ST 740: Linear Models and Multivariate Normal Inference

Alyson Wilson

Department of Statistics
North Carolina State University

November 4, 2013

Normal Linear Regression Model

We assume that

$$E[Y \mid \mathbf{x}] = \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p$$

and that

$$Y_i = \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p + \epsilon_i = \beta^T \mathbf{x} + \epsilon_i$$

where ϵ_i are i.i.d. Normal$(0, \sigma^2)$. This implies

$$\begin{aligned} p(y_1, \ldots, y_n \mid \mathbf{x}_1, \ldots, \mathbf{x}_n, \beta, \sigma^2) &= \prod_{i=1}^{n} p(y_i \mid \mathbf{x}_i, \beta, \sigma^2) \\ &= (2\pi\sigma^2)^{-n/2} \exp(-\frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - \beta^T \mathbf{x}_i)^2) \end{aligned}$$

Normal Linear Regression Model

Multivariate Normal

If we think of $\mathbf{y}$ as the n-dimensional column vector $(y_1, \ldots, y_n)^T$, and $\mathbf{X}$ as the $n \times p$ marix with ith row $\mathbf{x}_i$, then we can write the normal linear regression model as

$$\mathbf{y} \,|\, \mathbf{X}, \beta, \sigma^2 \sim \text{MVN}(\mathbf{X}\beta, \sigma^2 I)$$

Multivariate Normal Sampling Distribution

Assume that we have data $\mathbf{y}_1, \ldots, \mathbf{y}_n$ with a *multivariate normal* sampling distribution. If $\mathbf{y}_i$ is a p-dimensional vector, we have

$$f(\mathbf{y}_i \,|\, \theta, \Sigma) = (2\pi)^{-p/2} |\Sigma|^{-1/2} \exp(-(\mathbf{y}_i - \theta)^T \Sigma^{-1} (\mathbf{y}_i - \theta)/2)$$

Multivariate Normal Inference

Prior Distribution for θ

If our sampling distribution is univariate normal, we saw that it was often (computationally) convenient to specify the prior distribution for the mean as a normal distribution.

Suppose that we decide to specify the prior distribution for θ, the mean of our multivariate normal sampling distribution, as $\mathrm{MVN}(\mu_0, \Lambda_0)$.

What is the full conditional distribution of θ given $\mathbf{y}$ and Σ? To figure this out, we'll need the likelihood ($\mathrm{MVN}(\theta, \Sigma)$), the prior distribution for θ, and some information about the prior distribution for Σ.

Prior Distribution for θ

First, let's rewrite the prior distribution for θ.

$$\begin{aligned}
\pi(\theta \mid \mu_0, \Lambda_0) &= (2\pi)^{-p/2}|\Lambda_0|^{-1/2}\exp(-(\theta-\mu_0)^T\Lambda_0^{-1}(\theta-\mu_0)/2) \\
&= (2\pi)^{-p/2}|\Lambda_0|^{-1/2} \\
&\quad \exp\left[-\frac{1}{2}\theta^T\Lambda_0^{-1}\theta + \theta^T\Lambda_0^{-1}\mu_0 - \frac{1}{2}\mu_0^T\Lambda_0^{-1}\mu_0\right] \\
&\propto \exp\left[-\frac{1}{2}\theta^T\Lambda_0^{-1}\theta + \theta^T\Lambda_0^{-1}\mu_0\right] \\
&\propto \exp\left[-\frac{1}{2}\theta^T A_0\theta + \theta^T b_0\right]
\end{aligned}$$

where $A_0 = \Lambda_0^{-1}$ and $b_0 = \Lambda_0^{-1}\mu_0$.

Notice that the prior covariance matrix is A_0^{-1} and the prior mean is $A_0^{-1}b_0$.

