Sociology 7704: Regression Models for Categorical Data
Instructor: Natasha Sarkisian

Binary Logit: Introduction, Measures of Fit, and Diagnostics

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, and with regard to the resulting curves, the probit curve approaches 1 and -1 more quickly than the logit curve, 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 because logit has better interpretation than probit-- logistic regression can be interpreted as modeling log odds, also known as logits:

$$\log\left(\frac{p_i}{1-p_i}\right) = \alpha + \beta * X_i$$

Solving this equation back to get back to probabilities, we would get p = $e^{Xb}/(1+e^{Xb})$.

You could also use the log likelihood value from estimating both models or other measures of fit such as BIC or AIC (we will discuss them soon) to decide between logit or probit, but again, typically people just run one of them.

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

. 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
------------------------------------------------------------------------------
```

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 kids are less likely to support legalization. Divorced people are more likely than the married 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
------------------------------------------------------------------------------
```

In the probit model, residuals are assumed to be normally distributed, with a mean of zero and a variance of σ^2. However, while in OLS, we can get an actual unbiased estimate of σ^2, in probit (and logit), σ^2 is not identified – in fact we can only get estimates of ratios of coefficients to error variance (β/σ) but not independent estimates of each. That is, we know the effect of gender on one's views on marijuana legalization relative to the remaining (unexplained) dispersion of views on marijuana legalization on the population. To deal with that, in probit, we always make $\sigma^2 = 1$. In logit, the problem of model identification is the same, but the variance of residuals is fixed, also by convention, to $\pi^2/3$. And the distribution of residuals is assumed to be binomial rather than normal.

Hypothesis testing in logit models

In logit models, like in OLS models, we might need to test hypotheses about coefficients being jointly zero, or to compare if coefficients are equal to each other; once again, we can use test command:

```
. logit marijuana sex age educ childs i.marital
Iteration 0:   log likelihood =  -552.0232
Iteration 1:   log likelihood = -515.19453
Iteration 2:   log likelihood = -514.62744
Iteration 3:   log likelihood = -514.62716
Iteration 4:   log likelihood = -514.62716
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
               |
       marital |
      widowed  |   .0118099   .3568915     0.03   0.974    -.6876845    .7113043
     divorced  |   .9025573   .2053011     4.40   0.000     .5001746     1.30494
    separated  |   .0300665   .4239309     0.07   0.943    -.8008229    .8609558
never married  |   .2853992    .208832     1.37   0.172     -.123904    .6947024
               |
         _cons |   .2573784   .5195598     0.50   0.620    -.7609401    1.275697
--------------------------------------------------------------------------------

. test 2.marital 3.marital 4.marital 5.marital
 ( 1)  [marijuana]2.marital = 0
 ( 2)  [marijuana]3.marital = 0
 ( 3)  [marijuana]4.marital = 0
 ( 4)  [marijuana]5.marital = 0

           chi2(  4) =   20.55
         Prob > chi2 =    0.0004
```

When examining whether variables can be omitted as a group, we can also store our estimates and use likelihood ratio test:

```
. est store full

. logit marijuana sex age educ childs

Iteration 0:   log likelihood =  -552.0232
Iteration 1:   log likelihood = -525.10107
Iteration 2:   log likelihood = -524.84844
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   .1494797    -2.33   0.020    -.6409648   -.0550151
         age |  -.0183109   .0049147    -3.73   0.000    -.0279436   -.0086782
        educ |   .0401891   .0255531     1.57   0.116     -.009894    .0902722
      childs |  -.1696747   .0536738    -3.16   0.002    -.2748733    -.064476
       _cons |   .5412517   .4595611     1.18   0.239    -.3594716    1.441975
------------------------------------------------------------------------------

. lrtest . full

Likelihood-ratio test                                 LR chi2(4)  =     20.44
(Assumption: . nested in full)                        Prob > chi2 =    0.0004
```

Typically, these two approaches produce very similar results.

