MLE: Categorical and Limited Dependent Variables

Unit 1: Models for Dichotomous Dependent Variables
2. Maximum Likelihood Estimation

PS2730-2021

Week 2

Professor Steven Finkel

How Do We Estimate Logit/Probit Model Parameters?

- OLS cannot be used to estimate parameters in binary DV (and other non-continuous DV) models
- Why?

  1: If think of the dependent variable as Y*, it is unobserved

  2. If think of the dependent variable as P(Y=1), it is also unobserved, and using 0,1 in place as in logit model gives ln(1/0) or ln(0/1), each of which is undefined
- So we turn to another estimation procedure:

  Maximum Likelihood
- ML can also be used for continuous DV regression and many other models. And we will see below that if the OLS assumptions are satisfied, OLS=ML in the continuous DV case

Intuition of ML Estimation

- Find the β parameters (or other parameters you might be interested in) that give the highest likelihood of observing the data that were observed. Given a sample of observations, we search for the population parameters that *maximize the joint probability of having observed that sample*
- For logit/probit, that means finding β that maximize the P(Y=1|X) when there actually was an observed "1", and maximizing the P(Y=0|X) when there was actually an observed "0"
- In probit, for example, since P(Y=1|X)=Φ(XB), then we should attempt to find **B** that generate z-scores (XB) corresponding to high P(Y=1) for all of the 1s, and gives z-scores corresponding to low P (Y=1) for all of the 0s.
- **The parameter estimates that give the highest joint set of Ps according to this criteria are the ML estimates**

- ML estimation is not limited to estimating logit or probit parameters: it is a general principle extending to estimating any population parameter from (randomly selected) sample data.
- General steps in ML estimation:
  1. Assume a probability distribution for Y – e.g., normal, binomial (Bernoulli), poisson, etc.)
  2. Express the joint probability of the data (i.e., all of the Y) using the assumed probability distribution
  3. Calculate the joint probability of the data given the parameters–the "likelihood function" (taking the log of the likelihood to simplify)
  4. Maximize this function with respect to the unknown parameters (e.g., the **Bs** in a regression/logit/probit function)
- **This yields the parameter estimates that produced the observed data with the highest overall likelihood**

Likelihood Inference

- Assume we have a fair coin ($\boldsymbol{\pi}$=.5). We can then calculate the probability of getting 0 heads, 1 heads, 2 heads when we flip the coin 2 times (.25, .5, .25)
- Assume a population mean of 10. We can then calculate via the central limit theorem the probability of obtaining samples of given size N with means of 6, 7, 13, 20, etc.
- These examples assume a **known population parameter**, and we can then calculate the **exact probability** of obtaining samples with different characteristics
- But what we often (mostly) want to do is *estimate* model parameters (call the set of them $\boldsymbol{\theta}$), given the sample data that we've obtained
- Instead of obtaining the **P(Data|Model)** – or the probability of obtaining different samples with a known population parameter, what we are after is the **P(Model($\boldsymbol{\theta}$)|Data)** – that is, the sample *data* are fixed and known, but the model *parameters* are random and unknown

- It can be shown (following Bayes' Rule), that the quantity we are after – **p(Model($\boldsymbol{\theta}$)|Data)** – is *proportional* to **p(Data|Model ($\boldsymbol{\theta}$)).**

- We can't obtain the exact probability of a set of model $\boldsymbol{\theta}$ parameters being "true", but we can talk about the relative *likelihood* that one set of model parameters to another set of parameters

- More formally: $\mathcal{L}(\boldsymbol{\theta}|y) \propto p(y|\boldsymbol{\theta})$ (where $\propto$="is proportional to")

- **The likelihood of the parameters, given the data, is proportional to the probability of the data, given the parameters**

- So the parameters that maximize the joint probability of the data are the same ones that have the highest likelihood, given the data!

- Example: What is the Maximum Likelihood estimate of the population mean (μ) of a normally-distributed population, given the following sample of data values: 2, 4, 6, 8?
- Steps
  1. Assume a probability distribution for Y: here we assume the y are distributed normally with mean μ and standard deviation σ
  2. Express the joint probability of the data (i.e., all of the Y) using the assumed probability distribution

$$P(y_1 \mid \mu,\sigma^2) = f_N(y_1 \mid \mu,\sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{y_1-\mu}{\sigma}\right)^2}$$

$$P(y_1,y_2 \mid \mu,\sigma^2) = f_N(y_1 \mid \mu,\sigma^2) x\, f_N(y_1 \mid \mu,\sigma^2)$$

$$= \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{y_1-\mu}{\sigma}\right)^2} x \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{y_2-\mu}{\sigma}\right)^2}$$

$$P(Y \mid \mu, \sigma^2) = \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} \mathrm{e}^{-\frac{1}{2}\left(\frac{y_1-\mu}{\sigma}\right)^2}$$

- Assuming independent and identically distributed observations
- So joint overall probability is the product of the individual probabilities, denoted by the height of the normal curve for each y_i
- We see that the joint probability is the product of the heights of the normal pdf associated with the individual z-scores
- We can simplify further by assuming a σ of 1 without affecting the *relative* calculations of each p

3. Calculate the joint probability of the data given the parameters–the "likelihood function":

$$\mathcal{L}(\mu \mid Y) = (\propto) = P(Y \mid \mu) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}} \mathrm{e}^{-\frac{1}{2}(y_1-\mu)^2}$$

This yields an *extremely* small value, so it is easier to work with **log of the likelihood** – this maintains the relative likelihoods but makes the numbers more tractable

$$\ln \mathcal{L}(\mu \mid Y) \propto \ln P(Y \mid \mu) = \ln\left(\prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}} \mathrm{e}^{-\frac{1}{2}(y_1-\mu)^2}\right)$$

