PS2030

Political Research and Analysis

Unit 3: Models for Non-Continuous Dependent Variables

1. Introduction to Logit and Probit Models

Spring 2025, Week 9

WW Posvar Hall 3600

Professor Steven Finkel

Many important DVs in political and social sciences are not continuous and OLS is no longer appropriate. We need alternatives.

Examples of Categorical and Limited Dependent Variables

- Dichotomous Variables: two categories (e.g., vote/abstain, war/no war)
- Ordinal Variables: more than two ranked categories without necessarily equal distance between the categories (e.g. "low" "middle" or "upper" class)
- Multinomial Variables: more than two unranked categories (e.g., post-PhD career choice – academics, policy, government, other)
- Count Variables: more than two non-negative integer categories (e.g., political participation, terrorist attacks per year)
- Censored Variables: continuous up to (or down to) a threshold (e.g., demand for an undergraduate lecture class)
- Sample-Selected Variables: continuous but observed only when another variable is at specific values (e.g. wages observed only if individuals are employed; survey responses filtered by an initial income question to set a threshold for inclusion to have other questions asked)

Modeling Dichotomous Dependent Variables

- With a dichotomous dependent variable (coded as 0/1), you are modeling the probability that Y=1, or P(Y=1). This follows from the idea of regression as modeling the Expected Value or the conditional mean of Y, given the Xs:

$$E(Y \mid X) = X\mathrm{B}, \quad \textit{where}$$

$$X\mathrm{B} = \beta_0 + \beta_1 \mathrm{X}_1 + \beta_2 \mathrm{X}_2 + \ldots \beta_\mathrm{k} \mathrm{X}_\mathrm{k}$$

- This means that the average value of Y, given the Xs, is a linear function of the Xs.
- What is the “Average” Value of a Dichotomous Variable?
- If you have J cases on 1 and N-J cases on 0, then:

$$E(Y) = (J*1 + (N-J)*0)/N$$

$$E(Y) = J/N$$

- Which is equal to the proportion of cases on 1, or the probability that Y=1, or P(Y=1)

- So with dichotomous dependent variable regression, the initial formulation from the previous slide corresponds to:

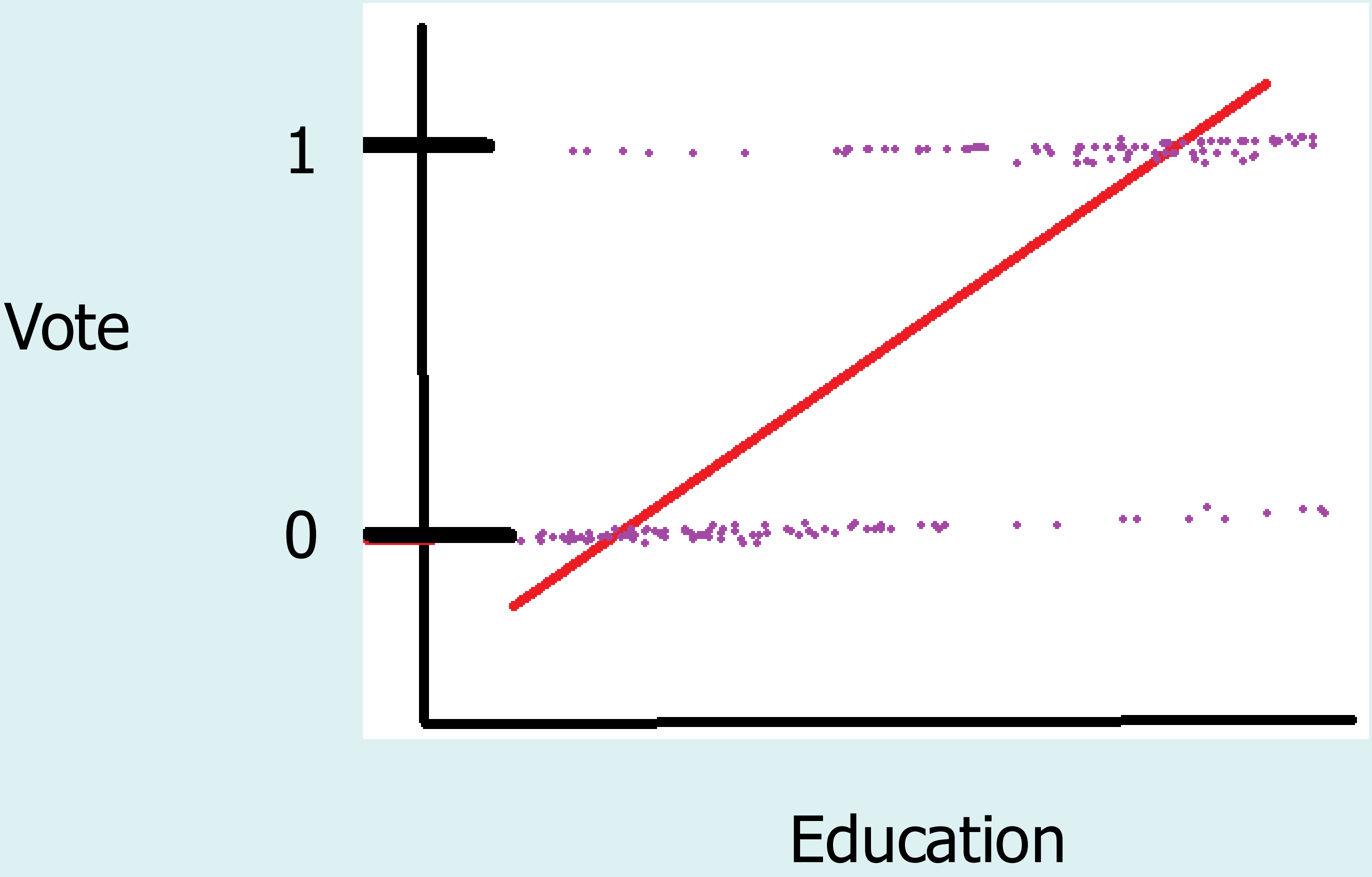
  $$P(Y=1|X)=X\mathrm{B}$$

  in bivariate case:

  $$P(Y=1|X_1)=\beta_0+\beta_1 X_1$$

- This is called the “Linear Probability Model,” as we assume that P(Y=1) is a linear function of the Xs – the effect of a unit change in X will have the same effect on P(Y=1) regardless of where on X the change takes place.
- This is just a regular old regression with the 0/1 Dichotomous variable as the dependent variable, predicted from a series of X independent variables
- Interpretation: For every unit change in X, the probability of Y being equal to 1 increases on average by β units

The Linear Probability Model (LPM)



Estimating the LPM

- Can we estimate the LPM with OLS?
- Error Term is odd: at any value of X_i, there are only 2 values for ε
  - If Y=0, then 0=XB+ε, and ε = -XB
  - If Y=1, the 1=XB+ε, and ε = 1-XB
  - So ε is *not* normally distributed
  - But with large N that still will not adversely affect estimation
- OLS would still be unbiased, no reason to think that, on average, E(ε)≠0, nor that E(Xε)≠0 (i.e., no reason to suspect endogeneity simply on the basis of the model choice itself)
- However, OLS will be *inefficient*, as there is intrinsic heteroskedasticity in the model
- You can see this from the graph: there is greater error variance at middle values of P(Y=1) than when P(Y=1) is very large or very small

- This corresponds to the variance of a dichotomous variable, which is P(1-P).
- So: Var(ε)=Var(Y|X)=P(Y=1|X)(1-P(Y=1|X)=XB(1-XB)
- This quantity is largest when P(Y=1|X)=.5, as can be seen from the graph, and will be small when, e.g. P(Y=1|X)=.1 or .9
- This problem can be overcome with a Weighted Least Squares procedure attributed to Arthur Goldberger in the 1960s
- Steps in Goldberger WLS
  - Estimate the LPM with OLS, generate the predicted probabilities $\hat{Y}_i$
  - Calculate weights as $\hat{w}_i = \sqrt{\frac{1}{X\hat{\mathrm{B}}(1-X\hat{\mathrm{B}})}}$ or $\sqrt{\frac{1}{\hat{Y}(1-\hat{Y})}}$
  - Run a weighted regression as:

$$\hat{w}_i Y = X\mathrm{B}\hat{w}_i + \hat{w}_i\varepsilon$$

  - Variance of this new error term is 1, so homoskedastic

Problems with LPM

- Possible predicted P(Y=1) outside of the 0-1 range of logical probabilities. There is no constraint or bound on Y in the LPM
- This affects the first stage of Goldberger's WLS procedure also, and would invalidate the construction of *w* for any case with Y-hat greater than 1 or less than 0, since there would a negative value in the denominator of *w* (no square root possible)
- Most important: theoretically it may not be the case that X has a constant effect on P(Y=1), rather there may be marginal decreasing effects on X as the prior P is very high or very low (e.g., one additional year of graduate school has less effect on voter turnout than changing from no high school degree to one year of college).

