A Workshop on Bayesian Nonparametric Regression Analysis

George Karabatsos
University of Illinois-Chicago

Methodological Illustration Presentation (90 min.)
Modern Modeling Methods (M³) Conference
University of Connecticut
Neag School of Education
Wednesday, May 21, 2014. 1:00 - 2:30pm
Supported by NSF-MMS Research Grant SES-1156372
Workshop Outline

I. Introduction (20 minutes).

A. Regression is ubiquitous in applied social science research.

B. Review linear regression by least-squares (LS).

C. Review Bayesian approach to linear regression: Prior, Data, Posterior (Denison et al. 2002).
   1. Bayes estimate of $(\beta, \sigma^2)$: Combination of LS estimate of $(\beta, \sigma^2)$, and information from the prior $\pi(\beta, \sigma^2)$.
   2. MCMC methods to estimate posterior, $\pi(\beta, \sigma^2 | \text{Data})$.
   3. Extensions of Bayesian linear model for binary regression (probit and logit), ordinal regression, and HLM/multilevel model.
Workshop Outline (continued)

I. Introduction (20 minutes) (Continued).

D. Why Bayesian nonparametric (BNP) regression?

*Compared to linear models, BNP regression models:*

1. Provides more modeling flexibility for real data sets that are typically more complex than what can be adequately described by a linear model.

2. Provides a richer approach to data analysis, compared to linear models which focus only on the mean of $Y \mid x$. A BNP regression model provides inference as to how the distribution of $Y \mid x$ depends on $x$, and/or provide flexible modeling of random effects.

3. BNP regression model has an infinite number of parameters. The estimation of the parameters of these models, from data, can be adequately handled by the Bayesian approach to statistical inference.
Workshop Outline (continued)

II. BNP regression models (30 min.)

A. General Overview of Bayesian Inference and MCMC.

1. Prior-to-posterior updating.
2. Posterior predictive inference.
3. Posterior predictive model fit evaluation.
4. MCMC methods, and convergence assessment.
II. BNP regression models (30 min.)

B. BNP Regression, based on infinite mixtures

1. General BNP (density) regression model
   a) General Dependent Dirichlet process (DDP)
   b) Dirichlet process (DP)
   c) General stick-breaking process
   d) Other normalized random measures

2. ANOVA/Linear DDP model (De Iorio et al. 2004)
3. Infinite probits model (Karabatsos & Walker, 2012b)
4. Infinite probits model, with automatic covariate selection
III. How to apply BNP regression models (40 minutes), for data analysis, using my Bayesian Regression Software.

A. PIRLS data (students nested within schools)
   1. Dependent variable: reading achievement. Covariates of the student, teacher, classroom, & school.
   2. Binary regression, and ordinal regression.
   3. Multilevel BNP regression analysis via the DDP.
   4. Causal analysis, regression discontinuity design (effect of class size on student achievement).

B. Larynx data: Censored survival data (survival analysis).
C. School Calendar data: Meta-analysis.
D. Item response data (NAEP). BNP-IRT analysis.

(All data applications above are time-permitting).

IV. Conclusions.
Regression and the Social Sciences

• Social science research questions usually have the form: How does $Y$ depend on $x = (x_1, \ldots, x_p)$?

• Examples of such questions:
  – Association of $Y$ and $X_1$, given $X_2 = x_2, \ldots, X_p = x_p$?
  – Association of $Y$ and $(X_1, X_2)$ given $X_3 = x_3, \ldots, X_p = x_p$?
  – Which of $(X_1, \ldots, X_p)$ are important predictors of $Y$?
  – How item responses $Y_{ij}$ relate to examinee ability and to item difficulty? As in Item Response Theory (IRT). ($x$: Indicators (0,1) of examinees; Indicators (0,−1) of items.)
  – How is $Y$ causally effected by $X_1$?
  – How is $Y$ causally effected by $X_1$, given $X_2 = x_0$?
  – …with some $Y$ observations censored (survival analysis)
Regression and the Social Sciences

- Social science research questions usually have the form:
  How does $Y$ depend on $x = (x_1, \ldots, x_p)$?

- Often of interest to investigate this dependence based on one or more features of the distribution of $Y$:
  - Mean of $Y$ (common approach; as in linear regression)
  - Variance of $Y$ (for variance regression)
  - Median of $Y$ (for median regression)
  - Quantiles of $Y$ (for quantile regression; incl. median reg.)
  - Survival function of $Y$ (for survival analysis)
  - Hazard function of $Y$ (for survival analysis)
  - p.d.f. of $Y$ (for density regression)
  - c.d.f. of $Y$ (for density regression)
**Linear Regression (least-squares)**

- **Linear model:** 
  \[ y_i = x_i^\top \beta + \varepsilon_i, \quad \varepsilon_i \sim N(0, \sigma^2), \quad i = 1, \ldots, n; \]

- i.e., 
  \[ f(y_i \mid x_i; \zeta) = n(y_i \mid x_i^\top \beta, \sigma^2), \quad i = 1, \ldots, n. \]

**Notation:**  
- \( x_i^\top \beta = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p \quad (x = (1, x_1, \ldots, x_p)^\top) \)

- \( \sigma^2 \) is variance of the regression errors (the \( \varepsilon_i \)).

- \( N(\mu, \sigma^2) \) is normal distribution, with mean and variance parameters, \( (\mu, \sigma^2) \), and with probability density function (p.d.f.) \( n(\cdot \mid \mu, \sigma^2) \) (p.d.f is a bell-shaped curve).

- \( f(y \mid x; \zeta) \) is the **likelihood** of \( y \) conditionally on \( x \), and given all model parameters, \( \zeta \).

For linear regression model, \( \zeta = (\beta, \sigma^2) \).
Linear Regression (least-squares)

• Linear model: \( y_i = x_i^\top \beta + \varepsilon_i, \quad \varepsilon_i \sim N(0, \sigma^2), \quad i = 1, \ldots, n; \)
  
i.e., \( f(y_i \mid x_i; \zeta) = n(y_i \mid x_i^\top \beta, \sigma^2), \quad i = 1, \ldots, n. \)

• Given sample data \( \mathcal{D}_n = \{(y_i, x_i)\}_{i=1:n} \), the maximum-likelihood estimate (MLE) of \((\beta, \sigma^2)\) is defined by:

\[
(\hat{\beta}, \hat{\sigma}^2) = \arg \max_{(\beta, \sigma^2) \in \mathbb{R}^{p+1} \times \mathbb{R}_+} \prod_{i=1}^n n(y_i | x_i^\top \beta, \sigma^2)
\]

• For linear model, MLE given by Ordinary Least-Squares (OLS)

\[
\hat{\beta} = (X^\top X)^{-1} X^\top y \quad \text{and} \quad \hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n (y_i - x_i^\top \hat{\beta})^2
\]

where: \( X = ((1, x_i^\top))_{n \times (p+1)} \) and \( y = (y_1, \ldots, y_n)^\top. \)
Bayesian Linear Regression Model

\[ f(y_i \mid x_i; \zeta) = n(y_i \mid x_i^T \beta, \sigma^2), \quad i = 1, \ldots, n \]

\[ \beta \mid \sigma^2 \sim N_{p+1}(m, \sigma^2 V) \]

\[ \sigma^2 \sim IG(a, b) \quad (a = \text{shape}, \ b = \text{rate}) \]

• **Prior density:** \( \pi(\beta, \sigma^2) = n_{p+1}(\beta \mid m, \sigma^2 V)ig(\sigma^2 \mid a, b) \)

\( n_{p+1}(\cdot \mid m, \Sigma) \) density function of \((p + 1)\)-variate normal \( N_{p+1}(0, \Sigma) \)

\( ig(\cdot \mid a, b) \) density funct. of inverse-gamma distribution \( IG(a, b) \).

