AMS 207: Intermediate
Bayesian Modeling

9: Bayesian Inference
in Linear Models

David Draper

Department of Applied Mathematics and Statistics
University of California, Santa Cruz

draper@ams.ucsc.edu
http://www.ams.ucsc.edu/~draper

© 2010 David Draper (all rights reserved)
The Linear Model

Probably the single most important (work-horse) model in statistics is the linear model (LM):

- you have a single quantitative (continuous) outcome variable $y$, and you want to relate it to $k$ predictor variables $x_1, \ldots, x_k$ (all of the $x_j$ could be quantitative, or all of them could be qualitative, or anything in between; note that in this part of the notes $k$ no longer stands for the length of the parameter vector $\theta$);

- there's a population $\mathcal{P}$ of individuals to which you wish to generalize from your data, so you take a random sample of $n$ of these individuals and measure $(y, x_1, \ldots, x_k)$ on all of them; and

- in the LM you pretend that the data has the simplest non-trivial geometry, namely that all the data points lie near a hyperplane:

$$y_i = \mu_i + e_i = \beta_0 + \sum_{j=1}^{k} \beta_j x_{ij} + e_i, \quad (1)$$

in which (given $\sigma^2$) the standard assumption about the discrepancies $e_i$ between the observed ($y_i$) and modeled ($\mu_i = \beta_0 + \sum_{j=1}^{k} \beta_j x_{ij}$) outcome values is $(e_i|\sigma^2) \sim \text{IID } N(0, \sigma^2)$.

This gives the vector $y = (y_1, \ldots, y_n)$ of outcome values a simple multivariate-normal structure:

$$y \sim N_n(\mu, \Sigma), \quad (2)$$

with $\mu = (\mu_1, \ldots, \mu_n)$ and $\Sigma = \sigma^2 I_n$, where $I_n$ is the $(n \times n)$ identity matrix.

This model is not necessarily linear in the $x$ variables but is linear in the regression parameters $\beta_j$, which is what the L in LM means.
LM (continued)

It's convenient to write equation (1) in matrix terms: let $X$ be the $(n \times (k+1))$ matrix with a vector of $n$ 1s in the first column, the vector $X_1 = (x_{11}, \ldots, x_{n1})$ of values of the variable $x_1$ in the second column, the vector $X_2 = (x_{12}, \ldots, x_{n2})$ of values of the variable $x_2$ in the third column, and so on down to the vector $X_k = (x_{1k}, \ldots, x_{nk})$ of values of the variable $x_k$ in column $(k+1)$; then (1) can be written

$$y = X \beta + e,$$

in which $\beta = (\beta_0, \beta_1, \ldots, \beta_k)$ and $e = (e_i, \ldots, e_n)$; the parameter vector in this model is $\theta = (\beta_0, \beta_1, \ldots, \beta_k, \sigma^2)$.

A reminder of some basic repeated-sampling facts about random vectors:

- You start with quantities $y_i$ ($i = 1, \ldots, n$), where $y_i$ is a random draw from a population with (repeated-sampling) mean $E_{RS}(y_i) = \mu_i$, (repeated-sampling) variance $V_{RS}(y_i) = \sigma_i^2$ and (repeated-sampling) covariances $C_{RS}(y_i, y_i) = c_{ii} = \sigma_i \sigma_i \rho_{ii}$, where $\rho_{ii}$ is the correlation between $y_i$ and $y_i$.

- A random vector $y = (y_1, \ldots, y_n)$ is just a way of collecting the quantities $y_i$ ($i = 1, \ldots, n$) into a column vector; similarly, let the $(n \times 1)$ vector $\mu = (\mu_i, \ldots, \mu_n)$ collect together the means and the $(n \times n)$ matrix $\Sigma = (c_{ii})$ — the matrix whose $(i, i')$ entry is $c_{ii'}$ — collect together the covariances; then it's natural to say that in repeated sampling $E_{RS}(y) = \mu$ and $V_{RS}(y) = \Sigma$.

