### Parameterization and Bayesian Modeling

Andrew Gelman Department of Statistics and Department of Political Science, Columbia University, New York

10 Oct 2011

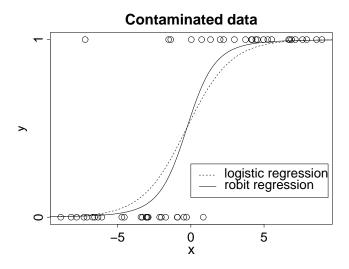
# Statistical computation and statistical modeling

- The folk theorem of statistical computing
- The Pinocchio principle
- Latent variables, transformations, and Bayes
- Examples (from Gelman, 2004, and Gelman and Hill, 2007):
  - 1. Robust logistic regression
  - 2. Truncated and censored data
  - 3. Modeling continuous data using an underlying discrete distirbution
  - 4. Modeling discrete data using an underlying continuous distirbution
  - 5. Parameter expansion for hierarchical models
  - 6. Iterative algorithms and time processes

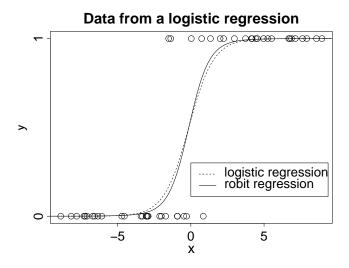
- ► Logistic regression: Pr(y=1) = logit<sup>-1</sup>(X<sub>i</sub>β)
- Try to fit in Bugs ... Crash!
- ► Work-around: bound probs between 0.01 and 0.99:  $Pr(y=1) = 0.01 + 0.98 \log t^{-1}(X_i\beta)$

- Logistic regression:
  - $\Pr(y=1) = \log it^{-1}(X\beta)$
- The problem of outliers
  - ▶ ??
  - ▶ 1's in regions where you expect 0's, or 0's where you expect 1's
  - The result: poor fit to the mass of the data
- Robust model:  $\Pr(y=1) = 0.01 + 0.98 \log it^{-1}(X\beta)$ 
  - ► Allows a 1% chance of random error in either direction
  - Motivating example: modeling Supreme Court decisions
- Alternative version uses the cumulative t distribution rather than the logistic curve

#### Robust logit works for contaminated data



### Robust logit also does ok when there is no contamination



#### Example 2: Truncated and censored data

- Sample of size N from a distribution f(y|θ). We only observe the measurments that are less than 200. Out of these N, we observe n = 91 cases where y<sub>i</sub> < 200.</p>
- Goal is inference about  $\theta$
- Scenario 1: N is unknown. Use the truncated-data likelihood:

$$p(\theta|y) \propto p(\theta) \left[1 - F(y|\theta)\right]^{-91} \prod_{i=1}^{91} f(y_i|\theta)$$

Scenario 2: N is known. Use the censored-data likelihood:

$$p( heta|y, N) \propto p( heta) F(y| heta)^{N-91} \prod_{i=1}^{91} f(y_i| heta)$$

## Truncated and censored data: things get weird

- ▶ Now suppose *N* is unknown. A Bayesian has two options:
- Use the truncated-data model, or
- ► Use the censored-data model, averaging over the unknown N:

$$p(\theta|y) \propto \sum_{N=91}^{\infty} p(N)p(\theta) \begin{pmatrix} N \\ 91 \end{pmatrix} F(y|\theta)^{N-91} \prod_{i=1}^{91} f(y_i|\theta)$$

If p(N) ∝ 1/N—but only with this prior—the above summation reduces to the truncated-data model:

$$p(\theta|y) \propto p(\theta) \left[1 - F(y|\theta)\right]^{-91} \prod_{i=1}^{91} f(y_i|\theta)$$

▶ ?!

# Example 3: Modeling continuous data using an underlying discrete distribution

A bimodal distribution modeled using a mixture:

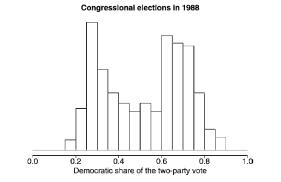


Figure 1. Histogram of Democratic Share of the Two-Party Vote in Congressional Elections in 1988. Only districts that were contested by both major parties are shown here.