• The data, \( D_n = \{(y_i, x_i)\}_{i=1:n} \), update the prior \( \pi(\beta, \sigma^2) \) to a …

**Posterior density:**

\[ \pi(\beta, \sigma^2 \mid D_n) = \frac{\prod_{i=1}^{n} n(y_i \mid x_i^T \beta, \sigma^2) \pi(\beta, \sigma^2)}{\int \prod_{i=1}^{n} n(y_i \mid x_i^T \beta, \sigma^2) d\Pi(\beta, \sigma^2)} \]

• Posterior distribution (density), \( \pi(\beta, \sigma^2 \mid D_n) \), conveys the probable values of the model parameters \( (\beta, \sigma^2) \), given the data \( (D_n) \) and prior \( \pi(\beta, \sigma^2) \).
Bayesian Linear Regression Model

- Can estimate posterior distribution (density) $\pi(\beta, \sigma^2 \mid D_n)$ by constructing a Markov (MCMC) chain $\{(\beta, \sigma^2)^{(s)}\}_{s=1:S}$.
- First, initialize with $\beta^{(0)} = \beta$, $\sigma^{2(0)} = \sigma^2$ and stage $s = 1$.
- Then this chain is constructed by sampling/drawing from the following full conditional posterior distributions (f.c.p.d.s), in the following Sampling Steps:

**Sampling Step 1.** Draw $\beta \mid X, y, \sigma^2 \sim N(m^*, \sigma^2 V^*)$.
**Sampling Step 2.** Draw $\sigma^2 \mid X, y, \beta \sim IG(a^*, b^*)$.
Then set $(\beta, \sigma^2)^{(s)} = (\beta, \sigma^2)$, $s = s + 1$, and repeat Steps if $s \leq S$.

- Notation: $V^* = (V^{-1} + X^\top X)^{-1}$, $m^* = V^*(V^{-1} m + X^\top y)$, $a^* = a + n/2$, $b^* = b + (1/2)(m^\top V^{-1} m + y^\top y - (m^*)^\top (V^*)^{-1} m^*)$.
- As $S \to \infty$, the chain $\{(\beta, \sigma^2)^{(s)}\}_{s=1:S}$ converges to samples from the posterior distribution (density), $\pi(\beta, \sigma^2 \mid D_n)$. 

---

12
Recall that the prior density of the linear model is:

$$\pi(\beta, \sigma^2) = n_{p+1}(\beta \mid \mathbf{m}, \sigma^2 \mathbf{V}) \text{ig}(\sigma^2 \mid a, b).$$

If the prior density is “non-informative” or “neutral,” i.e., with \(\mathbf{m} = \mathbf{0}, \mathbf{V}^{-1} \to \mathbf{0},\) \(a = b = 0,\)
so that the density \(\pi(\beta, \sigma^2)\) is flat over the range of \((\beta, \sigma^2),\)
then the mode of the posterior density \(\pi(\beta, \sigma^2 \mid \mathcal{D}_n)\)
coincides with the OLS/MLE estimator, \((\hat{\beta}, \hat{\sigma}^2),\) and

$$\mathbf{m}^* = \hat{\beta} = \mathbb{E}[\beta \mid \mathcal{D}_n] = \mathbf{V}^* \mathbf{X}^\top \mathbf{y}, \quad \text{with } \mathbf{V}^* = (\mathbf{X}^\top \mathbf{X})^{-1}.$$  

Then sampling distribution of \(\hat{\beta}\) (frequentist statistics)
coincides with the posterior distribution of \(\beta,\) given \(\hat{\sigma}^2\)

$$\hat{\beta} \mid \hat{\sigma}^2 \sim \mathcal{N}(\mathbf{m}^*, \hat{\sigma}^2 \mathbf{V}^*)$$
Extensions of Bayesian Linear Model

Extension to a binary dependent variable, \( Y \in \{0, 1\} \):

**Likelihood:**
\[
f(y \mid x; \zeta) = \Phi(x^\top \beta / \sigma)^y [1 - \Phi(x^\top \beta / \sigma)]^{1-y}
\]
with \( \Phi(\cdot) = \Pr[Y^* < \cdot] \).

**Probit model:** \( \Phi(\cdot) \) is c.d.f. of normal \( N(0,1) \) distribution.

**Logit model:** \( \Phi(\cdot) = \exp(\cdot) / [1 + \exp(\cdot)] \) is Logistic(0,1) c.d.f.

MCMC estimation methods are the same as in the linear model, after adding a MCMC sampling step that draws:

\[
y_i^* \mid \mathcal{D}_n, \ldots \sim N(x_i^\top \beta, \sigma^2) \mathbf{1}(y_i^* > 0)^y \mathbf{1}(y_i^* < 0)^{1-y}, \ i = 1, \ldots, n
\]
and after replacing \( y \) with \( y^* = (y_1^*, \ldots, y_i^*, \ldots, y_n^*)^\top \).

**Probit model:** Fix \( \sigma^2 \equiv 1 \) (Albert & Chib, 1992)

**Logit model:** Add another MCMC step to sample \( \sigma^2 \) from the Kolmogorov-Smirnov distribution (Holmes & Held, 2006).
Extensions of Bayesian Linear Model

Extension to ordinal dependent variable, \( Y \in \{ c = 0, 1, \ldots, C \} \):

**Likelihood, Cumulative logits or probits model:**

\[
f(y \mid x; \zeta) = \prod_c [\Phi(\{ w_c - x^\top \beta \}/\sigma) - \Phi(\{ w_{c-1} - x^\top \beta \}/\sigma)] \mathbb{1}(y = c)
\]

Add prior: \( \pi(w) \propto \mathbf{1}(\infty \equiv w_{-1} < w_0 = 0 < w_1 < w_2 < \ldots < w_C \equiv \infty) \) for the cutoff parameters, to the linear model.

Then MCMC estimation methods are the same as in the linear model, after adding MCMC sampling steps to draw from the following f.c.p.d.s:

\[
y_i^* \mid D_n, \ldots \propto N(x_i^\top \beta, \sigma^2) \prod_c \mathbb{1}(w_{c-1} < y_i^* < w_c) \mathbb{1}(y_i = c) \quad (\text{for } i = 1, \ldots, n)
\]

\[
w_c \mid D_n, \ldots \propto \prod_{i=1:n} N(y_i^* \mid x_i^\top \beta, \sigma^2) \mathbb{1}(w_{c-1} < w_c < w_{c+1}) \quad (c = 1, \ldots, C-1)
\]

and after replacing \( y \) with \( y^* = (y_1^*, \ldots, y_i^*, \ldots, y_n^*)^\top \).
Extension to a censored dependent variable response, $Y_i$

Suppose that a continuous-valued response $y_i$ (e.g. log survival time) is censored, i.e., known only to fall within a fixed interval $A_i = [Y_{LBi}, Y_{UBi}]$.

