Discrete Data Models
&
Likelihood Functions

GEOG 210B WINTER 2018

Konstadinos G. Goulias ©

Discrete & limited

- Binary Variables (0, 1)
- Ordinal Variables (small, medium, large, xlarge, xxlarge)
- Nominal Variables (red, pink, yellow, whatever)
- Censored Variables (lots of observations at a value)
- Count Variables (1, 2, 3, 4, …,many integer value)

Dichotomous or Binary (examples)

$$y_i = \begin{cases} 1 \; if \;\; person \; i \; is \; employed \\ 0 \; otherwise \end{cases}$$

$$y_i = \begin{cases} 1 \; if \;\; fishery \; i \; is \; depleted \\ 0 \; otherwise \end{cases}$$

$$y_i = \begin{cases} 1 \; if \;\; car \; i \; has \; an \; accident \\ 0 \; otherwise \end{cases}$$

We can imagine that each event above is "driven" by a process that we do not observe (latent) and can be represented by a latent variable y* that may be: the ability of a person to get a job, a fishery's resilience to overfishing, the propensity to engage in accidents. Choice of the functional form of y* is usually based on analytical convenience

A more complete model definition (we study insect's tolerance to DDT)

The event: insect dies when sprayed with q_i dose

$$y_i = \begin{cases} 1 \; if \;\; y_i^* < q_i \\ 0 \; otherwise \end{cases}$$

Event = dead insect

y* = insect's tolerance to DDT

$$prob(y_i = 1) = prob(y_i^* < q_i)$$

$$y_i^* = \mu + \varepsilon_i \qquad\qquad y_i^* \sim N(\mu, \sigma_\varepsilon^2)$$

Polychotomous Dependent Variables

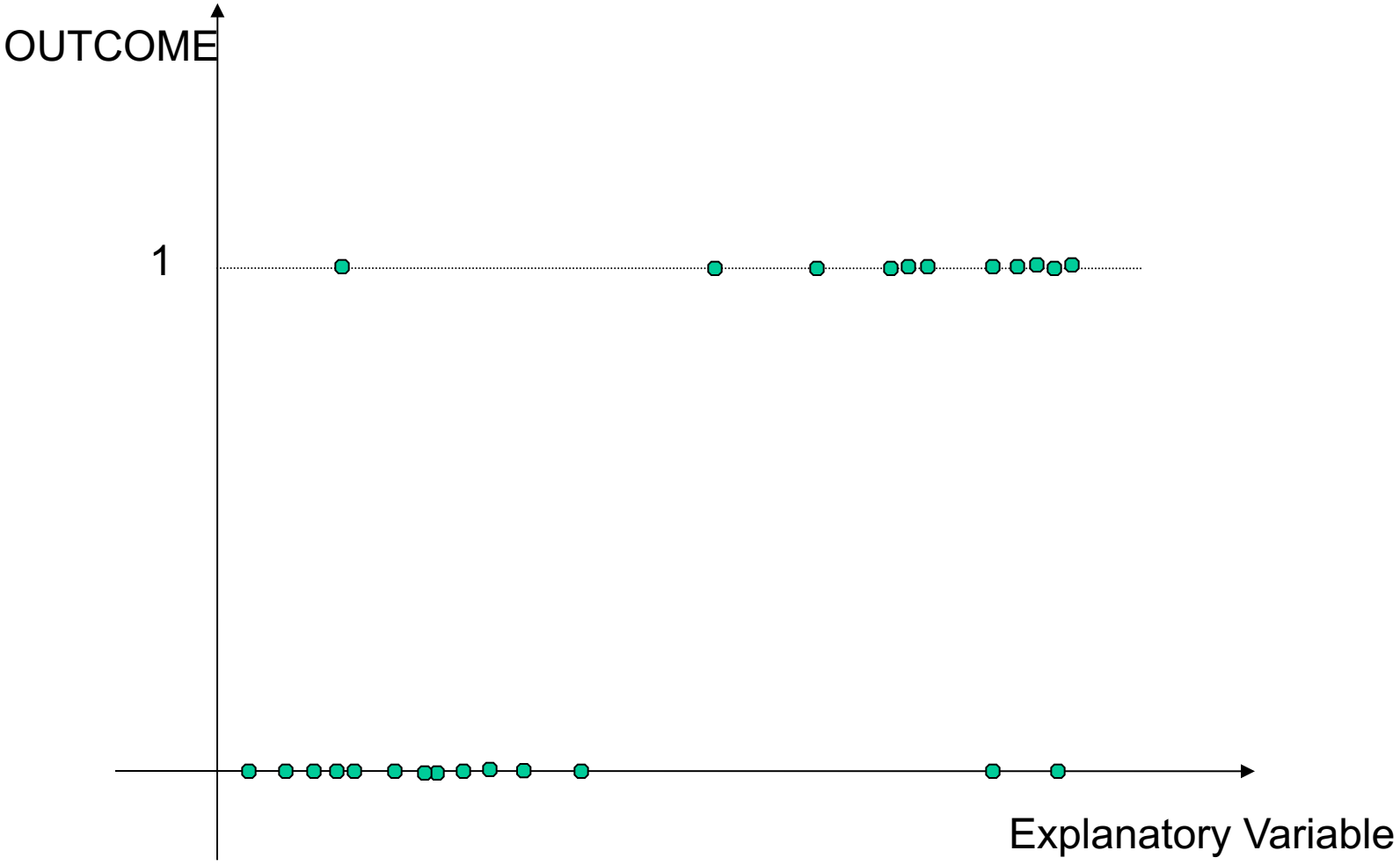
$$y_i = \begin{cases} 1 & person\ drives\ alone \\ 2 & person\ carpools \\ 3 & person\ goes\ by\ bike \\ 4 & person\ goes\ by\ other \end{cases}$$

Unordered, nominal
We assign numerical values to tell the software these are different categories

$$y_i = \begin{cases} 1 & person's\ job\ fit\ excellent \\ 2 & person's\ job\ fit\ very\ good \\ 3 & person's\ job\ fit\ good \\ 4 & person's\ job\ fit\ not\ good \end{cases}$$

Ordered, could be sequential (you need the first event for the second – school grades high school, college, masters, Ph.D.)

Would a linear model work in this case?



Probability Model

- We need a function that:

$$prob(y_i = 1) = F(X_i \beta)$$

This can tell us what happens for each observation as X changes from observation to observation. It also can tell us what is the sensitivity of the probability to each x variable

Examples in probability of employment?

