Maximum-Likelihood Estimation & QualDep Models

Maximum-Likelihood Estimation

\( y \sim f(X, \theta) \), conditional independence \( \Rightarrow \)

\[
L(y) = \prod_{i=1}^{N} f_i(x, \theta) \Rightarrow \ln(L(y)) \equiv \mathcal{L} = \sum_{i=1}^{n} \ln\left[f_i(x, \theta)\right]
\]

Maximum-likelihood Estimator (MLE):

1. Parameter estimates solve: \( \frac{\partial \mathcal{L}}{\partial \theta} = \nabla_{\theta} \mathcal{L} = 0 \)
2. Estimated variance of Estimated Parameters: \( -\mathcal{H}^{-1} \equiv -\left[\frac{\partial^2 \mathcal{L}}{\partial \theta \partial \theta'}\right]^{-1} \)

Properties of MLE: BANC

1. Asymptotically Normal
2. Consistent
3. Minimum-Variance
4. Invariance
Minimum Variance means Obtains Cramér-Rao Lower Bound

\[
\text{Cramér-Rao Lower Bound} \rightarrow \text{for any "well-behaved" pdf, any unbiased estimator of parameter } \theta, \text{ all } \hat{\theta}\text{'s will have variance at least as large as:}
\]

\[
[I(\theta)]^{-1} = (-E[\frac{\partial^2 \ln L(\theta)}{\partial \theta^2}])^{-1}
\]

\cdot \text{ } L(\theta) \text{ here is the "likelihood function" of the data, i.e., loosely, it is:}
\]

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

\[ X_i \sim f(X_i, \theta) = \theta e^{-\theta x_i} \] 
\text{\quad } \Theta = \text{hazard rate}
\]

\[ \text{\quad } \ln \Theta = \frac{n}{\Theta} \] 
\text{\quad } \frac{1}{\Theta} = \text{mean}
\]

\[ \text{\quad } \frac{1}{\Theta^2} = \text{var} \]

\text{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} \]

\text{thus, } \ln L(\theta) = n \ln \Theta - \Theta \sum X_i \]

\[ \frac{\partial}{\partial \theta} \ln L(\theta) = \frac{\partial^2 \ln L(\theta)}{\partial \theta^2} = \frac{\partial (\frac{n}{\Theta} - \sum X_i)}{\partial \Theta} = -\frac{n}{\Theta^2} \]

\[ \text{So, } \text{the C-R lower bound} = (-E(-\frac{\partial^2}{\partial \theta^2}))^{-1} = [E(\frac{n}{\Theta})]^{-1} = \frac{\Theta^2}{n} \]

\[ \nabla_{\theta} L = \frac{n}{\theta} - \sum x_i = 0 \Rightarrow \hat{\theta}_{ML} = \frac{\sum x_i}{n} = \bar{x} \]
Example 2: MLE for a Poisson Count Model

1. We view outcomes as being produced by some p.f. or p.d.f. \( f(x_i; \theta) \)

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:

\[
f(x_1, x_2, x_3, \ldots, x_n; \theta) = f(x_1; \theta) f(x_2; \theta) f(x_3; \theta) \ldots f(x_n; \theta)
\]

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

That is, given the data, \( X \), define the (joint) conditional distribution of \( \Theta \) given \( X \). Maximize this with respect to \( \Theta \).

Examples:

1. Each \( X_i \sim \text{Poisson}(\theta) \): 
   \[
f(x_i; \theta) = \frac{e^{-\theta} \theta^{x_i}}{x_i!}
\]
   \( (\text{A Poisson Distribution describes a count (RV) of events happening at average rate } \theta \text{ per fixed period}) \)

2. \[
\begin{align*}
&\text{Max } f(x_i; \theta) = \text{Max } \ln f(x_i; \theta) = \text{Max } \sum_{i=1}^{n} \left[ \ln \left( e^{-\theta} \theta^{x_i} \right) - \ln(x_i!) \right] \\
&= \text{Max } \sum_{i=1}^{n} \left( -\theta + x_i \ln \theta - \ln(x_i!) \right) \quad \text{First-Order Condition} \Rightarrow \frac{\partial L(\theta)}{\partial \theta} = 0
\end{align*}
\]

