Maximum-Likelihood Estimation & QualDep Models

Maximum-Likelihood Estimation

$\mathbf{y} \sim f(\mathbf{X}, \boldsymbol{\theta})$, conditional independence $\Rightarrow$

$$L(\mathbf{y}) = \prod_{i=1}^{N} f_i(\mathbf{x}, \boldsymbol{\theta}) \Rightarrow \ln\big(L(\mathbf{y})\big) \equiv \boldsymbol{\mathcal{L}} = \sum_{i=1}^{n} \ln\Big[f_i(\mathbf{x}, \boldsymbol{\theta})\Big]$$

Maximum-likelihood Estimator (MLE):

(1) Parameter estimates solve: $\dfrac{\partial \boldsymbol{\mathcal{L}}}{\partial \boldsymbol{\theta}} = \nabla_{\boldsymbol{\theta}} \boldsymbol{\mathcal{L}} = 0$

(2) Estimated variance of Estimated Parameters: $-\boldsymbol{\mathcal{H}}^{-1} \equiv -\left[\dfrac{\partial^2 \boldsymbol{\mathcal{L}}}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'}\right]^{-1}$

Properties of MLE: *BANC*

1. Asymptotically Normal
2. Consistent
3. Minimum-Variance
4. Invariance

Minimum Variance means Obtains Cramér-Rao Lower Bound

Cramér-Rao Lower Bound $\longrightarrow$ for any "well-behaved" p.d.f., any unbiased estimator of parameter θ, call it $\hat{\theta}$, will have variance at least as large as:

$$\left[I(\theta)\right]^{-1} \equiv \left(-E\left[\frac{\partial^2 \ln L(\theta)}{\partial \theta^2}\right]\right)^{-1}$$

- $L(\theta)$ here is the "likelihood function" of the data, i.e., loosely, it is

$L(\theta) = \prod_{i=1}^{n} f(X_i, \theta)$; e.g. suppose X_i drawn from exponential distribution,

$$X_i \sim f(X_i, \theta) = \theta e^{-\theta X_i}$$

θ = hazard rate

$\Rightarrow$ life-expectancy $= \frac{1}{\theta}$ = mean

$\frac{1}{\theta^2}$ = var

then, $L(\theta) = \prod_{i=1}^{n} f(X_i, \theta) = \prod_{i=1}^{n} \theta e^{-\theta X_i} = \theta^n e^{-\theta \sum X_i}$

thus, $\ln L(\theta) = n \ln \theta - \theta \sum_{i=1}^{n} X_i$ $\longrightarrow \hat{\theta}_{ML} = \left(\frac{\sum X_i}{n}\right)^{-1}$

so, $\frac{\partial^2 \ln L(\theta)}{\partial \theta^2} \equiv \frac{\partial\left(\frac{\partial \ln L(\theta)}{\partial \theta}\right)}{\partial \theta} = \frac{\partial\left(\frac{n}{\theta} - \sum X_i\right)}{\partial \theta} = -\frac{n}{\theta^2}$

So, the C-R lower bound $= \left(-E\left(-\frac{n}{\theta^2}\right)\right)^{-1} = \left[E\left(\frac{n}{\theta^2}\right)\right]^{-1} = \left[\frac{n}{\theta^2}\right]^{-1}$

$$= \frac{\theta^2}{n}$$

$$\nabla_{\theta} \mathcal{L} = \frac{n}{\theta} - \sum x_i = 0 \Rightarrow \hat{\theta}_{ML} = \frac{\sum x_i}{n} = \bar{x}$$

Example 2: MLE for a Poisson Count Model

1. We view outcomes as being produced by some p.f. or p.d.f., $f(x_i, \underset{\sim}{\theta})$
2. If each outcome is i.i.d. (i.e., is from a S.R.S.), then their joint distribution is the product of their marginal distributions:

$$f(X_1, X_2, X_3, \ldots, X_n, \underset{\sim}{\theta}) = f(x_1, \underset{\sim}{\theta}) f(x_2, \underset{\sim}{\theta}) f(x_3, \underset{\sim}{\theta}) \cdots f(x_n, \underset{\sim}{\theta})$$

3. What we do now is choose $\underset{\sim}{\theta}$ (i.e. estimate it) so that it maximizes the "probability" or likelihood of our having observed the data we actually have observed.

That is, given the data, X, & the model $f(\cdot)$, define the (joint) conditional distribution of θ given X. Maximize this with respect to θ.

Examples:

1. Each $X_i \sim \text{Poisson}(\theta)$: $f(x_i, \theta) = \dfrac{e^{-\theta}\theta^{x_i}}{x_i!}$ (A Poisson Distribution describes a count (R.V.) of events happening at average rate θ per fixed period)

a. $f(X_1, \ldots, X_n, \theta) = \prod_{i=1}^{n} \dfrac{e^{-\theta}\theta^{x_i}}{x_i!}$

b. $\underset{\theta}{\text{Max}}\ f(\underset{\sim}{x}, \theta) = \underset{\theta}{\text{Max}}\ \ln f(\underset{\sim}{x}, \theta) = \underset{\theta}{\text{Max}} \sum_{i=1}^{n} \left[\ln(e^{-\theta}\theta^{x_i}) - \ln(x_i!)\right]$