Simplest Model
Linear Probability Model

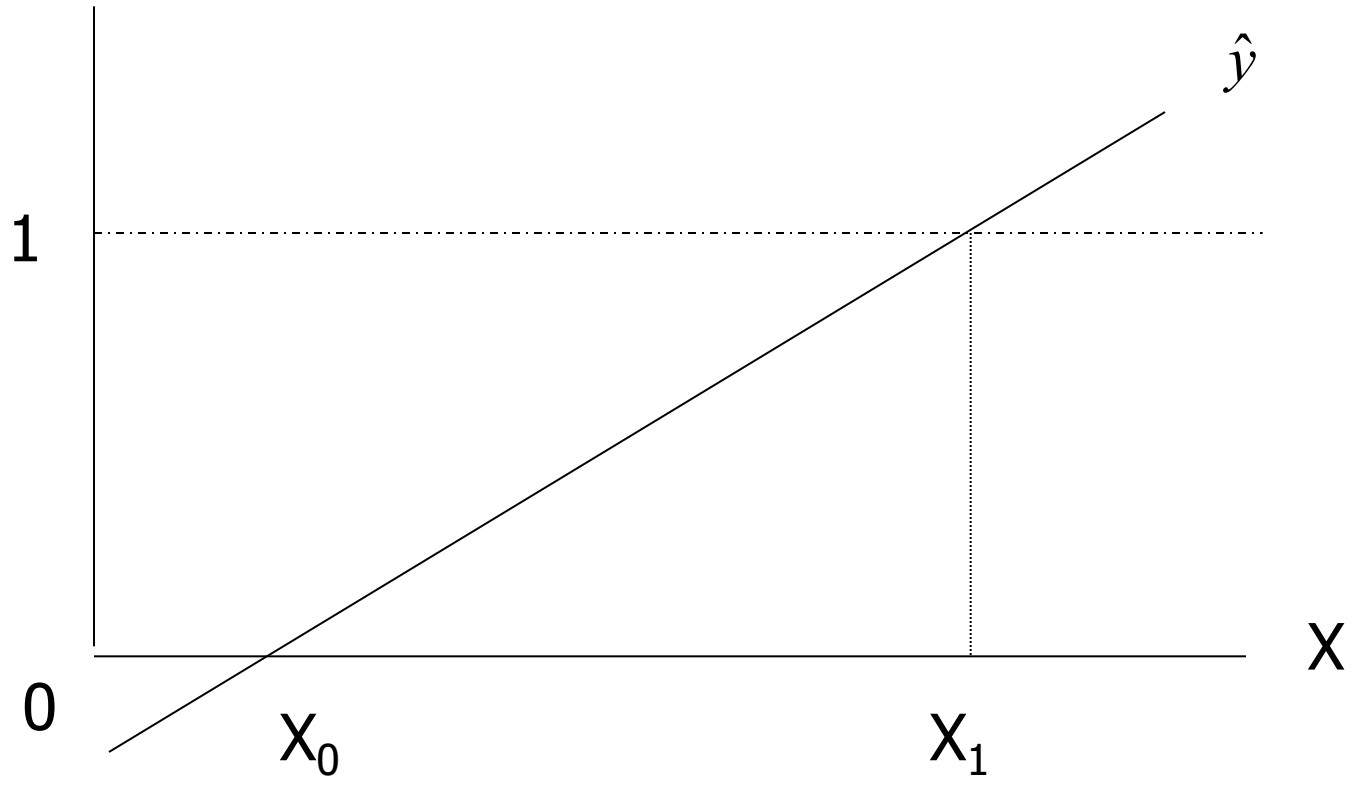
$$y_i = E[y_i] + \varepsilon_i$$

$$E[y_i] = P_i = \beta_1 + \beta_2 X_{i2} + \cdots + \beta_k X_{ik}$$

$$\text{var}(\varepsilon_i) = P_i(1 - P_i)$$

Heteroskedastic error term – need to use a weighted least squares

Picture of Linear Probability



Problems Using a Linear Model for Probability

- Predictions outside 0-1 range.
- Heteroscedasticity
  - This can be solved and a better estimator used.
- Coefficients have little meaning.
- Constant marginal effect.

$$In\ the\ linear\ \text{model}\ \frac{\partial E(y)}{\partial X_k} = \beta_k$$

Some Options

PROBIT

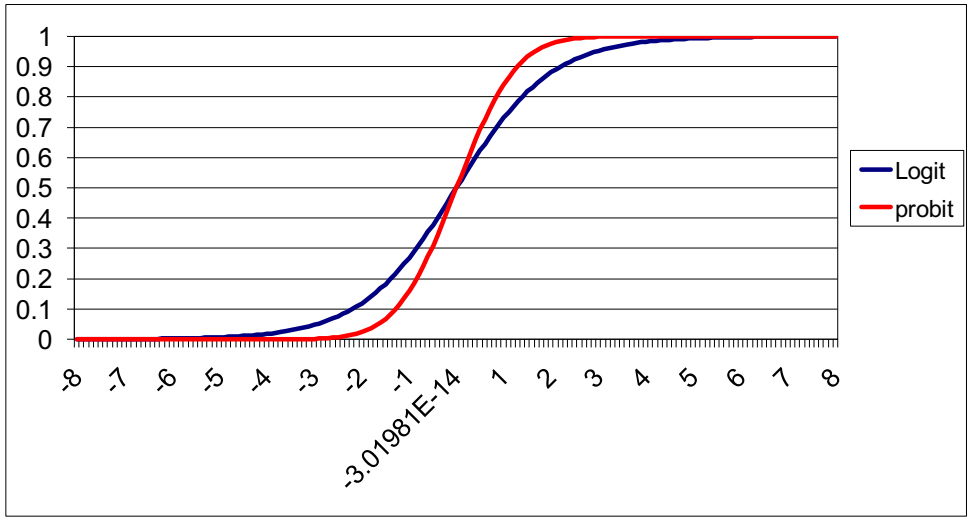
$$prob(y_i=1)=F(X_i\beta)=\Phi(X_i\beta)=\int_{-\infty}^{X_i\beta}\frac{1}{\sqrt{2\pi}}exp\left(\frac{-z^2}{2}\right)dz$$

Φ Is the standard normal cdf

LOGIT

$$prob(y_i=1)=F(X_i\beta)=\Lambda(X_i\beta)=\frac{\exp(X_i\beta)}{1+\exp(X_i\beta)}$$

Λ Is the Logistic cdf



THE PROBIT

Assume:

$$y_i^* = X_i\beta + \varepsilon_i$$

$$y_i = \begin{cases} 1 \; if \;\; y_i^* > 0 \\ 0 \; otherwise \end{cases}$$

$$\varepsilon_i \sim N(0, \sigma^2)$$

$$prob(y_i = 1|X) = prob(y_i^* > 0)$$

$$= prob(X_i\beta + \varepsilon_i > 0) = prob(\varepsilon_i > -X_i\beta)$$

$$= prob\left(\frac{\varepsilon_i}{\sigma} > -X_i\frac{\beta}{\sigma}\right) = prob\left(\frac{\varepsilon_i}{\sigma} < X_i\frac{\beta}{\sigma}\right) = \Phi\left(X_i\frac{\beta}{\sigma}\right)$$

This is standard normal with mean 0 and standard deviation 1

In estimation we cannot differentiate between beta and sigma but just their ratio. This is called lack of identification and we normalize by setting (saying) sigma is one and in this way we talk only about beta

