Sociology 704: Topics in Multivariate Statistics
Instructor: Natasha Sarkisian

Binary Logit

Binary models deal with binary (0/1, yes/no) dependent variables. OLS is inappropriate for this kind of dependent variable because we would violate numerous OLS assumptions (e.g., that the dependent variable is quantitative, continuous, and unbounded, or that the error terms should be homoscedastic and normally distributed).

Two main types of binary regression models are used most often – logit and probit. The two types differ in terms of the assumed variance of the error term, but in practice their results are usually very similar, and the choice between the two is mainly the matter of taste and discipline conventions. We'll mostly focus on logit models.

Binary logit and probit models as well as other models we'll discuss this semester are estimated using Maximum Likelihood estimation techniques – numerical, iterative techniques that search for a set of parameters with the highest level of the likelihood function (likelihood function tells us how likely it is that we would observe the data in hand for each set of parameters, and in fact what we maximize is the log of this likelihood function). This process is a trial and error process. Logit or probit output includes information on iterations – those iterations are the steps in that search process. Sometimes, with complicated models, the computer cannot find that maximum – then we get convergence problems. But this never happens with binary logit or probit models.

To run logit or probit models in Stata, the dependent variable has to be coded 0/1 -- it cannot be 1 and 2, or anything else. Let's generate a 0/1 variable:

```
. codebook grass
-------------------------------------------------------------------------------
grass
should marijuana be made legal
-------------------------------------------------------------------------------
                  type:  numeric (byte)
                 label:  grass
                 range:  [1,2]                        units:  1
         unique values:  2                        missing .:  1914/2765

            tabulation:  Freq.   Numeric  Label
                           306         1  legal
                           545         2  not legal
                          1914         .

. gen marijuana=(grass==1) if grass~=.
(1914 missing values generated)
. tab marijuana, miss
  marijuana |      Freq.     Percent        Cum.
------------+-----------------------------------
          0 |        545       19.71       19.71
          1 |        306       11.07       30.78
          . |      1,914       69.22      100.00
------------+-----------------------------------
      Total |      2,765      100.00
```

```
. xi: logit marijuana sex educ age childs
Iteration 0:   log likelihood =  -552.0232
Iteration 1:   log likelihood = -525.24385
Iteration 2:   log likelihood = -524.84887
Iteration 3:   log likelihood = -524.84843
Logistic regression                                     Number of obs   =        845
                                                        LR chi2(4)      =      54.35
                                                        Prob > chi2     =     0.0000
Log likelihood = -524.84843                             Pseudo R2       =     0.0492

------------------------------------------------------------------------------
   marijuana |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         sex |    -.34799   .1494796    -2.33   0.020    -.6409647   -.0550152
        educ |   .0401891    .025553     1.57   0.116     -.009894    .0902722
         age |  -.0183109   .0049147    -3.73   0.000    -.0279436   -.0086782
      childs |  -.1696747   .0536737    -3.16   0.002    -.2748733   -.0644762
       _cons |   .5412516   .4595609     1.18   0.239    -.3594713    1.441974
------------------------------------------------------------------------------
```

Basic interpretation: Women are less likely than men to support legalization of marijuana. The effect of education is not statistically significant. Those who are older and have more children are less likely to support legalization. Divorced people are more likely than married people to support legalization.

```
*Same with probit
. probit marijuana sex educ age childs
Iteration 0:   log likelihood =  -552.0232
Iteration 1:   log likelihood = -525.34877
Iteration 2:   log likelihood = -525.21781
Iteration 3:   log likelihood =  -525.2178
Probit regression                                       Number of obs   =        845
                                                        LR chi2(4)      =      53.61
                                                        Prob > chi2     =     0.0000
Log likelihood =  -525.2178                             Pseudo R2       =     0.0486

------------------------------------------------------------------------------
   marijuana |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         sex |  -.2101429   .0910856    -2.31   0.021    -.3886673   -.0316184
        educ |   .0229968   .0151532     1.52   0.129     -.006703    .0526965
         age |  -.0111514   .0029499    -3.78   0.000    -.0169331   -.0053696
      childs |  -.0984716   .0314167    -3.13   0.002    -.1600472    -.036896
       _cons |   .3374219   .2782445     1.21   0.225    -.2079273    .8827711
------------------------------------------------------------------------------
```

Goodness of fit

```
. estat gof
Logistic model for marijuana, goodness-of-fit test
       number of observations =       845
 number of covariate patterns =       748
            Pearson chi2(743) =       748.27
                  Prob > chi2 =         0.4389
```

The high p-value indicates that model fits well (there is no significant discrepancy between observed and predicted frequencies). But: this is a chi-square test that compares observed and predicted outcomes in cells defined by

covariate patterns – all possible combinations of independent variables. In this case, there are 770 covariate patterns, so it 770 cells for chi-square test, and therefore very few cases per cell. Not a good situation for a chi-square test.

Hosmer and Lemeshow suggested an alternative measure that solves the problem of too many covariate patterns. Rather than compare the observed and predicted frequencies in each covariate pattern, they divide the data into ten cells by sorting it according to the predicted probabilities and breaking it into deciles (i.e. the 10% of observations with lowest predicted probabilities form the first group, then next 10% the next group, etc.). This measure of goodness of fit is usually preferred over the Pearson chi-square. Here's how we obtain it:

```
. estat gof, group(10)
Logistic model for marijuana, goodness-of-fit test
  (Table collapsed on quantiles of estimated probabilities)
       number of observations =       845
             number of groups =        10
      Hosmer-Lemeshow chi2(8) =        10.55
                  Prob > chi2 =         0.2287
```

Again, the model appears to fit well. If it were not, we could rely on various diagnostics (discussed below) to improve model fit.

Other measures of fit can be obtained using fitstat. But first, we need to install it, along with other commands written by Scott Long, the author of our textbook:

```
. net search spost
[output omitted]
We need spostado from http://www.indiana.edu/~jslsoc/stata
```

Now let's obtain fit statistics for our last model

```
. fitstat, save
Measures of Fit for logit of marijuana
Log-Lik Intercept Only:       -552.023   Log-Lik Full Model:          -524.848
D(840):                       1049.697   LR(4):                         54.350
                                         Prob > LR:                      0.000
McFadden's R2:                   0.049   McFadden's Adj R2:              0.040
ML (Cox-Snell) R2:               0.062   Cragg-Uhler(Nagelkerke) R2:     0.085
McKelvey & Zavoina's R2:         0.090   Efron's R2:                     0.065
Variance of y*:                  3.615   Variance of error:              3.290
Count R2:                        0.669   Adj Count R2:                   0.079
AIC:                             1.254   AIC*n:                       1059.697
BIC:                         -4611.346   BIC':                         -27.392
BIC used by Stata:            1083.394   AIC used by Stata:           1059.697
```

See pp. 104-113 of Long and Freese for details on these measures of fit. McFadden's R2 is what's commonly reported as Pseudo-R2, although that tends to be fairly low.

Log likelihood value or deviance (-2LL) are also frequently reported. Examining the ratio of D/df to see how far from 1.0 it is gives us an idea of model fit (here: 1049.697/840=1.2496393).

Another very useful measure is BIC – based on the differences in BIC between models, we can select a model with a better fit more reliably than based on a

difference in Pseudo-R2 or even based on lrtest. Here's how we compare model fit using fitstat. We already saved the results of the previous model. Let's say, we consider adding the marital status dummies:

```
. xi: logit marijuana sex age educ childs i.marital
i.marital         _Imarital_1-5       (naturally coded; _Imarital_1 omitted)
Logistic regression                               Number of obs   =        845
                                                  LR chi2(8)      =      74.79
                                                  Prob > chi2     =     0.0000
Log likelihood = -514.62716                       Pseudo R2       =     0.0677
------------------------------------------------------------------------------
   marijuana |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         sex |  -.3620539   .1532607    -2.36   0.018    -.6624394   -.0616684
         age |  -.0177167   .0056026    -3.16   0.002    -.0286977   -.0067357
        educ |    .041343   .0263959     1.57   0.117    -.0103919    .0930779
      childs |  -.1614819   .0581657    -2.78   0.005    -.2754846   -.0474793
 _Imarital_2 |   .0118099   .3568915     0.03   0.974    -.6876845    .7113043
 _Imarital_3 |   .9025573   .2053011     4.40   0.000     .5001746     1.30494
 _Imarital_4 |   .0300665   .4239309     0.07   0.943    -.8008229    .8609558
 _Imarital_5 |   .2853992    .208832     1.37   0.172     -.123904    .6947024
       _cons |   .2573784   .5195598     0.50   0.620    -.7609401    1.275697
------------------------------------------------------------------------------

. fitstat, dif
Measures of Fit for logit of marijuana
                               Current            Saved        Difference
Model:                           logit            logit
N:                                 845              845                 0
Log-Lik Intercept Only        -552.023         -552.023             0.000
Log-Lik Full Model            -514.627         -524.848            10.221
D                            1029.254(836)    1049.697(840)       20.443(4)
LR                             74.792(8)        54.350(4)         20.443(4)
Prob > LR                          0.000            0.000             0.000
McFadden's R2                      0.068            0.049             0.019
McFadden's Adj R2                  0.051            0.040             0.011
ML (Cox-Snell) R2                  0.085            0.062             0.022
Cragg-Uhler(Nagelkerke) R2         0.116            0.085             0.031
McKelvey & Zavoina's R2            0.120            0.090             0.030
Efron's R2                         0.087            0.065             0.023
Variance of y*                     3.740            3.615             0.125
Variance of error                  3.290            3.290             0.000
Count R2                           0.673            0.669             0.005
Adj Count R2                       0.092            0.079             0.013
AIC                                1.239            1.254            -0.015
AIC*n                           1047.254         1059.697           -12.443
BIC                            -4604.831        -4611.346             6.515
BIC'                             -20.877          -27.392             6.515
BIC used by Stata               1089.908         1083.394             6.515
AIC used by Stata               1047.254         1059.697           -12.443
Difference of    6.515 in BIC' provides strong support for saved model.
Note: p-value for difference in LR is only valid if models are nested.
```

This suggests that adding marital status does not add enough to justify adding 4 extra variables. Again, we could consider adding just one dummy, divorced, and that would probably be "worth it" in terms of model fit.

