Module 12: Linear Regression, the g-prior, and model selection

Rebecca C. Steorts

Announcements

Agenda

Setup

- $X_{n \times p}$: regression features or covariates (design matrix)
- $x_{p \times 1}$: ith row vector of the regression covariates
- $y_{n \times 1}$: response variable (vector)
- $\beta_{p \times 1}$: vector of regression coefficients

Goal: Estimation of $p(y \mid x)$.

Dimensions: $y_i - \beta^T x_i = (1 \times 1) - (1 \times p)(p \times 1) = (1 \times 1)$.

Multivariate Setup

Let's assume that we have data points (x_i, y_i) available for all $i = 1, \ldots, n$.

- y is the response variable

$$y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}_{n \times 1}$$

- x_i is the ith row of the design matrix $X_{n \times p}$.

Consider the regression coefficients

$$\beta = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{pmatrix}_{p \times 1}$$

Multivariate Setup

$$y \mid X, \beta, \sigma^2 \sim MVN(X\beta, \sigma^2 I)$$
$$\beta \sim MVN(\beta_0, \Sigma_0)$$

Recall the posterior can be shown to be

$$\boldsymbol{\beta} \mid \boldsymbol{y}, \boldsymbol{X} \sim MVN(\boldsymbol{\beta_n}, \Sigma_n)$$

where

$$\boldsymbol{\beta_n} = E[\boldsymbol{\beta} \mid \boldsymbol{y}, \boldsymbol{X}, \sigma^2] = (\Sigma_o^{-1} + (X^T X)^{-1}/\sigma^2)^{-1}(\Sigma_o^{-1}\boldsymbol{\beta}_0 + \boldsymbol{X}^T \boldsymbol{y}/\sigma^2)$$

$$\Sigma_n = \mathrm{Var}[\boldsymbol{\beta} \mid \boldsymbol{y}, \boldsymbol{X}, \sigma^2] = (\Sigma_o^{-1} + (X^T X)^{-1}/\sigma^2)^{-1}$$

How do we specify $\boldsymbol{\beta}_0$ and Σ_0?

The g-prior

To do the *least amount of calculus*, we can put a *g-prior* on β

$$\beta \mid \boldsymbol{X}, \boldsymbol{z} \sim MVN(0, g\,\sigma^2(\boldsymbol{X}^T\boldsymbol{X})^{-1}).$$

$$\beta_n = E[\beta \mid \boldsymbol{y}, \boldsymbol{X}, \sigma^2] = \frac{g}{g+1}(\Sigma_o^{-1} + (X^T X)^{-1}/\sigma^2)^{-1} = \frac{g}{g+1}\hat{\beta}_{ols}$$

$$\Sigma_n = \text{Var}[\beta \mid \boldsymbol{y}, \boldsymbol{X}, \sigma^2] = \frac{g}{g+1}(X^T X)^{-1}/\sigma^2)^{-1} = \frac{g}{g+1}\text{Var}[\hat{\beta}_{ols}]$$

- g shrinks the coefficients and can prevent overfitting to the data
- if $g = n$, then as n increases, inference approximates that using $\hat{\beta}_{ols}$

Do an application and example now