Linear Regression

Rebecca C. Steorts
Predictive Modeling and Data Mining: STA 521

October 2015

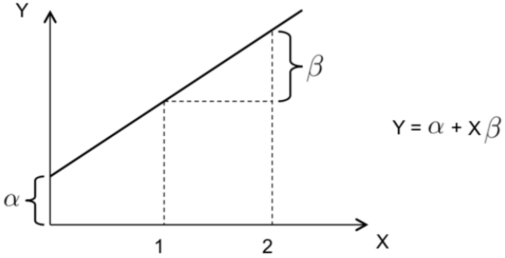
- Knowledge of linear regression is assumed.
- How can we do variable selection?
- Bayes factors.
- Gibbs sampling.

In a [Gaussian] linear regression,

$$y \mid \mathbf{x} \sim \mathrm{N}(\mathbf{x}'\boldsymbol{\beta}, \sigma^2)$$

Conditional mean is $\mathbb{E}[y|\mathbf{x}] = \mathbf{x}'\beta$.

With just one x, we have simple linear regression.



$\mathbb{E}[y]$ increases by β for every unit increase in x.

We assume $\boldsymbol{Y}$ observations and covariates $\boldsymbol{X}$.

Recall that

$$\boldsymbol{Y}_{n\times 1} = \boldsymbol{X}_{n\times p}\boldsymbol{\beta}_{p\times 1} + \epsilon_{n\times 1}$$

$$\epsilon \sim N(0, \sigma^2 I_n).$$

Recall that $E(\boldsymbol{Y} \mid \boldsymbol{X})$ is linear in it's parameter values.

For a thorough review see Hoff or your notes from STA 521.

Oxygen uptake

- Twelve healthy men that don't exercise recruited to study effects of 2 exercise programs on oxygen uptake
- Program one: 12 weeks of flat running
- Program two: 12 weeks of step aerobics

We estimate the coefficients $\hat{\beta} \in \mathbb{R}^p$ by least squares:

$$\hat{\beta} = \operatorname*{argmin}_{\beta \in \mathbb{R}^p} \|y - X\hat{\beta}\|_2^2$$

This gives

$$\hat{\beta} = (X^T X)^{-1} X^T y$$

(Check: does this match the expressions for univariate regression, without and with an intercept?)

The fitted values are

$$\hat{y} = X\hat{\beta} = X(X^T X)^{-1} X^T y$$

This is a linear function of y, $\hat{y} = Hy$, where $H = X(X^T X)^{-1} X^T$ is sometimes called the hat matrix

Let SSR denote sum of squared residuals.

$$\min_{\beta} SSR(\hat{\beta}) = \min_{\beta} \|y - X\hat{\beta}\|_2^2$$

Then

$$\frac{\partial SSR(\hat{\beta})}{\partial d\hat{\beta}} = \frac{\partial (y - X\hat{\beta})^T (y - X\hat{\beta})}{\partial d\hat{\beta}} \tag{1}$$

$$= \frac{\partial \boldsymbol{Y}^T \boldsymbol{Y} - 2\hat{\beta}^T \boldsymbol{X}^T \boldsymbol{Y} + \hat{\beta}^T \boldsymbol{X}^T \boldsymbol{X} \hat{\beta}}{\partial d\hat{\beta}} \tag{2}$$

$$= -2\boldsymbol{X}^T \boldsymbol{Y} + 2\boldsymbol{X}^T \boldsymbol{X} \hat{\beta} \tag{3}$$

This implies $-\boldsymbol{X}^T \boldsymbol{Y} + \boldsymbol{X}^T \boldsymbol{X} \hat{\beta} = 0 \implies \hat{\beta} = (\boldsymbol{X}^T \boldsymbol{X})^{-1} \boldsymbol{X}^T \boldsymbol{Y}$.

Called the ordinary least squares estimator. When is it unique?

