Homework: Binary Response Models

Exercise 1: MLE with Logit

tab3s.dta is a STATA dataset that contains information on all members of the house in 1990. The dependent variable is a dummy RETURNED - whether the members of the house were returned in 1992. The independent variables are all labeled in obvious ways. To see what the variables are, type DESC in STATA.

1. Run a logit of the dependent variable on the following two independent variables (Age, Marginal)

2. Now using the log-likelihood for a logit model (assuming independent observations), use STATA to manually estimate the same model via MLE. The code will look something like that shown below.

```
clear;
capture program drop mllogit;
program mllogit;
version 10.0;
args lnf xb;
quietly replace `lnf'=ln(1/(1+exp(-(`xb')))) if $ML_y1==1;
quietly replace `lnf'=ln(1-(1/(1+exp(-(`xb'))))) if $ML_y1==0;
end;
use "c:tab3s.dta", clear;
logit returned marginal age;
ml model lf mllogit (returned = marginal age);
ml maximize;
```

3. Put the results in a table like the one shown below and indicate what is the same, what is different, and why.

Table 1: The Determinants of Presidential Vote Choice in 1992

Dependent Variable: ClintonVote (1 if Clinton, 0 if Bush)

Regressor	Logit (STATA)	Logit (MLE)
Age		
Marginal		
Constant		
Log likelihood		
Observations		

* $p < 0.10$; ** $p < 0.05$; *** $p < 0.01$ (two-tailed)
(Standard errors are given in parentheses)

4. Why is there a second (σ:) entry in the OLS ML code from last week but not in the logit ML code?

5. In the class notes, I wrote that the gradient vector from the logit model is:

$$\mathrm{G} = \frac{\partial \ln \mathcal{L}}{\partial \beta} = \sum_{i=1}^{n} (y_i - \Lambda(x_i \beta)) x_i \tag{1}$$

   Illustrate how I got this i.e. show the steps involved in this calculation.

6. In the class notes, I wrote that the Hessian matrix from the logit model is:

$$\mathrm{H} = \frac{\partial^2 \ln \mathcal{L}}{\partial \beta \partial \beta} \quad = \quad -\sum_{i=1}^{n} \Lambda(x_i \beta)[1 - \Lambda(x_i \beta)] x_i x_i' \tag{2}$$

   Illustrate how I got this i.e. show the steps involved in this calculation.

7. What can you infer about how the independent variables affect the probability of an incumbent being releected just by looking at the coefficients?

Exercise 2: Interpretation

Use nes92.dta, which is a small subset of the 1992 National Election Study (cleaned up). The file should be reasonably self-documenting but you can also look at the codebook (nes92_codebook.pdf). For this exercise, use only Clinton or Bush voters i.e. drop the Perot voters.

1. Estimate the following probit model using:

$$\begin{aligned} \text{Clinton}_i^* = \beta_0 &+ \beta_1 \text{DistancetoClinton} + \beta_2 \text{DistancetoBush} + \beta_3 \text{Economyworse} \\ &+ \beta_4 \text{Education} + \beta_5 \text{Union} + \beta_6 \text{Income} + \beta_7 \text{Black} + \epsilon \end{aligned} \tag{3}$$

where DISTANCETOCLINTON and DISTANCETOBUSH measure the absolute distance from the respondent to Clinton and Bush respectively on the left-right 7 point scale, ECONOMYWORSE measures whether the respondent thinks that the economy is worse compared to a year ago on a five point scale, EDUCATION measures the respondent's years of education (0-17), UNION indicates whether the respondent is a union member, INCOME measures the family income of the respondent in thousands of dollars, and BLACK indicates whether the respondent is black.

Put the results in a table like the one below.

Table 2: The Determinants of Presidential Vote Choice in 1992

Dependent Variable: ClintonVote (1 if Clinton, 0 if Bush)

Regressor	Probit	Logit	Scobit	Heteroscedastic Probit
DistancetoClinton				
DistancetoBush				
EconomyWorse				
Education				
Union				
Income				
Black				
Constant				
α				
Log likelihood				
Observations				
Dependent Variable: $\ln\sigma^2$				
Education				
Black				

* $p < 0.10$; ** $p < 0.05$; *** $p < 0.01$ (two-tailed)
(Standard errors are given in parentheses)

2. By hand (using STATA), calculate the following. Always set the seed to 10101 so that we get the same answers.

   - The probability of voting for Clinton when the respondent is white, is a non-union member, thinks that the economy is in the same state as last year, has 12 years of education, has $20,000 in income, and has the mean distances on the left-right scale to Clinton and Bush. Remember to provide 95% confidence intervals.
   - The probability of voting for Clinton when the respondent is white, is a non-union member, thinks that the economy is in the same state as last year, has 12 years of education, **has $60,000 in income,** and has the mean distances on the left-right scale to Clinton and Bush. Remember to provide 95% confidence intervals.
   - The change in these probabilities and 95% confidence intervals.

3. Check what you did with CLARIFY.

4. Now redo question 1 with logit and put results in the same table as before. Compare the results. How do they differ, how are they the same, and why?

5. Repeat questions 2 and 3 for logit.

6. Interpret the logit results in terms of odds ratios.

7. Using the results from the probit model, test the hypothesis that the coefficient on DISTANCETOCLINTON is zero using (i) a Wald test, (ii) a z-test, and (iii) a likelihood-ratio test. Do the likelihood ratio test first by hand (that is, by computing twice the difference in the log-likelihoods etc.) and then by using the LRTEST command in STATA. For the Wald test, the easiest thing to do is TESTPARM VARNAME. What differs in the results from these tests and what is the same?

