Regression models cont'd

Dr. Olanrewaju Michael Akande

April 1, 2020

Announcements

- HW7 online.

Outline

- Linear regression cont'd
  - Recap
  - Weakly informative priors
- Bayesian model selection and averaging
  - Hypothesis testing
  - Model selection and averaging
- Example

Linear regression cont'd

Linear regression model recap

- Model:

$$\boldsymbol{Y} \sim \mathrm{N}_n(\boldsymbol{X}\boldsymbol{\beta}, \sigma^2 \boldsymbol{I}_{n \times n}).$$

where $\boldsymbol{I}$ is the identity matrix and

$$\boldsymbol{Y} = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} \boldsymbol{X} = \begin{bmatrix} 1 & x_{11} & x_{12} & \dots & x_{1(p-1)} \\ 1 & x_{21} & x_{22} & \dots & x_{2(p-1)} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n1} & x_{n2} & \dots & x_{n(p-1)} \end{bmatrix} \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_{p-1} \end{bmatrix} \boldsymbol{\epsilon} = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}$$

- Priors:

$$\pi(\boldsymbol{\beta}) = \mathrm{N}_p(\boldsymbol{\beta}_0, \Sigma_0)$$

$$\pi(\sigma^2) = \mathrm{IG}\left(\frac{\nu_0}{2}, \frac{\nu_0\sigma_0^2}{2}\right).$$

Bayesian estimation recap

- With those priors, we have

$$\pi(\boldsymbol{\beta} \,|\, \boldsymbol{y}, \boldsymbol{X}, \sigma^2) \equiv \mathrm{N}_p(\boldsymbol{\mu}_n, \Sigma_n),$$

where

$$\Sigma_n = \left[\Sigma_0^{-1} + \frac{1}{\sigma^2}\boldsymbol{X}^T\boldsymbol{X}\right]^{-1}$$

$$\boldsymbol{\mu}_n = \Sigma_n\left[\Sigma_0^{-1}\boldsymbol{\beta}_0 + \frac{1}{\sigma^2}\boldsymbol{X}^T\boldsymbol{y}\right].$$

- plus

$$\pi(\sigma^2 \,|\, \boldsymbol{y}, \boldsymbol{X}, \boldsymbol{\beta}) \equiv \mathrm{IG}\left(\frac{\nu_n}{2}, \frac{\nu_n\sigma_n^2}{2}\right),$$

where

Weakly informative priors

- Specifying hyperparameters that represent actual prior information can be challenging, especially for $\boldsymbol{\beta}$.
- It can therefore be desirable use weakly informative priors when possible. The Hoff book discusses a few different options, one of which is the Zellner's g-prior (there are other options but we will not cover them in class due to time restrictions).
- Note that we can also use Jefferys prior here to be completely non-informative.
- Zellner's g-prior is

$$\pi(\boldsymbol{\beta} \mid \sigma^2) = \mathrm{N}_p\left(\boldsymbol{\beta}_0 = \mathbf{0}, \Sigma_0 = g\sigma^2 \left[\boldsymbol{X}^T\boldsymbol{X}\right]^{-1}\right)$$

$$\pi(\sigma^2) = \mathrm{IG}\left(\frac{\nu_0}{2}, \frac{\nu_0\sigma_0^2}{2}\right)$$

for some positive value g, which is often commonly set to the sample size n.

Weakly informative priors

- Note that the g-prior uses a part of the data. As I have mentioned before, using your data to construct your prior is usually a no-no!

- However, the g-prior actually does not use the information in $\boldsymbol{y}$, the response variable of interest, just the information in $\boldsymbol{X}$.

- Observe that the prior specification actually looks like the conjugate prior we first used for the univariate normal model, that is, with

$$\sigma^2 \sim \mathrm{IG}\left(\frac{\nu_0}{2}, \frac{\nu_0 \sigma_0^2}{2}\right)$$

$$\mu \,|\, \sigma^2 \sim \mathrm{N}\left(\mu_0, \frac{\sigma^2}{\kappa_0}\right).$$

- Turns out that we also have conjugacy with the g-prior, so that we don't actually need Gibbs sampling to obtain posterior samples. $\pi(\boldsymbol{\beta} \,|\, \boldsymbol{y}, \boldsymbol{X}, \sigma^2)$ takes the same form as before but now we also have $\pi(\sigma^2 \,|\, \boldsymbol{y}, \boldsymbol{X})$.

