election, although not all the pre-election respondents replied to the post-election survey. The data is available in various formats at `http://cces.gov.harvard.edu/data`. I've uploaded two files I pulled from there:

- `CCES16_Common_OUTPUT_Feb2018_VV.RData`: An R workspace containing a data frame called x, containing 64,600 observations on 563 variables.
- `CCES Guide 2016.pdf` : the codebook.

Here are the variables we'll focus on.

Technical variables:

- commonweight_vv_post: The survey weights for people who took the post-election survey.
- tookpost: Whether the respondent took the post-election survey. Limit your study to those for whom this is "Yes."

Demographic variables:

- gender: Male or Female.
- educ: Education (an ordered factor with six levels.)
- race: A factor with eight levels.
- pid7: Party identification (an ordered factor with seven levels from "Strong Democrat" to "Strong Republican.") (One notable variable we omit is income, because the way it's coded in the CCES is hard to deal with.)

Voting variables:

- CC16_326: The respondent's vote in the 2012 Presidential election. Limit your study to those who voted for Barack Obama.
- CC16_410a: The respondent's vote in the 2016 Presidential election. "NA" could mean they didn't vote or that they didn't take the post-election survey. Do not limit your study to those who voted for Donald Trump; otherwise you won't be able to give probabilities.

Immigration variables:

Respondents were asked: "What do you think the U.S. government should do about immigration? Select all that apply."

- CC16_331_1: Grant legal status to all illegal immigrants who have held jobs and paid taxes for at least 3 years, and not been convicted of any felony crimes
- CC16_331_2: Increase the number of border patrols on the U.S.-Mexican border

- CC16_331_3: Grant legal status to people who were brought to the US illegal as children, but who have graduated from a U.S. high school
- CC16_331_7: Identify and deport illegal immigrants

(Some respondents were given additional options, but we'll omit these.)

The full documentation of the variables and question wording is in the PDF.

(a) Load the data in R. You'll need to pre-process your data before you can do anything. Create a data frame called `obama` that satisfies the following:

- Only keep respondents who responded to the post-election survey.
- Only keep respondents who voted for Obama in 2012.
- Create a binary variable that indicates whether the respondent voted for Trump or not.
- Create a *quantitative* variable that measures the respondent's attitude toward immigration using the four immigration variables described above. This variable should count the number of positive responses toward immigrants across the four items, i.e. granting legal status is a positive response, deportation and increasing border patrols are negative responses. (Make sure you add things the right way around.)
- The sample sizes for some of the racial categories is small, so recode this factor to have four levels: "White", "Black", "Hispanic", and "Other." The `recode()` function in `dplyr` will be useful for this.
- You might want to create numerical versions of the ordered categorical variables (party, education), though you can also do this later.

After doing this, I got a data frame consisting of data for 23,395 individuals who voted for Obama in 2012, of whom 2,121 said they voted for Trump in 2016.

Note: R code alone is sufficient for this question : it does not have to be part of your write-up.

(b) For each of the demographic categories, it could be that immigration attitude affects all groups in the same way, or it could affect different groups in different ways. (For example, if you compare white and black voters with the same attitudes toward immigration, it might be that these attitudes sway one group more than the other.) Using weighted logistic regression or otherwise, fit models using immigration attitudes as a predictor along with each demographic variable in turn (e.g. immigration and race, immigration and party, etc.) In each case, consider whether you need an interaction. With which of the demographic variables does immigration attitude interact with? Carefully explain the meaning of any large interaction effects you find.

Note: For this question, I'd advise you to use immigration attitude as a quantitative predictor. You might want to recode some of the demographic variables as quantitative variables as well.

(c) Fit TWO weighted logistic regression models to give the probability of an 2012 Obama voter switching to Trump in 2016: one without immigration attitude as a predictor, and one with immigration attitude as a predictor. Include interactions as necessary. State or display the coefficients of your models. Display the model probabilities for selected demographic groups. Compare the results of your models. Does including immigration attitudes make a substantive difference? Does it matter more for some demographic groups than others?