Goodness of fit

While in OLS we primarily rely on R^2 and adjusted R^2 to assess model fit, there are many alternative ways to assess fit for a logit model.

```
. qui logit marijuana sex educ age childs
. 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 to install `spost13_ado from http://www.indiana.edu/~jslsoc/stata`

Now let's obtain fit statistics for our last model:

```
. fitstat, save
                         |       logit
-------------------------+-------------
Log-likelihood           |
                   Model |    -524.848
          Intercept-only |    -552.023
-------------------------+-------------
Chi-square               |
       Deviance (df=840) |    1049.697
             LR (df=4)   |      54.350
               p-value   |       0.000
-------------------------+-------------
R2                       |
                McFadden |       0.049
     McFadden (adjusted) |       0.040
      McKelvey & Zavoina |       0.090
            Cox-Snell/ML |       0.062
  Cragg-Uhler/Nagelkerke |       0.085
                   Efron |       0.065
              Tjur's D   |       0.063
                   Count |       0.204
        Count (adjusted) |      -1.212
-------------------------+-------------
IC                       |
                     AIC |    1059.697
        AIC divided by N |       1.254
            BIC (df=5)   |    1083.394
-------------------------+-------------
Variance of              |
                       e |       3.290
                  y-star |       3.615
```

See pp. 120-130 of Long and Freese for details on these measures of fit. McFadden's R^2 is what's commonly reported as Pseudo-R^2 for logit, although that tends to be fairly low.

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

In addition to such absolute measures of fit, we are often interested in relative measures of fit that we use to select among two or more models--e.g., to decide whether to keep or omit a group of variables. We did that using test and lrtest commands above (to test joint statistical significance of a group of variables), but an alternative to that would involve comparing other measures of model fit (lrtest does that comparison by relying on log likelihoods as a measure of model fit). For this purpose, a 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 based on test and lrtest command results; BIC also allows us to compare non-nested models to each other (nested models are such that model 1 includes predictors A, B, and C, and model 2 includes predictors B and C – model 2 is nested in model 1; non-nested models are such that model 1 includes predictors A, B, and C, and model 2 includes predicts B, C, and D).

Here's how we compare model fit using fitstat. We already saved the fitstat results of the previous model. Let's say, we consider adding those marital status dummies:

```
. logit marijuana sex age educ childs i.marital

Iteration 0:   log likelihood =  -552.0232
Iteration 1:   log likelihood = -515.19453
Iteration 2:   log likelihood = -514.62744
Iteration 3:   log likelihood = -514.62716
Iteration 4:   log likelihood = -514.62716

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
             |
     marital |
     widowed  |  .0118099   .3568915     0.03   0.974    -.6876845    .7113043
    divorced  |  .9025573   .2053011     4.40   0.000     .5001746     1.30494
   separated  |  .0300665   .4239309     0.07   0.943    -.8008229    .8609558
never married  |  .2853992    .208832     1.37   0.172     -.123904    .6947024
             |
       _cons |  .2573784   .5195598     0.50   0.620    -.7609401    1.275697
------------------------------------------------------------------------------

. fitstat, dif

                         |     Current        Saved   Difference
-------------------------+---------------------------------------
Log-likelihood           |
                   Model |    -514.627     -524.848       10.221
          Intercept-only |    -552.023     -552.023        0.000
-------------------------+---------------------------------------
Chi-square               |
       D (df=836/840/-4) |    1029.254     1049.697      -20.443
           LR (df=8/4/4) |      74.792       54.350       20.443
                 p-value |       0.000        0.000        0.000
-------------------------+---------------------------------------
R2                       |
                McFadden |       0.068        0.049        0.019
     McFadden (adjusted) |       0.051        0.040        0.011
      McKelvey & Zavoina |       0.120        0.090        0.030
            Cox-Snell/ML |       0.085        0.062        0.022
  Cragg-Uhler/Nagelkerke |       0.116        0.085        0.031
                   Efron |       0.087        0.065        0.023
                Tjur's D |       0.086        0.063        0.023
                   Count |       0.206        0.204        0.001
        Count (adjusted) |      -1.208       -1.212        0.004
-------------------------+---------------------------------------
IC                       |
                     AIC |    1047.254     1059.697      -12.443
        AIC divided by N |       1.239        1.254       -0.015
          BIC (df=9/5/4) |    1089.908     1083.394        6.515
-------------------------+---------------------------------------
Variance of              |
```

```
                 e |       3.290        3.290        0.000
            y-star |       3.740        3.615        0.125

Note: Likelihood-ratio test assumes saved model nested in current model.

Difference of    6.515 in BIC provides strong support for saved model.
```

BIC suggests that adding marital status does not add enough to justify adding 4 extra variables (which is not what our LR test showed; but BIC is usually more conservative as it penalizes you more for adding additional parameters and losing parsimony). Of course, 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
different Ns between saved and current model (must use -force- option)
r(999);
```

Because our samples are not the same, 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~=.