Sampling Distribution

Now, let's take a look at the sampling distribution for $\mathbf{y}_1, \ldots, \mathbf{y}_n$.

$$\pi(\mathbf{y}_1, \ldots, \mathbf{y}_n \,|\, \theta, \Sigma) = \prod_{i=1}^{n} (2\pi)^{-p/2} |\Sigma|^{-1/2} \exp(-(\mathbf{y}_i - \theta)^T \Sigma^{-1} (\mathbf{y}_i - \theta)/2)$$

$$= (2\pi)^{-np/2} |\Sigma|^{-n/2} \exp\left[-\frac{1}{2}\sum_{i=1}^{n} (\mathbf{y}_i - \theta) \Sigma^{-1} (\mathbf{y}_i - \theta)\right]$$

$$\propto |\Sigma|^{-n/2} \exp\left[-\frac{1}{2}\sum_{i=1}^{n} \mathbf{y}_i^T \Sigma^{-1} \mathbf{y}_i\right] \exp\left[-\frac{n}{2}\theta^T \Sigma^{-1} \theta + \theta^T \Sigma^{-1} n\bar{\mathbf{y}}\right]$$

$$\propto |\Sigma|^{-n/2} \exp\left[-\frac{1}{2}\sum_{i=1}^{n} \mathbf{y}_i^T \Sigma^{-1} \mathbf{y}_i\right] \exp\left[-\frac{1}{2}\theta^T A_1 \theta + \theta^T b_1\right]$$

where $A_1 = n\Sigma^{-1}$ and $b_1 = n\Sigma^{-1}\bar{\mathbf{y}}$.

Posterior/Full Conditional Distribution for θ

$$
\begin{aligned}
\pi(\theta, \Sigma \,|\, \mathbf{y}_1, \ldots \mathbf{y}_n) \quad \propto \quad & \pi(\Sigma)|\Sigma|^{-n/2} \exp\left[-\frac{1}{2}\sum_{i=1}^{n} \mathbf{y}_i^T \Sigma^{-1} \mathbf{y}_i\right] \\
& \exp\left[-\frac{1}{2}\theta^T A_1 \theta + \theta^T b_1\right] \exp\left[-\frac{1}{2}\theta^T A_0 \theta + \theta^T b_0\right]
\end{aligned}
$$

Posterior/Full Conditional Distribution for θ

Assuming that the prior distribution for Σ does not depend on θ,

$$\begin{aligned}
\pi(\theta \,|\, \Sigma, \mathbf{y}_1, \ldots, \mathbf{y}_n) &\propto \exp\left[-\frac{1}{2}\theta^T A_1 \theta + \theta^T b_1\right] \exp\left[-\frac{1}{2}\theta^T A_0 \theta + \theta^T b_0\right] \\
&\propto \exp\left[-\frac{1}{2}\theta^T (A_0 + A_1)\theta + \theta^T (b_0 + b_1)\right]
\end{aligned}$$

This is a MVN with covariance matrix

$$(A_0 + A_1)^{-1} = (\Lambda_0^{-1} + n\Sigma^{-1})^{-1}$$

and mean

$$(A_0 + A_1)^{-1}(b_0 + b_1) = (\Lambda_0^{-1} + n\Sigma^{-1})^{-1}(\Lambda_0^{-1}\mu_0 + n\Sigma^{-1}\bar{\mathbf{y}})$$

Full Conditional Distribution for θ

This is a MVN with covariance matrix

$$(A_0 + A_1)^{-1} = (\Lambda_0^{-1} + n\Sigma^{-1})^{-1}$$

The precision (inverse variance) is the sum of the prior precision and the data precision.

The mean is

$$(A_0 + A_1)^{-1}(b_0 + b_1) = (\Lambda_0^{-1} + n\Sigma^{-1})^{-1}(\Lambda_0^{-1}\mu_0 + n\Sigma^{-1}\bar{\mathbf{y}})$$

which is a weighted average of the prior mean and the sample mean.

Multivariate Normal Inference

Prior Distribution for Σ

To construct a prior distribution for Σ, we need a distribution over symmetric, positive definite matrices.

The idea is a generalization of the chi-squared (gamma) distribution.

1. Sample $z_1, \ldots, z_{\nu_0}$ i.i.d. $\text{MVN}(0, \Phi_0)$
2. Calculate $Z^T Z = \sum_{i=1}^{\nu_0} z_i z_i^T$

The distribution of the matrices $Z^T Z$ is called a Wishart(ν_0, Φ_0) distribution.

Wishart Properties

- If $\nu_0 > p$ then $Z^T Z$ is positive definite with probability 1.
- $Z^T Z$ is symmetric with probability 1.
- $E[Z^T Z] = \nu_0 \Phi_0$

By analogy with the univariate normal case, we use the Wishart distribution as a prior for the precision matrix and the inverse Wishart distribution as a prior for the covariance matrix.