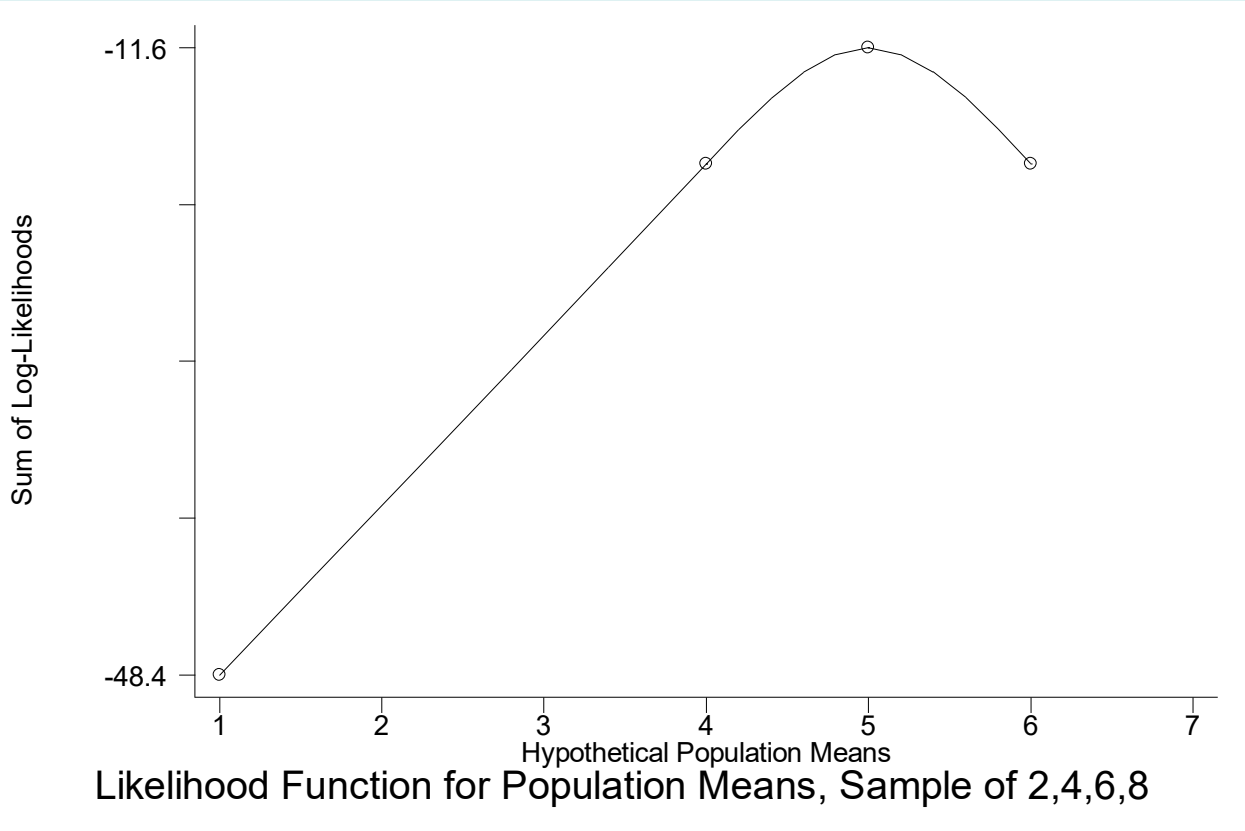
4. Last step: Maximize this function wrt the unknown parameters. That will give the "maximum likelihood estimates"!

- **There is a joint probability of having observed our data for all (infinite) hypothetical values of μ; these are equivalent (proportionately) to the likelihood of μ, given the data**
- E.g., if μ is 4, we can calculate the probability of observing a 2, a 4 a 6 and an 8; the product of these probabilities (or the addition of the log-probabilities) gives us the value of the likelihood function for μ=4.
- If μ is 6, there is another specific probability of observing a 2, a 4, a 6 and an 8; adding the log-probabilities gives us the value of the likelihood function for μ=6**;** if μ is -7, there is another value of the likelihood function, etc.
- So, for a given μ, we calculate ($P(y_i)$) for each *i* and take the product over the 4 cases in the sample, or $\ln(P(y_i))$ and **add** them together
- We then take the value of μ that gives the highest joint probability of observing a sample of {2, 4, 6, 8} from the (standard) normal distribution; this is the **maximum likelihood** estimate of μ

Now we can plot the likelihood function against different possible values of μ:

Stata gives you the pdf with "normalden(z-score)"

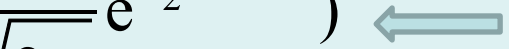
		μ =1 Z	pdf	μ =4 Z	pdf	μ =5 Z	pdf	μ =6 Z	pdf
Case 1:	2	1	.24	2	.05	3	.004	4	.00001
Case 2:	4	3	.004	0	.40	1	.24	2	.05
Case 3:	6	5	.000001	2	.05	1	.24	0	.40
Case 4:	8	7	.0000009	4	.00001	3	.004	2	.05
Sum of lns of each p			-48.5		-18.4		-11.6		-18.4



Can see that the μ that generated the highest joint likelihood for this sample is 5. This corresponds to the sample mean! So, given **Y** and a normally distributed population, the ML estimate of μ is $\overline{X}$

- How do you decide if there is a "maximum" without doing this kind of search for the millions of possible parameters?
- Take first derivative of the likelihood function and set it equal to 0 – that gives you the place where the slope of the tangent to the curve is 0, which is the maximum point on the curve

$$\ln \mathcal{L}(\mu \mid Y) = (\propto) = \ln\left(\prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(y_1-\mu)^2}\right)$$



For our purposes, we can ignore everything in this expression that does not depend on y_i

$$\ln \mathcal{L}(\mu \mid Y) = -\Sigma\left(y_1 - \mu\right)^2$$

$$\frac{\partial \ln \mathcal{L}(\mu \mid Y)}{\partial \mu} = 2\Sigma y_1 - 2N\mu$$

$$0 = \Sigma y_1 - N\mu$$

$$\mu = \frac{\Sigma y_1}{N} = \bar{Y}$$

ML Estimation of Linear Regression Parameters

- Next step: Don't assume a uniform or constant mean, but rather a mean that is conditioned on the Xs through a regression line Y=XB in linear fashion, as in $E(Y|X) = \mu = XB$
- Steps:

1. Assume a probability distribution for Y – e.g., normal in this case
2. Express the joint probability of the data (i.e., all of the Y) using the assumed probability distribution
3. Calculate the joint probability of the data given the parameters– the "likelihood function" (taking the log of the likelihood to simplify)
4. Maximize this function with respect to the unknown parameters (e.g., the **Bs** in a regression function)

$$\ln \mathcal{L}(\mu \mid Y) = (\propto) = \ln\left(\prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(y_1-\mu)^2}\right)$$

$$\ln \mathcal{L}(\mu \mid Y) = (\propto) = \ln\left(\prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(y_1-XB)^2}\right)$$

$$\ln \mathcal{L}(B \mid Y) = (\propto) = -\sum\left(y_1 - XB\right)^2$$

$$\ln \mathcal{L}(\beta_1 \mid Y) = (\propto) = -\sum\left(y_1 - \alpha - \beta_1 x\right)^2$$

What does this mean?