$= \underset{\theta}{\text{Max}} \underbrace{\sum_{i=1}^{n} (-\theta + x_i \ln\theta - \ln(x_i!))}_{\text{call this } L(\theta)}$

c. First-Order Condition $\dfrac{\partial L(\theta)}{\partial \theta} = 0$

$$L(\theta) = \sum_{i=1}^{n} (-\theta + x_i \ln\theta - \ln x_i) = -n\theta + \left(\sum_{i=1}^{n} x_i\right)\ln\theta - \sum_{i=1}^{n} \ln(x_i!)$$

F.O.C. $\dfrac{\partial L(\theta)}{\partial \theta} = 0 \Rightarrow -n + \dfrac{\sum_{i=1}^{n} x_i}{\theta^*} = 0$

$\Rightarrow \theta^* = \frac{1}{n}\sum_{i=1}^{n} x_i = \bar{X}$

e. Second-Order Condition:

$$0 > \frac{\partial\left(\frac{\partial L(\theta)}{\partial \theta}\right)}{\partial \theta} \qquad \frac{\partial(-n + \theta^{-1} \cdot \Sigma x_i)}{\partial \theta} = -\theta^{-2}\underbrace{(\Sigma x_i)}_{\text{greater than zero}}$$

$\Rightarrow \dfrac{\partial^2 L(\theta)}{\partial \theta^2} < 0$ ✓

Thus, MLE for Poisson parameter, θ, is $\hat{\theta} = \bar{X}$.

VI. C. 2. MLE Example 2: MLE for Normal

a. $f_n(x, \underset{\sim}{\theta}) = (2\pi\sigma^2)^{-1/2} e^{-(x_i-\mu)^2/2\sigma^2}$

b. $L(\underset{\sim}{\theta}) = \prod_{i=1}^{n} (2\pi\sigma^2)^{-1/2} e^{-(x_i-\mu)^2/(2\sigma^2)}$

$$\ln L(\underset{\sim}{\theta}) = \sum_{i=1}^{n} \ln\left[(2\pi\sigma^2)^{-1/2} e^{-(x_i-\mu)^2/(2\sigma^2)}\right]$$

$$= \sum_{i=1}^{n}\left[-\tfrac{1}{2}\ln(2\pi\sigma^2) - (x_i-\mu)^2/(2\sigma^2)\right]$$

$$= \sum_{i=1}^{n}\left(-\tfrac{1}{2}\ln 2 - \tfrac{1}{2}\ln\pi - \tfrac{1}{2}\ln\sigma^2 - \tfrac{1}{2}(x_i-\mu)^2/\sigma^2\right)$$

$$= -\tfrac{n}{2}\ln 2 - \tfrac{n}{2}\ln\pi - \tfrac{n}{2}\ln\sigma^2 - \tfrac{1}{2}\sum_{i=1}^{n}(x_i-\mu)^2(\sigma^2)^{-1}$$

c. F.O.C. $\dfrac{\partial \ln L(\underset{\sim}{\theta})}{\partial \underset{\sim}{\theta}} = 0$:

i.) $\dfrac{\partial \ln L(\underset{\sim}{\theta})}{\partial \mu} = (\sigma^2)^{-1}\sum_{i=1}^{n}(x_i-\mu) = 0$

ii.) $\dfrac{\partial \ln L(\underset{\sim}{\theta})}{\partial \sigma^2} = \dfrac{n}{2\sigma^2} + \tfrac{1}{2}\cdot(\sigma^2)^{-2}\sum_{i=1}^{n}(x_i-\mu)^2 = 0$

- Multiply both sides of (i) by $(\sigma^2)^{-1}$ $\Rightarrow \sum_{i=1}^{n}(x_i-\mu) = 0$

$$\Rightarrow \sum_{i=1}^{n} x_i - n\mu^* = 0$$

$$\Rightarrow \mu^* = \tfrac{1}{n}\sum_{i=1}^{n} x_i = \bar{x}$$

- Substitute that into (ii) & you'll find $\sigma^{*2} = \tfrac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2$

Of course, this all more interesting & useful when have a model of parameters…

$$y_i \overset{iid}{\sim} N(\mu_i, \sigma^2) \quad ; \mu_i = \beta X_i$$

$$f(y_i, \mu_i, \sigma^2) = (2\pi\sigma^2)^{-1/2} e^{-(y_i-\beta X_i)^2/(2\sigma^2)^{-1}}$$

$$L(\underset{\sim}{\mu}, \sigma^2 | X) = \prod_{i=1}^{n} \left[(2\pi\sigma^2)^{-1/2} e^{-(y_i-\beta X_i)^2/2\sigma^2} \right]$$

$$\ln L(\underset{\sim}{\mu}, \sigma^2 | \underset{\sim}{X}) = \sum_{i=1}^{n} \left[-\tfrac{1}{2} \ln(2\pi\sigma^2) - (y_i-\beta X_i)^2/2\sigma^2 \right]$$

$$= \sum_{i=1}^{n} \left[-\tfrac{1}{2}\ln 2 - \tfrac{1}{2}\ln\pi - \tfrac{1}{2}\ln(\sigma^2) - \tfrac{1}{2}(y_i-\beta X_i)^2/\sigma^2 \right]$$

$$= -\tfrac{n}{2}\ln 2 - \tfrac{n}{2}\ln\pi - \tfrac{n}{2}\ln(\sigma^2) - \tfrac{1}{2}\left[\sum_{i=1}^{n} (y_i-\beta X_i)^2 (\sigma^2)^{-1} \right]$$

$$\max_{\beta,\sigma^2} \ln L(\cdot)$$

F.O.C. i) $\dfrac{\partial \ln L}{\partial \beta} = -(\sigma^2)^{-1} \sum_{i=1}^{n} (y_i - \hat{\beta} x_i) \cdot (-x_i) = 0$

$$\Rightarrow \cancel{(\sigma^2)^{-1}} \sum_{i=1}^{n} (y_i - \hat{\beta} x_i) \cdot x_i = 0$$

$$(a) \quad \sum_{i=1}^{n} (y_i \cdot x_i - \hat{\beta} x_i^2) = 0$$

$$\Rightarrow \sum \hat{\beta} x_i^2 = \sum y_i x_i$$

$$\boxed{\hat{\beta} = \frac{\sum x_i y_i}{\sum x_i^2}}$$

check S.O.C.