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

   F.O.C.: \[
   \frac{\partial L(\theta)}{\partial \theta} = 0 \quad \Rightarrow \quad -n + \sum_{i=1}^{n} \frac{x_i}{\theta} = 0
   \]

   \[
   \Rightarrow \quad e^\theta = \frac{1}{n} \sum_{i=1}^{n} x_i = \bar{X}
   \]

   \[\text{Second-Order Condition:}\]

   \[
   0 > \frac{\partial^2 L(\theta)}{\partial \theta^2} = \frac{\partial (\bar{X})}{\partial \theta} = \frac{\partial (-n + \theta \sum x_i)}{\partial \theta} = -\theta (\sum x_i)
   \]

   \[\text{greater than}\]

   Thus, MLE for Poisson parameter, \( \theta \), is \( \hat{\theta} = \bar{X} \Rightarrow \frac{\bar{X}}{\theta} < 0 \)
VI. C. 2. MLE Example 2: MLE for Normal

a. \( f_n(x; \theta) = (2\pi\sigma^2)^{-\frac{1}{2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \)

b. \( L(\theta) = \prod_{i=1}^{n} \left(2\pi\sigma^2\right)^{-\frac{1}{2}} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}} \)

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

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

\[ = -\frac{n}{2} \ln(2\pi) - \frac{n}{2} \ln\sigma^2 - \frac{n}{2} \ln\sigma^2 - \frac{1}{2} \sum_{i=1}^{n} (x_i-\mu)^2 (\sigma^{-2})^{-1} \]

C. F.O.C. \( \frac{\partial \ln L(\theta)}{\partial \theta} = 0 \):

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

ii.) \( \frac{\partial \ln L(\theta)}{\partial \sigma^2} = \frac{n}{2\sigma^4} + \frac{1}{2} (\sigma^{-2})^{-2} \sum_{i=1}^{n} (x_i-\mu)^2 = 0 \)

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

\[ \Rightarrow \sum_{i=1}^{n} x_i = n \mu = n \mu^* = 0 \]

\[ \Rightarrow \mu^* = \frac{1}{n} \sum_{i=1}^{n} x_i = \bar{x} \]

- Substitute that into (ii) if you'll find \( \sigma^2 = \frac{1}{n} \sum_{i=1}^{n} (x_i-\bar{x})^2 \)
Of course, this all more interesting & useful when have a model of parameters...

\[ y_i \sim N(\mu_i, \sigma^2) \quad ; \quad \mu_i = \beta x_i \]

\[
\begin{align*}
    f(y_i, \mu_i, \sigma^2) &= (2\pi \sigma^2)^{-1/2} e^{-(y_i - \beta x_i)^2 / (2\sigma^2)} \\
    L(\beta, \sigma^2 | x) &= \prod_{i=1}^{n} (2\pi \sigma^2)^{-1/2} e^{-(y_i - \beta x_i)^2 / (2\sigma^2)} \\
    \ln L(\beta, \sigma^2 | x) &= \sum_{i=1}^{n} \left[ -\frac{1}{2} \ln (2\pi \sigma^2) - \frac{1}{2} \frac{(y_i - \beta x_i)^2}{\sigma^2} \right] \\
    &= \sum_{i=1}^{n} \left[ -\frac{1}{2} \ln 2 - \frac{1}{2} \ln \pi - \frac{1}{2} \ln (\sigma^2) - \frac{1}{2} (y_i - \beta x_i)^2 / \sigma^2 \right] \\
    &= -\frac{n}{2} \ln 2 - \frac{n}{2} \ln \pi - \frac{n}{2} \ln (\sigma^2) - \frac{1}{2} \sum_{i=1}^{n} (y_i - \beta x_i)^2 (\sigma^{-2})^{-1}
\end{align*}
\]
\[
\text{F.O.C. } \frac{\partial \ln L}{\partial \beta} = - (\sigma^2)^{-1} \sum_{i=1}^{n} (y_i - \beta_x i) \cdot (-x_i) = 0
\]

