

## Soc 63993, Advanced Social Statistics II

### Homework No. 9

### Logistic Regression

I. As we saw in the class handout on the PSI teaching example, 8 of the 14 students who were in PSI got A's compared to only 3 of the 18 students who were in a conventional classroom. Verify that those numbers are consistent with the following results that we get when GRADE is (logistically) regressed on PSI only. Recall that GRADE = 1 if grade is an A, 0 otherwise, PSI = 1 if in psi, 0 otherwise. [HINT: Compute the log odds for those in psi and those not in psi, and then take it from there.]

```
. use http://www.nd.edu/~rwilliam/xsoc63993/statafiles/logist.dta
. logit grade psi, nolog
```

```
Logistic regression                Number of obs   =           32
                                LR chi2(1)         =           5.84
                                Prob > chi2          =          0.0156
Log likelihood = -17.670815        Pseudo R2       =          0.1418
```

grade	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]
psi	1.89712	.831665	2.28	0.023	.2670865 3.527153
_cons	-1.609438	.6324555	-2.54	0.011	-2.849028 -.3698478

II. Download *lrb.dta* from the course web page. We use a sample of Southern Baptists from the GSS in this homework. General Social Surveys from 1973 to 1991 are used to make one big sample. All married Southern Baptists between the ages of 20 to 25 (all 61 of them!) are in the data file. The dependent variable is `happymar`, respondent's marital happiness (1 = Very Happy, 0 = Otherwise). `church`, Church attendance (1 = Often attends, 0 = other), `female` (1 = female, 0 = male), and `educ`, Years of education, are the DVs.

Use Stata to run the logistic regression of `happymar` on `church`, `female` and `educ`. Then answer the following questions.

1. What assumptions of OLS would be violated if OLS was used to approach this problem?
2. Interpret the logistic regression coefficients. What do the parameters tell you about the determinants of marital happiness? What can you say about the size and magnitude of effects?
3. Determine the log odds, odds and probability of marital happiness for:
  - (a) a male with 8 years of education who is not a regular churchgoer
  - (b) a male with 8 years of education who is a regular churchgoer
  - (c) a female with 16 years of education who is not a regular churchgoer
  - (d) a female with 16 years of education who is a regular churchgoer.

That is, complete the following table using the values above.

Church	Female	Educ	Log odds	Odds	P(Happy)

Do this first by hand. Then confirm your answers by using the `adjust` and/or `predict` and/or `margins` commands.

4. What are the values of  $DEV_0$ ,  $DEV_M$ , and  $G_M$ ? Explain what each of these parameters means and, in the case of  $G_M$ , what hypothesis it is testing and whether or not you should reject that hypothesis given the results. Also, what does McFadden's Pseudo  $R^2$  equal? (Note that some of these values are explicitly reported in the printout while others require minor computations.)
5. Run the following post-estimation commands and `extremes` command (you need to have the `extremes` command installed):

```
estat class
predict phappy
predict rstandard, rstandard
extremes rstandard happymar phappy church female educ
```

What is the proportion of cases that have been correctly classified? Of the cases that have been improperly classified, which ones appear to be the most problematic? [NOTE: SPSS and Stata handle standardized residuals somewhat differently and will give somewhat different results.]

6. The data set also includes a variable, `educx`, which is equal to education centered about its mean. Rerun the logistic regression using `educx`. Note that the value of the intercept (but no other coefficient) changes when you do this. Explain how to interpret the intercept once education is centered, and how that differs from the interpretation when education is not centered. Review your earlier notes on centering if necessary.
7. The data set also includes the interaction `cheducx = church * educx`. Add it to the model and use a likelihood ratio chi-square test (i.e. don't just rely on the Wald statistic) to test whether the effect of `cheducx` is significant. What is the value of the test statistic and what does it tell you?