Prior Distribution for Σ

Inverse Wishart Distribution

1. Sample $z_1, \ldots, z_{\nu_0}$ i.i.d. MVN$(0, S_0^{-1})$
2. Calculate $(Z^T Z)^{-1} = (\sum_{i=1}^{\nu_0} z_i z_i^T)^{-1}$

The distribution of the matrices $(Z^T Z)^{-1}$ is called an inverse Wishart(ν_0, S_0^{-1}) distribution.

Inverse Wishart Distribution

$$\begin{aligned} E[Z^T Z] &= \nu_0 S_0^{-1} \\ E[(Z^T Z)^{-1}] &= \frac{1}{\nu_0 - p - 1} S_0 \end{aligned}$$

The density function is

$$\pi(\Sigma) = \left[2^{\nu_0 p/2} \pi^{\binom{p}{2}/2} |S_0|^{-\nu_0/2} \prod_{j=1}^{p} \Gamma([\nu_0 + 1 - j]/2) \right]^{-1}$$
$$|\Sigma|^{-(\nu_0+p+1)/2} \exp[-\mathrm{tr}(S_0 \Sigma^{-1})/2]$$

Choosing Parameters

For the inverse Wishart distribution, think of ν_0 as the "prior sample size."

- If we are confident that the true covariance matrix is near some covariance matrix Σ_0, choose ν_0 large and set $S_0 = (\nu_0 - p - 1)\Sigma_0$.
- If we choose $\nu_0 = p + 2$ and $S_0 = \Sigma_0$, then the samples will be loosely centered around Σ_0.

Full Conditional Distribution for Σ

Useful Matrix Algebra Fact

$$\sum_{i=1}^{K} \mathbf{b}_k^T \mathbf{A} \mathbf{b}_k = \text{tr}(\mathbf{B}^T \mathbf{B} \mathbf{A})$$

where $\mathbf{B}$ is the matrix whose kth row is $\mathbf{b}_k^T$. Therefore,

$$\sum_{i=1}^{n} (\mathbf{y}_i - \theta)^T \Sigma^{-1} (\mathbf{y}_i - \theta) = \text{tr}(\mathbf{S}_\theta \Sigma^{-1})$$

where $\mathbf{S}_\theta = \sum_{i=1}^{n} (\mathbf{y}_i - \theta)(\mathbf{y}_i - \theta)^T$ is the *residual sum of squares matrix* for $\mathbf{y}_1, \ldots, \mathbf{y}_n$.

Full Conditional Distribution for Σ

$$
\begin{aligned}
\pi(\theta, \Sigma \mid \mathbf{y}_1, \ldots, \mathbf{y}_n) &\propto \pi(\theta)|\Sigma|^{-n/2} \exp\left(-\frac{1}{2}\sum_{i=1}^{n}(\mathbf{y}_i - \theta)^T \Sigma^{-1}(\mathbf{y}_i - \theta)\right) \\
&\quad |\Sigma|^{-(\nu_0+p+1)/2} \exp[-\mathrm{tr}(S_0\Sigma^{-1})/2] \\
\pi(\Sigma \mid \theta, \mathbf{y}_1, \ldots, \mathbf{y}_n) &\propto |\Sigma|^{-(n+\nu_0+p+1)/2} \exp\left(-\frac{1}{2}\mathrm{tr}((S_0 + S_\theta)\Sigma^{-1})\right)
\end{aligned}
$$

So the full conditional distribution of Σ is
Inverse Wishart$(\nu_0 + n, [S_0 + S_\theta]^{-1})$.

Noninformative Prior for Multivariate Normal Data

A commonly used noninformative prior is

$$\pi(\theta, \Sigma) \propto |\Sigma|^{-(p+1)/2}$$

MCMC for Multivariate Normal Model

To summarize, suppose that our model is

$$\begin{aligned}\mathbf{Y}_i \mid \theta, \Sigma &\sim \mathsf{MVN}(\theta, \Sigma)\\ \theta &\sim \mathsf{MVN}(\mu_0, \Lambda_0)\\ \Sigma &\sim \text{Inverse Wishart}(\nu_0, S_0^{-1})\end{aligned}$$

We can draw a random sample from our posterior distribution using Gibbs sampling.

$$\begin{aligned}\pi(\theta \mid \mathbf{y}_1, \ldots, \mathbf{y}_n, \Sigma) &\sim \mathsf{MVN}((\Lambda_0^{-1} + n\Sigma^{-1})^{-1}(\Lambda_0^{-1}\mu_0 + n\Sigma^{-1}\bar{\mathbf{y}}),\\ &\quad (\Lambda_0^{-1} + n\Sigma^{-1})^{-1})\\ \pi(\Sigma \mid \mathbf{y}_1, \ldots, \mathbf{y}_n, \theta) &\sim \text{Inverse Wishart}(\nu_0 + n, [S_0 + S_\theta]^{-1})\end{aligned}$$

The R package `MCMCpack` has a random inverse Wishart distribution `riwish(v,S)`. It seems to be built for Gibbs sampling—it generates only one random draw at a time.