Iteration 0:   log likelihood = -397.46953
Iteration 1:   log likelihood = -382.29137
Iteration 2:   log likelihood = -382.18666
Iteration 3:   log likelihood = -382.18666

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   .5479272     0.12   0.907    -1.009894    1.137941
------------------------------------------------------------------------------

. fitstat, save

                         |       logit
-------------------------+-------------
Log-likelihood           |
                   Model |    -382.187
          Intercept-only |    -397.470
-------------------------+-------------
Chi-square               |
       Deviance (df=594) |     764.373
             LR (df=4)   |      30.566
                 p-value |       0.000
-------------------------+-------------
R2                       |
                McFadden |       0.038
     McFadden (adjusted) |       0.026
      McKelvey & Zavoina |       0.069
            Cox-Snell/ML |       0.050
  Cragg-Uhler/Nagelkerke |       0.068
                   Efron |       0.053
               Tjur's D  |       0.051
                   Count |       0.140
        Count (adjusted) |      -1.270
-------------------------+-------------
IC                       |
                     AIC |     774.373
        AIC divided by N |       1.293
            BIC (df=5)   |     796.350
-------------------------+-------------
Variance of              |
                       e |       3.290
                  y-star |       3.534

. logit marijuana sex age educ childs rincom98

Iteration 0:   log likelihood = -397.46953
Iteration 1:   log likelihood = -379.96542
Iteration 2:   log likelihood = -379.82272
Iteration 3:   log likelihood = -379.82272

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   -.1600365
         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   -.0873306
    rincom98 |  -.0352966   .0162986    -2.17   0.030    -.0672413   -.0033519
       _cons |   .3036228   .5639178     0.54   0.590    -.8016358    1.408881
------------------------------------------------------------------------------

. fitstat, dif
                         |     Current        Saved   Difference
-------------------------+---------------------------------------
Log-likelihood           |
                   Model |    -379.823     -382.187        2.364
          Intercept-only |    -397.470     -397.470        0.000
-------------------------+---------------------------------------
Chi-square               |
       D (df=593/594/-1) |     759.645      764.373       -4.728
           LR (df=5/4/1) |      35.294       30.566        4.728
                 p-value |       0.000        0.000        0.030
-------------------------+---------------------------------------
R2                       |
                McFadden |       0.044        0.038        0.006
     McFadden (adjusted) |       0.029        0.026        0.003
      McKelvey & Zavoina |       0.078        0.069        0.009
            Cox-Snell/ML |       0.057        0.050        0.007
  Cragg-Uhler/Nagelkerke |       0.078        0.068        0.010
                   Efron |       0.060        0.053        0.008
                Tjur's D |       0.059        0.051        0.008
                   Count |       0.142        0.140        0.003
        Count (adjusted) |      -1.263       -1.270        0.008
-------------------------+---------------------------------------
IC                       |
                     AIC |     771.645      774.373       -2.728
        AIC divided by N |       1.288        1.293       -0.005
          BIC (df=6/5/1) |     798.017      796.350        1.667
-------------------------+---------------------------------------
Variance of              |
                       e |       3.290        3.290        0.000
                  y-star |       3.569        3.534        0.035

Note: Likelihood-ratio test assumes saved model nested in current model.

Difference of    1.667 in BIC provides weak support for saved model.
```

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
```