- Maximizing the (log)-likelihood is the same thing as minimizing the sum of squared errors in a linear regression equation!!
- So MLE = Least squares, assuming normally distributed Y (or ε)!
- Taking the first derivative wrt α and β and setting to 0 leads exactly to the two "normal equations" from OLS derivation and the same solutions

- More informally: the z score for each y_i gives the height of the normal pdf for that observation; z scores closer to 0 (less error) give larger pdfs;
- Data points that are far from (near) the hypothetical regression line get a very small (large) pdf corresponding to a very small (large) likelihood in a normal distribution
- The regression line (α and β) that is "closest" to all the points –which generates the overall smallest z-scores and largest summed log-probabilities -- will then be the **maximum likelihood** estimates of the intercept and slope parameters in the regression model.

```
. glm locpart groups, family(gaussian) link(identity)

Iteration 0:   log likelihood = -1480.3078

Generalized linear models                         Number of obs   =        940
Optimization     : ML                             Residual df     =        938
                                                  Scale parameter =   1.368676
Deviance         =  1283.818241                   (1/df) Deviance =   1.368676
Pearson          =  1283.818241                   (1/df) Pearson  =   1.368676

Variance function: V(u) = 1                       [Gaussian]
Link function    : g(u) = u                       [Identity]

                                                  AIC             =   3.153846
Log likelihood   = -1480.307821                   BIC             =  -5137.617

------------------------------------------------------------------------------
             |                 OIM
    locpart |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
------------+-----------------------------------------------------------------
     groups |   .3602997   .0241798    14.90   0.000     .3129083    .4076912
      _cons |   .3069086   .0715942     4.29   0.000     .1665865    .4472308
------------------------------------------------------------------------------
```

MLE bivariate linear regression

```
. regress locpart groups

      Source |       SS           df       MS      Number of obs   =       940
-------------+----------------------------------   F(1, 938)       =    222.04
       Model |  303.895589         1  303.895589   Prob > F        =    0.0000
    Residual |  1283.81824       938  1.36867616   R-squared       =    0.1914
-------------+----------------------------------   Adj R-squared   =    0.1905
       Total |  1587.71383       939  1.69085605   Root MSE        =    1.1699

------------------------------------------------------------------------------
    locpart |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
------------+-----------------------------------------------------------------
     groups |   .3602997   .0241798    14.90   0.000      .312847    .4077525
      _cons |   .3069086   .0715942     4.29   0.000     .1664052    .4474121
------------------------------------------------------------------------------
```

OLS bivariate linear regression

Properties of Maximum Likelihood Estimates

- **Consistency**—They are asymptotically consistent. As sample size increases, the estimates increasingly approach the actual population parameters. As a result, MLEs are good large sample estimators (N greater than 100, depending on number of parameters)

$$MLE(\hat{\theta}) \rightarrow \theta \text{ as } \mathrm{N} \rightarrow \infty$$

- **Asymptotic normalcy**—The MLE parameters are distributed according to the standard multivariate normal no mater what distribution assumptions you make in your model. This allows us to describe them using z-scores, construct confidence intervals, etc.
- **Asymptotic efficiency**— MLE has the smallest asymptotic variance of any estimators that are also consistent and asymptotically normal.

ML Estimation of Logit/Probit Parameters

- Principle: Given the probability distribution of Y_i and the function for P (Y=1) in either the logit or the probit model, find the Bs that maximize the overall probability of having observed the sample of 1s and 0s that were observed, given the values of the X_i.
- Steps:
  1. Assume a probability distribution for Y – e.g., binomial (Bernoulli) in this case
  2. Express the joint probability of the data (i.e., all of the Y) using the assumed probability distribution
  3. Calculate the joint probability of the data given the parameters–the "likelihood function" (taking the log of the likelihood to simplify)
  4. Maximize this function with respect to the unknown parameters (e.g., the **Bs** in the logit or probit function)

Step 1: Assume Y_i is a *binomial (Bernoulli)* distributed variable of 1s and 0s, with

$$P(y_i \mid \pi) = \pi^{y_i} * (1-\pi)^{1-y_i}$$

and with

$$\pi = \Phi(X\mathrm{B}) \qquad \text{Probit}$$

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

Step 2: Express the joint probability of the data

$$P(Y \mid \pi) = \prod_{i=1}^{n} (\pi^{y_i} * (1-\pi)^{1-y_i})$$

Step 3: Calculate (and simplify) the log-likelihood function

$$\ln \mathcal{L}(\mu \mid Y) = (\propto) = \ln(Y \mid \pi) = \ln(\prod_{i=1}^{n} \pi^{y_i} * (1-\pi)^{1-y_i})$$

$$\ln \mathcal{L}(\mu \mid Y) = \sum_{i=1}^{n} y_i \ln \pi + \sum_{i=1}^{n} (1-y_i) \ln(1-\pi)$$

$$\ln \mathcal{L}(B \mid Y) = \sum_{i=1}^{n} y_i \ln(\frac{e^{XB}}{1+e^{XB}}) + \sum_{i=1}^{n} (1-y_i) \ln(\frac{1}{1+e^{XB}}) \quad \text{Logit}$$

$$\ln \mathcal{L}(B \mid Y) = \sum_{i=1}^{n} y_i \ln(\Phi XB) + \sum_{i=1}^{n} (1-y_i) \ln(1-\Phi XB) \quad \text{Probit}$$

- Step 4: Maximize with respect to unknown parameters

Practically: add the logs of the predicted P(Y=1) for the 1s to the logs of the predicted (1-P(Y=1)) for the 0s. That is the sum of the log-likelihoods, and find the maximum value of the B which does this

Formally: set the derivative of the log-likelihood function to 0, and solve algebraically (if possible), or numerically (iteratively) if not

$$\frac{\partial \ln \mathcal{L}(B \mid Y)}{\partial B} = \sum X(y_i - \frac{e^{XB}}{1+e^{XB}}) \qquad \text{Logit}$$

$$\frac{\partial \ln \mathcal{L}(B \mid Y)}{\partial B} = \sum X(y_i - \Phi XB) \qquad \text{Probit}$$

There is no closed-form algebraic solution, but the function is still "well-behaved" with a single peak, so can be estimated iteratively