- **Interval censored:** $-\infty < Y_{LBi} < Y_{UBi} < \infty$
- **Right censored:** $-\infty < Y_{LBi} < Y_{UBi} = \infty$
- **Left censored:** $-\infty = Y_{LBi} < Y_{UBi} < \infty$

**Likelihood of censored response**

$f(y_i \mid x_i; \zeta) = \Phi(\{Y_{UBi} - x_i^\top \beta\} / \sigma) - \Phi(\{Y_{LBi} - x_i^\top \beta\} / \sigma)$

MCMC estimation methods are the same as in the linear model, after adding a MCMC sampling step to draw from the f.c.p.d.:

$y_i^* \mid \mathcal{D}_n, \ldots \sim N(x_i^\top \beta, \sigma^2) \mathbf{1}(y_i^* \in A_i)$

and after replacing $y_i$ with $y_i^*$
Extensions of Bayesian Linear Model

Bayesian HLM / Multilevel model:

Likelihood: \( f(y \mid x; \zeta) = n(y \mid x^\top \beta + z^\top u_h, \sigma^2), \)
for Level 2 groups \( h = 1, \ldots, H. \)

adding priors: \( u_h \mid T \sim iid \ N(0, T), \) and \( T \sim \pi(T), \)
to the linear model.

Then MCMC estimation methods are the same as in the linear model, after adding sampling steps in the MCMC algorithm to draw from the following f.c.p.d.s:

\[
\begin{align*}
    u_h \mid D_n, T, \ldots & \propto \prod_{i=1:n} N(y_i \mid x_i^\top \beta + z_i^\top u_h, \sigma^2) N(u_h \mid 0, T) \\
    T \mid D_n, \ldots & \propto \prod_{h=1:H} N(u_h \mid 0, T) \pi(T). \quad \text{(for } h = 1, \ldots, H) 
\end{align*}
\]
Why Nonparametric?

Why Bayesian nonparametric (BNP) regression?

• **Modeling flexibility needed for real data sets (typically complex)**
  - Simple linear/HLM model $\rightarrow$ stronger (less credible) data assumptions $\rightarrow$ inaccurate statistical inferences.
  - More flexible BNP regression model $\rightarrow$ weaker (more credible) data assumptions $\rightarrow$ more accurate statistical inferences.

• **Provides a richer approach to data analysis**
  - Linear model/HLM: How mean $Y$ changes with $x$?
  - BNP regression: How $Y$ changes with $x$?
    For mean, variance, p.d.f. (density regression), quantiles (quantile regression), survival functions of $Y$?
  - BNP-HLM: Supports all random effect distributions. Also provides a model-based cluster analysis.
Blue circles: Data points \((n = 565)\)

Solid red line: Normal linear model fit. OLS estimate of \( E[Y | x] \)
Dashed lines: +/- 2 time error s.d. \(\sigma\) estimate
Probability density function (p.d.f.) estimates:

**Red:** Normal linear model p.d.f.

**Blue:** Estimate of true p.d.f.
Why Bayes Nonparametric?

Why Bayes?

- A BNP regression model, provides a very flexible and rich approach to data analysis, by employing an infinite number of parameters.

- The estimation of the parameters of these models, from data, can be handled by the Bayesian approach to statistical inference.

- Hence, before we describe BNP regression models, we need to give a brief overview of the general approach to Bayesian statistical inference and MCMC.
Overview of Bayesian Modeling

• In general, consider a model \((m)\) with likelihood density function \(f(\cdot \mid x; \zeta_m)\), parameter \(\zeta_m \in \Omega_{\zeta_m} \subseteq \mathbb{R}^K\), and a prior probability density \(\pi_m(\zeta_m)\) assigned over the space \(\Omega_{\zeta_m}\).

• Data, \(D_n = \{(y_i, x_i)\}_{i=1:n}\), update the prior \(\pi_m(\zeta_m)\) to a …

**Posterior density:** \(\pi(\zeta_m \mid D_n) = \frac{\prod_{i=1}^{n} f(y_i \mid x_i; \zeta_m) \pi_m(\zeta_m)}{\int \prod_{i=1}^{n} f(y_i \mid x_i; \zeta_m) d\Pi_m(\zeta_m)}\)

with \(\Pi(\zeta_m \mid D_n) = \Pr[Z < \zeta]\) the c.d.f. of \(\pi(\zeta_m \mid D_n)\).

• Posterior density \(\pi_m(\zeta_m \mid D_n)\) conveys the probable values of \(\zeta_m\), given data \(D_n = \{(y_i, x_i)\}_{i=1:n}\), and prior, \(\pi_m(\zeta_m)\).
Overview of Bayesian Modeling

• Model predictions from the **prior predictive density**:

\[
f(y \mid x, m) = \int f(y \mid x; \zeta_m) d\Pi_m(\zeta_m)
\]

• \( f(y \mid x, m) \) is likelihood of \( y \) given model \( m \) and prior \( \Pi_m(\zeta_m) \).

• **Bayes factor** \((B_{12})\): Evidence in favor of model \( M_1 \) vs. model \( M_2 \):
  (Jeffreys, 1961; Kass & Raftery, 1995):

\[
B_{12} = \frac{\Pr(D_n \mid M_1)}{\Pr(D_n \mid M_2)} = \frac{\prod_{i=1}^{n} f(y_i \mid x_i; \zeta_1) d\Pi_1(\zeta_1)}{\prod_{i=1}^{n} f(y_i \mid x_i; \zeta_2) d\Pi_2(\zeta_2)} = \frac{\Pr(M_1 \mid D_n) / \Pr(M_2 \mid D_n)}{\Pr(M_1) / \Pr(M_2)}
\]

<table>
<thead>
<tr>
<th>Table of Evidence</th>
<th>( B_{12} )</th>
<th>( 2 \log B_{12} )</th>
<th>Evidence in favor of ( M_1 ) vs. ( M_2 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 to 3</td>
<td>0 to 2</td>
<td>Not worth more than a bare mention</td>
<td></td>
</tr>
<tr>
<td>3 to 20</td>
<td>2 to 6</td>
<td>Positive</td>
<td></td>
</tr>
<tr>
<td>20 to 150</td>
<td>6 to 10</td>
<td>Strong</td>
<td></td>
</tr>
<tr>
<td>&gt; 150</td>
<td>&gt; 10</td>
<td>Very strong</td>
<td></td>
</tr>
</tbody>
</table>
Overview of Bayesian Modeling

• Model predictions from the **posterior predictive density**:

\[
f_n(y \mid x, m) = \int f(y \mid x; \zeta_m) d\Pi(\zeta_m \mid D_n)
\]

\[
c.d.f. \quad F_n(y \mid x, m) = \int_{y \leq y} f(y \mid x; \zeta_m) d\Pi(\zeta_m \mid D_n)
\]

Model Predictive...

**Mean**  \[ E_n[Y \mid x, m] = \int y \, dF_n(y \mid x, m) \]

**Variance**  \[ \text{Var}_n[Y \mid x] = \int (y - E_n[Y \mid x])^2 dF_n(y \mid x, m) \]

**c.d.f.**  \[ F_n(y \mid x, m) = \Pr_n(Y \leq y \mid x, m) \]

**Median**  \[ F_n^{-1}(0.5 \mid x, m) \]

**Quantile**  \[ F_n^{-1}(u \mid x, m), \text{ for a choice } u \in [0, 1] \]

**Survival function**  \[ 1 - F_n(y \mid x, m) \]

**Hazard function**  \[ f_n(y \mid x, m) / [1 - F_n(y \mid x, m)] \]

etc.
Overview of Bayesian Modeling

• Partial dependence (PD) method (Friedman, 2001):
  Let \( x_S \in x \) denote 1 or 2 covariates of interest to study predictions, with \( x = x_S \cup x_C \) and \( x = x_S \cap x_C = \emptyset \).