What is the effect of each variable X on the dependent variable?

$$\frac{\partial E(y)}{\partial X_k} = \phi(X\beta)\beta_k$$

$$\phi(z) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}z^2\right)$$

The effect of X on the outcome probability depends (varies) with the level of X and the other variables in the model

For comparison:

$$In\ the\ linear\ \text{model}\ \frac{\partial E(y)}{\partial X_k} = \beta_k$$

THE LOGIT

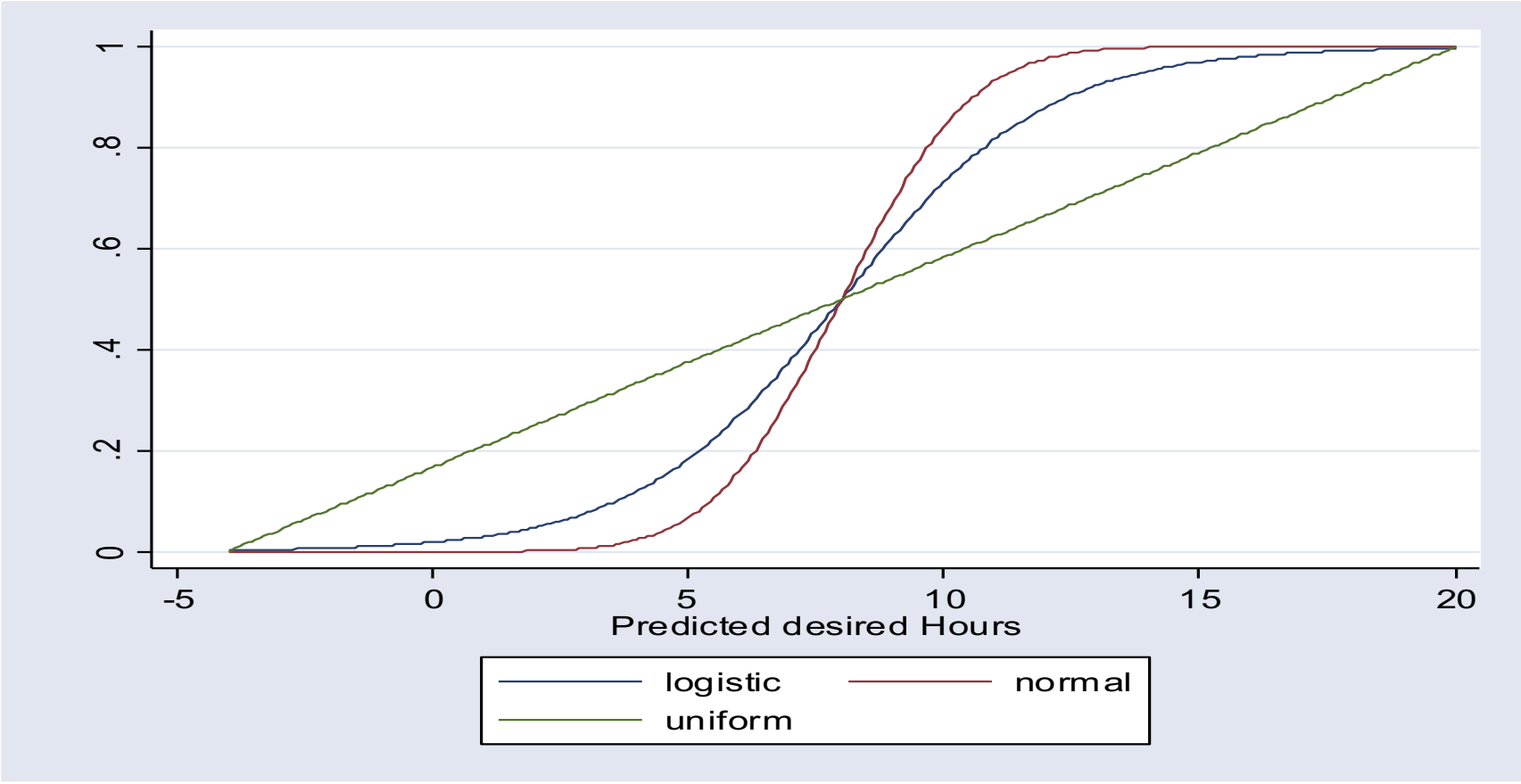
$$y_i^* = X_i\beta + \varepsilon_i$$

$$y_i = \begin{cases} 1 \; if \;\; y_i^* > 0 \\ 0 \; otherwise \end{cases}$$

$$\varepsilon_i \sim extreme$$

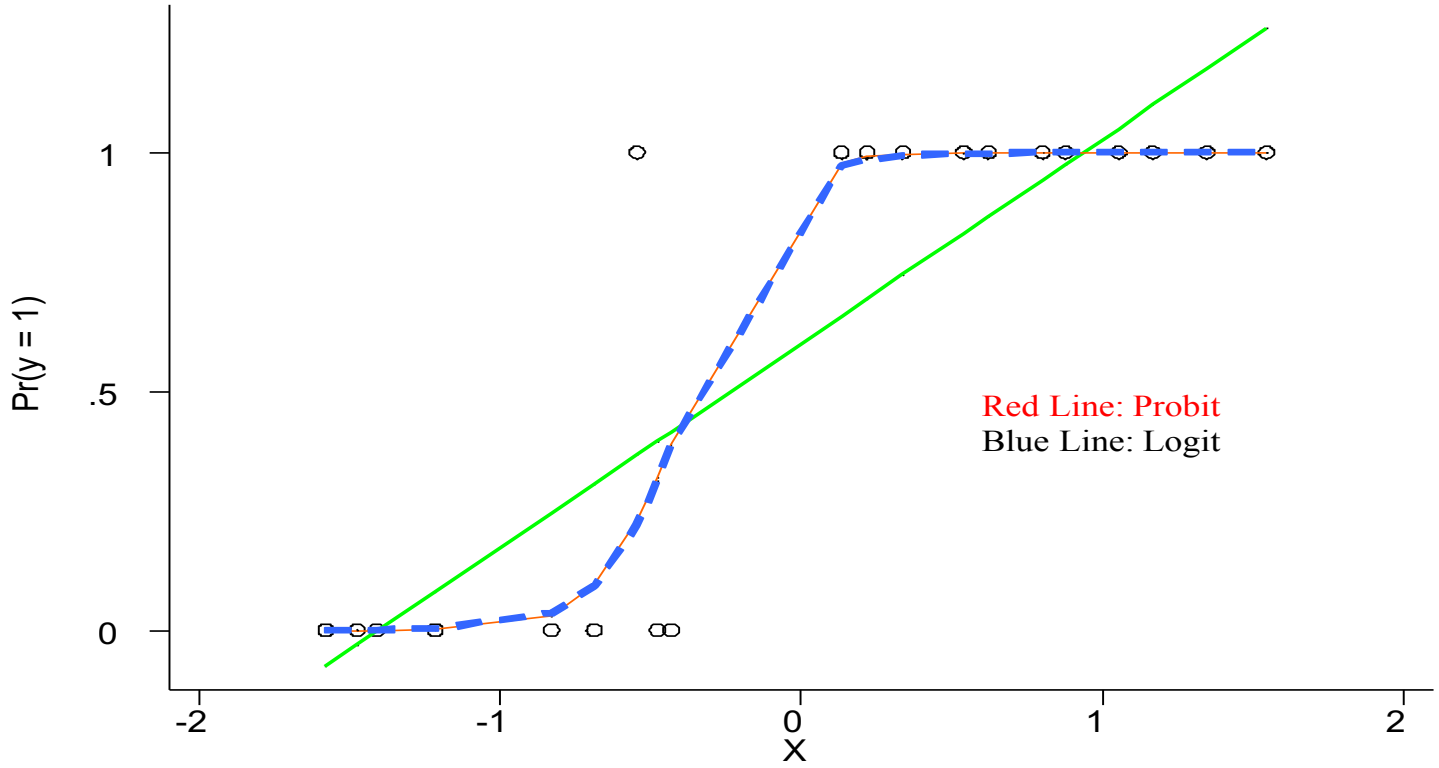
$$prob(y_i = 1|X) = \Lambda(X_i\beta) = \frac{\exp(X_i\beta)}{1+\exp(X_i\beta)}$$