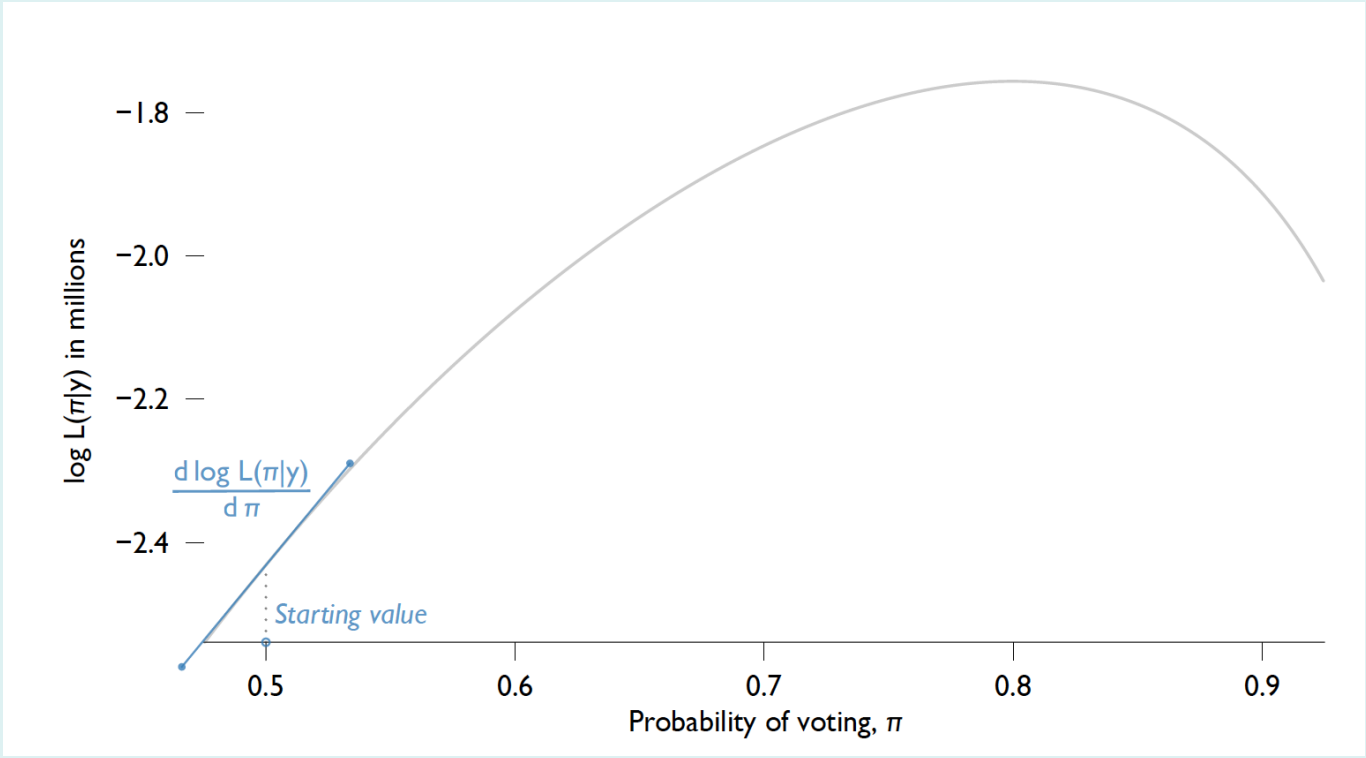
Analytic versus Numerical ML Estimation

- Earlier examples were cases where there was an analytic or clean algebraic solution, i.e., where setting the first derivative of the log-likelihood function yielded a nice arithmetic result
- Many instances the likelihood function is sufficiently complex that a clean result does not obtain; in those cases maximization is obtained through numerical methods
- Search and test many possible solutions iteratively
- Iterative search

    1. Start with an initial guess
    2. Use the current guess to seek a new best guess
    3. Repeat step 2 until "convergence": e.g., the local derivative of $\mathcal{L}(\boldsymbol{\theta}|\mathrm{Y})$ is *approximately* 0

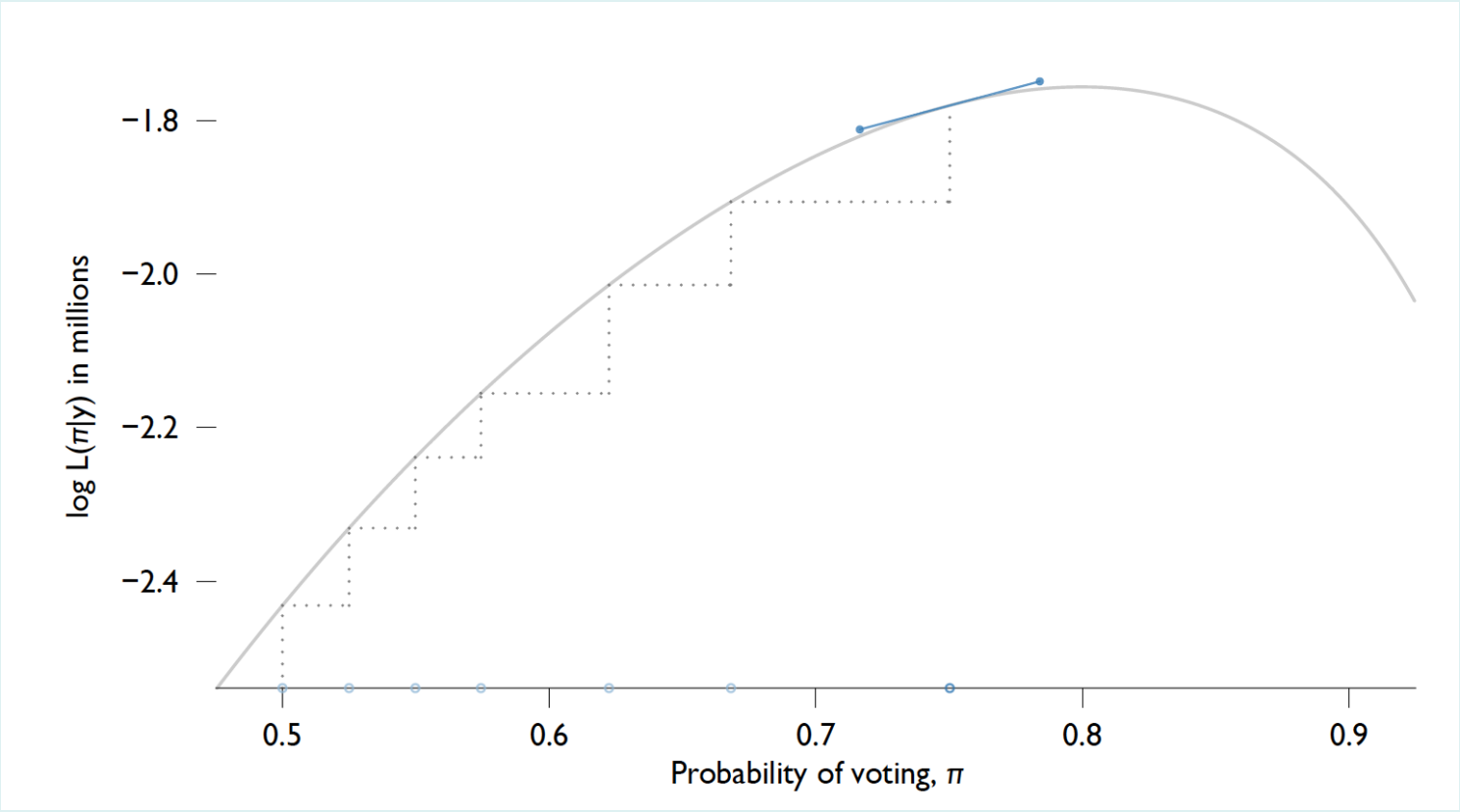
- Many search algorithms are available; default is "Newton-Raphson" – in Stata: options "technique (nr)" or others