8. Still using the results from the probit model, test that the effect of INCOME and BLACK are jointly zero using (i) a Wald test and (ii) a likelihood ratio test.

9. You are now going to compute various goodness of fit measures. For each measure, make sure that you only calculate the measures for the sample of observations actually used in the probit estimation (362 observations)

   - Calculate the percent in the modal category (PMC)
   - Calculate the percent correctly predicted (PCP) manually
   - Calculate the percent correctly predicted (PCP) using STATA:

     ```
     probit clintonvote distancetoclinton distancetobush economyworse
                 education union income black, tab;
     ```

   - Use these results to calculate the percentage reduction in error (PRE).
   - Now calculate the expected percent correctly predicted (ePCP) - no need to calculate confidence intervals around this right now.
   - Now draw a ROC curve. Pick a point on the ROC curve and tell me what it means. Use the LROC and ROCTAB commands.

- Compare the fit of the model where you include the covariates BLACK and INCOME to the model where you only include INCOME using ROC curves. Which model fits better and how do you know?

10. Estimate a probit model with the following covariates: DISTANCETOCLINTON DISTANCETOBUSH ECONOMYWORSE X15 UNION INCOME BLACK. Then estimate the same model but use X20 instead of X15. What happens and why? Graph the dependent variable against EDUCATION, then against X15, and then against X20. What is going on here?

11. I want you to take a look at Chris Zorn's (2005) article. On page 159 he describes how he came up with data to run simulated logistic regressions. I want you to follow his description and reproduce the first three panels in Figure 1 i.e. for $\alpha = 1$, 0.5, and 0.1. No need to put the parameter estimates in the top left corner of each panel.

12. Now estimate the same *full* model as before but using scobit. Put the results in the same table as before. Interpret the critical parameter α. You should also say whether we need to use logit or scobit and how you know this. What else changes?

13. Now estimate the following heteroskedastic probit model

$$\begin{aligned} \text{Clinton}_i^* = \beta_0 &+ \beta_1 \text{DistancetoClinton} + \beta_2 \text{DistancetoBush} + \beta_3 \text{Economyworse} \\ &+ \beta_4 \text{Union} + \beta_5 \text{Income} + \beta_6 \text{Black} + \epsilon_i \end{aligned} \tag{4}$$

where $\epsilon_i \sim N(0, \sigma^2)$ and $\sigma_i = e^{\gamma_1 \text{Education} + \gamma_2 \text{Black}}$.

- Put the results in the same table as before.
- Is this heteroskedastic probit preferred to the standard probit model? How do you know?
- What additional hypotheses are being tested by this model?
- How would you interpret the coefficient on EDUCATION?
- Calculate the probability of voting for Clinton when the respondent is white, is a non-union member, thinks that the economy is in the same state as last year, has 12 years of education, has $20,000 in income, and has the mean distances on the left-right scale to Clinton and Bush. Remember to provide 95% confidence intervals.
- Calculate the probability of voting for Clinton when the respondent is white, is a non-union member, thinks that the economy is in the same state as last year, has 12 years of education, **has $60,000 in income,** and has the mean distances on the left-right scale to Clinton and Bush. Remember to provide 95% confidence intervals.
- Calculate the change in these probabilities and 95% confidence intervals.
- Write down the equation for the marginal effect of UNION on the probability of voting for Clinton.
- Write down the equation for the marginal effect of BLACK on the probability of voting for Clinton.

Exercise 3: Rare Events Logit

Use orum.dta. The CODEBOOK command will give you information about the variables. This is data from Oneal and Russett (*ISQ*, 1997) on international disputes in all politically relevant dyads ($N = 21,844$).

1. Determine the percentage of 1s and 0s in the population
2. Generate a variable that is log of the capability ratio

   ```
   gen logcapratio = log(capratio)
   ```

3. Run a logit model of DISPUTE on DEM, GROWTH, ALLIES, CONTIG, LOGCAPRATIO TRADE. Put the results in column 1 of Table 3.

Table 3: Explaining International Disputes(Dependent Variable: Militarized Dispute)

	Logit	Logit Case Control Sample	Relogit
Democracy			
Growth			
Allies			
Contiguity			
ln(CapabilityRatio)			
Trade			
Prefail			
Constant			
Observations			
Log-Likelihood			

* $p < 0.10$; ** $p < 0.05$; *** $p < 0.01$ (two-tailed)

4. Now get predicted probabilities from this model.

   ```
   predict dumbprobs
   ```

5. Next, we'll select out all the 1s as well as a 10% random sample of the 0s.

   ```
   gen random=.

   replace random=invnorm(uniform()) if dispute==0

   gen select=1 if dispute==1
   ```

```
sum random, detail

replace select=1 if random>1.266858 & random~=.

tab dispute if select==1
```

6. Run the same logit model as before on the case control sample. Put the results in column 2 of Table 3. What's different to the previous model? What's the same?

7. You now need to use the RELOGIT command to correct for the biases that King and Zeng talk about. Recall that there are two ways of doing this. One uses weights and the other uses the prior correction. Use the prior correction first and then the weighted version:

```
relogit dispute dem growth allies contig logcapratio trade if
select==1, pc(tau)

relogit dispute dem growth allies contig logcapratio trade if
select==1, wc(tau)
```

   You'll have to substitute in the appropriate value for τ. Are there any differences between these two methods?

8. Use the SETX and RELOGITQ commands to calculate predicted probabilities for the following scenarios.

   - Scenario 1: dem=1, growth=mean, allies=0, contig=0, logcapratio=mean, trade=mean
   - Scenario 2: dem=1, growth=mean, allies=0, contig=1, logcapratio=mean, trade=mean

9. Now calculate the change in predicted probability between these two scenarios using RELOGITQ.