Statistics 225
Bayesian Statistical Analysis (Part 5)

Hal Stern

Department of Statistics
University of California, Irvine
sternh@uci.edu

March 28, 2019

Models for robust inference and sensitivity analysis

- Assume we have completed an analysis
  (i.e., we have obtained posterior simulations
  from a specified model)
- Often want to assess:
  - sensitivity of inferences (do the results change under other reasonable models)
  - robustness to outliers by considering overdispersed alternatives to our model (e.g., t rather than normal)
  - overdispersed version of model to address heterogeneity
  - effect of other small changes (e.g., deleting an observation)
- Computational approaches
  - exact posterior inference under new model (may be quite time consuming)
  - approximate posterior inference using importance ratios

Models for robust inference and sensitivity analysis

Example

- Recall SAT coaching example:
  - data: $y = (28, 8, -3, 7, -1, 1, 18, 12)$
  - model: $$\begin{aligned} & y_j|\theta_j \sim N(\theta_j, \sigma_j^2) \\ & \theta_j|\mu, \tau^2 \sim N(\mu, \tau^2) \\ & p(\mu, \tau) \propto 1 \end{aligned}$$
- What happens if we replace 12 for 100?
  - estimate of τ^2 gets bigger
  - estimate of θ_j moves towards y_j
  - if school 8 is an outlier, this affects conclusions for other seven schools
- Would a t-model at one or both stages of the model help?

Models for robust inference and sensitivity analysis
Overdispersed models

- We have seen that allowing for heterogeneity among units leads to "new" overdispersed models
- Binomial and Beta-Binomial
  – Standard model: $y_i \sim \text{Binomial}(n, p)$
  $$E(y_i) = np \qquad V(y_i) = np(1-p)$$
  – Overdispersed model:
  $$\left.\begin{array}{lcl} y_i & \sim & \text{Binomial}(n, p_i) \\ p_i & \sim & \text{Beta}(\alpha, \beta) \end{array}\right\} \Rightarrow y_i \sim \text{Beta-Binomial}(n, \alpha, \beta)$$
  $$E(y_i) = n \underbrace{\left(\frac{\alpha}{\alpha+\beta}\right)}_{p}$$
  $$V(y_i) = n \underbrace{\left(\frac{\alpha}{\alpha+\beta}\right)}_{p} \underbrace{\left(\frac{\beta}{\alpha+\beta}\right)}_{1-p} \underbrace{\left(\frac{\alpha+\beta+n}{\alpha+\beta+1}\right)}_{(*)}$$

$(*)$ = overdispersion factor

Models for robust inference and sensitivity analysis

Overdispersed models (cont'd)

- Poisson and Negative Binomial Poisson
  - Poisson: variance equals to mean
  - Negative Binomial: two-parameter distn allows the mean and variance to be fitted separately, with variance as least as great as the mean
  - Overdispersed model:

$$\left.\begin{array}{lcl} y_i & \sim & \text{Poisson}(\lambda_i) \\ \lambda_i & \sim & \text{Gamma}(\alpha, \beta) \end{array}\right\} \Rightarrow y_i \sim \text{Neg-Bin}(\alpha, \beta)$$

$$E(y_i) = \frac{\alpha}{\beta} \quad V(y_i) = \frac{\alpha}{\beta} \underbrace{\left(\frac{\beta+1}{\beta}\right)}_{(*)}$$

$(*)$ = overdispersion factor

Models for robust inference and sensitivity analysis

Overdispersed models (cont'd)

- Normal and t-distribution

  t has a longer tail than the normal and can be used for accommodating:

  (a) occasional unusual observations in the data distribution

  (b) occasional extreme parameters in the prior distribution or hierarchical model

  Overdispersed model:

$$\left.\begin{array}{rcl} y_i & \sim & \mathrm{N}(\mu, V_i) \\ V_i & \sim & \text{Inv-}\chi^2(\nu, \sigma^2) \end{array}\right\} \Rightarrow y_i \sim t_\nu(\mu, \sigma^2)$$

$$E(y_i) = \mu \qquad V(y_i) = \sigma^2 \underbrace{\left(\frac{\nu}{\nu - 2}\right)}_{(*)} \text{ for } \nu > 2$$

$(*)$ = overdispersion factor

Models for robust inference and sensitivity analysis

- Overdispersed (robust) models are "safer" in the sense that they include the non-robust models as a special case (e.g., normal is t with infinite d.f.)
- Why not start with robust (expanded) models?
  - non-robust models have special justification
    - normal justified by CLT
    - Poisson justified by Poisson process
  - non-robust models often computationally convenient

Models for robust inference and sensitivity analysis

Notation for model expansion

- $p_o(y|\theta) =$ sampling distribution for original model
- $p(y|\theta, \phi) =$ expanded sampling model for y
- $\phi =$ hyperparameter defining expanded model
- Normal/t example
  - $y|\mu, \sigma^2, \nu \sim t_\nu(\mu, \sigma^2)$ [i.e., $\theta = (\mu, \sigma^2)$ and $\phi = \nu$]
  - $p_o(y|\mu, \sigma^2) = N(y|\mu, \sigma^2)$ [$\nu = \infty$]
- Can be applied to data model (as above) or prior distribution for θ in a hierarchical model

Models for robust inference and sensitivity analysis
Computation

- Possible inferences
  - fit the model for one or more fixed ϕ's
    $$p(\theta|y,\phi) \propto p(\theta|\phi)p(y|\theta,\phi)$$
    e.g., $\phi = 4$ d.f. for t-distribution.
  - examine joint posterior of θ and ϕ
    $p(\theta,\phi|y) = p(\phi|y)p(\theta|y,\phi)$
- Computational approaches:
  - redo analysis for expanded model
    (use MCMC, especially Gibbs sampling)
  - approximations based on importance weights
  - approximations based on importance resampling