- If $A$ is an $(n \times n)$ matrix of constants, and $b$ is an $(n \times 1)$ vector of constants, then in repeated sampling

$$E_{RS}(Ay + b) = AE_{RS}(y) + b = A\mu + b$$

and

$$V_{RS}(Ay + b) = AV_{RS}(y)A' = A\Sigma A'.$$
Having written the model in matrix form it’s immediate that in repeated sampling

\[ E_{RS}(y) = E_{RS}(X \beta + e) = X \beta + E_{RS}(e) = X \beta, \]

which means that the multivariate-normal representation of \( y \) in (2) is

\[ y \sim N_n(X \beta, \sigma^2 I_n). \]  

Standard repeated-sampling estimates of the elements of \( \theta \) have been available for several centuries; for example, the least-squares estimate of the regression coefficient vector \( \hat{\beta} \) — the vector \( \hat{\beta} \) that minimizes

\[ \sum_{i=1}^{n} [y_i - (\beta_0 + \sum_{j=1}^{k} \beta_j x_{ij})]^2 \]

is

\[ \hat{\beta}_{LS} = (X' X)^{-1} X' y, \]

and the usual repeated-sampling-unbiased estimate of \( \sigma^2 \) is

\[ \hat{\sigma}^2_U = \frac{1}{n - (k + 1)} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2, \]

where \( \hat{y}_i \) is element \( i \) of the vector \( \hat{y} = X \hat{\beta}_{LS} \).

Maximum-likelihood estimation in the LM leads to the same estimate of \( \beta \) but a slightly different estimate of \( \sigma^2 \): the Bayesian model here is

\[ (\beta, \sigma^2) \sim p(\beta, \sigma^2) \]

\[ (y_i | \beta, \sigma^2) \overset{\text{indep}}{\sim} N \left( \beta_0 + \sum_{j=1}^{k} \beta_j x_{ij}, \sigma^2 \right) \]

for \( i = 1, \ldots, n \), in which it’s standard to condition on \( X \) because in and of itself it contains no information about \( \theta \) beyond that embodied in (9) with the \( x_{ij} \) as fixed known constants.
The joint sampling distribution in this model is evidently

\[
p(y|\beta, \sigma^2) = \prod_{i=1}^{n} p(y_i|\beta, \sigma^2) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2\sigma^2} \left[ y_i - \left( \beta_0 + \sum_{j=1}^{k} \beta_j x_{ij} \right) \right]^2 \right\} = (2\pi)^{-\frac{n}{2}} (\sigma^2)^{-\frac{n}{2}} \exp\left\{-\frac{1}{2\sigma^2} \sum_{i=1}^{n} \left[ y_i - \left( \beta_0 + \sum_{j=1}^{k} \beta_j x_{ij} \right) \right]^2 \right\},
\]

so the log-likelihood function is

\[
l(\beta, \sigma^2|y) = c - \frac{n}{2} \log \sigma^2 - \frac{1}{2\sigma^2} \sum_{i=1}^{n} \left[ y_i - \left( \beta_0 + \sum_{j=1}^{k} \beta_j x_{ij} \right) \right]^2.
\]

Without doing any calculating you can see how to maximize this log-likelihood function in \(\beta\): \(\beta\) appears only in the third term on the right-hand side of (11), and — because of the minus sign in front of this term — the log-likelihood will be maximized exactly where the sum of squares in this term is minimized, and that's the least-squares estimates, so in this model \(\hat{\beta}_{ML} = \hat{\beta}_{LS}\); but the MLE of \(\sigma^2\) is different:

\[
\frac{\partial}{\partial \sigma^2} l(\beta, \sigma^2|y) = -\frac{n}{2\sigma^2} + \frac{1}{\sigma^3} \sum_{i=1}^{n} \left[ y_i - \left( \beta_0 + \sum_{j=1}^{k} \beta_j x_{ij} \right) \right]^2,
\]

(12)
LM (continued)

and this is 0 (simultaneous with $\beta = \hat{\beta}_{ML}$) when

$$\hat{\sigma}_2^2 = \frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \left(1 - \frac{k + 1}{n}\right) \hat{\sigma}_U^2;$$  \hspace{1cm} (13)

this is another instance of the MLE being small-sample-biased in repeated sampling, with the bias $O\left(\frac{1}{n}\right)$.

What about an uncertainty assessment for $\hat{\theta}_{ML}$?

