STAT 431 — Applied Bayesian Analysis — Course Notes

Linear Regression

Fall 2022

Introduction

Regression is modeling the mean dependence of a **response** variable Y on **predictors** $X_1, \ldots, X_p$.

The response is generally called the **dependent variable** and predictors are **independent variables** (sometimes called **covariates**).

We observe

$$(Y_i, X_{i1}, \ldots, X_{ip}), \qquad i = 1, \ldots, n$$

but usually only Y_i is regarded as random. The values X_{ij} are regarded as constants (fixed and known).

Simple Linear Regression

Simple linear regression assumes the mean of Y is a straight-line function of a predictor X:

$$\mathrm{E}(Y \mid \beta_1, \beta_2) \;=\; \beta_1 + \beta_2 X$$

where β_1 and β_2 are parameters: the **regression coefficients**.

[Graph:]

We model the observed pairs (Y_i, X_i), $i = 1, \ldots, n$ as

$$Y_i \;\; = \;\; \beta_1 \, + \, X_i\beta_2 \, + \, \varepsilon_i$$

where the mean-zero **errors** ε_i are usually assumed to be (conditionally) iid normal:

$$\varepsilon_i \mid \beta_1, \beta_2, \sigma^2 \quad \sim \quad iid \;\, \mathrm{Normal}(0, \sigma^2)$$

with **error variance** σ^2 as an additional parameter.

Thus,

$$Y_i \mid \beta_1, \beta_2, \sigma^2 \quad \sim \quad indep \;\, \mathrm{Normal}(\beta_1 + X_i\beta_2, \sigma^2)$$

and we need a prior

$$\pi(\beta_1, \beta_2, \sigma^2)$$

To study the most common types of regression priors and their posteriors, we need a multivariate generalization of the normal distribution ...

Multivariate Normal

For the $m \times 1$ **random vector**

$$\boldsymbol{Z} \;=\; \begin{bmatrix} Z_1 \\ \vdots \\ Z_m \end{bmatrix}$$

we write

$$\boldsymbol{Z} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma} \;\sim\; \mathrm{Normal}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$$

for $m \times 1$ $\boldsymbol{\mu}$ and $m \times m$ symmetric invertible $\boldsymbol{\Sigma}$ when $\boldsymbol{Z}$ has (joint) PDF

$$f(\boldsymbol{z} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}) \;=\; \frac{1}{(2\pi)^{m/2}\, |\boldsymbol{\Sigma}|^{1/2}}\, e^{-\frac{1}{2}(\boldsymbol{z}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{z}-\boldsymbol{\mu})}$$

It turns out that

$$\mathrm{E}(\boldsymbol{Z}) \quad \equiv \quad \begin{bmatrix} \mathrm{E}(Z_1) \\ \vdots \\ \mathrm{E}(Z_m) \end{bmatrix} \quad = \quad \boldsymbol{\mu}$$

and

$$\mathrm{Cov}(\boldsymbol{Z}) \quad \equiv \quad \begin{bmatrix} \mathrm{Var}(Z_1) & \mathrm{Cov}(Z_1, Z_2) & \cdots & \mathrm{Cov}(Z_1, Z_m) \\ \mathrm{Cov}(Z_2, Z_1) & \mathrm{Var}(Z_2) & & \vdots \\ \vdots & & \ddots & \vdots \\ \mathrm{Cov}(Z_m, Z_1) & \cdots & \cdots & \mathrm{Var}(Z_m) \end{bmatrix}$$

$$= \quad \boldsymbol{\Sigma}$$

It also turns out that the marginal distribution of each Z_i is normal, and the conditional distribution of each Z_i given all of the others is normal.

Also,

$$Z_i \mid \mu_i, \sigma^2 \quad \sim \quad indep \ \mathrm{Normal}(\mu_i, \sigma^2)$$

if and only if

$$\boldsymbol{Z} \mid \boldsymbol{\mu}, \sigma^2 \quad \sim \quad \mathrm{Normal}(\boldsymbol{\mu}, \sigma^2 \boldsymbol{I}_m)$$

where $\boldsymbol{I}_m$ is the $m \times m$ identity matrix.