Weakly informative priors

- With the g-prior, we have

$$\pi(\boldsymbol{\beta} \,|\, \boldsymbol{y}, \boldsymbol{X}, \sigma^2) = \mathrm{N}_p(\boldsymbol{\mu}_n, \Sigma_n)$$

$$\pi(\sigma^2 \,|\, \boldsymbol{y}, \boldsymbol{X}) = \mathrm{IG}\left(\frac{\nu_n}{2}, \frac{\nu_n \sigma_n^2}{2}\right)$$

where

$$\Sigma_n = \left[\Sigma_0^{-1} + \frac{1}{\sigma^2}\boldsymbol{X}^T\boldsymbol{X}\right]^{-1} = \left[\frac{1}{g\sigma^2}\boldsymbol{X}^T\boldsymbol{X} + \frac{1}{\sigma^2}\boldsymbol{X}^T\boldsymbol{X}\right]^{-1} = \frac{g}{g+1}\sigma^2\left[\boldsymbol{X}^T\boldsymbol{X}\right]^{-1}$$

$$\begin{aligned}\boldsymbol{\mu}_n &= \Sigma_n\left[\Sigma_0^{-1}\boldsymbol{\beta}_0 + \frac{1}{\sigma^2}\boldsymbol{X}^T\boldsymbol{y}\right] = \frac{g}{g+1}\sigma^2\left[\boldsymbol{X}^T\boldsymbol{X}\right]^{-1}\left[\frac{1}{\sigma^2}\boldsymbol{X}^T\boldsymbol{y}\right] \\ &= \frac{g}{g+1}\left[\boldsymbol{X}^T\boldsymbol{X}\right]^{-1}\boldsymbol{X}^T\boldsymbol{y} = \frac{g}{g+1}\hat{\boldsymbol{\beta}}_{\text{ols}} \\ \nu_n &= \nu_0 + n; \qquad \sigma_n^2 = \frac{1}{\nu_n}\left[\nu_0\sigma_0^2 + \mathrm{SSR}(g)\right],\end{aligned}$$

Bayesian Model selection and averaging

Bayesian hypothesis testing/model selection

- How can we do model selection in a Bayesian framework? First let's quickly discuss Bayesian hypothesis testing for a simple model.

- Suppose we have univariate data $y_i \overset{iid}{\sim} \mathrm{N}(\mu, 1)$ and wish to test $\mathrm{H}_0 : \mu = 0; \quad \mathrm{vs.} \mathrm{H}_1 : \mu \neq 0$ under the Bayesian paradigm.

- Informal approach:

  1. Put a prior on μ, $\pi(\mu) = \mathrm{N}(\mu_0, \sigma_0^2)$.

  2. Compute posterior $\mu \,|\, Y = (y_1, \ldots, y_n) \sim \mathrm{N}(\mu_n, \sigma_n^2)$ for updated parameters μ_n and σ_n^2.

  3. Compute a 95% CI based on the posterior.

  4. Reject H_0 if interval does not contain zero.

Bayesian hypothesis testing

- Formal approach:

  1. Put a prior on the actual hypotheses/models, that is, on $\pi(\mathrm{H}_0) = \Pr[\mathrm{H}_0]$ and $\pi(\mathrm{H}_1) = \Pr[\mathrm{H}_1]$.

     For example, set $\Pr[\mathrm{H}_0] = 0.5$ and $\Pr[\mathrm{H}_1] = 0.5$, if apriori, we believe the two hypotheses are equally likely.

  2. Put a prior on the parameters in each model.

     In our simple normal model, the only unknown parameter is μ, so for example, our prior can once again be $\pi(\mu) = \mathrm{N}(\mu_0, \sigma_0^2)$.

  3. Compute marginal posterior probabilities for each hypothesis, that is, $\Pr[\mathrm{H}_0 | Y]$ and $\Pr[\mathrm{H}_1 | Y]$.

  4. Conclude based on the magnitude of $\Pr[\mathrm{H}_1 | Y]$ relative to $\Pr[\mathrm{H}_0 | Y]$.

Bayesian hypothesis testing

- Using Bayes theorem,

$$\Pr[\mathrm{H}_1 \mid Y] = \frac{\mathrm{L}[Y|\mathrm{H}_1]\Pr[\mathrm{H}_1]}{\mathrm{L}[Y|\mathrm{H}_0]\Pr[\mathrm{H}_0] + \mathrm{L}[Y|\mathrm{H}_1]\Pr[\mathrm{H}_1]},$$

  where $\mathrm{L}[Y|\mathrm{H}_0]$ and $\mathrm{L}[Y|\mathrm{H}_1]$ are the marginal likelihoods for the data under the null and alternative hypotheses respectively.

