Linear Regression

Statistics Honours 2020

February 25, 2024

References

Decision Theory and Bayesian Statistics: Chapters 5.3

Linear Regression: Simplest case

Consider a linear model,

$$y_i = \beta_0 + \beta_1 x_i + e_i$$

where we assume we know $e_i \sim N(0, 1)$.

The aim is to obtain the posterior distribution of β_0 and β_1 given observed data

$$\boldsymbol{y}^T = (y_1, \ldots, y_n)$$

and

$$\boldsymbol{x}^T = (x_1, \ldots, x_n).$$

Note that the linear model can be rewritten as

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e}$$

where $[\boldsymbol{y}|\boldsymbol{\beta}, \boldsymbol{X}] \sim \mathcal{N}_n\left(\boldsymbol{X}\boldsymbol{\beta}, \boldsymbol{I}_n\right)$. This is also the likelihood!

Linear Regression: Simplest case

Since the **regression coefficients** are location parameters we can assume an uninformative prior distribution to it! i.e.

$$[\beta] \propto 1.$$

This is not a proper pdf.

$$\begin{aligned}
[\beta|\boldsymbol{X}, \boldsymbol{y}] &\propto [\beta][\boldsymbol{y}|\beta, \boldsymbol{X}] \\
&\propto [\boldsymbol{y}|\beta, \boldsymbol{X}] \\
&\propto \exp\left(-\frac{1}{2}\sum_i (y_i - \beta_0 - \beta_1 x_i)^2\right) \\
&\propto \exp\left(-\frac{1}{2}(\boldsymbol{y} - \boldsymbol{X}\beta)^T(\boldsymbol{y} - \boldsymbol{X}\beta)\right)
\end{aligned}$$

We now need to complete the square in order to obtain the distribution of β.

Linear Regression: Simplest case

Recall that if $\boldsymbol{X} = (X_1, \ldots, X_p)^T \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ then

$$f(\boldsymbol{x}) \propto \frac{1}{|\boldsymbol{\Sigma}|^{1/2}} \exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right).$$

Linear Regression: Simplest case

Let's expand the exponent.

$$
\begin{aligned}
(\boldsymbol{y} - \boldsymbol{X}\beta)^T(\boldsymbol{y} - \boldsymbol{X}\beta) &= \boldsymbol{y}^T\boldsymbol{y} - \beta^T\boldsymbol{X}^T\boldsymbol{Y} - \boldsymbol{y}^T\boldsymbol{X}\beta + \beta^T\boldsymbol{X}^T\boldsymbol{X}\beta \\
&= \boldsymbol{\beta}\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{\beta} - 2\boldsymbol{y}^T\boldsymbol{X}\beta + \boldsymbol{y}^T\boldsymbol{y} \\
&= \boldsymbol{\beta}\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{\beta} - 2\boldsymbol{y}^T\boldsymbol{X}\beta + \ldots \text{(drop terms independent of } \boldsymbol{\beta}\text{)}
\end{aligned}
$$

Now we have to complete the square :(

Based on the MVN density, we're working towards:

$$
(\boldsymbol{\beta} - \boldsymbol{\mu}^*)^T\Sigma^{*-1}(\boldsymbol{\beta} - \boldsymbol{\mu}^*) = \boldsymbol{\beta}^T\Sigma^{*-1}\boldsymbol{\beta} - 2\boldsymbol{\mu}^{*T}\Sigma^{*-1}\boldsymbol{\beta} + \ldots
$$

Hence

$$
\Sigma^{*-1} = \boldsymbol{X}^T\boldsymbol{X} \Rightarrow \Sigma^* = \left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}
$$

and

$$
\boldsymbol{\mu}^{*T}\Sigma^{*-1} = \boldsymbol{y}^T\boldsymbol{X} \Rightarrow \boldsymbol{\mu}^* = \left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T\boldsymbol{y}
$$

Solve for what you need...

Linear Regression: Simplest case

- We have shown that the posterior distribution of the regression parameters are **bivariate normal**.
- The **mean equation** is the same as the maximum likelihood estimates.
- The covariance matrix is of a similar form as well!