$\dfrac{\partial (a)}{\partial \beta} < 0$ ✓

Holy Wa!

There it is again!

ii) $\frac{\partial \ln L}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{2}(\hat{\sigma}^2)^{-2} \sum (y_i - \beta x_i)^2 = 0$

$$\Rightarrow -n + (\hat{\sigma}^2)^{-1} \sum (y_i - \beta x_i)^2 = 0$$

$$\Rightarrow (\hat{\sigma}^2)^{-1} \sum (y_i - \beta x_i)^2 = n$$

$$\Rightarrow \frac{n}{1/\hat{\sigma}^2} = \sum (y_i - \beta x_i)^2$$

$$\hat{\sigma}^2 n = \sum (y_i - \beta x_i)^2$$

average squared residual

$$\hat{\sigma}^2 = \frac{1}{n} \sum (y_i - \beta x_i)^2$$

S.O.C. $-(\hat{\sigma}^2)^{-2} \sum (y_i - \beta x_i)^2 < 0$ ✓

One more time in more-compact notation?

$$f(\underset{\sim}{Y} \mid X, \underset{\sim}{\beta}, \sigma^2) = \prod_{i=1}^{n} (2\pi\sigma^2)^{-1/2} e^{-(y_i-\mu_i)^2/2\sigma^2}$$

$$L \equiv \text{Likelihood} = \prod_{i=1}^{n} (2\pi\sigma^2)^{-1/2} e^{-(y_i-X_i\beta)^2/2\sigma^2}$$

↳ likelihood is interpreted as "the probability of observing the data you actually observed, given the model"

4. <u>Maximum</u> Likelihood, then, proceeds by maximizing this probability. Sums being easier to maximize than products, this is most easily done by first taking the log of the likelihood.

$$LL \equiv \log \text{likelihood} = \sum_{i=1}^{n} \left\{ -\tfrac{1}{2} \ln(2\pi\sigma^2) - (y_i - X_i\beta)^2/2\sigma^2 \right\}$$

$$LL = \sum_{i=1}^{n} \left\{ -\tfrac{1}{2}(\ln 2 + \ln\pi + \ln\sigma^2) - (y_i - X_i\beta)^2/2\sigma^2 \right\}$$

$$\max_{\beta,\sigma^2} LL \Rightarrow \text{①}\ \frac{\partial LL}{\partial \beta} = 0 = \sum_{i=1}^{n} \left\{ +2(y_i - X_i\beta)X_i/2\sigma^2 \right\} = \sum_{i=1}^{n} +(y_i - X_i\beta)X_i/\sigma^2$$

$$\&\ \text{②}\ \frac{\partial LL}{\partial \sigma^2} = 0 = \sum_{i=1}^{n} \left\{ -\tfrac{1}{2}\cdot\frac{1}{\sigma^2} + (y_i - X_i\beta)^2 (2\sigma^2)^{-2}\cdot 2 \right\}$$

$$\text{①} \Rightarrow 0 = \frac{1}{\sigma^2}\sum_{i=1}^{n} (y_i - X_i\beta)X_i \Rightarrow \sum_{i=1}^{n} (y_i - X_i\beta)X_i = 0$$

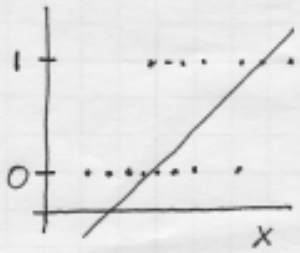
$$\text{②} \Rightarrow 0 = -\frac{n}{2\sigma^2} + \frac{2}{(2\sigma^2)^2}\sum_{i=1}^{n} (y_i - X_i\beta)^2 \Rightarrow \frac{n}{\sigma^2} = \frac{1}{(\sigma^2)^2}\sum (y_i - X_i\beta)^2$$

$$\text{①} \Rightarrow X'Y - X'X\underset{\sim}{\beta} = 0 \Rightarrow \underset{\sim}{\beta} = (X'X)^{-1}X'Y$$

$$\text{②} \Rightarrow n\sigma^2 = \sum_{i=1}^{n} (y_i - X_i\beta)^2 \Rightarrow \sigma^2 = \frac{1}{n}\sum_{i=1}^{n} \varepsilon_i^2 \Rightarrow \sigma^2 = \frac{\varepsilon'\varepsilon}{n}$$

5. So, OLS is the MLE of $\underset{\sim}{\beta}$ for the CNLRM.

Logit / Probit



$y \sim$ Bernoulli

$p(y)=1 = \pi$

$p(y)=0 = 1-\pi$

→ With a binary outcome, lin reg many problems

① $\hat{p}$, the $p(Y=1)$ can be >1 <0

reinterpret "Non-sense predictions"