$$\frac{\partial E(y)}{\partial X_k} = \frac{\exp(X\beta)}{(1+\exp(X\beta))^2}\beta_k$$

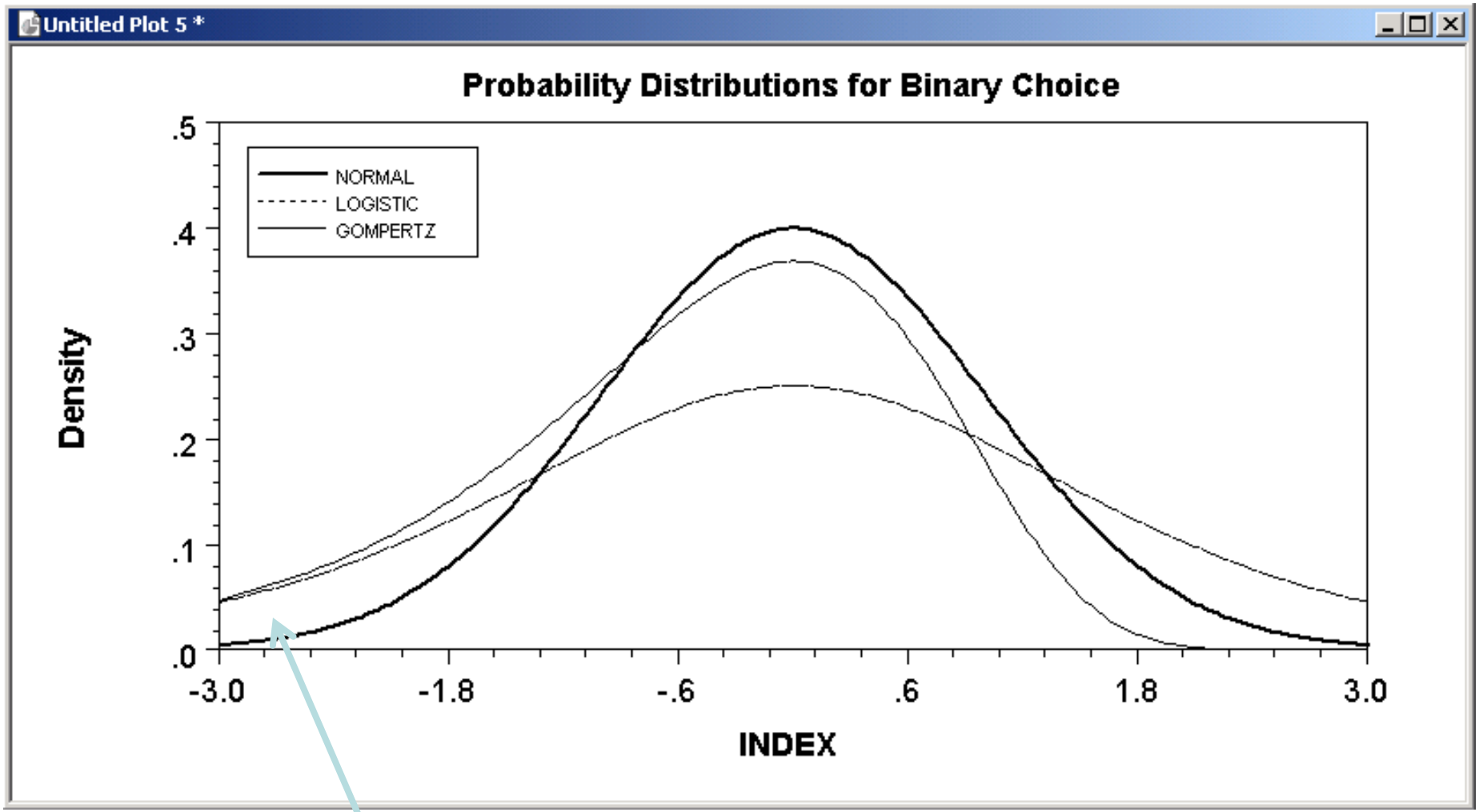

0
.2
.4
.6
.8
1
-5
0
5
10
15
20
Predicted desired Hours
logistic
normal
uniform

From Park

Probit, Logit and Linear Models



From Park



Logistic has thicker tails = more observations in extremes

Log of odds for Logit

A convenient transformation:

$$Odds(y=1|x)=\frac{P(y=1|x)}{P(y=0|x)}$$

$$Define$$

$$\log it=\log[odds(y=1|x)]$$

$$replace\ the\ probabilities\ with\ equations$$

$$\log[odds(y=1|x)]=\beta_0+\beta_1 x$$

When x changes by one unit the log odds changes by β_1 units

When x changes by one unit the odds change by exp(β_1) units

Maximum Likelihood & Related

See also: http://statgen.iop.kcl.ac.uk/bgim/mle/sslike_1.html

Overview

1. Univariate case and the Poisson model
2. Multivariate case ---> vector of parameters
3. First and second derivatives of the normal distribution
4. Cramer-Rao lower bound, the Information number, and Information matrix
5. Linear Regression Likelihood
6. Probit and Logit likelihoods
7. Notes on Non-linear Optimization

Definitions

- Consider a random sample of observations taking on the values X_i (i=1,...,N)

- The density of each observation i is $f(X_i, \beta)$.

- β is a parameter characterizing the distribution of the X_i

- After sampling of the observations has been performed the observations tell us nothing about β.

OBJECTIVE: Obtain values of estimates of β that are close to the true β in the population

FISHER'S CRITERION: After data have been collected, choose an estimate of β such that the probability of obtaining the sample actually observed is maximized

In reality, we are looking for the values of estimated β that are **the most likely** to have generated our data.