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

\Rightarrow \sum_{i=1}^{n} (y_i - \beta x_i) \cdot x_i = 0

\Rightarrow \sum \beta x_i^2 = \sum y_i x_i

\frac{\beta}{\sum x_i^2} = \frac{\sum x_i y_i}{\sum x_i^2}

\text{Holy Wa!}

\text{There it is again!}
\[(i) \quad \frac{\partial \ell}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{2} (\hat{\sigma}^2)^{-2} \sum (y_i - \beta x_i)^2 = 0\]

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

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

\[\Rightarrow \frac{n}{\hat{\sigma}^2} = \sum (y_i - \beta x_i)^2\]

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

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

average squared residual

\[s.o.c. \quad -(\hat{\sigma}^2)^{-2} \sum (y_i - \beta x_i)^2 < 0 \quad \checkmark\]
One more time in more-compact notation?

\[
L(\beta) = \prod_{i=1}^{n} \left( 2\pi \sigma^2 \right)^{-\frac{1}{2}} e^{-\frac{(y_i - x_i\beta)^2}{2\sigma^2}}
\]

Likelihood is interpreted as "the probability of observing the data you actually observed, given the model".

1. Maximum 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.

\[
\log L(\beta) = \sum_{i=1}^{n} \left\{ (x_i^2 + \log \sigma^2) - \frac{(y_i - x_i\beta)^2}{2\sigma^2} \right\}
\]

Maximizing \( \log L(\beta) \):\[\frac{\partial}{\partial \beta} \log L(\beta) = 0 \quad \Rightarrow \quad \sum_{i=1}^{n} (y_i - x_i\beta)x_i = 0 \]

\[\frac{\partial}{\partial \sigma^2} \log L(\beta) = 0 \quad \Rightarrow \quad \sum_{i=1}^{n} (y_i - x_i\beta)^2 = n \sigma^2 \]

5. So, OLS is the MLE of \( \beta \) for the CNLRM.
Many other options have been suggested, e.g.:

- **Logit / Probit**
  - With a binary outcome, lin reg many problems
  - \( \hat{p} \), \( \text{np}(Y=1) \) can be \( >1 \) \( <0 \)
  - “Non-sense predictions”
  - **Heteroskedasticity:** \( V(y) = \pi(1-\pi) \)
  - \( V(y) = V(\hat{p}) = \hat{p}(1-\hat{p}) \)
  - **Non-normality:** \( \hat{e} = y - x_b = \begin{cases} 1-x_b \\ 0-x_b \end{cases} \)
    - so bounded & related in particular way to \( x_b \)
      - \( \Rightarrow \) non-normal
  - **Non-linearity:** Substance strongly suggests
    - \( \lim_{x \beta \to \infty} p(y=1) = 1 \) & \( \lim_{x \beta \to \infty} = 0 \), \( \Rightarrow \) smoothly
    - \( \Rightarrow \) want \( p = \) s-shaped \( f(x_b) \), not linear, smoothly bounded \( 0, 1 \):
      - Logit: \( \frac{e^{x \beta}}{1+e^{x \beta}} = p \)
      - Probit: \( \Phi(x \beta) = p \)

- **Weibull:** \( p(x_b) = e^{x_b} e^{-e^{x_b}} \) reln symmetry
- **Compl. log-neg:** \( p = \frac{-e^{-x_b}}{1-e^{-x_b}} \) reln symmetry
MaxLike Estimation Logit/Probit:

\[ f(y | \eta) = \pi^Y (1-\pi)^{1-Y} \]

\[ L : f(y | \pi = F(x\beta)) = \prod_{i=1}^{n} \pi_i^{Y_i} (1-\pi_i)^{1-Y_i} \text{ independence} \]