Classical Approach

From a classical approach, we choose an estimator for β that minimzes $\sum_{i=1}^{n}(y_i - \beta^T \mathbf{x}_i)^2$. We will call $\widehat{\beta}_{OLS}$ the *ordinary least squares* estimate of β.

Bayesian Approach

Semi-Conjugate Prior Distribution for β

Suppose that we choose a prior distribution for β that is MVN(β_0, Σ_0).

$$
\begin{aligned}
&\pi(\beta \,|\, \mathbf{y}, \mathbf{X}, \sigma^2) \\
&\quad \propto p(\mathbf{y} \,|\, \mathbf{X}, \beta, \sigma^2)\pi(\beta)\pi(\sigma^2) \\
&\quad \propto \exp\left(-\frac{1}{2\sigma^2}(\mathbf{y}^T\mathbf{y} - 2\beta^T\mathbf{X}^T\mathbf{y} + \beta^T\mathbf{X}^T\mathbf{X}\beta)\right) \\
&\quad \exp\left(-\frac{1}{2}\beta^T\Sigma_0^{-1}\beta + \beta^T\Sigma_0^{-1}\beta_0\right) \\
&\quad \propto \exp\left(-\frac{1}{2}\beta^T\mathbf{X}^T\mathbf{X}\beta/\sigma^2 + \beta^T\mathbf{X}^T\mathbf{y}/\sigma^2\right) \\
&\quad \exp\left(-\frac{1}{2}\beta^T\Sigma_0^{-1}\beta + \beta^T\Sigma_0^{-1}\beta_0\right) \\
&\quad \propto \exp\left(-\frac{1}{2}\beta^T(\mathbf{X}^T\mathbf{X}/\sigma^2 + \Sigma_0^{-1})\beta + \beta^T(\mathbf{X}^T\mathbf{y}/\sigma^2 + \Sigma_0^{-1}\beta_0)\right)
\end{aligned}
$$

Bayesian Approach

Semi-Conjugate Prior Distribution for β

The full conditional distribution for β, $\pi(\beta \mid \mathbf{y}, \mathbf{X}, \sigma^2)$ is

$$\mathrm{MVN}\left((\mathbf{X}^T\mathbf{X}/\sigma^2 + \Sigma_0^{-1})^{-1}(\mathbf{X}^T\mathbf{y}/\sigma^2 + \Sigma_0^{-1}\beta_0), (\mathbf{X}^T\mathbf{X}/\sigma^2 + \Sigma_0^{-1})^{-1}\right)$$

Bayesian Approach

Semi-Conjugate Prior Distribution for σ^2

Suppose that we choose a prior distribution for σ^2 that is InverseGamma$(\nu_0/2, \nu_0\sigma_0^2/2)$.

$$
\begin{aligned}
&\pi(\sigma^2 \,|\, \mathbf{y}, \mathbf{X}, \beta) \\
&\quad \propto p(\mathbf{y} \,|\, \mathbf{X}, \beta, \sigma^2)\pi(\beta)\pi(\sigma^2) \\
&\quad \propto (\sigma^2)^{-n/2} \exp\left(-\frac{1}{2\sigma^2}(\mathbf{y}^T\mathbf{y} - 2\beta^T\mathbf{X}^T\mathbf{y} + \beta^T\mathbf{X}^T\mathbf{X}\beta)\right) \\
&\quad (1/\sigma^2)^{\nu_0/2+1} \exp(-\nu_0\sigma_0^2/2\sigma^2) \\
&\quad \propto (1/\sigma^2)^{(n+\nu_0)/2} \exp\left(-\frac{1}{2\sigma^2}(\nu_0\sigma_0^2 + (\mathbf{y} - \mathbf{X}\beta)^T(\mathbf{y} - \mathbf{X}\beta)\right)
\end{aligned}
$$

which is InverseGamma$((n + \nu_0)/2, (\nu_0\sigma_0^2 + (\mathbf{y} - \mathbf{X}\beta)^T(\mathbf{y} - \mathbf{X}\beta))$.

