Module 11: Linear Regression

Rebecca C. Steorts

Agenda

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

Regression models

How does an outcome Y vary as a function of the covariates $\boldsymbol{x} = (x_1, \ldots, x_p)$?

- Which x_j's have an effect?
- What are the effect sizes?
- Can we predict Y as a function of $\boldsymbol{x}$?

Such question can be assessed via a regression model $p(y \mid \boldsymbol{x})$.

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

Oxygen uptake experiment

Exercise is hypotheized to relate to O_2 uptake

What type of exercise is the most beneficial?

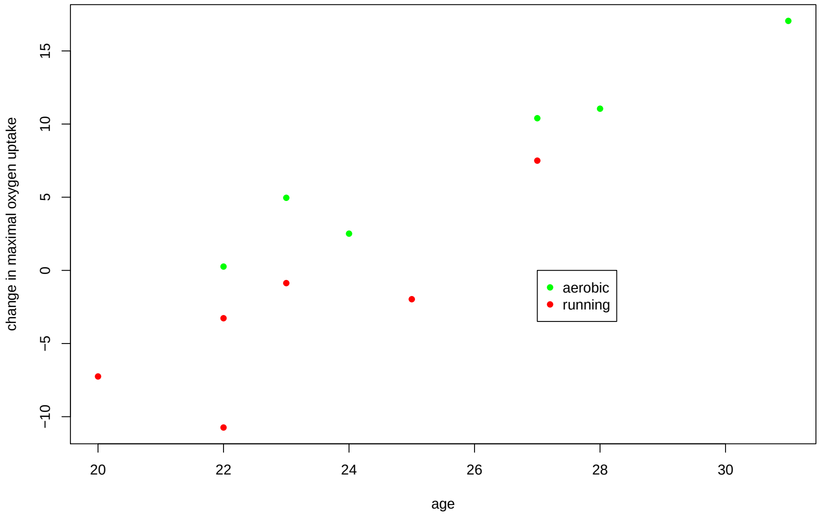
Experimental design: 12 male volunteers.

1. O_2 uptake measured at the beginning of the study.
2. 6 randomized to step aerobics program
3. 6 remaining men do a running program
4. O_2 uptake measured at end of study

Data

```
# aerobic versus running activity
x1<-c(0,0,0,0,0,0,1,1,1,1,1,1)
# age
x2<-c(23,22,22,25,27,20,31,23,27,28,22,24)
# change in maximal oxygen uptake
y<-c(-0.87,-10.74,-3.27,-1.97,7.50,
     -7.25,17.05,4.96,10.40,11.05,0.26,2.51)
```

Plot



aerobic
running
change in maximal oxygen uptake
age
−10
−5
0
5
10
15
20
22
24
26
28
30

Data analysis

y = change in oxygen uptake

x_1 = exercise indicator (0 for aerobic, 1 for running)

x_2 = age

How can we estimate $p(y \mid x_1, x_2)$?

Linear regression

Linear regression

Assume that smoothness is a function of age.

For each group,

$$y = \beta_o + \beta_1 x_2 + \epsilon$$

This is called a linear regression model.

Linearity means linear in the parameters.

We could also try the model

$$y = \beta_o + \beta_1 x_2 + \beta_2 x_2^2 + \beta_3 x_2^3 + \epsilon$$

which is also a linear regression model.

Multiple linear regression

We can estimate for both groups simulatenously.

$$Y_i = \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4} + \epsilon_i$$

where

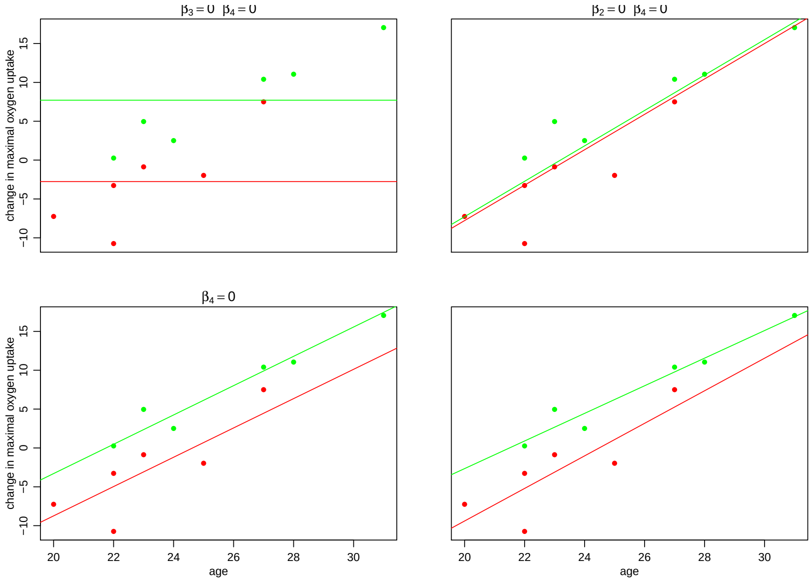
$$x_{i1} = 1 \text{ for subject } i \tag{1}$$
$$x_{i2} = 0 \text{ if running program for subject; 1 if aerobics} \tag{2}$$
$$x_{i3} = \text{age of subject i} \tag{3}$$
$$x_{i4} = x_{i2} \times x_{i3} \tag{4}$$