Example of Numerical MLE Optimization

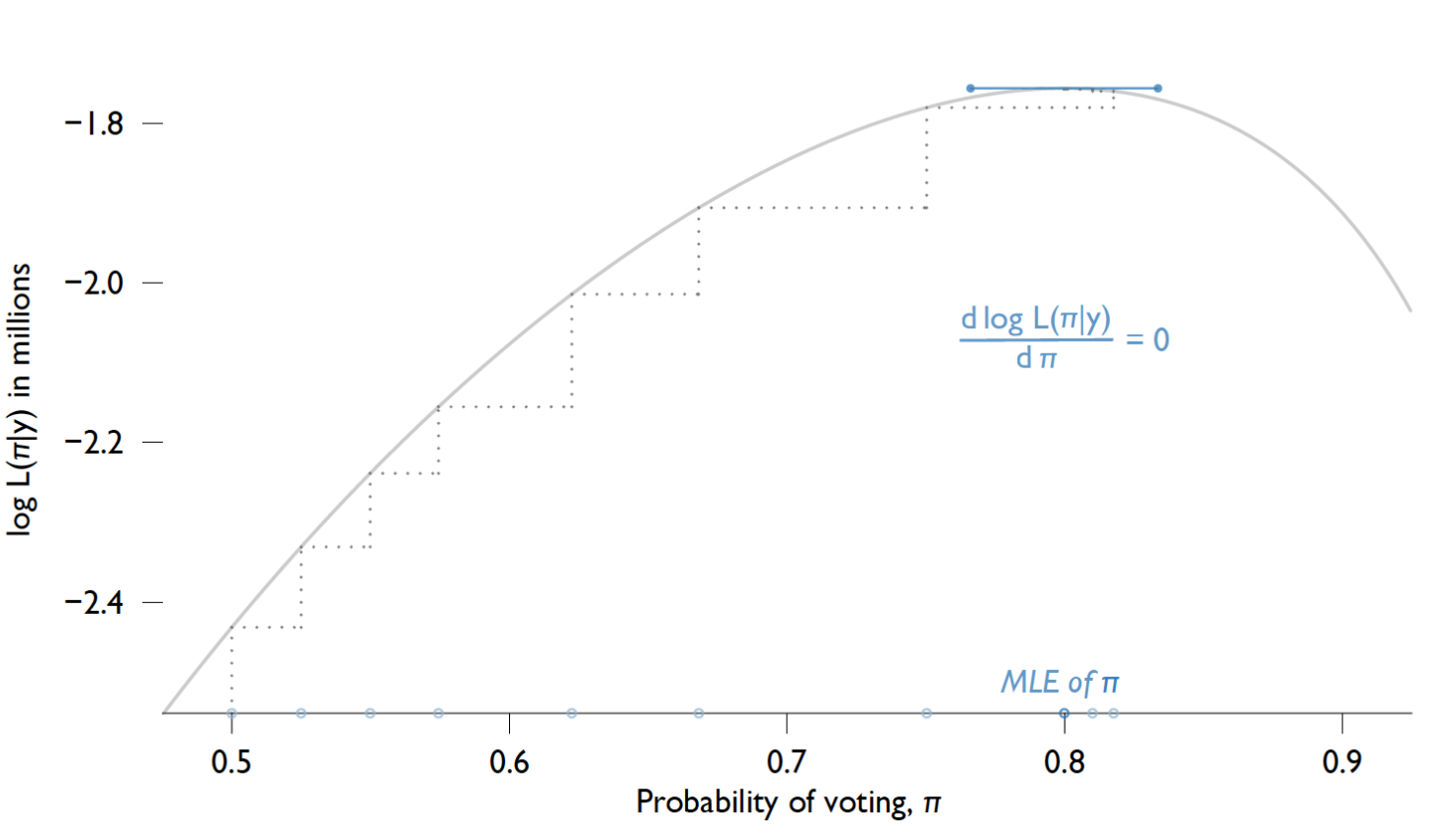


Climbing a log-likelihood hill!

Start the search with an initial "guess"; calculate first derivative of the curve at the starting value -- if positive we know we need to go to the right on the next guess; if negative we need to go the left



Each guess gets you closer to the top of the hill; if you "overstep and go too far (with a negative first derivative), you go backwards



Until, with successive zooming in and iterations, you reach the top where the first derivative equals, or is as close as possible to zero.

In some instances there will be local maxima that are not the true maximum, e.g. in multi-peaked log-likelihood functions. Be aware at least of this in more complex models

Probit Example (PS2730.maxlike.probit.xy.dta)

- We have 9 cases in a sample, 5 "1" and 4 "0".

	y	x
1	1	.3
2	0	-.3
3	1	.8
4	0	.2
5	1	.5
6	1	-.1
7	0	-.5
8	0	0
9	1	.6

Begin by assuming no slope effect whatsoever of X (so $\beta_1=0$). We get the default likelihood function w/out X and then can see whether knowledge of X improves things

$$\ln \mathcal{L}(\pi \mid Y) = \sum y_i \ln \frac{y_i}{N} + \sum (1 - y_i) \ln (1 - \frac{y_i}{N})$$

$$\ln \mathcal{L}(\pi \mid Y) = 5(\ln .55) + 4(\ln .45)$$

$$\ln \mathcal{L}(\pi \mid Y) = -6.183$$

We have a predicted P(Y=1) for all cases equal to 5/9, or .555515.
The z-score associated with P(Y=1) of .555515 is .1397
Stata: dis invnorm (5/9)
So in the "default", or "reduced model without X", everyone has a z-score of .1397, a P(Y=1) of .555515, and we take the sum of the log-likelihoods as ln(.55515) for all the 1s, and ln(1-.555515) for the 0s
This yields a sum of the log-likelihoods of -6.1826
That is the value of the likelihood function when $\beta_1=0$

