Module 10: Linear Regression

Rebecca C. Steorts

Agenda

- Oxygen uptake example
- Linear regression
- Multiple Linear Regression
- Ordinary Least Squares
- An application to swimmers

Oxygen uptake experiment

Exercise is hypotheized to relate to O_2 uptake

What type of exercise is the most beneficial?

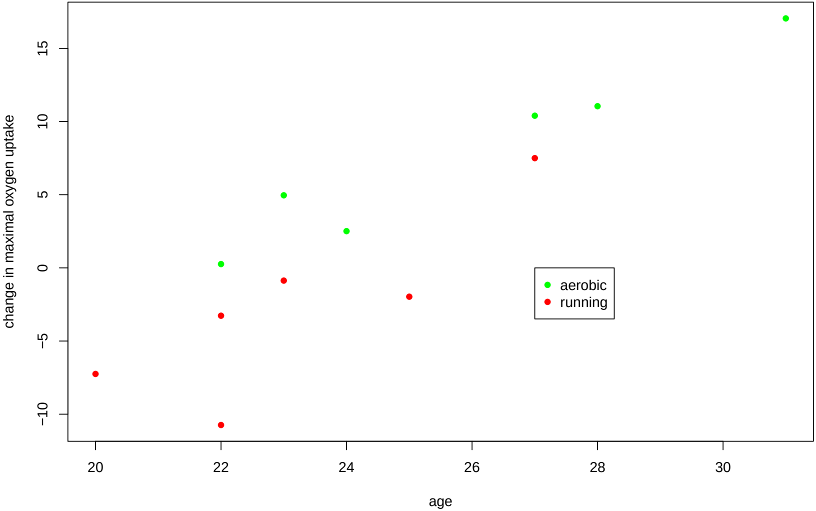
Experimental design: 12 male volunteers.

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

Data

```r
# running is 0, 1 is aerobic
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)
```

Exploratory Data Analysis



Data analysis

y = change in oxygen uptake (scalar)

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

x_2 = age

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

Linear regression

Assume that smoothness is a function of age.

For each group,

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

Linearity means linear in the parameters (β's).

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.

Notation

- $X_{n \times p}$: regression features or covariates (design matrix)
- $\boldsymbol{x}_i$: ith row vector of the regression covariates
- $\boldsymbol{y}_{n \times 1}$: response variable (vector)
- $\boldsymbol{\beta}_{p \times 1}$: vector of regression coefficients

Notation (continued)

$$\boldsymbol{X}_{n\times p} = \begin{pmatrix} x_{11} & x_{12} & \dots & x_{1p} \\ x_{21} & x_{22} & \dots & x_{2p} \\ x_{i1} & x_{i2} & \dots & x_{ip} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \dots & x_{np} \end{pmatrix}.$$

- A column of x represents a particular covariate we might be interested in, such as age of a person.
- Denote x_i as the ith row vector of the $X_{n\times p}$ matrix.

$$x_i = \begin{pmatrix} x_{i1} \\ x_{i2} \\ \vdots \\ x_{ip} \end{pmatrix}$$

Notation (continued)

$$\boldsymbol{\beta} = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{pmatrix}$$

$$\boldsymbol{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}$$

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

Regression models

How does an outcome $\boldsymbol{y}$ vary as a function of the covariates which we represent as $X_{n \times p}$ matrix?

- Can we predict $\boldsymbol{y}$ as a function of each row in the matrix $X_{n \times p}$ denoted by $\boldsymbol{x}_i$.
- Which $\boldsymbol{x}_i$'s have an effect?

