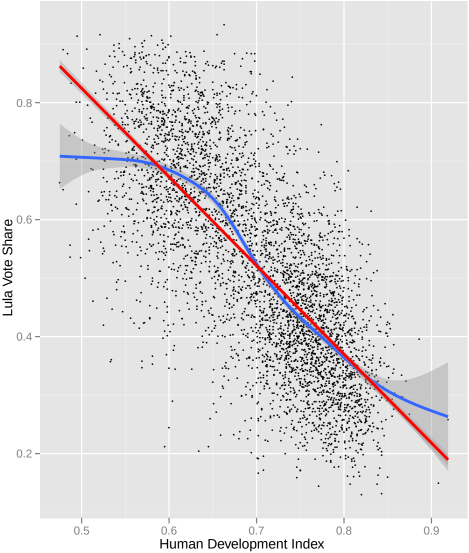
Maximum Likelihood

March 16, 2011

Two Models



Lula Vote Share
0.8
0.6
0.4
0.2
0.5
0.6
0.7
0.8
0.9
Human Development Index

Inference

- We want to estimate parametric models of the form:

$$y_i \sim f(\theta)$$

Inference

- We want to estimate parametric models of the form:

$$y_i \sim f(\theta)$$

- The $f(\cdot)$ is the part we assume (the model), θ is the part that we want to estimate (the parameters), and we observe y_i (the data).

Inference

- We want to estimate parametric models of the form:

$$y_i \sim f(\theta)$$

- The $f(\cdot)$ is the part we assume (the model), θ is the part that we want to estimate (the parameters), and we observe y_i (the data).
- What social scientists wish we could do:

Inference

- We want to estimate parametric models of the form:

$$y_i \sim f(\theta)$$

- The $f(\cdot)$ is the part we assume (the model), θ is the part that we want to estimate (the parameters), and we observe y_i (the data).
- What social scientists wish we could do:
  - Estimate $\mathrm{P}(\text{unknown}|\text{known})$ or $\mathrm{P}(f(\theta)|(y_i))$

Inference

- We want to estimate parametric models of the form:

$$y_i \sim f(\theta)$$

- The $f(\cdot)$ is the part we assume (the model), θ is the part that we want to estimate (the parameters), and we observe y_i (the data).
- What social scientists wish we could do:
  - Estimate $\mathrm{P}(\mathrm{unknown}|\mathrm{known})$ or $\mathrm{P}(f(\theta)|(y_i))$
  - This is the problem of inverse probability: what is the probability of the model given the data?

Inference

- We want to estimate parametric models of the form:

$$y_i \sim f(\theta)$$

- The $f(\cdot)$ is the part we assume (the model), θ is the part that we want to estimate (the parameters), and we observe y_i (the data).
- What social scientists wish we could do:
  - Estimate $\mathrm{P}(\text{unknown}|\text{known})$ or $\mathrm{P}(f(\theta)|(y_i))$
  - This is the problem of inverse probability: what is the probability of the model given the data?
- What about the reverse?

Inference

- We want to estimate parametric models of the form:

$$y_i \sim f(\theta)$$

- The $f(\cdot)$ is the part we assume (the model), θ is the part that we want to estimate (the parameters), and we observe y_i (the data).
- What social scientists wish we could do:
  - Estimate $\mathrm{P}(\text{unknown}|\text{known})$ or $\mathrm{P}(f(\theta)|(y_i))$
  - This is the problem of inverse probability: what is the probability of the model given the data?
- What about the reverse?
  - Estimate $\mathrm{P}(\theta|\{f(\cdot)y_i\})$ or, more simply, $\mathrm{P}(\theta|y_i)$

Inference

- We want to estimate parametric models of the form:

$$y_i \sim f(\theta)$$

- The $f(\cdot)$ is the part we assume (the model), θ is the part that we want to estimate (the parameters), and we observe y_i (the data).
- What social scientists wish we could do:
    - Estimate $\mathrm{P}(\text{unknown}|\text{known})$ or $\mathrm{P}(f(\theta)|(y_i))$
    - This is the problem of inverse probability: what is the probability of the model given the data?
- What about the reverse?
    - Estimate $\mathrm{P}(\theta|\{f(\cdot)y_i\})$ or, more simply, $\mathrm{P}(\theta|y_i)$
    - What is the probability of θ being a hypothesized value, given the assumed model and the observed data?

Can we estimate $\mathrm{P}(\theta|y_i)$?

Apply Bayes' theorem:

$$\mathrm{P}(\theta|y_i) = \frac{\mathrm{P}(\theta, y_i)}{P(y_i)} \tag{1}$$

$$= \frac{\mathrm{P}(\theta)\mathrm{P}(y|\theta)}{\mathrm{P}(y_i)} \tag{2}$$

We often focus on numerator, resulting in:

$$\mathrm{P}(\theta|y_i) \propto \mathrm{P}(\theta) \cdot \mathrm{P}(y_i|\theta)$$

- What is $\mathrm{P}(\theta)$? *The Prior.*

Can we estimate $\mathrm{P}(\theta|y_i)$?