- If for example we set $\Pr[\mathrm{H}_0] = 0.5$ and $\Pr[\mathrm{H}_1] = 0.5$ apriori, then

$$\begin{aligned}\Pr[\mathrm{H}_1 \mid Y] &= \frac{0.5\mathrm{L}[Y|\mathrm{H}_1]}{0.5\mathrm{L}[Y|\mathrm{H}_0] + 0.5\mathrm{L}[Y|\mathrm{H}_1]} \\ &= \frac{\mathrm{L}[Y|\mathrm{H}_1]}{\mathrm{L}[Y|\mathrm{H}_0] + \mathrm{L}[Y|\mathrm{H}_1]} = \frac{1}{\frac{\mathrm{L}[Y|\mathrm{H}_0]}{\mathrm{L}[Y|\mathrm{H}_1]} + 1}.\end{aligned}$$

- The ratio $\frac{\mathrm{L}[Y|\mathrm{H}_0]}{\mathrm{L}[Y|\mathrm{H}_1]}$ is known as the Bayes factor in favor of H_0, and often written as BF_{01}. Similarly, we can compute BF_{10}.

Bayes factors

- Bayes factor: is a ratio of marginal likelihoods and it provides a weight of evidence in the data in favor of one model over another.
- It is often used as an alternative to the frequentist p-value.
- **Rule of thumb**: $\mathrm{BF}_{01} > 10$ is strong evidence for H_0; $\mathrm{BF}_{01} > 100$ is decisive evidence for H_0.
- Notice that for our example,

$$\Pr[\mathrm{H}_1 | Y] = \frac{1}{\frac{\mathrm{L}[Y|\mathrm{H}_0]}{\mathrm{L}[Y|\mathrm{H}_1]} + 1} = \frac{1}{\mathrm{BF}_{01} + 1}$$

the higher the value of BF_{01}, that is, the weight of evidence in the data in favor of H_0, the lower the marginal posterior probability that H_0 is true.

- That is, here, as $\mathrm{BF}_{01} \uparrow$, $\Pr[\mathrm{H}_1 | Y] \downarrow$.

Bayes factors

- Let's look at another way to think of Bayes factors. First, recall that

$$\Pr[H_1 \mid Y] = \frac{L[Y|H_1]\Pr[H_1]}{L[Y|H_0]\Pr[H_0] + L[Y|H_1]\Pr[H_1]},$$

so that

$$\begin{aligned}
\frac{\Pr[H_0 \mid Y]}{\Pr[H_1 \mid Y]} &= \frac{L[Y|H_0]\Pr[H_0]}{L[Y|H_0]\Pr[H_0] + L[Y|H_1]\Pr[H_1]} \div \frac{L[Y|H_1]\Pr[H_1]}{L[Y|H_0]\Pr[H_0] + L[Y|H_1]\Pr[H_1]} \\
&= \frac{L[Y|H_0]\Pr[H_0]}{L[Y|H_0]\Pr[H_0] + L[Y|H_1]\Pr[H_1]} \times \frac{L[Y|H_0]\Pr[H_0] + L[Y|H_1]\Pr[H_1]}{L[Y|H_1]\Pr[H_1]} \\
\therefore \underbrace{\frac{\Pr[H_0 \mid Y]}{\Pr[H_1 \mid Y]}}_{\text{posterior odds}} &= \underbrace{\frac{\Pr[H_0]}{\Pr[H_1]}}_{\text{prior odds}} \times \underbrace{\frac{L[Y|H_0]}{L[Y|H_1]}}_{\text{Bayes factor } BF_{01}}
\end{aligned}$$

- Therefore, the Bayes factor can be thought of as the factor by which our prior odds change (towards the posterior odds) in the light of the data.

- In linear regression, **BIC** approximates the BF comparing a model to the null model.