First let’s look at the repeated-sampling mean story for $\hat{\beta}_{ML}$:

$$E_{RS}(\hat{\beta}_{ML}) = E_{RS} \left[ (X' X)^{-1} X' y \right] = \left[ (X' X)^{-1} X' \right] E_{RS}(y)$$

$$= \left[ (X' X)^{-1} X' \right] (X \beta) = \beta,$$  \hspace{1cm} (14)

so the $\hat{\beta}_{ML}$ part of $\hat{\theta}_{ML}$ is repeated-sampling unbiased; what about its repeated-sampling covariance matrix?

$$V_{RS}(\hat{\beta}_{ML}) = V_{RS} \left[ (X' X)^{-1} X' y \right]$$

$$= \left[ (X' X)^{-1} X' \right] V_{RS}(y) \left[ (X' X)^{-1} X' \right]' .$$  \hspace{1cm} (15)

A reminder of some basic matrix facts:

• For any $(m \times k)$ matrix $A$ and $(k \times n)$ matrix $B$, $(AB)' = B' A'$; this means here that

$$\left[ (X' X)^{-1} X' \right]' = (X')' \left[ (X' X)^{-1} \right]' ;$$  \hspace{1cm} (16)

• For any matrix $A$, $(A')' = A$;

• For any square matrix $A$ for which the inverses exist, $(A^{-1})' = (A')^{-1}$.
For any symmetric matrix $A$, $A' = A$; when you put all these facts together,

$$
\left[ (X' X)^{-1} X' \right]' = (X')' \left[ (X' X)^{-1} \right]' \\
= X \left[ (X'X)' \right]^{-1} = X(X'X)^{-1}.
$$

(17)

To complete the calculation, $V_{RS}(y) = \sigma^2 I_n$ and

$$
V_{RS}(\hat{\beta}_{ML}) = V_{RS} \left[ (X' X)^{-1} X'y \right] \tag{18}
$$

$$
= \left[ (X' X)^{-1} X' \right] \sigma^2 I_n \left[ X(X'X)^{-1} \right] \\
= \sigma^2 (X' X)^{-1} (X'X) (X'X)^{-1} \\
= \sigma^2 (X' X)^{-1}.
$$

(19)

Of course this also means that $V_{RS}(\hat{\beta}_{LS}) = \sigma^2 (X' X)^{-1}$; people tend to estimate this with $\hat{V}_{RS}(\hat{\beta}_{LS}) = \hat{\sigma}_U^2 (X' X)^{-1}$, which provides repeated-sampling standard errors, e.g., $\hat{SE}_{RS}(\hat{\beta}_j) = \sqrt{\hat{V}_{RS}(\hat{\beta}_j)} = \hat{\sigma}_U$ times the square root of the $(j + 1, j + 1)$ entry in $(X' X)^{-1}$.

Now it’s another basic fact about the multivariate normal distribution that if $y \sim N_d(\mu, \Sigma)$, $A$ is a $(q \times d)$ matrix of constants and $b$ is a $(q \times 1)$ vector of constants, then $(A y + b) \sim N_q(A \mu + b, A \Sigma A')$; applying this to

$$
\hat{\beta}_{LS} = \hat{\beta}_{ML} = \left[ (X' X)^{-1} X' \right] y
$$

we get that in repeated sampling

$$
\hat{\beta}_{LS} = \hat{\beta}_{ML} \sim N_{k+1} \left[ \beta, \sigma^2 (X' X)^{-1} \right].
$$

(20)

This provides the basis of small-sample (repeated-sampling) interval estimation: we know from this that in repeated sampling, for all $n$,

$$
\frac{\hat{\beta}_j - \beta_j}{\sqrt{V_{RS}(\hat{\beta}_j)}} \sim_{RS} N(0, 1),
$$

(21)
but of course as usual this doesn't mean that in repeated sampling, for small $n$,

$$\frac{\hat{\beta}_j - \beta_j}{\sqrt{\hat{V}_{RS}(\hat{\beta}_j)}} \sim_{RS} N(0, 1), \quad (22)$$

because we have to pay the price for estimating $V_{RS}(\hat{\beta}_j)$ by $\hat{V}_{RS}(\hat{\beta}_j)$.