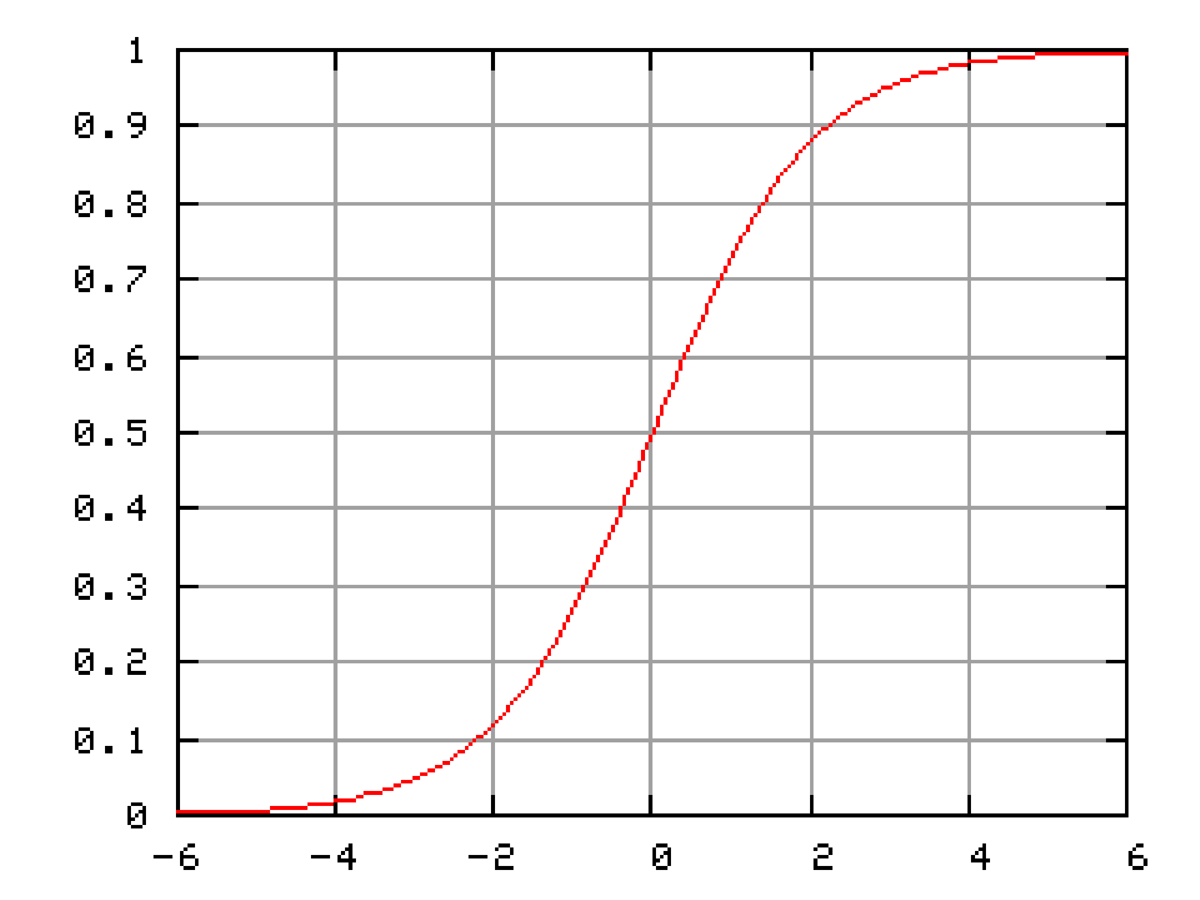
- This is an issue with the **functional form** of the LPM

So we want a model that has a **non-linear functional form** of the effects of X on the P(Y=1), where:

- the P(Y=1) are bounded by 0 and 1
- the Xs are unbounded, i.e., can take on any value
- the effects of the Xs are greater at middle levels of the distribution than at the tails

This is the justification for using the CUMULATIVE LOGISTIC FUNCTION -- or some other "sigmoid" function ("S-shaped") such as the cumulative standard normal distribution, as in probit analysis -- as the basic functional form of a binary or dichotomous variable model

The (Cumulative) Logistic Function



The Logit Model

$$P(Y=1|X)=\frac{\exp(X\mathrm{B})}{1+\exp(X\mathrm{B})}$$

So as XB goes to ∞, P(Y=1) goes to 1 but never gets there;

as XB goes to -∞, P(Y=1) goes to 0 but never gets there;

when XB is 0, P(Y=1)=.5.

So we have a perfectly symmetrical but non-linear functional form with the nice theoretical properties we wanted

Estimation of the Logit Model

- It can be shown that the probability of Y being "0" or (1-P(Y=1)) =

$$\frac{1}{1+\exp(X\mathrm{B})}$$

- Given this, we can construct the quantity **P(Y=1)/P(Y=0)** - what is called the **"odds"** of Y being 1 -- as:

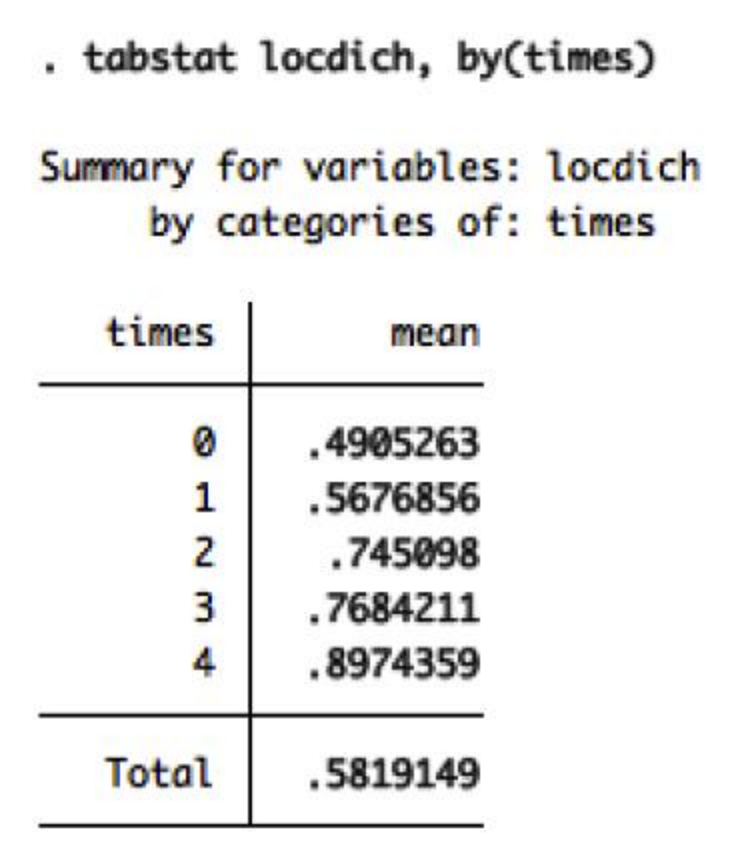
$$\frac{P(Y=1)}{P(Y=0)} = \frac{\frac{\exp(X\mathrm{B})}{1+\exp(X\mathrm{B})}}{\frac{1}{1+\exp(X\mathrm{B})}} = \exp(X\mathrm{B})$$

- And taking the natural logarithm of both sides (to the base "e") gives:

$$Ln\frac{P(Y=1|X)}{P(Y=0|X)}=(X\mathrm{B})=\beta_0+\beta_1\mathrm{X}_1+\beta_2\mathrm{X}_2+\ldots\beta_\mathrm{k}\mathrm{X}_\mathrm{k}$$

We call the log of the odds that P(Y=1) the **"logit"** of Y, and so we can say that the logit model is *linear in the logits*, such that an increase of a unit in X produces a constant change in the *logits* but is non-linear in the probabilities (and odds). This is how we interpret the estimated β effects, as **linear effects of a unit change in X on the change in the log-odds that P(Y=1)**.

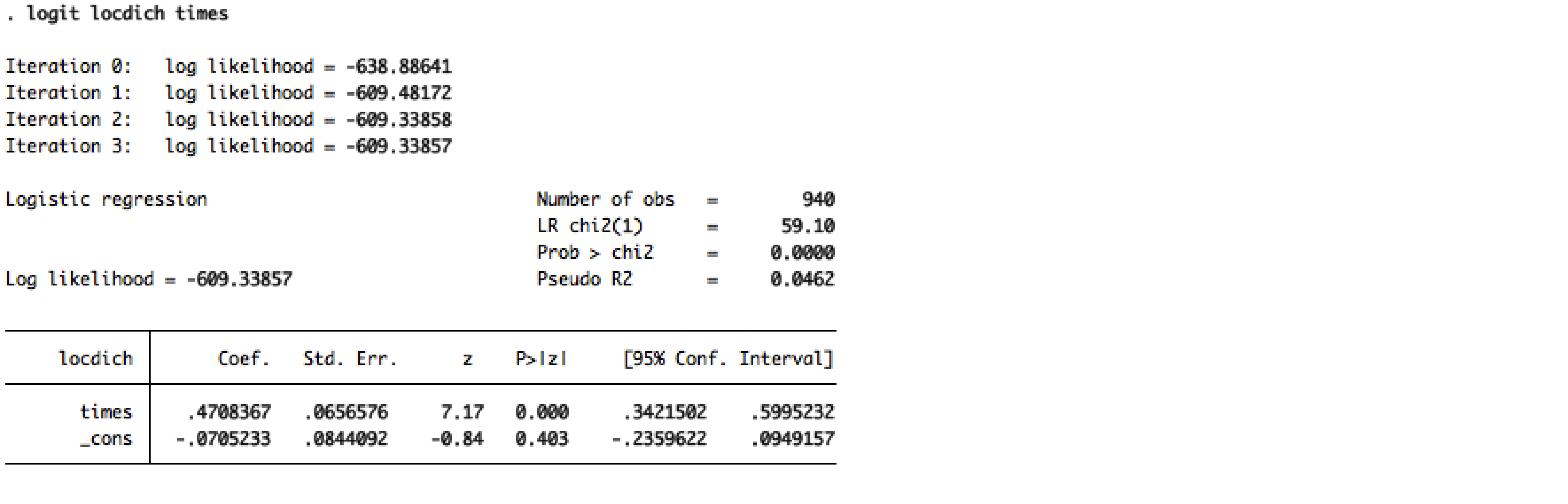
Example of Logits and Probabilities



```
. tabstat locdich, by(times)

Summary for variables: locdich
     by categories of: times

   times |      mean
---------+----------
       0 |  .4905263
       1 |  .5676856
       2 |   .745098
       3 |  .7684211
       4 |  .8974359
---------+----------
   Total |  .5819149
---------+----------
```