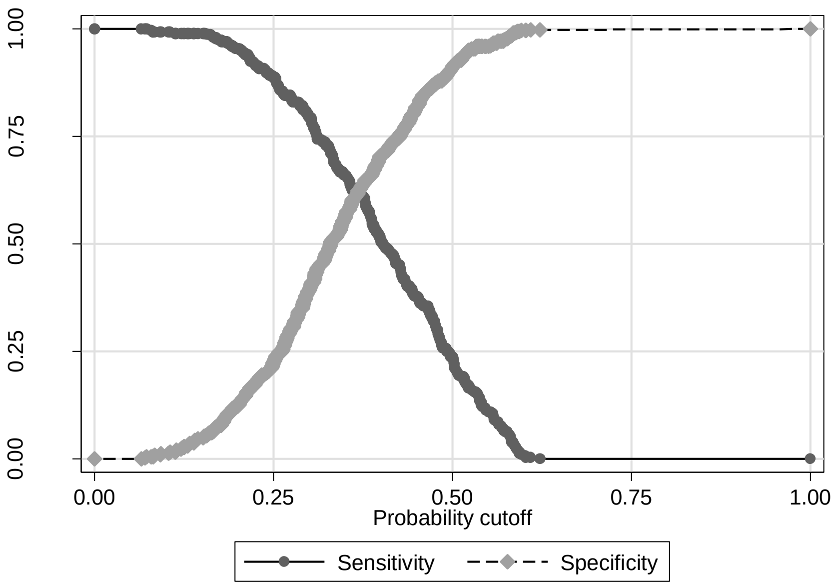
```
--------------------------------------------------
Sensitivity                     Pr( +| D)   23.68%
Specificity                     Pr( -|~D)   91.13%
Positive predictive value       Pr( D| +)   60.00%
Negative predictive value       Pr(~D| -)   68.00%
--------------------------------------------------
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 on the task at hand.

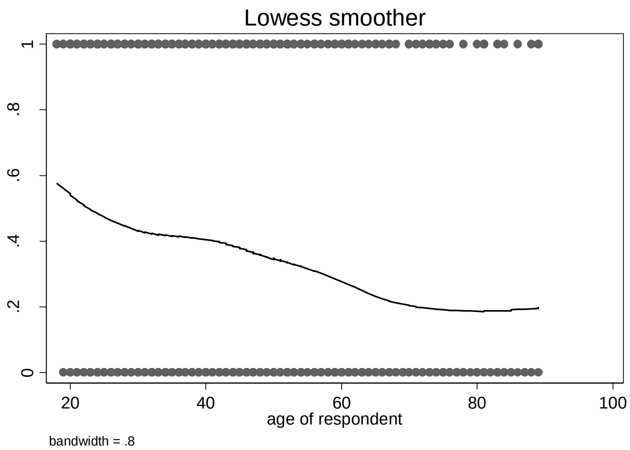
Diagnostics for binary logit

A. Data Screening

Before conducting logistic regression, you should do data screening (like we did for OLS). That is, it is 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. In logistic regression, we do not expect residuals to be normally distributed, but normally distributed predictors still help avoid problems. Also, look out for outliers and deal with those.

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

B. Multivariate Diagnostics

1. Linearity

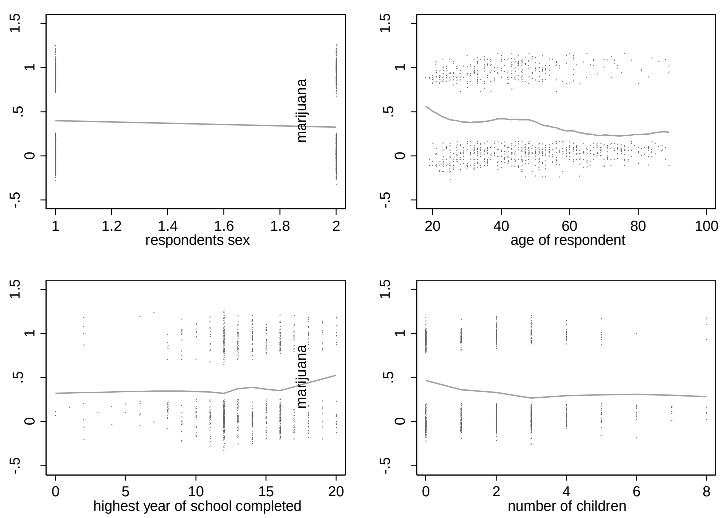
In multivariate context, you can use boxtid--don't forget to specify that you are using logit rather then reg when using boxtid, i.e. use:

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

Iteration 0:  Deviance =  1043.357
Iteration 1:  Deviance =  1042.752 (change = -.6045771)
Iteration 2:  Deviance =  1042.734 (change =  -.018392)
Iteration 3:  Deviance =  1042.733 (change = -.0012757)
Iteration 4:  Deviance =  1042.732 (change = -.0002699)
-> gen double Iage__1 = X^2.0968-25.22385401 if e(sample)
-> gen double Iage__2 = X^2.0968*ln(X)-38.83014807 if e(sample)
   (where: X = age/10)
-> gen double Ieduc__1 = X^7.1584-13.16861852 if e(sample)
-> gen double Ieduc__2 = X^7.1584*ln(X)-4.74218828 if e(sample)
   (where: X = (educ+1)/10)
-> gen double Ichil__1 = X^-0.8682-.4079980779 if e(sample)
-> gen double Ichil__2 = X^-0.8682*ln(X)-.4212880559 if e(sample)
   (where: X = (childs+1))
-> gen double Isex__1 = sex-1 if e(sample)

[Total iterations: 12]

Box-Tidwell regression model

Logistic regression                               Number of obs   =        845
                                                  LR chi2(7)      =      61.31
                                                  Prob > chi2     =     0.0000
Log likelihood = -521.36615                       Pseudo R2       =     0.0555

------------------------------------------------------------------------------
   marijuana |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     Iage__1 |   -.014519   .0496234    -0.29   0.770    -.1117792    .0827411
     Iage_p1 |  -.0002796    .022828    -0.01   0.990    -.0450217    .0444626
    Ieduc__1 |   .0037305   .0183905     0.20   0.839    -.0323143    .0397753
```

```
    Ieduc_p1 |      .00001   .0255285     0.00   1.000    -.0500249    .0500449
    Ichil__1 |      1.0601   .8716771     1.22   0.224    -.6483557    2.768556
    Ichil_p1 |  -.0007376   1.364119    -0.00   1.000    -2.674361    2.672886
     Isex__1 |  -.3217827   .1502772    -2.14   0.032    -.6163207   -.0272447
       _cons |  -.5952113   .1483855    -4.01   0.000    -.8860415   -.3043811
------------------------------------------------------------------------------
age      |  -.0184297   .0048878     -3.771   Nonlin. dev. 0.865   (P = 0.352)
      p1 |   2.096755   1.445354      1.451
------------------------------------------------------------------------------
educ     |   .0391444   .0254125      1.540   Nonlin. dev. 1.701   (P = 0.192)
      p1 |   7.158414   6.913701      1.035
------------------------------------------------------------------------------
childs   |  -.1810504   .0528152     -3.428   Nonlin. dev. 4.677   (P = 0.031)
      p1 |  -.8682125    1.28118     -0.678
------------------------------------------------------------------------------
Deviance: 1042.732.
```