9/16

# Separating into Republicans, Democrats, and open seats

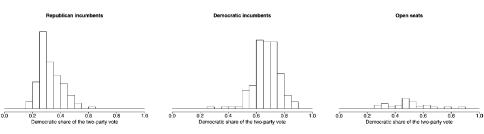


Figure 2. Histogram of Democratic Share of the Two-Party Vote in Congressional Elections in 1988, in Districts With (a) Republican Incumbents, (b) Democratic Incumbents, and (c) Open Seats. Combined, the three distributions yield the bimodal distribution in Figure 1.

#### We took the mixture components seriously ... and then they came to life!

# Example 4: Modeling discrete data using an underlying continuous distribution

Logistic regression model for voting:

 $\Pr(\text{vote Republican}) = \log t^{-1}(X\beta)$ 

Latent variable interpretation:

 $z = X\beta + \epsilon$ ; vote Republican if z > 0

- ► Take *z* seriously as a continuous measure of Republican-ness
- Can look at correlations with other z's, changes over time, ....

# Example 5: Parameter expansion for hierarchical models

► Hierarchical regression with *M* batches of coefficients:

$$y = \sum_{m=1}^{M} X^{(m)} \beta^{(m)} + ext{error}$$

Exchangeable prior distributions: within each batch m,

$$\beta_j^{(m)} \sim \mathsf{N}(0, \sigma_m^2), \text{ for } j = 1, \dots, J_m$$

- $\blacktriangleright$  Gibbs sampler for coefficients  $\beta$  and hyperparameters  $\sigma$  gets stuck near  $\sigma\approx$  0
- Parameter expansion:

$$y = \sum_{m=1}^{M} \xi_m X^{(m)} \beta^{(m)} + ext{error}$$

• Run Gibbs on  $\xi, \beta, \sigma$ 

- $y = \sum_{m=1}^{M} \xi_m X^{(m)} \beta^{(m)} + \text{error}$
- Factor-analysis model: ξ<sub>m</sub> is the importance of factor m, and the β<sub>i</sub><sup>(m)</sup>'s are the factor loadings
- Need informative prior distributions, for example  $\beta_j^{(m)} \sim N(1, 0.2^2)$
- The coefficients  $\beta_j^{(m)}$  get a new life as latent factor loadings

# New prior distributions inspired by parameter expansion

- ► Simple hierarchical model:  $y_{ij} \sim N(\theta_j, \sigma_y^2), \ \theta_j \sim N(\mu, \sigma_{\theta}^2)$
- What's a good prior distribution for  $\sigma_{\theta}$ ?
- ► Redundant parameterization:  $y_{ij} \sim N(\mu + \xi \eta_j, \sigma_y^2), \ \eta_j \sim N(0, \sigma_\eta^2)$ 
  - Same model:  $\theta_j = \mu + \xi \eta_j$  for each *j*, and  $\sigma_{\theta} = |\xi| \sigma_{\eta}$
  - Conditionally conjugate prior: normal distribution for  $\xi$  and inv- $\chi^2$  for  $\sigma_n^2$
  - Put these together to get a half-t prior for  $\sigma_{\xi}$  (Gelman, 2006)
  - It really works!
- Similar ideas for covariance matrices

# Example 6: Iterative algorithms and time processes

- Iterative simulation turns an intractable spatial process into a tractable space-time process
- MCMC algorithms such as Hamiltonian dynamics (in which "position" and "momentum" variables evlove over time)
- Simulation of "agent models" to determine equilibrium states in economics
- Connection to the Folk Theorem
- Latent variables and the cognitive revolution in psychology
- Mapping the Gibbs sampler or Metropolis algorithm to a real-time learning process
  - Burn-in = "learning"
  - Moving through the distribution = tracing of variation

- The folk theorem
- Often we add latent variables that improve computation but do not change the likelihood function
  - New parameters are sometimes partially identified
  - Sometimes completely nonidentified
- The Pinocchio principle
- New classes of prior distributions