Lecture 12: Estimation for linear regression

Ciaran Evans

Recap: optimization

Definition: *Optimization* is the problem of finding values that minimize or maximize some function.

Example:

$$RSS(\beta_0, \beta_1) = \sum_{i=1}^{n} (\text{Weight}_i - \beta_0 - \beta_1 \text{WingLength}_i)^2$$

- $RSS(\beta_0, \beta_1)$ is a function of β_0 and β_1
- We want to find the values of β_0 and β_1 that *minimize* this function

Previously: derivative-free methods

$$RSS(\beta_0, \beta_1) = \sum_{i=1}^{n} (\text{Weight}_i - \beta_0 - \beta_1 \text{WingLength}_i)^2$$

- **Compass search:** search along compass directions; move to points of lower RSS, shrink step size when needed
- **Nelder-Mead:** search through transformations of the triangle; allows both increasing and decreasing "step size"

Today: Beginning to use the *derivative* to optimize a function

Question: How do I use the derivative to find a maximum/minimum?

Take derivative, $\overset{\text{set}}{=} 0$

Preliminaries: linear regression in matrix form

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

Matrix form:

$$\begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} \beta_0 + \beta_1 X_1 \\ \beta_0 + \beta_1 X_2 \\ \vdots \\ \beta_0 + \beta_1 X_n \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

Preliminaries: linear regression in matrix form

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

contains a column of 1s, and a column for each explanatory variable

Matrix form:

$$\mathbf{y} = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} \beta_0 + \beta_1 X_1 \\ \beta_0 + \beta_1 X_2 \\ \vdots \\ \beta_0 + \beta_1 X_n \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix} = \underbrace{\begin{bmatrix} 1 & X_1 \\ 1 & X_2 \\ \vdots & \vdots \\ 1 & X_n \end{bmatrix}}_{X_D} \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

vector of responses

design matrix

β vector of coefficients

ε

In more concise form:

$$\mathbf{y} = \mathbf{X}_D \beta + \varepsilon$$

Derivatives for the linear regression model

$$Y = X_D \beta + \varepsilon$$

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

Want to minimize

$$RSS(\beta_0, \beta_1) = \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_i)^2$$

Goal: Take the derivative and set equal to 0

Question: We have *two* variables here – β_0 and β_1. What do I take the derivative with respect to?

Both of them!

Partial derivatives

Example:

$$f(x, y) = x^2 + 2xy + y^3$$

- Derivative *with respect to x*: ← treat y as a constant!

$$\frac{\partial f}{\partial x} = 2x + 2y$$

- Derivative *with respect to y*: ← treat x as a constant!

$$\frac{\partial f}{\partial y} = 2x + 3y^2$$

Derivatives for the linear regression model

$$RSS(\beta_0, \beta_1) = \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_i)^2$$

Partial derivatives:

$$\frac{\partial}{\partial \beta_0} RSS = \sum_{i=1}^{n} 2(Y_i - \beta_0 - \beta_1 X_i)(-1)$$

$$= -2 \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_i)$$

$$\frac{\partial}{\partial \beta_1} RSS = \sum_{i=1}^{n} 2(Y_i - \beta_0 - \beta_1 X_i)(-X_i)$$

$$= -2 \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_i) X_i$$

Gradient

Recall: $Y = X_D \beta + \varepsilon$

The **gradient** is the vector of partial derivatives:

$$\underbrace{\nabla RSS}_{\text{gradient}} = \begin{pmatrix} \frac{\partial}{\partial \beta_0} RSS \\ \frac{\partial}{\partial \beta_1} RSS \end{pmatrix} = -2 \begin{pmatrix} \sum_i (Y_i - \beta_0 - \beta_1 X_i) \\ \sum_i (Y_i - \beta_0 - \beta_1 X_i) X_i \end{pmatrix}$$

