

# Module 11: Linear Regression

Rebecca C. Steorts

# Agenda

- ▶ Oxygen uptake example
- ▶ Linear regression
- ▶ Multiple Linear Regression
- ▶ Ordinary Least Squares
- ▶ Back to oxygen uptake example<sup>1</sup>

---

<sup>1</sup>Much of this lecture comes from the Hoff book.

# Oxygen uptake experiment

Exercise is hypothesized to relate to  $O_2$  uptake

What type of exercise is the most beneficial?

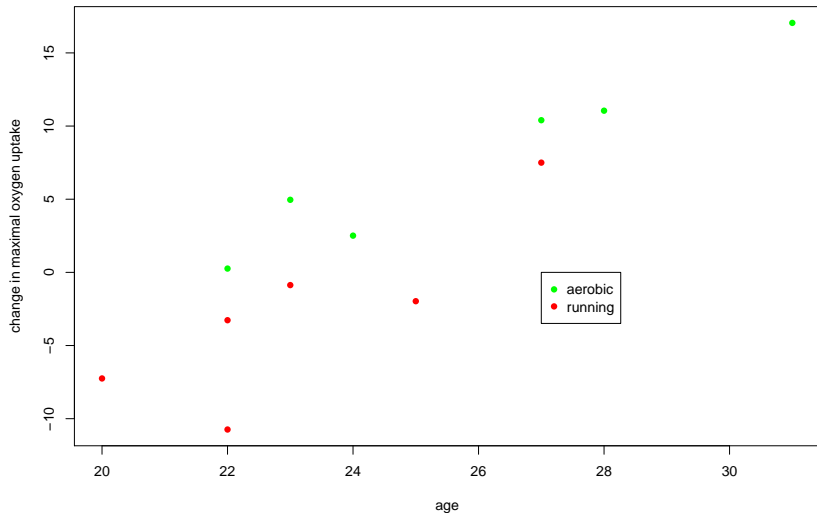
Experimental design: 12 male volunteers.