Here's how to interpret the difference in BIC (guidelines from Raftery 1995):

TABLE 6

Grades of Evidence Corresponding to Values of the Bayes Factor for M_2 Against M_1, the BIC Difference and the Posterior Probability of M_2

BIC Difference	Bayes Factor	$p(M_2\|D)(\%)$	Evidence
0–2	1–3	50–75	Weak
2–6	3–20	75–95	Positive
6–10	20–150	95–99	Strong
>10	>150	>99	Very strong

Note that if the variable you add to the second model changes the number of cases (because of missing data), BIC comparison won't work. E.g., add income:

```
. logit marijuana sex age educ childs  rincom98
Logistic regression                               Number of obs   =        599
                                                  LR chi2(5)      =      35.29
                                                  Prob > chi2     =     0.0000
Log likelihood = -379.82272                       Pseudo R2       =     0.0444
------------------------------------------------------------------------------
   marijuana |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         sex |  -.5153134    .181267    -2.84   0.004    -.8705902   -.1600366
         age |  -.0079214   .0072892    -1.09   0.277    -.0222079    .0063651
        educ |   .0849509   .0336502     2.52   0.012     .0189976    .1509041
      childs |  -.2199136   .0676456    -3.25   0.001    -.3524965   -.0873307
    rincom98 |  -.0352966   .0162986    -2.17   0.030    -.0672413    -.003352
       _cons |   .3036228   .5639177     0.54   0.590    -.8016357    1.408881
------------------------------------------------------------------------------

. fitstat, dif
Measures of Fit for logit of marijuana
                               Current             Saved        Difference
Model:                           logit             logit
N:                                 599               845              -246
N's do not match. To make the comparisons, use the force option.
```

Because our samples are not the same, so it's problematic to compare models. Do not use force option, however – such a comparison would not be correct. A better strategy is to limit both models to the same sample:

```
. logit marijuana sex age educ childs if rincom98~=.
Logistic regression                               Number of obs   =        599
                                                  LR chi2(4)      =      30.57
                                                  Prob > chi2     =     0.0000
Log likelihood = -382.18666                       Pseudo R2       =     0.0385
------------------------------------------------------------------------------
   marijuana |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         sex |  -.4295858   .1756775    -2.45   0.014    -.7739073   -.0852643
         age |  -.0096812   .0072661    -1.33   0.183    -.0239226    .0045601
        educ |   .0604882   .0312321     1.94   0.053    -.0007257     .121702
      childs |  -.2182796   .0678493    -3.22   0.001    -.3512617   -.0852974
       _cons |   .0640233   .5479271     0.12   0.907    -1.009894    1.137941
------------------------------------------------------------------------------

. fitstat, save
```

```
Measures of Fit for logit of marijuana
Log-Lik Intercept Only:       -397.470   Log-Lik Full Model:           -382.187
D(594):                        764.373   LR(4):                         30.566
                                         Prob > LR:                      0.000
McFadden's R2:                   0.038   McFadden's Adj R2:              0.026
ML (Cox-Snell) R2:               0.050   Cragg-Uhler(Nagelkerke) R2:     0.068
McKelvey & Zavoina's R2:         0.069   Efron's R2:                     0.053
Variance of y*:                  3.534   Variance of error:              3.290
Count R2:                        0.644   Adj Count R2:                   0.062
AIC:                             1.293   AIC*n:                        774.373
BIC:                         -3034.412   BIC':                          -4.985
BIC used by Stata:             796.350   AIC used by Stata:            774.373

. logit marijuana sex age educ childs rincom98
Logistic regression                              Number of obs   =        599
                                                 LR chi2(5)      =      35.29
                                                 Prob > chi2     =     0.0000
Log likelihood = -379.82272                      Pseudo R2       =     0.0444
------------------------------------------------------------------------------
   marijuana |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         sex |  -.5153134    .181267    -2.84   0.004    -.8705902   -.1600366
         age |  -.0079214   .0072892    -1.09   0.277    -.0222079    .0063651
        educ |   .0849509   .0336502     2.52   0.012     .0189976    .1509041
      childs |  -.2199136   .0676456    -3.25   0.001    -.3524965   -.0873307
    rincom98 |  -.0352966   .0162986    -2.17   0.030    -.0672413    -.003352
       _cons |   .3036228   .5639177     0.54   0.590    -.8016357    1.408881
------------------------------------------------------------------------------
. fitstat, dif
Measures of Fit for logit of marijuana
                              Current             Saved        Difference
Model:                          logit             logit
N:                                599               599                 0
Log-Lik Intercept Only        -397.470          -397.470             0.000
Log-Lik Full Model            -379.823          -382.187             2.364
D                              759.645(593)      764.373(594)        4.728(1)
LR                              35.294(5)         30.566(4)          4.728(1)
Prob > LR                        0.000             0.000             0.030
McFadden's R2                    0.044             0.038             0.006
McFadden's Adj R2                0.029             0.026             0.003
ML (Cox-Snell) R2                0.057             0.050             0.007
Cragg-Uhler(Nagelkerke) R2       0.078             0.068             0.010
McKelvey & Zavoina's R2          0.078             0.069             0.009
Efron's R2                       0.060             0.053             0.008
Variance of y*                   3.569             3.534             0.035
Variance of error                3.290             3.290             0.000
Count R2                         0.658             0.644             0.013
Adj Count R2                     0.097             0.062             0.035
AIC                              1.288             1.293            -0.005
AIC*n                          771.645           774.373            -2.728
BIC                          -3032.745         -3034.412             1.667
BIC'                            -3.317            -4.985             1.667
BIC used by Stata              798.017           796.350             1.667
AIC used by Stata              771.645           774.373            -2.728
Difference of    1.667 in BIC' provides weak support for saved model.
Note: p-value for difference in LR is only valid if models are nested.
```

It looks like based on BIC we wouldn't add income to the model. Another way to assess model fit is to concentrate on its predictive powers. This is especially important when we plan to use the model for prediction (e.g., we want to predict who would support legalization of marijuana for a sample that does not contain those data but contains all our independent variables). One way to assess predictive power is to look at prediction statistics:

```
. qui logit marijuana sex age educ childs
[output omitted]
. estat clas
Logistic model for marijuana
              -------- True --------
Classified |         D            ~D  |      Total
-----------+--------------------------+-----------
     +     |        72            48  |        120
     -     |       232           493  |        725
-----------+--------------------------+-----------
   Total   |       304           541  |        845
Classified + if predicted Pr(D) >= .5
True D defined as marijuana != 0
--------------------------------------------------
Sensitivity                     Pr( +| D)   23.68%
Specificity                     Pr( -|~D)   91.13%
Positive predictive value       Pr( D| +)   60.00%
Negative predictive value       Pr(~D| -)   68.00%
--------------------------------------------------
False + rate for true ~D        Pr( +|~D)    8.87%
False - rate for true D         Pr( -| D)   76.32%
False + rate for classified +   Pr(~D| +)   40.00%
False - rate for classified -   Pr( D| -)   32.00%
--------------------------------------------------
Correctly classified                        66.86%
```

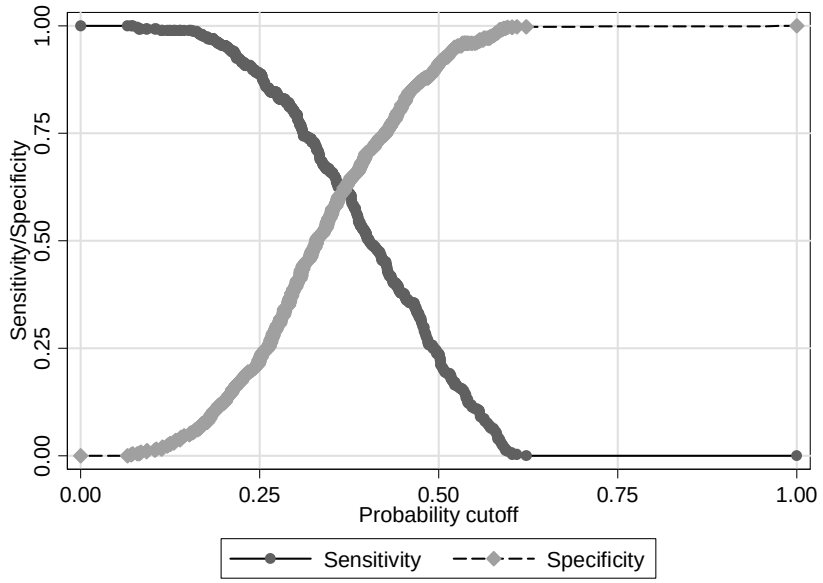
We can see that our model classified correctly 66.86% of cases. Note that it only classified 120 people out of 845 as supporters of marijuana legalization. The four cells in the table indicate how classification by the model compares to true status of each case. The statistics below reflect the percentage from the table above and indicate predictive success rates and rates of errors. Sensitivity indicates the percentage of cases with Y=1 that we identified correctly, and specificity indicates the percentages of cases with Y=0 that we classified correctly. We can see that our sensitivity is 23.68 but our specificity is much higher (91.13%). To alter that for a given model, we can change the cutoff point. In this table, the cutoff is 0.5 – this means that all observations with predicted probabilities of .5 and above get classified as 1 (i.e. supporters of legalization) and those observations with predicted probabilities below .5 are classified as 0 (against legalization). It appears that most cases have predicted probabilities below .5. Let's try to shift that cutoff to .3:

```
. estat clas, cutoff(.3)
Logistic model for marijuana
              -------- True --------
Classified |         D            ~D  |      Total
-----------+--------------------------+-----------
     +     |       242           329  |        571
     -     |        62           212  |        274
-----------+--------------------------+-----------
   Total   |       304           541  |        845
```

```
Classified + if predicted Pr(D) >= .3
True D defined as marijuana != 0
--------------------------------------------------
Sensitivity                     Pr( +| D)   79.61%
Specificity                     Pr( -|~D)   39.19%
Positive predictive value       Pr( D| +)   42.38%
Negative predictive value       Pr(~D| -)   77.37%
--------------------------------------------------
False + rate for true ~D        Pr( +|~D)   60.81%
False - rate for true D         Pr( -| D)   20.39%
False + rate for classified +   Pr(~D| +)   57.62%
False - rate for classified -   Pr( D| -)   22.63%
--------------------------------------------------
Correctly classified                        53.73%
--------------------------------------------------
```

Now our sensitivity and specificity are more balanced. We can further examine them and then select a cutoff point using the following command that graphs them against each other:

```
. lsens
```



Looks like the cutoff point of .4 would be close to the point where specificity and sensitivity are equal. But, the selection of the cutoff will depend on what's more important to us – correctly identify 0s or 1s, and what type of error is more problematic to us – this will depend in the task at hand.