Note: vector $v = \begin{pmatrix} v_1 \\ \vdots \\ v_n \end{pmatrix}$ then $1^T v = \sum_i v_i$ $\quad 1 = \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix}$

$$\Rightarrow \sum_i (Y_i - \beta_0 - \beta_1 X_i) = 1^T \begin{pmatrix} Y_1 - \beta_0 - \beta_1 X_1 \\ Y_2 - \beta_0 - \beta_1 X_2 \\ \vdots \\ Y_n - \beta_0 - \beta_1 X_n \end{pmatrix}$$

$$\Rightarrow \nabla RSS = -2 \begin{pmatrix} 1 & 1 & \cdots & 1 \\ X_1 & X_2 & \cdots & X_n \end{pmatrix} \begin{pmatrix} Y_1 - \beta_0 - \beta_1 X_1 \\ Y_2 - \beta_0 - \beta_1 X_2 \\ \vdots \\ Y_n - \beta_0 - \beta_1 X_n \end{pmatrix}$$

$$\nabla RSS = -2 \underbrace{\begin{pmatrix} 1 & 1 & \cdots & 1 \\ x_1 & x_2 & \cdots & x_n \end{pmatrix}}_{X_D^T} \underbrace{\begin{pmatrix} y_1 - \beta_0 - \beta_1 x_1 \\ y_2 - \beta_0 - \beta_1 x_2 \\ \vdots \\ y_n - \beta_0 - \beta_1 x_n \end{pmatrix}}_{\begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} - \begin{pmatrix} \beta_0 + \beta_1 x_1 \\ \vdots \\ \beta_0 + \beta_1 x_n \end{pmatrix}}$$

$$y \quad - \quad X_D \beta$$

$$X_D = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix}$$

$$\beta = \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix}$$

$$\Rightarrow \quad \nabla RSS = -2 X_D^T (y - X_D \beta)$$

Gradient

To minimize RSS, we set the gradient equal to 0 and solve for β:

$$\nabla RSS = \mathbf{X}_D^T(\mathbf{y} - \mathbf{X}_D\beta) \overset{\text{set}}{=} 0$$

0 vector

entries = # βs

e.g. $\beta = \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix}$

then $0 = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$

$$X_D^T y - X_D^T X_D \beta = 0$$

$$\Rightarrow X_D^T y = X_D^T X_D \beta$$

$$\Rightarrow (X_D^T X_D)^{-1} X_D^T y = \beta$$

So, our estimates are

$$\boxed{\hat{\beta} = (X_D^T X_D)^{-1} X_D^T y}$$

- closed form solution
- doesn't require searching

Least squares linear regression solution

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

Example: Regression with the sparrows data

```
lm(Sparrows$Weight ~ Sparrows$WingLength) |> coef()
```

```
##          (Intercept) Sparrows$WingLength
##             1.365490            0.467404
```

Question: How would we compute $(\mathbf{X}_D^T\mathbf{X}_D)^{-1}\mathbf{X}_D^T\mathbf{y}$ in R?

- make our design matrix X_D
- transpose in R: t( )
- matrix mult: %*%
- matrix inverse: solve(..)

Least squares linear regression solution

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

Example: Regression with the sparrows data

```
lm(Sparrows$Weight ~ Sparrows$WingLength) |> coef()
```

```
##          (Intercept) Sparrows$WingLength
##             1.365490            0.467404
```

Question: How would we compute $(\mathbf{X}_D^T \mathbf{X}_D)^{-1} \mathbf{X}_D^T \mathbf{y}$ in R?

```
y <- Sparrows$Weight
XD <- cbind(1, Sparrows$WingLength)
solve(t(XD) %*% XD) %*% t(XD) %*% y
```

```
##              [,1]
## [1,] 1.365490
## [2,] 0.467404
```

Optimization

Possibilities so far

- Derivatives are hard / expensive to find (or we don't want to calculate them)
  - Derivative-free optimization!
- Derivatives can be calculated and lead to a closed-form solution
  - Example: the usual linear regression model

Another possibility

- Derivatives can be calculated, but there is no closed-form solution to the system
  - Example: logistic regression
  - **Question:** what should we do if there is no closed-form solution?

iterate!

Optimization

Possibilities so far

- Derivatives are hard / expensive to find (or we don't want to calculate them)
  - Derivative-free optimization!
- Derivatives can be calculated and lead to a closed-form solution
  - Example: the usual linear regression model

Another possibility

- Derivatives can be calculated, but there is no closed-form solution to the system
  - Example: logistic regression

Next time: Begin iterative procedures using derivative information

Your turn

Practice questions on the course website:

https://sta379-s25.github.io/practice_questions/pq_12.html

- Fit a linear regression model
- Take derivatives for a logistic regression model
- Start in class. You are welcome to work with others
- Practice questions are to help you practice. They are not submitted and not graded
- Solutions are posted on the course website