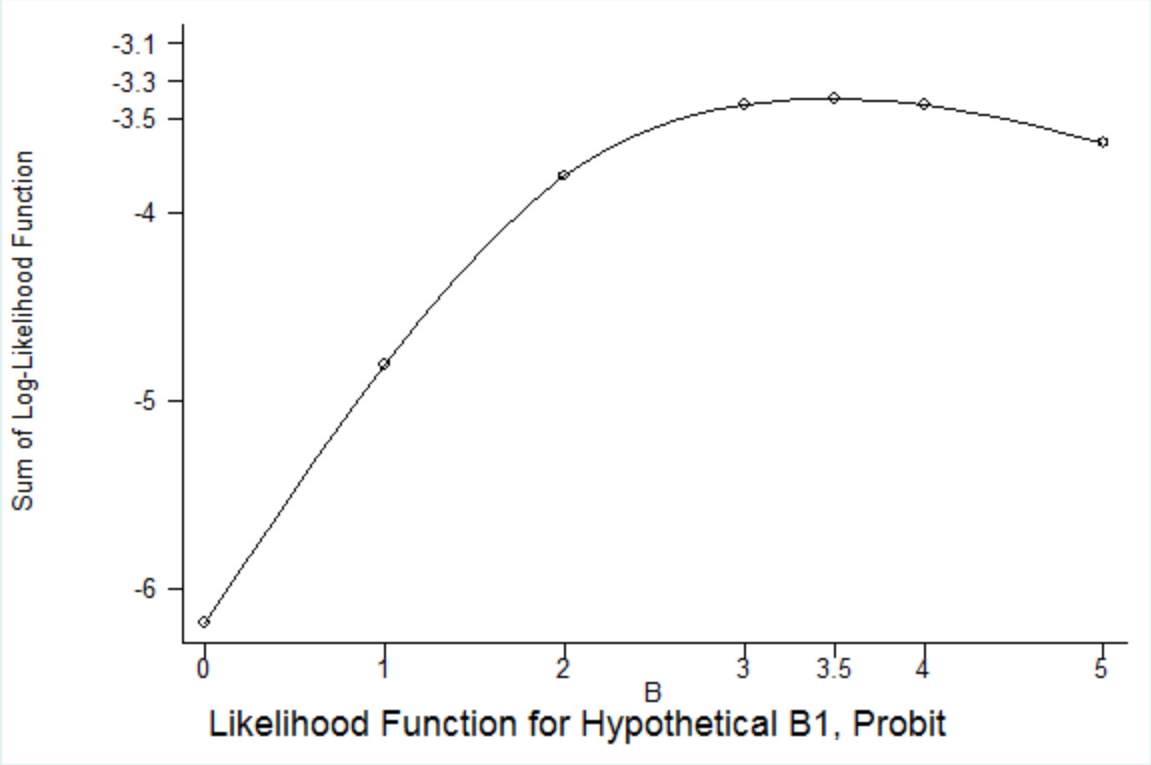
```
. probit y

Iteration 0:   log likelihood = -6.1826542
Iteration 1:   log likelihood = -6.1826542

Probit regression                                 Number of obs     =          9
                                                  LR chi2(0)        =       0.00
                                                  Prob > chi2       =          .
Log likelihood = -6.1826542                       Pseudo R2         =     0.0000

------------------------------------------------------------------------------
           y |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       _cons |   .1397103   .4192564     0.33   0.739    -.6820171    .9614377
------------------------------------------------------------------------------
```

- Then can try different values of βs, given the X, generate new z-scores, new sums of the log-likelihoods, and maximize
- See "ps2730-2021.maximum likelihood.probit.do"



- ML estimate for the slope looks to be just about 3.5

```
. probit y x

Iteration 0:   log likelihood = -6.1826542
Iteration 1:   log likelihood =  -3.433008
Iteration 2:   log likelihood = -3.3914407
Iteration 3:   log likelihood = -3.3913913
Iteration 4:   log likelihood = -3.3913913

Probit regression                                 Number of obs   =          9
                                                  LR chi2(1)      =       5.58
                                                  Prob > chi2     =     0.0181
Log likelihood = -3.3913913                       Pseudo R2       =     0.4515

------------------------------------------------------------------------------
           y |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           x |   3.482994   2.029595     1.72   0.086    -.4949396    7.460928
       _cons |  -.3043448   .5846979    -0.52   0.603    -1.450332     .841642
------------------------------------------------------------------------------
```

Iterations stopped at sum of the log-likelihoods =-3.39139

ML slope=3.483

MLE: Statistical Tests

- Is entire equation "significant"? We can arrive at the probit/logit equivalent of the least squares F test by comparing the log-likelihoods of the "full" (or "unconstrained") model" that includes X to a "reduced" (or "constrained" model that does not include X

$$Y_i^* = \beta_0 + \beta_1 X_i + \varepsilon_i \qquad \text{Full Model}$$

$$Y_i^* = \beta_0 + \varepsilon_i \qquad \text{Reduced Model}$$

- Each of the models has an associated log-likelihood
  - -3.3914 for the full; -6.1826 for the reduced
- Does inclusion of X significantly improve the log-likelihood? Calculate the "**LR Test**", also **called Likelihood Ratio Statistic**, or **G^2** in Long, or "**Model Chi-square**" in Stata, since it follows a χ^2 distribution
- LR Statistic G^2 = 2 lnL (Full Model) – 2 LnL(Reduced Model)
- Here LR Statistic=5.59, with 1 df (associated with 1 indep. variable)

- The test is based on **the likelihood ratio** principle, which expresses how many times more likely the data are under one model than the other (and, mathematically, the ratio of two logs is their difference)
- This likelihood ratio can then be used to compute a p-value, or compared to a critical value to decide whether to reject the null model in favor of the alternative model.