Models for robust inference and sensitivity analysis

Computation: complete analysis

- Consider t_ν distribution (ν specified) as a robust alternative to normal model
- Model:

$$y_i|\mu, Vi, \sigma^2 \sim N(\mu, V_i\sigma^2) \qquad V_i \sim \text{Inv-}\chi^2(\nu, 1)$$

$$p(\mu, \sigma^2) \propto \sigma^{-2}$$

- Posterior distribution

$$p(\mu, \sigma^2, V|y, \nu) \propto \frac{1}{\sigma^2}\prod_{i=1}^{n}\left[\frac{e^{-\nu/(2V_i)}}{V_i^{\nu/2+1}}\right] \times \prod_{i=1}^{n}\frac{e^{-\frac{1}{2}\frac{(y_i-\mu)^2}{V_i\sigma^2}}}{\sqrt{\sigma^2 V_i}}$$

- Computation via Gibbs sampler
  - $\mu|V_i, \sigma^2, y \sim N\left(\frac{\sum_{i=1}^n y_i/V_i}{\sum_{i=1}^n 1/V_i}, \frac{\sigma^2}{\sum_{i=1}^n 1/V_i}\right)$
  - $\sigma^2|V_i, \mu, y \sim \text{Inv-}\chi^2\left(n, \frac{1}{n}\sum_{i=1}^n \frac{(y_i-\mu)^2}{V_i}\right)$
  - $V_i|\mu, \sigma^2, y \sim \text{Inv-}\chi^2\left(\nu+1, \frac{\nu + \frac{(y_i-\mu)^2}{\sigma^2}}{\nu+1}\right)$

Models for robust inference and sensitivity analysis

Computation: complete analysis (cont'd)

- What if ν is unknown? Give it a prior distn $p(\nu)$ and include in the model as a parameter
- Posterior distribution

$$p(\mu, \sigma^2, V, \nu | y) \propto \frac{p(\nu)}{\sigma^2} \prod_{i=1}^{n} \left[\frac{e^{-\nu/(2V_i)} (\nu/2)^{\nu/2}}{\Gamma(\nu/2) V_i^{\nu/2+1}} \right] \times \prod_{i=1}^{n} \frac{e^{-\frac{1}{2}\frac{(y_i-\mu)^2}{V_i \sigma^2}}}{\sqrt{\sigma^2 V_i}}$$

- First three Gibbs steps are same as on previous slide
- Metropolis step to draw from conditional distn of ν, that is $p(\nu | \sigma^2, \mu, V, y)$

Models for robust inference and sensitivity analysis
Approximation based on importance weights

- Want to consider robust model without redoing the analysis
- Suppose interested in quantity of the form $E[h(\theta)|\phi, y]$
  - importance sampling review:

$$E(g) = \int g(y)f(y)dy \quad = \quad \int \frac{g(y)f(y)}{p(y)} p(y)dy$$
$$\approx \quad \frac{1}{N}\sum_{i=1}^{N} \frac{g(y_i)f(y_i)}{p(y_i)}$$

    where y_i's are sampled from $p(y)$
  - importance sampling approach
    (sometimes called importance weighting here)
    - $E[h(\theta)|\phi, y] = \int h(\theta)p(\theta|\phi, y)d\theta$
      $$= \underbrace{\frac{p_o(y)}{p(y|\phi)}}_{\text{unknown}} \int \frac{h(\theta)p(\theta|\phi)p(y|\theta,\phi)}{p_o(\theta)p_o(y|\theta)} p_o(\theta|y)d\theta$$
    - initial unknown term makes things slightly different
    - use draws θ^l, $l = 1, \ldots, L$ from $p_o(\theta|y)$ but not the usual importance sampling estimate

Models for robust inference and sensitivity analysis

Approximation based on importance weights

- Importance sampling (weighting) (cont'd)
  - estimate $E[h(\theta)|\phi, y]$ with

$$\hat{h} = \frac{\frac{1}{L}\sum_{l=1}^{L} \frac{h(\theta^l)p(\theta^l|\phi)p(y|\theta^l,\phi)}{p_o(\theta^l)p_o(y|\theta^l)} \frac{p_o(y)}{p(y|\phi)}}{\frac{1}{L}\sum_{l=1}^{L} \frac{p(\theta^l|\phi)p(y|\theta^l,\phi)}{p_o(\theta^l)p_o(y|\theta^l)} \frac{p_o(y)}{p(y|\phi)}}$$

    i.e.

$$\hat{h} = \frac{\frac{1}{L}\sum_{l=1}^{L} \omega_l h(\theta^l)}{\frac{1}{L}\sum_{l=1}^{L} \omega_l} = \frac{\sum_{l=1}^{L} \omega_l h(\theta^l)}{\sum_{l=1}^{L} \omega_l}$$

    where

$$\omega_l = \frac{p(\theta^l|\phi)p(y|\theta^l,\phi)}{p_o(\theta^l)p_o(y|\theta^l)}$$

  - note: denominator in $\hat{h}$ is same as numerator with $h = 1$ (essentially estimates reciprocal of unknown constant)

Models for robust inference and sensitivity analysis

Approximation based on importance weights

- Importance resampling
  - may be interested in quantities that are not posterior expectations
  - related idea of importance resampling is to obtain an "approximate sample" from $p(\theta|\phi, y)$
    - sample θ^l, $l = 1, \ldots, L$ from $p_o(\theta|y)$ with L large
    - calculate importance ratios

$$\frac{p(\theta^l|\phi)p(y|\theta^l, \phi)}{p_o(\theta^l)p_o(y|\theta^l)}$$

    - check distribution of importance ratios
    - subsample n draws without replacement from L draws with probability proportional to importance ratio.
    - why without replacement?
      to provide protection against the worst case scenario where one θ has enormous "importance"

Models for robust inference and sensitivity analysis
Approximation based on importance weights

