MLE: Categorical and Limited Dependent Variables

Unit 1: Models for Dichotomous Dependent Variables
3. Issues in the Estimation and Interpretation of Logit and Probit Models

PS2730-2021

Week 3

Professor Steven Finkel

UNIVERSITY OF PITTSBURGH

VERITAS 1787 VIRTUS

Interpretation of Logit and Probit Coefficients

- Given the non-linearities in the logit and probit models, it is not immediately clear how to interpret the regression coefficients and effects that are obtained. What do the β mean exactly, and how are they related to changes in P(Y=1) for different ranges and changes in X? This leads to the more general question of how best to interpret the effect of the X in these models
- General approaches
  1. Direct interpretation of the effect of β on the linear predictor η from GLM or on a transformed η (such as the "odds" in logit)
  2. Calculation of marginal effects or changes in P(Y=1) for different kinds of changes in X
     a. Average effect on P(Y=1) for "marginal change" in continuous X variables
     b. Average effect on P(Y=1) for "discrete changes" in all X variables
  3. Interpretation of the β on changes in Y* in latent variable framework

1. Direct interpretation of the β

- In logit/probit, as in any of the GLM models we'll consider, one can always interpret the β as the effect of a unit-change in X on the linear predictor η, whatever that is for the given model.
  - For logit, it is the "log-odds" that Y=1
  - For probit, it is the "inverse cumulative normal" or the z-score corresponding to the proportion of the cumulative normal curve cut off by P(Y=1)

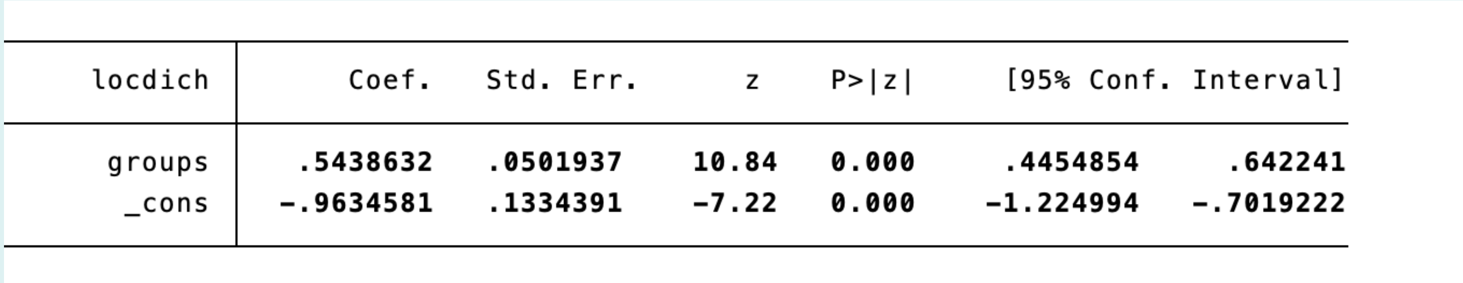
- The size, sign, and significance of β tell you something generally about the nature and magnitude of the effects
  - Bigger (smaller) means steeper (more gradual) changes in η for a unit change in X; positive/negative/significant/not significant: all self-explanatory.
  - These changes map onto P(Y=1) in a non-linear fashion, but they map nonetheless

- But: the exact numerical value of β is arbitrary; we chose (for identification purposes) the normal distribution with s.d. of 1 for probit or the logistic distribution with variance ($\boldsymbol{\pi}^2/3$) for logit (with "**π**" here being the irrational number 3.141……., not P(Y=1)!)
- Given the similarities of the normal and logistic distributions, can convert probit to approximate logit coefficients by multiplying probit by $\boldsymbol{\pi}/\sqrt{3}$, or approximately 1.81.
- But this also means that the β itself doesn't tell you much in and of itself; unlike linear regression, e.g., it doesn't tell you the average change in actual Y for a unit change in X (or even the average change in P(Y=1)
- Moreover, nobody intuitively understands what an effect on a log-odds (or a z-score) means anyway! So we generally want to use other ways to understand effects in these models

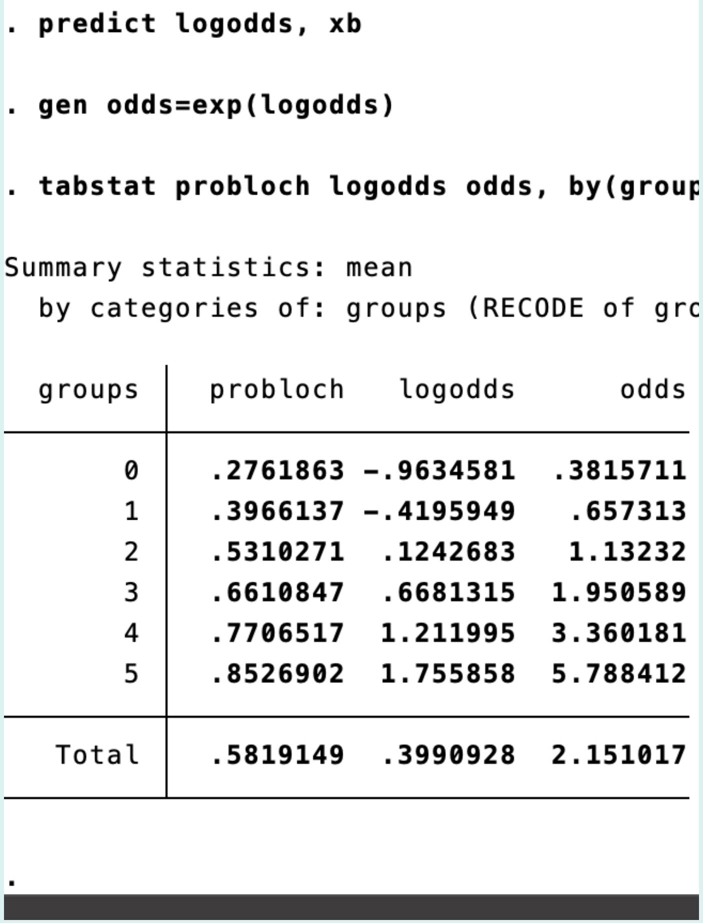
Odds Interpretations (in Logit)

- In logit models, there is a nice alternative interpretation of β: since β represents the change in the "log-odds" that P(Y=1) for every unit change in X, then exp(β) represents the **factor change in the odds that Y=1** for every unit change in X. Odds=(P(Y=1)/(1-P(Y=1))
- Exponentiating β gives you this information ("listcoef" in Stata)
- In our example, every unit increase in group memberships changes the odds that a person participates, i.e, (Y=1), by a factor of 1.723 (exp$^{.54}$)
- There is a *constant factor* or multiplicative change in the odds for every unit change in X. Going from 0-1 on X changes the odds by a factor of 1.723, going from 3-4 on X changes the odds by same factor, etc.
- **So in logit: X has linear effect on the log-odds that Y=1, and X has constant factor effect on the odds that Y=1**
- Can also say that a unit change in X increases the odds by 72.3%
- Can also calculate factor changes in odds by changing X by 1 SD and comparing across variables (see do file for today's session)

Our logistic regression of locdich against groups in the South African data (slide 15, week 1)



```
     locdich |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      groups |   .5438632   .0501937    10.84   0.000     .4454854     .642241
       _cons |  -.9634581   .1334391    -7.22   0.000    -1.224994   -.7019222
------------------------------------------------------------------------------
```

