Binary Logistic Regression
Instructor: Todd BenDor

Lab Preamble

For this lab you will use the 'bustrips_logistic.csv' file available under Lab #2.

Like lab #1, we will be analyzing data about bus trips taken in 2006, as reported by numerous transit agencies across the U.S. We wish to predict whether the reported total passenger miles taken by bus in 2006 (*passmile*) is based on the average trip length (*trplen*), the service area in square miles (*svmile*), the service area population (*spop*), and the urbanized area in square miles (*uzamile*). This involves fitting a statistical model to the data, but now we are using a binary dependent variable, *passmile*.

Here, you will perform several separate regression analyses, which should analyze the relationship between *passmile* (high vs. low passenger miles) and various attributes of the transit service area (independent vars).

Data dictionary:

passmile:
1 = high number of passenger miles (>median of 3,422,079 miles)
0 = low number of passenger miles (<= 3,422,079 miles)

trplen:
1 = long average trip length (>4 miles)
0 = shorter average trip length (<= 4 miles)

For each regression:

1. Determine the regression equation to estimate your data. What is y-hat?
2. Compute and interpret variable coefficients in terms of how the change in the independent variable impacts the odds of a high number of passenger miles by bus, being aware of the type of independent variable (ex: continuous versus dummy).
3. Does this observed association (or lack thereof) make sense under your hypothesized understanding of the relationship?
4. How does this coefficient differ from that obtained in linear regression?
5. What statistic do you get when you exponentiate the coefficient? (`logit depvar indvar, or`)
6. Interpret the pseudoR^2 value. Is there are better measure of the goodness of fit of a logistic regression model?
7. Graph the data and the predicted probabilities of high vs. low passenger miles based on each independent variable (use regression equation and given values of the independent variable to predict Y values).

Instructions

To obtain the regression table and results, start by plotting independent variables vs. the dependent variable. You can do this by going to "Graphics → Scatterplot Matrix" in STATA or by typing "graph matrix [vars]" in the STATA command window. What do you see? Could you draw a straight line through this relationship as we did in linear regression?

To determine if there is multicollinearity in our data, run a linear regression with **passmile** as the dependent variable, and the others as independent variables.

```
.regress passmile trplen uzamile svmile spop
.estat vif
```

[NOTE: Alternatively, a more advanced program called **lmcol** has been written as a 'plugin' to STATA. To install this plugin, "**findit lmcol**" and scroll to the **'lmcol'** program in the window that pops up, click it, and then click **"click here to install."**]

Does there appear to be multicollinearity in the independent variables above? How do we measure this?

Now, to run an actual regression, click on "Statistics → Binary Outcomes → Logistic Regression." Here, you will find a dialog box asking for independent and dependent variable inputs. You should run regressions on *passmile* using each of the independent variables above separately, as well as together.

What do the regressions tell you (in terms of the coefficients)? How well do each of the models fit? What does each of the coefficients do to the logit of high passenger miles? What about the odds?

What variables are significant? How could you have found this without using STATA?

Which is the best model?

Re-run each of your logit regressions (**Each command is Stata is saved in the "Review" window)**

> After **each new run**, predict the value of each observation having high passenger miles given your model using the command: **predict p_x** (use p1, p2, p3 for each model run)
>
> You can now open up the data browser and see how your predicted values fair against your actual values. See any systematic error? How did your model do?

You can look at this a bit more rigorously by running a post-estimation diagnostic through: "Statistics → Binary Outcomes → Postestimation →Classification Statistics after logistic/logit/probit."

What happens if you run the same regression using OLS? How well do the two methods compare?