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

$$E(\hat{\beta}) = \beta.$$

$$\begin{aligned}
\mathrm{Var}(\hat{\beta}) &= \mathrm{Var}\{(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y}\} && (4)\\
&= (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\sigma^2 I_n\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1} && (5)\\
&= \sigma^2(\boldsymbol{X}^T\boldsymbol{X})^{-1} && (6)
\end{aligned}$$

$$\hat{\beta} \sim MVN(\beta, \sigma^2(\boldsymbol{X}^T\boldsymbol{X})^{-1}).$$

Suppose

$$\boldsymbol{Y} \mid \boldsymbol{X}, \boldsymbol{\beta}, \sigma^2 \sim MVN(\boldsymbol{X}\boldsymbol{\beta}, \sigma^2 I) \tag{7}$$
$$\beta \sim MVN(\beta_o, \Sigma_o) \tag{8}$$

What is the form of the distribution of $\beta \mid \boldsymbol{Y}, \boldsymbol{X}, \sigma^2$?
Recall it's $MVN(\mu_n, \Sigma_n)$

Let's think about the covariance first.

$$\Sigma_n = [\Sigma_o^{-1} + \boldsymbol{X}^T \boldsymbol{X} \sigma^2]^{-1}.$$

Now let's think about the mean.

$$\mu_n = [\Sigma_o^{-1} + X^T X \sigma^2]^{-1} (\boldsymbol{X}^T \boldsymbol{Y} / \sigma^2 + \Sigma_o^{-1} \beta_o)$$

Suppose we don't know σ^2.
Let $\gamma = 1/\sigma^2$.

$$\boldsymbol{Y} \mid \boldsymbol{X}, \boldsymbol{\beta}, \sigma^2 \sim MVN(\boldsymbol{X}\boldsymbol{\beta}, \sigma^2 I) \tag{9}$$

$$\beta \sim MVN(\beta_o, \Sigma_o) \tag{10}$$

$$\gamma \sim IG(\nu_o/2, \nu_o\sigma_o^2/2). \tag{11}$$

Then

$$
\begin{aligned}
& p(\gamma \mid \boldsymbol{Y}, \boldsymbol{X}, \beta) && (12)\\
& = p(\gamma)p(\boldsymbol{Y} \mid bX, \boldsymbol{Y}, \beta) && (13)\\
& \propto \gamma^{(\nu_o+n)/2-1} \exp\{-\gamma \times \nu_o \sigma_o^2/2\} && (14)\\
& \qquad \times \gamma^{n/2} \exp\{-\gamma \times SSR(\beta)/2\} && (15)\\
& \propto \gamma^{(\nu_o+n)/2-1} \exp\{-\gamma(\nu_o \sigma_o^2/2 + SSR(\beta)/2)\} && (16)
\end{aligned}
$$

Thus,

$$
\gamma \mid \boldsymbol{Y}, \boldsymbol{X}, \beta \sim IG((\nu_o + n)/2, (\nu_o \sigma_o^2/2 + SSR(\beta))/2)
$$

In order to update $p(\beta, \sigma^2 \mid \boldsymbol{Y}, \boldsymbol{X})$ sample through a two-stage Gibbs sampler.

This is similar to other examples before, and the details can be found in Hoff. (Cycle through the MVN and the IG).

Model selection

- Often we have a large number of covariates.
- Using all of them induces poor statistical performance.
- How can we reduce the covariates and have good inference and prediction?
- Common method: Backwards and stepwise regression (slow).

Bayesian model comparison

Suppose that we believe some of the regression coefficients are 0.

Come up with a prior distribution that reflects the probability of this occuring.

Consider

$$y_i = z_1 b_1 x_{i,1} + \ldots z_p b_p x_{i,p},$$

where b_p is a real number and z_j indicate which regression coefficients are nonzero.

Note: $\beta_j = b_j \times z_j$.