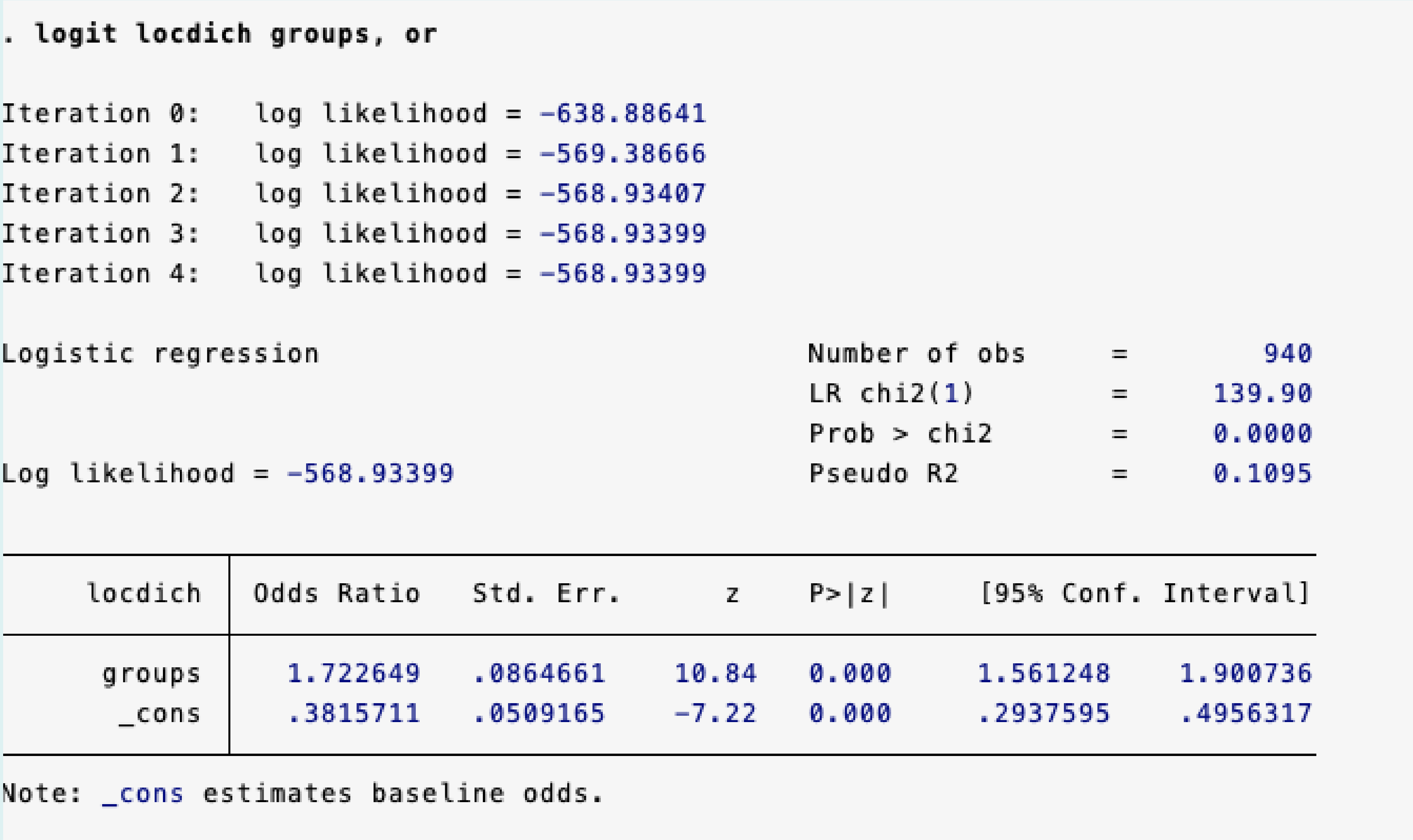


```
. predict logodds, xb

. gen odds=exp(logodds)

. tabstat probloch logodds odds, by(group

Summary statistics: mean
  by categories of: groups (RECODE of gro

  groups |  probloch    logodds       odds
---------+------------------------------
       0 |  .2761863 -.9634581   .3815711
       1 |  .3966137 -.4195949    .657313
       2 |  .5310271  .1242683    1.13232
       3 |  .6610847  .6681315   1.950589
       4 |  .7706517  1.211995   3.360181
       5 |  .8526902  1.755858   5.788412
---------+------------------------------
   Total |  .5819149  .3990928   2.151017
----------------------------------------

.
```

For 0 groups, P(Y=1)=.276, odds=.381
For 1 groups, P(Y=1)=.397 odds=.658
Do you see the 1.723 multiplicative change in the odds for every unit change in X?

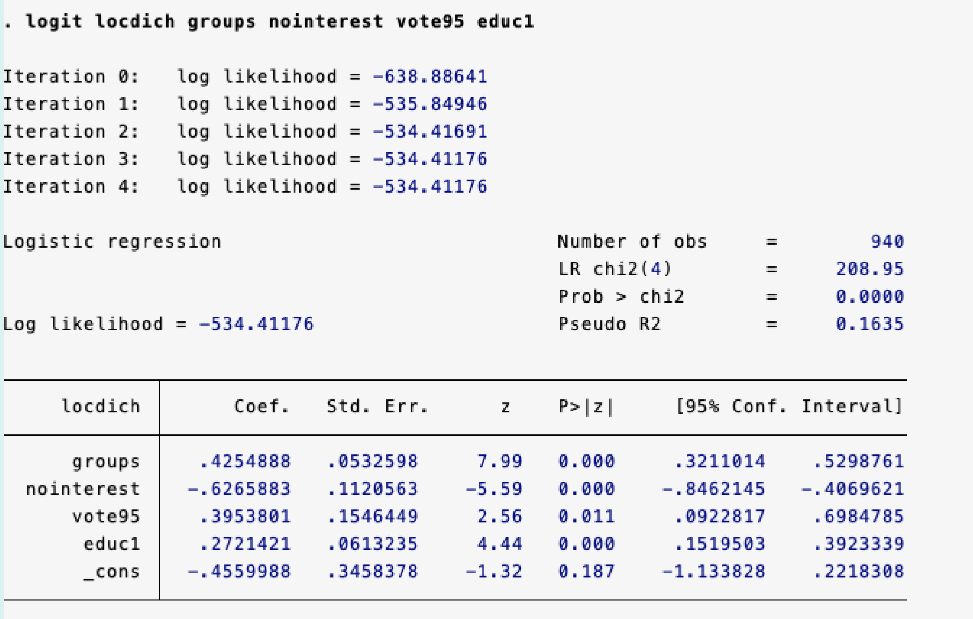
.38*1.723=.657 (from 0-1)
.657*1.723=1.132 (from 1-2)
3.360*1.723=5.789 (from 4-5)

Logit with “or” option (or “logistic” instead of “logit”) gives entire model in terms of odds-ratios instead of log-odds, and you get correct confidence intervals for the factor change in the odds for a unit change in X (by exponentiating the lower and upper bounds of the 95% confidence interval for the logit coefficient)

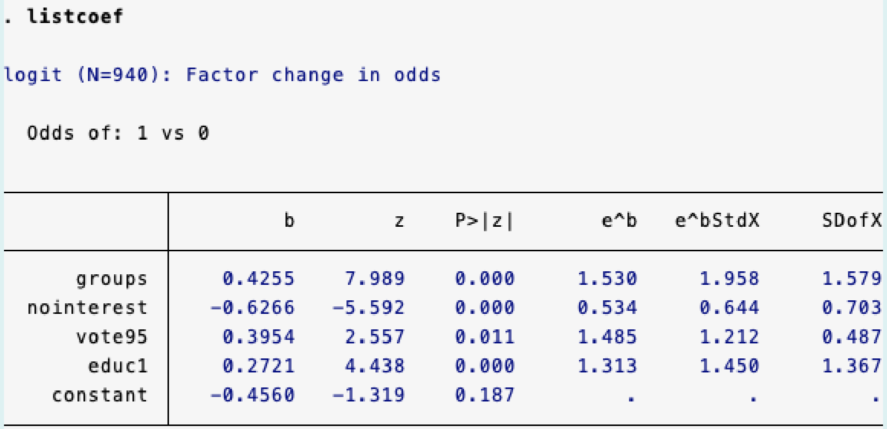


```
. logit locdich groups, or

Iteration 0:   log likelihood = -638.88641
Iteration 1:   log likelihood = -569.38666
Iteration 2:   log likelihood = -568.93407
Iteration 3:   log likelihood = -568.93399
Iteration 4:   log likelihood = -568.93399

Logistic regression                             Number of obs     =        940
                                                LR chi2(1)        =     139.90
                                                Prob > chi2       =     0.0000
Log likelihood = -568.93399                     Pseudo R2         =     0.1095

------------------------------------------------------------------------------
     locdich | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      groups |   1.722649   .0864661    10.84   0.000     1.561248    1.900736
       _cons |   .3815711   .0509165    -7.22   0.000     .2937595    .4956317
------------------------------------------------------------------------------
Note: _cons estimates baseline odds.
```

- And easy extension to multiple logistic regression: exp(logit) is the factor change in the odds that Y=1 for a unit change in X, *holding all other variables constant*. So it is a constant factor change regardless of the levels of other variables (which is also a nice feature of logit)



```
. logit locdich groups nointerest vote95 educ1

Iteration 0:   log likelihood = -638.88641
Iteration 1:   log likelihood = -535.84946
Iteration 2:   log likelihood = -534.41691
Iteration 3:   log likelihood = -534.41176
Iteration 4:   log likelihood = -534.41176

Logistic regression                             Number of obs     =        940
                                                LR chi2(4)        =     208.95
                                                Prob > chi2       =     0.0000
Log likelihood = -534.41176                     Pseudo R2         =     0.1635

------------------------------------------------------------------------------
     locdich |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      groups |   .4254888   .0532598     7.99   0.000     .3211014    .5298761
  nointerest |  -.6265883   .1120563    -5.59   0.000    -.8462145   -.4069621
      vote95 |   .3953801   .1546449     2.56   0.011     .0922817    .6984785
       educ1 |   .2721421   .0613235     4.44   0.000     .1519503    .3923339
       _cons |  -.4559988   .3458378    -1.32   0.187    -1.133828    .2218308
------------------------------------------------------------------------------
```