Other Prior Distributions

One common (and convenient) noninformative prior distribution is uniform on $(\beta, \log(\sigma))$,

$$\pi(\beta, \sigma^2) \propto \frac{1}{\sigma^2}$$

In this case, we can show that

$$\begin{aligned} \pi(\beta \,|\, \sigma^2, \mathbf{y}, \mathbf{X}) &\sim \mathrm{MVN}((\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}, (\mathbf{X}^T\mathbf{X})^{-1}\sigma^2) \\ \pi(\sigma^2 \,|\, \mathbf{y}, \mathbf{X}) &\sim \mathrm{InverseGamma}(\frac{n-p}{2}, \frac{1}{2}(\mathbf{y} - \mathbf{X}\hat{\beta}_{OLS})^T(\mathbf{y} - \mathbf{X}\hat{\beta}_{OLS})) \end{aligned}$$

It is straightforward to derive all of the full conditionals and marginals for this model, so there are many choices for calculating samples from the posterior distribution.

Other Prior Distributions

When we discussed noninformative prior distributions, we learned that prior distributions can be selected to capture many different kinds of "knowledge." If we do not have much knowledge about the values of the regression parameters, we may wish to choose a prior distribution that is somehow "minimally informative."

Other Prior Distributions

Unit Information Prior

One suggestion for the normal linear regression model comes from Kass and Wasserman (1995). The idea is to create a prior distribution that contains the same "amount of information" as one observation.

Recall that the variance of $\hat{\beta}_{OLS}$ is $\sigma^2(\mathbf{X}^T\mathbf{X})^{-1}$. If we think about the precision (inverse variance) of this estimate, we have $(\mathbf{X}^T\mathbf{X})/\sigma^2$. The idea is that the precision of $\hat{\beta}_{OLS}$ is the "amount of information" in n observations. This implies that the amount of information in one observation is "one n^{th} as much," or $(\mathbf{X}^T\mathbf{X})/n\sigma^2$.

Suppose that we set $\Sigma_0 = n\sigma^2(\mathbf{X}^T\mathbf{X})^{-1}$ in the prior distribution for β. They also suggest setting $\beta_0 = \hat{\beta}_{OLS}$, $\nu_0 = 1$ and $\sigma_0^2 = (\hat{\sigma}_0^2)_{OLS}$.

Other Prior Distributions

g-prior

The *g-prior* starts from the idea that parameter estimation should be invariant to the scale of the regressors.

Suppose that $\mathbf{X}$ is a given set of regressors and $\tilde{\mathbf{X}} = \mathbf{XH}$ for some $p \times p$ matrix H. If we obtain the posterior distribution of β from $\mathbf{y}$ and $\mathbf{X}$ and $\tilde{\beta}$ from $\mathbf{y}$ and $\tilde{\mathbf{X}}$, then, according to this principle of invariance, the posterior distributions of β and $\mathbf{H}\tilde{\beta}$ should be the same.

$$\begin{aligned} \pi(\beta \,|\, \sigma^2) &\sim \text{MVN}(0, k(\mathbf{X}^T\mathbf{X})^{-1}) \\ \pi(\sigma^2) &\sim \text{InverseGamma}(\nu_0/2, \nu_0\sigma_0^2/2) \end{aligned}$$

A popular choice for k is $g\sigma^2$. Choosing g large reflects less informative prior beliefs.

Other Prior Distributions

g-prior

In this case, we can show that

$$\begin{aligned}\pi(\beta \,|\, \sigma^2, \mathbf{y}, \mathbf{X}) &\sim \text{MVN}(\frac{g}{g+1}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}, \frac{g}{g+1}(\mathbf{X}^T\mathbf{X})^{-1}\sigma^2) \\ \pi(\sigma^2 \,|\, \mathbf{y}, \mathbf{X}) &\sim \text{InverseGamma}(\frac{n+\nu_0}{2}, \frac{\nu_0\sigma_0^2 + SSR_g}{2})\end{aligned}$$

where $SSR_g = \mathbf{y}^T(I - \frac{g}{g+1}\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T)\mathbf{y}$.

Another popular "noninformative" choice is $\nu_0 = 0$.