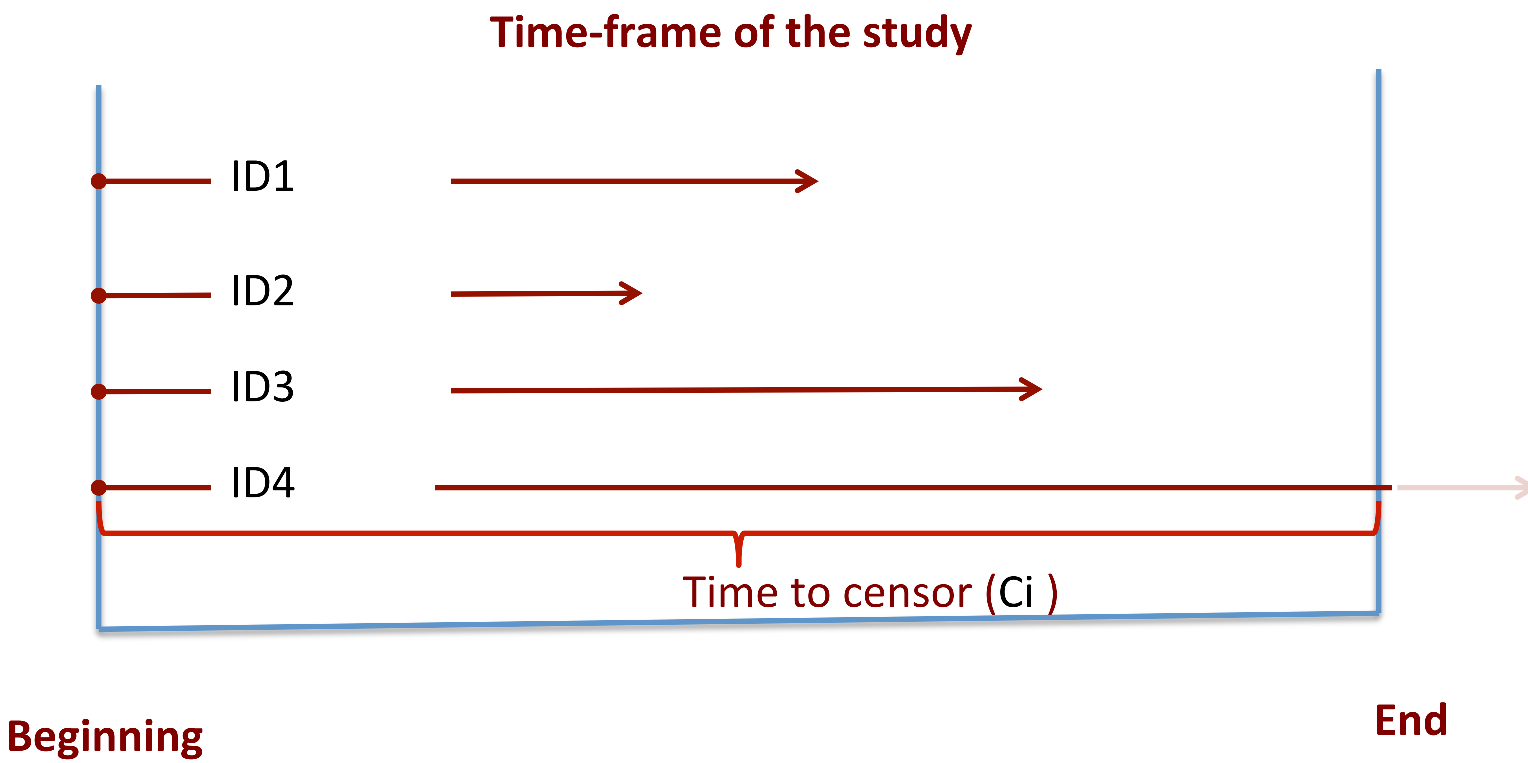
STT 465
Bayesian Multiple Linear Regression:

=> Regression With Censored Data

=> Regression with binary outcomes

Regression with Censored Data

Right-Censoring ($Y_i>C_i$)



IDs 1,2 and 3: time to event is observed.

