PS2030
Political Research and Analysis
Unit 3: Models for Non-Continuous Dependent Variables
2. Maximum Likelihood Estimation

Spring 2025, Week 10

WW Posvar Hall 3600

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) 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 that generate 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 "Maximum Likelihood" 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, 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**

- Example: What is the Maximum Likelihood estimate of the mean (μ) of a normally-distributed population, given the following sample of data values: 2, 4, 6, 8?

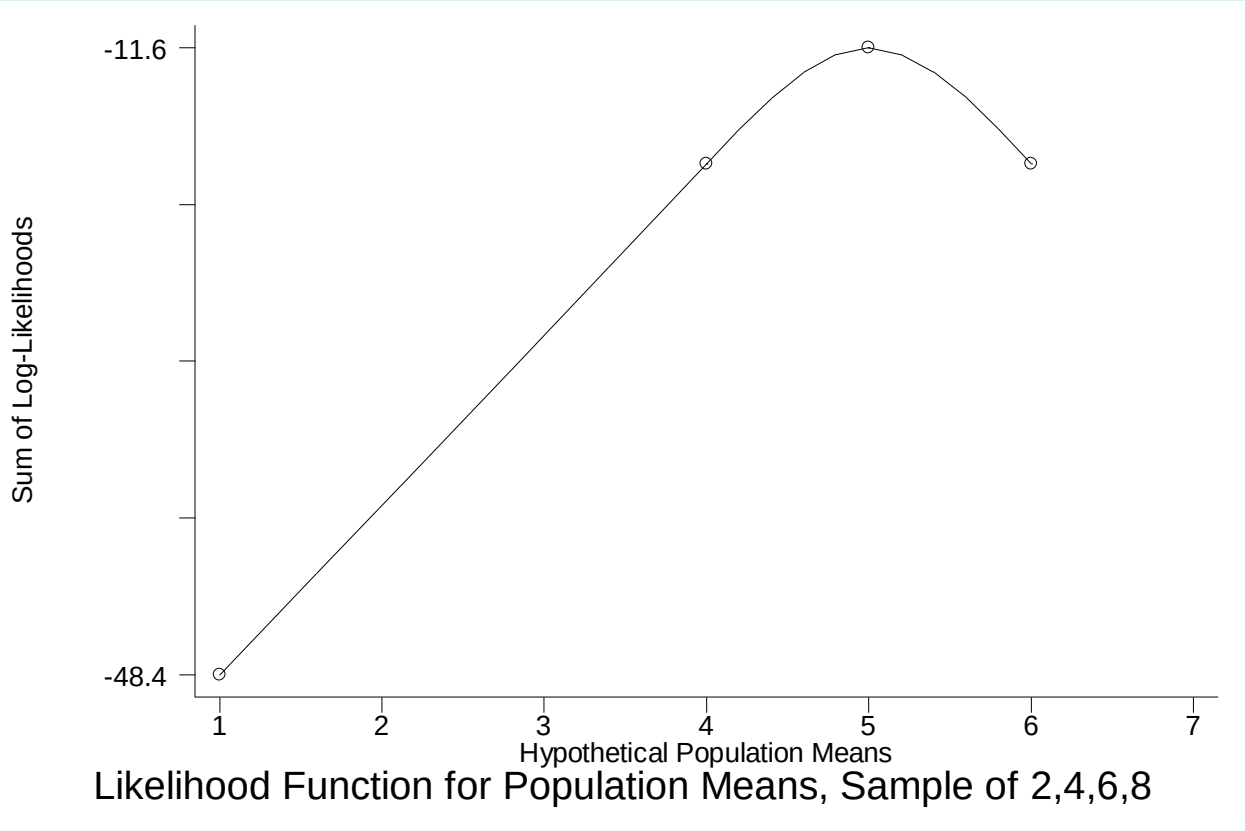
  - In ML language: given sample **s** with the assumption that Y is distributed normally, what is the *value of* μ that generated this **s** with highest likelihood?
  - Note that this is a related, but different, question to our usual "sample from a population" question: Given a μ of, say, 10, what is likelihood of observing sample values **s** and sample means $\overline{X}$ of different values, for a given sample size? We use the Central Limit Theorem and probability theory to generate a sampling distribution of sample means, etc.
  - Here we assume a fixed sample **s** and attempt to find the μ that gave the highest probability of having observed the specific **s** that we did observe

- **There is a specific likelihood of having observed our sample for all (infinite) hypothetical values of μ**
- E.g., if μ is 4, we can calculate the likelihood of observing a 2, a 4 a 6 and an 8. If μ is 6, there is another specific likelihood of observing a 2, a 4, a 6 and an 8; If μ is -7, there is another specific likelihood, etc.
- We want the value of μ that gives the highest joint likelihood of observing a sample of {2, 4, 6, 8}.
- The normal curve formula gives the likelihood, or pdf, of observing *any* value from a normal distribution as $P(y) = 1/\sqrt{2\pi\sigma^2} * e^{(-(y-\mu)2/2\sigma2}$
- So: we could fill in a bunch of possible μ values, get the ML estimate. (We lose nothing by assuming that the variance σ is 1, or any other value, as it does not change any of the **relative** calculations)
- What value of μ generates the highest joint probability of observing {2, 4, 6, 8} from a normal distribution?