ASSUMPTION: Nature cooperated with us to give us a sample that, based on the FISHER criterion, will provide us with **estimates of β** that are very close to β

1. There is some debate on the plausibility of this criterion.
2. It is the most popular method of estimation.

- The Probability we are talking about is a property of the SAMPLE.

- The joint density function associated with our sample can be viewed as a property of the UNKNOWN parameter β

The joint density for all the observations is:

$f(X_1,X_2,X_3,..., X_N, \beta)$ and if all the observations are independent then:

$f(X_1,X_2,X_3,..., X_N, \beta) = f(X_1, \beta)\, f(X_2, \beta)\, f(X_3,\beta)... f(X_N, \beta)$

We can write this as:

$$\prod_{i=1}^{n} f(X_i, \beta) = L(\beta \mid X_1, X_2, \ldots, X_N)$$

$$\prod_{i=1}^{n} f(X_i, \beta) = L(\beta \mid X_1, X_2, \ldots, X_N)$$

This is the LIKELIHOOD FUNCTION for β given the data $X_1, X_2, X_3, \ldots, X_N$.

We can consider the likelihood to be a joint "probability" type of function. It is a function that we can maximize to find the most LIKELY value (estimate of β) to associate with β.

LIKELIHOOD EXAMPLES

EXAMPLES
(Poisson distribution)

A sample of 10 observations (x_1, x_2, x_3, x_4,...,x_{10}) from a Poisson distribution with parameter β. This can be the number of accidents at an intersection or the number of trips we make in a day.

The density of the Poisson distribution is:

$$f(x_i, \beta) = \frac{e^{-\beta} \beta^{x_i}}{x_i!}$$

Then the likelihood function will be (Greene's textbook):

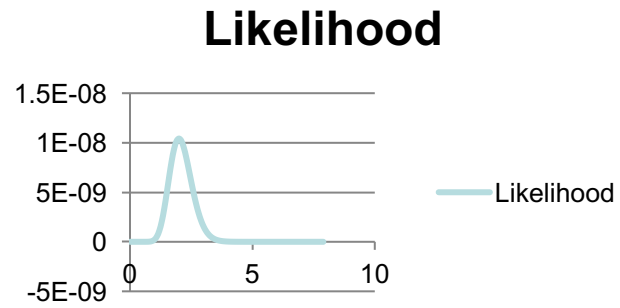
$$L(\beta) = \prod_{i=1}^{10} f(x_i, \beta) = \frac{e^{-10\beta} \beta^{\sum_{i=1}^{10} x_i}}{\prod_{i=1}^{10} x_i!}$$

My note: Greene – page 484
in 6th edition

$$L(\beta) = \prod_{i=1}^{10} f(x_i, \beta) = \frac{e^{-10\beta}\,\beta^{\sum_{i=1}^{10} x_i}}{\prod_{i=1}^{10} x_i!}$$

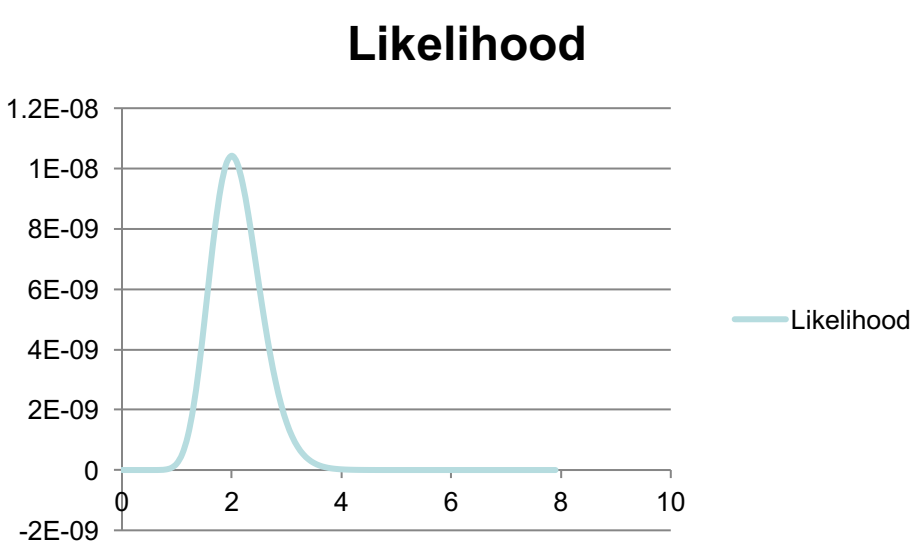
The parameter that we need to estimate is β and the data given to us are the values of the xs

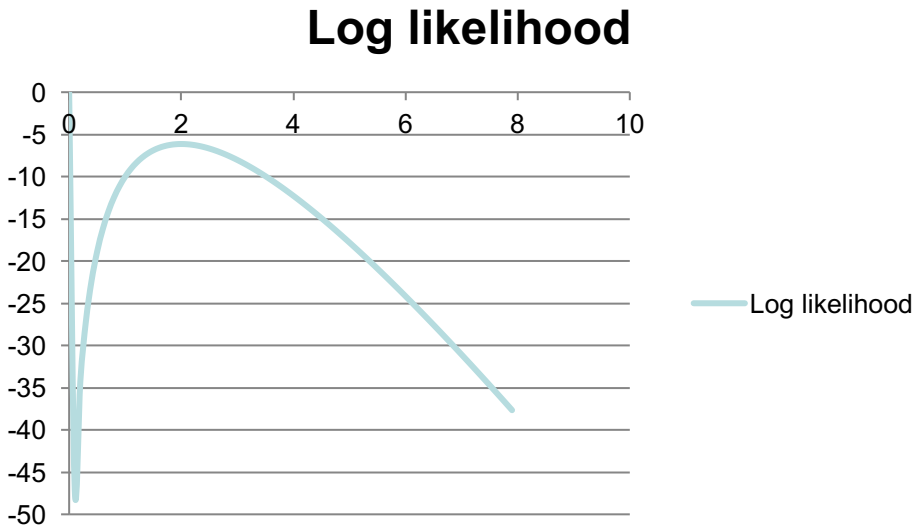
Given the data: $x_1 = 5$, $x_2 = 0$, $x_3 = 1$, $x_4 = 1$, $x_5 = 0$, $x_6 = 3$, $x_7 = 2$, $x_8 = 3$, $x_9 = 4$, $x_{10} = 1$



The likelihood becomes:

$$L(\beta) = \prod_{i=1}^{10} f(x_i, \beta) = \frac{e^{-10\beta}\,\beta^{20}}{207{,}360}$$