Diagnostics for binary logit

Before conducting logistic regression, it might be a good idea to check univariate distributions of independent variables and if some deviate substantially from normal and you can easily correct that with a transformation, then try those transformations. Although normality is not required, it may help avoid other problems. Obviously, this does not apply to your dependent variable. Also note that in logistic regression, we do not expect residuals to be normally distributed.

Further, before conducting multivariate analysis, you should also check the linearity of bivariate relationships (see below).

1. Multicollinearity

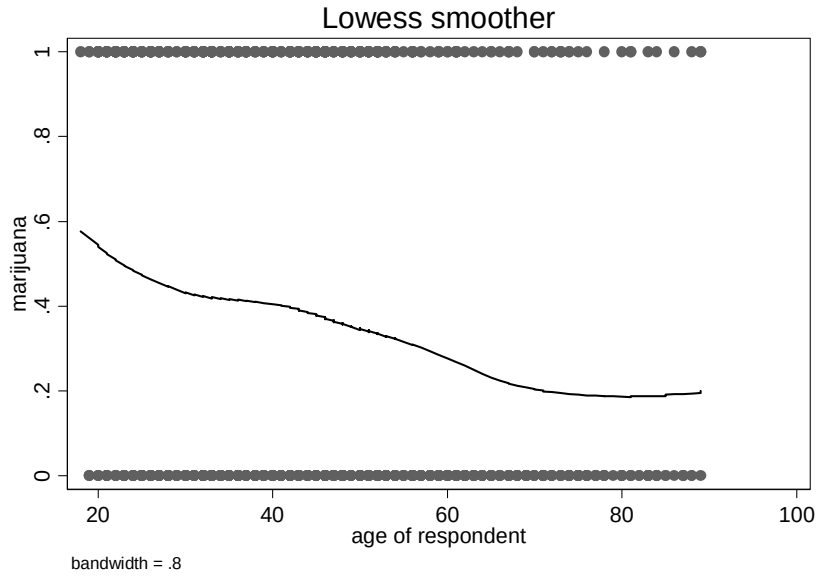
For multicollinearity, we can again use VIFs. But to obtain them, we need to run a regular OLS regression model with the same variables and then obtain VIFs – VIF command doesn't function after logit regression, even though VIF statistics don't depend on the dependent variable but rather on the correlations among the independent ones. So here's what we'd do:

```
. qui reg marijuana sex age educ childs  _Imarital_3
. vif

    Variable |       VIF       1/VIF
-------------+----------------------
      childs |      1.25    0.800429
         age |      1.21    0.823595
        educ |      1.04    0.959260
         sex |      1.01    0.985564
 _Imarital_3 |      1.01    0.989556
-------------+----------------------
    Mean VIF |      1.11
```

2. Linearity

In logistic regression, linearity and additivity in logits is expected (i.e. the relationships are nonlinear, but they should be linear in terms of the log odds). Bivariate graphical examination using lowess helps identify problems:

```
.lowess marijuana age
```



Note that we should not expect a straight line – after all, probability curve is not a straight line. But this can help you spot, for instance, a parabola.

In multivariate context, you can use boxtid--don't forget to specify that you are using logit rather then reg when using boxtid, i.e. use:

```
. boxtid logit marijuana sex age educ childs
```

3. Additivity

You can once again use fitint command to search for interactions; the syntax is

```
. fitint logit marijuana sex age educ childs, twoway(sex age educ childs)
factor(sex)
```

Note that interactions as a method to compare two or more groups can be problematic in logit or probit models because the coefficients are scaled according to the differences in residual dispersion. If you are interested in group comparisons, see:

Allison, Paul D. 1999. "Comparing Logit and Probit Coefficients Across Groups." *Sociological Methods and Research,* 28: 186-208.

Hoetker, Glenn. 2004. *"Confounded Coefficients: Extending Recent Advances in the Accurate Comparison of Logit and Probit Coefficients Across Groups."* http://www.business.uiuc.edu/Working_Papers/papers/03-0100.pdf

Long, Scott. 2006. Comparing Group Effects in Logit and Probit Models. http://www.umass.edu/family/conference/Long.htm

4. Outliers and influential data points

To detect influential observations and outliers, there are a few statistics you can obtain using predict command after logit

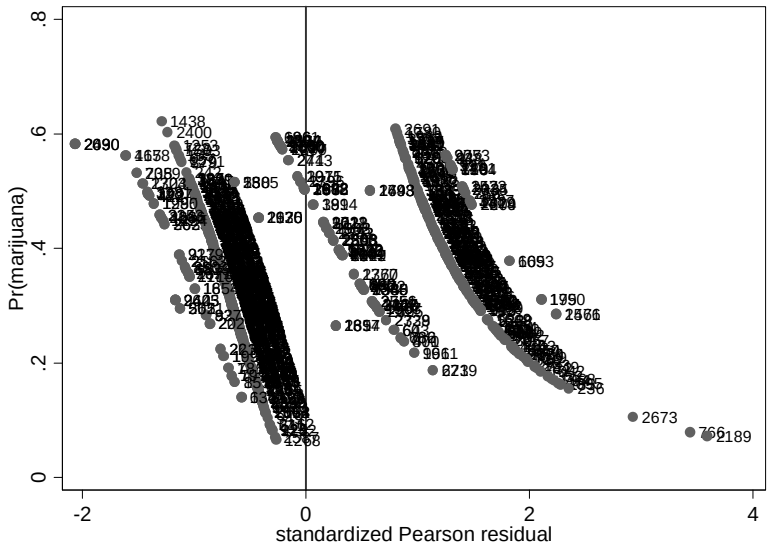
```
p           predicted probability of a positive outcome; the default
xb          linear prediction
stdp        standard error of the linear prediction
dbeta       Pregibon (1981) Delta-Beta influence statistic
deviance    deviance residual
dx2         Hosmer and Lemeshow (2000) Delta chi-squared infl. stat.
ddeviance   Hosmer and Lemeshow (2000) Delta-D influence statistic
hat         Pregibon (1981) leverage
number      sequential number of the covariate pattern
residuals   Pearson residual (adj. for # sharing covariate pattern)
rstandard   standardized Pearson residual (adj. for # sharing covariate
pattern)
```

To examine residuals, it is recommended to use standardized Pearson residual that accounts for in-built heteroscedasticity of residuals in the logit model.

```
. logit marijuana sex age educ childs
[Output omitted]

. predict rstandard, rs
(1920 missing values generated)
```

We can plot residuals against the predicted values and examine observations with residuals high in absolute value:

```
. predict prob
(option p assumed; Pr(marijuana))
(25 missing values generated)

. scatter prob rstandard, xline(0) mlabel(id)
```



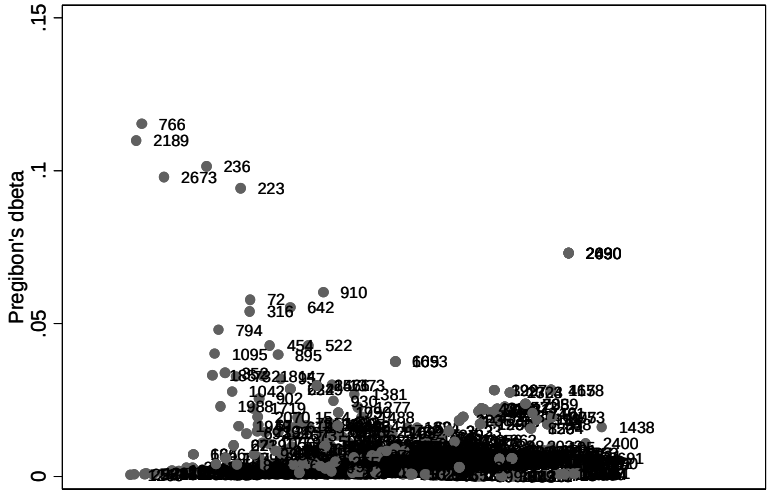
Observations on the far left or far right deserve further examination. Here, we would especially look at 766 and 2189, but also 2673.

