

# Module 9: Linear Regression

Rebecca C. Steorts

Hoff, Chapter 9

# Remainder of Semester

## **This week**

- ▶ Thursday, November 5: Linear Regression
- ▶ Friday, November 6: Lab 10 (Homework 8)

## **Next week**

- ▶ Tuesday, November 10: Linear and Logistic Regression
- ▶ Thursday, November 12: Logistic Regression + Final Exam
- ▶ Friday November 13, 5 PM EDT: Homework 8 (last homework + extra credit)

## **Reading period**

- ▶ All OH will be held. Mine will be during class to avoid any conflicts and be more friendly to international students.

# Final Exam

The final exam is **November 22, 2 PM - 5 PM EDT** (open note/open book)

- ▶ The material will be on modules 1 – 9.
- ▶ I will go over more details regarding the exam next week

# Agenda

- ▶ Oxygen uptake example
- ▶ Linear regression
- ▶ Multiple Linear Regression
- ▶ Ordinary Least Squares
- ▶ An application to swimmers (Lab 10 and Homework 8)

# Oxygen uptake case study

Experimental design: 12 male volunteers.

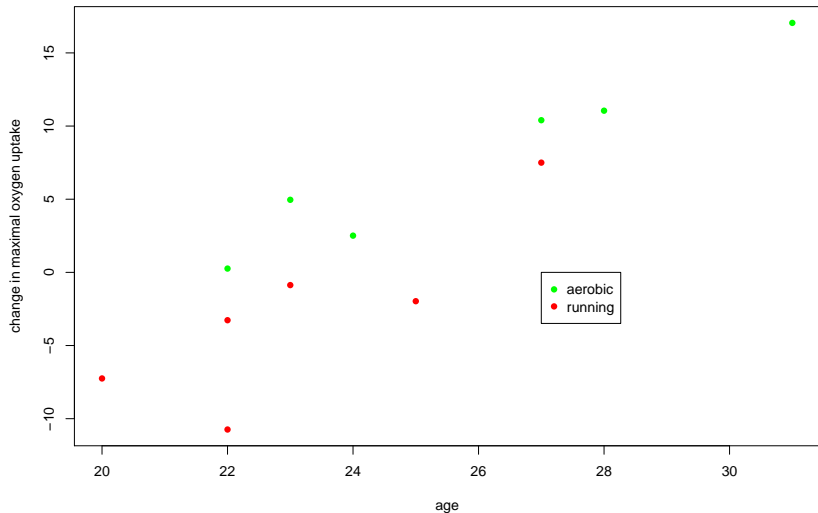
1.  $O_2$  uptake measured at the beginning of the study
2. 6 men take part in a randomized aerobics program
3. 6 remaining men participate in a running program
4.  $O_2$  uptake measured at end of study

What type of exercise is the most beneficial?