- P(case 1….case 2… case 3…case 4….case N) = P(case1)*(P(case 2)* P(case3)…P(caseN)
- $= \prod (P_i)$ -- this is the "likelihood function" with product operator $\prod$
- Gets extremely hard to work with such small numbers, so can also maximize the log of the joint Ps, and that means
  - $Ln(P1)+ ln(P2)+ln(P3)+ln(P4)= \Sigma ln(P_i)$
  - $ln(e^{(-(y-\mu)2/2\sigma2)})$ for case 1, plus ln $(e^{(-(y-\mu)2/2\sigma2)})$for case 2, ln $e^{(-(y-\mu)2/2\sigma2}$ for case 3, etc.

		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

  - Stata gives you the pdf with "normalden(z-score)"
- Now we can plot the likelihood function against different values of μ



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

- How do you find the "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 tangent to the curve (the slope at that point) is 0, which is the maximum point on the curve
- For the pdf of the normal curve, you start with $-\sum(y_i-\mu)^2$, which is called the "kernel" of the likelihood function, and take the first derivative with respect to the parameter of interest (in this case, μ)

  $2*\sum y_i - 2N\,\mu$
- Set that to 0 and you get $0=2*\sum y_i - 2N\,\mu$, then $0=\sum y_i - N\,\mu$, then $N\,\mu=\sum y_i$, then $\mu=\sum y_i/N$, or the sample mean $\overline{Y}$

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 $E(Y|X) = \mu = XB$
- Assume that Y is distributed normally with mean $\mu = XB$, and standard deviation of σ^2
- What is the ML estimate of B? We have:
- $P(y)= 1/\sqrt{2\pi\sigma^2} * e^{(-(y-\mu)2)/2\sigma2}$, with $\mu = XB$ or $\beta_0 + \beta_1 * X$
- That is the pdf for each observation, depending on Y_i and μ, or the predicted Y_i from the regression line (which depends on X) – it means that points that are far from (near) the line get a very small (large) pdf corresponding to a very small (large) likelihood in a normal distribution
- The regression line that the closest to all the points will then generate the highest joint likelihood. Can get a probability of observing each Y from this distribution, and find the β that gives the highest joint Ps

- The kernel of the likelihood function is still

  $-\sum(Y_i-\mu)^2$, which is now $-\sum(Y_i - \beta_0 - \beta_1 * X)^2$
- Take first derivative w/respect to the β, and set to 0
- You get the familiar OLS scenario, since you are minimizing the squared error term, exactly what we did in Week 1 of the class.

- **So ML and OLS are identical, if one assumes normally distributed error term (or distribution of Y_i)**

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)

$$MLE(\hat{q}) \to q \text{ as } N \to \infty$$

- **Asymptotic normality**—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., Bernoulli in this case (a single trial of the binomial distribution with $P(Y=1)= \pi$)
  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 *Bernoulli* distributed variable of 1s and 0s, with

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

and with $p = F(X\beta)$ Probit

$$p = \frac{\exp(X\beta)}{1 + \exp(X\beta)}$$ Logit

Step 2: Express the joint probability of the data

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

Step 3: Calculate the log-likelihood function

$$ln\mathcal{L}(\mathrm{B}|Y) = \sum_{i=1}^{N} y_i \, ln\pi + \sum_{i=1}^{N} (1 - y_i)\ln(1 - \pi)$$

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

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

- 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 B which maximizes this quantity