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)
```

# Exploratory Data Analysis



# 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

Assume that smoothness is a function of age.

For each group,

$$\mathbf{y} = \beta_0 + \beta_1 x_2 + \epsilon$$

Linearity means linear in the parameters ( $\beta$ 's).

We could also try the model

$$\mathbf{y} = \beta_0 + \beta_1 x_2 + \beta_2 x_2^2 + \beta_3 x_2^3 + \epsilon$$

which is also a linear regression model.

# Notation

- ▶  $X_{n \times p}$ : regression features or covariates (design matrix)
- ▶  $\mathbf{x}_{p \times 1}$ :  $i$ th row vector of the regression covariates
- ▶  $\mathbf{y}_{n \times 1}$ : response variable (vector)
- ▶  $\beta_{p \times 1}$ : vector of regression coefficients

Goal: Estimation of  $p(\mathbf{y} \mid X)$ .

Dimensions:  $\mathbf{y}_i - \beta^T \mathbf{x}_i = (1 \times 1) - (1 \times p)(p \times 1) = (1 \times 1)$ .



# Regression models

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

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

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

## Multiple linear regression

We can estimate for both groups simultaneously.

$$\mathbf{Y}_i = \beta_1 \mathbf{x}_{i1} + \beta_2 \mathbf{x}_{i2} + \beta_3 \mathbf{x}_{i3} + \beta_4 \mathbf{x}_{i4} + \epsilon_i$$

where

$$x_{i1} = 1 \text{ for subject } i \quad (1)$$

$$x_{i2} = 0 \text{ if running program for subject; } 1 \text{ if aerobics} \quad (2)$$

$$x_{i3} = \text{age of subject } i \quad (3)$$

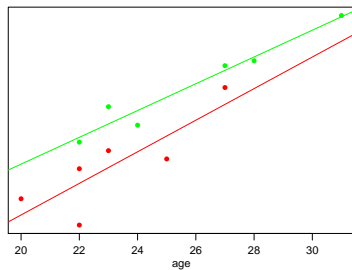
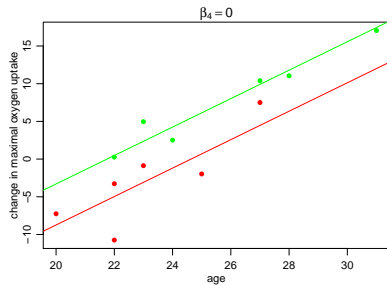
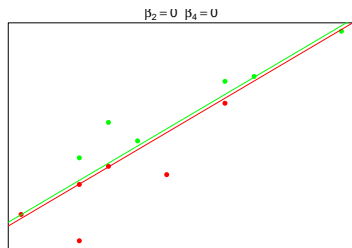
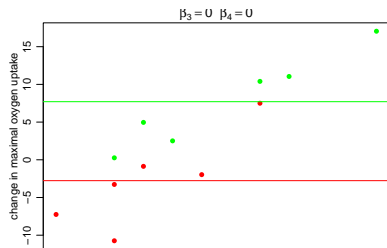
$$x_{i4} = x_{i2} \times x_{i3} \quad (4)$$

Under this model,

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

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

# Least squares regression lines



## Multivariate Setup

Let's assume that we have data points  $(\mathbf{x}_i, \mathbf{y}_i)$  available for all  $i = 1, \dots, n$ .

- ▶  $y$  is the response variable

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

- ▶  $\mathbf{x}_i$  is the  $i$ th row of the design matrix  $\mathbf{X}_{n \times p}$ .

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

- ▶  $E[Y | x]$  is linear and
- ▶ the sampling variability around the mean is independent and identically (iid) from a normal distribution

$$Y_i = \beta^T x_i + \epsilon_i \quad (5)$$

$$\epsilon_1, \dots, \epsilon_n \stackrel{iid}{\sim} \text{Normal}(0, \sigma^2) \quad (6)$$

We can specify a simple Bayesian model as the following:

$$\mathbf{y} | X, \beta, \sigma^2 \sim \text{MVN}(X\beta, \sigma^2 I)$$

$$\beta \sim \text{MVN}(0, \tau^2 I)$$

## Normal Regression Model (continued)

This specifies the density of the data:

$$p(y_1, \dots, y_n \mid x_1, \dots, x_n, \beta, \sigma^2) \quad (7)$$

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

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

$$= (2\pi\sigma^2)^{-n/2} \exp\{(\mathbf{y} - X\beta)^T (\mathbf{y} - X\beta)\} \quad (10)$$

# Ordinary Least Squares

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

$$\hat{\beta} = \arg \min_{\beta \in \mathbb{R}^p} \|\mathbf{y} - X\beta\|_2^2$$

This gives

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

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

The fitted values are

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

This is a linear function of  $\mathbf{y}$ ,  $\hat{\mathbf{y}} = H\mathbf{y}$ , 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.

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

Then

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

$$= \frac{\partial \mathbf{Y}^T \mathbf{Y} - 2\hat{\beta}^T X^T \mathbf{Y} + \hat{\beta}^T (X^T X) \hat{\beta}}{\partial d\hat{\beta}} \quad (12)$$

$$= -2X^T \mathbf{Y} + 2X^T X \hat{\beta} \quad (13)$$

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

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



## Recall data set up

```
# 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)