Bayes factors

- While Bayes factors can be appealing, calculating them can be computationally demanding!
- Why have we been "mildly obsessed" with MCMC sampling? To avoid computing any **marginal likelihoods**! Well, guess what? Bayes factors are ratios of marginal likelihoods, taking us back to the problem we always try to avoid.
- Of course this isn't all *"doom and gloom"*, there are various ways (once again!) of getting around computing those likelihoods analytically. Unfortunately, we will not have time to cover them in this course.
- As a teaser, one approach is to flip the relationship on the previous slide:

$$\underbrace{\frac{\mathrm{L}[Y|\mathrm{H}_0]}{\mathrm{L}[Y|\mathrm{H}_1]}}_{\text{Bayes factor BF}_{01}} = \underbrace{\frac{\Pr\,[\mathrm{H}_0\,|\,Y]}{\Pr\,[\mathrm{H}_1\,|\,Y]}}_{\text{posterior odds}} \times \underbrace{\frac{\Pr\,[\mathrm{H}_1]}{\Pr\,[\mathrm{H}_0]}}_{\text{prior odds}},$$

which is easy to compute as long as we can use posterior samples to compute/approximate the posterior odds.

Bayesian model selection

- Now that we have a general sense of how Bayesian hypothesis works, let's get back to model selection, and use some of the same ideas.

- General setting:

  1. Define a list of models. That is, let Γ be the "finite" set of different possible models.
  2. Each model γ is in Γ, including the "true" model. Also, let θ_γ represent the parameters in model γ.
  3. Put a prior over the set Γ. Let $\Pi_\gamma = \Pr[\gamma]$, for all $\gamma \in \Gamma$. Most common choice is the uniform prior, that is, $\Pi_\gamma = \frac{1}{\#\Gamma}$, for all $\gamma \in \Gamma$, where $\#\Gamma$ is the total number of models in Γ.
  4. Put a prior on the parameters in each model, that is, each $\pi(\theta_\gamma)$.
  5. Compute marginal posterior probabilities $\Pr[\gamma | Y]$ for each model.

Bayesian model selection

- For each model $\gamma \in \Gamma$, we need to compute $\Pr[\gamma \mid Y]$.
- Let $\mathrm{L}_\gamma(Y)$ denote the marginal likelihood of the data under model γ, that is, $\mathrm{L}[Y \mid \gamma]$ or $\mathrm{L}[Y; \gamma]$. As before,

$$\begin{aligned}\hat{\Pi}_\gamma = \Pr[\gamma \mid Y] &= \frac{\mathrm{L}_\gamma(Y)\Pi_\gamma}{\sum_{\gamma^\star \in \Gamma} \mathrm{L}_{\gamma^\star}(Y)\Pi_{\gamma^\star}} \\ &= \frac{\Pi_\gamma \cdot \left[\int_{\Theta_\gamma} \mathrm{L}_\gamma(Y \mid \theta_\gamma) \cdot \pi(\theta_\gamma) \mathrm{d}\theta_\gamma\right]}{\sum_{\gamma^\star \in \Gamma} \mathrm{L}_{\gamma^\star}(Y)\Pi_{\gamma^\star}}.\end{aligned}$$

- If we assume a uniform prior on Γ, that is, $\Pi_\gamma = \frac{1}{\#\Gamma}$, for all $\gamma \in \Gamma$, then

$$\begin{aligned}\hat{\Pi}_\gamma &= \frac{\mathrm{L}_\gamma(Y)}{\sum_{\gamma^\star \in \Gamma} \mathrm{L}_{\gamma^\star}(Y)} \\ &= \frac{\left[\int_{\Theta_\gamma} \mathrm{L}_\gamma(Y \mid \theta_\gamma) \cdot \pi(\theta_\gamma) \mathrm{d}\theta_\gamma\right]}{\sum_{\gamma^\star \in \Gamma} \mathrm{L}_{\gamma^\star}(Y)}.\end{aligned}$$

Bayesian model selection

- How should we choose the Bayes optimal model?
- First specify a loss function. The most natural is

$$L(\hat{\gamma}, \gamma) = \mathbf{1}(\hat{\boldsymbol{\gamma}} \neq \boldsymbol{\gamma}),$$

  that is,

  1. Loss equals zero if the correct model is chosen; and
  2. Loss equals one if incorrect model is chosen.

- Next, select $\hat{\gamma}$ to minimize Bayes risk. Here, Bayes risk (expected loss over posterior) is

$$R(\hat{\gamma}) = \sum_{\gamma \in \Gamma} \mathbf{1}(\hat{\boldsymbol{\gamma}} \neq \boldsymbol{\gamma}) \cdot \hat{\Pi}_\gamma = 0 \cdot \hat{\Pi}_{\gamma_{\text{true}}} + \sum_{\gamma \neq \gamma_{\text{true}}} \hat{\Pi}_\gamma = \sum_{\gamma \neq \hat{\gamma}} \hat{\Pi}_\gamma = 1 - \hat{\Pi}_{\hat{\gamma}}$$

- To minimize $R(\hat{\gamma})$, choose $\hat{\gamma}$ such that $\hat{\Pi}_{\hat{\gamma}}$ is the largest! That is, select the model with the largest posterior probability.