- Null hypothesis: All additional slopes=0; or, the full model does not significantly improve the log-likelihood over the reduced model
- Interpretation: The probability of getting a chi-square of the given magnitude, IF null hypothesis were true is .018, so we reject the null. Relaxing the constraint that $\beta_1=0$ improves the fit of the model

- Can also see this logic in terms of the "deviance" of the models from a *perfect or "saturated" model* where the predicted P for all 1s would be 1, and the predicted P for all 0s would be 0. [It is saturated because, in effect, we would have a dummy variable for each case to generate perfect predictions].
- This would generate a sum of the log-likelihoods of 0!!!! (ln(1)=0+ln(1)+ln(1)=0, etc.)
- A "Deviance" is calculated as -2* Model Log-Likelihood
- Smaller numbers for the Deviance are better (i.e., closer to the saturated (perfect) model's value of 0)
- 2*lnL(SATURATED) -(2*lnL "Our" Model) =

  2*ln(1)-2LnL (our Model) = 0-(2*-3.39)=**6.78**
- **So our full model has a deviance of 6.78**

- G^2, or the Likelihood Ratio chi-square, is the statistical test of whether a Full model represents an improvement in fit, is based on the difference of two Deviances, or the difference of the Intercept-only model Deviance and the Full model Deviance
- Deviance (Full Model) = -2*-3.39 = 6.78
- Deviance (Reduced Model) = -2*-6.18= 12.39
- **So the difference of the two deviances, or LR G^2 , is 5.58**

- We can apply this logic to any combination of additional independent variables from a (nested) "reduced" to a "full" model
- Stata: save estimates of M1 and M2 with "est store", and then "lrtest M2 M1" with the unconstrained (full) model as M2, then the constrained (reduced) M1
- Or use the Long-Freese SPOST routine "fitstat"

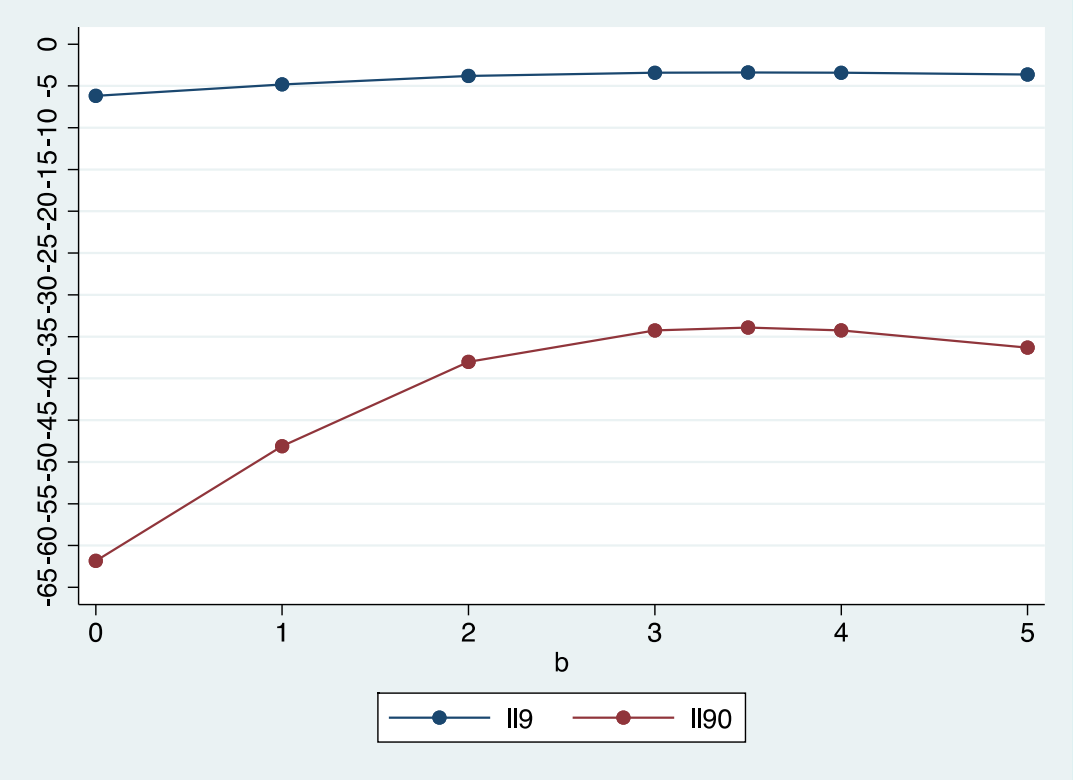
- The z test is used for the significance of individual coefficients

$$z_{\beta_1} = \frac{\beta_1}{\sigma_{\beta_1}}$$

where $\sigma_{\beta 1}$ is the inverse of the negative of the second derivative of the likelihood function with respect to β_1.

- What is the "second derivative"? – It is the rate of change of the rate of change (similar to ideas of "velocity" and "acceleration"). This quantity must be negative in order for it to be **maximum** likelihood estimation. (WHY?) It means that slope of the first derivative or the tangent of the likelihood function is getting smaller and smaller and will eventually level off (at the maximum) and then turn down
- The matrix of second derivatives for all β is called the "Hessian" matrix

- The lnL can be really flat (small Hessian) or really sharp (large Hessian). Which is better in terms of precision of estimates? If flat, not sure that curve at any given point is really the maximum or near the maximum since changing so slowly, so less precision
- Therefore: taking the inverse of the Hessian (what is called the "Information Matrix") gives the magnitude of precision and hence the standard errors for individual coefficients
- We take the negative of the inverse Hessian to ensure that the standard errors are positive
- Then use these standard errors in normal hypothesis testing
- These quantities, as would be expected, depend on the intrinsic nature of the likelihood function, given the data, as well as N, the number of cases in the sample. As N increases, the curvature of the likelihood function steepens as well



Can see how the curvature (2nd derivative) of our example on top log-likelihood is really flat when N=9; much more pronounced curvature when N=90. (Note that the ML estimate itself is the same in both cases). This means we are less certain about the maximum likelihood estimate in the N=9 than N=90 condition, and this uncertainty is reflected in the respective standard errors (2.03 versus .64). You can test this by running the same probit regression using "PS2730.maxlike.probit.n90.xy.dta".

