Statistical Methods - Bayesian Inference

Lecture 4

Mattias Villani

Sveriges Riksbank and Stockholm University

March 18, 2010

Overview Lecture 4

- The linear regression model
- Regression with dichotomous response
- The Metropolis algorithm
- Autoregressive processes (AR)

The linear regression model

- The ordinary linear regression model:

$$
\begin{aligned}
y_i &= \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_k x_{ik} + \varepsilon_i \\
& \varepsilon_i \overset{iid}{\sim} N(0, \sigma^2).
\end{aligned}
$$

- Parameters $\theta = (\beta_1, \beta_2, ..., \beta_k, \sigma^2)$.
- Assumptions:
  - $E(y_i) = \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_k x_{ik}$ (linear function)
  - $Var(y_i) = \sigma^2$ (homoscedasticity)
  - $Corr(y_i, y_j|X) = 0$, $i \neq j$.
  - Normality of ε_i.

The linear regression model, cont.

- The linear regression model in matrix form

$$\underset{(n\times 1)}{y} = \underset{(n\times k)(k\times 1)}{X\beta} + \underset{(n\times 1)}{\varepsilon}$$

$$\begin{aligned}
y &= \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix},\ \beta = \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_k \end{pmatrix},\ \varepsilon = \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix} \\
X &= \begin{pmatrix} x_1' \\ \vdots \\ x_n' \end{pmatrix} = \begin{pmatrix} x_{11} & \cdots & x_{1k} \\ \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{nk} \end{pmatrix}
\end{aligned}$$

- Usually $x_{i1} = 1$, for all i. β_1 becomes the intercept.

The linear regression model, cont.

- Likelihood:

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

- Standard non-informative prior: uniform on $(\beta, \log\sigma)$

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

- Joint posterior of β and σ^2:

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

- Conditional posterior of β :

$$\begin{aligned} \beta|\sigma^2, y &\sim N(\hat{\beta}, \sigma^2 V_\beta) \\ \hat{\beta} &= (X'X)^{-1}X'y \\ V_\beta &= (X'X)^{-1}. \end{aligned}$$

- Marginal posterior of σ^2 :

$$\begin{aligned} \sigma^2|y &\sim Inv\text{-}\chi^2(n-k, s^2) \\ s^2 &= \frac{1}{n-k}(y-X\hat{\beta})'(y-X\hat{\beta}). \end{aligned}$$

- Marginal posterior of β :

$$\beta|y \sim t_{n-k}(\hat{\beta}, \sigma^2 V_\beta).$$

which is proper if $n > k$ and X has full column rank.

- Simulate from the joint posterior by iteratively simulating from $p(\sigma^2|y)$ and $p(\beta|\sigma^2, y)$.
- Predictive distribution of response $\tilde{y}$ with known predictors $\tilde{X}$:

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

$$\begin{aligned} \text{Predictive Variance} &= s^2 I + \tilde{X} s^2 V_\beta \tilde{X}' \\ &= \varepsilon\text{-Variance} + \tilde{X}(\text{Posterior Variance of } \beta)\tilde{X}'. \end{aligned}$$

Informative prior - dummy variable approach

$$\beta_j \sim N(\beta_{j0}, \sigma^2_{\beta_j}).$$

- Typical regression observation

$$y_i | x_i \sim N(x_i\beta, \sigma^2) \propto \exp\left[-\frac{1}{2\sigma^2}(y_i - \sum_{j=1}^{k}\beta_j x_j)^2\right]$$

- The $N(\beta_{j0}, \sigma^2_{\beta_j})$ prior is proportional to

$$\exp\left[-\frac{1}{2\sigma^2_{\beta_j}}(\beta_j - \beta_{j0})^2\right] = \exp\left[-\frac{1}{2\sigma^2_{\beta_j}}(\beta_{j0} - \beta_j)^2\right],$$

which is identical to a regression observation with response β_{j0}, error variance $\sigma^2_{\beta_j}$ and predictors $x_j = 1$ and $x_i = 0$ for all $i \neq j$.

- The informative prior may therefore be implemented using a non-informative prior in the extended regression

$$y_* = X_* \beta$$

where

$$y_* = \begin{pmatrix} y_1 \\ \vdots \\ y_n \\ \beta_{j0} \end{pmatrix}, \; X_* = \begin{pmatrix} \underset{(n\times 1)}{x_1} & \cdots & \underset{(n\times 1)}{x_j} & \cdots & \underset{(n\times 1)}{x_k} \\ 0 & 0 & 1 & 0 & 0 \end{pmatrix}$$

$$\Sigma_{y*} = \begin{pmatrix} \sigma^2 I_n & \underset{(n\times 1)}{0} \\ \underset{(1\times n)}{0} & \sigma^2_{\beta_j} \end{pmatrix}.$$

Regression with dichotomous response

- Response is assumed to be dichotomous (0-1).
- Example: Spam data. Covariates: average word length, proportion of \$-symbols, is the word 'Mattias' present in the e-mail? etc.
- Logistic regression:

  $$\Pr(y_i = 1 \mid x_i) = \frac{\exp(x_i'\beta)}{1+\exp(x_i'\beta)}.$$

  Likelihood:

  $$p(y|X,\beta) = \prod_{i=1}^{n} \frac{\exp(x_i'\beta)^{y_i}}{1+\exp(x_i'\beta)}.$$

  Posterior is non-standard, but in most situation can be approximated well by a normal distribution. Numerical optimization.
- Probit regression: $\Pr(y_i = 1 \mid x_i) = \Phi(x_i'\beta)$. Also easy to handle by numerical optimization (or MCMC ...).