② Heteroskedasticity: $V(y_{bern}) = \pi(1-\pi)$

FWLS so $V(\hat{y}) = V(\hat{p}) = \hat{p}(1-\hat{p})$

③ Non-normality: $\hat{\varepsilon} = y - Xb = \begin{cases} 1-Xb \\ 0-Xb \end{cases}$

cut so bounded & related in particular way to Xb ⇒ non-normal

④ Nonlinearity: Substance strongly suggests

$\lim_{X\beta\to\infty} p(Y=1) = 1$ & $\lim_{X\beta\to\infty} = 0$, & smoothly

⇒ want $p =$ S-shaped $f(X\beta)$, not linear, smoothly bounded 0,1:

$$\boxed{\text{Logit: } \frac{e^{X\beta}}{1+e^{X\beta}} = p} \qquad \text{Probit: } \boxed{\Phi(X\beta) = p}$$

Many other options have been suggested, e.g.:

Weibull: $P(Y=1) = e[-e^{X\beta}]$ relaxes symmetry

compl. log-log: $P = 1 - e[e^{-X\beta}]$ also relaxes symm.

MaxLike Estimation Logit/Probit:

$$f(y \mid \pi) = \pi^{y}(1-\pi)^{1-y}$$

$$\mathcal{L} : \; f(\underset{\sim}{y} \mid \pi = F(x\beta)) = \prod_{i=1}^{n} \pi_i^{y_i}(1-\pi_i)^{1-y_i}$$

independence

$$\ln \mathcal{L} = \sum_{i=1}^{n} \left\{ y_i \ln F(\underset{\sim}{x}_i \underset{\sim}{\beta}) + (1-y_i) \ln\left[1-F(\underset{\sim}{x}_i \underset{\sim}{\beta})\right] \right\}$$

$$\frac{\partial \ln \mathcal{L}}{\partial \underset{\sim}{\beta}} = 0 \Rightarrow \sum_{i=1}^{n} \left\{ y_i \frac{F'(x\beta)}{F(\underset{\sim}{x}_i \underset{\sim}{\beta})} x + (1-y_i) \frac{-F'()}{F()} x \right\}$$

$$= \sum_{i=1}^{n} \left\{ y_i \frac{f(x\beta)}{F(x\beta)} - (1-y_i) \frac{f()}{F()} \right\} x$$

In more of its detailed glory…

B. Maximum Likelihood Estimation of Binary Dependent Variable Models:

1. Start with first principles (assumptions) just like in linear models

a. $y_i \sim$ independently, Bernoulli with probability p_i

i.e.: $y_i = 1$ with probability p_i

$y_i = 0$ with probability $1 - p_i$

b. $p_i = g(X_i, \beta)$ (i.e., the probability that y_i equals 1 is some function of X_i & the parameters, $\underset{\sim}{\beta}$)

(i) for logit:

$$p_i = [1 + e^{-X_i\beta}]^{-1} = \frac{e^{X_i\beta}}{1 + e^{X_i\beta}}$$

(ii) for probit:

$$p_i = \int_{-\infty}^{X_i\beta} f_N(0,1)\, dy_i^* = \Phi(X_i\underset{\sim}{\beta})$$

(the cumulative ~~standard~~ normal, variance 1, evaluated at $X_i\beta$)

N.b. Some comments about logit & probit:

① $X_i\beta = \infty \Rightarrow p_i = 0$

② $X_i\beta = \infty \Rightarrow p_i = 1$

③ $X_i\beta = 0 \Rightarrow p_i = .5$

④ both sigmoid shaped

⑤ logit approaches 1 & 0 more slowly as $X_i\beta \to \pm\infty$

C. X_i of Rank k (no perfect linear dependence among X's)

N.B. -- there is no separate variance parameter b/c for Bernoulli, it's just $V(y_i) = p \cdot (1-p)$

2. Then write down the probability function for each obs. i:

$$f(y_i \mid X_i, \beta) = p_i^{y_i} (1-p_i)^{1-y_i} \Rightarrow \text{probability } y_i = 1$$

3. Since each y_i is independent (given the X's) the joint p.f. is

$$L \equiv \text{likelihood} = \prod_{i=1}^{n} p_i^{y_i} (1-p_i)^{1-y_i}$$

4. Maximum Likelihood Proceeds by Maximizing this Likelihood over the parameters (β) given the data (X & y) -- using logs since this is easier:

$$LL \equiv \ln(L) = \sum_{i=1}^{n} \{ y_i \ln p_i + (1-y_i) \ln(1-p_i) \}$$

$$\max_{\beta} LL \Rightarrow \frac{\partial LL}{\partial \beta} = 0 \Rightarrow \sum_{i=1}^{n} \left\{ y_i \cdot \frac{1}{p_i} \cdot \frac{\partial p_i}{\partial \beta} - (1-y_i) \cdot \frac{1}{1-p_i} \cdot \frac{\partial p_i}{\partial \beta} \right\} = 0$$

$$\Rightarrow 0 = \sum_{i=1}^{n} \left\{ y_i \cdot (1+e^{-x_i\beta}) \cdot \frac{\partial [1+e^{-x_i\beta}]^{-1}}{\partial \beta} + (y_i - 1) \left(\frac{1+e^{-x_i\beta}}{e^{-x_i\beta}} \right) \frac{\partial [1+e^{-x_i\beta}]^{-1}}{\partial \beta} \right\} = 0$$

$$\Rightarrow 0 = \sum_{i=1}^{n} \left\{ y_i \cdot (1+e^{-x_i\beta}) \cdot \frac{\partial [1+e^{-x_i\beta}]^{-1}}{\partial \beta} + (y_i - 1)\left(\frac{1+e^{-x_i\beta}}{e^{-x_i\beta}}\right) \frac{\partial [1+e^{-x_i\beta}]^{-1}}{\partial \beta} \right\} = 0$$