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

\[ \frac{\partial \ln L}{\partial \beta} = 0 \Rightarrow \sum_{i=1}^{n} \{Y_i \frac{F'(x_i\beta)}{F(x_i\beta)} x + (1-Y_i) \frac{F'(x_i\beta)}{F(x_i\beta)} x \} \]

\[ = \sum_{i=1}^{n} \{Y_i \frac{F(x_i\beta)}{F(x_i\beta)} - (1-Y_i) \frac{F(x_i\beta)}{F(x_i\beta)}\} x \]
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 $p_i$
      $$y_i = 1 \text{ with probability } p_i$$
      $$y_i = 0 \text{ with probability } 1 - p_i$$
   b. $p_i = g(X_i, \beta)$ (i.e., the probability that $y_i$ equals 1 is some function of $X_i$ and the parameters, $\beta$)

   (i) For logit:
   $$p_i = \frac{1}{1 + e^{-X_i \beta}} = \frac{e^{X_i \beta}}{1 + e^{X_i \beta}}$$

   (ii) For probit:
   $$p_i = \int_{-\infty}^{X_i \beta} f_N(\sigma) \, dy_i = \Phi(X_i \beta)$$
   (the cumulative standard normal, evaluated at $X_i \beta$)

2. $X_i$ of rank $k$ (no perfect linear dependence among $X_i$s)
2. Then write down the probability function for each obs. i:

3. Since each \( Xi \) is independent (given the \( X_i \)'s) the joint pdf is:

4. Maximum likelihood proceeds by maximizing this likelihood function over the parameters (p) given the data (X) using logs since this is easier:

\[
f(y_1, X, \beta) = p \cdot (1-p)^{-y_1} = \sum_{x=y_1}^{\infty} \binom{x}{y_1} \cdot p^{y_1} \cdot (1-p)^{x-y_1}
\]

Therefore:

\[
L = \text{Likelihood} = \prod_{i=1}^{n} \binom{x_i}{y_i} \cdot p^{y_i} \cdot (1-p)^{x_i-y_i}
\]

5. N.B. there is no extra variance parameter for Bernoulli, it's just

\[
V(r) = p \cdot (1-p)
\]
\[
O = \sum_{i=1}^{n} \left\{ Y_i \cdot (1+e^{-X_i \beta}) \cdot \frac{\partial \left[ 1+e^{-X_i \beta} \right]}{\partial \beta} + (Y_i-1) \cdot \frac{(1+e^{-X_i \beta}) \cdot \partial \left[ 1+e^{-X_i \beta} \right]}{\partial \beta} \right\} = 0
\]

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

\[
O = \sum_{i=1}^{n} \left\{ Y_i \cdot \left( 1+e^{-X_i \beta} \right) \cdot \left( \frac{e^{-X_i \beta}}{(1+e^{-X_i \beta})^2} \right) \cdot X_i + (Y_i-1) \cdot \left( \frac{1+e^{-X_i \beta}}{e^{-X_i \beta}} \right) \cdot \left( \frac{e^{-X_i \beta}}{(1+e^{-X_i \beta})^2} \right) \cdot X_i \right\}
\]

\[
O = \sum_{i=1}^{n} \left\{ Y_i \cdot \left( \frac{e^{-X_i \beta}}{1+e^{-X_i \beta}} \right) \cdot X_i + (Y_i-1) \cdot \left( \frac{1}{1+e^{-X_i \beta}} \right) \cdot X_i \right\}
\]

\[
= \sum_{i=1}^{n} \left\{ Y_i \cdot X_i \cdot \left( \frac{1+e^{-X_i \beta}}{1+e^{-X_i \beta}} \right) - X_i \cdot \left( \frac{1}{1+e^{-X_i \beta}} \right) \right\} = \sum_{i=1}^{n} \left\{ Y_i \cdot X_i - X_i \cdot (1+e^{-X_i \beta})^{-1} \right\} = 0
\]

These are the normal equations for probit logistic regression:
\[
\Rightarrow \sum_{i=1}^{n} \left\{ Y_i - (1+e^{-X_i \beta})^{-1} \right\} \cdot X_i = 0
\]

For probit they are:
\[
\sum_{i=1}^{n} \left\{ Y_i - \Phi(X_i \beta) \right\} \cdot \frac{\phi(X_i \beta)}{\Phi(X_i \beta) \cdot (1-\Phi(X_i \beta))} \cdot X_i = 0
\]
Why are variance-covariance of MLE’s “Minus the Inverse of the Hessian”?