ID 4: time to event is unknown; however we know the time to censor, and we know that the event will happens after Y_4>C4 (right-censored data)

Notation

Data will be defined by a pair of variables

- Y_i : time to event or time to censoring.
- $C_{i:}$ 1 if event, 0 if censored

Likelihood Function

Observed Data-points

$$p\left(y_i \mid X,\beta,\sigma_\varepsilon^2,c_i=1\right)=\frac{Exp\left\{-\frac{\left(y_i-x_i'\beta\right)^2}{2\sigma_\varepsilon^2}\right\}}{\sqrt{2\pi\sigma_\varepsilon^2}}$$

Censored Data-points

$$p\left(y_i \mid X,\beta,\sigma_\varepsilon^2,c_i=0\right)=\int_{-\infty}^{y_i}\frac{Exp\left\{-\frac{\left(\tilde{y}_i-x_i'\beta\right)^2}{2\sigma_\varepsilon^2}\right\}}{\sqrt{2\pi\sigma_\varepsilon^2}}d\tilde{y}_{\boldsymbol{i}}$$

Likelihood Function

$$p\left(y \mid X,\beta,\sigma_\varepsilon^2\right)=\prod_{i:c_i=1}\frac{e^{\left\{-\frac{\left(y_i-x_i'\beta\right)^2}{2\sigma_\varepsilon^2}\right\}}}{\sqrt{2\pi\sigma_\varepsilon^2}}\times\prod_{i:c_i=0}\int_{-\infty}^{y_i}\frac{Exp\left\{-\frac{\left(\tilde{y}_i-x_i'\beta\right)^2}{2\sigma_\varepsilon^2}\right\}}{\sqrt{2\pi\sigma_\varepsilon^2}}d\tilde{y}_{\boldsymbol{i}}$$

- The above function can be used to derive Max. Likelihood Estimates (see survreg of survival package)
- ML and Bayesian analysis can be difficult because the integrals involved do not have closed forms.
- Instead we will use 'data augmentation'

Data Augmentation

Observed Data-points

$$\int_{-\infty}^{y_i} \frac{Exp\left\{-\frac{(\tilde{y}_i - x_i'\beta)^2}{2\sigma_\varepsilon^2}\right\}}{\sqrt{2\pi\sigma_\varepsilon^2}} d\tilde{y}_i = \int_{-\infty}^{\infty} \frac{Exp\left\{-\frac{(\tilde{y}_i - x_i'\beta)^2}{2\sigma_\varepsilon^2}\right\}}{\sqrt{2\pi\sigma_\varepsilon^2}} \times C_i \, d\tilde{y}_i$$

We will perform this integral using Monte Carlo Methods

Bayesian Likelihood

$$p(y \mid X, \beta, \sigma_\varepsilon^2) = \prod_{i:c_i=1} \frac{e^{\left\{-\frac{(y_i - x_i'\beta)^2}{2\sigma_\varepsilon^2}\right\}}}{\sqrt{2\pi\sigma_\varepsilon^2}} \times \prod_{i:c_i=0} \frac{Exp\left\{-\frac{(\tilde{y}_i - x_i'\beta)^2}{2\sigma_\varepsilon^2}\right\}}{\sqrt{2\pi\sigma_\varepsilon^2}} 1(\tilde{y}_i > y_i)$$

$\tilde{y}_i$: Unobserved time to event.

Fully Conditional Distribution

Likelihood Function

$$p\left(y \mid X,\beta,\sigma_\varepsilon^2\right) = \prod_{i:c_i=1} \frac{e^{\left\{-\frac{\left(y_i - x_i'\beta\right)^2}{2\sigma_\varepsilon^2}\right\}}}{\sqrt{2\pi\sigma_\varepsilon^2}} \times \prod_{i:c_i=0} \frac{Exp\left\{-\frac{\left(\tilde{y}_i - x_i'\beta\right)^2}{2\sigma_\varepsilon^2}\right\}}{\sqrt{2\pi\sigma_\varepsilon^2}} 1\left(\tilde{y}_i > y_i\right)$$

Prior

$$p\left(\beta,\sigma_\varepsilon^2,\sigma_\beta^2,\ldots\right)$$

Joint Posterior:

$$p\left(\beta,\sigma_\varepsilon^2,\sigma_\beta^2,\ldots \mid y,c\right) \propto \left[\prod_{i:c_i=1} \frac{e^{\left\{-\frac{\left(y_i - x_i'\beta\right)^2}{2\sigma_\varepsilon^2}\right\}}}{\sqrt{2\pi\sigma_\varepsilon^2}} \times \prod_{i:c_i=0} \frac{Exp\left\{-\frac{\left(\tilde{y}_i - x_i'\beta\right)^2}{2\sigma_\varepsilon^2}\right\}}{\sqrt{2\pi\sigma_\varepsilon^2}} 1\left(\tilde{y}_i > y_i\right)\right] \times p\left(\beta,\sigma_\varepsilon^2,\sigma_\beta^2,\ldots\right)$$