- DV: Dichotomized Local-level Political Participation (YES/NO)
- If no civic education exposure
  - Probability = .49
  - Odds = .49/.51=.94
  - Log-odds (logit)= ln(.94) = -.062
- If three times exposure
  - Probability=.77
  - Odds=.77/.23=3.35
  - Log-odds (logit)=ln(3.35)=1.21
- All log-odds (logits) less than 0 correspond to probabilities **less than** .5 (and **odds <1**)
- All log-odds (logits) more than 0 mean probabilities **greater than** .5 (and **odds>1**)
- Logistic regression models the logits as a linear function of the Xs, using maximum likelihood estimation methods

Bivariate Logistic Regression



```
. logit locdich times

Iteration 0:   log likelihood = -638.88641
Iteration 1:   log likelihood = -609.48172
Iteration 2:   log likelihood = -609.33858
Iteration 3:   log likelihood = -609.33857

Logistic regression                               Number of obs   =        940
                                                  LR chi2(1)      =      59.10
                                                  Prob > chi2     =     0.0000
Log likelihood = -609.33857                       Pseudo R2       =     0.0462

------------------------------------------------------------------------------
     locdich |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       times |   .4708367   .0656576     7.17   0.000     .3421502    .5995232
       _cons |  -.0705233   .0844092    -0.84   0.403    -.2359622    .0949157
------------------------------------------------------------------------------
```

- The logit coefficient here for the effect of "times exposed to civic education" on the dichotomous outcome "engaged in local participation" is .47.
- This means: for every additional CE exposure, the log-odds, or logit, of local participation changes, on average, by .47.
- This is a constant linear effect – it is the same when changing from 0-1 exposures, 1-2, 2-3, 3-4, etc.

- To convert into probabilities: we take the predicted logit for an individual with a given value of X, and plug it into the P(Y=1) expression for the logit model:

$$P(Y=1|X)=\frac{\exp(X\mathrm{B})}{1+\exp(X\mathrm{B})}$$

- So when X=0, the predicted logit: -.07 exp(-.07)/(1+exp(-.07))= .48
  X=1, the predicted logit: .40 exp(.40)/(1+exp(.40))=.60
  X=2, the predicted logit: .87 exp(.87)/(1+exp(.87))=.70
  X=3, the predicted logit: 1.34 exp(1.34)/(1+exp(1.34))=.79

- A unit change in X from 0-1 leads to a .12 change in predicted P(Y=1)
  A unit change in X from 1-2 leads to a .10 change in predicted P(Y=1)
  A unit change in X from 2-3 leads to a .09 change in predicted P(Y=1)

The Latent Variable Approach to Modeling Binary Dependent Variables

- Derivation of the logit model was done so far from the need for a non-linear probability model that was bounded by 0,1 with no bounds on the Xs
- Another way of deriving the non-linear functional form for predicting a 0,1 dependent variable is based on a "latent variable" approach. This usually ends up with the "probit" specification which makes use of the normal distribution (though one could also specify a logistic distribution using this approach and arrive again at the logit model)

- Idea is that you have a 0,1 **observed** variable: vote or not vote; protest or don’t protest, war/no war but there is an underlying, latent **“propensity”** to vote, to protest, to go to war which is a continuous, unobserved variable.
- So can imagine that the latent “propensity” variable might run from negative infinity to infinity, and that there is some threshold point beyond which we observe a voter, a protester, or a conflict. So can view the observed 0,1 variable as **mapped from a continuous latent, unobserved variable that has no bounds.**
- This also fits the notion of **“Expected Utility”** models of behavior perfectly: the utility derived from one behavioral choice versus another can be infinitely negative or positive, and at the threshold of (say) zero you observe behavior “1”, and below the threshold you observe behavior “0”
- Many discrete choice models -- ordinal variables, multinomial variables, and others you will see in PS2040 Maximum Likelihood Methods – are derived from this framework

- Model:

$$Y_i^* = \Sigma \beta X_i + \varepsilon_i$$

$$Y_i^* = X\mathrm{B} + \varepsilon_i \quad \text{in matrix type notation}$$

$$\mathrm{E}(Y_i^* \mid X) = X\mathrm{B}$$

- Y* is a continuous *unobserved* variable. It is mapped to the observed dichotomous variable Y through a "measurement equation" that says if Y* is above a certain threshold τ, then the observed Y will be 1; if Y* is below the threshold, then observed Y will be 0

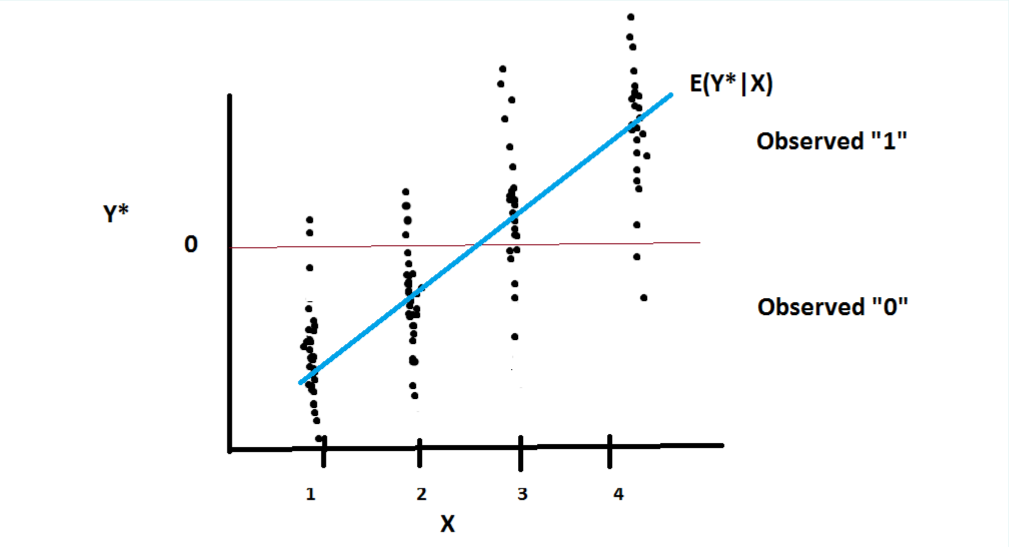
$$Y_i = 1 \text{ if } \mathrm{Y}_i^* > \tau$$

$$Y_i = 0 \text{ if } \mathrm{Y}_i^* \leq \tau$$

- Assuming τ to be 0 (following the logic above), then

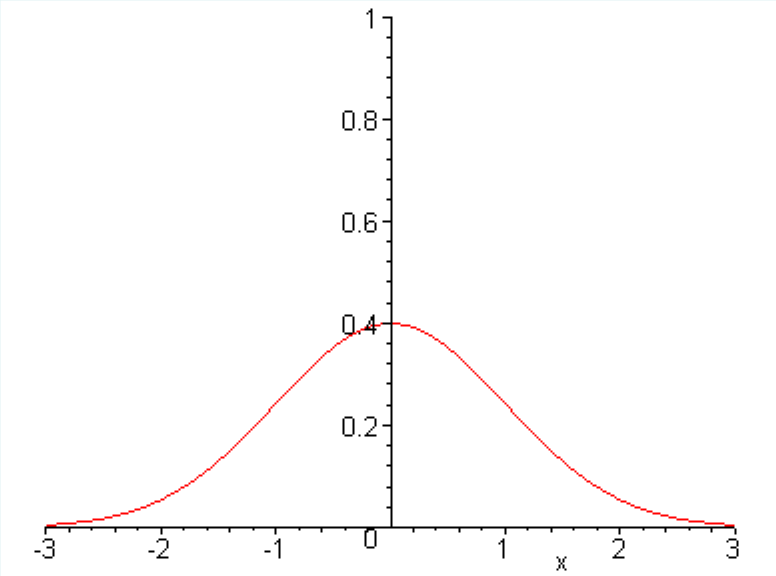
$$Y_i = 1 \text{ if } \mathrm{Y}_i^* > 0$$

$$Y_i = 0 \text{ if } \mathrm{Y}_i^* \leq 0$$



- Y* is unobserved, so we can't estimate with OLS. We use ML methods, which we will introduce next week. For now, need to make some assumptions about the error term ε in this model in order to identify the model parameters
- Assume that ε is a standard normal variable with mean of 0 and variance of 1
- This identifies the variance of Y*, which is unobserved. (And we can arbitrarily make this assumption with no substantive implications – it only changes the absolute value of the regression coefficient but not the substantive relationship)
- So ε~N(0,1)
- We could also say that ε is distributed logistically. Then var(ε)=π/3