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

\[
\left( \frac{\partial \text{likelihood}}{\partial \beta} \right) = 0 \quad \text{(first-order condition for a maximum)}
\]

\[
\left( \frac{\partial^2 \text{likelihood at } \beta}{\partial \beta^2} \right) < 0 \quad \text{(second-order condition)}
\]

Estimating the Variance:

Notice that L has likelihood changing more for movements of β from the optimum than the “flatter” previous likelihood function. In this sense comparing any β with some other
estimate of $\hat{\beta}$ leads to greater relative likelihood of $\hat{\beta}$ to $\hat{\beta}'$ in the former case than the latter.

For $\hat{\beta}$ and $\hat{\beta}'$ equally distant in the two cases,

\[
\frac{\text{likelihood} (\hat{\beta})}{\text{likelihood} (\hat{\beta}')} \quad \text{in case (a)} \quad \Rightarrow \text{more certain of } \hat{\beta} \text{ relative to alternatives } \hat{\beta}' \text{ in case (a) than in case (b)}
\]
So, the more the likelihood function “curves away” from the optimum \( \hat{\beta} \),
the more certain is our estimate of \( \hat{\beta} \).

\[ \Rightarrow \text{Variance estimates will be based on curvature of likelihood function at } \hat{\beta}. \]

To be specific:

\[ \text{var}(\hat{\beta}) = \left[ -\frac{\partial^2 \log L}{\partial \beta \partial \beta^T} \right] \]

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 \( L(L) \) 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 higher in case \( \hat{\beta} \), revealing more curvature: more certainty which means smaller standard errors (certainly standard errors are inversely related).

Why “second derivative at optimum”? Because that’s the mathematical definition of “curvature away from \( \hat{\beta} \)”.
Interpreting Logit/Probit:

\[
\hat{p} = \frac{e^{\hat{\beta} x}}{1 + e^{\hat{\beta} x}} = \left[1 + e^{\hat{\beta} x}\right]^{-1} \\
1 - \hat{p} = \frac{1}{1 + e^{\hat{\beta} x}}
\]

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

\[
= \frac{1}{1 + e^{\hat{\beta} x_2}} \cdot e^{-\hat{\beta} x_2} \cdot \hat{\beta}_2
\]

\[
= \frac{e^{\hat{\beta} x_2}}{1 + e^{\hat{\beta} x_2}} \cdot \frac{1}{1 + e^{\hat{\beta} x_2}} \cdot \hat{\beta}_2
\]

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

Probit: \( \hat{p} = \Phi(x_2 \hat{\beta}) \) \Rightarrow \( \frac{\partial \hat{p}}{\partial x_2} = \phi(x_2 \hat{\beta}) \beta_2 \)

\( \text{Probit:} \) Effect \( \rightarrow 0 \) as \( x_2 \hat{\beta} \rightarrow \infty \) or \( -\infty \)  \\
\( \text{Logit:} \) Effect \( \rightarrow \infty \) as \( x_2 \hat{\beta} \rightarrow 0 \)

\( \hat{p}_1 - \hat{p}_0 \) for given \( x_2 \) and \( x_1 \)

(First Diffs)

Alternatively, the first-difference approach is just:

\[
\frac{\Delta \hat{p}}{\Delta x} = \hat{p}\bigg|_{x^1} - \hat{p}\bigg|_{x^0}
\]
Standard Error of Effects:

\[ \Delta \text{ Delta Method: } \nabla f(\hat{\beta}) \approx \left[ f'(\hat{\beta}) \right] \nabla \hat{\beta} \left[ f'(\hat{\beta}) \right] \]

e.g. probit: \[ \nabla \hat{\beta} (\phi(x\hat{\beta})\beta) \nabla \hat{\beta} \left[ \nabla \phi(x\hat{\beta})\beta \right] \]