To identify influential observations, we can obtain a number of leverage statistics:

```
. predict dbeta, dbeta
(1920 missing values generated)
. predict hat, hat
(1920 missing values generated)
. predict dx2, dx2
(1920 missing values generated)
```

We can then examine these graphically to identify problematic observations:

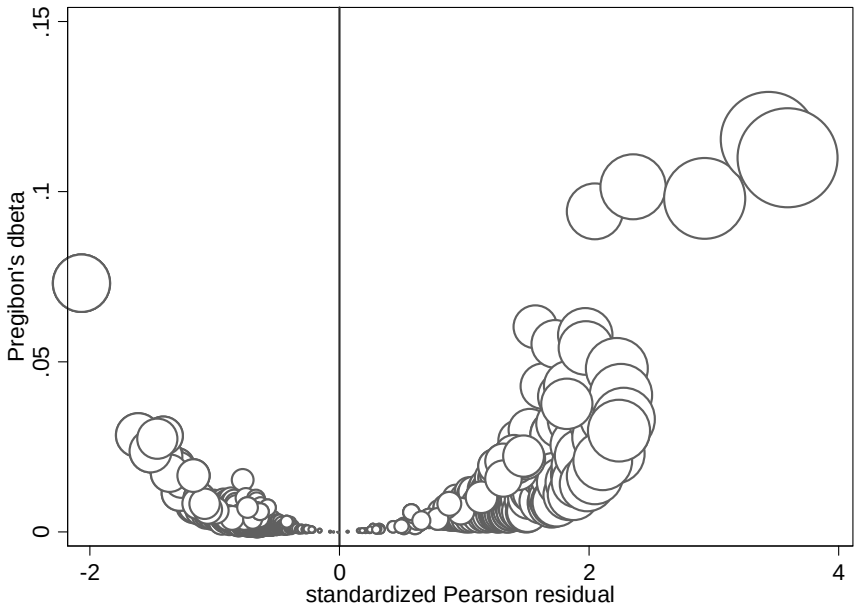
```
. scatter dbeta prob, mlabel(id)
```



Observations 766, 2189 stand out again as the ones with highest values of dbeta
Can similarly examine dx2 and hat values

We can also combine the information about multiple leverage statistics in one plot:

```
. scatter dbeta rs [w=dx2], mfc(white) xline(0)
```



Again those two observations (we can verify that they are the same ones by using mlabel option). These observations definitely warrant investigation – we need to figure out what's special about them and then decide how to deal with them.

5. Error term distribution

In terms of the error term distribution, we don't check for it directly (like with heteroscedasticity test in OLS). There is in-built heteroscedasticity in logit models – the variance of the error term is the greatest at the predicted probabilities around .5 and the smallest as we approach 0 or 1. But we still should be concerned whether the logit assumptions about the variance of the error term are correct. To test that, we can obtain robust standard error estimates and compare them with the regular standard error estimates. If they are similar, then our logistic results are fine. If they differ a lot, however, we would rather report robust standard errors as they do are correct even in the presence of assumptions violation.

```
. logit marijuana sex age educ childs
Logistic regression                                     Number of obs   =        845
                                                        LR chi2(4)      =      54.35
                                                        Prob > chi2     =     0.0000
Log likelihood = -524.84843                             Pseudo R2       =     0.0492
------------------------------------------------------------------------------
   marijuana |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         sex |    -.34799   .1494796    -2.33   0.020    -.6409647   -.0550152
         age |  -.0183109   .0049147    -3.73   0.000    -.0279436   -.0086782
        educ |   .0401891    .025553     1.57   0.116     -.009894    .0902722
      childs |  -.1696747   .0536737    -3.16   0.002    -.2748733   -.0644762
       _cons |   .5412516   .4595609     1.18   0.239    -.3594713    1.441974
------------------------------------------------------------------------------
```

```
. logit marijuana sex age educ childs, robust
Logistic regression                                     Number of obs   =        845
                                                        Wald chi2(4)    =      44.52
                                                        Prob > chi2     =     0.0000
Log pseudolikelihood = -524.84843                       Pseudo R2       =     0.0492
------------------------------------------------------------------------------
             |               Robust
   marijuana |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         sex |    -.34799    .149609    -2.33   0.020    -.6412182   -.0547617
         age |  -.0183109   .0048417    -3.78   0.000    -.0278003   -.0088214
        educ |   .0401891   .0269052     1.49   0.135    -.0125441    .0929223
      childs |  -.1696747   .0566388    -3.00   0.003    -.2806846   -.0586648
       _cons |   .5412516   .4677331     1.16   0.247    -.3754884    1.457992
------------------------------------------------------------------------------
```

The two sets of standard errors look the same – no violation of assumptions about error distribution.

6. Overdispersion

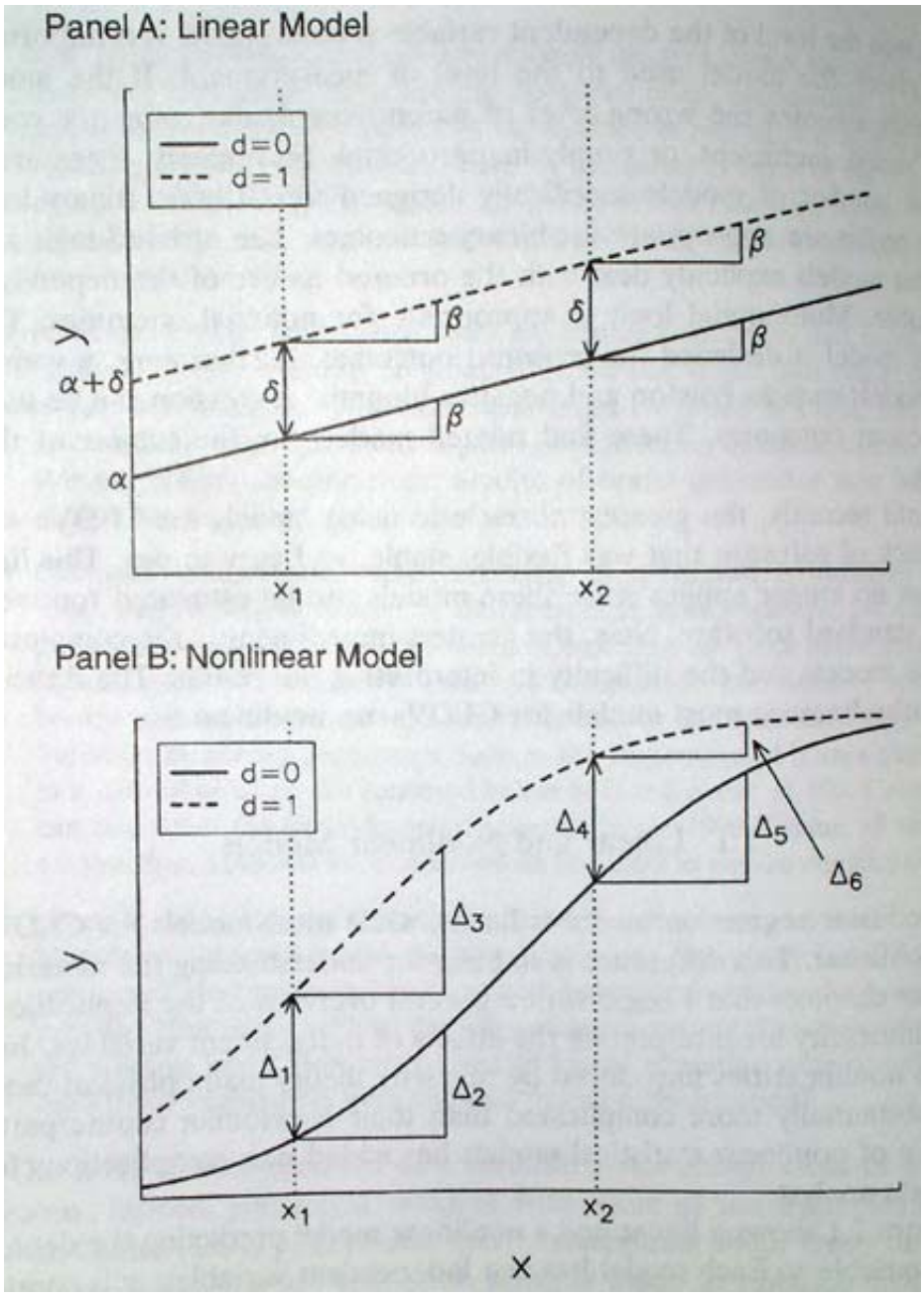
In logistic regression, the expected variance of the dependent variable can be compared to the observed variance, and discrepancies may be considered under- or overdispersion. If there is substantial discrepancy, standard errors will be over-optimistic. The expected variance is ybar*(1 - ybar), where ybar is the mean of the fitted values. This can be compared with the actual variance in observed DV to assess under- or overdispersion. We can see the extent of overdispersion by examining the ratio of D/df (where D is the deviance (-2LL) and df=N-k) -- given that we eliminated other reasons for deviance to be large (e.g., outliers, nonlinearities, other model specification errors like omitted variables). In the fitstat output, we find D(df=840) is 1049.697. The ratio is

```
. di 1049.697/840
1.2496393
```