Fully Conditional (truncated normal):

$$p\left(\tilde{y}_i \mid c_i =\right) \propto \frac{e^{\left\{-\frac{\left(\tilde{y}_i - x_i'\beta\right)^2}{2\sigma_\varepsilon^2}\right\}}}{\sqrt{2\pi\sigma_\varepsilon^2}} 1\left(\tilde{y}_i > y_i\right)$$

Regression with Binary Outcomes

Regression with Binary Outcomes

- Binary outcomes follow Bernoulli distributions

$$y_i \in \{0,1\} \qquad\qquad p\left(y_i \middle| \theta\right) = \theta^{y_i}\left(1-\theta\right)^{1-y_i}$$

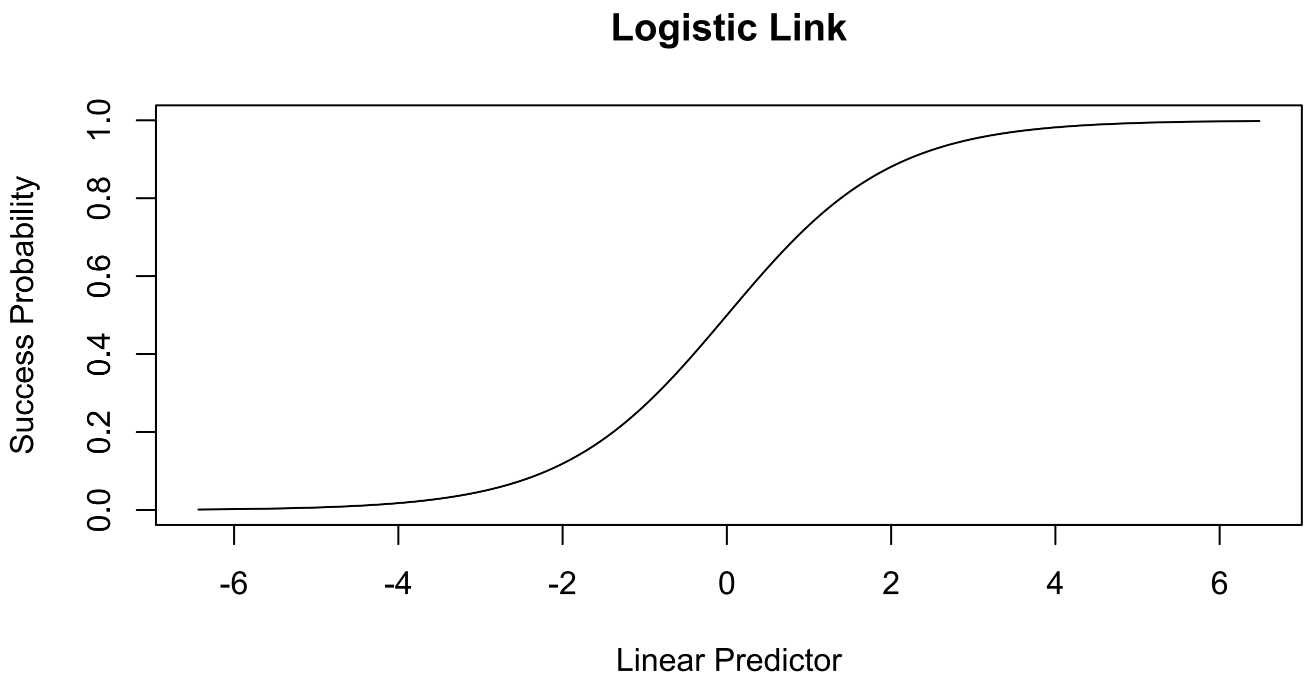
- Regression with Binary outcomes: we want to make θ a function of one or more predictors.
- Problem:
    - θ lives in the 0-1 interval,
    - while a regression function $\sum_{j=1}^{p} x_{ij}\beta_j$ lives in the real line.

- To get around this we need to introduce a link function that maps from the real line to the 0-1 interval.

- The most commonly used links are the logit and probit link.