The Metropolis Algorithm

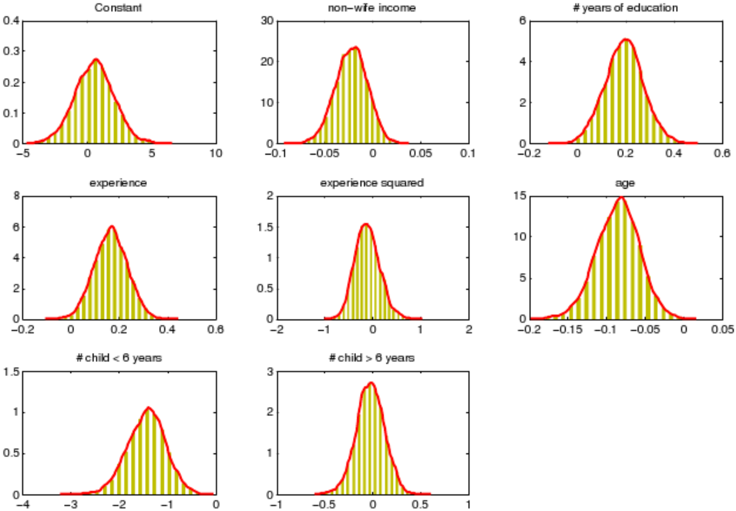
- General algorithm to simulate from the posterior $p(\theta|y)$.
- First: Optimize $p(\theta|y)$ to obtain posterior mode $\hat{\theta}$ and approximate covariance matrix $I^{-1}(\hat{\theta})$.
- Initialize with $\theta = \theta_0$
- For $t = 1, 2, ...$
- Sample a proposal draw $\theta^*|\theta^{(t-1)} \sim N_p[\theta^{(t-1)}, c \cdot I^{-1}(\hat{\theta})]$, where c is a tuning factor.
    - Accept θ^* with probability

$$r(\theta^*, \theta^{(t-1)}) = \min\left[\frac{p(\theta^*|y)}{p(\theta^{(t-1)}|y)}, 1\right].$$

    If the proposal is accepted, set $\theta^{(t)} = \theta^*$ (move), otherwise set $\theta^{(t)} = \theta^{(t-1)}$ (stay)

- Note that the draws are autocorrelated, but they still converge in distribution to $p(\theta|y)$.

Example: Participation of female spouse in labor market



Summary of Bayesian Inference with Variable Selection
Author: Mattias Villani, Stockholm University and Sveriges Riksbank
October 12, 2008. 22:28:07

Parameter	Mode	Mean	Stdev (Hess)	Stdev (MCMC)	t Ratio	Incl Prob
our	+0.434	+0.434	+0.062	+0.061	+6.953	+1.000
over	+0.912	+0.949	+0.155	+0.163	+5.895	+1.000
remove	+2.744	+2.738	+0.252	+0.256	+10.881	+1.000
internet	+0.901	+0.886	+0.133	+0.140	+6.768	+1.000
free	+0.689	+0.718	+0.083	+0.086	+8.311	+1.000
hpl	-0.657	-0.660	+0.143	+0.146	-4.599	+1.000
!	+0.680	+0.694	+0.102	+0.091	+6.665	+1.000
$	+6.129	+6.079	+0.423	+0.552	+14.498	+1.000
CapRunMax	+0.005	+0.005	+0.001	+0.001	+5.259	+1.000
CapRunTotal	+0.001	+0.001	+0.000	+0.000	+4.773	+0.998
Const	-1.329	-1.340	+0.063	+0.073	-21.085	+1.000
hp	-0.787	-0.797	+0.085	+0.096	-9.291	+1.000
george	-0.415	-0.406	+0.057	+0.055	-7.240	+1.000
1999	-0.586	-0.606	+0.129	+0.146	-4.559	+0.991
re	-0.547	-0.553	+0.089	+0.095	-6.157	+1.000
edu	-0.972	-0.975	+0.141	+0.143	-6.909	+1.000

- AR(p) process

$$x_t = \phi_1 x_{t-1} + ... + \phi_p x_{t-p} + \varepsilon_t, \quad \varepsilon_t \overset{iid}{\sim} N(0, \sigma^2).$$

- But this is just a linear regression of x_t on $(x_{t-1}, ..., x_{t-p})$.
- Random walk prior:

$$\begin{aligned} \mathrm{E}(\phi_1) &= 1 \\ \mathrm{E}(\phi_j) &= 0 \text{ for } j = 2, ..., p. \\ \mathrm{S}(\phi_j) &= \frac{\psi}{j}. \end{aligned}$$

Note how the prior shrinks longer lags more heavily toward zero.

- We can impose stationarity restrictions by restricting the domain of the prior. Posterior draws that imply non-stationarity behavior are removed from the posterior sample.
- Model with steady state:

$$x_t - \mu = \phi_1(x_{t-1} - \mu) + \ldots + \phi_p(x_{t-p} - \mu) + \varepsilon_t.$$

- $\mu = \mathrm{E}(x_t)$ is the unconditional mean or steady-state of the process. 'where the system goes to if the shocks (ε_t) are turned off'.
- μ is important as long-run forecasts (quickly) approach the steady state.
- Prior: $\mu \sim N(\theta_\mu, \psi_\mu^2)$, independent of ϕ's and σ.

Autoregressive processes, cont.

- The posterior can be simulated by Gibbs sampling:
  - $\mu | \phi, \sigma, x \sim$ Normal
  - $\phi | \mu, \sigma, x \sim$ Normal
  - $\sigma | \mu, \phi, x \sim$ Inverse Scaled χ^2
- Everything above can easily be extended to vector processes (VARs).

Example: Swedish macro data