- Importance sampling and importance resampling
  - importance sampling estimates $E(h(\theta)|y, \phi)$, importance resampling obtains approximate posterior sample
  - if there a small number of large importance weights, then both approximations are suspect

Models for robust inference and sensitivity analysis

Approximation based on importance weights

- Accuracy and efficiency of importance sampling estimates
  - no method exists for assessing how accurate the importance resampling (or reweighted) draws are as an approximation of the posterior distribution
  - check distribution of importance ratios to assess quality of estimate
  - performance depends on variability in importance ratios
  - estimates will often be poor if the largest ratios are too large relative to the others
  - note small importance ratios are not a problem
    (they have little influence on $E[h(\theta)|\phi, y]$)

Regression Models

- Notation
  Given a sample of size n
  - y_i - response or outcome variable for unit i
  - $y = (y_1, \ldots, y_n)'$
  - $x_i = (x_{i1}, \ldots, x_{ik})$ - explanatory variables for unit i; usually $x_{i1} \equiv 1 \ \forall i$
  - $X = n \times k$ matrix of predictors

Regression Models
Justification of conditional modeling

- Full model for (y, X)

$$p(y, X|\theta, \psi) = p(X|\psi)p(y|X, \theta)$$

- Posterior distribution for (θ, ψ)

$$p(\psi, \theta|X, y) \propto p(X|\psi)p(y|X, \theta)p(\psi, \theta)$$

- If ψ and θ are independent in their prior distribution, i.e. $p(\psi, \theta) = p(\psi)p(\theta)$, then

$$p(\psi, \theta|X, y) = p(\psi|X)p(\theta|X, y)$$

- We can analyze the second factor by itself with no loss of information

$$p(\theta|X, y) \propto p(\theta)p(y|X, \theta)$$

- Note: If the explanatory variables X are set by experimenter, then $p(X)$ is known, and there are no parameters ψ; this also justifies conditional modeling

Regression Models

- Goal: statistical inference for the parameters θ, conditional on X and y
- Since everything is conditional on X, we'll suppress it in subsequent notation
- Modeling issues
  1. defining X and y so that the conditional expectation of y given X is reasonably linear as a function of X
  2. setting up a prior distribution on the model parameters that acccurately reflects substantive knowledge,

Regression Models

Normal ordinary linear regression model

- Assumptions

$$y|\beta, \sigma^2 \sim N(X\beta, \sigma^2 I)$$

where I is the $n \times n$ identity matrix

  - the distribution of y given X is a normal r.v. whose mean is a linear function of X

$$E(y_i|\beta, X) = (X\beta)_i = \beta_1 x_{i1} + \ldots + \beta_k x_{ik}$$

  - $\text{Var}(y|\beta, \sigma^2) = \sigma^2 I$
    - can think of $y - X\beta$ as "errors"
    - observation errors are independent
    - observation errors are constant variance

Regression Models

Normal ordinary linear regression model

- Standard noninformative prior distribution

$$p(\beta, \sigma^2 | X) \propto \sigma^{-2}$$

- Posterior distribution

$$\begin{aligned}
p(\beta, \sigma^2 | y) &\propto p(y|\beta, \sigma^2) p(\beta, \sigma^2) \\
&\propto \left[\frac{1}{\sigma}\right]^n \exp\left\{-\frac{1}{2}\frac{(y-X\beta)'(y-X\beta)}{\sigma^2}\right\}\frac{1}{\sigma^2} \\
&= \left(\frac{1}{\sigma^2}\right)^{\frac{n}{2}+1} \exp\left\{-\frac{\beta' X' X\beta - 2\beta' X' y + y' y}{2\sigma^2}\right\}
\end{aligned}$$

Completing the square gives

$$p(\beta, \sigma^2 | y) \propto \left(\frac{1}{\sigma^2}\right)^{\frac{n}{2}+1} \exp\left\{-\frac{(\beta-(X'X)^{-1}X'y)'(X'X)(\beta-(X'X)^{-1}X'y)}{2\sigma^2}\right\} \times \exp\left\{-\frac{y'y - y'X(X'X)^{-1}X'y}{2\sigma^2}\right\}$$

Regression Models

Normal ordinary linear regression model

- Posterior distribution (cont'd)

  Completing the square gives

$$p(\beta, \sigma^2|y) \propto \left(\frac{1}{\sigma^2}\right)^{\frac{n}{2}+1} \exp\left\{-\frac{(\beta-(X'X)^{-1}X'y)'(X'X)(\beta-(X'X)^{-1}X'y)}{2\sigma^2}\right\} \times \exp\left\{-\frac{y'y-y'X(X'X)^{-1}X'y}{2\sigma^2}\right\}$$

  Thus

$$p(\beta, \sigma^2|y) = \underbrace{p(\beta|\sigma^2, y)}_{N(\beta|\hat{\beta},\sigma^2(X'X)^{-1})} \times \underbrace{p(\sigma^2|y)}_{\text{Inv-}\chi^2(\sigma^2|n-k,s^2)}$$

  where

$$\hat{\beta} = (X'X)^{-1}X'y \text{ and } s^2 = \frac{(y-X\hat{\beta})'(y-X\hat{\beta})}{n-k}$$

Regression Models

Normal ordinary linear regression model

- Posterior distribution, $p(\beta, \sigma^2|y)$, is proper as long as:
  1. $n > k$
  2. $\text{rank}(X) = k$
- Sampling from the posterior distribution of (β, σ^2) Recall: $\hat{\beta} = (X'X)^{-1}X'y$ and $s^2 = \frac{(y-\hat{\beta}X)'(y-\hat{\beta}X)}{n-k}$
  Also let $V_\beta = (X'X)^{-1}$
  1. compute $\hat{\beta}$ and V_β
     (note: these calculations are not usually done with traditional matrix calculations)
  2. compute s^2
  3. draw σ^2 from $p(\sigma^2|y) = \text{Inv-}\chi^2(\sigma^2|n-k, s^2)$
  4. draw β from $p(\beta|\sigma^2, y) = N(\beta|\hat{\beta}, \sigma^2 V_\beta)$