Logistic Regression

$$p(y_i = 1) = \theta \qquad \log\left(\frac{\theta}{1-\theta}\right) = \sum_{j=1}^{p} x_{ij}\beta_j \qquad \Rightarrow \theta = \frac{e^{\sum_{j=1}^{p} x_{ij}\beta_j}}{1 + e^{\sum_{j=1}^{p} x_{ij}\beta_j}}$$



Logistic Regression

- Likelihood Function

$$p(y) = \prod_{i=1}^{n} \theta^{y_i} (1-\theta)^{1-y_i}$$

$$\log\left(\frac{\theta}{1-\theta}\right) = \eta_i = \sum_{j=1}^{p} x_{ij}\beta_j \qquad \Rightarrow \theta_i = \frac{e^{\sum_{j=1}^{p} x_{ij}\beta_j}}{1+e^{\sum_{j=1}^{p} x_{ij}\beta_j}}$$

$$p(y) = \prod_{i=1}^{n} \left[\frac{e^{\eta_i}}{1+e^{\eta_i}}\right]^{y_i} \left[1 - \frac{e^{\eta_i}}{1+e^{\eta_i}}\right]^{1-y_i}$$

- The above function can be used to derive Max. Likelihood Estimates (see `glm(y~X, family=binomial)`)

Threshold Model (Probit Link)

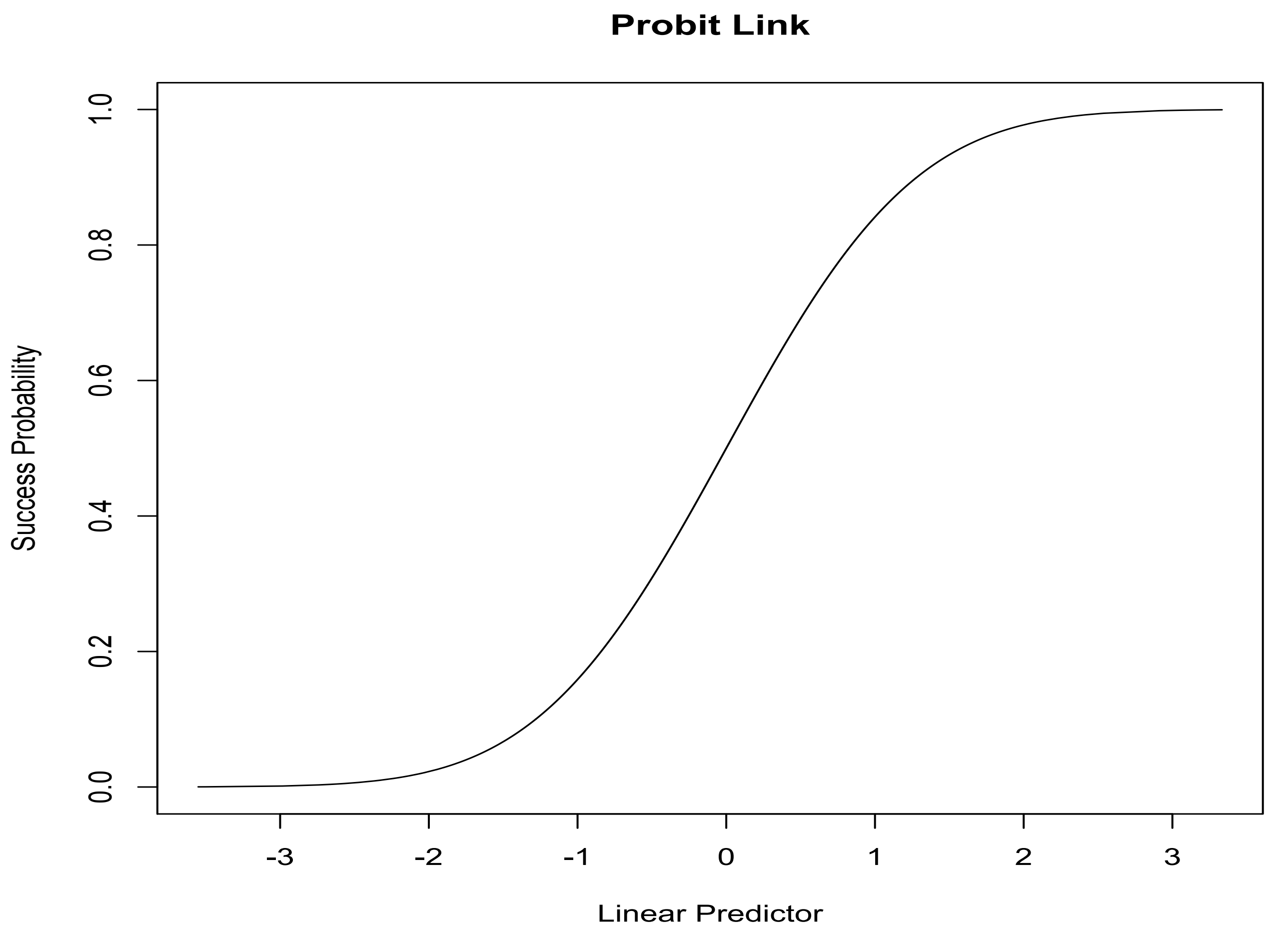
$$\tilde{y}_i = \sum_{j=1}^{p} x_{ij}\beta_j + \varepsilon_i \qquad y_i = \begin{cases} 0 & \tilde{y}_{\boldsymbol{i}} < 0 \\ 1 & \text{Otherwise} \end{cases}$$