Linear Regression

A **linear regression** model assumes

$$E(Y_i \mid \beta_1, \ldots, \beta_p) = \sum_{j=1}^{p} X_{ij} \beta_j \qquad i = 1, \ldots, n$$

for **(regression) coefficients** $\beta_1, \ldots, \beta_p$.

β_1 is usually an intercept:

$$X_{i1} \equiv 1 \qquad i = 1, \ldots, n$$

(Simple linear regression is the special case $p = 2$.)

Letting

$$\boldsymbol{Y} = \begin{bmatrix} Y_1 \\ \vdots \\ \vdots \\ Y_n \end{bmatrix} \qquad \boldsymbol{X} = \begin{bmatrix} X_{11} & \cdots & X_{1p} \\ \vdots & & \vdots \\ \vdots & & \vdots \\ X_{n1} & \cdots & X_{np} \end{bmatrix} \qquad \boldsymbol{\beta} = \begin{bmatrix} \beta_1 \\ \vdots \\ \beta_p \end{bmatrix}$$

we can write the linear regression as

$$\mathrm{E}(\boldsymbol{Y} \mid \boldsymbol{\beta}) = \boldsymbol{X}\boldsymbol{\beta}$$

We will assume that $\boldsymbol{X}^T\boldsymbol{X}$ is invertible (which makes $\boldsymbol{\beta}$ well-defined).

The usual normality assumption

$$Y_i \;=\; \sum_{j=1}^{p} X_{ij}\beta_j \;+\; \varepsilon_i$$

$$\varepsilon_i \mid \beta_1, \ldots, \beta_p, \sigma^2 \;\sim\; iid\ \mathrm{Normal}(0, \sigma^2)$$

is then equivalent to

$$\boldsymbol{Y} \mid \boldsymbol{\beta}, \sigma^2 \;\sim\; \mathrm{Normal}(\boldsymbol{X}\boldsymbol{\beta}, \sigma^2 \boldsymbol{I}_n)$$

We will need a prior $\pi(\boldsymbol{\beta}, \sigma^2)$.

Summary Statistics

The **(ordinary) least squares** estimator of $\boldsymbol{\beta}$ is

$$\hat{\boldsymbol{\beta}}_{LS} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y}$$

and typical estimators of σ^2 include

$$s^2 = \frac{1}{n-p}(\boldsymbol{Y}-\boldsymbol{X}\hat{\boldsymbol{\beta}}_{LS})^T(\boldsymbol{Y}-\boldsymbol{X}\hat{\boldsymbol{\beta}}_{LS}) \qquad (n > p)$$

$$\hat{\sigma}^2 = \frac{n-p}{n}s^2$$

Remark: $(\hat{\boldsymbol{\beta}}_{LS}, s^2)$ (or $(\hat{\boldsymbol{\beta}}_{LS}, \hat{\sigma}^2)$) is sufficient for $(\boldsymbol{\beta}, \sigma^2)$.

Priors

We will consider these kinds of prior:

- Jeffreys'

$$\pi(\boldsymbol{\beta}, \sigma^2) \quad \propto \quad \frac{1}{(\sigma^2)^{p/2+1}} \qquad (\sigma^2 > 0)$$

- "standard" noninformative

$$\pi(\boldsymbol{\beta}, \sigma^2) \quad \propto \quad \frac{1}{\sigma^2} \qquad (\sigma^2 > 0)$$

- conditional multivariate normal (zero-centered)

$$\boldsymbol{\beta} \mid \sigma^2 \quad \sim \quad \mathrm{Normal}(\mathbf{0}, \sigma^2 \boldsymbol{\Omega})$$

Jeffreys' Prior

If $\hat{\sigma}^2 > 0$, the prior

$$\pi(\boldsymbol{\beta}, \sigma^2) \quad \propto \quad \frac{1}{(\sigma^2)^{p/2+1}} \qquad (\sigma^2 > 0)$$

leads to proper posterior

$$\boldsymbol{\beta} \mid \sigma^2, \boldsymbol{y} \quad \sim \quad \mathrm{Normal}\big(\hat{\boldsymbol{\beta}}_{LS},\, \sigma^2(\boldsymbol{X}^T\boldsymbol{X})^{-1}\big)$$

$$\sigma^2 \mid \boldsymbol{y} \quad \sim \quad \mathrm{InvGamma}\left(\frac{n}{2},\, \frac{n}{2}\hat{\sigma}^2\right)$$

The marginal posterior distributions for the coefficients turn out to be

$$\beta_j \mid \boldsymbol{y} \;\; \sim \;\; \mathrm{t}_n\big(\hat{\beta}_{LS,j},\, \hat{\sigma}^2 c_{jj}\big)$$

where c_{jj} is the jth diagonal element of $(\boldsymbol{X}^T\boldsymbol{X})^{-1}$.