Apply Bayes' theorem:

$$\begin{aligned} \mathrm{P}(\theta|y_i) &= \frac{\mathrm{P}(\theta, y_i)}{P(y_i)} && (1) \\ &= \frac{\mathrm{P}(\theta)\mathrm{P}(y|\theta)}{\mathrm{P}(y_i)} && (2) \end{aligned}$$

We often focus on numerator, resulting in:

$$\mathrm{P}(\theta|y_i) \propto \mathrm{P}(\theta) \cdot \mathrm{P}(y_i|\theta)$$

- What is $\mathrm{P}(\theta)$? *The Prior.*
- What is $\mathrm{P}(\theta|y_i)$? *The Posterior.*

Can we estimate $\mathrm{P}(\theta|y_i)$?

Apply Bayes' theorem:

$$
\begin{aligned}
\mathrm{P}(\theta|y_i) &= \frac{\mathrm{P}(\theta, y_i)}{P(y_i)} && (1) \\
&= \frac{\mathrm{P}(\theta)\mathrm{P}(y|\theta)}{\mathrm{P}(y_i)} && (2)
\end{aligned}
$$

We often focus on numerator, resulting in:

$$\mathrm{P}(\theta|y_i) \propto \mathrm{P}(\theta) \cdot \mathrm{P}(y_i|\theta)$$

- What is $\mathrm{P}(\theta)$? *The Prior*.
- What is $\mathrm{P}(\theta|y_i)$? *The Posterior*.
- What is $\mathrm{P}(y_i|\theta)$? *the likelihood*.

Can we estimate $\mathrm{P}(\theta|y_i)$?

Apply Bayes' theorem:

$$\mathrm{P}(\theta|y_i) = \frac{\mathrm{P}(\theta, y_i)}{P(y_i)} \tag{1}$$

$$= \frac{\mathrm{P}(\theta)\mathrm{P}(y|\theta)}{\mathrm{P}(y_i)} \tag{2}$$

We often focus on numerator, resulting in:

$$\mathrm{P}(\theta|y_i) \propto \mathrm{P}(\theta) \cdot \mathrm{P}(y_i|\theta)$$

- What is $\mathrm{P}(\theta)$? *The Prior*.
- What is $\mathrm{P}(\theta|y_i)$? *The Posterior*.
- What is $\mathrm{P}(y_i|\theta)$? *the likelihood*.

Can we estimate $\mathrm{P}(\theta|y_i)$?

Apply Bayes' theorem:

$$\mathrm{P}(\theta|y_i) = \frac{\mathrm{P}(\theta, y_i)}{P(y_i)} \tag{1}$$

$$= \frac{\mathrm{P}(\theta)\mathrm{P}(y|\theta)}{\mathrm{P}(y_i)} \tag{2}$$

We often focus on numerator, resulting in:

$$\mathrm{P}(\theta|y_i) \propto \mathrm{P}(\theta) \cdot \mathrm{P}(y_i|\theta)$$

- What is $\mathrm{P}(\theta)$? *The Prior*.
- What is $\mathrm{P}(\theta|y_i)$? *The Posterior*.
- What is $\mathrm{P}(y_i|\theta)$? *the likelihood*.

The posterior is proportional to the prior times the likelihood.

The Likelihood Approach

- Like randomization as the "reasoned basis of inference", developed by R.A. Fisher (in his junior year).

The Likelihood Approach

- Like randomization as the "reasoned basis of inference", developed by R.A. Fisher (in his junior year).
- Following frequentist principles: θ is fixed and y is random.

The Likelihood Approach

- Like randomization as the "reasoned basis of inference", developed by R.A. Fisher (in his junior year).
- Following frequentist principles: θ is fixed and y is random.
- Define the likelihood as:

$$\mathrm{L}(\theta|y) \equiv \mathrm{k}(y)\mathrm{P}(y|\theta)$$

The Likelihood Approach

- Like randomization as the “reasoned basis of inference”, developed by R.A. Fisher (in his junior year).
- Following frequentist principles: θ is fixed and y is random.
- Define the likelihood as:

$$\mathrm{L}(\theta|y) \equiv \mathrm{k}(y)\mathrm{P}(y|\theta)$$

- $\mathrm{k}(y)$ is an unknown function of the data and can be treated as an unknown positive constant.

The Likelihood Approach

- Like randomization as the "reasoned basis of inference", developed by R.A. Fisher (in his junior year).
- Following frequentist principles: θ is fixed and y is random.
- Define the likelihood as:

$$\mathrm{L}(\theta|y) \equiv \mathrm{k}(y)\mathrm{P}(y|\theta)$$

- $\mathrm{k}(y)$ is an unknown function of the data and can be treated as an unknown positive constant.
- $\mathrm{L}(\theta|y)$ is a function: for y fixed at the observed values, it gives the "likelihood" of any value of θ.