• PD posterior predictive:

| Mean       | \( E_n(y|x_S) = \frac{1}{n} \sum_{i=1}^{n} E_n(y|x_S, x_{Ci}) \) |
|------------|---------------------------------------------------------------|
| Variance   | \( \text{Var}_n(y|x_S) = \frac{1}{n} \sum_{i=1}^{n} \text{Var}_n(y|x_S, x_{Ci}) \) |
| Density    | \( f_n(y|x_S) = \frac{1}{n} \sum_{i=1}^{n} f_n(y|x_S, x_{Ci}) \) |
| c.d.f.     | \( F_n(y|x_S) = \frac{1}{n} \sum_{i=1}^{n} F_n(y|x_S, x_{Ci}) \) |
| Median     | \( F_n^{-1}(0.5|x_S) = \frac{1}{n} \sum_{i=1}^{n} F_n^{-1}(0.5|x_S, x_{Ci}) \) |
| Quantile   | \( F_n^{-1}(u|x_S) = \frac{1}{n} \sum_{i=1}^{n} F_n^{-1}(u|x_S, x_{Ci}), \ u \in [0, 1] \) |
| Survival   | \( 1 - F_n(y|x_S) = \frac{1}{n} \sum_{i=1}^{n} [1 - F_n(y|x_S, x_{Ci})] \) |

etc.

• Then can study such PD predictions over a range of \( x_S \), via scatterplots, Trellis plots, etc.
Overview of Bayesian Modeling

- Measures of the predictive fit of a model \((m)\), for data \(D_n\)

**Response \(y_i\) fit:** Standardized residual \((z_i)\).

\[
z_i = \frac{y_i - E_n(Y \mid x_i, m)}{\sqrt{\text{Var}_n(Y \mid x_i, m)}}^2
\]

\(y_i\) is an outlier when \(|z_i| > 2\).

**Model fit:** \(D(m)\) criterion

\[
D(m) = \sum_{i=1}^{n} \left\{y_i - E_n(Y \mid x_i, m)\right\}^2 + \sum_{i=1}^{n} \text{Var}_n(Y \mid x_i, m)
\]

Predictive mean-square. First term measures data fit. Second term is penalty; increases with model complexity. (Laud & Ibrahim, 1995; Gelfand & Ghosh, 1998)

Among a set of \(M\) models compared, \(\{m = 1, \ldots, M\}\), the model with the best predictive fit for the data is identified as the one with the smallest value of \(D(m)\).
Bayesian Modeling and MCMC

- Posterior density of model:
  \[
  \pi(\zeta_m | \mathcal{D}_n) = \frac{\prod_{i=1}^{n} f(y_i | x_i; \zeta_m) \pi_m(\zeta_m)}{\int \prod_{i=1}^{n} f(y_i | x_i; \zeta_m) d\Pi_m(\zeta_m)}
  \]

- In typical applications of Bayesian modeling, \( \pi(\zeta | \mathcal{D}_n) \) cannot be solved directly, esp. when dimensionality of \( \zeta \) is high and/or when the prior \( \pi(\zeta) \) is not conjugate to likelihood \( f(\cdot | x; \zeta) \).
- To infer the posterior \( \pi(\zeta | \mathcal{D}_n) \), and all posterior functionals of interest, can use Markov chain Monte Carlo (MCMC),
- MCMC involves constructing a discrete-time Harris ergodic Markov chain \( \{\zeta^{(s)}\}_{s=1:S} \) having stationary distribution \( \Pi(\zeta | \mathcal{D}_n) \).
- Ergodicity ensured by a proper \( \pi(\zeta) \) (Robert & Casella, 2004).
- As \( S \to \infty \), chain \( \{\zeta^{(s)}\}_{s=1:S} \) converges to samples from the posterior \( \Pi(\zeta | \mathcal{D}_n) \).
Bayesian Modeling and MCMC

- Posterior density of model:
  \[
  \pi(\zeta_m \mid D_n) = \frac{\prod_{i=1}^n f(y_i \mid x_i; \zeta_m) \pi_m(\zeta_m)}{\int \prod_{i=1}^n f(y_i \mid x_i; \zeta_m) d\Pi_m(\zeta_m)}
  \]

- A MCMC chain \( \{\zeta^{(s)}\}_{s=1:S} \) can be constructed by sampling blocks \( \zeta_b \), for \( b = 1, \ldots, B \), of \( \zeta \), repeatedly for \( S \) times. (\( \bigcup_b \zeta_b = \zeta \) and \( \zeta_b \cap \zeta_c = \emptyset \) for all \( b \neq c \)).

- Each stage \( s \), sample \( \zeta^{(s)}_{m_b} \sim \pi(\zeta_{m_b} \mid D_n, \zeta_{m/b}) \), for \( b = 1, \ldots, B \), with
  \[
  \zeta_{m/b} = (\zeta^{(s)}_{m1}, \ldots, \zeta^{(s)}_{m,b-1}, \zeta^{(s-1)}_{m,b+1}, \ldots, \zeta^{(s-1)}_{mB}).
  \]

- **Gibbs sampler**: For sampling known \( \pi(\zeta_b \mid D_n, \zeta_{m/b}) \).

- **Rejection sampler**: For sampling \( \pi(\zeta_b \mid D_n, \zeta_{m/b}) \), when we only know:
  \[
  \pi(\zeta_{m_b} \mid D_n, \zeta_{m/b}) \propto \prod_{i=1:n} f(y_i \mid x_i; \zeta_{m/b}) \pi(\zeta_{m_b} \mid \zeta_{m/b}).
  \]
  Metropolis-Hastings, adaptive rejection, or slice sampling.
Bayesian Modeling and MCMC

• Given a sampled MCMC chain, \( \{ \zeta^{(s)} \}_{s=1:S} \), check whether \( S \) is “large enough” for the chain to provide samples from the posterior distribution (density), \( \pi(\zeta | D_n) \), by performing a:

MCMC Convergence Analysis (2 steps; Geyer 2011):

1. **Trace plots:** For each univariate parameter \( \zeta \in \zeta \) of interest, do a trace plot of MCMC samples \( \{ \zeta^{(s)} \}_{s=1:S} \) to verify _good mixing_; i.e., that the samples appear to adequately explore the posterior \( \pi(\zeta | D_n) \). Plot should look stable and hairy.

2. **95% Monte Carlo Confidence intervals (MCCIs):**
   For each parameter, and its posterior estimate, of interest, use overlapping batch means and/or sub-sampling methods (Flegal & Jones, 2011) to calculate and verify that the 95% MCCI is small enough for practical purposes.

• If trace plots and MCCIs both not satisfactory, then run additional MCMC samples until they are both satisfactory.
BNP Regression Modeling

• A flexible and rich approach to BNP regression is provided by Bayesian Density regression models.

• Bayesian density regression is a relatively new research area that has already seen many important developments.

• Nearly all proposed Bayesian density regression models have the general form:

\[
f(y|x; \xi) = \int f(y|x, \psi, \Theta) dG_x(\Theta) = \sum_{j=1}^{\infty} f(y|x, \psi, \theta_j(x)) \omega_j(x)
\]
General BNP Model

\[ f(y \mid x; \zeta) = \int f(y \mid x, \psi, \theta) \, dG_x(\theta) = \sum_{j=1}^{\infty} f(y \mid x, \psi, \theta_j(x)) \omega_j(x) \]

\{f(\cdot \mid x, \psi, \theta) : (\theta, \psi) \in \Theta\} \text{ chosen family of parametric densities}  
\text{(kernel densities)}

\[ G_x(\cdot) = \sum_{j=1}^{\infty} \omega_j(x) \delta_{\theta_j(x)}(\cdot) \]  
\text{mixing distribution}

\[ \omega_j(x) \]  
\text{mixing weights} \text{ that sum to 1 at every } x \in \mathcal{X}

\[ \delta_{\theta(x)}(\cdot) \]  
probability measure degenerate at \( \theta(x) \)

\{\omega_j(x)\}_{j=1,2,...} \text{ and } \{\theta_j(x)\}_{j=1,2,...} \text{ are infinite collections of processes indexed by } \mathcal{X}

\psi  
Other model parameters, not part of the mixture
The Bayesian density regression model is completed by the specification of a prior distribution on parameters:

\[ f(y|x; \zeta) = \int f(y|x, \psi, \theta)dG_x(\theta) = \sum_{j=1}^{\infty} f(y|x, \psi, \theta_j(x)) \omega_j(x) \]

- The Bayesian density regression model is completed by the specification of a prior distribution on parameters:

\[ \zeta = (\psi, (\omega_j(x), \theta_j(x)))_{j=1,2,...}, \; x \in X \]

- Current research aims to construct such prior distributions that guarantee large supports for \( f(y|x; \zeta) \).