Regression Models

Normal ordinary linear regression model

- Posterior predictive distribution for new data
  - consider new data to be collected with observed predictor matrix $\tilde{X}$; we wish to predict the outcomes, $\tilde{y}$
  - posterior predictive simulation
    - first draw (β, σ^2) from their joint posterior distn
    - then draw $\tilde{y} \sim N(\tilde{X}\beta, \sigma^2 I)$
  - posterior predictive distribution is

$$\tilde{y}|\sigma^2, y \sim N(\tilde{X}\hat{\beta}, (I + \tilde{X}V_\beta\tilde{X}')\sigma^2),$$

averaging over σ^2 gives

$$\tilde{y}|y \sim t_{n-k}(\hat{\beta}, (I + \tilde{X}V_\beta\tilde{X}')s^2)$$

Regression Models
Model checking

- Diagnostics
  - residual plots (traditional or Bayesian versions)
  - posterior predictive checks
- Problems/solutions
  - nonlinearity
    - wrong model so all inferences are suspect
    - fix by transformation and/or adding predictors (or polynomial terms)
  - nonnormality
    - inferences are not quite right (usually not terribly important since posterior distn can be nearly normal even if data are not)
    - fix by transformation or by using robust models

Regression Models

Model checking (cont'd)

- Problems/solutions (cont'd)
  - unequal variances
    - bad inferences (variance is the problem)
    - fix by generalizing the model (GLS: $y|\beta, \Sigma_y \sim N(X\beta, \Sigma_y)$ with $\Sigma_y \neq \sigma^2 I$)
    - can be solved by adding missing predictor
  - correlations
    - bad inferences (variance is the problem)
    - fix by generalizing the model (GLS)
    - can be solved by adding missing predictor (time,space)

Generalized Least Squares Model

$$y|\beta, \Sigma_y \sim N(X\beta, \Sigma_y)$$

- Possible choices for Σ_y
  - Σ_y known
  - $\Sigma_y = \sigma^2 Q_y$ with Q_y known
  - $\Sigma_y = f(\sigma^2, \phi)$ i.e. a function of some unknown parameters beyond σ^2
- Posterior distribution for special case Σ_y known
  Let $y^* = \Sigma_y^{-1/2} y$ then

  $$y^*|\beta \sim N(\Sigma_y^{-1/2} X\beta, I)$$

  Hence

  $$p(\beta|y) = N(\hat{\beta}, V_\beta) \quad \text{where} \quad \begin{aligned} \hat{\beta} &= (X'\Sigma_y^{-1}X)^{-1}X'\Sigma_y^{-1}y \\ V_\beta &= (X'\Sigma_y^{-1}X)^{-1} \end{aligned}$$

Generalized Least Squares Model

- Special case: $\Sigma_y = Q_y\sigma^2$ (with Q_y known)

$$p(\beta, \sigma^2|y) = p(\sigma^2|y)p(\beta|y, \sigma^2)$$

$$p(\sigma^2|y) = \text{Inv-}\chi^2\left(n-k, \frac{(y - X\hat{\beta})'Q_y^{-1}(y - X\hat{\beta})}{n-k}\right)$$

$$p(\beta|y, \sigma^2) = N(\hat{\beta}, \sigma^2 V_\beta)$$

  where

$$\hat{\beta} = (X'Q_y^{-1}X)^{-1}X'Q_y^{-1}y$$

  and

$$V_\beta = (X'Q_y^{-1}X)^{-1}$$

- Note: prediction can be harder in this case since must account for possible correlation between $\tilde{y}$ and y in $Q_{y,\tilde{y}}$

Generalized Least Squares Model

- General case (σ^2 included inside Σ_y perhaps with other parameters also)
  - prior distn:
    $p(\beta, \Sigma_y) = p(\Sigma_y) \underbrace{p(\beta|\Sigma_y)}_{\text{flat}} \propto p(\Sigma_y)$
  - joint posterior distn:
    $p(\beta, \Sigma_y|y) \propto p(\Sigma_y) N(y|X\beta, \Sigma_y)$
  - factor joint posterior distn:
    $p(\beta, \Sigma_y|y) = p(\Sigma_y|y) \underbrace{p(\beta|\Sigma_y, y)}_{N(\beta|\hat{\beta}, V_\beta)}$

    where $\hat{\beta} = (X'\Sigma_y^{-1}X)^{-1}X'\Sigma_y^{-1}y$ and $V_\beta = (X'\Sigma_y^{-1}X)^{-1}$
    - the hard part here is $p(\Sigma_y|y)$:

$$
\begin{aligned}
p(\Sigma_y|y) &= \frac{p(\beta, \Sigma_y|y)}{p(\beta|\Sigma_y, y)} \propto \left. \frac{p(\Sigma_y) N(y|X\beta, \Sigma_y)}{N(\beta|\hat{\beta}, V_\beta)} \right|_{\beta=\hat{\beta}} \\
&= |V_\beta|^{-\frac{1}{2}} p(\Sigma_y) |\Sigma_y|^{-\frac{1}{2}} e^{-\frac{1}{2}(y - X\hat{\beta})'\Sigma_y^{-1}(y - X\hat{\beta})}
\end{aligned}
$$

Prior information

- Suppose $y|\beta, \sigma^2 \sim N(X\beta, \sigma^2 I)$
- Conjugate analysis
  - conjugate prior distribution

$$p(\beta, \sigma^2) = p(\sigma^2)p(\beta|\sigma^2) = \text{Inv-}\chi^2(\sigma^2|n_0, \sigma_0^2) \times N(\beta|\beta_0, \sigma^2\Sigma_0)$$