The Likelihood Approach

- Like randomization as the "reasoned basis of inference", developed by R.A. Fisher (in his junior year).
- Following frequentist principles: θ is fixed and y is random.
- Define the likelihood as:

$$L(\theta|y) \equiv k(y)P(y|\theta)$$

- $k(y)$ is an unknown function of the data and can be treated as an unknown positive constant.
- $L(\theta|y)$ is a function: for y fixed at the observed values, it gives the "likelihood" of any value of θ.
- Likelihood-model of uncertainty is a *relative* measure appropriate for making comparisons of parameter estimates with the the same dataset, not for making comparisons across datasets.

OLS Using MLE

- The *likelihood* of the probability of the data given the model and inputs:

$$p(y|\beta, \sigma, X) = \prod_{i=1}^{n} \mathrm{N}(y_i|X_i\beta, \sigma^2)$$

where $\mathrm{N}(\cdot|\cdot, \cdot)$ represents the normal probability density function $\mathrm{N}(y|m, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma}}\exp(-\frac{1}{2}(\frac{Y-m}{\sigma})^2)$

OLS Using MLE

- The *likelihood* of the probability of the data given the model and inputs:

$$p(y|\beta, \sigma, X) = \prod_{i=1}^{n} \mathrm{N}(y_i|X_i\beta, \sigma^2)$$

where $\mathrm{N}(\cdot|\cdot,\cdot)$ represents the normal probability density function $\mathrm{N}(y|m, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma}} \exp(-\frac{1}{2}(\frac{Y-m}{\sigma})^2)$

- The more general form of the above expression is:

$$p(y_i|\theta, X_i) = \prod_{i=1}^{n} p(y_i|\theta, X_i)$$

Maximize the Likelihood

- Numerical: grid search

Maximize the Likelihood

- Numerical: grid search
- Analytical:

Maximize the Likelihood

- Numerical: grid search
- Analytical:
  - Write down the likelihood.

Maximize the Likelihood

- Numerical: grid search
- Analytical:
  - Write down the likelihood.
  - Differentiate the likelihood with respect to the paramters.

Maximize the Likelihood

- Numerical: grid search
- Analytical:
  - Write down the likelihood.
  - Differentiate the likelihood with respect to the paramters.
  - Set the derivative equal to zero.

Maximize the Likelihood

- Numerical: grid search
- Analytical:
    - Write down the likelihood.
    - Differentiate the likelihood with respect to the paramters.
    - Set the derivative equal to zero.
    - Solve for the parameters that make the derivative equal zero.

Maximize the Likelihood

- Numerical: grid search
- Analytical:
  - Write down the likelihood.
  - Differentiate the likelihood with respect to the paramters.
  - Set the derivative equal to zero.
  - Solve for the parameters that make the derivative equal zero.
  - Check that the second derivative matrix is negative definite, proving that you've found a *maximum* rather than a *minimum*.

Example

- Start with the simple case of the mean (i.e. where $X_i = 1$ for all i) and where we assume $\sigma = 1$.

Example

- Start with the simple case of the mean (i.e. where $X_i = 1$ for all i) and where we assume $\sigma = 1$.
- It's often easier to work with the "log" of the likelihood:

$$L_i(\beta|y_i) = (2\pi)^{-\frac{1}{2}}\exp(-\frac{1}{2}(y_i - \beta)^2)$$

$$\log L_i(\beta|y_i) = -\frac{1}{2}\log(2\pi) - \frac{1}{2}(y_i - \beta)^2$$

Example

- Start with the simple case of the mean (i.e. where $X_i = 1$ for all i) and where we assume $\sigma = 1$.
- It's often easier to work with the "log" of the likelihood:

$$L_i(\beta|y_i) = (2\pi)^{-\frac{1}{2}}\exp(-\frac{1}{2}(y_i - \beta)^2)$$

$$\log L_i(\beta|y_i) = -\frac{1}{2}\log(2\pi) - \frac{1}{2}(y_i - \beta)^2$$

- In the sample we have N observations, so sum over the N observations to get:

$$\log L(\beta|y_i) = -\frac{1}{2}\sum_{i=1}^{N}(y_i - \beta)^2 - n\log(\sqrt{2\pi})$$

Example, continued

- Next, take the derivative with respect to β.

$$\frac{d \log L}{d\beta} = \sum_{i=1}^{N} (y_i - \beta)$$

Example, continued

- Next, take the derivative with respect to β.

$$\frac{d \log L}{d\beta} = \sum_{i=1}^{N} (y_i - \beta)$$

- Set the derivative to zero and solve: $\hat{\beta} = \frac{\sum y_i}{N}$.