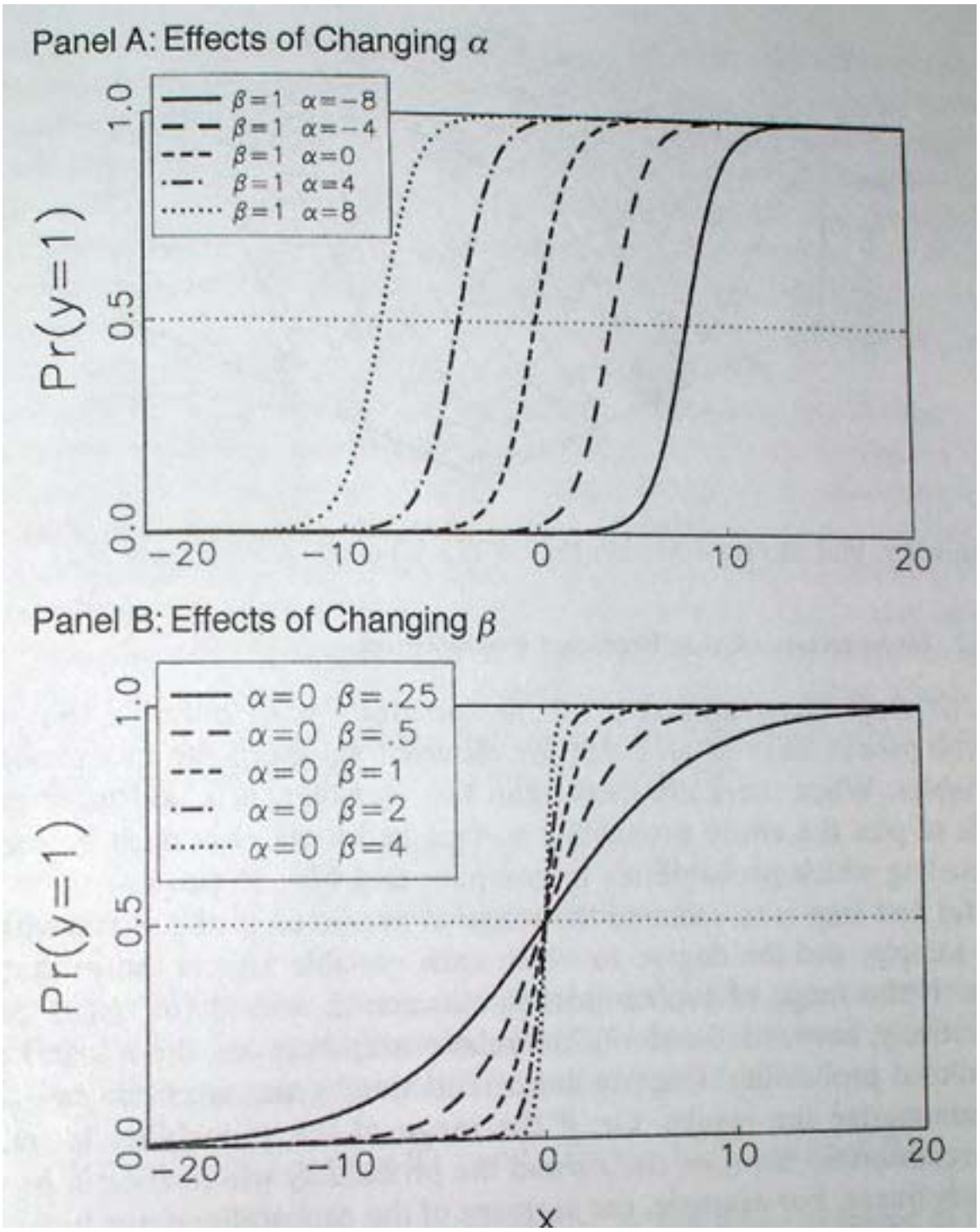
The ratio is close enough to 1 for us not to worry. If there is overdispersion (which is much more common than underdispersion), we can use adjusted standard errors. Adjusted standard errors will make the confidence intervals wider. Adjusted SE equals SE * sqrt(D/df), where D is the deviance (-2LL) and df=N-k. However, typically overdispersion reflects the fact that we need to respecify the model (i.e. we omitted an important variable), or that our observations are not independent – i.e., data over time or clusters of observations. We'll discuss methods to deal with clusters of observation later in the course.

Binary Logit Interpretation

As logistic regression models (whether binary, ordered, or multinomial) are nonlinear, they pose a challenge for interpretation. The increase in the dependent variable in a linear model is constant for all values of X. Not so for logit models – probability increases or decreases per unit change in X is nonconstant, as illustrated in this picture.



When interpreting logit regression coefficients, we can interpret only the sign and significance of the coefficients – cannot interpret the size. The following picture can give you an idea how the shape of the curve varies depending on the size of the coefficient, however. Note that, similarly to OLS regression, the constant determines the position of the curve along the X axis and the coefficient (beta) determines the slope.


Panel A: Effects of Changing α
β=1 α=−8
β=1 α=−4
β=1 α=0
β=1 α=4
β=1 α=8
Pr(y=1)
1.0
0.5
0.0
−20
−10
0
10
20
Panel B: Effects of Changing β
α=0 β=.25
α=0 β=.5
α=0 β=1
α=0 β=2
α=0 β=4
Pr(y=1)
1.0
0.5
0.0
−20
−10
0
10
20
X

Next, we'll examine various ways to interpret logistic regression results.

1. Coefficients and Odds Ratios

We'll use another model, focusing now on the probability of voting.

```
. codebook vote00
-------------------------------------------------------------------------------
vote00
did r vote in 2000 election
-------------------------------------------------------------------------------
                  type:  numeric (byte)
                 label:  vote00

                 range:  [1,4]                        units:  1
         unique values:  4                        missing .:  14/2765

            tabulation:  Freq.   Numeric  Label
                         1780         1  voted
                          822         2  did not vote
                          138         3  ineligible
                           11         4  refused to answer
                           14         .

. gen vote=(vote00==1) if vote00<3
(163 missing values generated)
. gen married=(marital==1)

. logit vote age sex born married childs educ
Iteration 0:   log likelihood = -1616.8899
Iteration 1:   log likelihood = -1365.9814
Iteration 2:   log likelihood = -1353.4091
Iteration 3:   log likelihood = -1353.2224
Iteration 4:   log likelihood = -1353.2224
Logistic regression                               Number of obs   =       2590
                                                  LR chi2(6)      =     527.33
                                                  Prob > chi2     =     0.0000
Log likelihood = -1353.2224                       Pseudo R2       =     0.1631
------------------------------------------------------------------------------
        vote |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   .0466321    .003337    13.97   0.000     .0400917    .0531726
         sex |   .1094233     .09552     1.15   0.252    -.0777924     .296639
        born |  -.9673683   .1859278    -5.20   0.000     -1.33178   -.6029564
     married |   .4911099   .0983711     4.99   0.000     .2983062    .6839136
      childs |  -.0391447   .0327343    -1.20   0.232    -.1033028    .0250133
        educ |   .2862839   .0197681    14.48   0.000     .2475391    .3250287
       _cons |  -4.352327   .3892601   -11.18   0.000    -5.115263   -3.589391
------------------------------------------------------------------------------
```

These are regular logit coefficients; so we can interpret the sign and significance but not the size of effects. So we can say that age increases the probability of voting but we can't say by how much – that's because a 1 year increase in age will not affect the probability the same way for a 30 year old and for a 40 year old.

To be able to interpret effect size, we turn to odds ratios. Note that odds ratios are only appropriate for logistic regression – they don't work for probit models.

Odds are ratios of two probabilities – probability of a positive outcome and a probability of a negative outcome (e.g. probability of voting divided by a probability of not voting). But since probabilities vary depending on values of X, such a ratio varies as well. What remains constant is the ratio of such odds – e.g. odds of voting for women divided by odds of voting for men will be the same number regardless of the values of other variables. Similarly, the odds ratio for age can be a ratio of the odds of voting for someone who is 31 y.o. to the odds of a 30 y.o. person, or of a 41 y.o. to a 40 y.o. person's odds – these will be the same regardless of what age values you pick, as long as they are one year apart. So let's examine the odds ratios.

```
. logit vote age sex born married childs educ, or
Iteration 0:   log likelihood = -1616.8899
Iteration 1:   log likelihood = -1365.9814
Iteration 2:   log likelihood = -1353.4091
Iteration 3:   log likelihood = -1353.2224
Iteration 4:   log likelihood = -1353.2224
Logistic regression                               Number of obs   =       2590
                                                  LR chi2(6)      =     527.33
                                                  Prob > chi2     =     0.0000
Log likelihood = -1353.2224                       Pseudo R2       =     0.1631
------------------------------------------------------------------------------
        vote | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   1.047736   .0034963    13.97   0.000     1.040906    1.054612
         sex |   1.115634   .1065654     1.15   0.252     .9251564     1.34533
        born |    .380082   .0706678    -5.20   0.000     .2640069    .5471915
     married |   1.634129    .160751     4.99   0.000     1.347574    1.981618
      childs |   .9616115   .0314777    -1.20   0.232     .9018538    1.025329
        educ |    1.33147   .0263207    14.48   0.000     1.280869     1.38407
------------------------------------------------------------------------------
```

Another way to obtain odds ratios would be to use "logistic" command instead of "logit" – it automatically displays odds ratios instead of coefficients. But yet another, more convenient way is to use listcoef command (that's one of the commands written by Scott Long that we downloaded as a part of spost package):

```
. listcoef
logit (N=2590): Factor Change in Odds
  Odds of: 1 vs 0
------------------------------------------------------------------------
        vote |      b         z     P>|z|    e^b    e^bStdX      SDofX
-------------+----------------------------------------------------------
         age |   0.04663   13.974   0.000   1.0477   2.2297    17.1953
         sex |   0.10942    1.146   0.252   1.1156   1.0559     0.4972
        born |  -0.96737   -5.203   0.000   0.3801   0.7885     0.2457
     married |   0.49111    4.992   0.000   1.6341   1.2777     0.4990
      childs |  -0.03914   -1.196   0.232   0.9616   0.9365     1.6762
        educ |   0.28628   14.482   0.000   1.3315   2.3108     2.9257
------------------------------------------------------------------------
```

The advantage of listcoef is that it reports regular coefficients, odds ratios, and standardized odds ratios in one table.