  - posterior distribution

$$p(\beta|\sigma^2, y) = N(\beta|\tilde{\beta}, V_\beta)$$

$$p(\sigma^2|y) = \text{Inv-}\chi^2(\sigma^2|n + n_0, \phi)$$

  where

$$\begin{aligned}
\tilde{\beta} &= (\Sigma_0^{-1} + X'X)^{-1}(\Sigma_0^{-1}\beta_0 + (X'X)\hat{\beta}) \\
V_\beta &= \sigma^2(\Sigma_0^{-1} + X'X)^{-1} \\
\phi &= (n-k)s^2 + n_0\sigma_0^2 \\
&\quad +(\hat{\beta} - \beta_0)'\Sigma_0^{-1}(\Sigma_0^{-1} + X'X)^{-1}X'X(\hat{\beta} - \beta_0) \\
\hat{\beta} &= (X'X)^{-1}(X'y) \\
s^2 &= (y - X\hat{\beta})'(y - X\hat{\beta})/(n-k)
\end{aligned}$$

Prior information

- Suppose $y|\beta,\sigma^2 \sim N(X\beta,\sigma^2 I)$
- Semi-conjugate analysis
  - prior distribution

$$p(\beta,\sigma^2) = p(\sigma^2)p(\beta) = \text{Inv-}\chi^2(\sigma^2|n_0,\sigma_0^2) \times N(\beta|\beta_0,\Sigma_0)$$

  - posterior distribution

$$\begin{aligned} p(\beta|\sigma^2,y) &= N(\beta|\tilde{\beta},V_\beta) \\ p(\sigma^2|y) &= p(\beta,\sigma^2|y)/p(\beta|\sigma^2,y) \quad \text{(a 1-dim grid)} \end{aligned}$$

    where

$$\begin{aligned} \tilde{\beta} &= (\Sigma_0^{-1} + \sigma^{-2}X'X)^{-1}(\Sigma_0^{-1}\beta_0 + \sigma^{-2}(X'X)\hat{\beta}) \\ V_\beta &= (\Sigma_0^{-1} + \sigma^{-2}X'X)^{-1} \end{aligned}$$

New view of prior information

- Consider prior information for a single regression coefficient β_j of the form $\beta_j \sim N(\beta_{j0}, \sigma^2_{\beta_j})$ with β_{j0} and $\sigma^2_{\beta_j}$ known
- Mathematically equivalent to $\beta_{j0} \sim N(\beta_j, \sigma^2_{\beta_j})$
- Prior can be viewed as "additional data"
- Can write $y^*|\beta, \Sigma^* \sim N(X^*\beta, \Sigma^*)$ with

$$y^* = \begin{pmatrix} y \\ \beta_{j0} \end{pmatrix} \quad X^* = \begin{bmatrix} X \\ J \end{bmatrix} \quad \Sigma^* = \begin{bmatrix} \Sigma_y & 0 \\ 0 & \sigma^2_{\beta_j} \end{bmatrix}$$

where $J = (0, \ldots, 0, \underbrace{1}_{j}, 0, \ldots, 0)$

- Posterior distn is $p(\beta, \Sigma^*|y) \propto p(\Sigma_y) N(y^*|\beta, \Sigma^*)$ (last term is product of two normal distns)
- If $\sigma^2_{\beta_j} \to +\infty$, the added "data point" has no effect on inference
- If $\sigma^2_{\beta_j} = 0$, the added "data point" fixes β_j exactly to β_{j0}

New view of prior information

- Same idea works for prior distn for the whole vector β if $\beta \sim N(\beta_0, \Sigma_\beta)$ with β_0, Σ_β known
- Treat the prior distribution as k prior "data points"
- Write $y_* \sim N(X_*\beta, \Sigma_*)$ with

$$y_* = \begin{pmatrix} y \\ \beta_0 \end{pmatrix} \quad X_* = \begin{bmatrix} X \\ I_k \end{bmatrix} \quad \Sigma_* = \begin{bmatrix} \Sigma_y & 0 \\ 0 & \Sigma_\beta \end{bmatrix}$$

- Posterior distn is

$$p(\beta, \Sigma_*|y) \propto p(\Sigma_y) \times \underbrace{N(y_*|X_*\beta, \Sigma_*)}_{N(y|X\beta,\Sigma_y)N(\beta|\beta_0,\Sigma_\beta)}$$

- If some of the components of β have infinite variance (i.e. noninformative prior distributions), they should be excluded from these added "prior" data points
- The joint prior distribution for β is proper if all k components have proper prior distributions; i.e. $\text{rank}(\Sigma_\beta) = k$

Hierarchical Linear Models

- Motivation - combine hierarchical modeling ideas with regression framework
- Useful way to handle
  - random effects
  - units that can be considered at two or more levels (students in classes in schools)
- General Notation
  - Likelihood for n data points

    $$y|\beta, \Sigma_y \sim N(X\beta, \Sigma_y)$$

    (often $\Sigma_y = \sigma^2 I$)
  - Prior distn on J regression coefficients

    $$\beta|\alpha, \Sigma_\beta \sim N(X_\beta \alpha, \Sigma_\beta)$$

    (often $X_\beta = 1$ and $\Sigma_\beta = \sigma^2_\beta I$)
  - Hyperprior distribution on K parameters α

    $$\alpha|\alpha_0, \Sigma_\alpha \sim N(\alpha_0, \Sigma_\alpha)$$

    with α_0, Σ_α known (often assume $p(\alpha) \propto 1$)