Bayesian model selection works by obtaining a posterior distribution for $\boldsymbol{z}$.

Suppose a prior $p(\boldsymbol{z})$ over models

Then

$$p(\boldsymbol{z} \mid \boldsymbol{Y}, \boldsymbol{X}) = \frac{p(\boldsymbol{z})p(\boldsymbol{Y} \mid \boldsymbol{X}, \boldsymbol{z})}{\sum_{z} p(\boldsymbol{z})p(\boldsymbol{Y} \mid \boldsymbol{X}, \boldsymbol{z})}$$

Suppose we want to compare two models. Consider

$$odds(z_a, z_b \mid \boldsymbol{Y}, \boldsymbol{X}) = \frac{p(z_a \mid \boldsymbol{Y}, \boldsymbol{X})}{p(z_b \mid \boldsymbol{Y}, \boldsymbol{X})} = \frac{p(z_a)}{p(z_b)} \times \frac{p(\boldsymbol{Y} \mid \boldsymbol{X}, z_a)}{p(\boldsymbol{Y} \mid \boldsymbol{X}, z_b)}$$

This is posterior odds $=$ prior odds $\times$ “Bayes factor”

“Bayes factor”: how much the data favor model z_a over model z_b

To obtain a posterior distribution over models, we must compute $p(\boldsymbol{Y} \mid \boldsymbol{X}, \boldsymbol{z})$ for *each* model under consideration.

We must compute

$$p(\boldsymbol{Y} \mid \boldsymbol{X}, \boldsymbol{z}) = \int\int p(\boldsymbol{Y}, \beta, \sigma^2, \mid \boldsymbol{X}, \boldsymbol{z}) \tag{17}$$

$$\int\int p(\boldsymbol{Y} \mid \boldsymbol{X}, \boldsymbol{z}) p(\beta \mid \boldsymbol{X}, \boldsymbol{z}) p(\sigma^2). \tag{18}$$

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

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

What is a g-prior?

(Hoff, Section 9.2, p. 156-157). Based upon the prior should be invariant to changes in the scale of the covariates.

- Defined $\tilde{\boldsymbol{X}} = H\boldsymbol{X}$ for some matrix H.
- Suppose we obtain a posterior of β from $\boldsymbol{Y}$ and $\boldsymbol{X}$.
- According to the idea of invariance above, then the posterior of β and $H\beta$ should be the same.
- Homework: Condition is met if $\beta_o = 0$ and $\Sigma_o = k(\boldsymbol{X}^T\boldsymbol{X})^{-1}$ for $k > 0$.

Popular choice: let $k = g\sigma^2$ for $g > 0$. This is the g-prior.

Given the g-prior

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

$p(\boldsymbol{Y} \mid \boldsymbol{X}, \boldsymbol{z})$ can be worked out in closed form (details p. 165).

Go through the details on your own.

This results in being able to compute

$$\frac{p(\boldsymbol{Y} \mid \boldsymbol{X}, \boldsymbol{z}_a)}{p(\boldsymbol{Y} \mid \boldsymbol{X}, \boldsymbol{z}_b)} = (1+n)^{(p_{z_b} - p_{z_a})/2} \times \left(\frac{s_{z_a}^2}{s_{z_b}^2}\right)^{1/2} \tag{19}$$

$$\times \left(\frac{s_{z_b}^2 + SSR_g^{z_b}}{s_{z_b}^2 + SSR_g^{z_a}}\right)^{(n+1)/2} \tag{20}$$

We have a ratio of the marginal probabilities, giving us a balance between model complexity and model fit.

Suppose p_{z_b} is large compared to p_{z_a}.

This causes a penalization of model z_b

Note that a large value of $SSR_g^{z_b}$ compared to $SSR_g^{z_a}$ will penalize model z_a.