Odds ratios are exponentiated logistic regression coefficients. They are sometimes called factor coefficients, because they are multiplicative coefficients. Odds ratios are equal to 1 if there is no effect, smaller than 1 if the effect is negative and larger than 1 if it is positive. So for example, the odds ratio for married indicates that the odds of voting for those who are

married are 1.63 times higher than for those who are not married. And the odds ratio for education indicates that each additional year of education makes one's odds of voting 1.33 times higher -- or, in other words, increases those odds by 33%. To get percent change directly, we can use percent option:

```
. listcoef, percent
logit (N=2590): Percentage Change in Odds
  Odds of: 1 vs 0
------------------------------------------------------------------------
        vote |      b         z     P>|z|      %      %StdX      SDofX
-------------+----------------------------------------------------------
         age |   0.04663   13.974   0.000      4.8    123.0    17.1953
         sex |   0.10942    1.146   0.252     11.6      5.6     0.4972
        born |  -0.96737   -5.203   0.000    -62.0    -21.2     0.2457
     married |   0.49111    4.992   0.000     63.4     27.8     0.4990
      childs |  -0.03914   -1.196   0.232     -3.8     -6.4     1.6762
        educ |   0.28628   14.482   0.000     33.1    131.1     2.9257
------------------------------------------------------------------------
```

Beware: if you would like to know what the increase would be per, say, 10 units increase in the independent variable – e.g. 10 years of education, you cannot simply multiple the odds ratio by 10! The coefficient, in fact, would be odds ratio to the power of 10. Or alternatively, you could take the regular logit coefficient, multiply it by 10 and then exponentiate it -- e.g. for education:

```
. di exp(0.28628*10)
17.510488
. di 1.3315^10
17.515063
```

Standardized odds ratios (presented under e^bStdX) are similar to regular odds ratios, but they display the change in the odds of voting per one standard deviation change in the independent variable. The last column in the table generated by listcoef shows what one standard deviation for each variable is. So for age the standardized odds ratio indicates that 17 years of age increase one's odds of voting 2.23 times, or by 123%. Standardized odds ratios, like standardized coefficients in OLS, allow us to compare effect sizes across variables regardless of their measurement units. But, beware of comparing negative and positive effects – odds ratios of 1.5 and .5 are not equivalent, even though the first one represents a 50% increase in odds and the second one represents a 50% decrease. This is because odds ratios cannot be below zero (there cannot be a decrease more than 100%), but they do not have an upper bound – i.e. can be infinitely high. In order to be able to compare positive and negative effects, we can reverse odds ratios and generate odds ratios for odds of not voting (rather than odds of voting).

```
. listcoef, reverse
logit (N=2590): Factor Change in Odds
  Odds of: 0 vs 1
------------------------------------------------------------------------
        vote |      b         z     P>|z|    e^b    e^bStdX      SDofX
-------------+----------------------------------------------------------
         age |   0.04663   13.974   0.000   0.9544   0.4485    17.1953
         sex |   0.10942    1.146   0.252   0.8964   0.9470     0.4972
        born |  -0.96737   -5.203   0.000   2.6310   1.2682     0.2457
     married |   0.49111    4.992   0.000   0.6119   0.7826     0.4990
      childs |  -0.03914   -1.196   0.232   1.0399   1.0678     1.6762
        educ |   0.28628   14.482   0.000   0.7510   0.4328     2.9257
```

We can see for example that the odds ratio of 0.3801 for born is a negative effect corresponding in size to a positive odds ratio of 2.6310.

Listcoef also has a help option that explains what's what in the table:

```
. listcoef, reverse help
logit (N=2590): Factor Change in Odds
  Odds of: 0 vs 1
------------------------------------------------------------------------------
        vote |      b         z     P>|z|    e^b    e^bStdX      SDofX
-------------+----------------------------------------------------------------
         age |   0.04663   13.974   0.000   0.9544   0.4485    17.1953
         sex |   0.10942    1.146   0.252   0.8964   0.9470     0.4972
        born |  -0.96737   -5.203   0.000   2.6310   1.2682     0.2457
     married |   0.49111    4.992   0.000   0.6119   0.7826     0.4990
      childs |  -0.03914   -1.196   0.232   1.0399   1.0678     1.6762
        educ |   0.28628   14.482   0.000   0.7510   0.4328     2.9257
------------------------------------------------------------------------------
       b = raw coefficient
       z = z-score for test of b=0
   P>|z| = p-value for z-test
     e^b = exp(b) = factor change in odds for unit increase in X
 e^bStdX = exp(b*SD of X) = change in odds for SD increase in X
   SDofX = standard deviation of X
```

2. Predicted Probabilities

In addition to regular coefficients and odds ratios, we also should examine predicted probabilities – both for the actual observations in our data and for strategically selected hypothetical cases. Predicted probabilities are always calculated for a specific set of independent variables' values. One thing we can calculate is predicted probabilities for the actual data that we have – for each case, we take the values of all independent variables and plug it into the equation:

```
. predict prob
(option p assumed; Pr(vote))
(26 missing values generated)

. sum prob if e(sample)

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
        prob |      2590    .6833977     .204702   .0205784   .9926677
```

Mean of predicted probabilities represents the average proportion in the sample:

```
. sum vote if e(sample)

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
        vote |      2590    .6833977    .4652406          0          1
```

These are predicted probabilities for the actual cases in our dataset. It can be useful, however, to calculate predicted probabilities for hypothetical sets of values – some interesting combinations that we could compare and contrast.

```
. prvalue
logit: Predictions for vote
Confidence intervals by delta method
                                95% Conf. Interval
```

```
  Pr(y=1|x):          0.7249   [ 0.7052,    0.7446]
  Pr(y=0|x):          0.2751   [ 0.2554,    0.2948]
          age        sex       born    married     childs       educ
x=  46.935907  1.5532819  1.0644788  .46756757  1.8389961  13.394595
```

This calculates a predicted probability for a case with all values set at the mean. So an "average" person has 72.5% chance of voting. We can also see what these averages are. Clearly, for some variables they don't make sense – we don't want to use averages for dummy variables; rather, we'd want to specify what values to use. Here are some examples of specifying values:

```
. prvalue, x(age=30 born=1 sex=2 married=0)
logit: Predictions for vote
Confidence intervals by delta method
                               95% Conf. Interval
  Pr(y=1|x):          0.5152   [ 0.4722,    0.5582]
  Pr(y=0|x):          0.4848   [ 0.4418,    0.5278]
          age        sex       born    married     childs       educ
x=         30          2          1          0  1.8389961  13.394595
```

This is the predicted value for someone who is 30, native born, female, and unmarried (and has average number of children and average education).

Note that if you have a set of dummy variables, you should always specify values for each of them in prvalue command. E.g. if we were using 4 marital status dummies, we'd have to specify all of them, otherwise, some of them will be assigned their mean values and your calculation will be unrealistic.

```
. xi: qui logit vote age sex born i.marital childs educ
. prvalue, x( _Imarital_2=1  _Imarital_3=0  _Imarital_4=0  _Imarital_5=0)
logit: Predictions for vote
Confidence intervals by delta method
                               95% Conf. Interval
  Pr(y=1|x):          0.6736   [ 0.5908,    0.7565]
  Pr(y=0|x):          0.3264   [ 0.2435,    0.4092]
      age          sex         born        _Imarital_2  _Imarital_3  _Imarital_4
_Imarital_5   childs   educ
x=    46.935907    1.5532819    1.0644788            1            0            0
0    1.8389961    13.394595
Note: to get the predicted probability for the omitted category, we need to
specify all zeros.
```

We can also use prtab to obtain values of predicted probabilities for various combinations of categorical variables – we can select one variable at a time or up to four variables in this command – but note that we need to specify what values to use for all other variables – e.g. in this case, all other variables are set at the mean.

```
. qui logit vote age sex born married childs educ
. prtab born married, rest(mean)
logit: Predicted probabilities of positive outcome for vote
--------------------------
was r     |
born in   |
this      |    married
country   |      0       1
----------+---------------
      yes | 0.6903  0.7846
```

```
       no | 0.4587  0.5806
--------------------------
          age        sex       born    married     childs       educ
x=  46.935907  1.5532819  1.0644788  .46756757  1.8389961  13.394595
```

This allows us to see that the effect of one variable depends on the level of the other – for native born individuals, marriage increases chances of voting by 9.5%, but for the foreign born, marriage increases these chances by 12.2%.

And we can use conditions:

```
. prtab childs born if married ==1
logit: Predicted probabilities of positive outcome for vote
------------------------------
              | was r born in
number of     |  this country
children      |    yes      no
--------------+---------------
         none | 0.8153  0.6265
          one | 0.8093  0.6173
          two | 0.8032  0.6080
        three | 0.7969  0.5987
         four | 0.7905  0.5892
         five | 0.7840  0.5797
          six | 0.7773  0.5702
        seven | 0.7704  0.5605
eight or more | 0.7634  0.5509
------------------------------
          age        sex       born    married     childs       educ
x=  48.010735  1.5111478  1.0817506          1  2.1965318  13.654831
```

But note that the means used in this case are the means for the subgroup specified by these conditions (in this case, for the married). If you want to use the means for the whole sample, you'd have to specify them using x option:

```
. prtab childs born if married ==1, x(age=46.935907 sex=1.5532819 educ=
13.394595)
logit: Predicted probabilities of positive outcome for vote
------------------------------
              | was r born in
number of     |  this country
children      |    yes      no
--------------+---------------
         none | 0.7965  0.5981
          one | 0.7901  0.5886
          two | 0.7835  0.5791
        three | 0.7768  0.5695
         four | 0.7700  0.5599
         five | 0.7630  0.5502
          six | 0.7558  0.5405
        seven | 0.7485  0.5308
eight or more | 0.7411  0.5210
------------------------------
          age        sex       born    married     childs       educ
x=  46.935907  1.5532819  1.0817506          1  2.1965318  13.394595
```

Note that it only makes sense to create such tables of predicted probabilities for variables that have significant effects – otherwise, you'll see no differences. And if you have sets of dummy variables, you are better off using

prvalue to obtain your predicted values (see above); prtab can be quite confusing for such cases.

