

**Advanced Regression
Problem Set 2
Due: October 5, 2009**

I. Voter turnout among registered voters has declined dramatically among U.S. voters since the late-1960s. While a large fraction of registered voters still turn out to vote—over 80 percent—there is wide variation in voter turnout rates across jurisdictions and states. To some extent, these differences are due to the presence of contested elections; in cases where a race is close, voter turnout is likely to be higher.

However, there are other reasons that may explain cross-state variation in voter turnout. Consider, for example, states' policies regarding the ease with which voters can make their selection. Some states open polls early on Election Day (12 states), and some keep them open late (19 states). Other states mail out sample ballots for an election (7 states) and others mail information about polling places to registered voters (7 states). Do these policy differences have an impact on voter turnout?¹

Consider the dataset `voterturnout2000.dta`. These data are from the Census Bureau's Current Population Survey (CPS), and contain detailed information on individual's background characteristics, as well as voter behavior. The data were collected within two weeks of the November's 2000 Presidential election, with interviewers inquiring about voter registration and actual participation. For this problem, sample data have been limited to registered voters.

For the analyses in this problem, you must weight the data to reflect the properties of the sampling method used and the universe (or population) of individuals from which they were drawn. To weight your analyses, use the variable `weight2`. For example, to obtain a weighted mean, you would type:

```
sum didvote [aw = weight2]
```

Note that the appropriate weight variable (provided by the CPS) is denoted in brackets and is preceded by `aw =`. The `aw` stands for "analytic weight." To run a weighted probit regression, you would perform your analysis as follows:

```
probit didvote black hispanic [aw = weight2], robust
```

Notice that `robust` is listed after the bracketed weight. This tells Stata to estimate the regression model with standard errors that are robust to heteroskedasticity.

a) Begin by examining the dataset. What proportion voted in the 2000 election? What are the voter turnout rates for states with and without early opening hours? Late opening hours? Mail poll information? Mail sample ballots? Are the differences statistically significant? Conduct the appropriate empirical tests.

¹ This problem is based on the work of Raymond Wolfinger of the University of California, Berkeley.

b) Next, estimate three models for voter turnout that only control for policy differences in election laws. For the first model, estimate a linear probability model (LPM). For the second model, estimate a probit regression that displays all coefficients as marginal effects. Finally, estimate a logistic regression model. Interpret the coefficient on each policy variable in the probit model. How close are your results for the LPM and probit models?

c) For a moment, let's concentrate on the probit model. Add controls for individual's demographic characteristics (i.e., race, gender, age, and marital status, educational attainment, metropolitan status, and employment status). Conduct a likelihood ratio test to see whether these controls add significant explanatory power to your model. What is the hypothesis being tested? Interpret your results. What do you find?²

d) Using the full model estimated in question (c), assess its goodness-of-fit. In particular, how well does your model's prediction of voting match what is observed in the data? Use the `lstat` command and describe your results. Also, use the `fitstat` command to check the AIC. How do you interpret this fit statistic?³

e) Now, let's return to the logit model. Estimate a logit regression of voting status on the policy variables, race, gender, age, and marital status, educational attainment, metropolitan status, and employment status. Be sure to weight your regression and use robust standard errors. Once you have these results, let's try some interpretations of the coefficients using odds ratios. To get these from the logit model, type this command:

```
listcoef early late poll sampbal black hispanic asian amin
female age marr o_marr hs somecoll baplus metro unemploy,
factor help
```

Interpret the odds ratio coefficient on the variables `late` and `asian` (e^b). To convince yourself that Stata has calculated these odds ratios correctly, go back to the original logit regression and transform the logit coefficients into odds ratios. What do you find?

f) Finally, let's make predictions about voter turnout based on the characteristics of individuals in this sample. To do this, we will make use of the `prvalue` command, which produces predicted probabilities of the outcome variable based on values we specify for the independent variables. Begin

²To perform a likelihood ratio test, you will need to do the following. First, estimate the full model, without robust standard errors:

```
probit didvote female hispanic black asian native
```

Then, after your output is displayed, enter the following command:

```
lrtest, saving(0)
```

Next, estimate the restricted model:

```
logit didvote female
```

Finally, after this output is displayed, enter `lrtest`. Stata will produce an estimate of the likelihood test statistic, which is distributed as a chi-squared with the degrees of freedom equal to the number of restrictions.

³To use the `fitstat` command, you will need to download a suite of Stata post-estimation commands, called SPOST. Type `net search spost` in the command box, find the relevant program from the list, and follow the instructions to download the program to your computer. Once SPOST is installed, you will be able to use a number of new programs that allow you to summarize results from logit, probit, and other qualitative dependent variable models.

by re-estimating the logit model in (e). Now, generate a predicted probability that a white, married female with a high school degree voted in the 2000 election (holding all other variables at the mean). What do you find? Generate the probability that a white, married female with a BA+ voted in the 2000 election. What changed and does it make sense? How is Stata calculating these predicted probabilities? Provide the intuition for this through all necessary formulas (but do not worry about making the actual calculations). Hint: You'll need to know what the logit model looks like.

2. This brief exercise will give you some practice working with a multinomial logit model (MNL). Use the dataset called **nomocc2.dta** for all analyses. These data contain information on job occupations for a sample of 337 individuals. The job occupation variable is called `occ`, and we will use this as the dependent variable in the regressions. Our independent variables will be race, years of education, and years of experience.

a) Begin by describing the data and producing some simple descriptive statistics as a verification. In particular, produce a tabulation of `occ`. How many categories are included in this variable? Does the variable take a nominal or ordinal scale?

b) Now, let's run a multinomial logit regression of occupational category on race, education, and experience. The command you will want to enter is the following:

```
mlogit occ white ed exper, basecategory(5)
```

Notice we set `basecategory(5)`. What does this imply about the interpretation of the results? Stata's default is to set as the base category as the one with the greatest number of observations, but you can set it to be anything. Alternatively, you can use Stata's `listcoef` command to get estimates for all combinations of base/comparison categories. Try this on your own.

c) As noted in the Stata textbook, an MNL is essentially identical to simultaneously estimating all combinations of binary logit models. Therefore, we should be able to determine the impact of each independent variable for *any* combination of base/comparison categories by working with the results from just a single MNL. To do this, we need to *subtract* the coefficients associated with a variable across two distinct categories. Let's try this. First, estimate:

```
mlogit occ white ed exper, basecategory(5)
```

Suppose we want to know the impact of race on choosing the menial occupation, but with blue collar as the base category. We could estimate another MNL with `basecategory(2)`. But we could simply subtract the white logit coefficient in the current model's blue collar category from the white coefficient in the menial category. Try this. What do you find? Now, to convince yourself that this is in fact the correct coefficient, estimate this model:

```
mlogit occ white ed exper, basecategory(2)
```

What is the coefficient on white in the menial category? Does this match your calculation?

d) Let's experiment with interpreting the coefficients in this model, beginning with a transformation of the logit coefficients to marginal effects. To do this, you will need to enter this command after estimating the model:

```
mfx compute, predict(outcome(1))
```

Note that the `mfx compute` command allows you specify marginal effects one category at a time. Interpret the marginal effect on each variable (race, education, and experience) for the first outcome (menial). Now, let's practice some interpretations using odds ratios. To produce an odds ratio transformation of the logit coefficients, use the `listcoef` command. Again, interpret the odds ratio on each variable for the first outcome.

e) Finally, let's check whether our model passes the test for the assumption of independence of irrelevant alternatives (IIA). What does this assumption specify? Use the Hausman test. What is the null hypothesis being tested? What do you find?