- The general model encompasses linear models, normal LMMs/HLMs, GLMs, GLMMs, finite mixture latent class models, hierarchical mixtures-of-experts regression models, and mixtures of Gaussian process regressions.
Examples, General BNP Model

$$f(y | x; \zeta) = \int f(y | x, \psi, \theta) dG_x(\theta) = \sum_{j=1}^{\infty} f(y | x, \psi, \theta_j(x)) \omega_j(x)$$

- DDP prior is denoted by: $$G_x \sim \text{DDP}(\alpha, G_{0x})$$.
- A random distribution $$G_x$$ from a DDP is constructed by:
  $$G_x(\cdot) = \sum_{j=1}^{\infty} \omega_j(x) \delta_{\theta_j(x)}(\cdot)$$

  with stick-breaking weights:
  $$\omega_j(x) = v_j(x) \prod_{l=1}^{j-1} (1 - v_l(x)), \quad j = 1, 2, \ldots,$$

  $$v_j \sim Q_j, \quad v_j : \mathcal{X} \to [0,1],$$

  and atoms $$\theta_j(x) \sim_{\text{ind}} G_{0x}$$ (Sethuraman, 1994)
The ordinary Dirichlet process (DP; Ferguson, 1973), is not covariate-dependent, and is denoted by:

$$G \sim DP(\alpha, G_0),$$

where the random distribution function:

$$G(\cdot) = \sum_{j=1}^{\infty} \omega_j \delta_{\theta_j}(\cdot)$$

is constructed by "covariate independent", stick breaking weights:

$$\omega_j = v_j \prod_{l=1}^{j-1} (1 - v_j), \quad v_j \sim iid \ Beta(1, \alpha),$$

and atoms

$$\theta_j \sim iid \ G_0, \quad j = 1,2,\ldots$$

Under DP(\alpha, G_0),

$$E[G(\cdot)] = G_0,$$

$$\text{Var}[G(\cdot)] = \left\{ G(\cdot)[1-G(\cdot)] \right\} / (\alpha + 1).$$
Examples, General BNP Model

\[ f(y|\mathbf{x}; \zeta) = \int f(y|\mathbf{x}, \psi, \theta) dG(\theta) = \sum_{j=1}^{\infty} f(y|\mathbf{x}, \psi, \theta) \omega_j \]

- The more general nonparametric, *stick-breaking process*,
  \[ G \sim \text{Stick-Breaking}((a_j, b_j)_{j=1,2,...}, G_0), \]
  (Ishwaran & James, 2001)

for the stick-breaking mixture weights,

\[ \omega_j = v_j \prod_{l=1}^{j-1} (1 - v_j) \]

lets:

\[ v_j \sim \text{ind} \ \text{Beta}(a_j, b_j), \ j = 1,2,... \]

- Special cases of the stick-breaking process:
  \[ G \sim \text{DP}(\alpha, G_0), \ \text{when} \ a_j = 1 \ \text{and} \ b_j = \alpha > 0. \]
  \[ G \sim \text{PitmanYor}(a, b, G_0) \ \text{when} \]
  \[ a_j = 1 - a \ \text{and} \ b_j = b + ja, \ 0 \leq a < 1 \ \text{and} \ b > -a. \]
ANOV A/Linear DDP Mixture Model

\[ f(y|x; \zeta) = \int f(y|x, \psi, \theta) dG_x(\theta) = \sum_{j=1}^{\infty} f(y|x, \psi, \theta_j(x)) \omega_j \]

A multi-level model

\[ (y_{i(h)})_{i(h)=1}^{n_h} | X_h \sim f(y_h | X_h), \ h = 1, \ldots, N_h \]

\[ f(y_h | X_h) = \sum_{j=1}^{\infty} \left\{ \prod_{i(h)=1}^{n_h} n(y_{i(h)} | x_{i(h)}^T \beta_j, \sigma^2) \right\} \omega_j \]

\[ \omega_j = v_j \prod_{l=1}^{j-1} (1 - v_l) \]

\[ v_j | \alpha \sim \text{Beta}(1 - a, \alpha + aj) \]

\[ \beta_j | \mu, T \sim \text{N}(\mu, T) \]

\[ \sigma^2 \sim \text{IG}(a_0/2, a_0/2) \]

\[ \mu, T \sim \text{N}(\mu | 0, r_0 I_{p+1}) \IW(T | p + 3, s_0 I_{p+1}) \]

\[ \alpha \sim \text{Ga}(a_\alpha, 1/b_\alpha) \]

De Iorio, Müller, Rosner, & MacEachern (2004)
While the ANOVA/linear DDP model, as represented in the previous slide, is a mixture model with mixing distribution:

\[ G \sim \text{Stick-Breaking}((a_j, b_j)_{j=1,2,...}, G_0). \]

the model is identical to a model with mixing distribution

\[ G_x \sim \text{ANOVA-DDP}((a_j, b_j)_{j=1,2,...}, G_0) \]

with \( G_x(\theta) = \sum_{j=1}^{\infty} \omega_j \delta_{\theta_j(x)}(\theta), \quad \theta_j(x) = x^\top \beta_j, \)

\[ \beta_j \mid \mu, \Sigma \sim_{\text{iid}} G_0 = N(\mu, \Sigma), \quad \text{and normal kernel } n(y \mid \theta, \sigma^2). \]

De Iorio, Müller, Rosner, & MacEachern (2004, p. 208).
Infinite Probits (IP) Model (v1)

\[
f(y|\mathbf{x}; \zeta) = \int f(y|\mathbf{x}, \psi, \theta) dG_{\mathbf{x}}(\theta) = \sum_{j=1}^{\infty} f(y|\mathbf{x}, \psi, \theta_j) \omega_j(\mathbf{x})
\]

\[
y_i|\mathbf{x}_i \sim f(y|\mathbf{x}_i), \quad i = 1, \ldots, n
\]

\[
f(y|\mathbf{x}) = \sum_{j=-\infty}^{\infty} n(y|\mu_j, \sigma_j^2) \omega_j(\mathbf{x})
\]

\[
\omega_j(\mathbf{x}) = \Phi\left( \frac{j - \mathbf{x}^T \beta_\omega}{\exp(\mathbf{x}^T \lambda_\omega)^{1/2}} \right) - \Phi\left( \frac{j - 1 - \mathbf{x}^T \beta_\omega}{\exp(\mathbf{x}^T \lambda_\omega)^{1/2}} \right)
\]

\[
\mu_j|\mu, \sigma_\mu^2 \sim N(\mu, \sigma_\mu^2)
\]

\[
\sigma_j^2|\tau \sim IG(1, \tau)
\]

\[
\mu, \sigma_\mu \sim N(\mu|0, \sigma_0^2 \to \infty)U(\sigma_\mu|0, b_{\sigma\mu})
\]

\[
\tau \sim Ga(a_0/2, a_0/2)
\]

\[
(\beta_\omega, \lambda_\omega) \sim N(0, v\mathbf{I})
\]

Karabatsos & Walker (2012b)
Infinite Probits (IP) Model

- $\sigma(x)$ controls the level of multimodality of $f(y \mid x)$.

  x more informative $\rightarrow f(y \mid x)$ becomes more unimodal

  x less informative $\rightarrow f(y \mid x)$ becomes more multimodal.