The Standard Normal Distribution

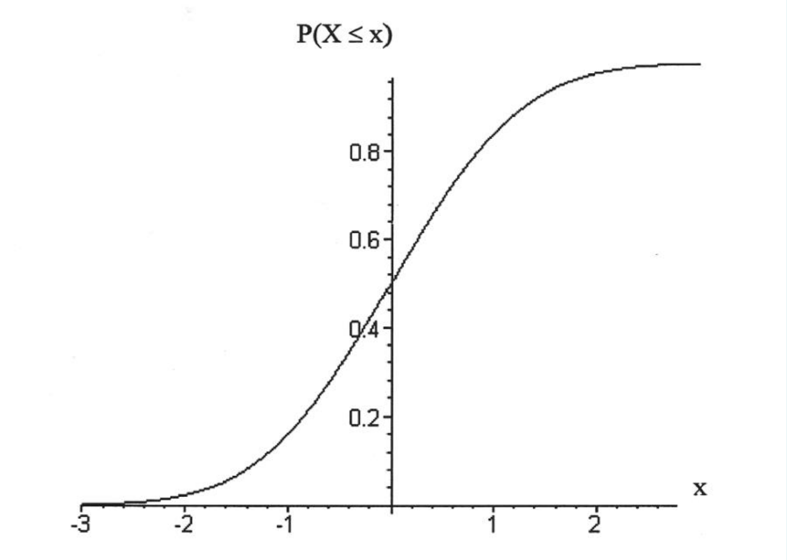


$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-u}{\sigma}\right)^2}$$

for $-\infty < x < \infty$

- This is a graph of the frequency distribution (technically, the probability density function or pdf) of the values in a standard normal variable.
- On the X axis are the "z-scores" of *any* normally distributed variable
- The f(x) gives you the height, or "pdf" of the curve at any given point on X
- There is actually not a probability value associated with observing X at any single discrete point since this is a continuous function, but there are probabilities associated with observing Xs *between* two points on the pdf

The Cumulative Normal Distribution Function



z-score

Probit models will give coefficients that indicate the average change in the z-score of Y for a given unit change in X, and this translates into changes in the probability that Y=1 via the cdf

- The cumulative normal distribution function (the "cdf") gives the proportion of the standard normal curve at or below a given value
- So a z-score of 0 corresponds to a cdf of .5; this means that .5 of the normal curve is at or below a z-score of 0
- A z-score of -1 corresponds to a cdf value of .16, or 16% of the function
- A z-score of 1 corresponds to a cdf value of .84. or 84% of the function
- These cdf values correspond also to the **cumulative probabilities** of observing values in the distribution at or below the given value. So a z-score of 1 has a cumulative probability of .84, e.g.
- We represent this as, e.g., $\Phi(0)=.5$
- $\Phi(-1)=.16$ $\Phi(1)=.84$ $\Phi(.5)=.69$

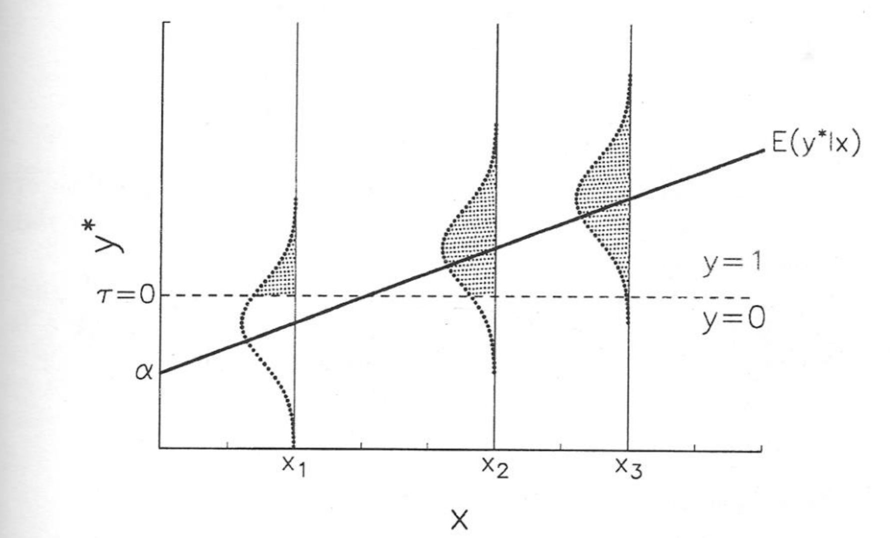


Figure 3.2. The Distribution of y^* Given x in the Binary Response Model

- Back to our probit derivation. Assuming normally distributed errors in the equation for Y*:
- Can see that sometimes, for a given level of X (and thus a given value of predicted Y*), the person will have an **observed** value of 1 if the error term is sufficiently **large** to push her over the threshold τ (0). If not, we observe 0.
- Can also see that the probability of obtaining an error term large enough to push a person over the threshold is greater, as X increases (given the positive relationship here between XB and Y*). When X=3, e.g., it will be very unlikely to get an error sufficiently negative to push the person *under* the τ (0) threshold, so the probability of observing a 1 will be very high
- But even if Y* is very low (high), there is still a chance that a very high (low) ε leads to an observed 1 or 0.
- With the assumption of a normally distributed error term, can calculate the probabilities exactly!

The Probit Model

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

$$P(Y = 1 | X) = P(Y^* > 0 | X)$$

$$P(Y = 1 | X) = P(X\mathrm{B} + \varepsilon > 0 | X)$$

$$P(Y = 1 | X) = P(\varepsilon > -X\mathrm{B} | X)$$

- So the probability that Y=1 is equal to the probability of obtaining an error term greater than –XB, which will push (or keep) the person over the threshold of 0
- If XB puts the person at -1, for example, we will observe a "1" only if the error term is greater than 1 (i.e., greater than –XB), which would make Y* greater than 0.
- Given normal curve probabilities, we know that will happen with .16 probability. (How?).
- If XB puts the person at 1 on Y*, we will observe a "1"on the outcome if the error term is greater than -1 (i.e., greater than –XB). We know that will happen with .84 probability (since there is only a .16 chance of ε<-1)

- Notice, though, that the probability of observing an error term *greater than –XB* is the same as the probability of observing an error term *less than XB*; this follows from the symmetry of the normal distribution

  Previous slide, example 1 : XB=-1

  - $P(Y^*>0|XB=-1)$ = probability that $\varepsilon>1$ = .16, which is also the probability that $\varepsilon<-1$ =.16. **That is the cdf associated with an XB of -1 = $\Phi(XB)$!!!!**

  Previous slide, example 2: XB=1

  - $P(Y^*>0|XB=1)$ = probability that $\varepsilon>-1$ = .84, which is also the probability that $\varepsilon<1$ =.84. **That is the cdf associated with an XB of 1 = $\Phi(XB)$!!!!**

- So Probability that $\varepsilon>-XB$= Probability that $\varepsilon<XB$ = $\Phi(XB)$
- So the Probability of obtaining an error term greater than –XB, which will push Y* above the threshold of 0 so that observed Y will be 1, is equal to the cumulative probability associated with the z-score XB
- **This is the probit model!**

$$\mathbf{P(Y=1|XB)=P(\varepsilon>-XB)=P(\varepsilon<XB)=\Phi(XB)}$$

- So the probit model for binary dependent variables is:

$$P(Y=1|X)=\Phi(X\mathrm{B})$$

- The probability that Y=1 is equal to 1 is equal to the cdf – the value of the cumulative normal distribution function – associated with the z-score quantity XB
- Probit is "linear in the z-scores" and non-linear in the probabilities, just like logit was "linear in the logits" and non-linear in the probabilities

Examples of Probit XB and P(Y=1). In STATA: display normal(XB)

- Model: $Y^*=-1+.2^*X$

– $X=-1$	$XB=-1.2$	$P(Y=1)=\Phi(-1.2)=.12$
– $X=0$	$XB=-1$	$P(Y=1)=\Phi(-1)=\ .16$
– $X=5$	$XB=0$	$P(Y=1)=\Phi(0)=\ .5$
– $X=10$	$XB=1$	$P(Y=1)=\Phi(1)=\ .84$

- Change intercept to +1: $Y^*=1+.2X$

– $X=-1$	$XB=-.8$	$P(Y=1)=\Phi(.8)=.79$
– $X=0$	$XB=1$	$P(Y=1)=\Phi(1)=\ .84$
– $X=5$	$XB=2$	$P(Y=1)=\Phi(2)=\ .98$
– $X=10$	$XB=3$	$P(Y=1)=\Phi(3)=\ .99$

- Change slope to .5: $Y^*=-1+.5X$

– $X=-1$	$XB=-1.5$	$P(Y=1)=\Phi(-1.5)=.07$
– $X=0$	$XB=-1$	$P(Y=1)=\Phi(-1)=\ .16$
– $X=5$	$XB=1.5$	$P(Y=1)=\Phi(1.5)=.93$
– $X=10$	$XB=4$	$P(Y=1)=\Phi(4)=\ .99$