These can be used to form individual credible intervals and perform individual tests for the β_js, but they will not match the usual confidence intervals and frequentist tests.

"Standard" Noninformative Prior

If $s^2 > 0$, the prior

$$\pi(\boldsymbol{\beta}, \sigma^2) \quad \propto \quad \frac{1}{\sigma^2} \qquad (\sigma^2 > 0)$$

leads to proper posterior

$$\begin{aligned} \boldsymbol{\beta} \mid \sigma^2, \boldsymbol{y} \quad &\sim \quad \mathrm{Normal}\big(\hat{\boldsymbol{\beta}}_{LS},\, \sigma^2(\boldsymbol{X}^T\boldsymbol{X})^{-1}\big) \\ \sigma^2 \mid \boldsymbol{y} \quad &\sim \quad \mathrm{InvGamma}\left(\frac{n-p}{2},\, \frac{n-p}{2}s^2\right) \end{aligned}$$

The marginal posterior distributions for the coefficients turn out to be

$$\beta_j \mid \boldsymbol{y} \;\;\sim\;\; \mathrm{t}_{n-p}\big(\hat{\beta}_{LS,j},\, s^2 c_{jj}\big)$$

where c_{jj} is the jth diagonal element of $(\boldsymbol{X}^T\boldsymbol{X})^{-1}$.

The credible intervals and tests obtained from this do turn out to match frequentist confidence intervals and tests.

Conditionally Normal Prior

Conditioning on σ^2 (as if it is fixed) and letting $p \times p$ symmetric matrix $\mathbf{\Omega}$ be invertible, the conditional prior

$$\boldsymbol{\beta} \mid \sigma^2 \quad \sim \quad \text{Normal}(\mathbf{0}, \sigma^2 \mathbf{\Omega})$$

leads to conditional posterior

$$\boldsymbol{\beta} \mid \sigma^2, \boldsymbol{y} \quad \sim$$

$$\text{Normal}\Big(\big(\boldsymbol{X}^T \boldsymbol{X} + \mathbf{\Omega}^{-1}\big)^{-1} \boldsymbol{X}^T \boldsymbol{Y},\ \sigma^2 \big(\boldsymbol{X}^T \boldsymbol{X} + \mathbf{\Omega}^{-1}\big)^{-1} \Big)$$

(Further allowing $\boldsymbol{\beta}$ an arbitrary prior mean would make the multivariate normal semi-conjugate for $\boldsymbol{\beta}$.)

Remark: This does not require $n > p$ or invertible $\boldsymbol{X}^T\boldsymbol{X}$.

The typical effect of such a proper prior is to "shrink" the coefficient estimates toward zero.

When the non-intercept predictors in $\boldsymbol{X}$ have been standardized (by subtracting sample means and dividing by sample standard deviations) and we take

$$\boldsymbol{\Omega} \;=\; \begin{bmatrix} \omega_{11} & \\ & \boldsymbol{I}_{p-1}/\lambda \end{bmatrix}$$

there is a connection with *ridge regression* — see BSM, Sec. 4.2.2.

We can further give σ^2 some kind of prior, such as inverse gamma (to be conjugate) or

$$\pi(\sigma^2) \quad \propto \quad \frac{1}{\sigma^2} \qquad\qquad (\sigma^2 > 0)$$

(to be noninformative).

Remarks

- Other kinds of priors are available that “shrink” the coefficients differently, based on the data.

  If interested, see BSM, Sec. 4.2.3.

- Posterior prediction (at “new” predictor values) can be performed by simulation (BSM, Sec. 4.2.4) or sometimes with exact formulas.