Aside: $$\frac{\partial [1+e^{-x_i\beta}]^{-1}}{\partial \beta} = (-1)[1+e^{-x_i\beta}]^{-2} e^{-x_i\beta}(-x_i) = \frac{e^{-x_i\beta}}{(1+e^{-x_i\beta})^2} \cdot x_i$$

$$\Rightarrow 0 = \sum_{i=1}^{n} \left\{ y_i (1+e^{-x_i\beta}) \left(\frac{e^{-x_i\beta}}{(1+e^{-x_i\beta})^2}\right) x_i + (y_i - 1)\left(\frac{1+e^{-x_i\beta}}{e^{-x_i\beta}}\right)\left(\frac{e^{-x_i\beta}}{(1+e^{-x_i\beta})^2}\right) x_i \right\}$$

$$0 = \sum_{i=1}^{n} \left\{ y_i \left(\frac{e^{-x_i\beta}}{1+e^{-x_i\beta}}\right) x_i + (y_i - 1)\left(\frac{1}{1+e^{-x_i\beta}}\right) x_i \right\}$$

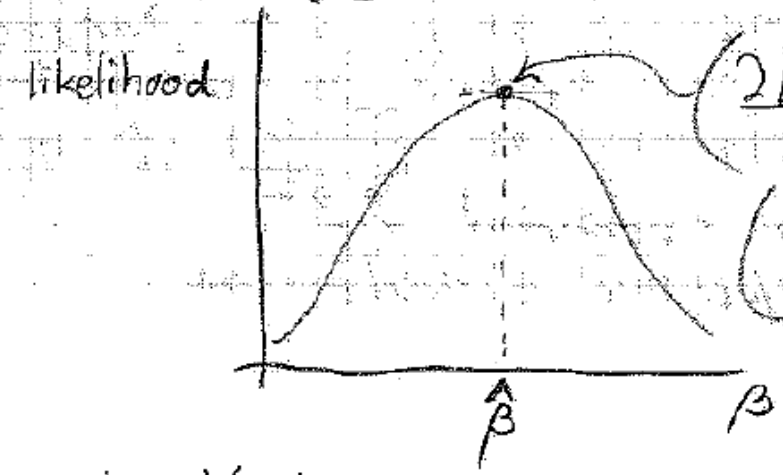
$$= \sum_{i=1}^{n} \left\{ y_i x_i \left(\frac{1+e^{-x_i\beta}}{1+e^{-x_i\beta}}\right) - x_i \left(\frac{1}{1+e^{-x_i\beta}}\right) \right\} = \sum \left\{ y_i x_i - x_i \cdot (1+e^{-x_i\beta})^{-1} \right\} = 0$$

these are the "normal" equations for logit $\Rightarrow \sum_{i=1}^{n} \left\{ y_i - (1+e^{-x_i\beta})^{-1} \right\} x_i = 0$

for probit they are : $$\sum_{i=1}^{n} \left\{ y_i - \Phi(x_i\beta) \right\} \cdot \frac{\phi(x_i\beta)}{\Phi(x_i\beta)(1-\Phi(x_i\beta))} \cdot x_i = 0$$

Why are variance-covariance of MLE's "Minus the Inverse of the Hessian"?

For a single X variable (& one associated β) it looks like this



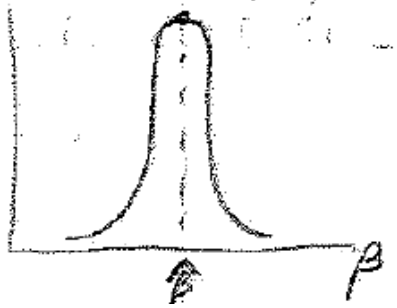
$\left(\frac{\partial \text{Likelihood}}{\partial \beta} \text{ at } \hat{\beta}\right) = 0$ (first-order condition for a maximum)

$\left(\frac{\partial^2 \text{likelihood}}{\partial \beta^2} \text{ at } \hat{\beta}\right) < 0$ (second order condition)

→ refers to fact that likelihood function bends away from the point $\left(\frac{\partial L}{\partial \beta} = 0\right)$ in a downward direction.

i. Estimating the Variance:

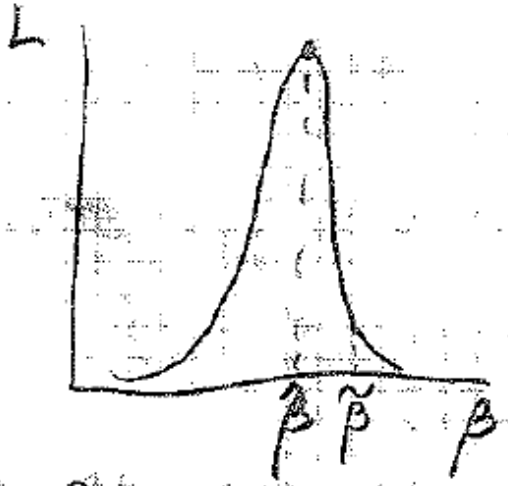
- Notice that



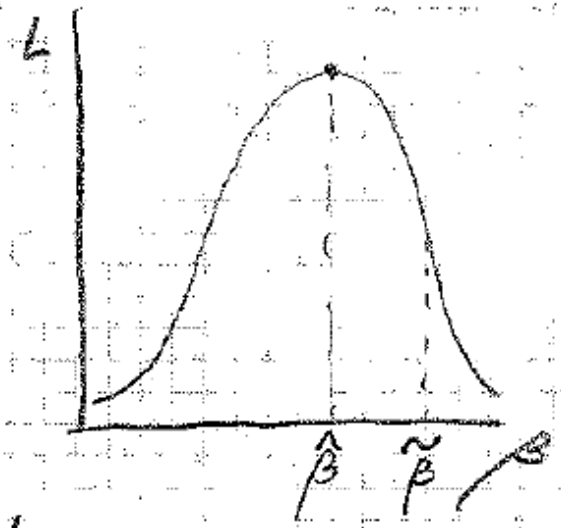
has likelihood changing more for movements of $\hat{\beta}$ from the optimum than the "flatter" previous likelihood function. In this sense, comparing any $\hat{\beta}_0$ with some other

estimate of $\tilde{\beta}$ leads to greater relative likelihood of $\hat{\beta}$ to $\tilde{\beta}$ in the former case than the latter:

case a)



case b)



- for $\hat{\beta}$ & $\tilde{\beta}$ equally distant in the two cases,

$$\left.\frac{\text{likelihood}(\hat{\beta})}{\text{likelihood}(\tilde{\beta})}\right|_{\text{in case (a)}} > \left.\frac{\text{likelihood}(\hat{\beta})}{\text{likelihood}(\tilde{\beta})}\right|_{\text{in case (b)}}$$

$\Rightarrow$ more <u>certain</u> of $\hat{\beta}$ relative to alternatives $\tilde{\beta}$ in case (a) than in case (b)

- So, the more the likelihood function "curves away" from the optimum $\hat{\beta}$, the more certain is our estimate of $\hat{\beta}$

$\Longrightarrow$ Variance estimates will be based on curvature of likelihood function at $\hat{\beta}$. To be specific:

c) $$V(\hat{\beta}) = -\left[\frac{1}{\left.\frac{\partial^2 \ln(L)}{\partial \beta^2}\right|_{\hat{\beta}}}\right]$$

i.e. negative the inverse of the second derivative of the log-likelihood function, evaluated at $\hat{\beta}$.

- why "negative"? Second derivative is negative because $\ln(L)$ curves down at $\hat{\beta}$ (see above figures). We want positive signs here -- we're concerned about magnitude of curvature, we know it's direction is down but that's unimportant here.
- why "inverse"? b/c absolute value of second derivative is <u>higher</u> in case (a), revealing <u>more</u> curvature $\therefore$ <u>more</u> certainty which means <u>smaller</u> standard errors (certainty & standard errors are <u>inversely</u> related)
- why "second derivative at optimum"? Because that's the mathematical definition of "curvature away from $\hat{\beta}$"

Interpreting Logit / Probit: (Marginal Effects)

logit: $\hat{p} = \dfrac{e^{x\hat{\beta}}}{1+e^{x\hat{\beta}}} = \left[1+e^{-x\hat{\beta}}\right]^{-1}$ (div by $e^{x\hat{\beta}}/e^{x\hat{\beta}}$)

$1-\hat{p} = \dfrac{1}{1+e^{x\hat{\beta}}}$

$$\frac{\partial \hat{p}}{\partial x_2} = +\left[1+e^{-x\hat{\beta}}\right]^{-2} e^{-x\hat{\beta}} \cdot \hat{\beta}_2$$

$$= \frac{1}{1+e^{-x\hat{\beta}}} \cdot \frac{e^{-x\hat{\beta}}}{1+e^{-x\hat{\beta}}} \cdot \hat{\beta}_2$$

(div by $e^{-x\hat{\beta}}/e^{-x\hat{\beta}}$)

$$= \frac{e^{x\hat{\beta}}}{1+e^{x\hat{\beta}}} \cdot \frac{1}{1+e^{x\hat{\beta}}} \cdot \hat{\beta}_2$$

$$= \hat{p} \cdot (1-\hat{p}) \cdot \hat{\beta}_2$$

probit: $\hat{p} = \Phi(x\hat{\beta})$ $\Rightarrow \dfrac{\partial \hat{p}}{\partial x_2} = \phi(x\hat{\beta})\beta_2$

probit or logit:
- Effect $\to 0$ as $x\hat{\beta} \to \infty$ or $-\infty$
- Effect $\to$ max as $x\hat{\beta} \to 0$

- Effects vary with $\hat{p}$ & so with all X

$\hat{p}_1 - \hat{p}_0$ for given x_0 & x_1 (First Diff's)

Alternatively, the first-difference approach is just: $\dfrac{\Delta \hat{p}}{\Delta \mathbf{x}} = \hat{p}\big|_{\mathbf{x}^1} - \hat{p}\big|_{\mathbf{x}^0}$

Standard Error of Effects:

Delta Method: $V(f(\hat{\underset{\sim}{\beta}})) \approx \left[f'(\hat{\underset{\sim}{\beta}})\right]' V(\hat{\underset{\sim}{\beta}}) \left[f'(\hat{\underset{\sim}{\beta}})\right]$

e.g. probit: from above $\left[\nabla_{\underset{\sim}{\beta}} (\phi(x\hat{\beta})\underset{\sim}{\beta}_2)\right]' V(\hat{\underset{\sim}{\beta}}) \left[\nabla_{\underset{\sim}{\beta}} \phi(\,)\underset{\sim}{\beta}_2\right]$

....