Example, continued

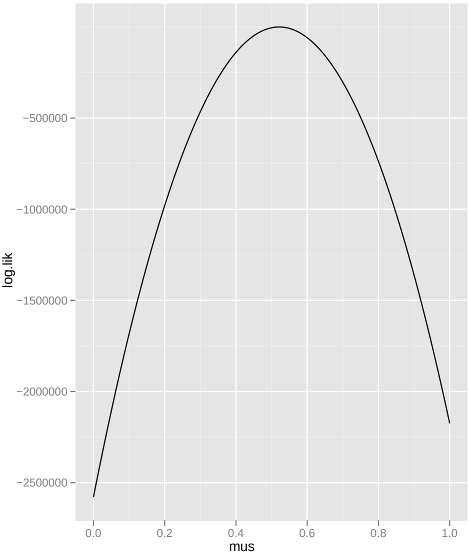
- Next, take the derivative with respect to β.

$$\frac{d \log L}{d\beta} = \sum_{i=1}^{N} (y_i - \beta)$$

- Set the derivative to zero and solve: $\hat{\beta} = \frac{\sum y_i}{N}$.
- Check if negative definite:

$$\frac{d^2 \log L}{d\beta^2} = -N$$

Likelihood Function for One Parameter



OLS in the MLE Framework, continued

- Likelihood for one unit:

$$L_i(\beta|y_i) = (2\pi\sigma^2)^{-\frac{1}{2}} \exp\left(-\frac{1}{2}\frac{(y_i - x_i\beta)^2}{\sigma^2}\right)$$

OLS in the MLE Framework, continued

- Likelihood for one unit:

$$L_i(\beta|y_i) = (2\pi\sigma^2)^{-\frac{1}{2}} \exp\left(-\frac{1}{2}\frac{(y_i - x_i\beta)^2}{\sigma^2}\right)$$

- Likelihood for all the data:

$$L(\beta|y) = \prod_{i=1}^{n}(2\pi\sigma^2)^{-\frac{1}{2}} \exp\left(-\frac{1}{2}\frac{(y_i - x_i\beta)^2}{\sigma^2}\right)$$

OLS in the MLE Framework, continued

- Likelihood for one unit:

$$L_i(\beta|y_i) = (2\pi\sigma^2)^{-\frac{1}{2}} \exp\left(-\frac{1}{2}\frac{(y_i - x_i\beta)^2}{\sigma^2}\right)$$

- Likelihood for all the data:

$$L(\beta|y) = \prod_{i=1}^{n}(2\pi\sigma^2)^{-\frac{1}{2}} \exp\left(-\frac{1}{2}\frac{(y_i - x_i\beta)^2}{\sigma^2}\right)$$

- Log likelihood:

$$\log L = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\sigma^2) - \frac{1}{2}\sum_{i=1}^{n}\frac{(y_i - x_i\beta)^2}{\sigma^2}$$

OLS in the MLE Framework, continued

- Likelihood for one unit:

$$L_i(\beta|y_i) = (2\pi\sigma^2)^{-\frac{1}{2}} \exp\left(-\frac{1}{2}\frac{(y_i - x_i\beta)^2}{\sigma^2}\right)$$

- Likelihood for all the data:

$$L(\beta|y) = \prod_{i=1}^{n}(2\pi\sigma^2)^{-\frac{1}{2}} \exp\left(-\frac{1}{2}\frac{(y_i - x_i\beta)^2}{\sigma^2}\right)$$

- Log likelihood:

$$\log L = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\sigma^2) - \frac{1}{2}\sum_{i=1}^{n}\frac{(y_i - x_i\beta)^2}{\sigma^2}$$

- Log likelihood (in matrix form):

$$\log L = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\sigma^2) - \frac{1}{2}\frac{(\mathbf{y} - \mathbf{x}\beta)'(\mathbf{y} - \mathbf{x}\beta)}{\sigma^2}$$

OLS in the MLE framework, continued

- Doing a bit of algebra, we get

$$\log L = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}(\mathbf{y}'\mathbf{y} - \mathbf{2}\beta'\mathbf{x}'\mathbf{y} + \beta'\mathbf{x}'\mathbf{x}\beta)$$

OLS in the MLE framework, continued

- Doing a bit of algebra, we get

$$\log L = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}(\mathbf{y}'\mathbf{y} - \mathbf{2}\beta'\mathbf{x}'\mathbf{y} + \beta'\mathbf{x}'\mathbf{x}\beta)$$

- Take the derivative with respect to β and we get:

$$\frac{\partial \log L}{\partial \beta} = \frac{1}{\sigma^2}(\mathbf{x}'\mathbf{y} - \mathbf{x}'\mathbf{x}\beta)$$