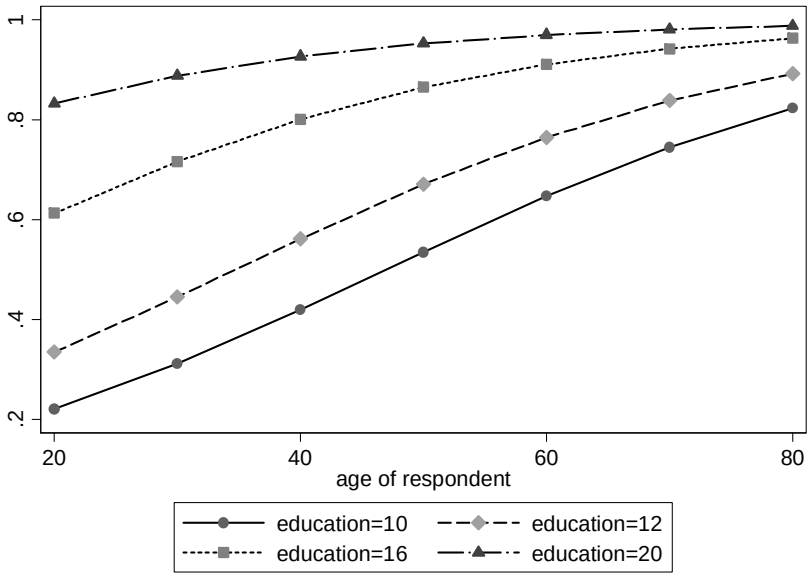
Further, we can use prgen to generate new variables containing probabilities for certain sets of values. This is useful with continuous variables, as it allows us to see how predicted probability changes across values of one variable (given that the rest of them are set at some specific values).

In the following example, we generate predicted values for 7 different ages -- 20, 80, and 5 more points in between. We generate these for four groups defined by education (10, 12, 16, 20). The rest of the variables are set at mean. We'll add labels to the new variables containing predicted probabilities.

```
. for num 10 12 16 20: prgen age, from (20) to (80) gen(preducX) x(educ=X)
rest(mean) n(7) \ lab var preducXp1 "education=X"

->  prgen age, from (20) to (80) gen(preduc10) x(educ=10) rest(mean) n(7)

logit: Predicted values as age varies from 20 to 80.

          age        sex       born    married     childs       educ
x=  46.935907  1.5532819  1.0644788  .46756757  1.8389961         10

->  lab var preduc10p1 `"education=10"'

->  prgen age, from (20) to (80) gen(preduc12) x(educ=12) rest(mean) n(7)

logit: Predicted values as age varies from 20 to 80.

          age        sex       born    married     childs       educ
x=  46.935907  1.5532819  1.0644788  .46756757  1.8389961         12

->  lab var preduc12p1 `"education=12"'

->  prgen age, from (20) to (80) gen(preduc16) x(educ=16) rest(mean) n(7)

logit: Predicted values as age varies from 20 to 80.

          age        sex       born    married     childs       educ
x=  46.935907  1.5532819  1.0644788  .46756757  1.8389961         16

->  lab var preduc16p1 `"education=16"'

->  prgen age, from (20) to (80) gen(preduc20) x(educ=20) rest(mean) n(7)

logit: Predicted values as age varies from 20 to 80.

          age        sex       born    married     childs       educ
x=  46.935907  1.5532819  1.0644788  .46756757  1.8389961         20

->  lab var preduc20p1 `"education=20"'
```

Now we can plot four curves that show how probability of voting changes by age for an average person who has 10, 12, 16, or 10 years of education.

```
. graph twoway connected preduc10p1 preduc12p1 preduc16p1 preduc20p1 preduc20x
```

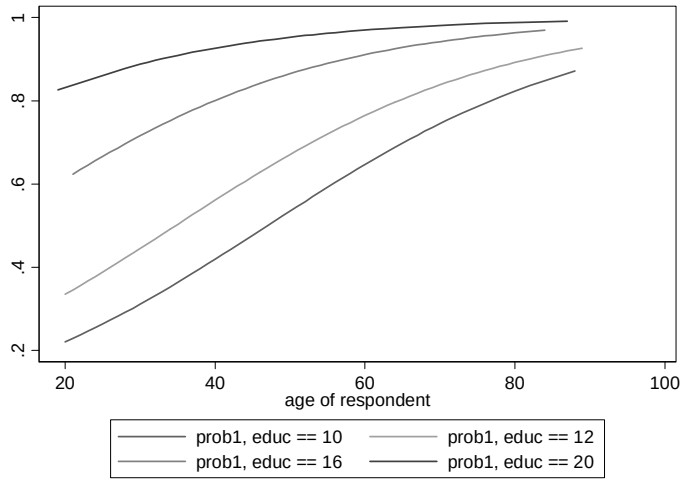


If there are interactions or nonlinearities that required that you entered a variable more than once (e.g. X and X squared), you can use adjust command to do the graphs. This is done in the same manner as we did in OLS, but we need to use pr option to get probabilities rather than linear prediction (xb). This is the best way to examine what interactions mean in logit models, because their value For example we can replicate our previous graph. We run adjust command omitting age and educ:

```
. adjust sex born married childs if e(sample), gen(prob1) pr
------------------------------------------------------------------------------
Dependent variable: vote     Command: logit
       Created variable: prob1
   Variables left as is: age, educ
 Covariates set to mean: sex = 1.5532819, born = 1.0644788, married = .46756756,
childs = 1.8389962
------------------------------------------------------------------------------
------------------------------
      All |         pr
----------+-------------------
          |    .724903
----------------------
     Key:  pr  =  Probability

. separate prob1, by(educ)
              storage  display     value
variable name   type   format      label      variable label
------------------------------------------------------------------------------
prob10          float  %9.0g                  prob1, educ == 0
prob11          float  %9.0g                  prob1, educ == 1
prob12          float  %9.0g                  prob1, educ == 2
prob13          float  %9.0g                  prob1, educ == 3
prob14          float  %9.0g                  prob1, educ == 4
prob15          float  %9.0g                  prob1, educ == 5
prob16          float  %9.0g                  prob1, educ == 6
prob17          float  %9.0g                  prob1, educ == 7
prob18          float  %9.0g                  prob1, educ == 8
```

```
prob19          float  %9.0g                 prob1, educ == 9
prob110         float  %9.0g                 prob1, educ == 10
prob111         float  %9.0g                 prob1, educ == 11
prob112         float  %9.0g                 prob1, educ == 12
prob113         float  %9.0g                 prob1, educ == 13
prob114         float  %9.0g                 prob1, educ == 14
prob115         float  %9.0g                 prob1, educ == 15
prob116         float  %9.0g                 prob1, educ == 16
prob117         float  %9.0g                 prob1, educ == 17
prob118         float  %9.0g                 prob1, educ == 18
prob119         float  %9.0g                 prob1, educ == 19
prob120         float  %9.0g                 prob1, educ == 20
```

```
. line prob110 prob112 prob116 prob120 age, sort
```



3. Changes in Predicted Probabilities

Another way to interpret logistic regression results is using changes in predicted probabilities. These are changes in probability of the outcome as one variable changes, holding all other variables constant at certain values. There are two ways to measure such changes – discrete change and marginal effect.

A. Discrete change

Discrete change is a change in predicted probabilities corresponding to a given change in the independent variable. To obtain these, we calculate two probabilities and then calculate the difference between them. These can be obtained using prvalue command, but it is much easier to do using prchange:

```
. prchange
logit: Changes in Probabilities for vote
         min->max      0->1     -+1/2    -+sd/2  MargEfct
    age    0.5320    0.0083    0.0093    0.1591    0.0093
    sex    0.0219    0.0229    0.0218    0.0109    0.0218
   born   -0.2212   -0.1435   -0.1914   -0.0474   -0.1929
married    0.0970    0.0970    0.0977    0.0489    0.0979
 childs   -0.0647   -0.0076   -0.0078   -0.0131   -0.0078
   educ    0.8920    0.0166    0.0571    0.1661    0.0571
              0         1
Pr(y|x)  0.2751  0.7249

            age      sex     born  married   childs     educ
```

```
    x=  46.9359  1.55328  1.06448  .467568    1.839  13.3946
sd(x)=  17.1953  .497249  .245651  .499043  1.67616  2.92567
```

Here we can see how probability changes when we go from the minimum value of each variable, e.g. education, to its maximum, how it changes when we go from 0 to 1, how it changes per one unit at the mean (that is displayed as -+1/2 because it calculates the differences between mean-1 and mean+1, and then divides it by 2. Then there is the change per one standard deviation, also around the mean. We can also get a clear explanation of what's what using help option:

```
. prchange, help
logit: Changes in Probabilities for vote
         min->max      0->1     -+1/2    -+sd/2  MargEfct
    age    0.5320    0.0083    0.0093    0.1591    0.0093
    sex    0.0219    0.0229    0.0218    0.0109    0.0218
   born   -0.2212   -0.1435   -0.1914   -0.0474   -0.1929
married    0.0970    0.0970    0.0977    0.0489    0.0979
 childs   -0.0647   -0.0076   -0.0078   -0.0131   -0.0078
   educ    0.8920    0.0166    0.0571    0.1661    0.0571

              0       1
Pr(y|x)  0.2751  0.7249

            age      sex     born  married   childs     educ
    x=  46.9359  1.55328  1.06448  .467568    1.839  13.3946
sd(x)=  17.1953  .497249  .245651  .499043  1.67616  2.92567

 Pr(y|x): probability of observing each y for specified x values
Avg|Chg|: average of absolute value of the change across categories
Min->Max: change in predicted probability as x changes from its minimum to
          its maximum
    0->1: change in predicted probability as x changes from 0 to 1
   -+1/2: change in predicted probability as x changes from 1/2 unit below
          base value to 1/2 unit above
  -+sd/2: change in predicted probability as x changes from 1/2 standard
          dev below base to 1/2 standard dev above
MargEfct: the partial derivative of the predicted probability/rate with
          respect to a given independent variable
```