Under this model,

$$E[Y \mid \boldsymbol{x}] = (\beta_1 + \beta_3) \times age \text{ if } x_2 = 0$$

$$E[Y \mid \boldsymbol{x}] = (\beta_1 + \beta_2) + (\beta_3 + \beta_4) \times age \text{ if } x_2 = 1$$

Least squares regression lines



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

$$Y_i = \beta^T x_i + e_i \tag{5}$$

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

Normal Regression Model (continued)

This full specifies the density of the data:

$$p(y_1, \ldots, y_n \mid x_1, \ldots x_n, \beta, \sigma^2) \tag{6}$$

$$= \prod_{i=1}^{n} p(y_i \mid x_i, \beta, \sigma^2) \tag{7}$$

$$(2\pi\sigma^2)^{-n/2} \exp\{\frac{-1}{2\sigma^2} \sum_{i=1}^{n} (y_i - \beta^T x_i)^2\} \tag{8}$$

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

$$p(y_1, \ldots, y_n \mid x_1, \ldots x_n, \beta, \sigma^2) \tag{9}$$
$$(2\pi\sigma^2)^{-n/2} \exp\{\frac{-1}{2\sigma^2} \sum_{i=1}^{n} (y_i - \beta^T x_i)^2\} \tag{10}$$
$$(2\pi\sigma^2)^{-n/2} \exp\{\frac{-1}{2\sigma^2} (y - X\beta)^T (y - X\beta)\} \tag{11}$$

Posterior computation

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

$$
\begin{align}
p(\beta \mid y, X) &\propto p(y \mid X, \beta) p(\beta) \tag{12} \\
&\propto \exp\{-a/2(y - X\beta)^T (y - X\beta)\} \times \exp\{-b/2\beta^T\beta)\} \tag{13}
\end{align}
$$

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)

$$
\begin{aligned}
& p(\beta \mid y, X) && (14)\\
& \propto \exp\{-\frac{a}{2}(y - X\beta)^T(y - X\beta)\} \times \exp\{-\frac{b}{2}\beta^T\beta)\} && (15)\\
& \propto \exp\{-\frac{a}{2}[y^T y - 2\beta^T X^T y + \beta^T X^T X\beta] - \frac{b}{2}\beta^T\beta\} && (16)\\
& \propto \exp\{a\beta^T X^T y - \frac{a}{2}\beta^T X^T X\beta - b/2\beta^T\beta\} && (17)\\
& \propto \exp\{a\beta^T[X^T y] - 1/2\beta^T(aX^T X + bI)\beta\} && (18)
\end{aligned}
$$

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

Multivariate inference for regression models

$$\boldsymbol{y} \mid \beta \sim MVN(\boldsymbol{X}\beta, \sigma^2 \boldsymbol{I})\beta \qquad \sim MVN(\beta_0, \Sigma_0) \tag{19}$$

The posterior can be shown to be

$$\beta \mid \boldsymbol{y}, \boldsymbol{X} \sim MVN(\beta_n, \Sigma_n)$$

where

$$\beta_n = E[\beta \ mid \boldsymbol{y}, \boldsymbol{X}, \sigma^2] = (\Sigma_o^{-1} + (X^T X)^{-1}/\sigma^2)^{-1}(\Sigma_o^{-1}\beta_0 + \boldsymbol{X}^T \boldsymbol{y}/\sigma^2)$$

$$\Sigma_n = \mathrm{Var}[\beta \ mid \boldsymbol{y}, \boldsymbol{X}, \sigma^2] = (\Sigma_o^{-1} + (X^T X)^{-1}/\sigma^2)^{-1}$$

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

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

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.