Summary of Logit/Probit Models

$$P(Y=1|X)=\Phi(X\mathrm{B}) \qquad \text{Probit}$$

$$P(Y=1|X)=\frac{\exp(X\mathrm{B})}{1+\exp(X\mathrm{B})} \qquad \text{Logit}$$

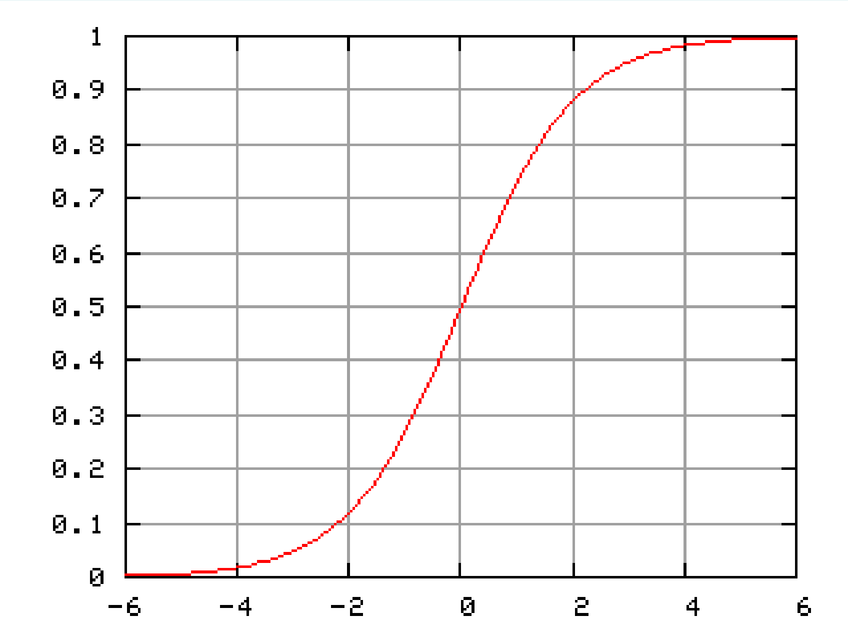
- These are both non-linear probability models bounded by 0 and 1
- In each case, though, *something* has a linear relationship to X as well
  - Probit: The Y and X relationship is linear in the z-scores
  - Logit is linear in the log-odds
- We can go from probabilities to z-scores or log-odds via the "inverse" of the P(Y=1) functions above

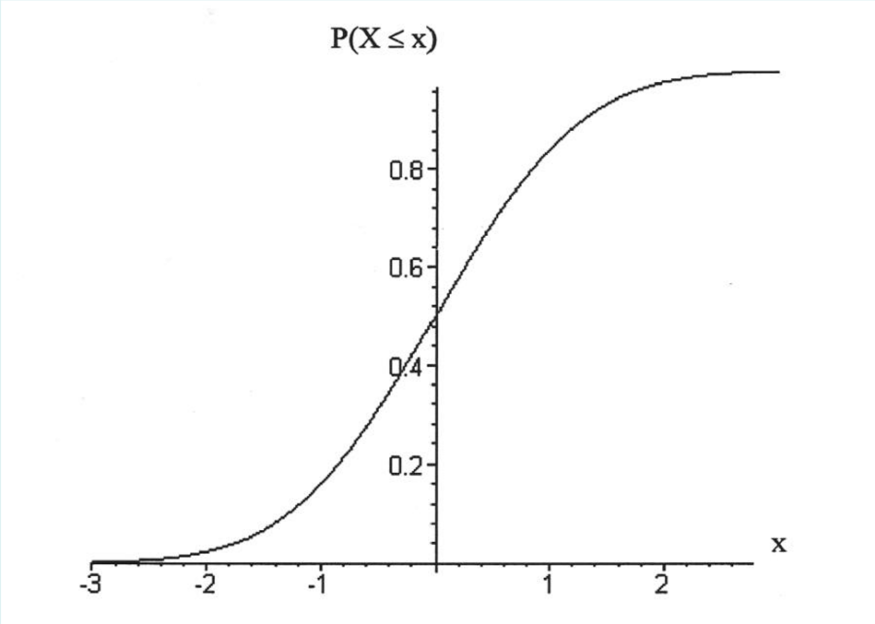
$$X\mathrm{B}=\Phi^{-1}P(Y=1|X) \qquad \text{Probit}$$

$$X\mathrm{B}=\ln(\frac{P(Y=1|X)}{1-P(Y=1|X)} \qquad \text{Logit}$$

Logit

Probit





$$\mathrm{Ln}\left(\frac{P(Y=1)}{1-P(Y=1)}\right)=\mathrm{XB}$$

$$\boldsymbol{\Phi}^{-1}(\mathrm{P}(\mathrm{Y}=1))=\mathrm{XB}$$

$$\mathrm{P}(\mathrm{Y}=1)=\frac{exp^{XB}}{1+exp^{XB}}$$

$$\mathrm{P}(\mathrm{Y}=1)=\boldsymbol{\Phi}(\mathrm{XB})$$

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. 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 Xs in these models
- General approaches
  1. Direct interpretation of the effect of β
  2. Interpretation of β in terms of changes in **odds** that Y=1 in logit
  3. Calculation of effects of "marginal" change in X on P(Y=1), or of "discrete" changes in X on P(Y=1)
  4. Interpretation of the β on changes in Y* in latent variable framework

1. Direct Interpretation of the β

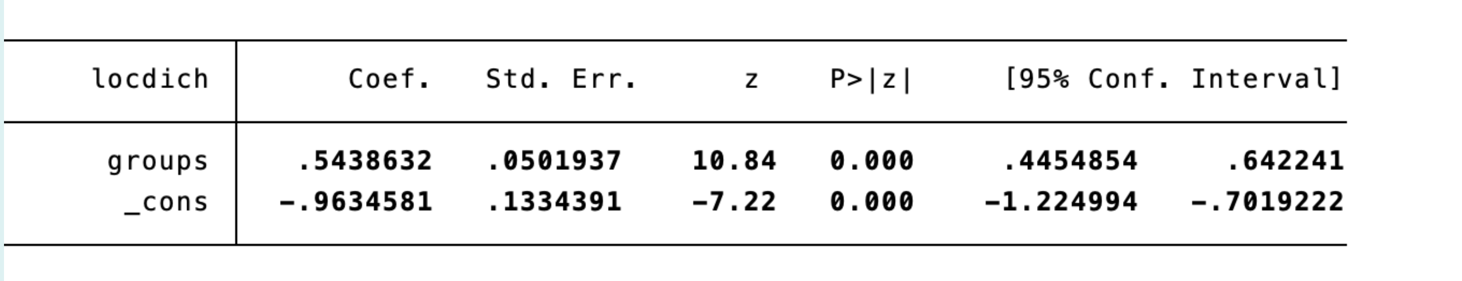
- In logit/probit one can always interpret the β as the effect of a unit-change in X on the linear quantity associated with that given model.
  - For logit, it is the "log-odds" that Y=1
  - For probit, it is the "inverse cumulative normal" or the z-score corresponding to the proportion of the cumulative 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 log-odds or z-score 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

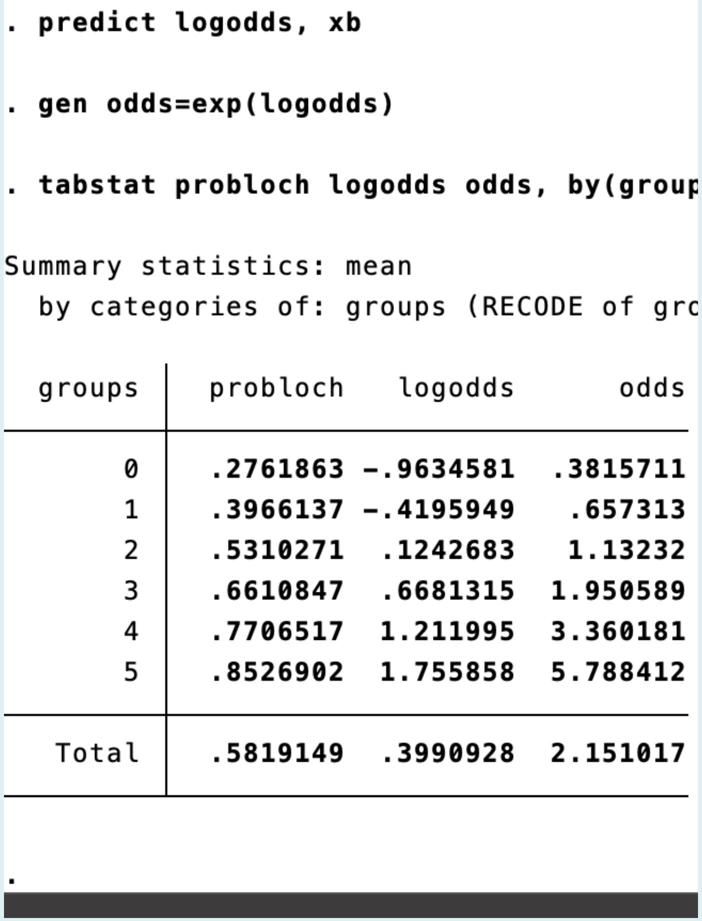
2. Odds Interpretations (in Logit)

- 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)
- Next slide: a unit increase in group memberships changes the odds that a person participates, i.e, (Y=1), by a factor of 1.723, which is ($\exp^{.54}$)
- 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 (like a Beta coefficient comparison in OLS)

Logistic regression of locdich against groups in the SA data



```
     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
----------------------------------------

.
```