You can also try mrunning but it is based on OLS regression so it is a less precise tool here. Still, it can identify potential problems.

```
. mrunning marijuana sex age educ childs

845 observations, R-sq = 0.0829
```



2. Outliers and influential data points

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

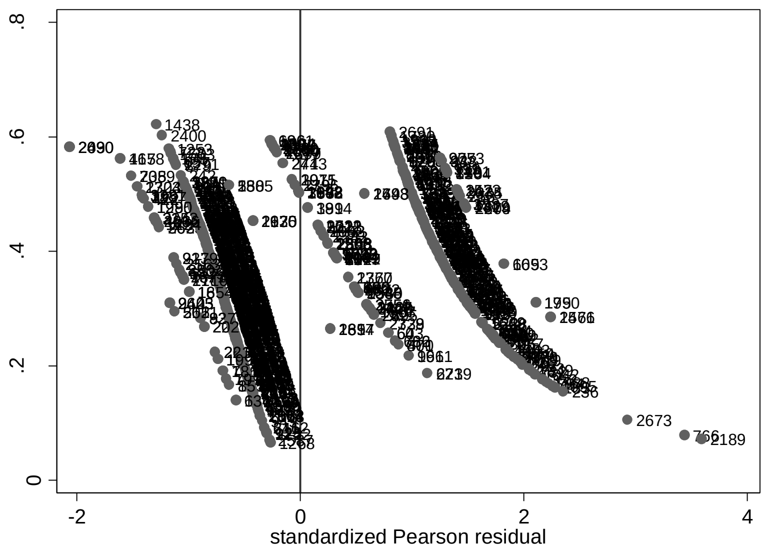
```
p            predicted probability of a positive outcome; the default
xb           linear prediction
stdp         standard error of the linear prediction
dbeta        Pregibon (1981) Delta-Beta influence statistic
deviance     deviance residual
dx2          Hosmer and Lemeshow (2000) Delta chi-squared 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)
```



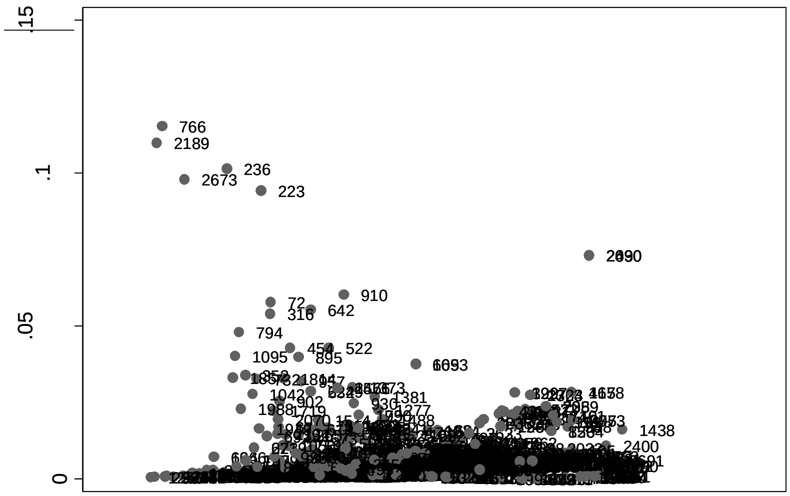
Observations on the far left or far right deserve further examination. Here, we would especially look at 766 and 2189, but also 2673.