... or simulate: ghking.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 on several possible values which can be rank-ordered, but which cannot be spaced in any meaningful way, i.e., the interval between values of $Y$ is not meaningfully captured by the numerical distance between them. Any ranking is an example. A commonly occurring ranking is a Likert Scale:
   - Strongly Disagree - Disagree - Somewhat Disagree
   - Neither Agree nor Dis - Somewhat Agree - Agree - Strongly Agree
   \[ 1 - 2 - 3 - 4 - 5 - 6 - 7 \]

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

2. Option 1: Treat them as continuous & interval anyway
   \[ \Rightarrow \text{We don't know how observed } Y \in \{1, 2, 3, 4, 5, 6, 7\} \text{ relates to the underlying } Y^* \text{ which properly spaces the distances } \Delta \text{ the options, but perhaps it can be thought of as difference } \Delta \text{ & } Y^* \text{ is measurement error}. \]
   \[ \Rightarrow \text{Meas. Err. in Dep Var } \Rightarrow \text{Inefficient} \]

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:

$Y_i^* = f(Y_i^* | \mu_i)$

b) Theorize that $\mu_i = X\beta$ - actual position of $Y_i^*$ is a function of $X$ with coeffs $\beta$

c) Postulate a set of thresholds

$2.0 < \tau_1 < \tau_2 < \ldots < \tau_{m-1} < \tau_m$

so that $Y_i$ observed as $j$ iff

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

$\Rightarrow \Pr(Y_{ji} = 1) = \Pr(\tau_{j-1} < Y_i^* < \tau_j)$
The joint likelihood of all the obs. (category i) is

\[ L = \prod_{i=1}^{n} \left[ F(x_i | \mu, \sigma^2) - F(x_i | \mu, \sigma^2) \right] \]

This is the likelihood of one observation evaluated.

To get the likelihood function for obs. and all categories we get:

\[ L = \prod_{i=1}^{n} \left[ F(x_i | \mu, \sigma^2) - F(x_i | \mu, \sigma^2) \right] \]

and under a cumulative normal with mean \( \mu \) and variance \( \sigma^2 \), the distribution is

\[ \int_{-\infty}^{x_i} f(y) \, dy \]

so probability between \( \mu \) and \( x_i \) is

\[ \text{prob. between } \mu \text{ and } x_i \]
C. Counts of Uncorrelated Events (we don't know N or there is no upper bound on how often the events can occur within one obs.)

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

Key here is: Events are occurring at some mean rate (that is what we'll explain - the rate of event occurrence)
if we observe counts of "event occurs"

Assumptions:
- Within an observation, the rate of occurrence, \( \lambda_i \), is constant
- No two events can occur in precisely the same instant

\[ \Rightarrow \quad \lambda_i \text{, the count of events, is Poisson Distributed} \]

\[ Pr(\gamma_i | \lambda_i) = \frac{\lambda_i^{\gamma_i} e^{-\lambda_i}}{\gamma_i!} \]

likelihood: \( \gamma_i \) = some \( \gamma_i \) given \( \lambda_i \); the occurrence rate

Model: \( \lambda_i = e^{x \beta} \) (\( \beta \) has to be positive)

\[ \Rightarrow \text{Joint Likelihood: } \prod_{i=1}^{n} \frac{e^{-\lambda_i} \lambda_i^{\gamma_i}}{\gamma_i!} = \prod_{i=1}^{n} \frac{e^{-e^{x \beta}} e^{x \beta \gamma_i}}{\gamma_i!} \]

D. Extensions of Count Distributions

1. Grouped, correlated events with possibly differing probabilities
   \[ \Rightarrow \text{Beta-Binomial, see King 5.5} \]
2. Counts of uncorrelated events with unequal observation intervals
   \[ \Rightarrow \text{Quasi-Poisson with variable extrema, see King 5.8} \]
3. Counts of correlated events
   \[ \Rightarrow \text{Generalized Event Count distribution, due to King, see 5.9} \]
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 $\lambda$, Expected Duration is $\frac{1}{\lambda}$

