STAT 525 Lecture 16

October 26, 2017

1 Topics

1. Linear regression
2. Multiple testing
3. Prediction

Question: How does y vary as a function of X?

Interested in $p\left(\boldsymbol{y} \mid \boldsymbol{X}, \theta\right)$

Assume $(y_i, \boldsymbol{x}_i)_{i=1}^n$ are exchangeable.

$n = \#$ of observations, $p = \#$ of predictors

2 Linear model

$$\begin{aligned}\mathbb{E}\left[y_i \mid \boldsymbol{x}, \boldsymbol{\beta}\right] &= \beta_1 x_1 + \cdots + \beta_p x_p \\ &= \boldsymbol{x}^T \boldsymbol{\beta}\end{aligned}$$

Ordinary linear model

$$\begin{aligned}\mathrm{Var}\left[y_i \mid \boldsymbol{x}, \boldsymbol{\theta}\right] &= \sigma^2 \\ \boldsymbol{\theta} &= \begin{bmatrix} \boldsymbol{\beta} & \sigma \end{bmatrix}\end{aligned}$$

Normal linear model

$$\boldsymbol{y}_{n\times 1} = \boldsymbol{X}_{n\times p}\boldsymbol{\beta}_{p\times 1} + \boldsymbol{\epsilon}_{n\times 1}$$

where

$$\begin{aligned}\epsilon_i &\overset{\text{iid}}{\sim} N\left(0, \sigma^2\right) \\ \boldsymbol{\epsilon} &\sim MVN\left(0, \sigma^2 \mathbb{I}\right)\end{aligned}$$

Key points

1. Flexible choices for $\boldsymbol{X}$
2. Transformation of $\boldsymbol{X}$ and / or $\boldsymbol{y}$ for linearity and Gaussian of $\boldsymbol{\epsilon}$.
3. Generation for other variance model.

3 Bayesian Conditional Modeling

Data $(\boldsymbol{y}, \boldsymbol{X})$

Likelihood:

$$p(\boldsymbol{y}, \boldsymbol{X} \mid \boldsymbol{\theta}, \psi)$$

Priors

$$p(\boldsymbol{\theta}, \psi)$$

Posterior

$$p(\boldsymbol{\theta}, \psi \mid \boldsymbol{y}, \boldsymbol{X}) = p(\psi \mid \boldsymbol{\theta}, \boldsymbol{y}, \boldsymbol{X})\, p(\boldsymbol{\theta} \mid \boldsymbol{y}, \boldsymbol{X})$$

Assume

$$p(\psi \mid \boldsymbol{\theta}, \boldsymbol{y}, \boldsymbol{X}) = p(\psi \mid \boldsymbol{X})$$

- Prior independence: $p(\boldsymbol{\theta}, \psi) = p(\boldsymbol{\theta})\, p(\psi)$
- ψ conditional independent of $\boldsymbol{y}$ given $\boldsymbol{X}$

$$p(\boldsymbol{\theta}, \psi \mid \boldsymbol{y}, \boldsymbol{X}) = p(\psi \mid \boldsymbol{X})\, p(\boldsymbol{\theta} \mid \boldsymbol{y}, \boldsymbol{X})$$

Posterior inference for $\boldsymbol{\theta}$ only need the density $p(\boldsymbol{\theta} \mid \boldsymbol{y}, \boldsymbol{X})$ or

$$p(\boldsymbol{\theta} \mid \boldsymbol{y}, \boldsymbol{X}) \propto p(\boldsymbol{y} \mid \boldsymbol{\theta}, \boldsymbol{X})\, p(\boldsymbol{\theta} \mid \boldsymbol{X})$$

Commentary on the intercept:

1. if $\boldsymbol{x}_{i1} = 1$ for all i, usually assume $p(\beta_1) \propto 1$
2. Usually center and scale $\boldsymbol{y}$ and columns of $\boldsymbol{X}$

$$y_i = \frac{y_i - \text{mean}(\boldsymbol{y})}{\text{sd}(\boldsymbol{y})}$$

4 General Case

The prior on $\boldsymbol{\beta}$ is

$$\boldsymbol{\beta} \sim \text{MVN}\left(\boldsymbol{\mu}_\beta, \boldsymbol{\Sigma}_\beta\right)$$

Then the likelihood is

$$\boldsymbol{y} \mid \boldsymbol{X}, \boldsymbol{\beta}, \sigma^2 \sim MVN\left(\boldsymbol{X}\boldsymbol{\beta}, \sigma^2 \mathbb{I}\right)$$