Or simulate: gking.harvard.edu/docs/clarify.html

Or, in stata:

```
. help mfx
. help dprobit
. help inteff   [install if necessary]
. help predictnl
```

Further MLE for QualDep Examples

Ordered Categorical Variables

1. Suppose our dependent variable takes several possible values which can be rank-ordered but which cannot be spaced in any meaningful way. I.e. the substantive interval between values of Y is not meaningfully captured by the numerical distance between them. Any ranking is an example. A commonly occuring ranking is a Likert scale:

   Strongly Disagree - Disagree - Somewhat Disagree

   Neither Agree nor Dis - Somewh. Agree - Agree - Strongly Agree

   $\Rightarrow 1 - 2 - 3 - 4 - 5 - 6 - 7$

   but who knows if & whether these should be taken as equally distant?

2. Option 1: Treat them as continuous & interval anyway

   $\Rightarrow$ We don't know how observed $Y \in \{1,2,3,4,5,6,7\}$ relates to the underlying Y^* which properly spaces the distances b/w the options, but perhaps it can be ~~thought of~~ as difference b/w Y & Y^* is just measurement error.

   $\Rightarrow$ Meas. Err. in Dep Var $\Rightarrow$ Inefficient

   But: absent more info we have little idea how inefficient. Could be a lot.

3. Option 2: MLE

a) Postulate an underlying Y_i^* reflecting the actual position of obs. i on the interval scale

$$Y_i^* \sim f_{sn}(y_i^* \mid \mu_i)$$

~~standard~~ normal

meaningful spacing

This is the model of the mean. It is this from which our hypotheses derive

b) Theorize that $\boxed{\mu_i = X\beta}$

actual position of y_i^* is a function of X with coeff's β

c) Postulate a set of thresholds

$$\tau_0 < \tau_1 < \tau_2 < \ldots < \tau_{m-1} < \tau_m$$

where m is the number of categories for Y (7 in a Likert scale)

so that Y_i observed as j iff

$$\tau_{j-1} < y_i^* < \tau_j$$

Probability Y_i is in group j

$$\Rightarrow \Pr(Y_{ji} = 1) = \Pr(\tau_{j-1} < Y_i^* < \tau_j)$$

d. $Pr(Y_{ji} = 1) = Pr(\tau_{j-1} < Y_i^* < \tau_j)$

area under cum norm from τ_{j-1} to τ_j $= \int_{\tau_{j-1}}^{\tau_j} f_{stn}(y_i^* | \mu_i)\, dy_i^*$

• we postulated $Y^* \sim f_{stn}(y_i^* | \mu_i)$ so, prob. Y^* between τ_{j-1} & τ_j is integral from τ_{j-1} to τ_j of the p.d.f.

area under cum norm from τ_{j-1} to τ_j $= F_n(\tau_j | \mu_i, 1) - F_n(\tau_{j-1} | \mu_i, 1)$

cumulative normal with mean $\mu_i = X_i\beta$ and variance 1 evaluated at τ_j minus same cum. normal evaluated at τ_{j-1}

• This is the likelihood of one observation being category j

• To get the likelihood function for obs i and all m categories we get:

$$\prod_{j=1}^{m} \left[F_n(\tau_j | X_i\beta, 1) - F_n(\tau_{j-1} | X_i\beta, 1) \right]^{Y_{ji}}$$

($Y_{ji} = 1$ if obs i is in category j) **(=0 else)**

e) The joint likelihood of all the obs (if they're indep.) is product of all i=1 to n of them:

$$\prod_{i=1}^{n} \left\{ \prod_{j=1}^{m} \left[F_n(\tau_j | X_i\beta, 1) - F_n(\tau_{j-1} | X_i\beta, 1) \right]^{Y_{ji}} \right\}$$

C. Counts of Uncorrelated Events (we don't know N &/or there is no upper bound on how often the individual events can occur w/in one obs.)

Examples: # wars in a year
medical consultations for each patient
of anything in some time period or some group

Key here is: Events are occuring at some mean rate (& this is what we'll explain - the rate of event occurence) & we observe counts of "event occurs"

Assumptions:

- Within an observation, the rate of occurence, λ_i, is constant
- No two events can occur in precisely the same instant

$\Rightarrow$ Y_i, the count of events, is Poisson Distributed

$$\Pr(Y_i \mid \lambda_i) = \frac{e^{-\lambda_i}\lambda_i^{y_i}}{y_i!}$$

likelihood Y_i = some y_i given λ_i, the occurence rate

this is what we use to interpret effects of X

$$\boxed{\text{Model: } \lambda_i = e^{X_i\beta}}$$

(λ has to be positive)

$\Rightarrow$ Joint Likelihood: $$\prod_{i=1}^{n} \frac{e^{-\lambda_i}\lambda_i^{y_i}}{y_i!} = \prod_{i=1}^{n} \frac{e^{-e^{X_i\beta}} e^{X_i\beta y_i}}{y_i!}$$

D. Extensions of Count Distributions

1. Grouped, correlated events w/ possibly differing probabilities
   $\Rightarrow$ Beta-Binomial, see King 5.5
2. Counts of uncorrelated events w/ unequal observation intervals
   $\Rightarrow$ Quasi-Poisson w/ variable extrema, see King 5.8
3. Counts of correlated events
   $\Rightarrow$ Generalized Event Count distribution, due to King, see 5.9