$\boldsymbol{z}$	model	$\log p(\boldsymbol{y}\|\mathbf{X}, \boldsymbol{z})$	$p(\boldsymbol{z}\|\boldsymbol{y}, \mathbf{X})$
(1,0,0,0)	β_1	-44.33	0.00
(1,1,0,0)	$\beta_1 + \beta_2 \times \text{group}_i$	-42.35	0.00
(1,0,1,0)	$\beta_1 + \beta_3 \times \text{age}_i$	-37.66	0.18
(1,1,1,0)	$\beta_1 + \beta_2 \times \text{group}_i + \beta_3 \times \text{age}_i$	-36.42	0.63
(1,1,1,1)	$\beta_1 + \beta_2 \times \text{group}_i + \beta_3 \times \text{age}_i + \beta_4 \times \text{group}_i \times \text{age}_i$	-37.60	0.19

Table 9.1. Marginal probabilities of the data under five different models.

Figure 1: The most probable model is the one corresponding $\boldsymbol{z} = (1, 1, 1, 0)$

What is the biggest downside of this approach?

How do we fix it easily using what we've learned so far in the course?

Suppose p is large. Then 2^p models to consider.

Instead let's use a Gibbs sampler to search through the space of models for values where $\boldsymbol{z}$ has a high posterior probability.

Generate a new value of $\boldsymbol{z}$ via

$$p(z_j \mid \boldsymbol{Y}, \boldsymbol{X}, \boldsymbol{z}_{-j}).$$

The full conditional that $z_j = 1$ can be written as $o_j/(o_j + 1)$.

$$o_j = \frac{p(z_j = 1 \mid \boldsymbol{Y}, \boldsymbol{X}, \boldsymbol{z}_{-j})}{p(z_j = 0 \mid \boldsymbol{Y}, \boldsymbol{X}, \boldsymbol{z}_{-j})} \tag{21}$$

$$= \frac{p(z_j = 1)p(\boldsymbol{Y} \mid \boldsymbol{X}, \boldsymbol{z}_{-j}, z_j = 1)}{p(z_j = 0)p(\boldsymbol{Y} \mid \boldsymbol{X}, \boldsymbol{z}_{-j}, z_j = 0)} \tag{22}$$

Note: we may also want to obtain posterior samples of β and σ^2.

Using the conditional distributions from Section 9.2, we can sample from these directly.

The Gibbs sampling scheme requires using Section 9.2 and 9.3 (covered in lab).

$$
\begin{tikzcd}
\boldsymbol{z}^{(s)} \arrow[r] \arrow[d] & \sigma^{2(s)} \arrow[r] & \boldsymbol{\beta}^{(s)} \\
\boldsymbol{z}^{(s+1)} \arrow[r] & \sigma^{2(s+1)} \arrow[r] & \boldsymbol{\beta}^{(s+1)}
\end{tikzcd}
$$

Figure 2: Start with $\boldsymbol{z}^{(s)}$. Then in random order update z_j from its full conditional.

Generate

$$\{\boldsymbol{z}^{(s+1)}, \sigma^{2(s+1)}, \beta^{(s+1)}\}:$$

1. Set $\boldsymbol{z} = \boldsymbol{z}^{(s)}$
2. For $j \in \{1, \ldots, p\}$ in random order, replace z_j with a sample from
$$p(z_j \mid \boldsymbol{z}_{-j}, \boldsymbol{Y}, \boldsymbol{X})$$
3. Set $\boldsymbol{z}^{(s+1)} = \boldsymbol{z}^{(s)}$
4. Sample $\sigma^{2(s)} \sim p(\sigma^2 \mid \boldsymbol{z}^{(s+1)}, \boldsymbol{Y}, \boldsymbol{X})$
5. Sample $\beta^{(s+1)} \sim p(\beta \mid \boldsymbol{z}^{(s+1)}, \sigma^{2(s+1)}, \boldsymbol{Y}, \boldsymbol{X})$

Lab this week: Linear regression and understanding model selection using the diabetes data.