The posterior is

$$\begin{aligned}
p\left(\boldsymbol{\beta} \mid \boldsymbol{y}, \boldsymbol{X}, \sigma^2\right) &\propto p\left(\boldsymbol{y} \mid \boldsymbol{X}, \boldsymbol{\beta}, \sigma^2\right) p(\boldsymbol{\beta}) \\
&\propto \left|2\pi\sigma^2\mathbb{I}\right|^{-\frac{1}{2}} \exp\left(-\frac{1}{2}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^T \left(\sigma^2\mathbb{I}\right)^{-1} (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})\right) \times \\
&\left|2\pi\Sigma_{\boldsymbol{\beta}}\right|^{-\frac{1}{2}} \exp\left(-\frac{1}{2}\left(\boldsymbol{\beta} - \boldsymbol{\mu}_\beta\right)^T (\Sigma_\beta)^{-1} \left(\boldsymbol{\beta} - \boldsymbol{\mu}_\beta\right)\right) \\
&\propto \exp\left(-\frac{1}{2}\sigma^{-2}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^T (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) + \left(\boldsymbol{\beta} - \boldsymbol{\mu}_\beta\right)^T \boldsymbol{\Sigma}_\beta \left(\boldsymbol{\beta} - \boldsymbol{\mu}_\beta\right)\right) \\
&\propto \exp\left(-\frac{1}{2}\left(\boldsymbol{\beta}^T \boldsymbol{Q}_\beta \boldsymbol{\beta} - 2\boldsymbol{\beta}\boldsymbol{l}_\beta\right)\right)
\end{aligned}$$

where

$$\boldsymbol{Q}_\beta = \sigma^{-2}\boldsymbol{X}^T\boldsymbol{X} + \boldsymbol{\Sigma}_\beta^{-1}$$
$$\boldsymbol{l}_\beta = \sigma^{-2}\boldsymbol{X}\boldsymbol{y} + \boldsymbol{\Sigma}_\beta^{-1}\boldsymbol{\mu}_\beta$$

Therefore, the posterior for $\boldsymbol{\beta}$ is

$$\left[\boldsymbol{\beta} \mid \boldsymbol{y}, \boldsymbol{X}, \sigma^2\right] \sim MVN\left(\boldsymbol{Q}_\beta^{-1}\boldsymbol{l}_\beta, \boldsymbol{Q}_\beta^{-1}\right)$$

5 g-prior (Zeller 1986)

$$\left[\boldsymbol{\beta} \mid \sigma^2, \boldsymbol{X}, g\right] \sim MVN\left(\boldsymbol{\mu}_\beta, \underbrace{g\sigma^2\left[\boldsymbol{X}^T\boldsymbol{X}\right]^{-1}}_{\boldsymbol{\Sigma}_\beta}\right)$$

We plug the new $\boldsymbol{\mu}_\beta$ and $\boldsymbol{\Sigma}_\beta$ into the section 4, then

$$\begin{aligned}
\hat{\boldsymbol{Q}}_\beta &= \sigma^{-2}\boldsymbol{X}^T\boldsymbol{X} + g^{-1}\sigma^{-2}\boldsymbol{X}^T\boldsymbol{X} \\
&= \sigma^{-2}\boldsymbol{X}^T\boldsymbol{X}\left[1 + g^{-1}\right] \\
\hat{\boldsymbol{l}}_\beta &= \sigma^{-2}\boldsymbol{X}^T\boldsymbol{y} + g^{-1}\sigma^{-2}\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{\mu}_\beta \\
&= \sigma^{-2}\boldsymbol{X}^T\left(\boldsymbol{y} + g^{-1}\boldsymbol{X}\boldsymbol{\mu}_\beta\right)
\end{aligned}$$

Therefore we have

$$\begin{aligned}
\boldsymbol{Q}_\beta^{-1} &= \frac{g}{g+1}\sigma^2\left[\boldsymbol{X}^T\boldsymbol{X}\right]^{-1} \\
\boldsymbol{Q}_\beta^{-1}\boldsymbol{l}_\beta &= \frac{g}{g+1}\cancel{\sigma^2}\left[\boldsymbol{X}^T\boldsymbol{X}\right]^{-1}\cancel{\sigma^{-2}}\boldsymbol{X}^T\left(\boldsymbol{y} + g^{-1}\boldsymbol{X}\boldsymbol{\mu}_\beta\right) \\
&= \frac{g}{g+1}\underbrace{\left[\boldsymbol{X}^T\boldsymbol{X}\right]^{-1}\boldsymbol{X}^T\boldsymbol{y}}_{\hat{\boldsymbol{\beta}}_{\text{ols}}} + \frac{1}{g+1}\boldsymbol{\mu}_\beta
\end{aligned}$$

The posterior expectation for $\boldsymbol{\beta}$ is

$$\mathbb{E}\left[\boldsymbol{\beta} \mid \boldsymbol{X}, \boldsymbol{y}, \sigma^2, g\right] = (1-\kappa)\hat{\boldsymbol{\beta}} + \kappa\boldsymbol{\mu}_\beta$$
$$\kappa = \frac{1}{g+1}$$

