Logistic Regression

Logistic regression is a variation of the regression model. It is used when the dependent response variable is binary in nature. Logistic regression predicts the probability of the dependent response, rather than the value of the response (as in simple linear regression).

In this example, the dependent variable is frequency of sex (less than once per month versus more than once per month).

```
. logistic freqdum age marital racenew attend happy

Logistic regression                                 Number of obs   =       1052
                                                    LR chi2(5)      =     302.11
                                                    Prob > chi2     =     0.0000
Log likelihood = -567.40591                         Pseudo R2       =     0.2102b

------------------------------------------------------------------------------
     freqdum | Odds Ratioa  Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   .9408818   .0047102   -12.17   0.000     .9316952     .950159
     marital |     .17496   .0290688   -10.49   0.000     .1263327    .2423048
     racenew |   .8854209   .1564181    -0.69   0.491      .626292    1.251765
      attend |   .9309624   .0262984    -2.53   0.011     .8808194    .9839599
       happy |   .7156043    .087177    -2.75   0.006     .5636079    .9085918
------------------------------------------------------------------------------
```

a. The “Odds Ratio” is the predicted change in odds for a unit increase in the predictor. When the Odds Ratio is less than 1, increasing values of the variable correspond to decreasing odds of the event's occurrence. When the Odds Ratio is greater than 1, increasing values of the variable correspond to increasing odds of the event's occurrence.

If you subtract 1 from the odds ratio and multiply by 100, you get the percent change in odds of the dependent variable having a value of 1. For example, for age:
= 1 - (.941) = .059
= .059 * 100 = 5.9%
The odds ratio for age indicates that every unit increase in age is associated with a 5.9% decrease in the odds of having sex more than once a month.

b. The R-Square statistic cannot be exactly computed for logistic regression models, so these approximations are computed instead. Larger pseudo r-square statistics indicate that more of the variation is explained by the model, to a maximum of 1.

Interpretation

Recall: When Exp(B) is less than 1, increasing values of the variable correspond to decreasing odds of the event's occurrence. When Exp(B) is greater than 1, increasing values of the variable correspond to increasing odds of the event's occurrence.

Constant = Not interpretable in logistic regression.

Age = Increasing values of age correspond with decreasing odds of having sex more than once a month.

Marital = Increasing values of marital status (married to unmarried) correspond with decreasing odds of having sex more than once a month.

Race = Increasing values of race correspond with increasing odds of having sex more than once a month. Notice that this variable, however, is not significant.

Church Attendance = Increasing values of church attendance correspond with decreasing odds of having sex more than once a month.

Happiness = Increasing values of general happiness correspond with decreasing odds of having sex more than once a month. Recall that happiness is coded such that higher values indicate less happiness.

Logistic Regression (with non-linear variables)

It is known that some variables are often non-linear, or curvilinear. Such variables may be age or income. In this example, we include the original age variable and an age squared variable.

```
. logistic freqdum age marital racenew attend happy agesquar

Logistic regression                               Number of obs   =       1052
                                                  LR chi2(6)      =     314.11
                                                  Prob > chi2     =     0.0000
Log likelihood = -561.40465                       Pseudo R2       =     0.2186

------------------------------------------------------------------------------
     freqdum | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   1.029322   .0274165     1.09   0.278     .9769654    1.084485
     marital |   .1876667   .0319891    -9.82   0.000     .1343674    .2621081
     racenew |   .8854594   .1559968    -0.69   0.490     .6269129    1.250633
      attend |   .9311651   .0264615    -2.51   0.012     .8807195    .9845002
       happy |   .7026621    .085654    -2.89   0.004     .5533319    .8922928
    agesquar |   .9990557   .0002795    -3.38   0.001      .998508    .9996037
------------------------------------------------------------------------------
```

The age squared variable is significant, indicating that age is non-linear.

Logistic Regression (with interaction term)

To test for two-way interactions (often thought of as a relationship between an independent variable (IV) and dependent variable (DV), moderated by a third variable), first run a regression analysis, including both independent variables (IV and moderator) and their interaction (product) term. It is highly recommended that the independent variable and moderator are standardized before calculation of the product term, although this is not essential. For this example, two dummy variables were created, for ease of interpretation. Sex was recoded such that 1=Male and 0=Female. Marital status was recoded such that 1=Currently married and 0=Not currently married. The interaction term is a product of these two dummy variables.

Regression Model (without interactions)

```
. logistic freqdum age racenew happy attend male married

Logistic regression                               Number of obs   =       1052
                                                  LR chi2(6)      =     311.10
                                                  Prob > chi2     =     0.0000
Log likelihood = -562.91057                       Pseudo R2       =     0.2165

------------------------------------------------------------------------------
     freqdum | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   .9407224   .0047616   -12.07   0.000     .9314359    .9501014
     racenew |   .8615719   .1535544    -0.84   0.403     .6075547    1.221793
       happy |   .7273111   .0893135    -2.59   0.010     .5717327    .9252252
      attend |   .9422746   .0269726    -2.08   0.038     .8908649     .996651
        male |   1.558383   .2310822     2.99   0.003     1.165347    2.083978
     married |   5.464756   .9139101    10.16   0.000     3.937479    7.584435
------------------------------------------------------------------------------
```

Regression Model (with interactions)

```
. logistic freqdum age racenew happy attend male married interact

Logistic regression                               Number of obs   =       1052
                                                  LR chi2(7)      =     313.87
                                                  Prob > chi2     =     0.0000
Log likelihood =  -561.5232                       Pseudo R2       =     0.2184

------------------------------------------------------------------------------
     freqdum | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   .9417151   .0047935   -11.80   0.000     .9323667    .9511572
     racenew |   .8412302   .1508375    -0.96   0.335      .591956    1.195474
       happy |   .7244743   .0892968    -2.61   0.009     .5689921    .9224436
      attend |   .9452654   .0271617    -1.96   0.050     .8935009    1.000029
        male |   1.913216   .3698852     3.36   0.001     1.309784    2.794655
     married |    6.93103   1.542039     8.70   0.000     4.481461    10.71953
    interact |   .6043344   .1827447    -1.67   0.096     .3341046    1.093131
------------------------------------------------------------------------------
```

The product term should be significant in the regression equation in order for the interaction to be interpretable. In this example, the interaction term is significant at the 0.1 level.

Interpretation

Main Effects

The married coefficient represents the main effect for females (the 0 category). The effect for females is then 1.94, or the “marital” coefficient. The effect for males is 1.94 - .50, or 1.44.

The gender coefficient represents the main effect for unmarried persons (the 0 category). The effect for unmarried is then .65, or the “sex” coefficient. The effect for married is .65 - .50, or .15.

Odds Ratios

Using “married” as the focus variable, we can say that the effect of being married on having sex more than once per month is greater for females.

Females: $e^{1.936} = 6.93$
Males: $e^{1.432} = 4.20$

Using “gender” as the focus variable, we can say that the effect of being male on having sex more than once per month is greater for marrieds.

Marrieds: $e^{0.15} = 1.16$
Unmarrieds: $e^{0.65} = 1.92$