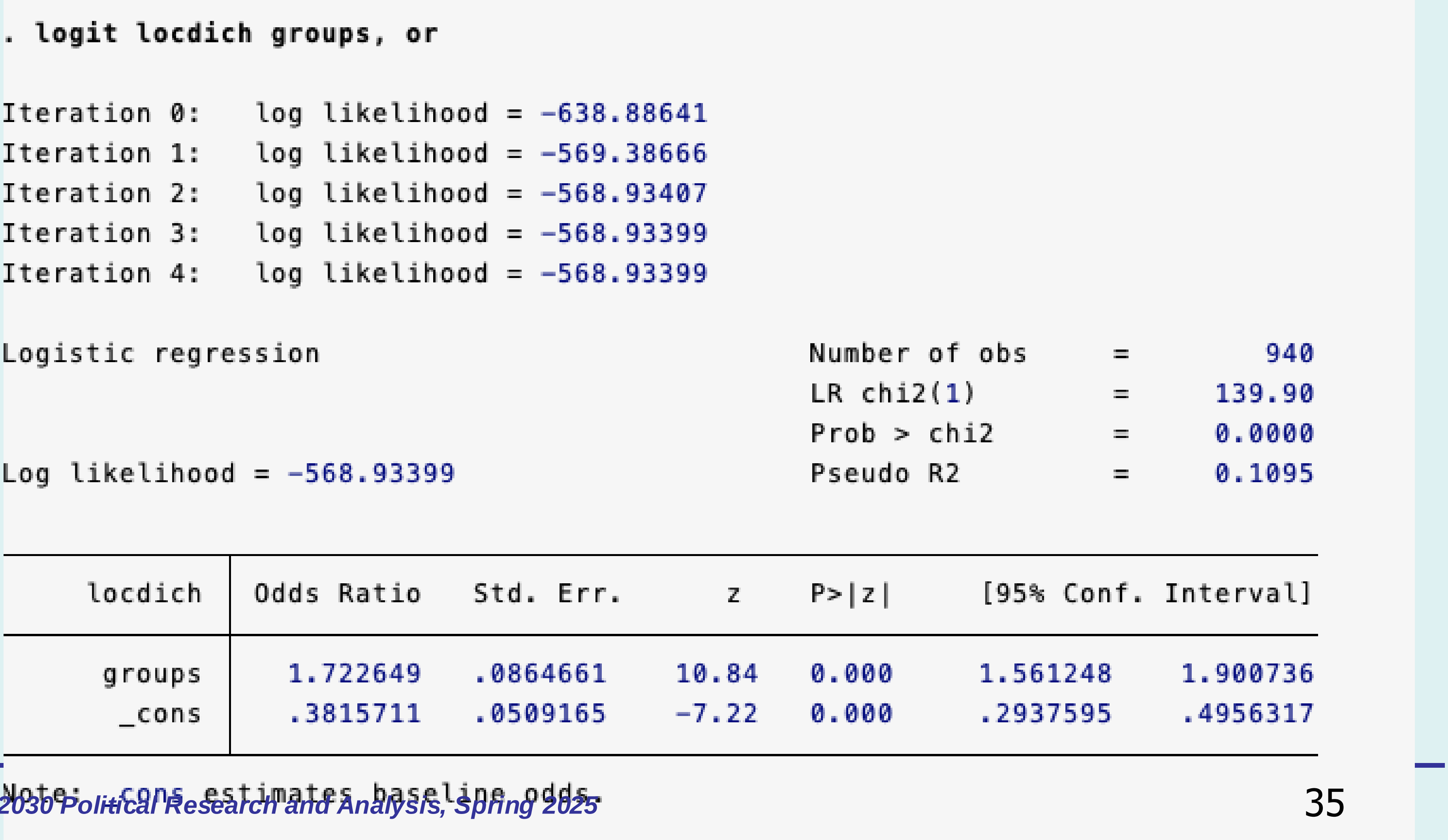
For 0 groups, ln(odds)=-.963 and P(Y=1)=.276, odds=.381

For 1 groups, ln(odds)=-.420 and P(Y=1)=.397 odds=.658

Do you see the 1.723 multiplicative change in the odds for every unit change in X?

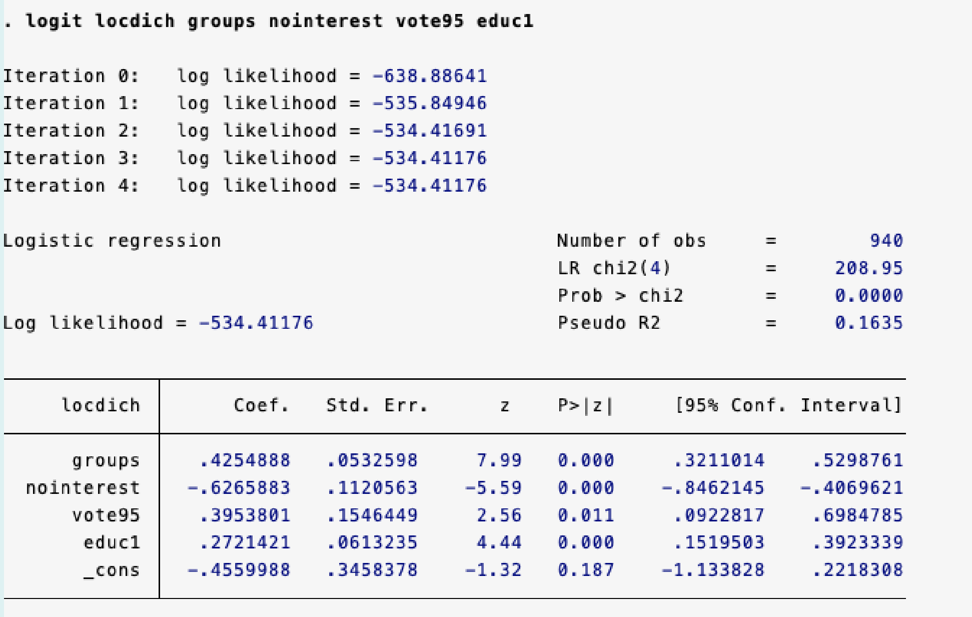
.38*1.723=.657   (from 0-1)
.657*1.723=1.132 (from 1-2)
3.360*1.723=5.789 (from 4-5)

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)

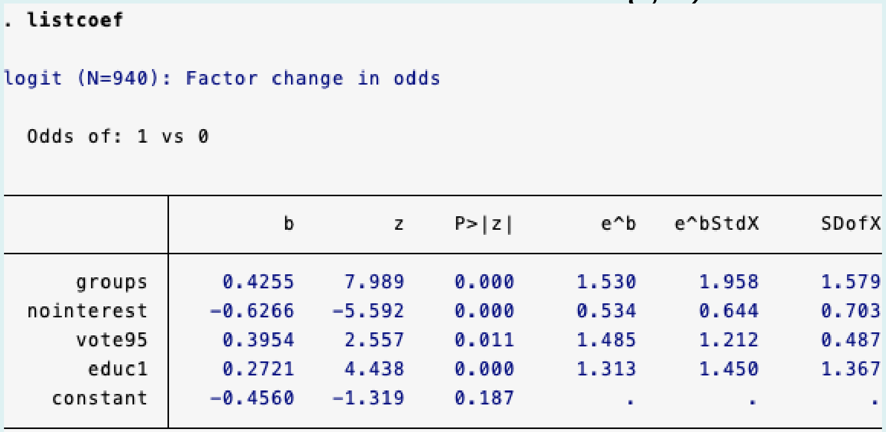


```
. logit locdich groups, or

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
------------------------------------------------------------------------------
```



