Module XX: Linear Regression

Rebecca C. Steorts

Based Upon Hoff, Chapter 9

Agenda

- Motivation: oxygen uptake example
- Linear regression
- Multiple and Multivariate Linear Regression
- Background on the Euclidean norm and argmin
- Ordinary Least Squares + Exercises

Oxygen uptake case study

Experimental design: 12 male volunteers.

1. 6 men take part in a randomized aerobics program
2. 6 remaining men take part in a randomized running program
3. The maximal O_2 uptake measured before and after the 12 week program
4. The change in maximal O_2 uptake is then calculated for each individual

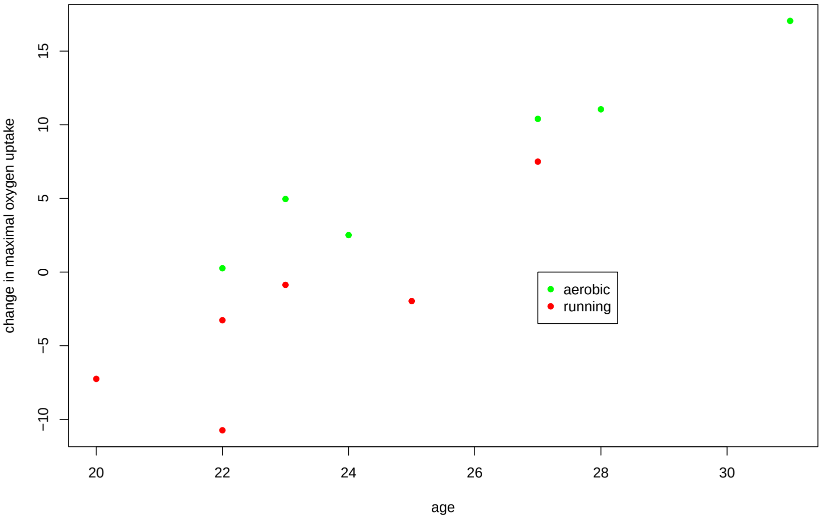
What type of exercise is the most beneficial?

Full details of the study can be found in Hoff, page 149-151.