When there are no predictor variables ($k = 0$) we know the small-$n$ correction that's needed: the (repeated-sampling) unbiased estimate of $\sigma^2$,

$$\hat{\sigma}_U^2 = \frac{1}{n-1} \sum_{i=1}^{n} (y_i - \bar{y}) = \frac{1}{n - (k + 1)} \sum_{i=1}^{n} (y_i - \hat{y}_i), \quad (23)$$

is said to have $(n - (k + 1)) = (n - 1)$ degrees of freedom for estimating $\sigma^2$, and

$$\frac{\bar{y} - \mu}{\sqrt{\hat{V}_{RS}(\bar{y})}} = \frac{\hat{\beta}_j - \beta_j}{\sqrt{\hat{V}_{RS}(\hat{\beta}_j)}} \sim_{RS} t_{n-(k+1)} = t_{n-1}; \quad (24)$$

by analogy with this the regression result is

$$\frac{\hat{\beta}_j - \beta_j}{\sqrt{\hat{V}_{RS}(\hat{\beta}_j)}} \sim_{RS} t_{n-(k+1)}, \quad (25)$$

and a (small-sample-exact, under the model) 95% confidence interval for $\beta_j$ is then of the form

$$\hat{\beta}_j \pm t^{0.95}_{n-(k+1)} \sqrt{\hat{V}_{RS}(\hat{\beta}_j)}.$$ 

A (small-sample-exact, under the model) 95% confidence interval for $\sigma^2$ can then be based on the fact (in parallel with what happens with $k = 0$) that in repeated sampling

$$[n - (k + 1)] \hat{\sigma}_U^2 / \sigma^2 \sim_{RS} \chi^2_{n-(k+1)}. \quad (26)$$
(\beta, \sigma^2) \sim p(\beta, \sigma^2) \quad \text{(27)}

(y_i | \beta, \sigma^2) \sim \text{indep} \quad N\left( \beta_0 + \sum_{j=1}^{k} \beta_j x_{ij}, \sigma^2 \right)

For **Bayesian estimation** in this model we need a **prior distribution** on \( \theta = (\beta, \sigma^2) \); with a **flat prior** you’d expect to get results **similar to those from maximum-likelihood estimation**, but is there a **conjugate prior** for this model, and if so what’s the **posterior**?

First let’s talk about the **flat-prior situation**: as we’ve discussed, one **general way** to get a **diffuse prior** is to use **Jeffrey’s invariance idea**; here this comes out

\[
p(\beta, \sigma^2) = \frac{c}{\sigma^2}. \quad \text{(28)}
\]

[Raquel’s slides]

[Bernardo-Smith conjugate analysis:]

442. A. Summary of Basic Formulae

Linear Regression

- \( z = (y, X), \quad y = (y_1, \ldots, y_n)^T \in \mathbb{R}^n, \ x_i = (x_{i1}, \ldots, x_{ik}) \in \mathbb{R}^k, \ X = (x_{ij}) \)
- \( p(y | X, \theta, \lambda) = N_n(y | X\theta, \lambda I_n), \quad \theta \in \mathbb{R}^k, \quad \lambda > 0 \)
- \( t(z) = (X^T X, X^T y) \)
- \( p(\theta, \lambda) = N_k(\theta, \theta_0, n_0, \alpha, \beta) = N_k(\theta | \theta_0, n_0 \lambda) Ga(\lambda | \alpha, \beta) \)
- \( p(\theta) = St_k(\theta | \theta_0, n_0 \alpha \beta^{-1}, 2\alpha) \)
- \( p(\lambda) = Ga(\lambda | \alpha, \beta) \)
- \( p(y | x) = St(y | x\theta, f(x) \alpha \beta^{-1}, 2\alpha) \)
- \( f(x) = 1 - x(x^T x + n_0)^{-1} x^T \)
- \( p(\theta | z) = St_k(\theta | \theta_n, (n_0 + X^T X)(\alpha + \frac{1}{2}n)\beta_n^{-1}, 2\alpha + n) \)
- \( \theta_n = (n_0 + X^T X)^{-1}(n_0 \theta_0 + X^T y) \)
- \( \beta_n = \beta + \frac{1}{2}(y - X\theta_n)^T y + \frac{1}{2}(\theta_0 - \theta_n)^T n_0 \theta_0 \)
- \( p(\lambda | z) = Ga(\lambda | \alpha + \frac{1}{2}n, \beta_n) \)
- \( p(y | x, z) = St(y | x\theta_n, f_n(x)(\alpha + \frac{1}{2}n)\beta_n^{-1}, 2\alpha + n) \)
- \( f_n(x) = 1 + x(x^T x + n_0 + X^T X)^{-1} x^T \)

\[ \sum_{i=1}^{10} \alpha_i \cdot f_i \]

\[ \Sigma = \begin{pmatrix} 2 \\ \vdots \\ 1 \end{pmatrix} \]