Department of Computer Science
UNIVERSITY OF COLORADO **BOULDER**

Linear Regression

Introduction to Data Science Algorithms

Jordan Boyd-Graber and Michael Paul

SLIDES ADAPTED FROM FEDERICO

Deriving OLS

- Common theme in data science:
  - Build model
  - Write error model
  - Derive how to minimize error
- Practice for OLS (other models next week)

Model and Objective

Model

$$y_i = b_0 + b_1 x_i + e_i \tag{1}$$

Error

$$e_i = y_i - b_1 x_i - b_0 = e_i \tag{2}$$

Objective

$$\ell \equiv \sum_i e_i^2 \tag{3}$$

Partial Derivatives

Intercept

$$\frac{\partial \ell}{\partial b_0} = \frac{\partial \sum_i (y_i - b_0 - b_1 x_i)^2}{\partial b_0} =$$

Partial Derivatives

Intercept

$$\frac{\partial \ell}{\partial b_0} = \frac{\partial \sum_i (y_i - b_0 - b_1 x_i)^2}{\partial b_0} = -2 \sum_i (y_i - b_0 - b_1 x_i) \tag{4}$$

Partial Derivatives

Intercept

$$\frac{\partial \ell}{\partial b_0} = \frac{\partial \sum_i (y_i - b_0 - b_1 x_i)^2}{\partial b_0} = -2 \sum_i (y_i - b_0 - b_1 x_i) \tag{4}$$

Slope

$$\frac{\partial \ell}{\partial b_1} = \frac{\partial \sum_i (y_i - b_0 - b_1 x_i)^2}{\partial b_1} =$$

Partial Derivatives

Intercept

$$\frac{\partial \ell}{\partial b_0} = \frac{\partial \sum_i (y_i - b_0 - b_1 x_i)^2}{\partial b_0} = -2 \sum_i (y_i - b_0 - b_1 x_i) \tag{4}$$

Slope

$$\frac{\partial \ell}{\partial b_1} = \frac{\partial \sum_i (y_i - b_0 - b_1 x_i)^2}{\partial b_1} = -2 \sum_i x_i (y_i - b_0 - b_1 x_i) \tag{5}$$

System of Equations with Two Unknowns

Solve for Intercept

(6)

System of Equations with Two Unknowns

Solve for Intercept

$$0 = -2\sum_i (y_i - b_0 - b_1 x_i) \tag{6}$$

$$\tag{7}$$

System of Equations with Two Unknowns

Solve for Intercept

$$0 = -2\sum_i (y_i - b_0 - b_1 x_i) \tag{6}$$

$$0 = \sum_i y_i - \sum_i b_0 - b_i \sum_i x_i \tag{7}$$

$$\tag{8}$$

Multiply by $-\frac{1}{2}$, distribute sum

System of Equations with Two Unknowns

Solve for Intercept

$$0 = -2\sum_i (y_i - b_0 - b_1 x_i) \tag{6}$$

$$0 = \sum_i y_i - \sum_i b_0 - b_i \sum_i x_i \tag{7}$$

$$Nb_0 = \sum_i y_i - b_i \sum_i x_i \tag{8}$$

$$\tag{9}$$

b_0 is constant, so $\sum_i b_0 = Nb_0$, move to LHS

System of Equations with Two Unknowns

Solve for Intercept

$$0 = -2\sum_i (y_i - b_0 - b_1 x_i) \tag{6}$$

$$0 = \sum_i y_i - \sum_i b_0 - b_i \sum_i x_i \tag{7}$$

$$Nb_0 = \sum_i y_i - b_i \sum_i x_i \tag{8}$$

$$b_0 = \left(\frac{\sum_i y_i}{N}\right) - b_1 \left(\frac{\sum_i x_i}{N}\right) \tag{9}$$

$$\tag{10}$$

Divide by N

System of Equations with Two Unknowns

Solve for Intercept

$$0 = -2\sum_i (y_i - b_0 - b_1 x_i) \tag{6}$$

$$0 = \sum_i y_i - \sum_i b_0 - b_i \sum_i x_i \tag{7}$$

$$Nb_0 = \sum_i y_i - b_i \sum_i x_i \tag{8}$$

$$b_0 = \left(\frac{\sum_i y_i}{N}\right) - b_1 \left(\frac{\sum_i x_i}{N}\right) \tag{9}$$