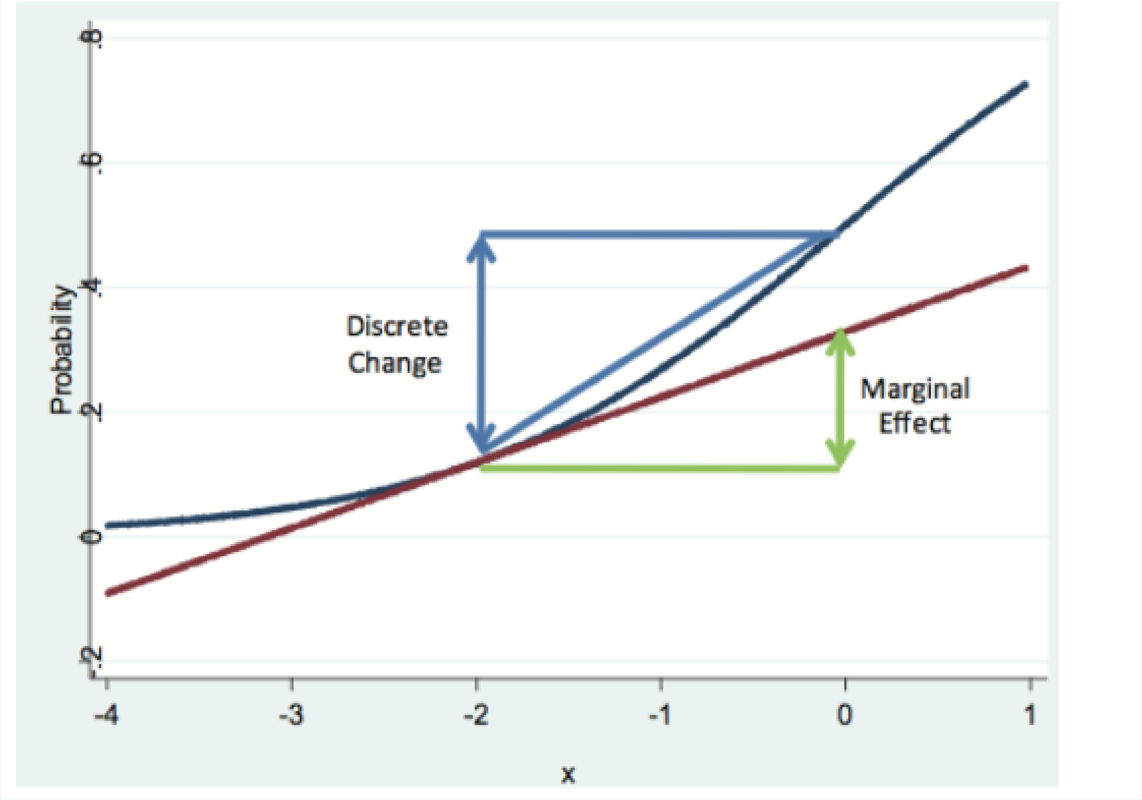
```
. listcoef

logit (N=940): Factor change in odds

  Odds of: 1 vs 0

--------------------------------------------------------------------------
             |         b         z     P>|z|       e^b    e^bStdX     SDofX
-------------+------------------------------------------------------------
      groups |    0.4255     7.989     0.000     1.530      1.958     1.579
  nointerest |   -0.6266    -5.592     0.000     0.534      0.644     0.703
      vote95 |    0.3954     2.557     0.011     1.485      1.212     0.487
       educ1 |    0.2721     4.438     0.000     1.313      1.450     1.367
    constant |   -0.4560    -1.319     0.187         .          .         .
--------------------------------------------------------------------------
```

Note: Factor changes on the odds that are less than 1 can still be *very* powerful effects. "Nointerest" has a logit effect of -.63, and a factor change in the odds effect of .534. (I created "nointerest" by generating a new variable as (5-interest)). This is the same as a 1/.534, or a 1.87 **decrease** in the odds of participating, for every increase of 1 unit on "nointerest". It is a larger absolute change on the odds than a unit change in any other variable!

2. Marginal effects on changes in P(Y=1)

- Alternative to direct interpretation of β is to examine the effects of changes in X on changes in the P(Y=1), holding other variables constant. In linear regression, this effect is simply β, since the partial derivative of the regression line wrt X is β. In non-linear models, these quantities change as X and other independent variables change.
- Several implications: we can first choose (for continuous variables), whether to calculate the effects of *marginal change* or *discrete changes* in X on P(Y=1) (see next slide)
- And since the effect of changes in X on P(Y=1) will depend on the values of all the other variables, we need to take this into account to see how changes in the variable(s) of interest affect P(Y=1).
- Finally, it means, for discrete change analysis, we can pick specific interesting/meaningful quantities of change in X to display the effects on the P(Y=1), for example changing X from minimum to maximum values, unit increases, etc.

Marginal Change Versus Discrete Change



- A “marginal effect” or “marginal change” is the effect of an infinitesimally small change in X on P(Y=1). It is calculated as the slope of the tangent to the probability curve for P(Y=1), or the first (partial) derivative at a given **value** of X, or ∂ P(Y=1)/ ∂X_k where “k” is the specific value of X
- A “discrete change” is the change in the P(Y=1) for a given amount of **change** in X. It is calculated as

  P(Y=1|X, X_k =end)- P(Y=1|X, X_k= start), or ΔP(Y=1|X)/ ΔX
- You can see that both changes will vary, depending on where on the probability curve X is (which depends on all other variables *and* X)

  This can be seen formally for marginal change by taking the partial derivative for a logit curve as:

$$\frac{\exp(X_k\beta)}{(1+\exp(X_k\beta))^2}\beta = P(Y=1|X_k)(1-P(Y=1|X_k)\beta$$

  which means that the marginal effect is greatest when closest to P=.5
- It also means that dividing the **logit coefficient** by 4 gives you the “maximum marginal effect” (since the expression is maximized at .5*.5)

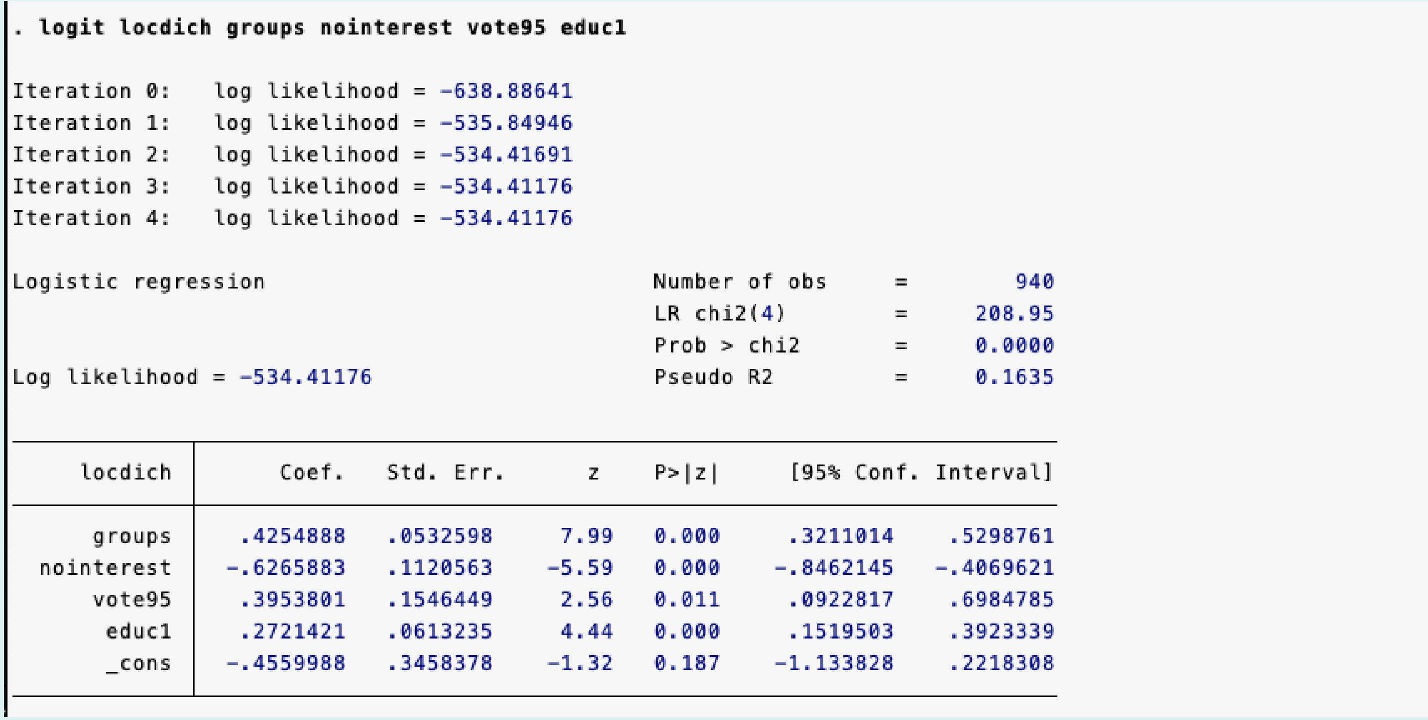
- Where shall you set the values of the other independent variables in calculating marginal or discrete changes?
- **Marginal Effect at the Mean (MEM)**: set all other variables equal to their sample means, and calculate either marginal or discrete changes. This used to be the standard method.
  - Advantage: it provides the baseline probability for an "otherwise average" unit (and it is easy to calculate); Disadvantage: no unit might be at or near the mean on all the other independent variables, so it may not correspond to a "typical" case
- **Average Marginal Effect (AME)**: for marginal change, allow all other variables to remain at their observed sample values, calculate the marginal change for a unit based on its value of X, and average this quantity across all units. For discrete change, calculate P(Y=1) for a given change in X, allowing all other variables to remain at their observed sample values. This is **now** Stata's default! See Hanmer and Kalkan (2013) for a recent treatment.

- Marginal changes and discrete changes can be very different, depending on the non-linearities of the probability curve for the points on X that you are concerned with
- Differences in displaying "marginal" versus "discrete" changes depend mainly on personal or disciplinary preference (sociology likes discrete change (following Long), economics likes marginal change)
- Marginal changes can be calculated for dummy/categorical independent variables, but not really meaningful to talk of "instantaneous" change in a dummy variable. There is a more complicated interpretation but often advised just to use discrete change.
- Both kinds of marginal effects are estimated quantities, so it is useful to compute standard errors and confidence intervals for both
- All of this can now be done via the MARGINS and SPOST commands in Stata

- **Marginal Effect at Representative Values (MER**): Pick substantively interesting values of other Xs and calculate the associated marginal or discrete changes in P(Y=1). Used anecdotally more than systematically. Some values typically chosen: minimum/maximum (range), quartiles, or sometimes other theoretically compelling values of covariates
- Following the earlier discussion of marginal effects, one combination of all other independent variables can be especially interesting: when, taken together, they put the unit at the point on the probability curve corresponding to .5 (i.e., *a priori* logits or z-scores of 0). This is the place where a unit (or standard unit) change in X has its maximum impact!
- Which is best, **MEM**, **AME**, or **MER**? Discipline is converging on AME since it is the average marginal effect for the given sample, and MEM may not be "representative". They differ depending on how big P(Y=1) is when all variables are at their means; high/low values mean AME is larger; medium values means MEM is larger. See excellent discussion in Long and Freese pp. 244-246.