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

Multiple linear regression

Consider the following:

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

where

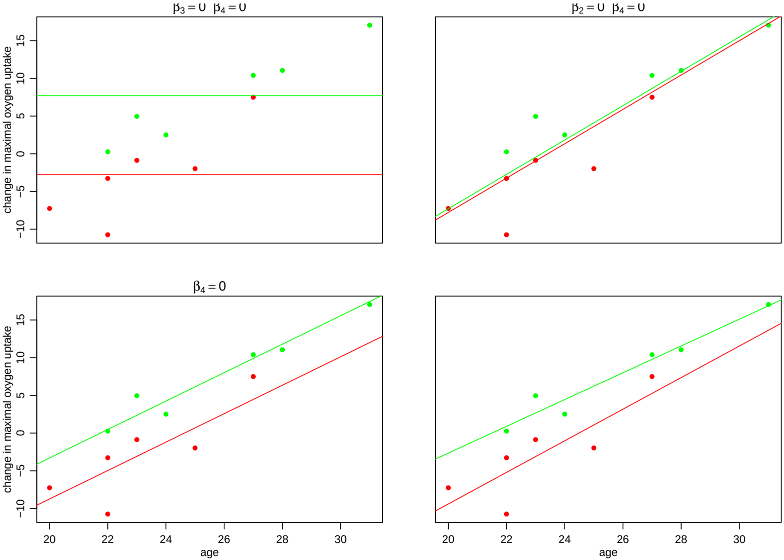
$$x_{i1} = 1 \text{ for subject } i \tag{1}$$
$$x_{i2} = 0 \text{ for running; } 1 \text{ for aerobics} \tag{2}$$
$$x_{i3} = \text{age of subject i} \tag{3}$$
$$x_{i4} = x_{i2} \times x_{i3} \tag{4}$$

Under this model,

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

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

Least squares regression lines



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

$$\boldsymbol{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}_{n \times 1}$$

- $\boldsymbol{x}_i$ is the ith row of the design matrix $X_{n \times p}$.

Consider the regression coefficients

$$\boldsymbol{\beta} = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{pmatrix}_{p \times 1}$$

Normal Regression Model

The Normal regression model specifies that

- $E[Y \mid x]$ is linear and
- the sampling variability around the mean is independently and identically (iid) drawn from a normal distribution

$$Y_i = \beta^T x_i + \boldsymbol{\epsilon}_i \tag{5}$$

$$\epsilon_1, \ldots, \epsilon_n \overset{iid}{\sim} Normal(0, \sigma^2) \tag{6}$$

We can specify a simple Bayesian model as the following:

$$\boldsymbol{y} \mid X, \boldsymbol{\beta}, \sigma^2 \sim MVN(X\boldsymbol{\beta}, \sigma^2 I)$$
$$\boldsymbol{\beta} \sim MVN(0, \tau^2 I)$$

Normal Regression Model (continued)

This specifies the density of the data:

$$p(y_1, \ldots, y_n \mid x_1, \ldots x_n, \boldsymbol{\beta}, \sigma^2) \tag{7}$$

$$= \prod_{i=1}^{n} p(\boldsymbol{y}_i \mid \boldsymbol{x}_i, \boldsymbol{\beta}, \sigma^2) \tag{8}$$

$$(2\pi\sigma^2)^{-n/2} \exp\{\frac{-1}{2\sigma^2} \sum_{i=1}^{n} (\boldsymbol{y}_i - \beta^T \boldsymbol{x}_i)^2\} \tag{9}$$

$$= (2\pi\sigma^2)^{-n/2} \exp\{-\frac{1}{2}(\boldsymbol{y} - X\beta)^T (\sigma^2)^{-1} I(\boldsymbol{y} - X\beta)\} \tag{10}$$

Ordinary Least Squares

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

$$\hat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta} \in \mathbb{R}^p}{\arg\min} \|\boldsymbol{y} - X\boldsymbol{\beta}\|_2^2$$

This gives

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

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

The fitted values are

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

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

Ordinary Least squares estimation

Let SSR denote sum of squared residuals.

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

Then

$$\frac{\partial SSR(\boldsymbol{\beta})}{\partial d\boldsymbol{\beta}} = \frac{\partial(\boldsymbol{y} - X\boldsymbol{\beta})^T(\boldsymbol{y} - X\boldsymbol{\beta})}{\partial d\boldsymbol{\beta}} \tag{11}$$

$$= \frac{\partial \boldsymbol{y}^T\boldsymbol{y} - 2\boldsymbol{\beta}^T X^T\boldsymbol{y} + \hat{\beta}^T(X^TX)\boldsymbol{\beta}}{\partial d\boldsymbol{\beta}} \tag{12}$$

$$= -2X^T\boldsymbol{y} + 2X^TX\boldsymbol{\beta} \tag{13}$$

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

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

Ordinary Least squares estimation

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

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

$$\begin{aligned} \mathrm{Var}(\hat{\boldsymbol{\beta}}) &= \mathrm{Var}\{(X^T X)^{-1} X^T \boldsymbol{Y}\} && (14) \\ &= (X^T X)^{-1} X^T \sigma^2 I_n X (X^T X)^{-1} && (15) \\ &= \sigma^2 (X^T X)^{-1} && (16) \end{aligned}$$