$$\varepsilon_i \overset{iid}{\sim} N(0,1)$$

$$\begin{aligned} p(y_i = 1) = p(\tilde{y}_{\boldsymbol{i}} > 0) &= p\left(\sum_{j=1}^{p} x_{ij}\beta_j + \varepsilon_i > 0\right) \\ &= p\left(\varepsilon_i > -\sum_{j=1}^{p} x_{ij}\beta_j\right) \\ &= p\left(-\varepsilon_i < \sum_{j=1}^{p} x_{ij}\beta_j\right) \\ &= \text{pnorm}\left(q = \sum_{j=1}^{p} x_{ij}\beta_j\right) \end{aligned}$$

Threshold Model (Probit link)

$$p(y_i = 1) = \theta \qquad \theta = pnorm\left(\sum_{j=1}^{p} x_{ij}\beta_j\right)$$



Threshold Model (Probit Link)

- Likelihood Function

$$p(y)=\prod_{i=1}^{n}\theta^{y_i}\left(1-\theta\right)^{1-y_i} \qquad \theta = pnorm\left(\sum_{j=1}^{p}x_{ij}\beta_j\right)$$

$$p(y)=\prod_{i=1}^{n}\left[pnorm\left(\sum_{j=1}^{p}x_{ij}\beta_j\right)\right]^{y_i}\left[1-pnorm\left(\sum_{j=1}^{p}x_{ij}\beta_j\right)\right]^{1-y_i}$$

- The above function can be used to derive Max. Likelihood Estimates (see `glm(y~X,family=binomial(link=probit))`)
- ML and Bayesian analysis can be difficult because the integrals involved do not have closed forms.
- Instead we will use 'data augmentation'

Bayesian Model For Binary Outcomes

(Bayesian) Likelihood Function

$$p\left(\tilde{y} \mid X, y, \beta, \sigma_\varepsilon^2\right) = \prod_{i:c_i=1} \frac{e^{\left\{-\frac{(\tilde{y}_i - x_i'\beta)^2}{2}\right\}} 1\left(\tilde{y}_i > 0\right)^{y_i} \times 1\left(\tilde{y}_i < 0\right)^{1-y_i}}{\sqrt{2\pi}}$$

$\tilde{y}_i$: Unobserved liability.

Prior $\quad p\left(\beta, \sigma_\varepsilon^2, \sigma_\beta^2, \ldots\right)$

Joint Posterior

$$p\left(\tilde{y}, \beta, \sigma_\varepsilon^2, \sigma_\beta^2, \ldots \mid X, y,\right) \propto \prod_{i:c_i=1} \frac{e^{\left\{-\frac{(\tilde{y}_i - x_i'\beta)^2}{2}\right\}} 1\left(\tilde{y}_i > 0\right)^{y_i} \times 1\left(\tilde{y}_i < 0\right)^{1-y_i}}{\sqrt{2\pi}} \times p\left(\beta, \sigma_\varepsilon^2, \sigma_\beta^2, \ldots\right)$$

Fully Conditional Distributions

$$p\left(\tilde{y}_i \mid ELSE\right) \propto \frac{e^{\left\{-\frac{\left(\tilde{y}_i - x_i'\beta\right)^2}{2}\right\}} 1\left(\tilde{y}_i > 0\right)^{y_i} \times 1\left(\tilde{y}_i < 0\right)^{1-y_i}}{\sqrt{2\pi}}$$

- This is the kernel of a truncated normal density, right or left depending on whether y is equal to zero or one.

- Therefore, we sample, for each subject in the sample, the unobserved liability from a truncated normal distribution, truncated below zero if y[i]=0, otherwise truncated above zero.

- Once we sample the un-observed liability, we treat liability as observed, therefore, all the other fully conditional densities are as in the standard linear regression model.