3) Model the Hazard Rate:

   $\lambda_i = f(X_i, \beta)$

   Then Survival function is:

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

   i.e., Probability of surviving to $t$ or longer is $e^{-\lambda t}$.
4. Model II: Event as (possibly correlated) binary outcome.

\[ P(Y_i = 1 | X_i) \]

\[ = \frac{e^{x_i \beta}}{1 + e^{x_i \beta}} \]

\[ \text{Logit} = \log \left( \frac{P(Y_i = 1)}{1 - P(Y_i = 1)} \right) = x_i \beta \]

5. Survival:

\[ S(t) = 1 - F(t) = 1 - (1 - e^{-x_i \beta})^t \]

\[ \text{Hazard rate: } \frac{\text{number of events}}{\text{total person-time}} \]

\[ \Rightarrow \text{Hazard} = -\lambda e^{-x_i \beta} \]

\[ \text{Expected event time} = \frac{1}{\lambda} \]

\[ \text{Survival} = e^{-\lambda t} \]

6. Possible correlation between repeated events:

\[ \text{Logistic regression with correlated outcomes} \]

7. Example:

\[ i = 1, 2, \ldots, n \]

\[ k = 1, 2, \ldots, K_i \]

\[ y_{ik} \sim \text{Bernoulli}(p_{ik}) \]

\[ p_{ik} = \frac{e^{x_{ik} \beta}}{1 + e^{x_{ik} \beta}} \]

\[ \text{Logit} = \log \left( \frac{p_{ik}}{1 - p_{ik}} \right) = x_{ik} \beta \]

\[ \text{Expected event time} = \frac{1}{\lambda} \]

\[ \text{Survival} = e^{-\lambda t} \]
Multinomial-Logit Model for Polychotomous Data, \( m \) categories

- Model \( \frac{p_j}{p_m} = \frac{F(XB_j)}{1 - F(XB_m)} \)
  \( m-1 \) models of prob. category \( j \) v. base \( m \).

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

\[ p_m = \frac{1}{1 + \sum_{j=1}^{m-1} G(XB_j)} \]

\[ \Rightarrow p_m = \frac{1}{1 + \frac{1}{p_m}} = \frac{p_m}{1} \]

\[ \Rightarrow p_m = \frac{1}{1 + \sum_{j=1}^{m-1} e^{XB_j}} \]

let \( G(XB_j) = e^{XB_j} \)

\[ p_j = \frac{e^{XB_j}}{1 + \sum_{j=1}^{m-1} e^{XB_j}} \]

\[ p_m = \frac{1}{1 + \sum_{j=1}^{m-1} e^{XB_j}} \]
\[ P_i = \frac{e^{X_i \beta}}{1 + e^{X_i \beta}} \quad \text{;} \quad P_m = \frac{1}{1 + e^{X_m \beta}} \]

Interpret as set of logits, cat by cat.

Likelihood Function:

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

\[ L = \prod_{i=1}^{n} P_{i1}^{y_{i1}} P_{i2}^{y_{i2}} \cdots P_{im}^{y_{im}} \]
\[ = P_{i2}^{y_{i2}} \quad \text{(where } \frac{y_{i2}}{\text{true category}} \text{)} \]

\[ \ln L = \sum_{i=1}^{n} \sum_{j=1}^{m} y_{ij} \ln P_{ij} \quad (p_{ij} \text{ given above}) \]

\[ \frac{\partial \ln L}{\partial \beta_k} = \left[ \sum_{i=1}^{n} \left( \frac{y_{ik} - P_{ik}(\frac{y_{ik}}{\text{true category}})}{P_{ik}(1-P_{ik})} \right) X_i \right] \]

\[ = \sum_{i=1}^{n} (y_{ik} - P_{ik}) X_i \]

\[ \text{...i.e., our friends the normal equations again.} \]

\[ \text{...interpret as logit, except } m-1 \text{ binary models, each against base } m^{\text{th}} \text{ case.} \]