39
Infinite Probits (IP) Model (v2)

\[ f(y|x; \zeta) = \int f(y|x, \psi, \theta) dG_x(\theta) = \sum_{j=1}^{\infty} f(y|x, \psi, \theta_j) \omega_j(x) \]

\[ y_i|x_i \sim f(y|x_i), \ i = 1, \ldots, n \]

\[ f(y|x) = \sum_{j=-\infty}^{\infty} n(y|\mu_j, \sigma_j^2) \omega_j(x) \]

\[ \omega_j(x) = \Phi\left( \frac{j - x^T \beta_\omega}{\exp(x^T \lambda_\omega)^{1/2}} \right) - \Phi\left( \frac{j - 1 - x^T \beta_\omega}{\exp(x^T \lambda_\omega)^{1/2}} \right) \]

\[ \mu_j|\mu, \sigma_\mu^2 \sim N(\mu, \sigma_\mu^2) \]

\[ \sigma_j^2|\tau \sim IG(1, \tau) \]

\[ (\mu, \sigma_\mu) \sim N(\mu|0, \sigma_0^2 \to \infty) U(\sigma_\mu|0, b_{\sigma\mu}) \]

\[ \tau \sim Ga(a_0/2, a_0/2) \]

\[ (\beta_\omega, \lambda_\omega)|\gamma \sim N(0, diag(\nu_1 \gamma + \nu_0(1 - \gamma))) \]

\[ (\gamma_{\beta_k}, \gamma_{\lambda_k}) \sim Ber(1/2)Ber(1/2), \ k = 0, 1, \ldots, p \]

Karabatsos & Walker (2012b)
Infinite Probits (IP) Model (v3)

\[ f(y|x; \zeta) = \int f(y|x, \psi, \theta) dG_x(\theta) = \sum_{j=1}^{\infty} f(y|x, \psi, \theta_j(x)) \omega_j \]

\[ y_i | x_i \sim f(y|x_i), \ i = 1, \ldots, n \]

\[ f(y|x) = \sum_{j=-\infty}^{\infty} n(y|\mu_j + x^T \beta, \sigma^2) \omega_j(x) \]

\[ \omega_j(x) = \Phi\left( \frac{j - x^T \beta}{\sigma_{\omega}} \right) - \Phi\left( \frac{j - 1 - x^T \beta}{\sigma_{\omega}} \right) \]

\[ \mu_j | \sigma^2_{\mu} \sim N(0, \sigma^2_{\mu}); \quad \sigma_{\mu} \sim U(0, b_{\sigma\mu}); \]

\[ \beta_0 | \sigma^2 \sim N(0, \sigma^2 v_{\beta_0} \rightarrow \infty) \]

\[ \beta_k | \sigma^2, \gamma_k \sim N(0, \sigma^2 \{v_1 \gamma_k + v_0(1 - \gamma_k)\}) \]

\[ \gamma_k \sim Ber(w_{\gamma}), \ k = 1, \ldots, p \]

\[ \sigma^2 \sim IG(a_0/2, a_0/2) \]

\[ \beta_{\omega} | \sigma^2_{\omega} \sim N(0, \sigma^2_{\omega} v I) \]

\[ \sigma^2_{\omega} \sim IG(a_{\omega}/2, a_{\omega}/2) \]

(\(\Phi(\cdot)\) is N(0,1) c.d.f.)

Karabatsos & Walker (2012b)
MCMC Sampling of BNP models

- The standard Gibbs samplers of Kalli et al. (2010) can be used to perform MCMC sampling of BNP regression models, combined with standard Gibbs sampling methods for sampling the f.c.p.d.s of parameters of ordinary linear/HLM models that are assigned conjugate priors; and possibly combined with rejection sampling methods (e.g., Metropolis or slice sampling) for sampling the f.c.p.d.s of parameters that are assigned non-conjugate priors.

- MCMC extensions to the linear DDP model for binary, ordinal, or censored $Y$, are similar to that for the linear model, after replacing $N(x_i^T\beta, \sigma^2)$ with $N(x_i^T\beta_{zi}, \sigma^2)$ for the f.c.p.d.s for the $y_i^*$s, with along with f.c.p.d. $z_i \sim N(x_i^T\beta_{zi}, \sigma^2) \omega_{zi}$.

- MCMC extensions to v1 or v2 (or v3) of the infinite probits model would instead replace with $N(\mu_{zi}, \sigma_{zi}^2)$ and f.c.p.d. $z_i \sim N(x_i^T\beta_{zi}, \sigma^2) \omega_{zi}(x)$ (with $N(\mu_{zi} + x_i^T\beta, \sigma^2)$ and $z_i \sim N(\mu_{zi} + x_i^T\beta, \sigma^2) \omega_{zi}(x)$, respectively). See Karabatsos & Walker (2012a,b).
Applying BNP regression

Applying BNP regression analysis, for data analysis, using the *Bayesian Regression* Software.

A. PIRLS data  (students nested within schools)
   1. **Dependent variable:** reading achievement. 
      **Covariates** of the student, teacher, classroom, & school.
   2. **Binary regression, and ordinal regression.**
   3. **Multilevel BNP regression analysis** via the DDP.
   4. **Causal analysis, regression discontinuity design** (effect of class size).

B. Larynx data: Censored survival data (survival analysis).

C. School Calendar data: Meta-analysis.

D. Item response data (NAEP). BNP-IRT analysis.

(All data applications above are time-permitting).
Bayesian Regression Software

- Website for *Bayesian Regression Software*:
  [www.uic.edu/~georgek/HomePage/BayesSoftware.html](http://www.uic.edu/~georgek/HomePage/BayesSoftware.html)

- **Software Requirements:**
  The software can run on either a Windows (PC) computer, a Linux computer, or on a Macintosh computer, which has installed either the
  (1) (free) MATLAB Compiler Runtime (MCR) R2013b (8.2) software for Windows 64-bit;
  [Click here if you need to download and install the free MCR R2013b (8.2) software for Windows on your computer.](http://www.uic.edu/~georgek/HomePage/BayesSoftware.html)
  (2) or has installed MATLAB version R2013b.

  [Click this link to download the Bayesian Regression software package (file BayesRegression_2014b_pkg.exe).](http://www.uic.edu/~georgek/HomePage/BayesSoftware.html) (8 MB)
Bayesian Regression Software

• For a Macintosh or Linux computer, the Bayesian Regression software can be run under a free software program that can run executable (.exe) files, such as

Wine, Wine Bottler, Darwine, VirtualBox, or Bochs PC Emulator.
Bayesian Regression Software

• After downloading the **Bayesian Regression** software onto your computer desktop, click the file (icon): BayesRegression_2014b_pkg.exe
to unpack several files, including the software file BayesRegression_2014b.exe, a readme file, a folder containing several example data files, and the software user's manual.

• The software user's manual provides step-by-step instructions on how to analyze a given data set, using a model of your choice. Any one of the example data files may be used to illustrate the Bayesian Regression Software.

