Module 11: Linear Regression

Rebecca C. Steorts

Announcements

- Today is the last class
- Homework 7 has been extended to Thursday, April 20, 11 PM.
- There will be no lab tomorrow.
- There will be office hours this week.
- Optional review class next Tuesday, April 25th.
- Module 9 has been updated to match the Hoff book's notation.

Agenda

- What is linear regression
- Motivating Example
- Application from Hoff

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

Health Insurance Example

- We want to predict whether or not a patient has health insurance based upon one covariate or predictor variable, income.
- Typically, we have many predictor variables, such as income, age, education level, etc.
- We store the predictor variables in a matrix $X_{n \times p}$.

Normal Regression Model

The Normal regression model specifies that

- $E[Y \mid x]$ is linear and
- the sampling variability around the mean is independent and identically (iid) from a normal distribution

$$Y_i = \beta^T x_i + e_i \tag{1}$$

$$e_1, \ldots, e_n \overset{iid}{\sim} Normal(0, \sigma^2)$$

Normal Regression Model (continued)

This allows us to write down

$$p(y_1, \ldots, y_n \mid x_1, \ldots x_n, \beta, \sigma^2) \tag{2}$$

$$= \prod_{i=1}^{n} p(y_i \mid x_i, \beta, \sigma^2) \tag{3}$$

$$(2\pi\sigma^2)^{-n/2} \exp\{\frac{-1}{2\sigma^2} \sum_{i=1}^{n} (y_i - \beta^T x_i)^2\} \tag{4}$$

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

$$y \mid X, \beta, \sigma^2 \sim MVN(X\beta, \sigma^2 I)$$
$$\beta \sim MVN(0, \tau^2 I)$$

The likelihood in the multivariate setting simpifies to

$$p(y_1, \ldots, y_n \mid x_1, \ldots x_n, \beta, \sigma^2) \tag{5}$$
$$(2\pi\sigma^2)^{-n/2} \exp\{\frac{-1}{2\sigma^2} \sum_{i=1}^{n} (y_i - \beta^T x_i)^2\} \tag{6}$$
$$(2\pi\sigma^2)^{-n/2} \exp\{\frac{-1}{2\sigma^2} (y - X\beta)^T (y - X\beta)\} \tag{7}$$

Posterior computation

Let $a = 1/\sigma^2$ and $b = 1/\tau^2$.

$$
\begin{aligned}
p(\beta \mid y, X) &\propto p(y \mid X, \beta) p(\beta) && (8)\\
&\propto \exp\{-a/2(y - X\beta)^T(y - X\beta)\} \times \exp\{-b/2\beta^T\beta)\} && (9)
\end{aligned}
$$

Just like in the Multivariate modules, we just simplify. (Check these details on your own).

$$p(\beta \mid y, X) \propto MVN(\beta \mid y, X, \Lambda^{-1})$$

where $\Lambda = aX^TX + bI$ and $\mu = a\Lambda^{-1}X^Ty$.

Posterior computation (details)

$$
\begin{aligned}
& p(\beta \mid y, X) && (10)\\
& \propto \exp\{-\frac{a}{2}(y - X\beta)^T(y - X\beta)\} \times \exp\{-\frac{b}{2}\beta^T\beta)\} && (11)\\
& \propto \exp\{-\frac{a}{2}[y^T y - 2\beta^T X^T y + \beta^T X^T X\beta] - \frac{b}{2}\beta^T\beta\} && (12)\\
& \propto \exp\{a\beta^T X^T y - \frac{a}{2}\beta^T X^T X\beta - b/2\beta^T\beta\} && (13)\\
& \propto \exp\{a\beta^T[X^T y] - 1/2\beta^T(aX^T X + bI)\beta\} && (14)
\end{aligned}
$$

Then $\Lambda = aX^T X + bI$ and $\mu = a\Lambda^{-1}X^T y$.

Linear Regression Applied to Swimming

- We will consider Exercise 9.1 in Hoff very closely to illustrate linear regression.
- The data set we consider contains times (in seconds) of four high school swimmers swimming 50 yards.
- There are 6 times for each student, taken every two weeks.
- Each row corresponds to a swimmer and a higher column index indicates a later date.

Data set

```r
read.table("https://www.stat.washington.edu/~pdhoff/Book/Da
```

```
##      V1   V2   V3   V4   V5   V6
## 1 23.1 23.2 22.9 22.9 22.8 22.7
## 2 23.2 23.1 23.4 23.5 23.5 23.4
## 3 22.7 22.6 22.8 22.8 22.9 22.8
## 4 23.7 23.6 23.7 23.5 23.5 23.4
```

Full conditionals (Task 1)

We will fit a separate linear regression model for each swimmer, with swimming time as the response and week as the explanatory variable. Let $Y_i \in \mathbb{R}^6$ be the 6 recorded times for swimmer i. Let

$$X_i = \begin{bmatrix} 1 & 1 \\ 1 & 3 \\ \dots & \\ 1 & 9 \\ 1 & 11 \end{bmatrix}$$

be the design matrix for swimmer i. Then we use the following linear regression model:

$$\begin{aligned} Y_i &\sim \mathcal{N}_6\left(X\beta_i, \tau_i^{-1}\mathcal{I}_6\right) \\ \beta_i &\sim \mathcal{N}_2\left(\beta_0, \Sigma_0\right) \\ \tau_i &\sim \text{Gamma}(a, b). \end{aligned}$$

Derive full conditionals for β_i and τ_i.

Solution (Task 1)

The conditional posterior for β_i is multivariate normal with

$$\mathbb{V}[\beta_i \mid Y_i, X_i, \tau_i] = (\Sigma_0^{-1} + \tau X_i^T X_i)^{-1}$$
$$\mathbb{E}[\beta_i \mid Y_i, X_i, \tau_i] = (\Sigma_0^{-1} + \tau_i X_i^T X_i)^{-1}(\Sigma_0^{-1}\beta_0 + \tau_i X_i^T Y_i).$$

while

$$\tau_i \mid Y_i, X_i, \beta \sim \text{Gamma}\left(a + 3\,,\ b + \frac{(Y_i - X_i\beta_i)^T(Y_i - X_i\beta_i)}{2}\right).$$

These can be found in in Hoff in section 9.2.1.

Task 2

Complete the prior specification by choosing a, b, β_0, and Σ_0. Let your choices be informed by the fact that times for this age group tend to be between 22 and 24 seconds.

Solution (Task 2)

Choose $a = b = 0.1$ so as to be somewhat uninformative.

Choose $\beta_0 = [23 \;\; 0]^T$ with

$$\Sigma_0 = \begin{bmatrix} 5 & 0 \\ 0 & 2 \end{bmatrix}.$$

This centers the intercept at 23 (the middle of the given range) and the slope at 0 (so we are assuming no increase) but we choose the variance to be a bit large to err on the side of being less informative.

Gibbs sampler (Task 3)

Code a Gibbs sampler to fit each of the models. For each swimmer i, obtain draws from the posterior predictive distribution for y_i^*, the time of swimmer i if they were to swim two weeks from the last recorded time.

Posterior Prediction (Task 4)

The coach has to decide which swimmer should compete in a meet two weeks from the last recorded time. Using the posterior predictive distributions, compute $\Pr\left\{y_i^* = \max\left(y_1^*, y_2^*, y_3^*, y_4^*\right)\right\}$ for each swimmer i and use these probabilities to make a recommendation to the coach.

- This is left as an exercise.