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

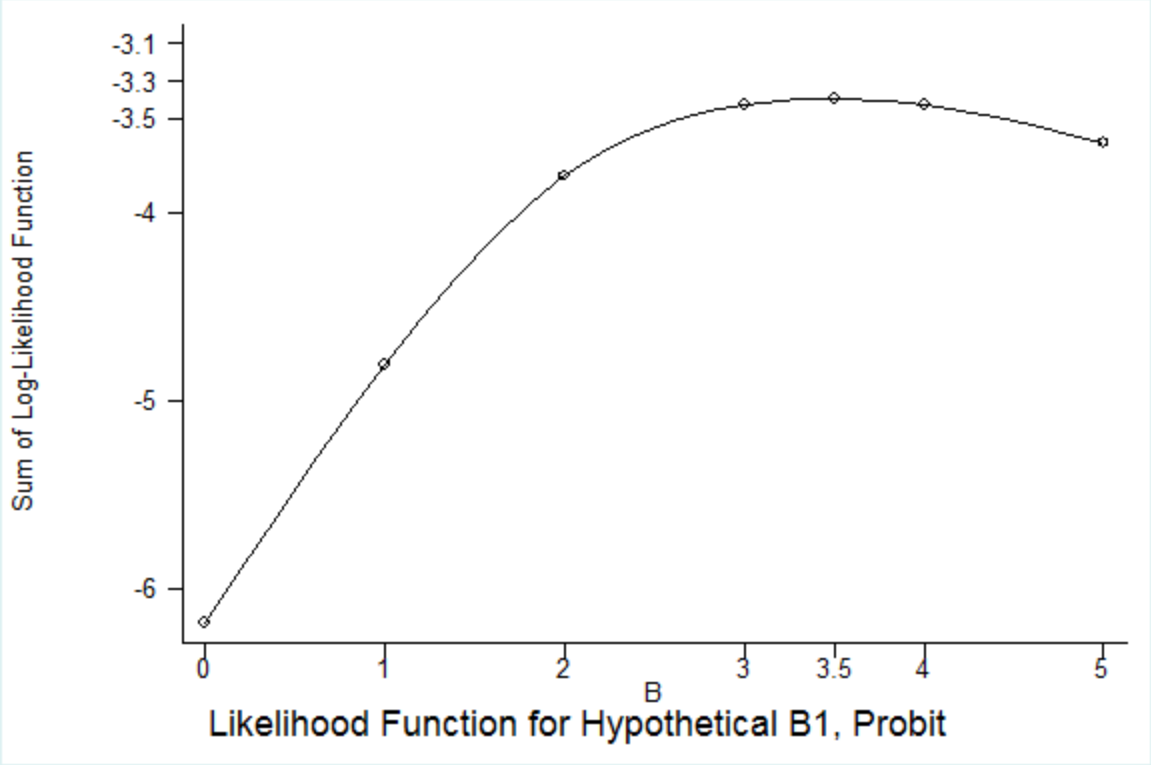
Probit Example

- We have 9 cases in a sample. 5 "1" and 4 "0".
- See "maxlike.probit.xy.dta"
- 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

So Ln L = $\sum Y_i$ ln $\Phi(\mathbf{XB})$+ $\sum(1- Y_i)$ln $(1-\Phi(\mathbf{XB}))$, where the β_1 for the slope is 0. We just have a predicted P(Y=1) for all cases equal to 5/9, or .555515.

- What is the z-score associated with P(Y=1) of .555515?
  - Stata: display invnorm(5/9)= .1397
- 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 of the 1s, and ln(1-.555515) for the 0s
- This yields a sum of the log-likelihoods of -6.1826

- Then can try different values of βs, generate new z-scores, new sums of the log-likelihoods, and try to maximize
- See "ps2030-2025.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

```
. 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
```

Statistical Tests

- Is entire equation "significant"? We can arrive at the probit/logit equivalent of F by comparing the log-likelihoods of a "full model" that includes X to a "reduced 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 $G^2 = 2 \ln L$ (Full Model) – 2 LnL(Reduced Model)
- Here LR Statistic=5.59, with 1 df (associated with 1 indep. variable)

- Null hypothesis: All 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 see this logic in terms of the "deviance" of the full model 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].
- 2*lnL(SATURATED) -(2*lnL Full Model) =

  0-2LnL (our Model) = 0-(2*-3.39)=**6.78**

- So a "Deviance" is calculated as -2* Model Log-Likelihood
- Smaller numbers for the Deviance are better (i.e., closer to 0)
-

- Deviance (Full Model) = -2*-3.39 = 6.78
- Deviance (Reduced Model) = -2*-6.18= 12.39
- So the difference of the two deviances is 5.58
- This is also the Model Chi-Square or G^2, so 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 (Reduced) model Deviance and the Full model Deviance
- This is same logic as the F test in OLS regression!
- Can compare any two nested models in terms of improvement to LnL (or Deviance), and assess whether the improvement is statistically significant