Inference vs prediction

- What if the goal is prediction? Then we should care more about predictive accuracy, rather than selecting specific variables.

- For predictions, we care about the predictive distribution, that is

$$
\begin{aligned}
p(y_{n+1} \mid Y = (y_1, \ldots, y_n)) &= \int\int p(y_{n+1} \mid \gamma, \theta_\gamma) \cdot \pi(\gamma, \theta_\gamma \mid Y) \mathrm{d}\theta_\gamma \mathrm{d}\gamma \\
&= \int\int p(y_{n+1} \mid \gamma, \theta_\gamma) \cdot \pi(\theta_\gamma \mid Y, \gamma) \cdot \ \Pr\left[\gamma \mid Y\right] \mathrm{d}\theta_\gamma \mathrm{d}\gamma \\
&= \sum_{\gamma \in \Gamma} \int p(y_{n+1} \mid \gamma, \theta_\gamma) \cdot \pi(\theta_\gamma \mid Y, \gamma) \cdot \hat{\Pi}_\gamma \mathrm{d}\theta_\gamma \\
&= \sum_{\gamma \in \Gamma} \hat{\Pi}_\gamma \cdot \int p(y_{n+1} \mid \gamma, \theta_\gamma) \cdot \pi(\theta_\gamma \mid Y, \gamma) \mathrm{d}\theta_\gamma \\
&= \sum_{\gamma \in \Gamma} \hat{\Pi}_\gamma \cdot \int p(y_{n+1} \mid Y, \gamma) \mathrm{d}\theta_\gamma,
\end{aligned}
$$

which is just averaging out the predictions from each model, over all possible models in Γ, with the posterior probability of each model, and this is known as Bayesian model averaging (BMA).

Back to Bayesian linear regression

- So what does this mean specifically in the context of linear regression?
- First, recall that for model γ, the posterior probability that the model is the right model is

$$\hat{\Pi}_\gamma = \frac{\Pi_\gamma L_\gamma(Y)}{\sum_{\gamma^\star \in \Gamma} \Pi_{\gamma^\star} L_{\gamma^\star}(Y)}.$$

- *Practical issues*
  - We need to calculate marginal likelihoods for ALL models in Γ.
  - In general for, we cannot calculate the marginal likelihoods unless we have a proper or conjugate priors.
  - For linear regression, that would mean looking to priors like Zellner's g-prior, the horseshoe prior you were introduced to in the lab, and so on.

Bayesian variable selection

- To explore Bayesian variable selection, rewrite each model $\gamma \in \Gamma$ as

$$\boldsymbol{Y} \sim \mathrm{N}_n(\boldsymbol{X}_\gamma \boldsymbol{\beta}_\gamma, \sigma^2 \boldsymbol{I}_{n \times n}).$$

- γ represents the set of predictors we want to throw into our model.

- Using the notation as before, each $\gamma = (\gamma_0, \gamma_1, \ldots, \gamma_{p-1}) \in \{0, 1\}^p$, so that the cardinality of Γ is 2^p, that is, the number of models in Γ.

- That is,
  - $\gamma_j = 1$ means the j'th predictor is included in the model, but $\gamma_j = 0$ means it is not;
  - $\boldsymbol{X}_\gamma$ is the matrix of predictors with $\gamma_j = 1$;
  - $\boldsymbol{\beta}_\gamma$ is the corresponding vector of predictors with $\gamma_j = 1$.

- Set $p_\gamma = \sum_{j=1}^{p} \gamma_j$, so that p_γ is the number of predictors included in model γ, then $\boldsymbol{X}_\gamma$ is $n \times p_\gamma$ and $\boldsymbol{\beta}_\gamma$ is $p_\gamma \times 1$.

Bayesian variable selection

- Recall that we can also write each model as

$$Y_i = \boldsymbol{\beta}_\gamma^T \boldsymbol{x}_{i\gamma} + \epsilon_i; \quad \epsilon_i \overset{iid}{\sim} \text{N}(0, \sigma^2).$$

- As an example, suppose we had data with 6 predictors including the intercept, so that each $\boldsymbol{x}_i = (1, x_{i1}, x_{i2}, x_{i3}, x_{i4}, x_{i5})$, and $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2, \beta_3, \beta_4, \beta_5)$.