Neg. Binomial 3.3, pp 51-52

The Exponential Duration Model

1) We are interested in explaining how long something lasts, or how long until something happens

e.g.
- Gov't Duration
- Peace Duration (War Duration)
- Unemployment spell lengths (Empl. spell lengths)
- Strike Lengths
- Recovery Lengths (Illness lengths)

etc. (Comes from Physics/Chemistry/Bio on survival rates)

2) Notice that it's symmetric: can think of
   a) Hazard Rate: probability of an end to the ongoing process
   b) Survival Function: Probability process lasts some length

- For Constant Hazard Rate, λ, Expected Duration is $1/\lambda$

3) Model the Hazard Rate:

$$\lambda_i = f(\underset{\sim}{X}, \underset{\sim}{\beta})$$

i.e. Prob. of ending in time t given by $f(X)$

Then Survival function is:

$$S(t) = e^{-\lambda t}$$

i.e. Probability of surviving to t or longer is $e^{-\lambda t}$

Option 1: Model it as (possibly correlated) Bernoulli observations

n.b. λ & t must be positive

$$\left.\begin{array}{l} 0 \leftarrow \text{survives} \\ 0 \leftarrow \text{"} \\ 0 \leftarrow \text{"} \\ 0 \leftarrow \text{"} \\ 1 \leftarrow \text{ends} \end{array}\right\} \Rightarrow \lambda_i = \text{prob(Process Ends)} = f(X_i, \beta)$$

$\Rightarrow$ use logit or probit

4) Model the Survival:

p.d.f. $\Pr(Y_i = t) = \lambda e^{-\lambda t} = \lambda e^{-\lambda y_i}$

$$\boxed{\lambda = \text{hazard rate}, \quad E(\text{survival}) = \frac{1}{\lambda} = e^{x\beta}}$$

this is what you interpret

$$\text{p.d.f.} = \prod_{i=1}^{n} \lambda_i e^{-\lambda_i y_i}$$

$$\text{Likelihood} = \underbrace{\prod_{i=1}^{n} e^{-x_i\beta} e^{-(e^{-x_i\beta} y_i)}}$$

This is what you maximize to find $\hat{\beta}_{MLE}$

Multinomial-Logit Model for Polychotomous Data, *m* categories

- Model $p_j/p_m = \dfrac{F(XB_j)}{1-F(XB_j)} = G(XB_j)$ n.b. $\dfrac{p_m}{p_m} = \dfrac{F(XB_m)}{1-F(XB_m)} = 1$

  $\Rightarrow m-1$ models of prob. category j v. base (m).

- (n.b., for 2 categories $\Rightarrow$ odds)

$$\Rightarrow p_m = \frac{1}{1+\sum_{j=1}^{m-1} G(XB_j)} \quad \text{b/c} \quad \sum_{j=1}^{m-1} \frac{p_j}{p_m} = \frac{1-p_m}{p_m} = \frac{1}{p_m} - 1$$

$$\Rightarrow 1 + \sum_{j=1}^{m-1} \frac{p_j}{p_m} = \frac{1}{p_m}$$

$$\Rightarrow p_m = 1 \Big/ 1 + \sum_{j=1}^{m-1} G(XB_j)$$

$$\Rightarrow p_j = \frac{G(XB_j)}{1+\sum_{j=1}^{m-1} G(XB_j)}$$

let $G(XB_j) = e^{XB_j} \Rightarrow p_j = \dfrac{e^{XB_j}}{1+\sum_{j=1}^{m-1} e^{XB_j}}$ & $p_m = \dfrac{1}{1+\sum_{j=1}^{m-1} e^{XB_j}}$

$$p_i = \frac{e^{x\beta_i}}{1+\sum_{j=1}^{m-1} e^{x\beta_j}} \quad ; \quad p_m = \frac{1}{1+\sum_{j=1}^{m-1} e^{x\beta_j}}$$

$\Rightarrow$ interp. as set of logits, cat. by cat. against base cat.

<u>Likelihood Function:</u>

$$y_{ij} = 1 \quad \text{if obs } i \text{ in category } j$$
$$= 0 \quad \text{otherwise}$$

$$L = \prod_{i=1}^{n} p_{i1}^{y_{i1}} p_{i2}^{y_{i2}} \cdots p_{im}^{y_{im}} \quad \left(= p_{i\ell}^{y_{i\ell}} \text{ where } \ell = \text{true category obs } i \right)$$

$$\ln L = \sum_{i=1}^{n} \sum_{j=1}^{m} y_{ij} p_{ij} \qquad (p_{ij} \text{ given above})$$

$$\frac{\partial \ln L}{\partial \beta_k} = \begin{cases} = \sum_{i=1}^{n} \left[\frac{y_{ik}}{p_{ik}} p_{ik}(1-p_{ik}) + \sum_{\substack{j=1 \\ j \neq k}}^{m} \frac{y_{ij}}{p_{ij}} (-p_{ij} p_{ik}) \right] x_i \\ = \sum_{i=1}^{n} \left[y_{ik} - p_{ik} \left(\sum_{j=1}^{m} y_{ij} \right) \right] x_i \\ = \sum_{i=1}^{n} (y_{ik} - p_{ik}) x_i \end{cases}$$

…i.e., our friends the normal equations again.

…interpret as logit, except m-1 binary models, each against base m^{th} case.