OLS in the MLE framework, continued

- Doing a bit of algebra, we get

$$\log L = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}(\mathbf{y}'\mathbf{y} - \mathbf{2}\beta'\mathbf{x}'\mathbf{y} + \beta'\mathbf{x}'\mathbf{x}\beta)$$

- Take the derivative with respect to β and we get:

$$\frac{\partial \log L}{\partial \beta} = \frac{1}{\sigma^2}(\mathbf{x}'\mathbf{y} - \mathbf{x}'\mathbf{x}\beta)$$

- Set derivative equal to zero and solve for the MLE:

$$0 = \frac{1}{\sigma^2}(\mathbf{x}'\mathbf{y} - \mathbf{x}'\mathbf{x}\beta) \quad (3)$$

$$\mathbf{x}'\mathbf{x}\beta = \mathbf{x}'\mathbf{y} \quad (4)$$

$$\hat{\beta} = \mathbf{x}'\mathbf{x}^{-1}\mathbf{x}'\mathbf{y} \quad (5)$$

OLS in the MLE framework, continued

- What about the variance, σ^2? Take the derivative of the log-likelihood with respect to σ^2

$$\frac{\partial \log L}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}[(\mathbf{y} - \mathbf{x}\beta)'(\mathbf{y} - \mathbf{x}\beta))]$$

OLS in the MLE framework, continued

- What about the variance, σ^2? Take the derivative of the log-likelihood with respect to σ^2

$$\frac{\partial \log L}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}[(\mathbf{y} - \mathbf{x}\beta)'(\mathbf{y} - \mathbf{x}\beta))]$$

- Set this to zero and rearrange terms:

$$\frac{1}{\sigma^2}[(\mathbf{y} - \mathbf{x}\beta'(\mathbf{y} - \mathbf{x}\beta))] = n$$

OLS in the MLE framework, continued

- What about the variance, σ^2? Take the derivative of the log-likelihood with respect to σ^2

$$\frac{\partial \log L}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}[(\mathbf{y} - \mathbf{x}\beta)'(\mathbf{y} - \mathbf{x}\beta))]$$

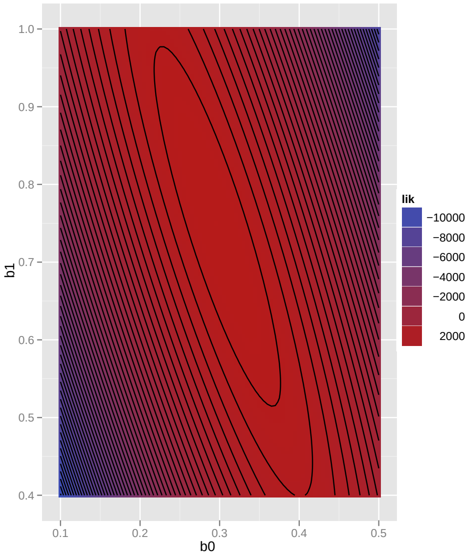
- Set this to zero and rearrange terms:

$$\frac{1}{\sigma^2}[(\mathbf{y} - \mathbf{x}\beta'(\mathbf{y} - \mathbf{x}\beta))] = n$$

- Since we've solved for $\hat{\beta}$, we can replace β with its estimate:

$$\begin{aligned}
\frac{1}{\sigma^2}[(\mathbf{y} - \mathbf{x}\hat{\beta})'(\mathbf{y} - \mathbf{x}\hat{\beta}))] &= n \\
\frac{1}{\sigma^2}[(\mathbf{y} - \hat{\mathbf{y}})'(\mathbf{y} - \mathbf{x}\hat{\mathbf{y}}))] &= n \\
\frac{1}{\sigma^2}[\mathbf{e}'\mathbf{e}] &= n \\
\hat{\sigma}^2 &= \frac{\mathbf{e}'\mathbf{e}}{n}
\end{aligned}$$

Likelihood Function for Two Parameters



Example: Binomial

- Binomial(1, p) with $0 < p < 1$. Let X_i be independent. Each X_i is 1 with probability p and 0 with remaining probability $1 - p$.

Example: Binomial

- Binomial(1, p) with $0 < p < 1$. Let X_i be independent. Each X_i is 1 with probability p and 0 with remaining probability $1 - p$.
- The probability that $X_i = x_i$ for $i, \cdots, n$ is

$$\prod_{i=1}^{n} p^{x_i}(1-p)^{1-x_i}$$

Example: Binomial

- Binomial(1, p) with $0 < p < 1$. Let X_i be independent. Each X_i is 1 with probability p and 0 with remaining probability $1 - p$.
- The probability that $X_i = x_i$ for $i, \cdots, n$ is

$$\prod_{i=1}^{n} p^{x_i}(1-p)^{1-x_i}$$

- Let $S = X_1 + \cdots + X_n$. The likelihood function is as follows:

$$\begin{aligned} L_n(p) &= \sum_{i=1}^{n} [X_i \log p + (1 - X_i) \log(1-p)] && (6) \\ &= S \log p + (n - S) \log(1-p) && (7) \end{aligned}$$