- Last issue: what changes in X are most informative to use when calculating **discrete change** in P(Y=1)?
- Many choices:
  - Show P(Y=1) associated with minimum, mean, and maximum values of X
  - Set X at its mean value, and show the effect of a unit change in X at that point. This is the effect of a unit change in X for an otherwise "average" person on X.
  - (Centered) change in X of 1 unit at the mean of X $=\bar{X} \pm .5$
  - Set X at its mean value, and show the effect of a *standard deviation* change in X at that point. This is the effect of one standard deviation change in X for an otherwise "average" person on X. Or centered change in X of 1 standard unit at $\bar{X} \pm .5\sigma$

Multivariate Logit Example

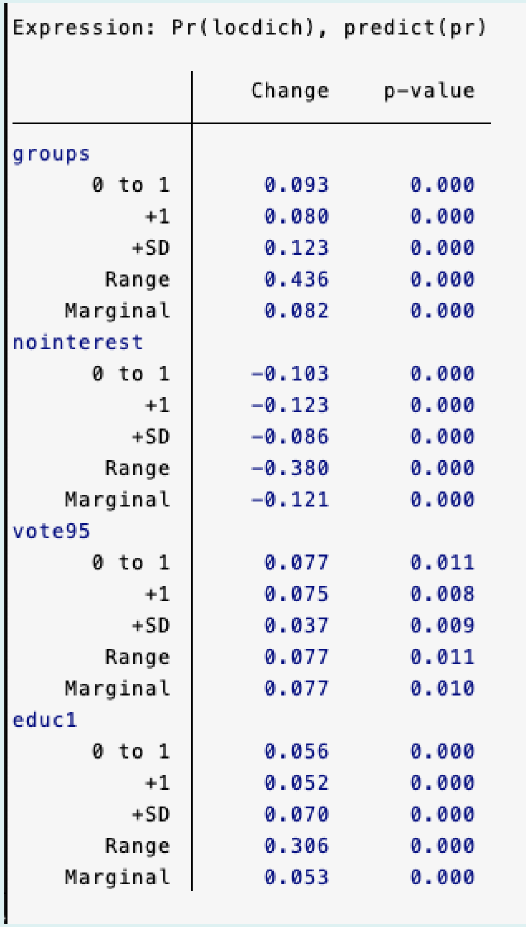


```
. logit locdich groups nointerest vote95 educ1

Iteration 0:   log likelihood = -638.88641
Iteration 1:   log likelihood = -535.84946
Iteration 2:   log likelihood = -534.41691
Iteration 3:   log likelihood = -534.41176
Iteration 4:   log likelihood = -534.41176

Logistic regression                               Number of obs     =        940
                                                  LR chi2(4)        =     208.95
                                                  Prob > chi2       =     0.0000
Log likelihood = -534.41176                       Pseudo R2         =     0.1635

------------------------------------------------------------------------------
     locdich |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      groups |   .4254888   .0532598     7.99   0.000     .3211014    .5298761
  nointerest |  -.6265883   .1120563    -5.59   0.000    -.8462145   -.4069621
      vote95 |   .3953801   .1546449     2.56   0.011     .0922817    .6984785
       educ1 |   .2721421   .0613235     4.44   0.000     .1519503    .3923339
       _cons |  -.4559988   .3458378    -1.32   0.187    -1.133828    .2218308
------------------------------------------------------------------------------
```

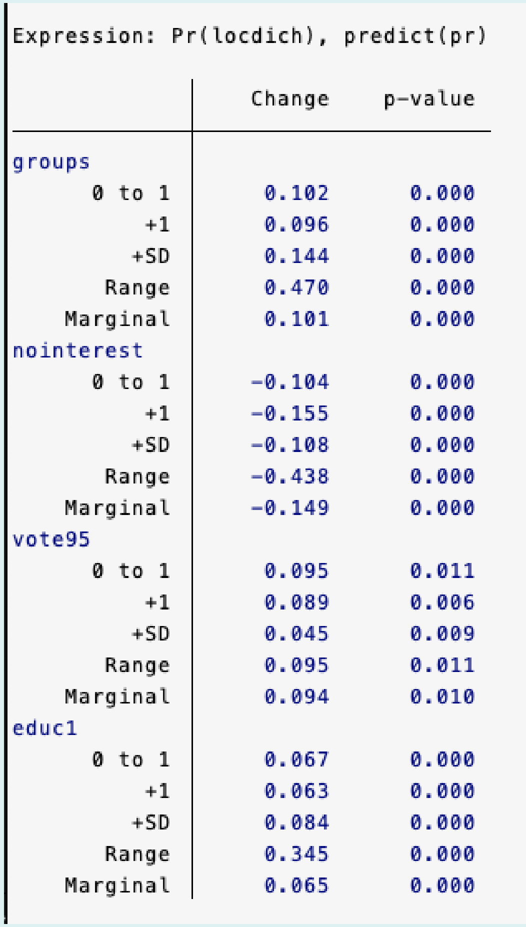
- So: unit change in groups leads to a constant .425 change in the logit of Y
- Unit change in education leads to a constant .272 change in the logit of Y
- Voters from 1995 have logits that are .395 higher than non-voters
- So calculating the effect of each variable on the probability that Y=1 means different effects for different kinds of people, depending on where they are on the cdf from all other variables, which determines their “otherwise existing probabilities”

- Effect of going from 0 to 4 groups for voters, for otherwise “average” person
  - Logit= -.456 + -.627*1.92+.3954+.2721*2.880=-.481 P(Y=1)=.382
  - Logit= -.456 + .426*4+ -.627*1.92+.3954+.2721*2.880=1.22 P(Y=1)=.772
- Effect of going from 0 to 4 groups for non-voters, for otherwise “average” person
  - Logit= -.456 + -.627*1.92+.2721*2.880 =-.876 P(Y=1)=.294
  - Logit= -.456 +.426*4+ -.627*1.92+.2721*2.880 =.828 P(Y=1)=.696
- Effect for Voters: .772-.382=.310 Effect for Non-Voters: .696-.294=.402
- Effect of one group membership for a person with prior probability of .5 =.105 (i.e. p goes from .5 to .605)
- You can do these kinds of calculations for any combination of independent variables, what Long and Freese call ”Marginal Effects at Representative Values” or MER.

- This is the output from “mchange”; it provides all the marginal and discrete changes, holding all other variables at their *observed* sample values on left, at their mean values on right. They are similar but *not* identical



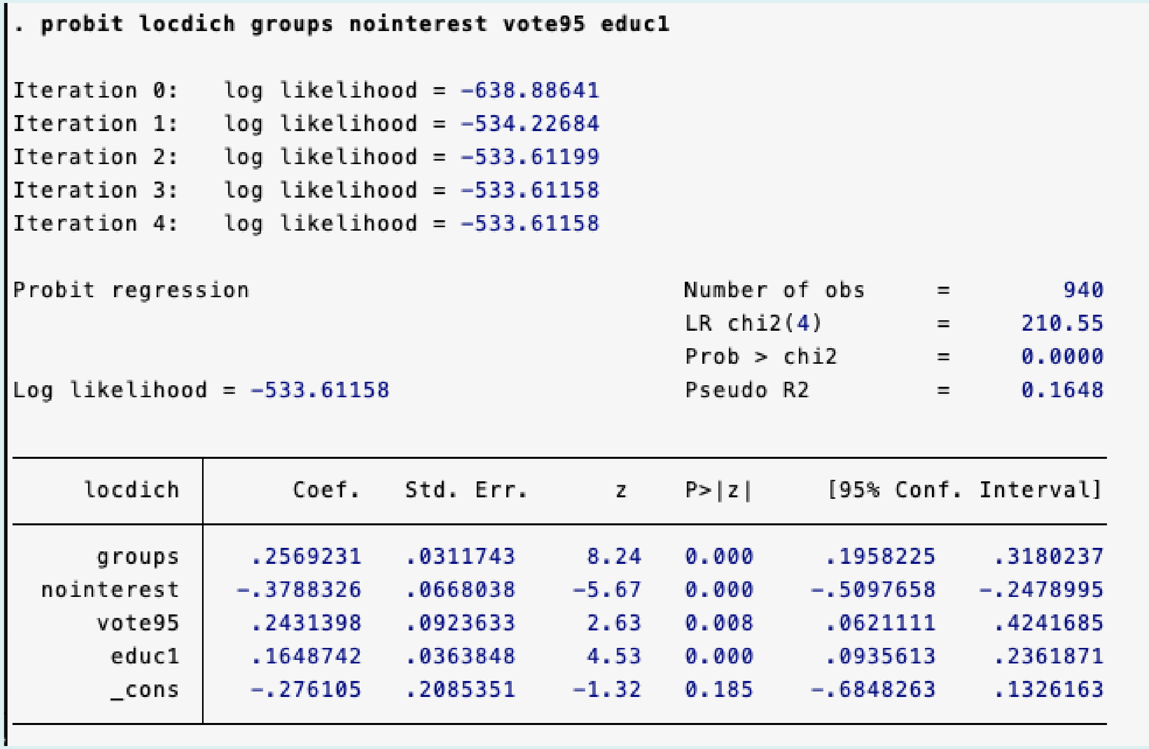
```
Expression: Pr(locdich), predict(pr)

                        |    Change    p-value
------------------------+---------------------
groups                  |
                 0 to 1 |     0.093      0.000
                     +1 |     0.080      0.000
                    +SD |     0.123      0.000
                  Range |     0.436      0.000
               Marginal |     0.082      0.000
nointerest              |
                 0 to 1 |    -0.103      0.000
                     +1 |    -0.123      0.000
                    +SD |    -0.086      0.000
                  Range |    -0.380      0.000
               Marginal |    -0.121      0.000
vote95                  |
                 0 to 1 |     0.077      0.011
                     +1 |     0.075      0.008
                    +SD |     0.037      0.009
                  Range |     0.077      0.011
               Marginal |     0.077      0.010
educ1                   |
                 0 to 1 |     0.056      0.000
                     +1 |     0.052      0.000
                    +SD |     0.070      0.000
                  Range |     0.306      0.000
               Marginal |     0.053      0.000
```