Linear Regression: Non-informative priors

What if the residual variance is unknown???
That means our linear model is now:

$$\boldsymbol{y} = \boldsymbol{X}\beta + \boldsymbol{e}$$

where $\left[\boldsymbol{y}|\boldsymbol{\beta}, \sigma^2, \boldsymbol{X}\right] \sim \mathcal{N}_n\left(\boldsymbol{X}\boldsymbol{\beta}, \sigma^2 \boldsymbol{I}_n\right)$.
In the case where the regression variance is unknown we require the posterior distribution:

$$[\boldsymbol{\beta}, \sigma^2|\boldsymbol{x}] \propto [\boldsymbol{\beta}, \sigma^2][\boldsymbol{y}|\boldsymbol{\beta}, \sigma^2, \boldsymbol{X}]$$

Linear Regression: Non-informative priors

Lets assume $[\beta] \propto 1$ and $[\sigma^2] \propto \frac{1}{\sigma^2}$ and are independent from each other, then it can be shown that

$$
\begin{aligned}
[\boldsymbol{\beta}, \sigma^2 | \boldsymbol{x}] &\propto [\beta][\sigma^2][\boldsymbol{y}|\beta, \sigma^2, \boldsymbol{X}] \\
&\propto \left(\sigma^2\right)^{-1} \frac{1}{\left|\sigma^2 \boldsymbol{I}_n\right|^{1/2}} \exp\left(-\frac{1}{2\sigma^2}(\boldsymbol{y} - \boldsymbol{X}\beta)^T(\boldsymbol{y} - \boldsymbol{X}\beta)\right) \\
&\propto \left(\sigma^2\right)^{-1} \left(\sigma^2\right)^{-n/2} \exp\left(-\frac{1}{2\sigma^2}(\boldsymbol{y} - \boldsymbol{X}\beta)^T(\boldsymbol{y} - \boldsymbol{X}\beta)\right) \\
&\propto (\sigma^2)^{-(n/2+1)} \exp\left(-\frac{1}{2\sigma^2}(\boldsymbol{y} - \boldsymbol{X}\beta)^T(\boldsymbol{y} - \boldsymbol{X}\beta)\right)
\end{aligned}
$$

Can you identify the ‘name’ of this joint density function??? ... NO!

However, the two conditional distributions have simple forms namely

1. $\pi(\boldsymbol{\beta}|\boldsymbol{x}, \sigma^2) \sim N\left(\boldsymbol{\mu}^*,\ \sigma^2\Sigma^*\right)$ and
2. $\pi(\sigma^2|\boldsymbol{x}, \beta) \sim \mathcal{IG}\left(\frac{n}{2},\ \frac{1}{2}(\boldsymbol{y} - \boldsymbol{X}\beta)^T(\boldsymbol{y} - \boldsymbol{X}\beta)\right)$

The **Gibbs algorithm** proceeds by sampling from

1. $\pi(\boldsymbol{\beta}|\boldsymbol{x}, \sigma^2)$ and
2. $\pi(\sigma^2|\boldsymbol{x}, \beta)$

in turn.

These draws can be viewed as being samples from the joint distribution, $\pi(\boldsymbol{\beta}, \sigma^2|\boldsymbol{x})$.

Gamma and inverse gamma distribution

If $X \sim \mathcal{G}(a, b)$ then

$$f(x) \propto x^{a-1} e^{-bx} \quad \text{for } x > 0, a > 0, b > 0.$$

If $Y = \frac{1}{X}$ then Y is said to be an inverse gamma random variable with

$$f(y) \propto y^{-(a+1)} e^{-b/y}.$$

Sampling from multivariate normal distributions in R

```
#sample from a bivariate normal distribution
set.seed(10)
require(MASS)
sigma<- matrix(c(1, -.5, -.5, 1), 2,2) #the covariance matrix
sigma

##      [,1] [,2]
## [1,]  1.0 -0.5
## [2,] -0.5  1.0

y<-mvrnorm(n=1000, mu = c(0,10), Sigma = sigma) #mu = mean vector
```

Sampling from a Gamma distribution in R

Read ?rgamma. R uses $f(x) \propto x^{a-1}e^{-x/s}$ where a is the **shape parameter** and s is the **scale parameter**.

```
set.seed(10)

s2<- rgamma(10000, shape = 2, rate = 4)
mean(s2) #2/4

## [1] 0.4997123

var(s2) #2/16

## [1] 0.1244175
```

Sampling from an Inverse Gamma distribution in R

```
set.seed(10)
s2<- rgamma(10000, shape = 10, rate = 4)

invs2<-1/s2
```