Example: Binomial

- Binomial(1, p) with $0 < p < 1$. Let X_i be independent. Each X_i is 1 with probability p and 0 with remaining probability $1 - p$.
- The probability that $X_i = x_i$ for $i, \cdots, n$ is

$$\prod_{i=1}^{n} p^{x_i}(1-p)^{1-x_i}$$

- Let $S = X_1 + \cdots + X_n$. The likelihood function is as follows:

$$\begin{aligned} L_n(p) &= \sum_{i=1}^{n}[X_i \log p + (1-X_i)\log(1-p)] && (6) \\ &= S\log p + (n-S)\log(1-p) && (7) \end{aligned}$$

- Take the derivative $L_n'(p) = \frac{S}{p} - \frac{n-S}{1-p}$ and the second derivative $L_n''(p) = -\frac{S}{p^2} - \frac{n-S}{(1-p)^2}$. The MLE is $\hat{p} = \frac{S}{n}$.

Fisher Information

Theorem

Suppose $X_i, \cdots, X_n$ are IID with probability distribution governed by the parameter θ. Let θ_0 be the true value of θ. Under regularity conditions (which are omitted here), the MLE for θ is asymptotically normal. The asymptotic mean of the MLE is θ_0. The asymptotic variance can be computed as follows:

$$[-L_n''(\hat{\theta})]^{-1}$$

If $\hat{\theta}$ is the MLE and v_n is the asymptotic variance, the theorem says that $\frac{\hat{\theta}-\theta_0}{\sqrt{v_n}}$ is nearly $N(0,1)$ when the sample size n is large.

The Hessian

- Let θ be a vector containing the paramters being estimated. For example, in the regression $y = \alpha + \beta x + \epsilon$ with variance σ, $\theta : \{\alpha, \beta, \sigma\}$.

The Hessian

- Let θ be a vector containing the paramters being estimated. For example, in the regression $y = \alpha + \beta x + \epsilon$ with variance σ, $\theta : \{\alpha, \beta, \sigma\}$.
- The *Hessian* is a matrix of second derivatives defined as

$$\mathbf{H}(\theta) : \frac{\partial^2 \log L(\theta)}{\partial\theta\partial\theta'}$$

which in our example is:

$$\mathbf{H}(\theta) = \begin{pmatrix} \frac{\partial^2 \log L(\theta)}{\partial\alpha\partial\alpha} & \frac{\partial^2 \log L(\theta)}{\partial\alpha\partial\beta} & \frac{\partial^2 \log L(\theta)}{\partial\alpha\partial\sigma} \\ \frac{\partial^2 \log L(\theta)}{\partial\beta\partial\alpha} & \frac{\partial^2 \log L(\theta)}{\partial\beta\partial\beta} & \frac{\partial^2 \log L(\theta)}{\partial\beta\partial\sigma} \\ \frac{\partial^2 \log L(\theta)}{\partial\sigma\partial\alpha} & \frac{\partial^2 \log L(\theta)}{\partial\sigma\partial\alpha} & \frac{\partial^2 \log L(\theta)}{\partial\sigma\partial\sigma} \end{pmatrix}$$

The Information Matrix

- The *information matrix* is defined as the negative of the expected value of the Hessian: $-E[\mathbf{H}(\theta)]$. Under very general conditions, the covariance matrix for the ML estimator is the inverse of the information matrix:

$$\mathrm{Var}(\hat{\theta}) = -E[\mathbf{H}(\theta)]^{-1}$$

. In our example,

$$\mathrm{Var}((\theta)) = \begin{pmatrix} -E\left(\frac{\partial^2 \log L(\theta)}{\partial\alpha\partial\alpha}\right) & -E\left(\frac{\partial^2 \log L(\theta)}{\partial\alpha\partial\beta}\right) & -E\left(\frac{\partial^2 \log L(\theta)}{\partial\alpha\partial\sigma}\right) \\ -E\left(\frac{\partial^2 \log L(\theta)}{\partial\beta\partial\alpha}\right) & -E\left(\frac{\partial^2 \log L(\theta)}{\partial\beta\partial\beta}\right) & -E\left(\frac{\partial^2 \log L(\theta)}{\partial\beta\partial\sigma}\right) \\ -E\left(\frac{\partial^2 \log L(\theta)}{\partial\sigma\partial\alpha}\right) & -E\left(\frac{\partial^2 \log L(\theta)}{\partial\sigma\partial\alpha}\right) & -E\left(\frac{\partial^2 \log L(\theta)}{\partial\sigma\partial\sigma}\right) \end{pmatrix}^{-1}$$

The Probit Model

- Let Y_i be a 0-1 response variable Y_i for subject i in terms of a row vector of covariates X_i.