We can also run prchange with fromto option to get starting and ending probabilities in addition to the amount of change:

```
. prchange, fromto
logit: Changes in Probabilities for vote
          from:      to:      dif:     from:      to:      dif:     from:      to:      dif:     from:      to:      dif:
          x=min     x=max  min->max       x=0      x=1      0->1     x-1/2     x+1/2     -+1/2   x-1/2sd   x+1/2sd    -+sd/2
   age   0.4173    0.9493    0.5320    0.2280    0.2363    0.0083    0.7202    0.7295    0.0093    0.6383    0.7974    0.1591
   sex   0.7127    0.7345    0.0219    0.6897    0.7127    0.0229    0.7139    0.7357    0.0218    0.7194    0.7303    0.0109
  born   0.7372    0.5160   -0.2212    0.8807    0.7372   -0.1435    0.8104    0.6190   -0.1914    0.7480    0.7006   -0.0474
married  0.6768    0.7739    0.0970    0.6768    0.7739    0.0970    0.6733    0.7711    0.0977    0.6998    0.7487    0.0489
 childs  0.7390    0.6743   -0.0647    0.7390    0.7314   -0.0076    0.7288    0.7210   -0.0078    0.7314    0.7183   -0.0131
   educ  0.0539    0.9458    0.8920    0.0539    0.0705    0.0166    0.6955    0.7525    0.0571    0.6342    0.8002    0.1661

         MargEfct
    age    0.0093
    sex    0.0218
   born   -0.1929
married    0.0979
```

```
 childs   -0.0078
   educ    0.0571

             0       1
Pr(y|x)  0.2751  0.7249

            age      sex     born  married   childs     educ
    x=  46.9359  1.55328  1.06448  .467568    1.839  13.3946
sd(x)=  17.1953  .497249  .245651  .499043  1.67616  2.92567
```

We can customize the amount of change in X using delta option, set the value of X to whatever we want, and we can also select uncentered option if we don't want our selected interval to be centered at X but would rather prefer it to start at X. For example, with and without uncentered option:

```
. prchange educ, x(educ=16) delta(4) uncentered

logit: Changes in Probabilities for vote

(Note: delta = 4)

      min->max      0->1    +delta       +sd  MargEfct
educ    0.8920    0.0166    0.0984    0.0803    0.0370

             0       1
Pr(y|x)  0.1525  0.8475

            age      sex     born  married   childs     educ
    x=  46.9359  1.55328  1.06448  .467568    1.839       16
sd(x)=  17.1953  .497249  .245651  .499043  1.67616  2.92567

. prchange educ, x(educ=16) delta(4)
logit: Changes in Probabilities for vote

(Note: d = 4)

      min->max      0->1     -+d/2    -+sd/2  MargEfct
educ    0.8920    0.0166    0.1497    0.1090    0.0370

             0       1
Pr(y|x)  0.1525  0.8475

            age      sex     born  married   childs     educ
    x=  46.9359  1.55328  1.06448  .467568    1.839       16
sd(x)=  17.1953  .497249  .245651  .499043  1.67616  2.92567
```

B. Marginal effects.

The last column of prchange output presents marginal effects – these are partial derivatives, slopes of probability curve at a certain set of values of independent variables. Marginal effects, of course, vary along X; they are the largest at the value of X that corresponds to P(Y=1|X)=.5 – this can be seen in the graph.

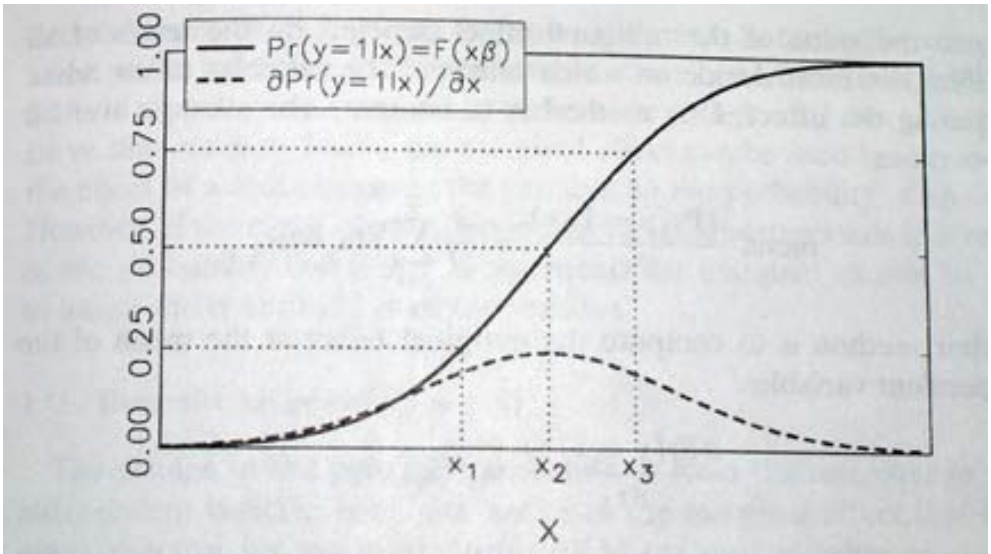


Figure 3.12. Marginal Effect in the Binary Response Model

Usually, if marginal effects are presented in journal articles, they are evaluated with all variables held at their means. In case of logistic regression, marginal effect for X can be calculated as P(Y=1|X)*P(Y=0|X)*b; For example, we can replicate the last result,

```
di 0.1525*0.8475*0.28628
.0369999
```

The following graph compares a marginal change and a discrete change at a specific point:

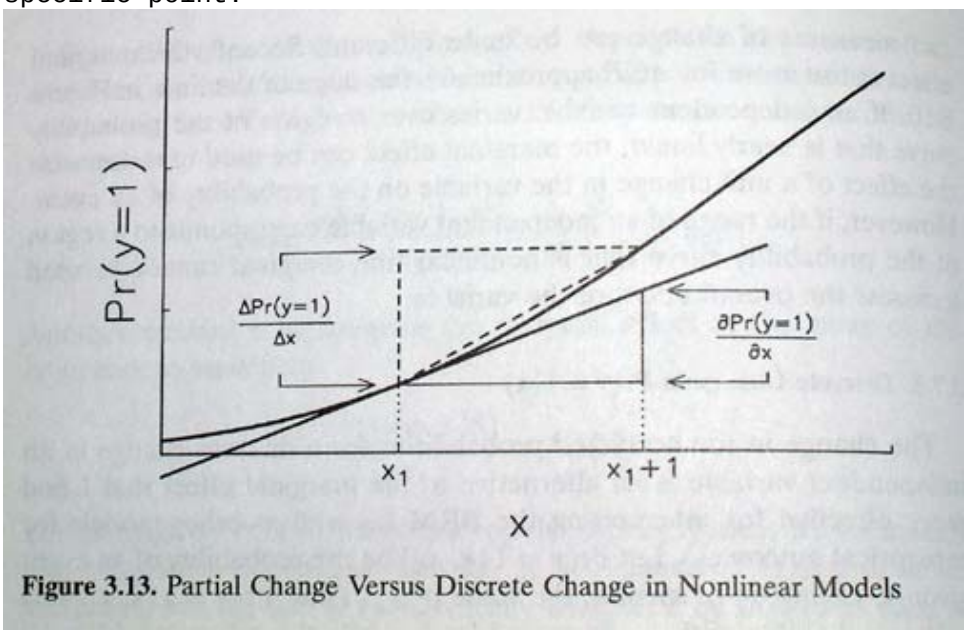


Figure 3.13. Partial Change Versus Discrete Change in Nonlinear Models

We can also generate marginal effects with standard errors using mfx compute. Computing those standard errors can take a while, however.

```
. mfx compute
Marginal effects after logit
      y  = Pr(vote) (predict)
         =  .72490265
------------------------------------------------------------------------------
variable |      dy/dx    Std. Err.     z    P>|z|  [    95% C.I.   ]      X
---------+--------------------------------------------------------------------
     age |   .0092993      .00064   14.50   0.000   .008042  .010556   46.9359
     sex |   .0218211      .01905    1.15   0.252  -.015522  .059164   1.55328
    born |  -.1929114      .03711   -5.20   0.000  -.265655 -.120167   1.06448
 married*|   .0970482       .0192    5.05   0.000   .059412  .134684   .467568
  childs |  -.0078062      .00653   -1.20   0.232  -.020596  .004984     1.839
    educ |   .0570904      .00382   14.96   0.000    .04961  .064571   13.3946
------------------------------------------------------------------------------
(*) dy/dx is for discrete change of dummy variable from 0 to 1
```

Marginal effects are inappropriate for binary independent variables; that's why discrete changes are reported for those instead.

We could also specify other values of X for this computation using "at" option:

```
. mfx compute, at(age=30)
[output omitted]
```

Note: For binary dependent variables, though, marginal effects are not very useful – discrete changes are more easily interpretable.

Also note that marginal effects in models with interactions or higher order terms are complicated to estimate. To learn more about that, you can consult http://www.stata.com/support/faqs/stat/mfx_interact.html
and
http://www.unc.edu/~enorton/NortonWangAi.pdf

And to learn more about interactions in logistic models:
http://www.ats.ucla.edu/stat/stata/seminars/stata_vibl/

Binary Logit Article Example:

Alba, Richard, John Logan, Amy Lutz, and Brian Stults. 2002. "Only English by the Third Generation? Loss and Preservation of the Mother Tongue among the Grandchildren of Contemporary Immigrants." *Demography, 39*: 467-484.

Questions to answer about the article:
1. What are the dependent and the independent variables in this analysis?
2. What is reported in Table 4? How can we interpret these results? How do the authors discuss these results in the text?
3. What is reported in Table 5? How can we interpret these results?
4. In addition to what the authors chose to present, how else could they have presented their results?
5. What measures of model fit and model diagnostics are presented? What diagnostics and potential problems did the authors not address?