- Then for model with $\gamma = (1, 1, 0, 0, 0, 0)$, $Y_i = \boldsymbol{\beta}_\gamma^T \boldsymbol{x}_{i\gamma} + \epsilon_i$

$$\Longrightarrow \; Y_i = \beta_0 + \beta_1 x_{i1} + \epsilon_i; \quad \epsilon_i \overset{iid}{\sim} \text{N}(0, \sigma^2),$$

  with $p_\gamma = 2$.

- Whereas for model with $\gamma = (1, 0, 0, 1, 1, 0)$, $Y_i = \boldsymbol{\beta}_\gamma^T \boldsymbol{x}_{i\gamma} + \epsilon_i$

$$\Longrightarrow \; Y_i = \beta_0 + \beta_3 x_{i3} + \beta_4 x_{i4} + \epsilon_i; \quad \epsilon_i \overset{iid}{\sim} \text{N}(0, \sigma^2),$$

  with $p_\gamma = 3$.

Bayesian variable selection

- The outline for variable selection would be as follows:

1. Write down likelihood under model γ. That is,

$$p(\boldsymbol{y} \mid \boldsymbol{X}, \gamma, \boldsymbol{\beta}_\gamma, \sigma^2) \propto (\sigma^2)^{-\frac{n}{2}} \exp\left\{ -\frac{1}{2\sigma^2} (\boldsymbol{y} - \boldsymbol{X}_\gamma \boldsymbol{\beta}_\gamma)^T (\boldsymbol{y} - \boldsymbol{X}_\gamma \boldsymbol{\beta}_\gamma) \right\}$$

2. Define a prior for γ, $\Pi_\gamma = \Pr[\gamma]$. For example, (i) uniform over all 2^p possible models, or even (ii) beta prior (since each $\gamma_j \in \{0, 1\}$).

3. Put a prior on the parameters in each model. Using the g-prior, we have

$$\pi(\boldsymbol{\beta}_\gamma \mid \sigma^2) = \mathrm{N}_p\left(\boldsymbol{\beta}_{0\gamma} = \mathbf{0}, \Sigma_{0\gamma} = g\sigma^2 \left[\boldsymbol{X}_\gamma^T \boldsymbol{X}_\gamma \right]^{-1} \right)$$

$$\pi(\sigma^2) = \mathrm{IG}\left(\frac{\nu_0}{2}, \frac{\nu_0 \sigma_0^2}{2} \right)$$

Bayesian variable selection

- With those pieces, the conditional posteriors are straightforward.
- We can then compute marginal posterior probabilities $\Pr[\gamma | Y]$ for each model and select model with the highest posterior probability.
- We can also compute posterior $\Pr[\gamma_j | Y]$, the posterior probability of including the j'the predictor, often called marginal inclusion probability (MIP), allowing for uncertainty in the other predictors.
- Also straightforward to do model averaging once we all have posterior samples.
- The Hoff book works through one example and you can find the Gibbs sampler for doing inference there. I strongly recommend you go through it carefully!
- In class however, let's focus on using R packages for doing the same.

Example

- Health plans use many tools to try to control the cost of prescription medicines.

- For older drugs, generic substitutes that are the equivalent to name-brand drugs are available at considerable savings.

- Another tool that may lower costs is restricting drugs that the physician may prescribe.

- For example if three similar drugs for treating the same condition are available, a health plan may require the physician to prescribe only one of them, allowing the plan to negotiate discounts based on a higher volume of sales.

- We have data from 29 health plans can be used to explore the effectiveness of these two strategies in controlling drug costs.

- The response is COST, the average cost of the prescriptions to the plan per day (in dollars).

Example

- Potential explanatory variables are:
    - RXPM: Average number of prescriptions per member per year
    - GS: Percent generic substitute used by the plan
    - RI: Restrictiveness Index, from 0 (no restrictions) to 100 (total restrictions on the physician)
    - COPAY: Average member copay on prescriptions
    - AGE: Average member age
    - F: percent female members
    - MM: Member months, a measure of the size of the plan
    - ID: an identifier for the name of the plan
- Since we do not have so many data points, let's use Bayesian model selection and model averaging to explore the relationship of GS and RI to COST, adjusting for the other variables.
- The data is in the file `costs.txt` on Sakai.

In-class analysis: move to the R script here.