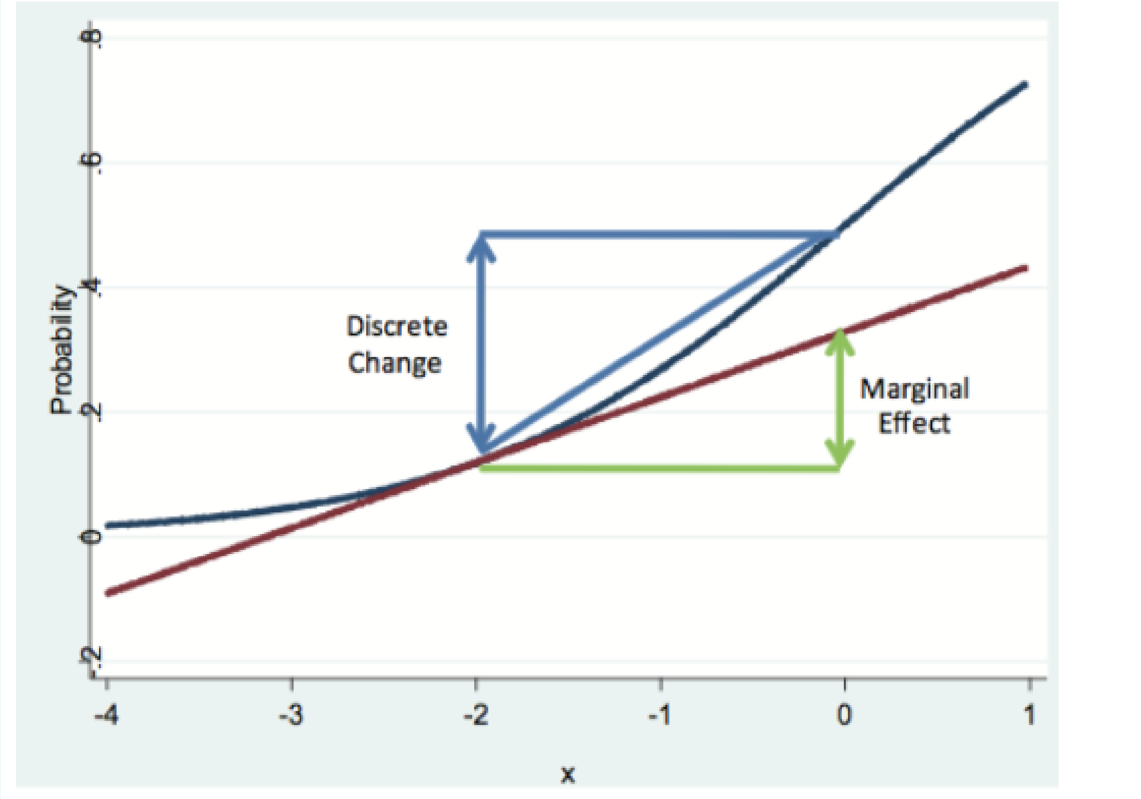
```
. listcoef

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!

3. Marginal or Discrete Effects Interpretation

- Another way to interpret effects in logit/probit models is to examine the effects of changes in X on changes in the P(Y=1), holding other variables constant. In linear regression the effect of a unit or marginal change in X on Y is β, and it is the same regardless of the levels of the other variables In non-linear models, the effect of changes in X on P(Y=1) differs, depending on the level of the other variables.
- So: where should you "hold the other variables constant" at?
- Another issue is that, in non-linear models, the "marginal" effect of changes in X is not necessarily the same as an effect of a one-unit change in X. (See next slide). There is a possible difference between "marginal" and "discrete" changes in X on P(Y=1).
- Final issue is that, in discrete change estimation, it isn't obvious what kind of "change" in X would be most informative to show changes in P(Y=1). Minimum to maximum values on X? One unit change (from where, though)? One standard deviation change?

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 or specified amount of change in X. It is calculated as

  $P(Y=1|X, X_k = \text{end}) - P(Y=1|X, X_k = \text{start})$, or $\Delta P(Y=1|X)/ \Delta X$
- 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:

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

  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)

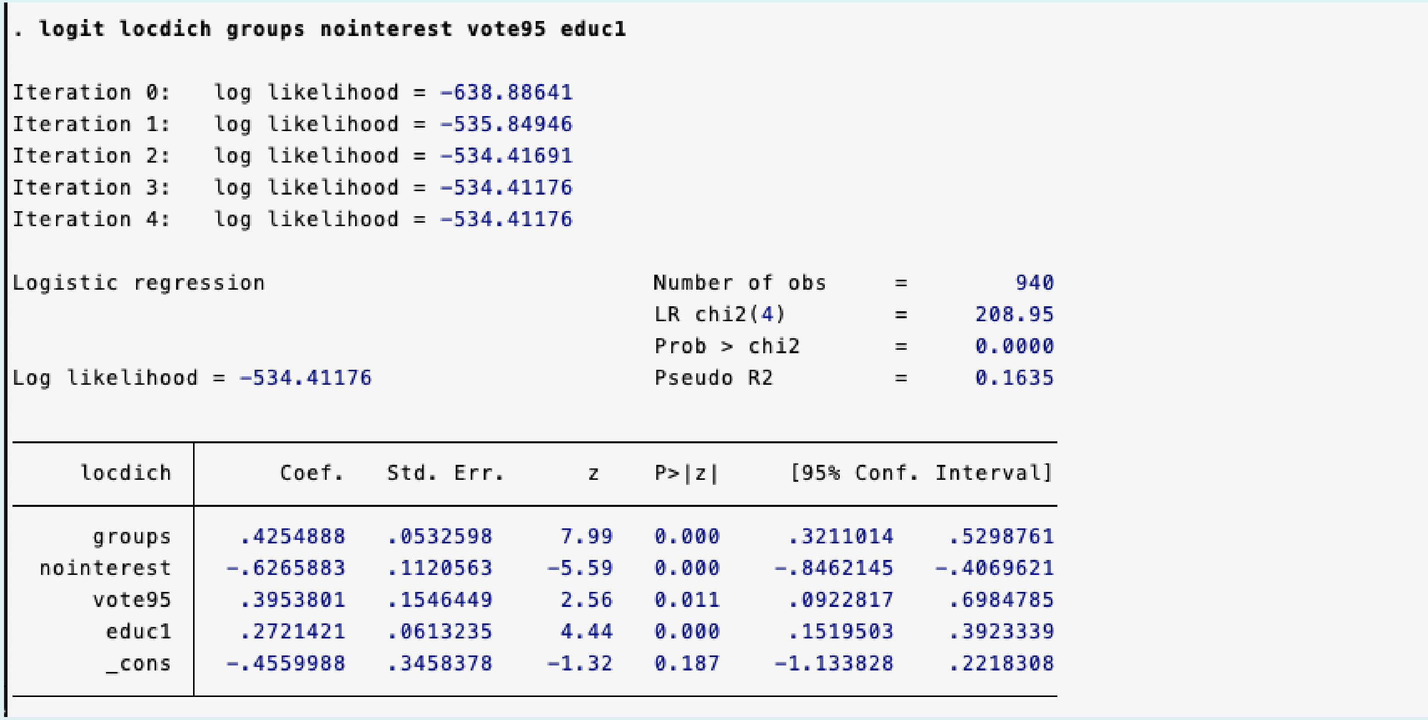
- Where shall you set the values of the other independent variables in calculating marginal or discrete changes?
- **Marginal Effect at the Mean (MEM**): set all other variables equal to their sample means and calculate either marginal or discrete changes. This used to be the standard method.
  - Advantage: 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 the estimated slope for 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 can 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 $\bar{X} \pm .5s$

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
------------------------------------------------------------------------------
```

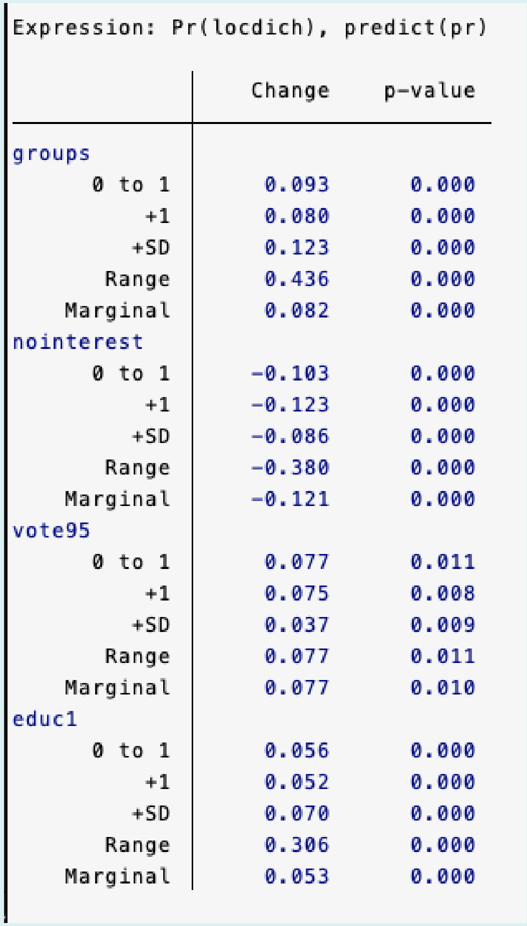
- So: unit change in groups leads to a constant .425 change in the logit of Y
- Unit change in education leads to a constant .272 change in the logit of Y
- Voters from 1995 have logits that are .395 higher than non-voters
- So calculating the effect of each variable on the probability that Y=1 means 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”

Example of “Marginal Effects at Representative Values” or **MER**

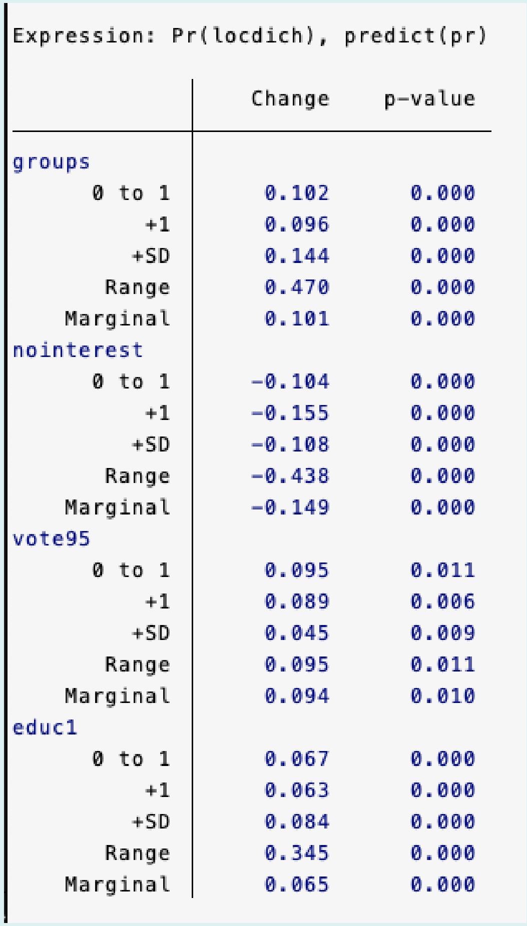
- Effect of going from 0 to 4 groups for voters, for otherwise “average” person
  - Logit= -.456 + -.627*1.92+.3954+.2721*2.880=-.481 P(Y=1)=.382
  - Logit= -.456 + .426*4+ -.627*1.92+.3954+.2721*2.880=1.22 P(Y=1)=.772
- Effect of going from 0 to 4 groups for non-voters, for otherwise “average” person
  - 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)
- Effect of going from 0 to 4 groups for a person with prior probability of .5=.35 (i.e, p goes from .5 to .85)