III. (Optional. This is the old SPSS version of the problem. You can do part or all of this problem if you want experience with SPSS or want to double-check your Stata work.) You need to copy `lrb.sps`, `lrb.sav` and `lrcalc.sps` from my web page. We use a sample of Southern Baptists from the GSS in this homework. General Social Surveys from 1973 to 1991 are used to make one big sample. All married Southern Baptists between the ages of 20 to 25 (all 61 of them!) are in the data file. The dependent variable is `HAPPYMAR`, respondent's marital happiness (1 = Very Happy, 0 = Otherwise). `CHURCH`, Church attendance (1 = Often attends, 0 = other), `FEMALE` (1 = female, 0 = male), and `EDUC`, Years of education, are the DVs.

*NOTE:* Both the NOMREG and LOGISTIC REGRESSION routines are used in lrb.sps. NOMREG is designed for multinomial logistic regression (i.e categorical DVs with more than 2 categories) but can also be used with a dichotomous DV. The NOMREG and LOGISTIC REGRESSION routines have different options in them, so which is best for your purposes depends on what output you want. LOGISTIC REGRESSION is better if you want to test a hierarchy of models, but in other cases NOMREG's output strikes me as being a little easier to read and a bit more informative.

There are some differences between NOMREG and LOGISTIC REGRESSION. NOMREG uses a different approach for computing deviances; sometimes you'll get the same results using both NOMREG and LOGISTIC REGRESSION and sometimes you won't. Nomreg's formulas for the deviances are based on the number of unique combinations of values (which it calls subpopulations), rather than just the number of cases. I wouldn't say that it is "wrong," but it is different than the way SPSS LOGISTIC REGRESSION or STATA's MLOGIT handle things.

Also, if you use the default settings, the signs of effects are flipped between the two routines; in NOMREG, the DV category with the highest value gets treated as the reference, which, in the case of a dichotomy, basically reverses the dichotomy's coding. (Again, whatever its other virtues, SPSS's internal inconsistencies can be pretty maddening at times.)

I've set the program up to get around these inconsistencies. If you want to use SPSS for your own work, you should pay attention to how I did this.

1. What assumptions of OLS would be violated if OLS was used to approach this problem?
2. Run lrb.sps. Interpret the logistic regression coefficients from Part I. What do the parameters tell you about the determinants of marital happiness? What can you say about the size and magnitude of effects?
3. Determine the log odds, odds and probability of marital happiness for:
  - (a) a male with 8 years of education who is not a regular churchgoer
  - (b) a male with 8 years of education who is a regular churchgoer
  - (c) a female with 16 years of education who is not a regular churchgoer
  - (d) a female with 16 years of education who is a regular churchgoer.

That is, complete the following table using the values above.

Church	Female	Educ	Log odds	Odds	P(Happy)

Do this first by hand. Confirm your answers by modifying lrcalc.sps. (I suggest you make a copy, call it lrcalc2.sps, and work off of it).

4. What are the values of  $DEV_0$ ,  $DEV_M$ , and  $G_M$ ? Explain what each of these parameters means and, in the case of  $G_M$ , what hypothesis it is testing and whether or not you should reject that

hypothesis given the results. Also, what does McFadden's Pseudo  $R^2$  equal? (Note that some of these values are explicitly reported in the printout while others require minor computations.)

5. What is the proportion of cases that have been correctly classified? Of the cases that have been improperly classified, which ones appear to be the most problematic?
6. Part II of lrb.sps uses centered variables.
  - a. The program creates a new variable, EDUCX, which is equal to education centered about its mean. It reruns the logistic regression using Educx. Note that the value of the intercept (but no other coefficient) changes when you do this. Explain how to interpret the intercept once education is centered, and how that differs from the interpretation when education is not centered.
  - b. Next the program computes the interaction CHEDUCX = Church \* Educx and adds it to the model. It uses a likelihood ratio chi-square test (i.e. it doesn't just rely on the Wald statistic) to test whether the effect of CHEDUCX is significant. What is the value of the test statistic and what does it tell you?