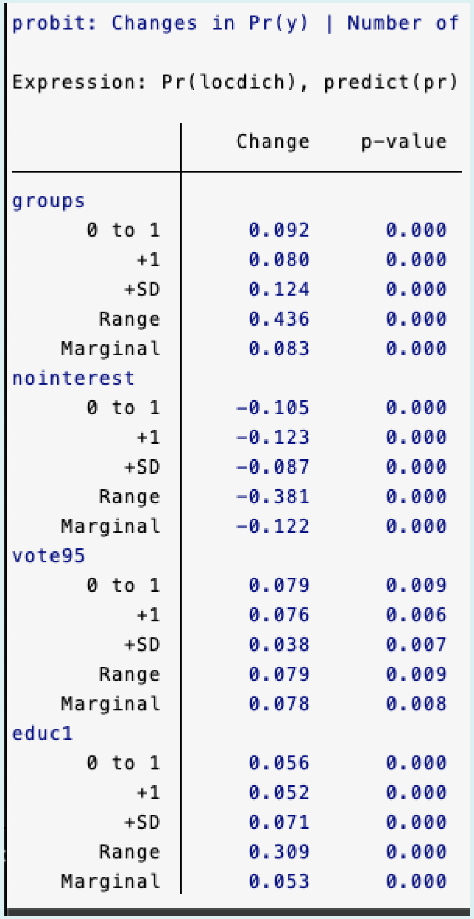
```
Expression: Pr(locdich), predict(pr)

                        |    Change    p-value
------------------------+---------------------
groups                  |
                 0 to 1 |     0.102      0.000
                     +1 |     0.096      0.000
                    +SD |     0.144      0.000
                  Range |     0.470      0.000
               Marginal |     0.101      0.000
nointerest              |
                 0 to 1 |    -0.104      0.000
                     +1 |    -0.155      0.000
                    +SD |    -0.108      0.000
                  Range |    -0.438      0.000
               Marginal |    -0.149      0.000
vote95                  |
                 0 to 1 |     0.095      0.011
                     +1 |     0.089      0.006
                    +SD |     0.045      0.009
                  Range |     0.095      0.011
               Marginal |     0.094      0.010
educ1                   |
                 0 to 1 |     0.067      0.000
                     +1 |     0.063      0.000
                    +SD |     0.084      0.000
                  Range |     0.345      0.000
               Marginal |     0.065      0.000
```

3. Interpretation of β on Y* in Latent Variable Framework

- As noted, probit coefficients can be interpreted similarly to logits in terms of the constant change on some quantity related to P(Y=1). In logit, it is the “log-odds”. In probit, it is the “z-score” that corresponds to a point on the cumulative normal distribution associated with the probability that Y=1
- Everything from previous slides holds for calculating effects in bivariate and multivariate models *except* for the odds constant factor change interpretation in logit, which is not applicable in probit
- You will get (almost) exactly the same P(Y=1) for a person with a given set of values on the IVs in logit or probit.
  And logit $\beta \cong$ probit $\beta * \boldsymbol{\pi}/\sqrt{3}$, or probit*1.81
- So the choice of logit or probit in this respect is pretty much a matter of personal preference



```
. probit locdich groups nointerest vote95 educ1

Iteration 0:   log likelihood = -638.88641
Iteration 1:   log likelihood = -534.22684
Iteration 2:   log likelihood = -533.61199
Iteration 3:   log likelihood = -533.61158
Iteration 4:   log likelihood = -533.61158

Probit regression                               Number of obs     =        940
                                                LR chi2(4)        =     210.55
                                                Prob > chi2       =     0.0000
Log likelihood = -533.61158                     Pseudo R2         =     0.1648

------------------------------------------------------------------------------
     locdich |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      groups |   .2569231   .0311743     8.24   0.000     .1958225    .3180237
  nointerest |  -.3788326   .0668038    -5.67   0.000    -.5097658   -.2478995
      vote95 |   .2431398   .0923633     2.63   0.008     .0621111    .4241685
       educ1 |   .1648742   .0363848     4.53   0.000     .0935613    .2361871
       _cons |   -.276105   .2085351    -1.32   0.185    -.6848263    .1326163
------------------------------------------------------------------------------
```



```
probit: Changes in Pr(y) | Number of

Expression: Pr(locdich), predict(pr)

             |      Change     p-value
-------------+------------------------
groups       |
      0 to 1 |       0.092       0.000
          +1 |       0.080       0.000
         +SD |       0.124       0.000
       Range |       0.436       0.000
    Marginal |       0.083       0.000
nointerest   |
      0 to 1 |      -0.105       0.000
          +1 |      -0.123       0.000
         +SD |      -0.087       0.000
       Range |      -0.381       0.000
    Marginal |      -0.122       0.000
vote95       |
      0 to 1 |       0.079       0.009
          +1 |       0.076       0.006
         +SD |       0.038       0.007
       Range |       0.079       0.009
    Marginal |       0.078       0.008
educ1        |
      0 to 1 |       0.056       0.000
          +1 |       0.052       0.000
         +SD |       0.071       0.000
       Range |       0.309       0.000
    Marginal |       0.053       0.000
```

- E.g. effect of going from 0 groups to 4 groups for voters in 1995, otherwise average

0: z= -.276+-.379*1.92+.243+.165* 2.880 = -.285   PHI(-.285)=.388

4: z= -.276+4*.256+-.379*1.92+.243+.165* 2.880 = .739   PHI(.739)= .770

- But given the derivation of probit in terms of the latent variable approach, we can also interpret probit coefficients another way: as the effect of a unit change in X on Y*, the latent "propensity" of the latent "utility" for the behavior or the choice that is being modeled. This is a nice linear effect (see the graph in week 1, slide 20)!
- So every additional group to which an individual belongs changes Y* (the propensity to participate in local politics) by .257 units.
- But there is a big problem in interpreting this value: the variance of Y* is not fixed, it is determined by the model (given that it is unobserved and we had to fix the variance of ε to identify the model in the first place). This is fundamentally different from linear regression, where the variance of Y was observed and is independent from the Xs in a regression model.
- In non-linear latent variable models, the variance of Y* and the β are not separately identifiable! And adding Xs makes bigger variance in Y*, so the β mean different things depending on which Xs are included