It is easier to work with the logarithm of the likelihood (called the log-likelihood):

$$\ln L(\beta) = -10\beta + 20 \ln \beta - \ln k$$

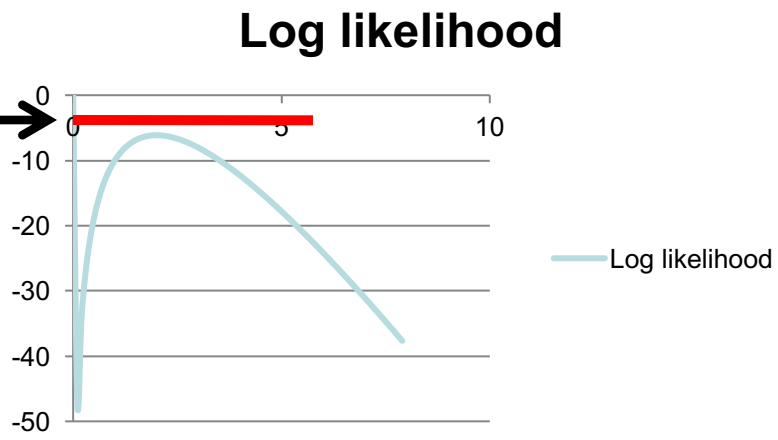
- Maximization can be performed by taking the first derivatives
- Concavity of the function is verified by the second derivative

To maximize a function we need to take the first derivative and verify using the second derivative

$$\frac{d \ln L(\beta)}{d\beta} = -10 + \frac{20}{\beta} = 0$$

we solve this to get:

$$\hat{\beta} = 2$$



- The second derivative is negative (concave down) -----> the value 2 is a maximum

The β in a Poisson distribution is also the mean of the Poisson distribution. This means that on average at an intersection (based on the data) we should expect the occurrence of 2 accidents.

$$\frac{d(\frac{20}{\beta})}{d\beta} = \frac{d(20\beta^{-1})}{d\beta} = 20\frac{d(\beta^{-1})}{d\beta} = 20(-1)\beta^{0} = -20$$

EXAMPLES
(normal distribution)

A sample of 4 observations (x_1, x_2, x_3, x_4) from a normal distribution with parameters μ (the mean) and σ (the standard deviation).

The density of the normal distribution is:

$$\frac{1}{\sigma\sqrt{(2\pi)}} e^{\left(-\frac{(x_i-\mu)^2}{2\sigma^2} \right)}$$

Then the likelihood function will be:

$$L(\mu,\sigma^2 \mid x_1, x_2, x_3, x_4) = \prod_{i=1}^{4} \left[\frac{1}{\sqrt{(2\pi\sigma^2)}} e^{\left(-\frac{(x_i-\mu)^2}{2\sigma^2}\right)} \right.$$

With some algebraic manipulation this becomes

$$\ln L(\mu,\sigma^2 \mid x_1, x_2, x_3, x_4) = -\frac{4}{2}\ln(2\pi) - \frac{4}{2}\ln\sigma^2 - \frac{1}{2}\sum_{i=1}^{4}\left[\frac{(x_i-\mu)^2}{\sigma^2} \right.$$

The parameters that we would need to estimate are μ and σ.
The values that we will find are conditional on the data!!

Talk about negative loglikelihood and maximization

We proceed as before (i.e., take the first and second derivatives with respect to the unknown parameters) and we can find that the maximum likelihood estimates for μ and σ are:

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} x_i = \bar{x}$$

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 = s^2$$

All of the above, Poisson and Normal, are valid for the Univariate case (i.e., one random variable with one mean and one variance).

MULTIVARIATE DISTRIBUTIONS

- We want to study K random variables that are related with each other.

- Assume the K random variables are multivariate normally distributed

- Population mean vector $M = (\mu_1, \mu_2, \mu_3, \ldots, \mu_K)$

- Variance matrix $\sigma^2\mathbf{I}$

- The sample consists of $X_1, \ldots, X_n$ multivariate observations

The density for each multivariate observation is:

$$f(X_i) = (2\pi)^{-M/2} \left|\sigma^2 I\right|^{-1/2} e^{-1/2(X_i - M)'[\sigma^2 I]^{-1}(X_i - M)}$$

Taking the product for all the n observations (which forms the likelihood) and then taking the logarithms we get the Log-likelihood:

$$\ln L = -\frac{nK}{2}\ln(2\pi) - \frac{nK}{2}\ln(2\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(X_i - \mu)'(X_i - \mu)$$

The first part of this does not change with the observations or with different values of the parameters to estimate.

- **Discussion**:
- How do we maximize this?
- What are the unknown parameters?
- How many are they?
- What is considered fixed?

The first derivatives are:

$$\frac{\partial \ln L}{\partial \mu} = \frac{1}{\sigma^2} \sum_{i=1}^{n} (x_i - \mu) = 0$$

These are K equations, one for each μ and for the variance

$$\frac{\partial \ln L}{\partial \sigma^2} = -\frac{nK}{2\sigma^2} - \frac{1}{2\sigma^4} \sum_{i=1}^{n} (x_i - \mu)'(x_i - \mu) = 0$$

The solution of this system of equations is:

$$\hat{\mu}_j = \bar{x}_j \quad with \quad j = 1,...,K$$

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{n} \sum_{j=1}^{k} (x_{ij} - \bar{x}_j)^2}{nK}$$

The second derivatives of the log-likelihood function are very useful quantities

The information matrix of a vector of random variables **Γ** is:

$$Inform.(\Gamma) = -E\left[\frac{\partial^2 \ln L(\Gamma \mid X)}{\partial\Gamma\partial\Gamma}\right]$$

- This is the negative of the expectation of the second order derivatives of the log likelihood function of the random variable Γ given the data X.

- The inverse of the information matrix provides a lower bound for the sampling precision for unbiased estimators of Γ.

- Then, maybe we can use the second derivatives of the log-likelihood function as estimates of the variance of the estimates

- Note: When we have one random variable the Information matrix is called Information Number.

The second derivatives of the likelihood for the multivariate normal vector are:

$$\frac{\partial^2 \ln L}{\partial \mu \partial \mu'} = -\frac{n}{\sigma^2} I$$

$$\frac{\partial^2 \ln L}{\partial (\sigma^2)^2} = \frac{nK}{2\sigma^4} + \frac{1}{\sigma^6} \sum_{i=1}^{n} (x_i - \mu)'(x_i - \mu)$$

$$\frac{\partial^2 \ln L}{\partial \mu \partial \sigma^2} = -\frac{1}{\sigma^4} \sum_{i=1}^{n} (x_i - \mu)$$

Page 493 – Greene 6th edition.

If we place all these elements in a matrix, invert it, and take the expectation we obtain:

Estimator 1:

$$[Inform.(M)\,]^{-1} = \left(- E\left[\frac{\partial^2 \ln L(M \mid X)}{\partial M \partial M} \right] \right)^{-1}$$

- If we knew the value of the expectation of the derivatives (from theory) then we could evaluate the variance elements directly.

- In most models we don't have the expectations available so we need another estimator.