The Probit Model

- Let Y_i be a 0-1 response variable Y_i for subject i in terms of a row vector of covariates X_i.
- Given X, the responses Y_i are assumed to be independent random variables taking values 0 or 1, with

$$P(Y_i = 1|X) = \Phi(X_i\beta)$$

where Φ is a standard normal distribution function.

The Probit Model

- Let Y_i be a 0-1 response variable Y_i for subject i in terms of a row vector of covariates X_i.
- Given X, the responses Y_i are assumed to be independent random variables taking values 0 or 1, with

$$P(Y_i = 1|X) = \Phi(X_i\beta)$$

  where Φ is a standard normal distribution function.
- Likelihood function:

$$\prod_i^n Y_i \cdot \Phi(X_i\beta) \times (1 - Y_i) \cdot 1 - \Phi(X_i\beta)$$

The Probit Model

- Let Y_i be a 0-1 response variable Y_i for subject i in terms of a row vector of covariates X_i.
- Given X, the responses Y_i are assumed to be independent random variables taking values 0 or 1, with

$$P(Y_i = 1|X) = \Phi(X_i\beta)$$

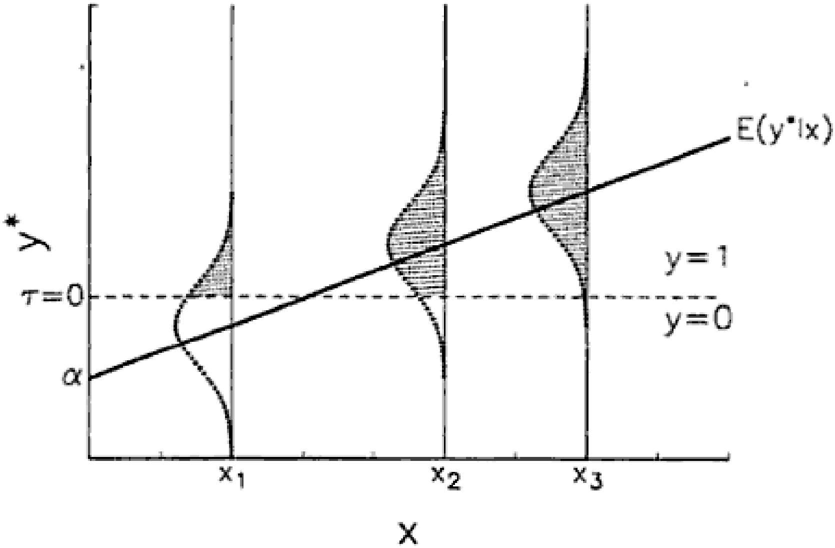
  where Φ is a standard normal distribution function.
- Likelihood function:

$$\prod_i^n Y_i \cdot \Phi(X_i\beta) \times (1 - Y_i) \cdot 1 - \Phi(X_i\beta)$$

- Log likelihood function:

$$L(\beta) = \sum_{i=1}^n (Y_i \log[\Phi(X_i\beta)] + (1 - Y_i) \log[1 - \Phi(X_i\beta)])$$

The Latent Variable Formulation



Latent Variables

- There is an *unmeasured* **latent variable** that determines whether or not an observation is observed as $y_i = 1$ or as $y_i = 0$.

Latent Variables

- There is an *unmeasured* **latent variable** that determines whether or not an observation is observed as $y_i = 1$ or as $y_i = 0$.
- The latent Y^* is assumed to be linearly related to the covariates through the equation:

$$Y_i^* = X_i\beta + U_i$$

Latent Variables

- There is an *unmeasured* **latent variable** that determines whether or not an observation is observed as $y_i = 1$ or as $y_i = 0$.
- The latent Y^* is assumed to be linearly related to the covariates through the equation:

$$Y_i^* = X_i\beta + U_i$$

- The latent variable y^* is linked to the observed binary variable y by the measurement equation:

$$y_i = \begin{cases} 1 & \text{if } Y_i^* > \tau \\ 0 & \text{if } Y_i^* \leq \tau \end{cases}$$

where τ is the *threshold* or *cutpoint*.

The “error” term

- We never observe Y^*, hence we can't estimate the variance of U as we could with ordinary regression.

The “error” term

- We never observe Y^*, hence we can’t estimate the variance of U as we could with ordinary regression.
- What do we do? We assume! U_i is independent of the X_i’s and IID across subjects.

The "error" term

- We never observe Y^*, hence we can't estimate the variance of U as we could with ordinary regression.
- What do we do? We assume! U_i is independent of the X_i's and IID across subjects.
- For probit, assume $E(U) = 0$, $\sigma = 1$ and is distributed normally.