• The Bayesian Regression software is opened/run by clicking the icon (file) BayesRegression_2014b.exe (it may take a few moments to load on screen).
Bayesian Regression: Nonparametric and Parametric Models

\[ y|x; \theta \sim f(y|x; \theta) \]
\[ \theta \sim \pi(\theta) \]

U.S. Copyright to George Karabatsos
Professor, University of Illinois-Chicago
Supported by NSF grant SES-1156372

Data File: Please open or import a data set from the File menu.
You may use the Import data menu option.

This will import your data set for analysis. The data set must be a comma delimited (.csv) file (variable names in first line; all data numeric).

After data is imported, the software will automatically rename the data file to have the *.DAT extension.

If you already have a data file that has the *.DAT extension, then you may just select the Open Data (*.DAT) file menu option (the first menu option).
Imported (or opened) data file appears below.
You may initially inspect the data, by running basic descriptive statistics (summaries, frequencies, and/or correlations of variables) and/or by constructing plots of the data (e.g., histograms, Q-Q plots, scatter plots, box plots, etc.).
## Univariate Descriptive Statistics

Results:

<table>
<thead>
<tr>
<th>Variable</th>
<th>VarID</th>
<th>#obs</th>
<th>#miss</th>
<th>Mean</th>
<th>SD</th>
<th>Med</th>
<th>Min</th>
<th>25%</th>
</tr>
</thead>
<tbody>
<tr>
<td>zREAD</td>
<td>2</td>
<td>100</td>
<td>0</td>
<td>-0.467</td>
<td>0.931</td>
<td>-0.525</td>
<td>-2.824</td>
<td>-1.143</td>
</tr>
<tr>
<td>MALE</td>
<td>5</td>
<td>100</td>
<td>0</td>
<td>0.510</td>
<td>0.502</td>
<td>1.000</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>AGE</td>
<td>6</td>
<td>100</td>
<td>0</td>
<td>10.214</td>
<td>0.689</td>
<td>10.080</td>
<td>9.170</td>
<td>9.830</td>
</tr>
<tr>
<td>CLSIZE</td>
<td>7</td>
<td>100</td>
<td>0</td>
<td>22.130</td>
<td>4.184</td>
<td>21.000</td>
<td>17.000</td>
<td>18.000</td>
</tr>
</tbody>
</table>

Sample size, mean, median, s.d., min, max, percentiles, etc.
Univariate descriptive statistics for variables: zREAD, MALE, AGE, CLSIZE.

Bivariate correlations (Pearson $r$ statistics).

<table>
<thead>
<tr>
<th>Variable</th>
<th>(V1)</th>
<th>(V2)</th>
<th>(V3)</th>
<th>(V4)</th>
</tr>
</thead>
<tbody>
<tr>
<td>zREAD</td>
<td>1.00</td>
<td>-0.138</td>
<td>-0.246</td>
<td>0.235</td>
</tr>
<tr>
<td>MALE</td>
<td>-0.138</td>
<td>1.00</td>
<td>0.031</td>
<td>0.117</td>
</tr>
<tr>
<td>AGE</td>
<td>-0.246</td>
<td>0.031</td>
<td>1.00</td>
<td>-0.201</td>
</tr>
<tr>
<td>CLSIZE</td>
<td>0.235</td>
<td>0.117</td>
<td>-0.201</td>
<td>1.00</td>
</tr>
</tbody>
</table>
Also, with any of these menu options, you may reformat the data:

- Perform simple variable transformations such as constructing z-scores;
- Create new variables by dummy coding, creating interaction variables;
- Impute missing values of variables;
- etc.
We will construct z-scores of 8 variables, which will be used as covariates for a BNP regression analysis that we will perform shortly.
After easily clicking through the Simple variable transformations > Z-score menu options, new variables, the z-scores of 8 variables, are instantly created.
Select a BNP regression model for data analysis.

We select the infinite probits regression model (v1)

(one of 43 possible choices of models, from the software)
Select Dependent Variable for the model:

zRead  Student literacy score. (a z-standardized version of READSC).
Select Covariates for the model:

We select the z-standardized covariates: 
Enter the Parameters of the Prior Distribution for the model:

We select the default choices.
After menu selections, the Infinite Probits regression model appears (middle screen); along with a list of the choices of prior parameters (upper left), and a list of the choices of the dependent variables and covariates (upper right of screen).
When necessary (optional):

1. If you have censored data for your dependent variable, then you may select variables that indicate the nature of the censoring; (e.g., for a survival analysis)

2. If your data observations are weighted, then you may select the variable for the observation weights (e.g., for a meta-analysis).

See the software user’s manual for more details.
Click Run Posterior Analysis button. This will generate 50K MCMC samples from the model’s posterior distribution.

Before clicking button, we entered 50000 in the MCsamples box.
The analysis is in progress:

A wait bar appears, indicating the current number of MCMC sampling iterations run (out of 50K), the time elapsed, the time left to 50K MCMC iterations, and the current value of the model fit (\(D(m)\) statistic).
After end of analysis run, this text output file is generated and opened automatically.

Text display of BNP regression model.
Prior Parameters of Model

\[
\begin{align*}
    b_s_{\text{mu}} &= 5 \\
    a_0 &= 0.001 \\
    v &= 1000000
\end{align*}
\]

Variables in Model

Dependent Variable (Y):
Variables in Model

Dependent Variable (Y):
zREAD

Data Sample Size: \( n = 100 \)

Covariates:
Z:MALE
Z:AGE
Z:CLSIZE
Z:ELL
Z:TEXP4
Z:EDLEVEL
Z:ENROL
Z:SCHSAFE

Text display of chosen dependent variable and covariates for BNP regression model.
Reports of fit of the BNP regression model.