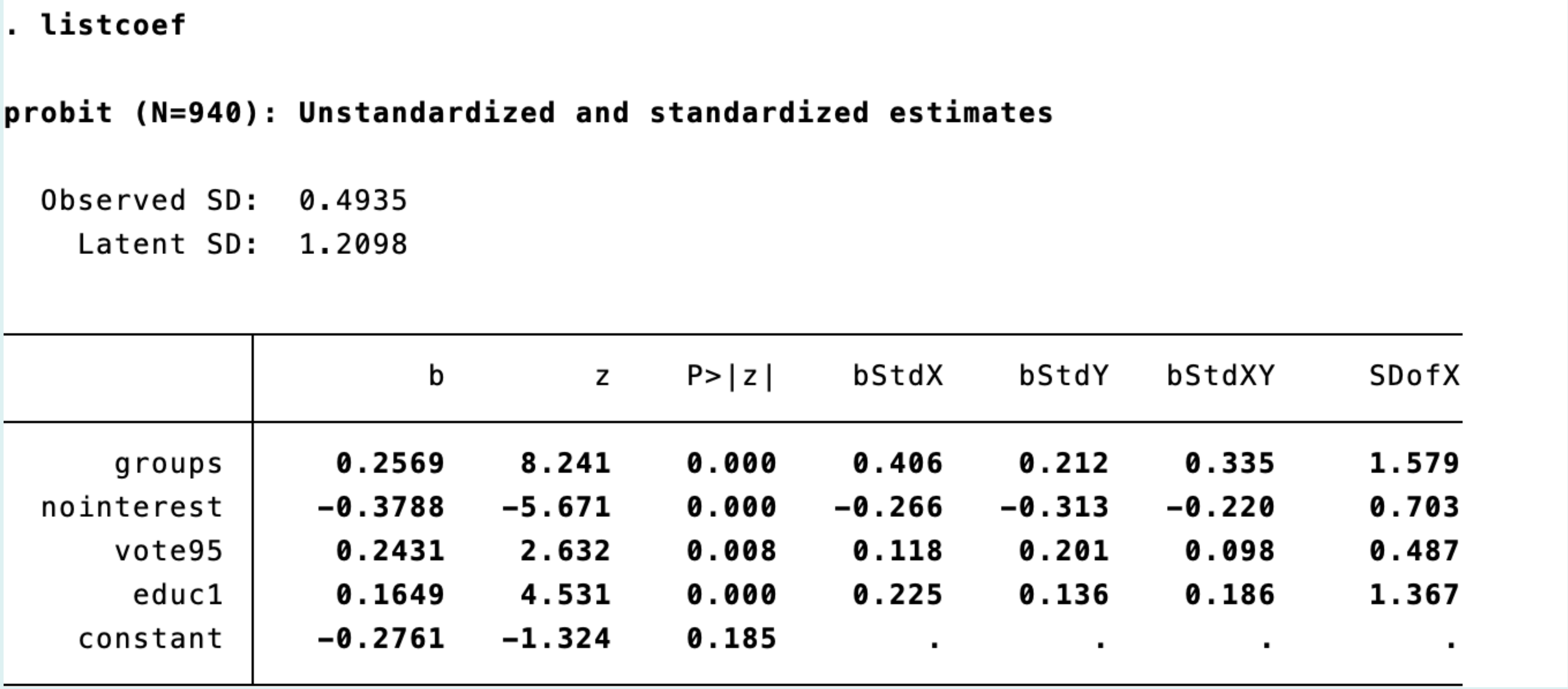
$$Y^* = X\mathrm{B} + \varepsilon$$

$$Var(Y^*) = \beta^2 Var(X) + Var(\varepsilon)$$

$$Var(Y^*) = \beta^2 Var(X) + 1$$

- This shows that adding more Xs changes the variance of Y*, since the variance of epsilon is fixed at 1 by assumption
- Long suggests *standardizing* Y* and fixing it onto a SD of 1 scale. Then you can interpret the effects of a given unit change in X, or a given standard deviation change in X, on a standard deviation change in Y. Then you will have exactly the same kind of standardized beta coefficients you have in regular regression!

- In Stata SPOST: "listcoef" after probit gives you effects of X, and standardized X on standardized Y*
  - bStdY: effect of a unit change in X on standardized Y*
  - bStdXY: effect of an SD change in X on standardized Y*
  - bStdX: effect of an SD change in X on Y*



```
. listcoef

probit (N=940): Unstandardized and standardized estimates

  Observed SD:  0.4935
    Latent SD:  1.2098

-------------------------------------------------------------------------------------
             |          b          z      P>|z|      bStdX      bStdY     bStdXY      SDofX
-------------+-----------------------------------------------------------------------
      groups |     0.2569      8.241      0.000      0.406      0.212      0.335      1.579
  nointerest |    -0.3788     -5.671      0.000     -0.266     -0.313     -0.220      0.703
      vote95 |     0.2431      2.632      0.008      0.118      0.201      0.098      0.487
       educ1 |     0.1649      4.531      0.000      0.225      0.136      0.186      1.367
    constant |    -0.2761     -1.324      0.185          .          .          .          .
-------------------------------------------------------------------------------------
```

- There are deeper consequences of the indeterminacy of the variance of Y* -- it means that group comparisons (say, between treatment and control) may be confounded by those groups having intrinsically different variances, and hence the coefficients may differ for non-causal reasons
- See Breen *et al* (2018) for further discussion; they recommend relying on Y-standardized comparisons and discrete/marginal changes in P(Y=1)
- It also makes it difficult to interpret interaction effects in latent variable models, whereby the effect of X differs for different groups represented by categories of Z. If the categories of Z have different intrinsic variances, then the interaction effects may also differ for non-causal reasons
- See Breen *et al* .(2018) for further discussion, and Rainey (2016) for other issues in the estimation of interaction effects in probit/logit

- A final consequence of the indeterminacy of the variance of Y*: it becomes very difficult to compare β for the same variable in two different models using the same data set. But: this is a very important part of the research process! We often want to include additional variables in a multiple regression format to see how the effect of a given variable changes when we "control" for other variables that may confound the process.
- In the latent variable framework, adding new variables can change the β simply because the scale (variance) of Y* changes, not because we've controlled for confounders and have better isolated the "causal" effect.
- What to do? Can we separate changes in β that are due to scale changes from changes in β that are due to confounding?

- Karlson, Holm and Breen (2011; described in Breen *et al.* (2018)) suggest an ingenious method, implemented as KHB in Stata ("net search KHB" and install)
  - Assume you want to estimate the effect of X on Y, controlling for Z
  - Comparing the bivariate to the multivariate effect is not possible as it is in a linear regression, due to the scaling issue we are discussing
  - So KHB method: regress Z against X, then take the residuals of this regression as the proxy for Z -- by construction it is uncorrelated with X but it has the same scale as Z!!!
  - Compare the probit coefficient for X in a "reduced model" with X and the Z-residuals included and a "full model" with X and Z included. The reduction in the size of β *must* be due to confounding, not changes in the variance of Y*, since the "reduced model" already took the scale changes into account
- See the KBH example section in the do file

Rare Events and “Separation” Problems in ML Estimation of Logit/Probit Models

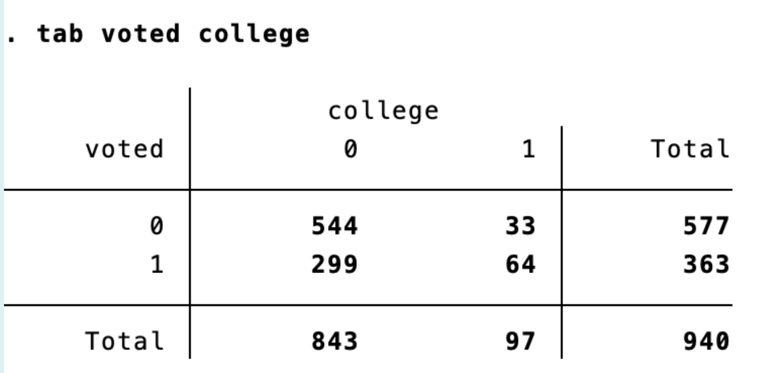
- ML estimation of logit and probit models is problematic when events are “rare”, i.e., when either the “0” or “1” outcome is infrequent
- The problem is exacerbated in small samples, so if you have a sample size of 1000 with only 20 “1”s, there will likely be problems. But the same percentage of 1s (2%) in a sample of 100,000 gives 200 1s and there likely won’t be a problem
- And even if you have a sample of 100,000, having only 20 1s *will* likely produce problems.
- So the issue is more the frequency of the “rare” event, but in small N studies the problem will likely intensify

- In extreme cases, you have what is termed **“separation”** in the data. That means you could generate a function which would “separate” the 1s and 0s perfectly under some conditions
- For example, if all college graduates voted then when X (education)>high school, Y (voted)=1. There would be 0 respondents in the cell “college graduate non-voter”
- In this case ML estimates of the effect of “college” would fail; the estimate of β for “college” could be infinitely large and still not converge to the P(Y=1) of 1 (due to the bounds on p)
- Stata would terminate the iterations and drop the college graduates from the estimation process. Not ideal!
- This example would be an instance of **“quasi-complete separation”** since knowing one category of X predicts Y perfectly; if all categories of X predict Y perfectly there is **“complete separation”**

- Another way to look at it: in a two-by-two table (say, college/no college versus vote/no vote), you can arrive at the ML estimate in a grouped data analysis as:

	No College	College
Didn't Vote	f11	f12
Voted	f21	f22

- β =ln(f11*f22/f21*f12), or what is termed the "cross-product ratio"



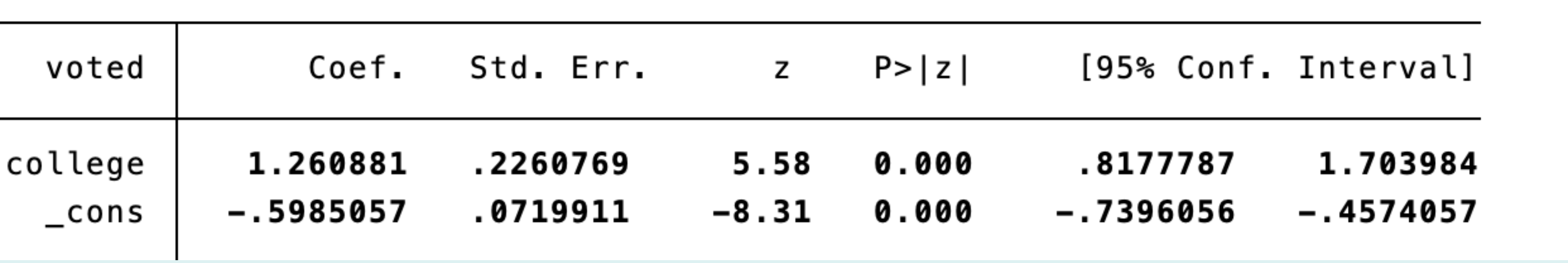
```
. tab voted college

           |       college
     voted |         0          1 |     Total
-----------+----------------------+----------
         0 |       544         33 |       577 
         1 |       299         64 |       363 
-----------+----------------------+----------
     Total |       843         97 |       940 
```

