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?

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

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

Matrix form:

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

In more concise form:

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

Derivatives for the linear regression model

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

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?

Partial derivatives

Example:

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

- Derivative *with respect to* x:

$$\frac{\partial f}{\partial x} =$$

- Derivative *with respect to* y:

$$\frac{\partial f}{\partial y} =$$

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:

$$\frac{\partial}{\partial \beta_0} RSS =$$

$$\frac{\partial}{\partial \beta_1} RSS =$$

Gradient

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

$$\nabla RSS = \begin{pmatrix} \frac{\partial}{\partial \beta_0} RSS \\ \frac{\partial}{\partial \beta_1} RSS \end{pmatrix} =$$

Gradient

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

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

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?

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?

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