Lab Assignment 12

I. Objectives: To learn the key components of logistic regression. We will

estimate a logistic regression

interpret coefficients (as they affect odds and probability)

make predictions

II. Datasets: “lending\_club\_cleaned.csv”

III. Packages: lmtest; separationplot

IV. Preparation

1) Open RStudio by double-clicking the icon or selecting RStudio from the Windows Start menu.

2) Clear any data in memory by typing rm(list=ls())

3) Download R script “POLS 6480 Lab 12.R” and place it in your working directory.

4) Open the R script by typing *Ctrl*+*O* or by clicking on File in the upper-left corner, using the dropdown menu, and navigating to the script in your working directory.

V. Instructions for Lab 11

In this lab we will apply logistic regression. Let’s begin with the Lending Club data. This examines whether individuals repaid loan. The main Dv is the good loan/bad loan indicator named good. Explanatory variables include an average credit score,  “fico”, a categorical variable which indicates the purpose the of the loan (“purpose”), a variable “income” which equals income if either the source of income or income itself is verified, otherwise it equals NA, a debt-to-income variable “dti”, and loan amount (“loan\_amount”).

1. Load the dataset by typing the following line:

* loan <- read.csv("https://www.dropbox.com/s/89g1yyhwpcqwjn9/lending\_club\_cleaned.csv?raw=1")

2. Look at the data

* Head(loan)

3. We will use function glm() (part of base R) to estimate a logit model. GLM stands for generalized linear model. Its first argument is a formula in the Y ~ X1 + X2 format where Y is the variable we try to predict (the dependent variable) and the X1, X1 etc are the predictor (independent) variables. The second argument specifies the data and in the third argument we specify we want a “binomial” model which tells gml() to fit a logistic function to the data.

> logit1 <- glm(good ~ fico, data = loan, family = "binomial")

> summary(logit1)

Note that good is a factor variable with two levels (“bad” and “good”). gml() will treat the first level as 0 and the second level as 1. Thus, in our case we are estimating the probability of “good”.

4. The output includes summary of deviance residuals and AIC which are both measures of fit. It also includes coefficients associated with each independent variable and the intercept. The coefficient on fico is positive and statistically significant. It tells us that a one point increase in fico score raises the log of odds ratio by 0.01. Log of odds is rather hard to interpret, therefore, we often take the exponential of the coefficients.

* exp(coef(logit1))

The exponential of the slope coefficient tells us the factor by which odds increase if the independent variable increases by one. So, if a fico score increases by 1 point, the odds ratio of loan being good increases by factor of 1.012 which means odds increase 1.2 percent.

5. Run a second logit model that include “dti” in the model (call this model “logit2”). Interpret the odds of “dti.”

It is important to keep in mind that once you include more than one independent variable, the interpretation of the coefficient on an independent variables is the effect of that independent variable holding all other independent variables constant.

6. While the interpretation in terms of odds is relatively easy, sometimes interpretation in terms of probability is desired. However, calculating how a change in fico score affects the probability of a good loan is a bit complicated. The relationship is non-linear and therefore the effect depends on the value of fico score (and other variables if they are included). Therefore, we need to specify the levels of fico we want to see the effect of. For example, suppose we wanted to know the effect of fico going from 700 to 750 on the probability of loan being good. We could create a test data set with these two values. We will then feed these values to our logit1 model using the function predict(). As option we specify type="response"which tells predict to return probabilities.

* test <- data.frame(fico=c(700,750))
* test$pred <- predict(logit1,test, type="response")
* test

The fico score going from 700 to 750 increases the probability of a good loan by seven percentage points.

7. What is the effect of fico going from 750-800 in model “logit1”?

8. Let’s see how gml() handles categorical/factor variables. For example, let’s include “purpose” into a model, with fico and loan\_amnt.

> logit3 <- glm(good ~ fico + loan\_amnt + purpose, data = loan, family = "binomial")

> summary(logit3)

Note that purpose is a factor with eight levels, but glm() is showing us only seven coefficients associated with purpose. The interpretation of the coefficient on each of the included dummies is the effect of that category relative to the ‘left out’ category. In our case the category that is left out is debt\_consolidation so the coefficients on eductional, home\_improvement etc are relative to debt\_consolidation loans. Let’s take an exponent of the coefficients so we can interpret the magnitude of these coefficients.

* round(exp(coef(logit3)),3)

The coefficients show that the odds of a loan being good are lower for educational loans relative to debt consolidation loans by a factor of 0.57, or about 43% lower. As another example, let’s take vacation and wedding loans. These have shockingly 12% higher odds of being good than debt consolidation loans.

9. Suppose we wanted to include income as an explanatory variable. Income is verified in only about half of the cases and therefore about half of the values for this variable are missing.

* logit4 <- glm(good ~ fico + loan\_amnt + income + purpose, data = loan, family = "binomial")
* summary(logit4)

Logit omits observations/rows with missing values for any of the independent variables. Thus, the model above is estimated on only half of the observations in our data set. Perhaps most importantly, this model may not be used for predicting the status of loans that have missing income.

10. Let’s compare models 1 and 3 using a log-likelihood test.

* lrtest(logit1,logit3)

11. While model “logit3” is a better fit than “logit1” is it a good model? One way to evaluate this is see if the predicted probabilities from the model are associated with outcomes. Separation plots sort predicted probabilities from a logit, and then plot vertical red lines for DV’s coded as 1 (in this case in good loans). The ideal model will have red lines associated with higher levels of predicted probabilities.

* separationplot(pred=logit3$fitted.values, actual=logit3$y, type="rect",

line=TRUE, show.expected=TRUE, width = 9, height = 1.2, heading="Probability of Y", zerosfirst)

How does the model look? Not great.