- 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



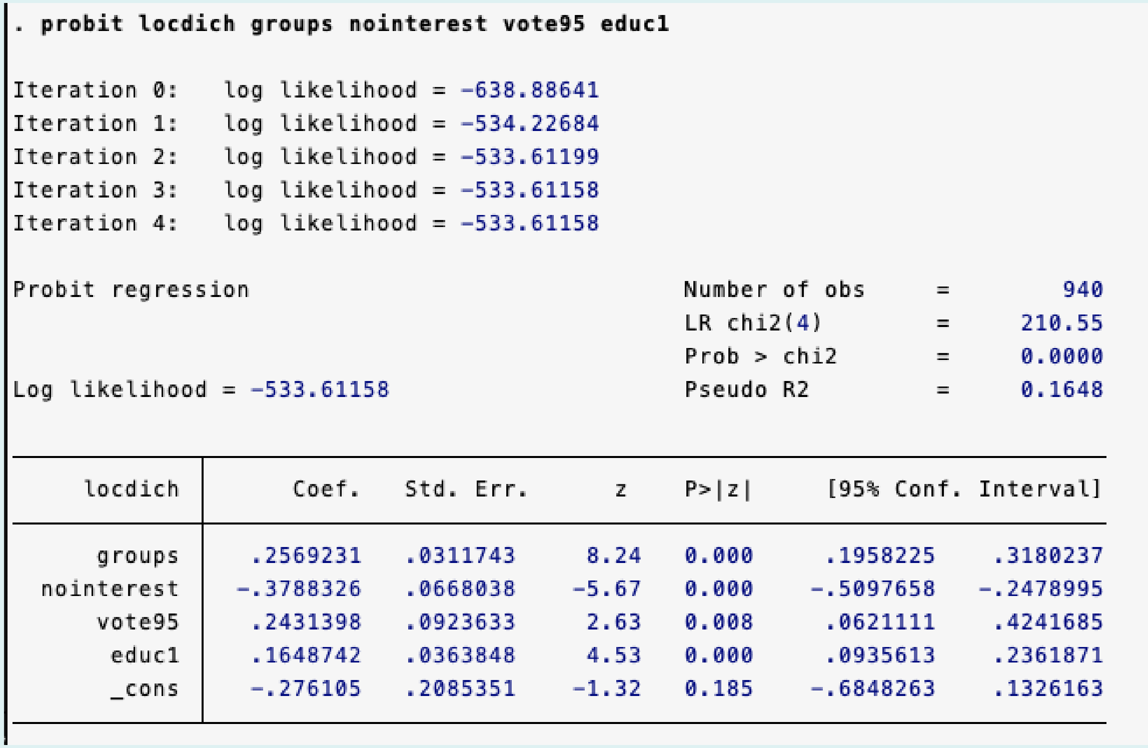
```
Expression: Pr(locdich), predict(pr)

                  |    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
```



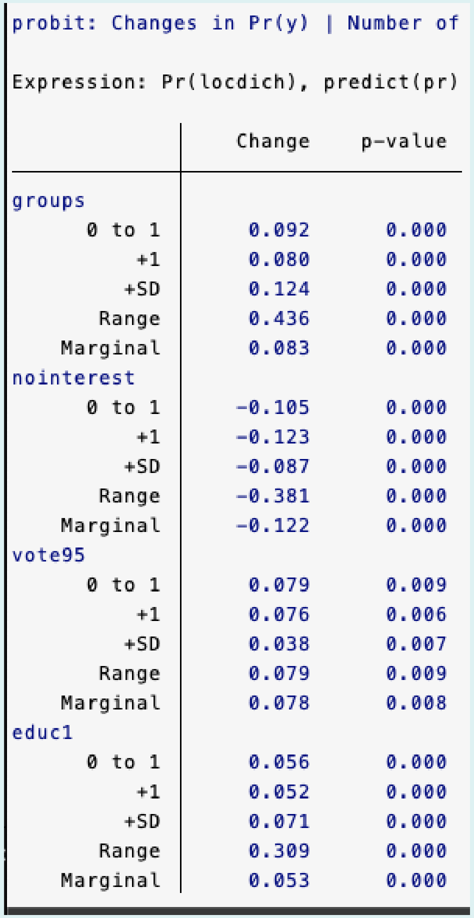
```
Expression: Pr(locdich), predict(pr)

                  |    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 logit interpretations holds for probit *except* for the constant factor change in odds 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
```

- So can interpret the effects using marginal change, discrete change of various amounts, and/or MER, AME, and MEM settings of other independent variables

- 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 slide 20 from the Logit-Probit lecture)!
- 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 coefficient: 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 means different things depending on which Xs are included

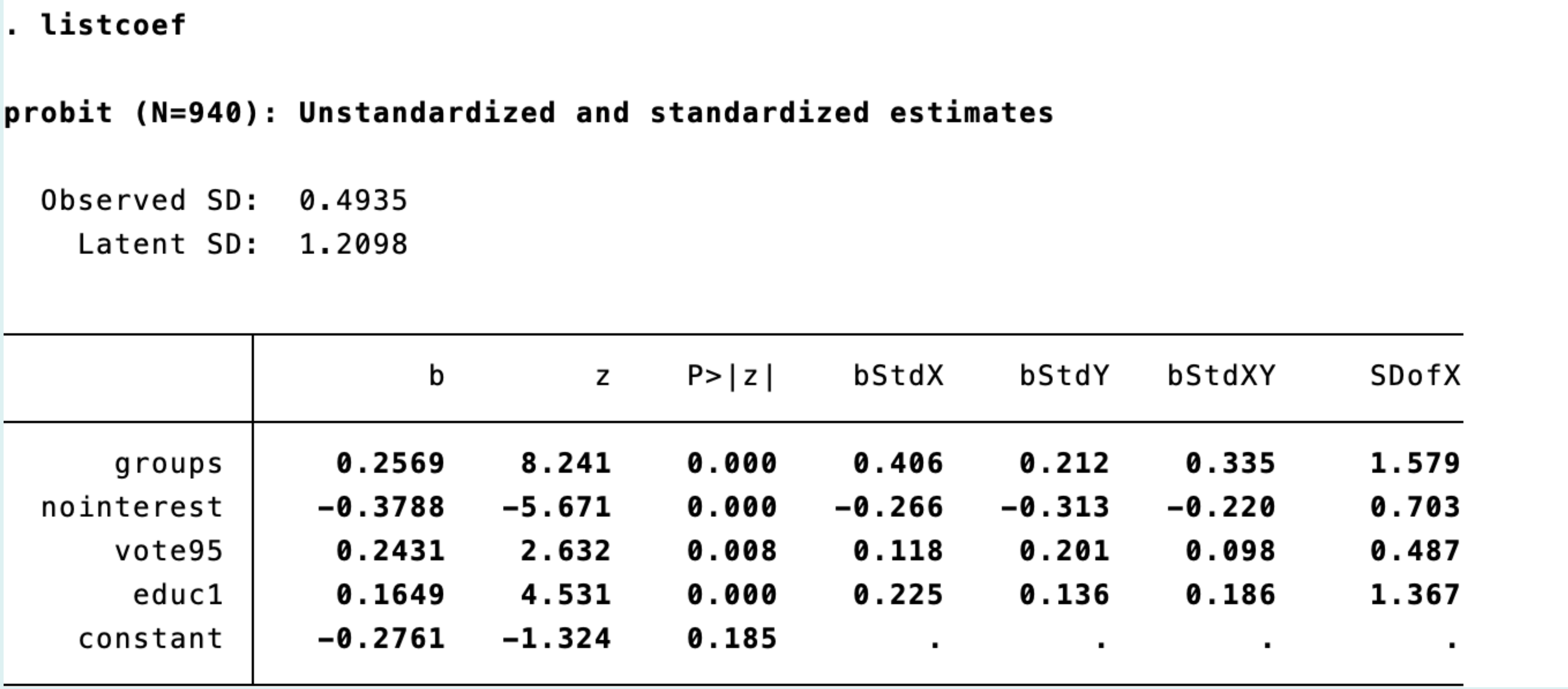
$$Y* = X\mathrm{B} + e$$

$$Var(Y*) = b^2 Var(X) + Var(e)$$

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

- This shows that adding more Xs changes the variance of Y*, since the variance of epsilon is fixed at 1 by assumption
- To facilitate interpretations of the estimated probit coefficients, 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*.
- If you use standard deviation changes in X as the independent variable, then the estimated coefficient will be exactly the same kind of standardized beta coefficient you have in regular OLS regression!
- This would represent the effect of one standard deviation (z-score) change in X on standardized (z-score) Y*

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

- The indeterminacy of the variance of Y* also means 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. [It is like changing the scale of Y in one model from dollars to "dollars and a half" – that would make the β smaller even if the effect was exactly the same!]
- 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" do file

- Another R-squared for probit models is also suggested by the latent variable Y* approach: how much variance in Y* is explained by the independent variables in the model? This is the “McKelvey and Zavoina R-squared”
- In regular regression one calculates R^2 as:

$$\frac{\text{Explained Variance}}{\text{Total Variance}} = \frac{\text{Explained Variance}}{\text{ExplainedVariance} + \text{Error Variance}}$$

$$\frac{\beta^2 (VarX)}{\beta^2 (VarX) + Var(\varepsilon)} \quad \text{or} \quad \frac{Var(\hat{Y})}{Var(Y)}$$

- In probit, can get an analogue by estimating Var(Y*) as β^2(VarX)+Var(ε)= β^2(VarX)+1 and Var($\hat{Y}$) as β^2(VarX) or Var(Y*-1)

$$\frac{\beta^2 Var(X)}{\beta^2 Var(X) + 1} = \frac{Var(Y^*) - 1}{Var(Y^*)}$$

In our ML probit example, the estimated coefficient was 3.48, and we can estimate the standard deviation of X -- .43.

=((3.48^2)*(.43^2))/((3.48^2)*(.43^2)+1)=2.24/3.34=.69