Estimator 2:

$$[Estimate\ Inform.(\hat{M})\ ]^{-1} = -\left[\frac{\partial^2 \ln L(\hat{M} \mid X)}{\partial \hat{M} \partial \hat{M}}\right]^{-1}$$

- First we maximize the log-likelihood function (via the first derivatives) and then we evaluate the second derivatives of the log-likelihood function.

Note: These are not the expectations!
The second derivatives may be complicated to compute

Estimator 3:

Called the BHHH estimator: Berndt, E.K., B.H. Hall,T.E. Hall, J.A. Hausman, "Estimation and Inference in Non- linear Structural Models, Annals of Economic and Social Measurement, 3 (1974), 653-666

$$VAR(\hat{\beta}) = [Estimate\ Inform.(\hat{M})\ ]^{-1} = (G'G)^{-1}$$

$$\hat{G}_i = \frac{\partial \ln f(X_i, \hat{M})}{\partial \hat{M}}$$

- This is "theoretically" nice because the variance matrix of the first derivatives is the expected matrix of the second derivatives
- This is "practically" nice because we don't have to compute more than the first derivatives
- Different variance estimators produce different numbers by substantial amounts

My note: come back later on this

1. MAXIMUM LIKELIHOOD FOR THE **LINEAR** REGRESSION MODEL

The Data given are Y, and the matrix of X. The random variables to estimate are the β and σ.

The model is the typical linear regression model written in a different way:

$$\varepsilon_i = y_i - \beta' X_i$$

Our assumption: These are n independent, identically distributed, and normally distributed random variables

Based on what we have seen above, the likelihood function can be written as:

$$L = (2\pi\sigma^2)^{-n/2} e^{\left(\frac{-1}{2\sigma^2}\right)} \sum_{i=1}^{n} (y_i - \beta' x_i)^2$$

Remember that the Sum can also be written in matrix notation as:

$$\sum_{i=1}^{n} (y_i - \beta' x_i)^2 = (y - X\beta)'(y - X\beta)$$

If we take the logarithm of the likelihood we obtain the log-likelihood function:

$$\ln L = -\frac{n}{2}\ln 2\pi - \frac{n}{2}\ln\sigma^2 - \frac{1}{2\sigma^2}(y - X\beta)'(y - X\beta)$$

This begins to look like the form of least squares. The derivatives of this with respect to the unknown variables β, and σ are:

$$\frac{\partial \ln L}{\partial \beta} = \frac{1}{\sigma^2} X'(Y - X\beta) = 0$$

and

$$\frac{\partial \ln L}{\partial \sigma} = -\frac{n}{2\sigma^2} + \frac{1}{2}\sigma^4 (Y - X\beta)'(Y - X\beta) = 0$$

The solution of this gives:

$$\hat{\beta} = (X'X)^{-1} X'Y$$

Which is exactly the least squares solution. Solving a similar equation for the variance, we get:

$$\hat{\sigma}^2 = \frac{e'e}{n}$$

Note: This time the variance estimate is not the same as the one computed in least squares regression. The least squares was divided by n-k. Asymptotically (as $n \to \infty$), least squares and max lik. coincide.

ESTIMATION FOR DISCRETE DATA MODELS

MAXIMUM LIKELIHOOD FOR DISCRETE DEPENDENT BINARY VARIABLES

- Each observation is considered as a single draw from a Bernoulli distribution
- The probability of each occurrence is $F(\beta' x_i)$
- All observations are independent from each other
- The joint probability is the likelihood function

$$Prob(y_1, y_2, y_3, \ldots, y_n) = \prod_{for\ y=1} F(\beta' x_i) \prod_{for\ y=0} (1 - F(\beta' x_i))$$

Equivalently we can write this as:

$$L = Prob(y_1, y_2, y_3, \ldots, y_n) = \prod_{i=1}^{N} [F(\beta' x_i)]^{y_i} [1 - F(\beta' x_i)]^{1-y_i}$$

$$\ln L = \sum_{i=1}^{n} [y_i \ln F(\beta' x_i) + (1 - y_i) \ln(1 - F(\beta' x_i))]$$

My Note: talk about the values of y and what happens for y=1, y=0

2. MAXIMUM LIKELIHOOD FOR THE **LOGIT** REGRESSION MODEL

The logit model is

$$F(\beta'X) = \frac{e^{\beta'X}}{1+e^{\beta'X}}$$

We can replace this in the log-likelihood function and then take the first derivatives to obtain:

$$\frac{\partial \ln L}{\partial \beta} = \sum_{i=1}^{n} (y_i - \frac{e^{\beta'X}}{1+e^{\beta'X}}) x_i$$

This is non-linear and to obtain estimates of β we need some kind of iterations that make these derivatives zero

3. MAXIMUM LIKELIHOOD FOR THE **PROBIT** REGRESSION MODEL

The Probit model is

$$F(X,\beta) = \int_{-infinity}^{\beta'X} \phi(t)dt = \Phi(\beta'X)$$

Similarly with Logit we get the first derivatives as:

$$\frac{\partial \ln L}{\partial \beta} = \sum_{y=0} \frac{-\phi_i}{1-\Phi_i} x_i + \sum_{y=1} \frac{\phi_i}{\Phi_i} x_i$$

This is also highly non linear and we need numerical solution with iterations because we have integrals to deal with.

Iterative Max/Min procedures

Common elements to all methods: 1) initial values for βs; 2) search direction; 3) step size; 4) at each step compute the first derivatives

Some methods: 1) use the second derivatives or 2) an approximation of the second derivatives