- **Wald test** provides a more general test of whether coefficient(s) in an estimated model are statistically different from those in a "constrained" model
- This differs conceptually from the LR test which tests whether the additional parameters in a full model improve the log-likelihood (reduce the deviance) compared to the reduced model
- But you should arrive at the same conclusions either way (asymptotically)!
- Look at top graph of the log-likelihood and examine the difference between the estimate of 3.48 and 0, given the curvature of the function. Are we sure that 3.48 is greater than 0?
- Do the same for the bottom graph. Much more confidence!
- Calculation (for comparison to constrained model with $\beta=0$:

$$\frac{(\beta_1 - \beta_c)^2}{\hat{\sigma}^2} = \frac{\beta_1^2}{\hat{\sigma}^2}$$

Stata: "test VARNAME(S)"

Goodness of Fit Statistics

- What are some R-Squared analogues in ML models?
- Several measures exist based on comparisons of likelihood ratios of the "constrained" (reduced) and "unconstrained" (full) models. Basic idea: How much did we improve the LnLikelihood, compared to how much we *could* have improved it? A "perfect" model would go all the way to 1 – we improved 100% of what we could have improved, or we achieved complete perfection in the unconstrained model's Log-Likelihood. This happens as LnLikelihood, Full →0 !!!!
- "McFadden" R-squared or "Pseudo R-squared":

  1-(LnL (Unconstrained)/LnL(Constrained)) = 1 – (-3.39/-6.18) = 1-.55=.45

  We improved the log-likelihood by 45% through including X
- Alternative Calculation: G^2/(-2*Ln(Constrained))=5.58/(-2*-6.18) = .45
- Model Chi-Square divided by -2 * LnL of Constrained Model or Model Chi-Square divided by Deviance of Constrained Model

- "Adjusted" McFadden: 1-((LnL(Unconstrained)-K)/lnL(Constrained)) =

  1-((-3.39-2)/-6.18)=.128

  where k is number of parameters to be estimated
- Same logic as Adjusted R-squared – you can't decrease McFadden by adding new variables, so there should be a penalty for too many IVs
- **Adjusted McFadden will only increase if the LnL of the unconstrained model increases by more than 1 for each parameter added to the model**

- Another way to look at R-squared is the "Explained Variance in **Y***", the latent variable in the probit formulation.
- In OLS 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

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

=((3.4829^2)*(.4301^2))/((3.4829^2)*(.4301^2)+1)

= 2.24/3.24=.69

- This is also called the "McKelvey and Zavoina R-squared"

- Other measures of fit are based on the idea of "correct predictions" of Y. Do we predict Y to be 1 when it is 1 and Y to be 0 when it is 0? These predictions are based on whether the probit or logit predicted probabilities are greater or less than .5
- Stata "lstat" shows .78 predicted correctly. Seems good. BUT:

```
. lstat

Probit model for y

                 True
Classified |         D          ~D |     Total
-----------+----------------------+----------
     +     |         4           1 |         5
     -     |         1           3 |         4
-----------+----------------------+----------
Correctly classified               |    77.78%
-------------------------------------------------
```

- 5 cases are on 1, so we would predict .55 correctly by chance, or simply by predicting "1" for everybody. Need to compare .78 to this.
- So can calculate (7-5)/(9-5) = 2/4 = .50 as the **"Adjusted" Count R2**
- (Total Number of Correct – Correct Predictions from Marginals)/(Total Number of Cases – Correct Predictions from Marginals). This is Adjusted Count R^2. **VERY IMPORTANT!!!!**

- Alternative measure proposed by Danish statistician Tue Tjur (2009). It is based on the simple logic that a good model will produce high predicted probabilities for the cases where P(Y=1), and low predicted probabilities for the cases where P(Y=0).
- So take the difference between the average predicted probability for cases where P(Y=1) and cases where P(Y=0)
- Tjur's R-squared or Tjur's D", the **"Coefficient of Discrimination"**:

  $D = \bar{\pi}_1 - \bar{\pi}_0$

  where π_Y=P(Y=1|XB) for Y=1,0
- Simple! It is intuitive, and Tjur (2009) shows that it has many attractive properties. One of them is being (asympotically) very close to the squared correlation between actual Y (1 or 0) and the predicted probability that Y=1, another common measure of R-squared

- Final kind of summary statistic: **"entropy-based measures"** which can be used to compare models that may or may not be nested
- Idea is that log-likelihoods of models, relative to their degrees of freedom, provide general indication of "fit"; we can compare some summary quantity (like a modified "deviance") from one model to another and decide which to prefer
- **Akaike Information Criterion (AIC)**

$$AIC = \frac{-2 \ln L(M) + 2(k+1)}{N}$$

- First term in the numerator is the Deviance of the model, second term is the penalty for the number of parameters
- We want *smaller* values for AIC; that indicates less deviance and better fit

- **Bayesian Information Criterion (BIC)** compares two models in terms of their relative probability or likelihood, given the data. We prefer M2 over M1 if the ratio of P(M2|Data) is greater than P (M1|Data)
- For comparing M2 to a saturated M1 model (with 0 Deviance):

$$BIC = df(M2) * \ln(N) - 2lnL(M2)$$

- This value can then be calculated for any other model (M3) and compared to M2: BIC_{m2} - BIC_{m3} with smaller (more negative) values preferred
- Our example: BIC(Unconstrained) = 11.18 BIC (Constrained) = 14.56
- Rule of thumb (Long, p. 112): absolute differences between models should be greater than 5 to provide "strong" evidence in favor of one or the other. So we are not sure our model is "better" according to BIC, or at least not "strongly" better

Summary Model Goodness of Fit Measures with "Fitstat"

```
. fitstat
```

	probit
Log-likelihood	
Model	**-3.391**
Intercept-only	**-6.183**
Chi-square	
Deviance(df=7)	**6.783**
LR(df=1)	**5.583**
p-value	**0.018**
R2	
McFadden	**0.451**
McFadden(adjusted)	**0.128**
McKelvey & Zavoina	**0.692**
Cox-Snell/ML	**0.462**
Cragg-Uhler/Nagelkerke	**0.619**
Efron	**0.464**
Tjur's D	**0.496**
Count	**0.778**
Count(adjusted)	**0.500**
IC	
AIC	**10.783**
AIC divided by N	**1.198**
BIC(df=2)	**11.177**
Variance of	
e	**1.000**
y-star	**3.244**

Summary: R-squared in Models with Discrete Outcomes

- Reduced Error Variation (the analog of (1-SSE)/SST)=
  $1-(LnL_{full}-LnL_{reduced})$
  - McFadden's R-squared or Adjusted McFadden's R-squared

- Explained Variation in Y* (in probit):
  - McKelvey-Zavoina's R-squared $\frac{Var(Y^*)-1}{Var(Y^*)}$

- Accuracy in Prediction of Y
  - Percent Predicted Correctly, Count R-squared, Adjusted Count R-squared
  - Tjur's R-squared or Coefficient of Discrimination

- Entropy-based measures for possibly non-nested models: AIC and BIC