β = ln((544*64)/(299*33)) = 1.261

Log-odds of voting, no college: ln(299/544)=-.598

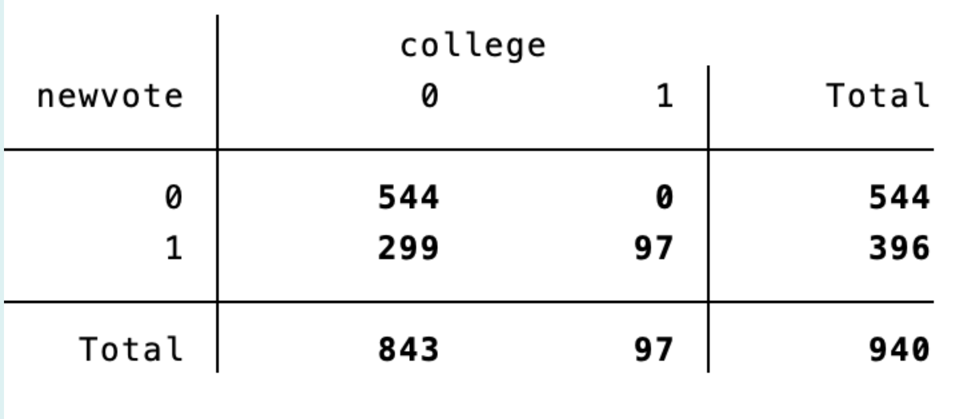
Log-odds of voting, college: ln(64/33)=.662

Difference in log-odds = 1.259 = β



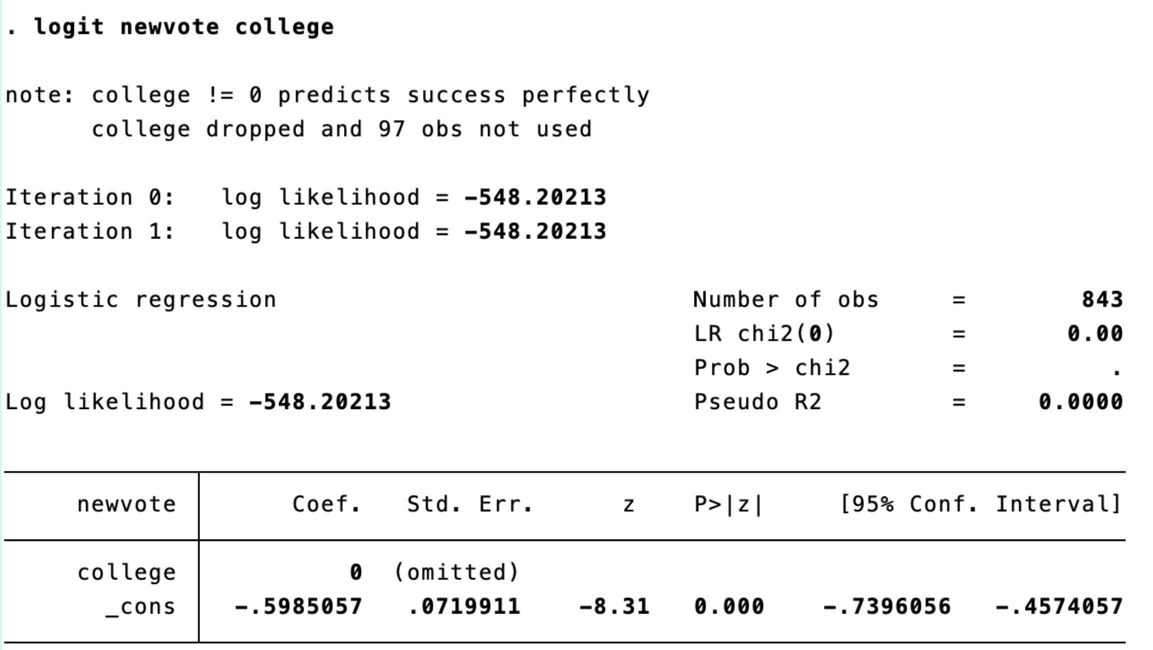
```
------------------------------------------------------------------------------
       voted |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     college |   1.260881   .2260769     5.58   0.000     .8177787    1.703984
       _cons |  -.5985057   .0719911    -8.31   0.000    -.7396056   -.4574057
```

- What happens under “separation”? Let’s say there were no college non-voters



```
           |        college
   newvote |         0          1 |     Total
-----------+----------------------+----------
         0 |       544          0 |       544
         1 |       299         97 |       396
-----------+----------------------+----------
     Total |       843         97 |       940
```

- ML estimate would then have a 0 in the denominator!!! Stata drops the cases

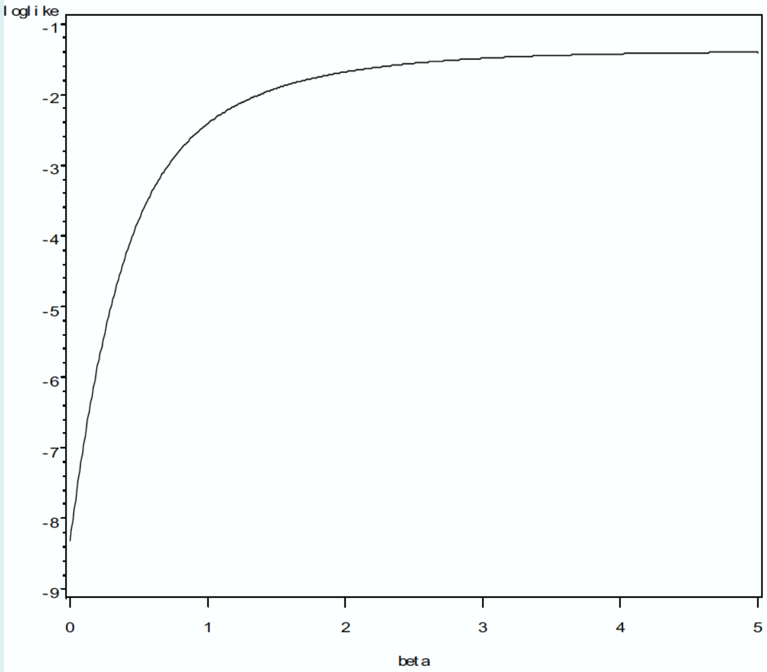


```
. logit newvote college

note: college != 0 predicts success perfectly
      college dropped and 97 obs not used

Iteration 0:   log likelihood = -548.20213
Iteration 1:   log likelihood = -548.20213

Logistic regression                             Number of obs     =        843
                                                LR chi2(0)        =       0.00
                                                Prob > chi2       =          .
Log likelihood = -548.20213                     Pseudo R2         =     0.0000

------------------------------------------------------------------------------
     newvote |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     college |          0  (omitted)
       _cons |  -.5985057   .0719911    -8.31   0.000    -.7396056   -.4574057
------------------------------------------------------------------------------
```

General rule: Whenever there is a zero in a 2x2 table there is no ML estimate!!

Separation and ML Estimation

Fig. 9.2 Log Likelihood as a Function of the Slope, Quasi-Complete Separation



If Stata didn't drop the cases, it would have generated a likelihood function looking like this – the beta for the variable goes higher and higher and never converges to a "maximum"

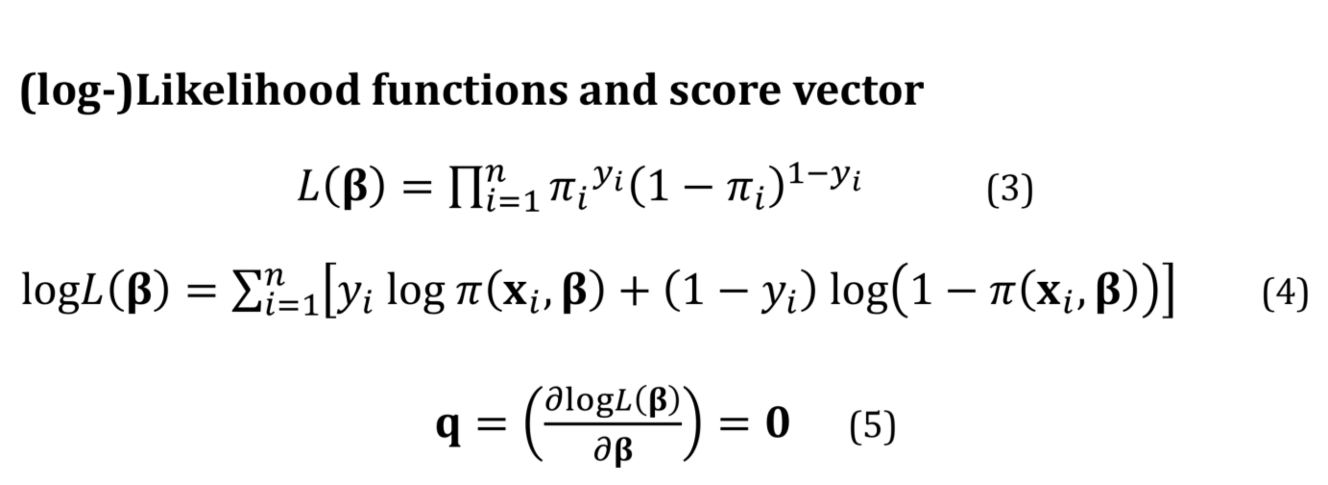
General problem: ML estimates are systematically biased away from 0 as N and the number of events (1s) get smaller and smaller. This leads to the overestimation of β!

