APPLIED STATISTICS FOR SOCIAL SCIENTISTS (SOCIOLOGY)

WEEK 8: CATEGORICAL DEPENDENT VARIABLES II

Topic: Logistic regression. Predicted probabilities.

STATA commands and features: `logit, logistic, test, lfit, fitstat, margins`

Data set: ghs72_04.dta, taken from the General Household Survey 1972 - 2004.

Readings: A. Agresti and B. Finlay (1997). [CHAPTER 15].

T.H. Wonnacott and R.J. Wonnacott. 1990. [CHAPTER 17].

Further readings: J. Long and J. Freese (2005) *Regression Models for Categorical Outcomes Using Stata*. Second Edition. College Station, TX: Stata Press.

INTRODUCTION

1. LOGISTIC REGRESSION MODELS WITH CONTINUOUS AND CATEGORICAL DEPENDENT VARIABLES

Working with categorical and continuous independent variables in logistic regression is not that much different from what we have done earlier with OLS regression. The interpretation of the coefficients is just a bit more complicated because we have to think in log odds (and odds ratios and predicted probabilities), but otherwise the same logic applies.

When you have a categorical variable with *k* categories, you have to construct *k* dummy variables. In the regression model, we will exclude one of the dummies, this is the reference category. The logit coefficient (b_j) for dummy *j* now tells us how much larger or smaller the log odds is for category *j* compared to the reference category. The odds ratio ($\exp^b$) tells us how much more likely outcome y is in group *j* compared to the reference group.

When the independent variable is continuous the interpretation is per unit change in x. The log odds increase or decrease by b for each unit change in x. If x is age, we would say that the log odd of y changes by b for each unit change in age. Say the logit coefficient for age is: b=0.03. The log odds increase by 0.03 for each additional year a respondent is older. In terms of odds ratios, we would say that the odds are exp(0.03)=1.03 times higher for each age compared to the previous age. In other words, the odds of Y=1/Y=0 for age 30 is 3% higher compared to the odds at age 29. Often the relationship becomes clearer by calculating predicted probabilities.

In stata11, you can use **`margins`** (type "help margins" in stata to see all the options – quite a lot!). The **`Spost`** package by Scott Long also provides very useful tools for calculating predicted probabilities. See: www.indiana.edu/~jslsoc/spost.htm

The clearest way, however, is to use the coefficients from the output and calculate probabilities yourself. See last week's hand-out and do-file for an example of how to do this.

EXERCISE 1

Use last week's data and do file on smoking (BHPS2001). Predict smoking status with age, sex and marital status.

1. Controlling for age and sex, what is the odds ratio of smoking for divorced respondents compared to married respondents?

2. Interpret the effect of age in terms of log odds, odds ratios and predicted probabilities. For the latter, present predicted probabilities at three different ages (30, 50 and 70) keeping all other co-variates at their mean. You can either calculate these thee by hand or you can use the margins code below. Check to see whether your own calculations give the same results.

```
margins , at(age=(30(20)70))
```

2. INTERACTIONS IN LOGISTIC REGRESSION

Interactions in logistic regression work very much like they do in OLS models, but again: the interpretation is just a bit more complicated.

Most straightforward is the interpretation in terms of log odds. Take as an example this model:
Employed= a + b1*_woman_ + b2*_degree_ + b3*_age_ + b4*_child_ + b5*_woman_*_child_
Where *child* is a dummy variable (1=children in household, 0=no children in hh)

We can no longer interpret the main effects – child and woman – on their own, but only in combination with each other. The effects of *woman* and *child* on housework should only be interpreted together with the interaction term, when other variables, i.e. *age* and *degree* are held constant. The interaction term indicates whether the effect of children is different for men and women. This is the output we obtain:

```
. logit employed woman degree age child woman_child if age<64

Logistic regression                               Number of obs   =       4406
                                                  LR chi2(5)      =     447.88
                                                  Prob > chi2     =     0.0000
Log likelihood = -2166.0667                       Pseudo R2       =     0.0937

------------------------------------------------------------------------------
    employed |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       woman |   -.583784   .0952007    -6.13   0.000    -.7703738   -.3971941
      degree |   .6186437     .12723     4.86   0.000     .3692774      .86801
         age |  -.0609504   .0046262   -13.18   0.000    -.0700176   -.0518832
       child |   .5143039   .1641942     3.13   0.002     .1924893    .8361185
 woman_child |  -1.181734   .1847289    -6.40   0.000    -1.543797   -.8196724
       _cons |   4.387373   .2504983    17.51   0.000     3.896405    4.87834
------------------------------------------------------------------------------
```