Hierarchical Linear Models

Example: J regression expts

- Model for jth experiment is $y_j|\underline{\beta}_j, \sigma_j^2 \sim N(X_j\underline{\beta}_j, \sigma_j^2 I)$ where $y_j = (y_{1j}, y_{2j}, \ldots, y_{n_j j})$
- The regressions can be viewed as a single model

$$y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_J \end{pmatrix} \quad X = \begin{pmatrix} X_1 & 0 & \cdots & 0 \\ 0 & X_2 & \cdots & 0 \\ \vdots & & \ddots & \vdots \\ 0 & \cdots & \cdots & X_J \end{pmatrix} \quad \beta = \begin{pmatrix} \underline{\beta}_1 \\ \underline{\beta}_2 \\ \vdots \\ \underline{\beta}_J \end{pmatrix}$$

- Hierarchy involves setting a prior distn for $\underline{\beta}_j$'s, often $\underline{\beta}_j|\underline{\alpha}, \Sigma_\beta \sim N(\underline{\alpha}, \Sigma_\beta)$
- Also need hyperpriors, e.g., $p(\underline{\alpha}, \Sigma_\beta) \propto 1$, $\sigma_j^2 \sim \text{Inv-}\chi^2(c, d)$
- Implied model is

$$y_j|\underline{\alpha}, \sigma_j^2, \Sigma_\beta \sim N(X_j\underline{\alpha}, \sigma_j^2 I + X_j'\Sigma_\beta X_j)$$

- The hierarchy introduces correlation in the distn of y_j

Hierarchical Linear Models

Other examples

- SAT coaching example (a.k.a. 8 schools)

$$y|\underline{\beta}, \sigma^2 \sim N\left[I_8\underline{\beta}, \begin{pmatrix} \sigma_1^2 & & \\ & \ddots & \\ & & \sigma_8^2 \end{pmatrix}\right]$$

$$\underline{\beta}|\alpha, \sigma_\beta^2 \sim N(\underline{1}\alpha, \sigma_\beta^2 I_8)$$

- Animal breeding

$$y|\beta, u, \sigma^2 \sim N(\underbrace{X\beta + Zu}_{\begin{pmatrix} X \\ Z \end{pmatrix}' \begin{pmatrix} \beta \\ u \end{pmatrix}}, \sigma^2 I)$$

$$u|\sigma_\alpha^2 \sim N(0, \sigma_\alpha^2 A) \qquad p(\beta) \propto 1$$

Hierarchical Linear Models

Random effects to introduce correlation

- More about how random effects introduce correlation by considering two models
- Model 1 - introduces correlation directly
- Model 2 - introduces correlation through hierarchical model

Hierarchical Linear Models

Random effects to introduce correlation (cont'd)

- Model 1
  - n_j obs from group/cluster j
  - expect objects in a group to be correlated
  - assume $y_j = (y_{1j}, y_{2j}, \ldots, y_{n_j j})' | \alpha, A_j \sim N(\alpha \underline{1}, A_j)$ where

$$A_j = \begin{pmatrix} \sigma^2 & \rho\sigma^2 & \cdots & \rho\sigma^2 \\ \rho\sigma^2 & \sigma^2 & \cdots & \rho\sigma^2 \\ \vdots & & \ddots & \vdots \\ \rho\sigma^2 & \rho\sigma^2 & & \sigma^2 \end{pmatrix}_{n_j \times n_j}$$

  - combine data into single model with correlated observations so that $y = (y_1, y_2, \ldots, y_J)' | \alpha, \Sigma_y \sim N(\alpha \underline{1}, \Sigma_y)$ where

$$\Sigma_y = \begin{pmatrix} A_1 & 0 & \cdots & 0 \\ 0 & A_2 & \cdots & 0 \\ \vdots & & \ddots & \vdots \\ 0 & 0 & & A_J \end{pmatrix}$$

Hierarchical Linear Models

Random effects to introduce correlation (cont'd)

- Model 2 – Let

$$X = \begin{pmatrix} 1_{n_1} & 0 & \cdots & 0 \\ 0 & 1_{n_2} & \cdots & 0 \\ & & \ddots & \vdots \\ & & & 1_{n_J} \end{pmatrix} \quad \text{and} \quad \beta = \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_J \end{pmatrix}$$

  and assume

$$\left.\begin{array}{l} y|\beta, \tau^2 \sim N(X\beta, \tau^2 I) \\ \\ \beta|\alpha, \tau_\beta^2 \sim N(\alpha\underline{1}, \tau_\beta^2 I) \end{array}\right\} \Rightarrow y|\alpha, \tau^2, \tau_\beta^2 \sim N\left(\alpha\underline{1}, \begin{pmatrix} B_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & B_J \end{pmatrix}\right)$$

  where

$$B_j = \begin{pmatrix} \tau^2 + \tau_\beta^2 & \cdots & \tau_\beta^2 \\ \vdots & \ddots & \vdots \\ \tau_\beta^2 & \cdots & \tau^2 + \tau_\beta^2 \end{pmatrix}_{n_j \times n_j}$$

- Model 1 = Model 2 (with $\sigma^2 = \tau^2 + \tau_\beta^2$ and $\rho = \frac{\tau_\beta^2}{\tau^2 + \tau_\beta^2}$)

Hierarchical Linear Models
Computation

- Recall

$$
\begin{array}{ll}
\text{“likelihood”} & y|\beta, \Sigma_y \sim N(X\beta, \Sigma_y) \\
\text{“population distribution”} & \beta|\alpha, \Sigma_\beta \sim N(X_\beta \alpha, \Sigma_\beta) \\
\text{“hyperprior distribution”} & \alpha|\alpha_0, \Sigma_\alpha \sim N(\alpha_0, \Sigma_\alpha)
\end{array}
$$