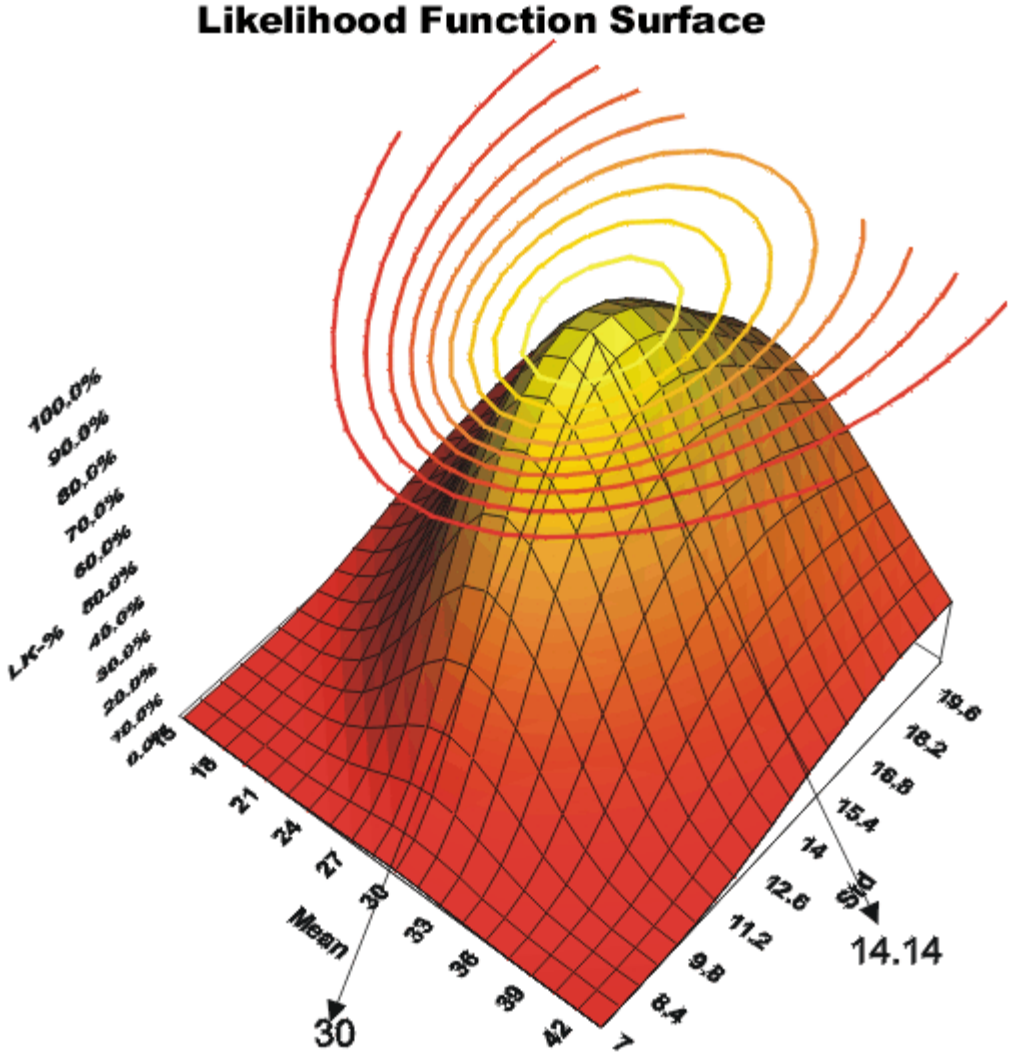
Remember: We get many derivatives because we maximize a function with respect to all the unknown parameters. When our algorithm converges to a solution ALL the first derivatives should be equal to zero.

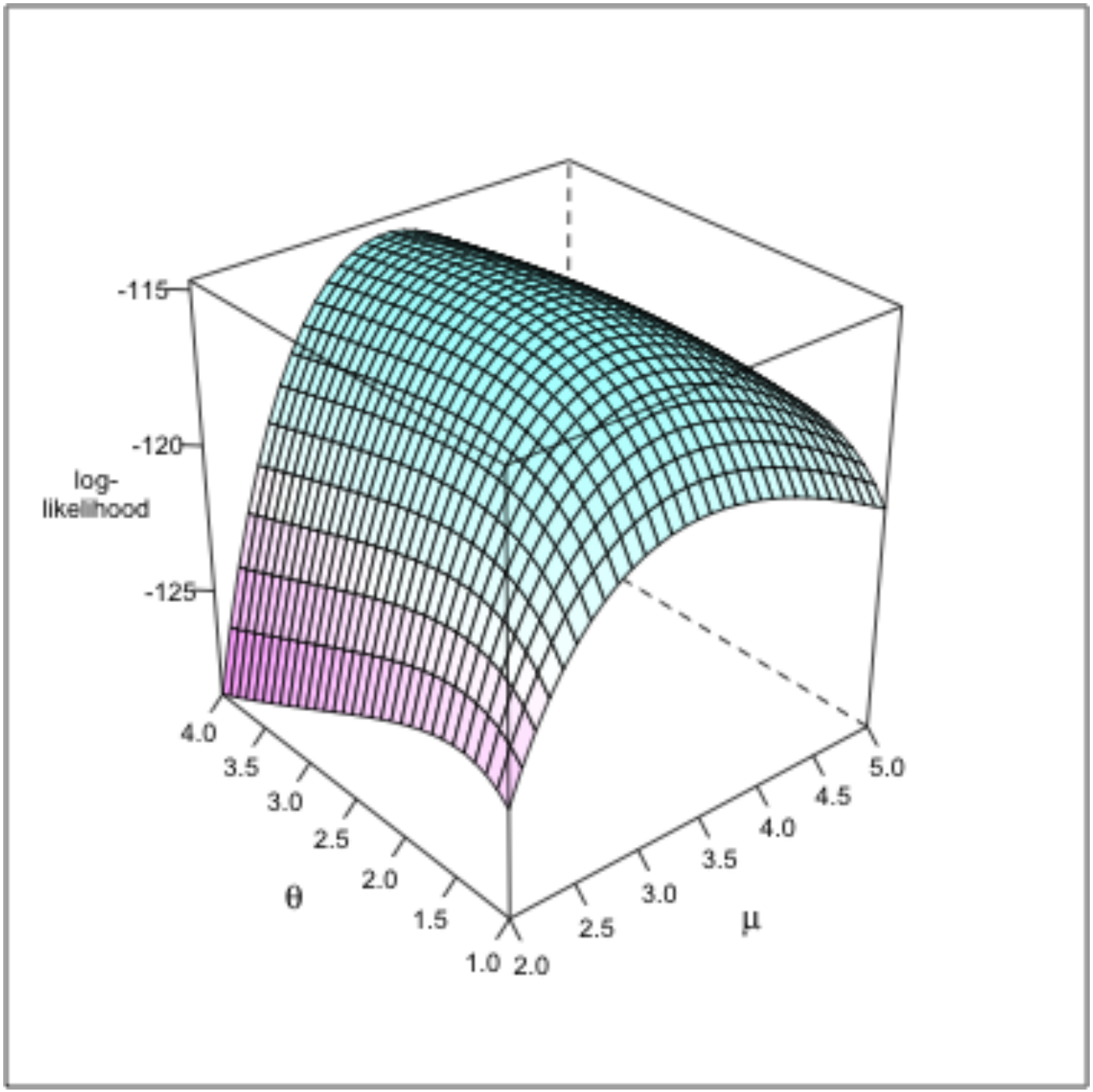
The beta values in a sequence of iterations should start becoming increasingly similar as we move toward the maximum of the likelihood

For the usual regression models we estimate we do not need many iterations (tens are many)

More coefficients require more iterations

From
http://www.weibull.com/LifeDataWeb/maximum_likelihood_estimation_appendix.htm




-115
-120
log-
likelihood
-125
4.0
3.5
3.0
2.5
2.0
1.5
1.0
θ
2.0
2.5
3.0
3.5
4.0
4.5
5.0
μ

Likelihood Ratio Test

- Can't use F-test ---- because there are no residuals to sum and subtract from TSS
- Likelihood Ratio (LR) test is the equivalent
- Intuition -- see if the restriction changes the likelihood significantly
- Test Statistic : $LR = -2\left[\ln L_R - \ln L_U\right]$
- Less variables = Restricted
- Critical Value :χ^2 with d.f. equal to no. of restrictions