Having a child in the household increases the log odds of being employed:
for men with b4 + b5*0 = b4 = .51
for women with b4 + b5*1 = b4+b5 = .51-1.18=-0.67

The odds ratio for men is exp(b4) and for women it is exp(b4+b5), which equals exp(b4)*exp(b5):
Odds ratio for men exp(.51) =1.67
Odds ratio for women exp(.51)*exp(-1.18)=exp(-0.67) =0.51

Predicted probabilities, keeping age and education at the mean:

Remember that: $\Pr(employed) = \frac{1}{1+\exp(-\log\text{odds of employed})}$

We first calculate the log odds of being employed for the groups we are interested in and then we transform those to probabilities. The means for degree=.15, for age=44.4

Log odds men with child: .618*.15 + -.0609 * 44.4 + .51 + 4.387=2.288
```
(di _b[degree]*.15 + _b[age]*44.4 + _b[child] + _b[_cons])
```
The predicted probability now is: 1/(1+exp(-2.288)) = .91

Log odds men without child: .618*.15 + -.0609 * 44.4 + + 4.387=1.77
```
(di _b[degree]*.15 + _b[age]*44.4 + _b[_cons])
```
The predicted probability now is: 1/(1+exp(-1.77)) = .85

For women you have to take the interaction into account!
Log odds women with child: -.58 +.618*.15 + -.0609 * 44.4 + .51 - 1.18+ 4.387=.52
```
(di _b[woman] +_b[degree]*.15 + _b[age]*44.4 + _b[child] +
_b[woman_child]  + _b[_cons])
```
The predicted probability now is: 1/(1+exp(-.52)) = .63

Log odds women without child: -.58 +.618*.15 + -.0609 * 44.4 + + 4.387=1.19
```
(di _b[woman] +_b[degree]*.15 + _b[age]*44.4 + _b[child] +
_b[woman_child]  + _b[_cons])
```
The predicted probability now is: 1/(1+exp(-1.19)) = .77

(a quicker way to get these results would have been:
```
logit employed i.degree age i.woman#i.child if age<64
margins woman#child, atmeans)
```

EXERCISE 2

Use last week's data on voting for the *conservatives*. Select respondents with valid information on `sclass`. Combine the first two classes. Interact sex and class in a model that predicts voting for the *conservatives*. Control for age.

Is the relationship between voting preferences and class similar among men and women? Give the odds ratios for voting *conservatives* by gender in the three classes.

3. WALD TESTS

In Exercise 1, when we use married as a reference category, we observed that the coefficients for other marital dummies are statistically significant. Given that we are comparing categories, what we might actually be interested in is whether any one category is different from any other category. For example, we might want to find out if smoking differs between widowed and divorced people. One way of checking this is to change the reference category. Another way is to use the **`test`** command to check whether two or more coefficients are equal to each other. To test whether the coefficient for var1 is different from coefficient for var2, we type:

```
test var1=var2
```

We are testing here whether we can distinguish statistically between the size of the two coefficients. Our null hypothesis here is that the coefficients are the same. If the p-value is smaller than 0.05, we can reject this hypothesis, and therefore conclude that the coefficients are statistically different from each other.

EXERCISE 3

As observed earlier, smoking is associated with marital status. Do separated and divorced respondents really differ from each other in the likelihood to smoke? What about widowed and divorced?

As an extra exercise, rerun the models to get the odds ratios with different reference groups and calculate the predicted probabilities.

4. GOODNESS OF FIT OF LOGISTIC REGRESSION MODELS

We often use a likelihood ratio test to compare nested models and as an initial model fit measure. The likelihood-ratio test uses the ratio of the maximized value of the likelihood function for the full model (L_1) over the maximized value of the likelihood function for the simpler model (L_0). The likelihood-ratio test statistic equals:

$$-2\log\left(\frac{L_0}{L_1}\right) = -2[\log(L_0) - \log(L_1)] = -2(L_0 - L_1)$$

It follows a chi-squared distribution with p - 1 degrees of freedom, where p is the number of parameters in the model. In other words, the likelihood-ratio test is the chi-square difference between the null model (the simplest model) and the model containing one or more predictors. It is an assessment of the improvement of fit between the predicted and observed values on the dependent variable Y by adding predictor(s).

The lr-test (against the intercept-only model) and -2LL are the most useful pieces of information to report about your logit model. You can get a whole range of fit measures after running a logit model with the command: `fitstat`