Data

```r
# 0 is running
# 1 is aerobic exercise
x1<-c(0,0,0,0,0,0,1,1,1,1,1,1)
# x2 is 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 maximal 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).

Linear regression

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 questions 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

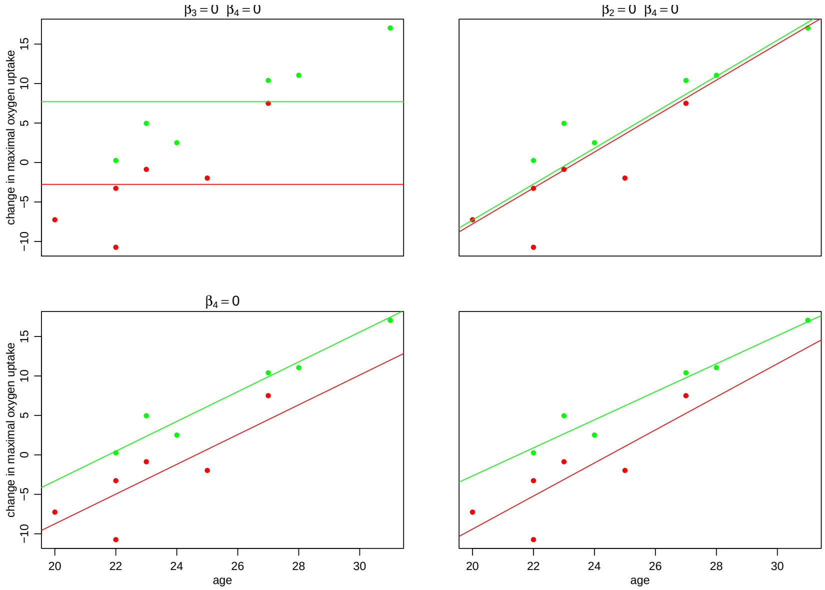
$$x_{i1} = 1 \text{ for subject } i \tag{1}$$
$$x_{i2} = 0 \text{ for running; } 1 \text{ for aerobics} \tag{2}$$
$$x_{i3} = \text{age of subject i} \tag{3}$$
$$x_{i4} = x_{i2} \times x_{i3} \tag{4}$$

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

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

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

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

This implies $Y_i \mid \boldsymbol{\beta}, \boldsymbol{x}_i \sim \text{Normal}(\boldsymbol{\beta}^T \boldsymbol{x}_i, \sigma^2)$.

Multivariate Bayesian Normal Regression Model

We can re-write this as a multivariate regression model as:

$$\boldsymbol{y} \mid X, \boldsymbol{\beta}, \sigma^2 \sim \text{MVN}(X\boldsymbol{\beta}, \sigma^2 I_p).$$

We can specify a multivariate Bayesian model as:

$$\begin{gathered}\boldsymbol{y} \mid X, \boldsymbol{\beta}, \sigma^2 \sim \text{MVN}(X\boldsymbol{\beta}, \sigma^2 I_p) \\ \boldsymbol{\beta} \sim \text{MVN}(0, \tau^2 I_p),\end{gathered}$$

where σ^2, τ^2 are known.

Likelihood

The likelihood is

$$
\begin{aligned}
& p(y_1, \ldots, y_n \mid x_1, \ldots x_n, \boldsymbol{\beta}, \sigma^2) && (7)\\
&= \prod_{i=1}^{n} p(\boldsymbol{y}_i \mid \boldsymbol{x}_i, \boldsymbol{\beta}, \sigma^2) && (8)\\
& (2\pi\sigma^2)^{-n/2} \exp\{\frac{-1}{2\sigma^2} \sum_{i=1}^{n} (\boldsymbol{y}_i - \boldsymbol{\beta}^T \boldsymbol{x}_i)^2\} && (9)\\
&= (2\pi\sigma^2)^{-n/2} \exp\{-\frac{1}{2}(\boldsymbol{y} - X\beta)^T (\sigma^2)^{-1} I_p (\boldsymbol{y} - X\beta)\} && (10)
\end{aligned}
$$

Background

The Euclidean norm (L^2 norm or square root of the sum of squares) of $\boldsymbol{y} = (y_1, \ldots, y_n)$ is defined by

$$\|\boldsymbol{y}\|_2 = \sqrt{y_1^2 + \ldots + y_n^2}.$$

It follows that

$$\|\boldsymbol{y}\|_2^2 = y_1^2 + \ldots + y_n^2.$$

Why do we use this notation? It's compact and convenient!

Background

We would like to find

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

where the arg min (the arguments of the minima) are the points or elements of the domains of some function as which the functions values are minimized.

Ordinary Least Squares

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

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

One can show that

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

The fitted values are

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

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**.

Exercise 1 (OLS)

Let SSR denote sum of squared residuals.

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

Show that

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

Ordinary Least squares estimation

Proof: Observe

$$\frac{\partial SSR(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} := \frac{\partial \|\mathbf{y} - X\beta\|_2^2}{\partial \boldsymbol{\beta}} \tag{11}$$

$$= \frac{\partial (\mathbf{y} - X\boldsymbol{\beta})^T(\mathbf{y} - X\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} \tag{12}$$

$$= \frac{\partial \mathbf{y}^T\mathbf{y} - 2\boldsymbol{\beta}^T X^T \mathbf{y} + \hat{\beta}^T (X^T X)\boldsymbol{\beta}}{\partial \boldsymbol{\beta}} \tag{13}$$

$$= -2X^T\mathbf{y} + 2X^T X\beta \tag{14}$$

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

This is called the **ordinary least squares estimator**. How do we know it is unique?

Exercise 2 (OLS)

Show that

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

Ordinary Least squares estimation

Proof: Recall

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

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

$$\begin{aligned}
\text{Var}(\hat{\boldsymbol{\beta}}) &= \text{Var}\{(X^T X)^{-1} X^T \boldsymbol{Y}\} && (15)\\
&= (X^T X)^{-1} X^T \sigma^2 I_n X (X^T X)^{-1} && (16)\\
&= \sigma^2 (X^T X)^{-1} && (17)
\end{aligned}$$

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

Recall data set up

```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)
```

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

```
(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
```