Example

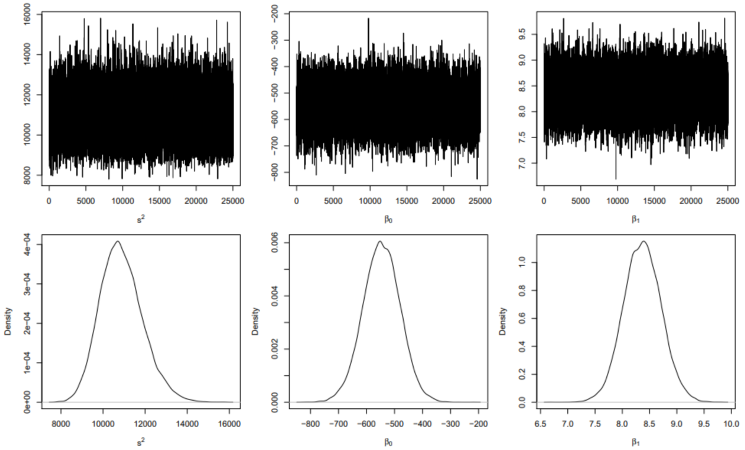
Mann et al. (2006) studied angulate tortoises. This tortoise is sometimes called the fighting tortoise, because males often attack rival males and try to overturn them. The researchers were interested in morphology and sexual selection. They compared morphology in males that lost and won fights. (Taken from STA2007H course, UCT)

Reference

GKH Mann, MJ ORiain & MD Hofmeyr. 2006. Shaping up to fight: sexual selection influences body shape and size in the fighting tortoise (Chersina angulata). Journal of Zoology 269:373-379.

Using Bayesian regression with non-informative priors, fit the model Weight $\sim$ Length.

Bayesian regression



- The trace plots indicate potential convergence of Markov Chain.
- **Posterior distributions displayed!**

Bayesian regression

```
#Some quantiles of the posterior distribution
apply(post_reg, 2, function(x){quantile(x, c(0.025, 0.5, .975))})

##             [,1]      [,2]     [,3]
## 2.5%    9080.848 -677.5829 7.700899
## 50%    10807.897 -546.5273 8.367104
## 97.5% 13056.413 -418.1254 9.051761

#posterior mean and sd
apply(post_reg, 2, mean)

## [1] 10879.895568  -546.461332     8.368508

apply(post_reg, 2, sd)

## [1] 1013.0020660   65.8763430    0.3443709
```

- Compare to OLS results

Linear Regression: Conjugate priors

Consider the linear model:

$$\boldsymbol{y} = \boldsymbol{X}\beta + \boldsymbol{e}$$

where $\left[\boldsymbol{y}|\boldsymbol{\beta}, \sigma^2, \boldsymbol{X}\right] \sim \mathcal{N}_n\left(\boldsymbol{X}\boldsymbol{\beta}, \sigma^2\boldsymbol{I}_n\right)$.
Lets assume

$$\boldsymbol{\beta}|\sigma^2 \sim \mathcal{N}_{k+1}\left(\tilde{\boldsymbol{\beta}}, \sigma^2\boldsymbol{M}^{-1}\right),$$

where $\boldsymbol{M}$ is a $k+1$ positive definite symmetric matrix and

$$\sigma^2 \sim \mathcal{IG}\left(a, b\right), \quad a, b > 0.$$

Then

$$\left[\boldsymbol{\beta}|\sigma^2, \boldsymbol{y}, \boldsymbol{X}\right] \sim \mathcal{N}_{k+1}\left(\boldsymbol{\mu}_{\boldsymbol{\beta}}, \sigma^2\left(\boldsymbol{M} + \boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\right)$$
$$\left[\sigma^2|\boldsymbol{y}, \boldsymbol{X}\right] \sim \mathcal{IG}\left(a + \frac{n}{2}, b + \frac{A_2}{2}\right)$$

where

$$\mu_{\boldsymbol{\beta}} = \left(\boldsymbol{M} + \boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\left(\left(\boldsymbol{X}^T\boldsymbol{X}\right)\hat{\boldsymbol{\beta}} + \boldsymbol{M}\tilde{\boldsymbol{\beta}}\right)$$
$$A_2 = \boldsymbol{y}^T\boldsymbol{y} + \tilde{\boldsymbol{\beta}}^T\boldsymbol{M}\tilde{\boldsymbol{\beta}} - \mu_{\boldsymbol{\beta}}^T\left(\boldsymbol{M} + \boldsymbol{X}^T\boldsymbol{X}\right)\mu_{\boldsymbol{\beta}}.$$