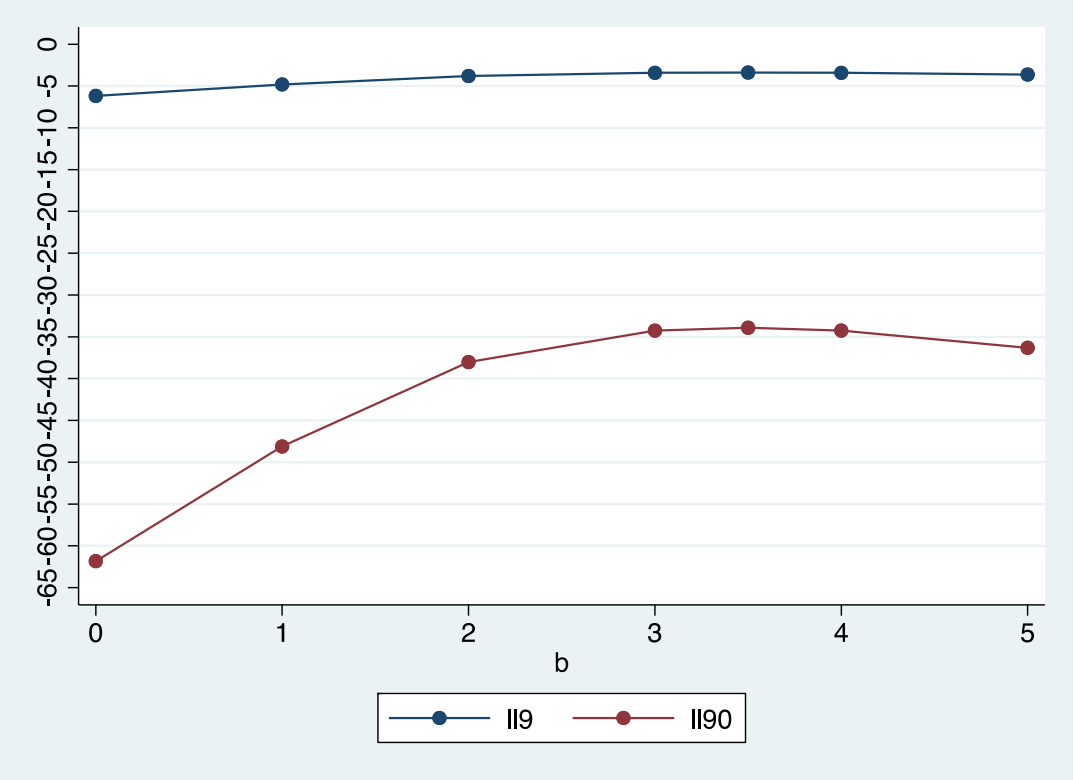
- The z test is used for the significance of individual coefficients

$$z_{b_1} = \frac{b_1}{s_{b_1}}$$

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
  - Sharp curves = large Hessian= small inverse Hessian = small standard errors
  - Flat curves = small Hessian = large inverse Hessian = large standard errors
- 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 "PS2030.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 $\beta=0$:

$$\frac{(b_1 - b_c)^2}{\hat{s}^2} = \frac{b_1^2}{\hat{s}^2}$$

Stata: "test VARNAME(S)"

Goodness of Fit Statistics in Logit/Probit

- 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. (We call the reduced model the "constrained" model because we *constrain* the β to be equal to 0.)
- 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, i.e., we achieved complete perfection in the unconstrained model's Log-Likelihood. This happens as LnLikelihood, Full or Unconstrained Model →0 !!!!
- **"McFadden" R-squared or "Pseudo R-squared":**

  1-(LnL (Unconstrained)/Ln(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

- But: as in linear regression, you can't decrease McFadden by adding new variables, so there should be a penalty for too many IVs. This results in the "Adjusted" McFadden Pseudo R-squared.
- "Adjusted" McFadden: 1-((LnL(Unconstrained)-*k*)/lnL(Constrained)) =

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

- **Adjusted McFadden will only increase if the LnL of the unconstrained model increases by more than 1 for each parameter added to the model**
- Can compare any two nested models in terms of improvement to McFadden or Adjusted McFadden and interpret the difference

- Another way to look at R-squared is the "Explained Variance in **Y***". In regular regression one calculates R^2 as:

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

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

- 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!!!!**

- Interesting 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"**:

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

  where p_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 we have used throughout the course

- 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)**

$$\mathbf{AIC = -2}\mathit{LnL}\mathbf{(M) + 2(k+1)}$$

- First term in the numerator is the Deviance of the model, second term is the penalty for the number of parameters ***k***
- We want *smaller* values for AIC; that indicates less deviance and better fit
- Sometimes you see analyses where AIC is divided by N to compare across models with different sample sizes

- **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):

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

- Here we have 2*ln(9) – (-6.78)=11.18
- This value can then be calculated for any other model (M3) and compared to M2: $BIC_{m2} - BIC_{m3}$ with more negative values preferred
- Our example: BIC(Unconstrained) = 14.56 - 11.18 BIC (Constrained) = 3.38
- Rule of thumb (Long, p. 112): BIC absolute differences between models should be greater than 5 to provide "strong" evidence in favor of one or the other. Our model is "better", but not "strongly" better

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