# Data

```
# 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 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,

$$y = \beta_0 + \beta_1 x_2 + \epsilon$$

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

# Linear regression

We could also try the model

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

which is also a linear regression model.

# Notation

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

## Notation (continued)

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

- ▶ A column of  $\mathbf{x}$  represents a particular covariate we might be interested in, such as age of a person.
- ▶ Denote  $x_i$  as the  $i$ th **row vector** of the  $\mathbf{X}_{n \times p}$  matrix.

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

## Notation (continued)

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

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

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

# 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  $\mathbf{x}_i$ .
- ▶ Which  $\mathbf{x}_i$ 's have an effect?

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

## 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

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

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

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

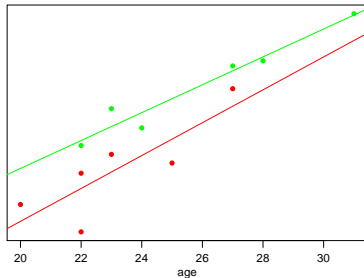
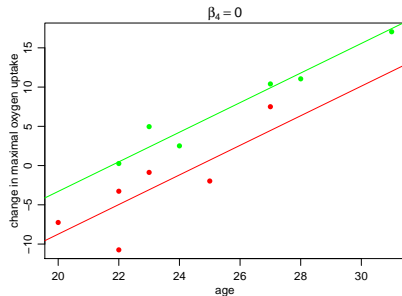
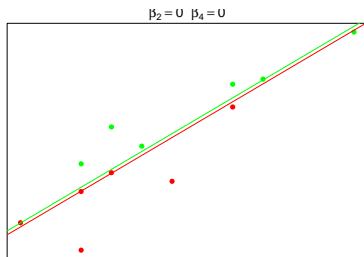
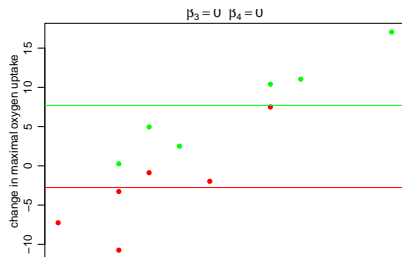
$$x_{i4} = x_{i2} \times x_{i3} \quad (4)$$

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  $(x_i, y_i)$  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  $X_{n \times p}$ .

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

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

$$Y_i = \beta^T \mathbf{x}_i + \epsilon_i \tag{5}$$

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

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

# Multivariate Bayesian Normal Regression Model

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

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

We can specify a multivariate Bayesian model as:

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

where  $\sigma^2, \tau^2$  are known.

# Bayesian Normal Regression Model

The likelihood is

$$p(y_1, \dots, y_n \mid \mathbf{x}_1, \dots, \mathbf{x}_n, \boldsymbol{\beta}, \sigma^2) \quad (7)$$

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

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

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

# Background

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

$$\|\mathbf{y}\|_2 = \sqrt{y_1^2 + \dots + y_n^2}.$$

It follows that

$$\|\mathbf{y}\|_2^2 = y_1^2 + \dots + y_n^2.$$

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

# Background

We would like to find

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

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  $\hat{\beta} \in \mathbb{R}^p$  by least squares:

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

One can show that

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

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**.

## Exercise 1 (OLS)

Let SSR denote sum of squared residuals.

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

Show that

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



## Ordinary Least squares estimation

Proof: Observe

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

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

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

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

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

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

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

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

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

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

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

## 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 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

## Multivariate inference for regression models

$$\mathbf{y} \mid \beta \sim \text{MVN}(X\beta, \sigma^2 I) \quad (17)$$

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

The posterior can be shown to be

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

where

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

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



# 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

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

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

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  $\beta_n \approx \beta_0$

## 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.

I would like for your to continue this on your own.

## Linear Regression Applied to Swimming (Lab 10)

- ▶ 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.
- ▶ This corresponds with Lab 10 and Homework 8 (the last homework)!

## Data set

```
read.table("data/swim.dat",header=FALSE)
```

```
## Warning in read.table("data/swim.dat", header = FALSE):  
## found by readTableHeader on 'data/swim.dat'
```

```
##      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:

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

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

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

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

## Solution (Task 1)

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

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

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

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.

**I highly recommend that you derive these as practice for the final exam.**

## 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.

# Final Grades

I am proposing to drop your lowest exam grade (out of Exam I, Exam II, and the Final Exam).

- ▶ Homework: 30 percent
- ▶ Highest Exam: 35
- ▶ Lowest Exam: 35
- ▶ So your two highest exam scores will be weighted evenly and your lowest exam score will be completely dropped.
- ▶ Yes, you still must take the final exam.

# Course Evaluations

- ▶ I would be very appreciative if you would fill out the course evaluations
- ▶ They are located on DukeHub
- ▶ Directions: <https://assessment.trinity.duke.edu/students-course-evaluations>
- ▶ If there is a 100 percent response rate, I will give everyone in the course 1 point on their final exam grade.

## Exam II

- ▶ Students did extremely well on problems 1 and 2.
- ▶ You should feel very proud of how you did on these problems.  
Great job!

## Exam II

There were issues on both problems 3 and 4.

Due to this, I'm releasing these problems so that you can go over these for the final.

No solutions will be posted to encourage you to work through these derivations on your own.