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)
```

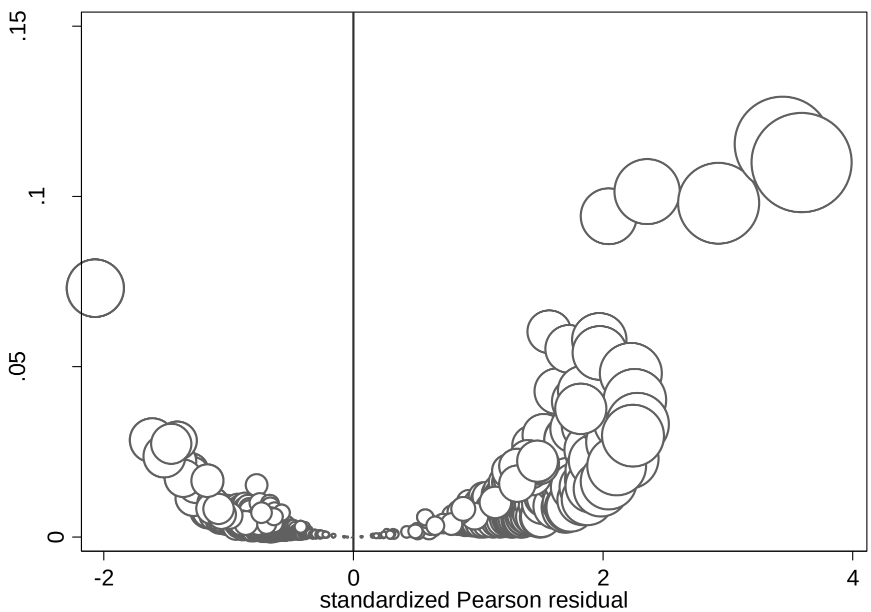
We can then examine these graphically to identify problematic observations:

```
. 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.

2. Additivity

You can once again use fitint command to search for checking for interactions; the syntax for testing all interactions in the same combined model is

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

But it is also a good idea to test interactions one by one as well, like we did in OLS.

Note, however, 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 – as I mentioned earlier, residual variance in both logit and probit models is always fixed to the same number, regardless of how much variance your predictors actually explain. That is, if you are trying to compare the effect of a predictor in two groups – e.g., men and women—the coefficients for one of the groups could be "scaled up" and therefore larger because the residual variance is smaller (i.e., we explain the variance better than in the other group), and such difference will end up incorporated in the residual term because the variance is fixed to be the same for both groups (and that will still be the case if we estimate separate models for the two groups rather than use interaction terms). This problem was originally noted in: Allison, Paul D. 1999. "Comparing Logit and Probit Coefficients Across Groups." *Sociological Methods and Research,* 28: 186-208.

The best way to explore group comparisons under these circumstances is by creating graphs of predicted probabilities with confidence intervals, or better yet, a graph for the difference in predicted probabilities, also with confidence intervals:
http://www.indiana.edu/~jslsoc/files_research/rm4cldv/group_compare/long_group_nd_2007-04-16.pdf
We will deal with that later, when discussing the interpretation of results. You may also want to look into heterogenous choice models implemented in oglm:
https://www3.nd.edu/~rwilliam/stats/Oglm.pdf
https://www3.nd.edu/~rwilliam/oglm/RW_Hetero_Choice.pdf

3. 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
. vif

    Variable |       VIF       1/VIF
-------------+----------------------
      childs |      1.24    0.803381
         age |      1.21    0.825046
        educ |      1.04    0.961375
         sex |      1.01    0.985827
-------------+----------------------
    Mean VIF |      1.13
```

4. 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 binomial distribution of the error term implies that 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 are more appropriate in the presence of assumption violations.

```
. 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.

5. 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 towards the end of this course, when talking about survey data.