(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
```

# 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

## Ordinary Least squares estimation

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

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

$$\text{Var}(\hat{\beta}) = \text{Var}\{(X^T X)^{-1} X^T \mathbf{Y}\} \quad (14)$$

$$= (X^T X)^{-1} X^T \sigma^2 I_n X (X^T X)^{-1} \quad (15)$$

$$= \sigma^2 (X^T X)^{-1} \quad (16)$$

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

## Posterior computation

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

$$p(\beta \mid y, X) \propto p(y \mid X, \beta)p(\beta) \quad (17)$$

$$\propto \exp\{-a/2(y - X\beta)^T(y - X\beta)\} \times \exp\{-b/2\beta^T\beta\} \quad (18)$$

Just like in the Multivariate modules, we just simplify. (Check these details on your own).

$$p(\beta \mid \mathbf{y}, X) \propto \text{MVN}(\beta \mid \mathbf{y}, X, \Lambda^{-1})$$

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

## Posterior computation (details)

$$p(\boldsymbol{\beta} \mid \mathbf{y}, X) \quad (19)$$

$$\propto \exp\left\{-\frac{a}{2}(\mathbf{y} - X\boldsymbol{\beta})^T(\mathbf{y} - X\boldsymbol{\beta})\right\} \times \exp\left\{-\frac{b}{2}\boldsymbol{\beta}^T\boldsymbol{\beta}\right\} \quad (20)$$

$$\propto \exp\left\{-\frac{a}{2}[\mathbf{y}^T\mathbf{y} - 2\boldsymbol{\beta}^T X^T \mathbf{y} + \boldsymbol{\beta}^T X^T X \boldsymbol{\beta}] - \frac{b}{2}\boldsymbol{\beta}^T \boldsymbol{\beta}\right\} \quad (21)$$

$$\propto \exp\left\{a\boldsymbol{\beta}^T X^T \mathbf{y} - \frac{a}{2}\boldsymbol{\beta}^T X^T X \boldsymbol{\beta} - b/2\boldsymbol{\beta}^T \boldsymbol{\beta}\right\} \quad (22)$$

$$\propto \exp\left\{a\boldsymbol{\beta}^T [X^T \mathbf{y}] - 1/2\boldsymbol{\beta}^T (aX^T X + bl)\boldsymbol{\beta}\right\} \quad (23)$$

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

## Multivariate inference for regression models

$$\mathbf{y} \mid \boldsymbol{\beta} \sim \text{MVN}(\mathbf{X}\boldsymbol{\beta}, \sigma^2 \mathbf{I}) \quad (24)$$

$$\boldsymbol{\beta} \sim \text{MVN}(\boldsymbol{\beta}_0, \boldsymbol{\Sigma}_0) \quad (25)$$

The posterior can be shown to be

$$\boldsymbol{\beta} \mid \mathbf{y}, \mathbf{X} \sim \text{MVN}(\boldsymbol{\beta}_n, \boldsymbol{\Sigma}_n)$$

where

$$\boldsymbol{\beta}_n = E[\boldsymbol{\beta} \mid \mathbf{y}, \mathbf{X}, \sigma^2] = (\boldsymbol{\Sigma}_0^{-1} + (\mathbf{X}^T \mathbf{X})^{-1} / \sigma^2)^{-1} (\boldsymbol{\Sigma}_0^{-1} \boldsymbol{\beta}_0 + \mathbf{X}^T \mathbf{y} / \sigma^2)$$

$$\boldsymbol{\Sigma}_n = \text{Var}[\boldsymbol{\beta} \mid \mathbf{y}, \mathbf{X}, \sigma^2] = (\boldsymbol{\Sigma}_0^{-1} + (\mathbf{X}^T \mathbf{X})^{-1} / \sigma^2)^{-1}$$

# Multivariate inference for regression models

The posterior can be shown to be

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

where

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

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

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

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

# 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

```
read.table("https://www.stat.washington.edu/~pdhoff/Book/Data")
```

##		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

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

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

$$Y_i \sim \mathcal{N}_6 \left( X\beta_i, \tau_i^{-1} \mathcal{I}_6 \right)$$

$$\beta_i \sim \mathcal{N}_2 (\beta_0, \Sigma_0)$$

$$\tau_i \sim \text{Gamma}(a, b).$$

Derive full conditionals for  $\beta_i$  and  $\tau_i$ .

## Solution (Task 1)

The conditional posterior for  $\beta_i$  is multivariate normal with

$$\mathbb{V}[\beta_i | Y_i, X_i, \tau_i] = (\Sigma_0^{-1} + \tau X_i^T X_i)^{-1}$$

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

while

$$\tau_i | Y_i, X_i, \beta \sim \text{Gamma} \left( a + 3, b + \frac{(Y_i - X_i \beta_i)^T (Y_i - X_i \beta_i)}{2} \right).$$

These can be found in in Hoff in section 9.2.1.

## Task 2

Complete the prior specification by choosing  $a$ ,  $b$ ,  $\beta_0$ , and  $\Sigma_0$ . 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  $\Pr\{y_i^* = \max(y_1^*, y_2^*, y_3^*, y_4^*)\}$  for each swimmer  $i$  and use these probabilities to make a recommendation to the coach.

- This is left as an exercise.