AA502 – Logistic Regression

Homework 1

Matthew Austin

Background

You work for a major bank in the retail department. You are in charge of predicting which customers will buy a variable rate annuity insurance product.

A variable annuity is a contract between you and an insurance company/bank under which the insurer agrees to make periodic payments to you, beginning immediately or at some future date. You purchase a variable annuity contract by making either a single purchase payment or a series of purchase payments.

A variable annuity offers a range of investment options. The value of your investment as a variable annuity owner will vary depending on the performance of the investment options you choose. The investment options for a variable annuity are typically mutual funds that invest in stocks, bonds, money market instruments, or some combination of the three. If you are interested in more information, see:

https://www.sec.gov/reportspubs/investor-publications/investorpubsvaranntyhtm.html

Data

We have a sample of 10,619 customers who have been offered the product in the dataset under the variable *INS*, which takes a value of 1 if they bought and 0 if not. The data was split into training (**Insurance_T**) and validation (**Insurance_V**) sets. There are 47 other variables describing the customers' attributes **before** they were offered the new insurance product. There is a mix of categorical and continuous predictors with labels that should serve as a data dictionary. Except for *Branch of Bank*, consider anything with more than 10 distinct values as continuous.

If you are using `R`, you can open `.sas7bdat` files by installing the `haven` package and using the `read_sas()` function.

Assignment

Using only the training dataset, develop a report about your initial analysis of different factors that influence whether or not the customer purchased the insurance product. **For any hypothesis testing**, pick a value of α between 0.2 and 0.001, state it, and use it for this entire assignment. If you want to just use $\alpha = 0.05$, go for it. Make sure that the report addresses the following issues:

- **Before you use the response for anything**, look at the distributions of all of your predictors. Are there any with a large proportion of missing values? Ignore these variables. Are there any that have a very narrow distribution (e.g., almost entirely 0s or entirely 1s)? Consider ignoring these or transforming/combining them in some sensible way if you can think of one. Feel free to examine any crosstabulation tables between some sets of two predictors as well to see if anything jumps out.
- Which of your predictors (continuous and categorical) do you think might be important to your problem? Why? This can be based on subject knowledge, literature, test results, or whatever you feel might be important. Fit a logistic regression model with these variables. (If you have no idea and are only going off test results to decide what goes into your model, that's fine.) Give an interpretation (including the confidence interval) of the odds ratio for the predictor with the largest estimate (in magnitude).

- Think of an interesting comparison involving multiple predictors. Compute and interpret the odds ratio for these two subjects.
- The dataset has several variables that might have redundant information (e.g., money market account and money market balance) or might be indicative of the same underlying phenomenon (e.g., teller visits and phone number banking could represent something like actual human contact with the bank). Is anything like this in your model? If so, why do you feel like you need to keep both? (There's no right or wrong answer.)
- How many of your predictors have missing values? Earlier, you ignored predictors with a large number of missing values, which is a perfectly valid thing to do—the idea being that they might be likely to be missing in the future as well and thus may not be useful for the application of your model.[1] How many observations have missing values? You should keep in mind and make a note of how much of your sample is being discarded when we only do a complete case analysis. Dealing with missing values is challenging to do accurately and beyond the scope of this class, so for now we won't worry about it aside from noting it here.[2]

Questions/topics to know

- What is the relationship between odds and probability? Specifically, how are they related in a logistic regression model?
- What is the response in a logistic regression model?
- How to compute and interpret odds and odds ratios from a logistic regression model
- How estimates for logistic regression are obtained (i.e., what method is used)
- What linear separation is, how to detect it, and how to remedy it

[1] You don't have to do this, but if you'd like to see clusters of the predictors with the same missing observations, get the `Hmisc` package in `R`. Put the variables from your model into a dataset and do `plot(naclus(yournewdataset))`. In `SAS`, you can do something similar with `proc varclus`, but it's taken me more than 30 seconds to figure out how, so I gave up. If you really want help with it, come see me. I also think that a table with this information is in the output given by `proc mi`.

[2] This is completely optional and I will never ask you about it, but if you are interested in how this is done, it's called `multiple imputation`. See `proc mi` in `SAS` or the `mice` package in `R` for more information and how to implement it.