- Interpretation as a single linear regression

$$y_*|X_*, \gamma, \Sigma_* \sim N(X_*\gamma, \Sigma_*)$$

where

$$y_* = \begin{pmatrix} y \\ 0 \\ \alpha_0 \end{pmatrix} \quad X_* = \begin{pmatrix} X & 0 \\ I & -X_\beta \\ 0 & I \end{pmatrix} \quad \gamma = \begin{pmatrix} \beta \\ \alpha \end{pmatrix}$$

$$\Sigma_* = \begin{pmatrix} \Sigma_y & 0 & 0 \\ 0 & \Sigma_\beta & 0 \\ 0 & 0 & \Sigma_\alpha \end{pmatrix}$$

$$p(\alpha, \beta, \Sigma_y, \Sigma_\beta|y) = N(y|X\beta, \Sigma_y)N(\beta|X_\beta\alpha, \Sigma_\beta)N(\alpha|\alpha_0, \Sigma_\alpha)$$

Hierarchical Linear Models
Computation (cont'd)

- Interpretation on previous slide builds on the fact that the two sides of each equality below are the same distn statement

$$N(\alpha|\alpha_0, \Sigma_\alpha) = N(\alpha_0|\alpha, \Sigma_\alpha)$$

$$N(\beta|X_\beta\alpha, \Sigma_\beta) = N(0|\beta - X_\beta\alpha, \Sigma_\beta)$$

- Drawing samples from the posterior distribution
  - $p(\alpha, \beta|\Sigma_y, \Sigma_\beta, y)$ is the posterior distn for a linear regression model with known error variance matrix which is

$$N((X_*'\Sigma_*^{-1}X_*)^{-1}(X_*'\Sigma_*^{-1}y_*), (X_*'\Sigma_*^{-1}X_*)^{-1})$$

  - need $p(\Sigma_y, \Sigma_\beta|y)$ to complete the joint posterior distn or $p(\Sigma_y, \Sigma_\beta|\alpha, \beta, y)$ for Gibbs sampling
  - hard to describe this last step in general because of the many possible models
- Presidential Election example

Study design in Bayesian analysis

- Naive view: data collection doesn't matter for Bayesian inference
- Example where data collection doesn't matter
  - observe 9 successes in 24 trials
    design 1: 24 Bernoulli trials
    design 2: sample until you get 9 successes
  - $p(\theta|y) \propto \theta^9(1-\theta)^{15}p(\theta)$ is the same for both designs
- Example where data collection does matter
  - observe 9 successes, unknown number of trials
    design 1: 24 Bernoulli trials
    design 2: wait for 100 failures
  - $p(\theta|y)$ surely depends on design

Study design in Bayesian analysis

- Study design is important
  - pattern of what is observed can be informative
  - ignorable designs (studies where design doesn't effect inference) are likely to be less sensitive to assumptions. note: randomization is useful to Bayesians as a tool for producing ignorable designs
  - data one *could have* observed can help us to build models (causality)

Study design in Bayesian analysis
General framework

- View the world in terms of observed data and complete data, where complete data includes observed and "missing" values

	"Observed data"	"Complete data"
Sampling	Values for the n units in the sample	Values for all N units in the population
Experiment	Outcomes under the observed treatment for each unit treated	Outcomes under all treatments for all units
Rounded data	Rounded observations	Precise values of all observations
Unintentional missing data	Observed data values	Complete data, both observed and missing

Formal models for data collection

Notation

- Data

$$y = (y_1, \ldots, y_N)$$

where $y_i = (y_{i1}, y_{i2}, \ldots, y_{in})$ = data for the ith unit.

- Indicators for observed values

$$I = (I_1, \ldots, I_N)$$

where

$$I_{ij} = \begin{cases} 1 & \text{if } y_{ij} \text{ is observed} \\ 0 & \text{otherwise} \end{cases}$$

where y_{ij} is the jth variable for the ith unit.
Let $obs = \{i, j : I_{ij} = 1\}$ index the observed components of y and $mis = \{i, j : I_{ij} = 0\}$ index the unobserved components of y. Then y can be writen as $y = (y_{obs}, y_{mis})$.

Formal models for data collection

- Stability assumption
  Measurement process (I) doesn't effect the data (y) (this assumption fails if, for example, there are carryover effects or treatments in soil leak out)
- Fully observed covariates x
  We use the notation x for variables that are fully observed for all units. We might want to include x in an analysis for the following reasons:
  - we may be interested in some aspect of the joint distribution of (x, y)
  - we may be interested in some of the distribution of y, but x provides information about y
  - even if we are only interested in y, we must include x in the analysis if x is involved in the data collection mechanism

Formal models for data collection

- Complete-data model

$$p(y, I|x, \theta, \phi) = p(y|x, \theta)p(I|x, y, \phi)$$

  - $p(y|x, \theta)$ models the underlying data without reference to the data collection process
  - The estimands of primary interest are
    - functions of the complete data y (finite-population estimands)
    - functions of the parameters θ (superpopulation estimands)
  - The parameters ϕ that index the missingness are not generally of scientific interest
  - θ and ϕ can be related but this is rare

Formal models for data collection

- We don't observe all of y
- Observed-data likelihood

$$\begin{aligned} p(y_{obs}, I|x, \theta, \phi) &= \int p(y, I|x, \theta, \phi)\, d\, y_{mis} \\ &= \int p(y|x, \theta) p(I|x, y, \phi)\, d\, y_{mis} \end{aligned}$$

- Posterior distributions
  - joint posterior distribution of (θ, ϕ)

$$\begin{aligned} p(\theta, \phi|x, y_{obs}, I) &\propto p(\theta, \phi|x) p(y_{obs}, I|x, \theta, \phi) \\ &= p(\theta, \phi|x) \int p(y, I|x, \theta, \phi)\, d\, y_{mis} \\ &= p(\theta, \phi|x) \int p(y|x, \theta) p(I|x, y, \phi)\, d\, y_{mis} \end{aligned}$$

  - marginal posterior distribution of θ

$$p(\theta|x, y_{obs}, I) = p(\theta|x) \int \int p(\phi|x, \theta) p(y|x, \theta) p(I|x, y, \phi)\, d\, y_{mis}\, d\, \phi$$

Note: We don't have to perform the integrals above; as usual we can simulate treating y_{mis}, θ, and ϕ as unknowns

Ignorability

It is tempting to ignore data collection issues I and focus on

$$\begin{aligned} p(\theta|x, y_{obs}) &= p(\theta|x)p(y_{obs}|x, \theta) \\ &= p(\theta|x) \int p(y|x, \theta)\, d\, y_{mis} \end{aligned}$$

When the missing data pattern supplies no information; that is, when

$$p(\theta|x, y_{obs}) = p(\theta|x, y_{obs}, I)$$

we say that the study design or data collection mechanism is *ignorable* (with respect to the proposed model)

Ignorability

When do we get ignorability?