The posterior for $\boldsymbol{\beta}$ is

$$\left[\boldsymbol{\beta} \mid \boldsymbol{X}, \boldsymbol{y}, \sigma^2, g\right] \sim MVN\left((1-\kappa)\boldsymbol{\beta} + \kappa\boldsymbol{\mu}_\beta, \frac{g}{g+1}\sigma^2\left[\boldsymbol{X}^T\boldsymbol{X}\right]^{-1}\right)$$

Invariance of g prior

Let $\boldsymbol{X}^* = \boldsymbol{X}\boldsymbol{H}$ for some $p \times p$ $\boldsymbol{H}$

Posterior for $\boldsymbol{\beta} \mid \boldsymbol{X}, \boldsymbol{y}$ shall be the same as $\boldsymbol{H}\boldsymbol{\beta}^* \mid y, \boldsymbol{X}^*$

stratify when $\boldsymbol{\mu}_\beta \to 0$ and $\boldsymbol{\Sigma}_\beta = \kappa\left(\boldsymbol{X}^T X\right)^{-1}$ for $\kappa > 0$.

Flat prior

A common non-informative choice is

$$p\left(\boldsymbol{\beta}, \sigma^2\right) = p\left(\boldsymbol{\beta} \mid \sigma^2\right) p\left(\sigma^2\right) \propto \frac{1}{\sigma^2}$$

Special case of g prior (in the posterior) when $g \to \infty$ and $p\left(\sigma^2\right) \propto \dfrac{1}{\sigma^2}$

when $g \to \infty$ then

$$\boldsymbol{\beta} \mid \boldsymbol{y}, \boldsymbol{X}, \sigma^2 \sim \left(\hat{\boldsymbol{\beta}}_{\text{ols}}, \sigma^2 \left[\boldsymbol{X}^T \boldsymbol{X}\right]^{-1}\right)$$

6 Penalized regression

6.1 Ridge regression

Suppose $\beta_j \mid \sigma_\beta \overset{\text{iid}}{\sim} N\left(0, \sigma_\beta^2\right)$ for $j = 1, \cdots, p$

The log-posterior (up to an additive constant) is

$$\begin{aligned}
\log p\left(\boldsymbol{\beta} \mid \boldsymbol{y}, \boldsymbol{X}, \sigma^2, \sigma_\beta^2\right) &\propto \log p\left(\boldsymbol{y} \mid \boldsymbol{X}, \boldsymbol{\beta}, \sigma^2, \sigma_\beta^2\right) + \log p\left(\boldsymbol{\beta}\right) \\
&\propto -\frac{1}{2\sigma^2} \sum_{i=1}^{n} \left\|y_i - \boldsymbol{x}_i^T \boldsymbol{\beta}\right\|^2 + \sum_{j=1}^{p} \log p\left(\beta_j\right) \\
&\propto -\frac{1}{2\sigma^2} \left\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\right\|^2 - \frac{1}{2\sigma_\beta^2} \sum_{j=1}^{p} \beta_j^2 \\
&\propto -\frac{1}{2\sigma^2} \underbrace{\left\{\left\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\right\|^2 + \lambda \sum_{j=1}^{p} \beta_j^2\right\}}_{\text{Ridge Regression}}
\end{aligned}$$

where

$$\lambda = \frac{\sigma^2}{\sigma_\beta^2}$$

The posterior expectation for $\boldsymbol{\beta}$ is

$$\mathbb{E}\left[\boldsymbol{\beta} \mid \boldsymbol{y}, \boldsymbol{X}, \sigma^2, \lambda\right] = \underbrace{\left[\boldsymbol{X}^T \boldsymbol{X} + \lambda \mathbb{I}\right]^{-1} \boldsymbol{X}^T \boldsymbol{y}}_{\text{Ridge Estimator}}$$

λ is the tuning parameter for freq. For Bayesian $\lambda = \frac{\sigma^2}{\sigma_\beta^2}$. We just need to estimate σ_β^2 and put a prior on σ_β^2.

Key points

1. Posterior mode (and here posterior mean) solves the optimization problem

$$\begin{aligned}
\hat{\boldsymbol{\beta}}_R &= \arg\min_{\boldsymbol{\beta}} \left\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\right\|^2 + \lambda \sum_{j=1}^{p} \beta_j^2 \\
&= \arg\min_{\boldsymbol{\beta}} \left\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\right\|^2 \text{ s.t. } \left\|\boldsymbol{\beta}\right\|^2 \leq s
\end{aligned}$$

(a) for $\lambda > 0$, allowed to have $p \geq n$

(b) Ridge regression is useful for $p > n$ and / or correlated predictions.

2. Lasso regression

$$\hat{\boldsymbol{\beta}}_L = \arg\min_{\boldsymbol{\beta}} \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|^2 + \lambda \sum_{j=1}^{p} |\beta_j|$$

(a) often have $\hat{\beta}_{L,j} = 0$ for many j provides sparse solution.

(b) overshrinkage large β_j