D(m) = 44.5
R-squared = .97

No outliers; all standardized residuals < |1|
Posterior estimates of the parameters of the BNP regression model.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Mean</th>
<th>Med</th>
<th>SD</th>
<th>2.5%</th>
<th>97.5%</th>
<th>Var</th>
</tr>
</thead>
<tbody>
<tr>
<td>mu_mu</td>
<td>-0.45</td>
<td>-0.45</td>
<td>3.07</td>
<td>-6.78</td>
<td>5.83</td>
<td>9.41</td>
</tr>
<tr>
<td>sigma^2_mu</td>
<td>5.56</td>
<td>3.77</td>
<td>5.19</td>
<td>0.52</td>
<td>20.49</td>
<td>26.97</td>
</tr>
<tr>
<td>tau</td>
<td>0.07</td>
<td>0.05</td>
<td>0.09</td>
<td>0.01</td>
<td>0.21</td>
<td>0.01</td>
</tr>
<tr>
<td>beta_w0</td>
<td>0.67</td>
<td>0.80</td>
<td>2.61</td>
<td>-5.23</td>
<td>5.62</td>
<td>6.81</td>
</tr>
<tr>
<td>beta_w:Z:MALE</td>
<td>-0.27</td>
<td>-0.23</td>
<td>1.91</td>
<td>-3.91</td>
<td>3.04</td>
<td>3.67</td>
</tr>
<tr>
<td>beta_w:Z:AGE</td>
<td>0.00</td>
<td>0.05</td>
<td>1.37</td>
<td>-2.71</td>
<td>2.64</td>
<td>1.87</td>
</tr>
<tr>
<td>beta_w:Z:CLSIZE</td>
<td>0.02</td>
<td>-0.08</td>
<td>1.51</td>
<td>-2.64</td>
<td>3.06</td>
<td>2.27</td>
</tr>
<tr>
<td>beta_w:Z:ELL</td>
<td>-0.15</td>
<td>-0.09</td>
<td>1.74</td>
<td>-3.47</td>
<td>2.91</td>
<td>3.03</td>
</tr>
<tr>
<td>beta_w:Z:TEXP4</td>
<td>-0.32</td>
<td>-0.19</td>
<td>1.40</td>
<td>-3.40</td>
<td>2.08</td>
<td>1.93</td>
</tr>
<tr>
<td>beta_w:Z:EDLEVEL</td>
<td>-0.13</td>
<td>-0.17</td>
<td>1.21</td>
<td>-2.36</td>
<td>2.22</td>
<td>1.48</td>
</tr>
<tr>
<td>beta_w:Z:ENROL</td>
<td>0.53</td>
<td>0.68</td>
<td>2.14</td>
<td>-3.21</td>
<td>4.18</td>
<td>4.57</td>
</tr>
<tr>
<td>beta_w:Z:SHSAFE</td>
<td>-0.16</td>
<td>-0.12</td>
<td>1.38</td>
<td>-3.05</td>
<td>2.59</td>
<td>1.90</td>
</tr>
<tr>
<td>lambda_w0</td>
<td>1.90</td>
<td>1.89</td>
<td>0.50</td>
<td>0.97</td>
<td>2.98</td>
<td>0.25</td>
</tr>
<tr>
<td>lambda_w:Z:MALE</td>
<td>0.33</td>
<td>0.34</td>
<td>0.44</td>
<td>-0.56</td>
<td>1.16</td>
<td>0.20</td>
</tr>
<tr>
<td>lambda_w:Z:AGE</td>
<td>0.03</td>
<td>0.01</td>
<td>0.49</td>
<td>-0.87</td>
<td>1.01</td>
<td>0.24</td>
</tr>
</tbody>
</table>
We need to evaluate the accuracy of the posterior estimates (reported in text output), through a MCMC convergence analysis of the 50K MCMC samples that we generated.

We can do so by clicking the Posterior Summaries button.

Recall that in analyzing the MCMC convergence, that we need to generate, for all model parameters of interest: (1) trace plots to evaluate MCMC mixing; and (2) 95% Monte Carlo confidence intervals, for posterior estimates of interest.
After clicking Posterior Summaries button, select menu option Trace Plots of MC samples.
Select the parameters that you want to view in the trace plots.
Now we evaluate the 95% Monte Carlo confidence intervals (MCCIs).
After clicking Posterior Summaries button, select menu option:
Table: Posterior Summaries, with 95% MCCIs.
For the model parameters, the 95% MCCI half-widths are near .00 to 2 decimals, for nearly all parameter estimates. Given this result, and trace plots, we conclude convergence. If convergence cannot be confirmed by either the trace plots or the 95% MCCIs, then we may generate additional MCMC samples, beyond the 50K MCMC samples already generated, in order to attain better MCMC convergence. Additional MCMC samples can be generated by clicking again the Run Posterior Analysis button of the software.
Under the model, we can investigate the posterior predictions of \( Y \) (zREAD), as a function of one or more covariates of interest.

We can do so by clicking the Posterior Predictive button.
We will choose to investigate how the mean, variance, p.d.f., survival function, and various quantiles, namely, the 10%ile, 25%ile, 50%ile (median), 75%ile, and 90%ile, of $Y$ (zREAD), varies as a function of one covariate, school enrollment (Z:ENROL).
Now select the covariate, school enrollment (Z:ENROL).
We now choose to investigate how posterior predictions of \( Y \) (zREAD) varies as a function of covariate Z:ENROL, for values of this covariate ranging from \(-1.7\) to \(1.7\), with the values in this range separated by \(0.1\). Therefore, in the text box above, we enter \(-1.7 : 0.1 : 1.7\).
When investigating how posterior predictions of $Y$ (zREAD) varies as a function of covariate Z:ENROL, we choose to make posterior predictions using the partial dependence method (Friedman, 2001).
Plot of posterior predictive $Y$ (zREAD), as a function of zENROL, in terms of the mean, 10\%ile, 25\%, 50\%, 75\%, and 90\%iles of $Y$. 
Plot of posterior predictive variance of $Y$ (zREAD), as a function of zENROL.
Plot of posterior predictive p.d.f. of $Y \ (z\text{READ})$, as a function of $z\text{ENROL}$. 
Plot of posterior predictive survival function of $Y$ (zREAD), as a function of zENROL.
Output table of specific numerical estimates of the posterior predictives of $Y$ (zREAD), as a function of zENROL. Comma-delimited output files of these results are also generated.

| ID | E[Y|x] | varE[Y|x] | 2.5%E[Y|x] | 97.5%E[Y|x] | V[Y|x] | varV[Y|x] | 2.5%V[Y] |
|----|--------|-----------|------------|-------------|--------|-----------|----------|
| 1  | -0.56  | 0.33      | -1.67      | 0.51        | 0.34   | 2.63      | 0.65     |
| 2  | -0.56  | 0.30      | -1.63      | 0.49        | 0.37   | 10.41     | 0.45     |
| 3  | -0.56  | 0.27      | -1.59      | 0.46        | 0.35   | 7.29      | 0.32     |
| 4  | -0.56  | 0.25      | -1.52      | 0.42        | 0.34   | 6.25      | 0.38     |
| 5  | -0.55  | 0.23      | -1.46      | 0.36        | 0.37   | 28.30     | 0.41     |
| 6  | -0.55  | 0.22      | -1.43      | 0.33        | 0.34   | 5.19      | 0.35     |
| 7  | -0.55  | 0.20      | -1.41      | 0.27        | 0.32   | 3.83      | 0.36     |
| 8  | -0.55  | 0.19      | -1.34      | 0.26        | 0.32   | 2.95      | 0.32     |
| 9  | -0.54  | 0.17      | -1.31      | 0.24        | 0.33   | 4.59      | 0.40     |
| 10 | -0.53  | 0.17      | -1.29      | 0.23        | 0.31   | 3.02      | 0.31     |
| 11 | -0.52  | 0.16      | -1.25      | 0.19        | 0.32   | 5.99      | 0.38     |

Also, comma-delimited output files of all posterior predictive are generated, in case that the data analyst want to further analyze or plot these results on her/his own.
Run another posterior predictive analysis, by clicking again the Posterior Predictive button of the software.

Here, plot of posterior predictive p.d.f. of $Y$ (zREAD), by gender group.

Using partial dependence method.

$z_{MALE} = -1.02$ is female, and $z_{MALE} = .98$ is male.
Plot of posterior predictive $Y$ (zREAD), for male vs. female, in terms of the mean, 10%ile, 25%, 50%, 75%, and 90%iles of $Y$. 
Conclusions

• Linear models, while often used in social science research, do not adequately describe many data sets, as they usually give rise to complex relationships between the $Y$ and $x$ variables. Then, linear models can provide misleading inferences.

• Also, ordinary linear models can only convey how the mean of $Y$ varies with $x$.

• We proposed some BNP regression models that are:
  (1) flexible enough to adequately capture any complex relationships between the $Y$ and $x$ variables;
  (2) and which provide a richer approach to data analysis, in terms of how the mean, variance, quantiles, p.d.f., survival function, etc., of $Y$, changes with $x$.

• We illustrated these models on several real data sets, using my free software: *Bayesian Regression: Nonparametric and Parametric Models* (Karabatsos, 2014a, 2014b).


   Available from http://www.uic.edu/~georgek/HomePage/BayesSoftware.html

   Available from http://www.uic.edu/~georgek/HomePage/BayesSoftware.html