The "error" term

- We never observe Y^*, hence we can't estimate the variance of U as we could with ordinary regression.
- What do we do? We assume! U_i is independent of the X_i's and IID across subjects.
- For probit, assume $E(U) = 0$, $\sigma = 1$ and is distributed normally.
- For logit, assume $E(U) = 0$, $\sigma = \pi^2/3$ which results in the following PDF:

$$\lambda(U) = \frac{\exp(U)}{(1 + \exp(U))^2}$$

and the following CDF:

$$\Lambda(U) = \frac{\exp(U)}{1 + \exp(U)}$$

Logistic vs Normal: PDF

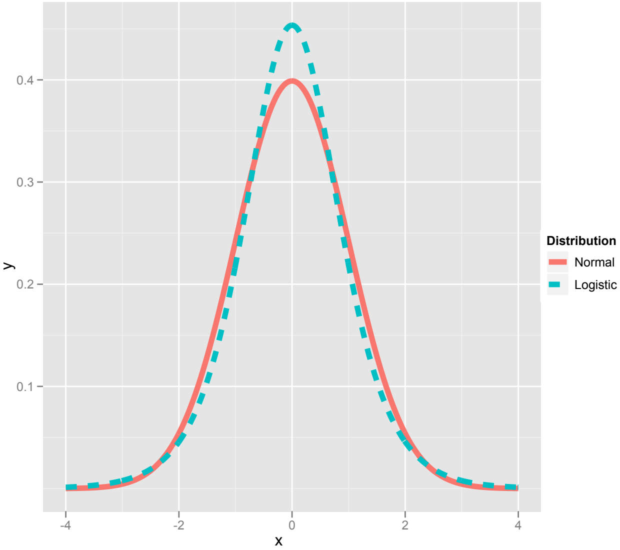
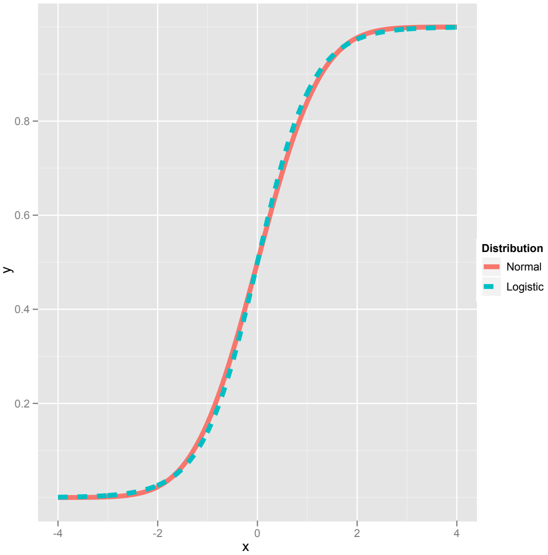
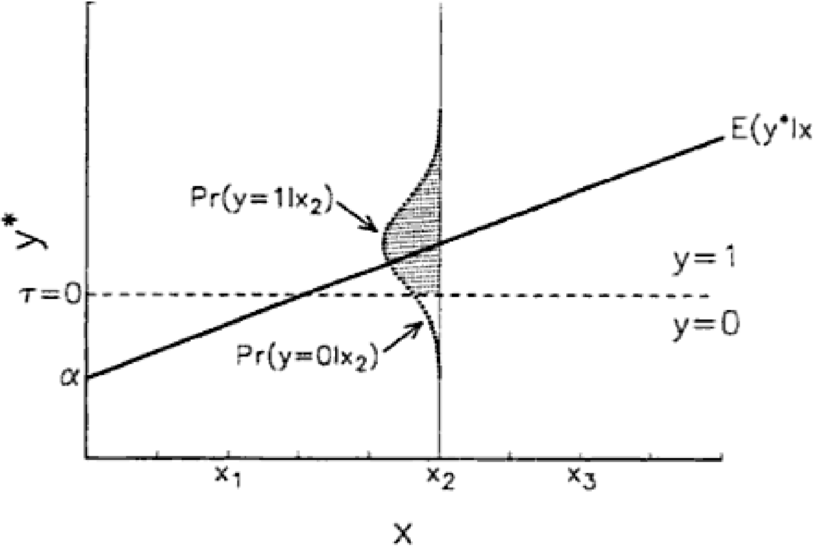


Figure: PDFs for Logistic and Normal Distribution

Logistic vs Normal: CDF



Probability of Observed Values



From Latent Variables to MLE

- Write $P(X_i\beta + U_i > 0) = P(U_i > -X_i\beta) = P(-U_i < X_i\beta)$ and because U_i is distributed symmetrically around 0, $P(-U_i < X_i\beta) = P(U_i < X_i\beta) = \Phi(X_i\beta)$, in the probit case.

From Latent Variables to MLE

- Write $P(X_i\beta + U_i > 0) = P(U_i > -X_i\beta) = P(-U_i < X_i\beta)$ and because U_i is distributed symmetrically around 0, $P(-U_i < X_i\beta) = P(U_i < X_i\beta) = \Phi(X_i\beta)$, in the probit case.
-