First, some terminology

- Missing at random (MAR)

$$p(I|x, y, \phi) = p(I|x, y_{obs}, \phi)$$

  - whether a value is missing doesn't depend on value it would have had
  - the state of being missing is allowed to depend on observed values but not on unobserved values
- Missing completely at random (MCAR)

$$p(I|x, y, \phi) = p(I|\phi)$$

- Distinct parameters

$$p(\phi|\theta, x) = p(\phi|x)$$

Ignorability

- If MAR and distinct parameters, then

$$
\begin{aligned}
p(\theta|x, y_{obs}, I) &= p(\theta|x) \int\int p(\phi|x,\theta) p(y|x,\theta) p(I|x,y,\theta) d\, y_{mis}\, d\,\phi \\
&= p(\theta|x) \int p(y|x,\theta) d\, y_{mis} \underbrace{\int p(\phi|x) p(I|x, y_{obs}, \phi) d\,\phi}_{\text{no info about } \theta} \\
&\propto p(\theta|x) p(y_{obs}|x,\theta) \\
&\propto p(\theta|y_{obs}, x)
\end{aligned}
$$

we get ignorability

Formal models for data collection
Example 1

Weigh object 100 times with $y|\theta \sim N(\theta, 1)$
Scale works with probability ϕ so that $p(I_i = 1|y, \phi) = \phi$

- Complete data

$$p(y, I|\theta, \phi) = \prod_{i=1}^{100} N(y|\theta, 1) \prod_{i=1}^{100} \phi^{I_i}(1-\phi)^{1-I_i}$$

- Observed data

$$\begin{aligned} p(y_{obs}, I|\theta, \phi) &= \int \prod_{i=1}^{100} N(y|\theta, 1) \prod_{i=1}^{100} \phi^{I_i}(1-\phi)^{1-I_i} d\, y_{mis} \\ &= \phi^{\sum_i I_i}(1-\phi)^{100-\sum_i I_i} \prod_{i=1}^{100} N(y_i|\theta, 1)\chi(\{I_i = 1\}) \end{aligned}$$

where $\chi(A)$ is the indicator function of the event A.

Formal models for data collection

Example 1 (cont'd)

- Observed data

$$p(y_{obs}, I|\theta,\phi) = \int \prod_{i=1}^{100} N(y|\theta,1) \prod_{i=1}^{100} \phi^{I_i}(1-\phi)^{1-I_i} d\, y_{mis}$$

$$= \phi^{\sum_i I_i}(1-\phi)^{100-\sum_i I_i} \prod_{i=1}^{100} N(y_i|\theta,1)\chi(\{I_i=1\})$$

because

$$\prod_{i=1}^{100} \int N(y_i|\theta,1)\,\chi(\{I_i=0\})\,dy_i = 1$$

where $\chi(A)$ is the indicator function of the event A.

- The data collection mechanism is ignorable

Formal models for data collection

Example 2

Weigh object 100 times with $y|\theta \sim N(\theta, 1)$
Scale fails if weight $> \phi$ with ϕ unknown

- Complete data

$$p(y, I|\theta, \phi) = \prod_{i=1}^{100} N(y_i|\theta, 1) \prod_{i=1}^{100} \chi(A_i)$$

where $A_i = \{\{I_i = 1\} \cap \{y_i < \phi\}\} \cup \{\{I_i = 0\} \cap \{y_i > \phi\}\}$

- Observed data

$$\begin{aligned} p(y_{obs}, I|\theta, \phi) &= \int p(y, I|\theta, \phi) d\, y_{mis} \\ &= \prod_{i=1}^{100} N(y_i|\theta, 1)\chi(\{I_i = 1\}) \\ &\quad \times \prod_{i=1}^{100} \chi(\{I_i = 0\}) \underbrace{\int N(y_i|\theta, 1)\chi(\{y_i > \phi\}) d\, y_i}_{\Phi(\theta - \phi) = P(y_i > \phi)} \end{aligned}$$

- This censored data collection mechanism is not ignorable

Bayesian Statistics - Summary

- Model building
  - basic probability distns as building blocks
  - hierarchical structure
  - condition on covariates to get ignorable designs
- Posterior inference
  - the power of simulation
  - flexible inference for any quantity of interest
  - use of decision for formal problem-solving
- Model checking
  - model checking/model selection
  - importance of checking with all available info
  - sensitivity analysis

Bayesian Statistics - Pro/Con

- Advantages
  - account for uncertainty
  - combine information from multiple sources
  - probability is the language of uncertainty
  - usual straightforward how to proceed with model development
  - flexible inference and model extensions
- Disadvantages
  - need for prior distn (importance of sensitivity analysis)
  - always requires a formal model (except for Bayesian nonparametrics)
  - high dimensional nuisance parameters (e.g., in survival analysis)
  - communication with practitioners

Bayesian Statistics - Final thoughts

- There are differences between Bayesian methods and traditional procedures
- Both will give reasonable data analyses in good hands
- Bayesians can be interested in frequency properties of procedures
- No need to declare as a Bayesian or Frequentist now (or ever)
- Goal of course has been exposure to the fundamental concepts and methods of Bayesian data analysis