Solutions

- Several solutions in the literature:
  - exact logistic regression (for very small samples; computationally intensive),
  - King and Zeng's (2001) "rare events logit" (Stata: relogit)
- According to the most recent literature , the preferred correction is the "Firth logit" method which relies on "penalized maximum likelihood estimation" (PMLE)
- Stata: firthlogit ("net search firthlogit" and install)

Firth Logit and PMLE

- Instead of maximizing



(log-)Likelihood functions and score vector

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{n} \pi_i^{\,y_i}(1-\pi_i)^{1-y_i} \qquad (3)$$

$$\log L(\boldsymbol{\beta}) = \sum_{i=1}^{n}\left[y_i \log \pi(\mathbf{x}_i, \boldsymbol{\beta}) + (1-y_i)\log\left(1-\pi(\mathbf{x}_i, \boldsymbol{\beta})\right)\right] \qquad (4)$$

$$\mathbf{q} = \left(\frac{\partial \log L(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}\right) = \mathbf{0} \qquad (5)$$

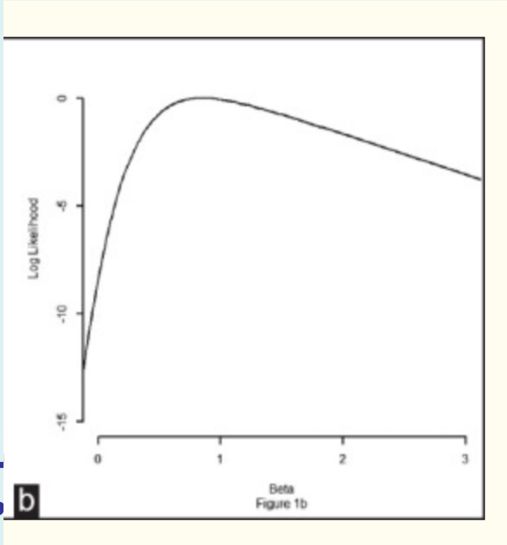
- We penalize the likelihood function by a term that is inversely related to the degree of sparseness in the data, or the "information" that is contained in the inverse of the matrix of second derivatives (i.e., flat functions lead to a larger penalty since there is less information in the data)

$$L_{PML}(\boldsymbol{\beta}) = L_{ML}(\boldsymbol{\beta})|\mathbf{i}(\boldsymbol{\beta})|^{1/2} \qquad (9)$$

$$\log L_{PML}(\boldsymbol{\beta}) = \log L_{ML}(\boldsymbol{\beta}) + 1/2 \log|\mathbf{i}(\boldsymbol{\beta})| \quad (10)$$

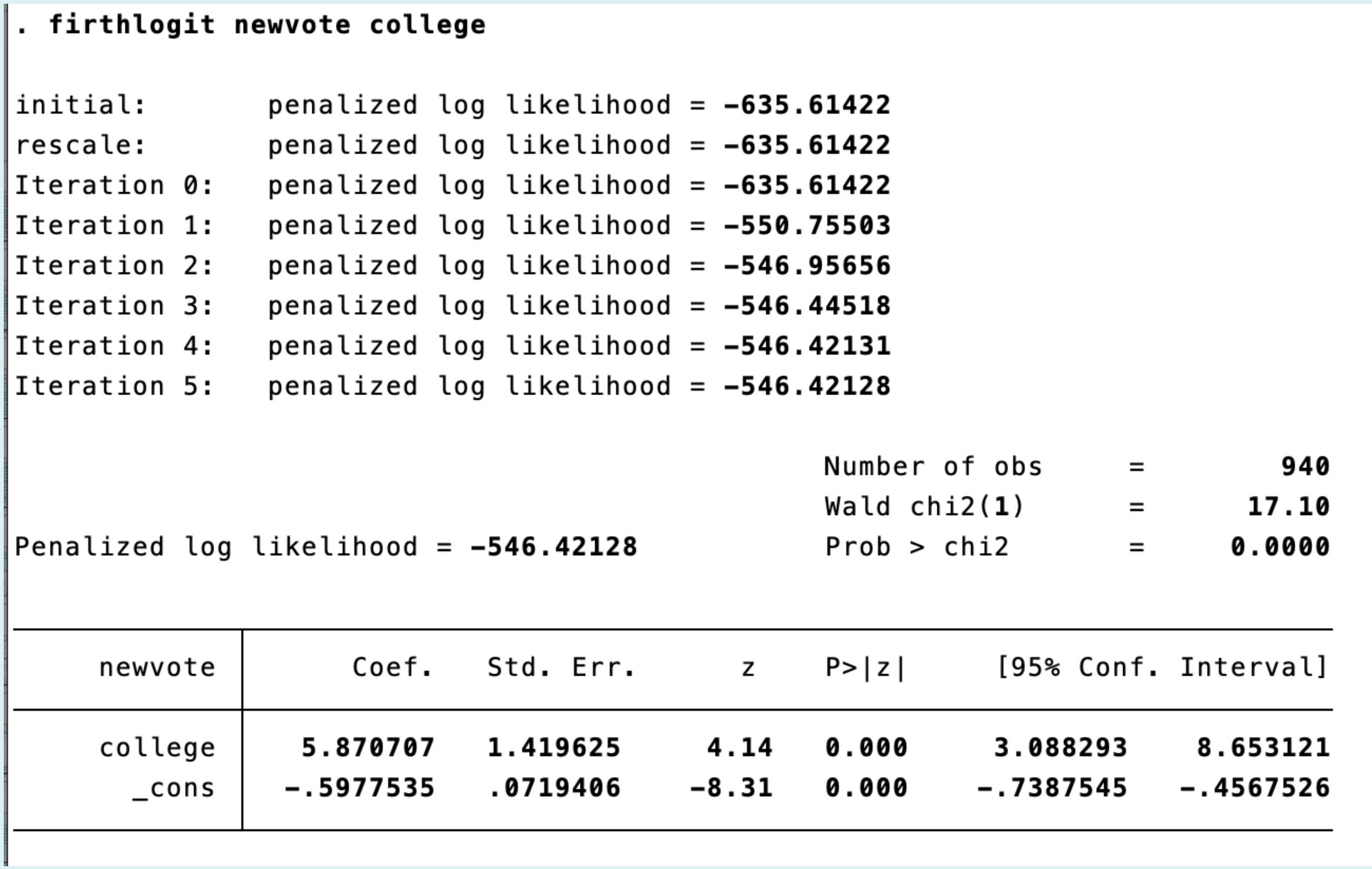
$$\mathbf{q}_{PML} = \mathbf{q}_{ML} + 1/2\, tr\left[\mathbf{i}^{-1}\left(\frac{\partial \mathbf{i}}{\partial \boldsymbol{\beta}}\right)\right] \qquad (11)$$

- We maximize this penalized function, with the ½ log|i(β)| term being the penalty related to the determinant of the "information" matrix (the inverse of the matrix of second derivatives)

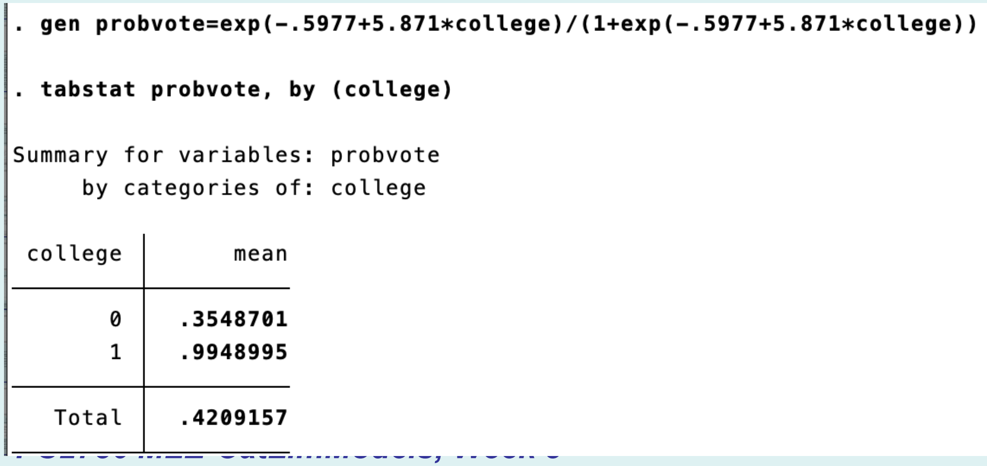


Here is a "penalized" likelihood function – it does have a maximum!

It is sometimes argued that one should **always** use PMLE – no harm done if there are no sparseness problems



```
. firthlogit newvote college

initial:       penalized log likelihood = -635.61422
rescale:       penalized log likelihood = -635.61422
Iteration 0:   penalized log likelihood = -635.61422
Iteration 1:   penalized log likelihood = -550.75503
Iteration 2:   penalized log likelihood = -546.95656
Iteration 3:   penalized log likelihood = -546.44518
Iteration 4:   penalized log likelihood = -546.42131
Iteration 5:   penalized log likelihood = -546.42128

                                                Number of obs     =        940
                                                Wald chi2(1)      =      17.10
Penalized log likelihood = -546.42128           Prob > chi2       =     0.0000

------------------------------------------------------------------------------
     newvote |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     college |   5.870707   1.419625     4.14   0.000     3.088293    8.653121
       _cons |  -.5977535   .0719406    -8.31   0.000    -.7387545   -.4567526
------------------------------------------------------------------------------
```



```
. gen probvote=exp(-.5977+5.871*college)/(1+exp(-.5977+5.871*college))

. tabstat probvote, by (college)

Summary for variables: probvote
     by categories of: college

 college |      mean
---------+----------
       0 |  .3548701
       1 |  .9948995
---------+----------
   Total |  .4209157
--------------------
```