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

Recall data set up

```
# running is 0, 1 is aerobic
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)
```

Recall data set up

```
(x3 <- x2) #age
```

```
##  [1] 23 22 22 25 27 20 31 23 27 28 22 24
```

```
(x2 <- x1) #aerobic versus running
```

```
##  [1] 0 0 0 0 0 0 1 1 1 1 1 1
```

```
(x1<- seq(1:length(x2))) #index of person
```

```
##  [1]  1  2  3  4  5  6  7  8  9 10 11 12
```

```
(x4 <- x2*x3)
```

```
##  [1]  0  0  0  0  0  0 31 23 27 28 22 24
```

Recall data set up

```r
(X <- cbind(x1,x2,x3,x4))
```

```
##       x1 x2 x3 x4
##  [1,]  1  0 23  0
##  [2,]  2  0 22  0
##  [3,]  3  0 22  0
##  [4,]  4  0 25  0
##  [5,]  5  0 27  0
##  [6,]  6  0 20  0
##  [7,]  7  1 31 31
##  [8,]  8  1 23 23
##  [9,]  9  1 27 27
## [10,] 10  1 28 28
## [11,] 11  1 22 22
## [12,] 12  1 24 24
```

OLS estimation in R

```
## using the lm function
fit.ols<-lm(y~ X[,2] + X[,3] +X[,4])
summary(fit.ols)$coef
```

```
##               Estimate Std. Error    t value    Pr(>|t|)
## (Intercept) -51.2939459 12.2522126 -4.1865047 0.003052321
## X[, 2]       13.1070904 15.7619762  0.8315639 0.429775106
## X[, 3]        2.0947027  0.5263585  3.9796120 0.004063901
## X[, 4]       -0.3182438  0.6498086 -0.4897500 0.637457484
```

Multivariate inference for regression models

$$\boldsymbol{y} \mid \beta \sim MVN(X\beta, \sigma^2 I) \tag{17}$$

$$\beta \sim MVN(\beta_0, \Sigma_0) \tag{18}$$

The posterior can be shown to be

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

where

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

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

Multivariate inference for regression models

The posterior can be shown to be

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

where

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

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

Remark: If $\Sigma_o^{-1} << (X^T X)^{-1}$ then $\beta_n \approx \hat{\beta}_{ols}$

If $\Sigma_o^{-1} >> (X^T X)^{-1}$ then $\beta_n \approx \beta_0$

Posterior inference applied to Oxygen uptake

To continue the rest of the oxygen uptake example, please refer to 9.2 in Hoff (commentary and code). Pages 157 – 159 in Hoff.

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

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

be the design matrix for swimmer i. Then we use the following linear regression model:

$$\begin{aligned} Y_i &\sim \mathcal{N}_6\left(X_i\beta_i, \tau_i^{-1}\mathcal{I}_6\right) \\ \beta_i &\sim \mathcal{N}_2\left(\beta_0, \Sigma_0\right) \\ \tau_i &\sim \text{Gamma}(a, b). \end{aligned}$$

Derive full conditionals for β_i and τ_i.

Solution (Task 1)

The conditional posterior for β_i is multivariate normal with

$$\mathbb{V}[\beta_i \mid Y_i, X_i, \tau_i] = (\Sigma_0^{-1} + \tau_i X_i^T X_i)^{-1}$$
$$\mathbb{E}[\beta_i \mid Y_i, X_i, \tau_i] = (\Sigma_0^{-1} + \tau_i X_i^T X_i)^{-1}(\Sigma_0^{-1}\beta_0 + \tau_i X_i^T Y_i).$$

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.