$$b_0 = \bar{y} - b_1 \bar{x} \tag{10}$$

System of Equations with Two Unknowns

Solve for Intercept

$$b_0 = \bar{y} - b_1 \bar{x} \tag{6}$$

Solve for Slope

$$\tag{7}$$

System of Equations with Two Unknowns

Solve for Intercept

$$b_0 = \bar{y} - b_1 \bar{x} \tag{6}$$

Solve for Slope

$$0 = -2 \sum_i x_i (y_i - b_0 - b_1 x_i) \tag{7}$$

$$\tag{8}$$

System of Equations with Two Unknowns

Solve for Intercept

$$b_0 = \bar{y} - b_1 \bar{x} \tag{6}$$

Solve for Slope

$$0 = -2 \sum_i x_i (y_i - b_0 - b_1 x_i) \tag{7}$$

$$0 = \sum_i x_i y_i - b_0 \sum_i x_i - \sum_i b_1 x_i^2 \tag{8}$$

$$\tag{9}$$

Multiply by $-\frac{1}{2}$, distribute sum and x_i

System of Equations with Two Unknowns

Solve for Intercept

$$b_0 = \bar{y} - b_1 \bar{x} \tag{6}$$

Solve for Slope

$$0 = -2 \sum_i x_i (y_i - b_0 - b_1 x_i) \tag{7}$$

$$0 = \sum_i x_i y_i - b_0 \sum_i x_i - \sum_i b_1 x_i^2 \tag{8}$$

$$b_1 \sum_i x_i^2 = \sum_i x_i y_i - b_0 \sum_i x_i \tag{9}$$

$$\tag{10}$$

Move last term to RHS

System of Equations with Two Unknowns

Solve for Intercept

$$b_0 = \bar{y} - b_1 \bar{x} \tag{6}$$

Solve for Slope

$$0 = -2 \sum_i x_i (y_i - b_0 - b_1 x_i) \tag{7}$$

$$0 = \sum_i x_i y_i - b_0 \sum_i x_i - \sum_i b_1 x_i^2 \tag{8}$$

$$b_1 \sum_i x_i^2 = \sum_i x_i y_i - b_0 \sum_i x_i \tag{9}$$

$$b_1 \sum_i x_i^2 = \sum_i x_i y_i - \left[\left(\frac{\sum_i y_i}{N} \right) - b_1 \left(\frac{\sum_i x_i}{N} \right) \right] \sum_i x_i \tag{10}$$

Solve for Slope (continued)

$$b_1 \sum_i x_i^2 = \sum_i x_i y_i - \left[\left(\frac{\sum_i y_i}{N} \right) - b_1 \left(\frac{\sum_i x_i}{N} \right) \right] \sum_i x_i$$

Solve for Slope (continued)

$$b_1 \sum_i x_i^2 = \sum_i x_i y_i - \left[\left(\frac{\sum_i y_i}{N} \right) - b_1 \left(\frac{\sum_i x_i}{N} \right) \right] \sum_i x_i$$

$$b_1 \sum_i x_i^2 = \sum_i x_i y_i - \left(\frac{\sum_i y_i \sum_i x_i}{N} \right) - b_1 \left(\frac{(\sum_i x_i)^2}{N} \right)$$

Multiplying out the last term

Solve for Slope (continued)

$$b_1 \sum_i x_i^2 = \sum_i x_i y_i - \left[\left(\frac{\sum_i y_i}{N} \right) - b_1 \left(\frac{\sum_i x_i}{N} \right) \right] \sum_i x_i$$

$$b_1 \sum_i x_i^2 = \sum_i x_i y_i - \left(\frac{\sum_i y_i \sum_i x_i}{N} \right) - b_1 \left(\frac{(\sum_i x_i)^2}{N} \right)$$

$$b_1 \sum_i x_i^2 + b_1 \left(\frac{(\sum_i x_i)^2}{N} \right) = \sum_i x_i y_i - \left(\frac{\sum_i y_i \sum_i x_i}{N} \right)$$

Move last term to LHS

Solve for Slope (continued)

$$b_1 \sum_i x_i^2 = \sum_i x_i y_i - \left[\left(\frac{\sum_i y_i}{N} \right) - b_1 \left(\frac{\sum_i x_i}{N} \right) \right] \sum_i x_i$$

$$b_1 \sum_i x_i^2 = \sum_i x_i y_i - \left(\frac{\sum_i y_i \sum_i x_i}{N} \right) - b_1 \left(\frac{(\sum_i x_i)^2}{N} \right)$$

$$b_1 \sum_i x_i^2 + b_1 \left(\frac{(\sum_i x_i)^2}{N} \right) = \sum_i x_i y_i - \left(\frac{\sum_i y_i \sum_i x_i}{N} \right)$$

$$b_1 \left[\sum_i x_i^2 + \left(\frac{(\sum_i x_i)^2}{N} \right) \right] = \sum_i x_i y_i - \left(\frac{\sum_i y_i \sum_i x_i}{N} \right)$$

Factor out b_1

Solve for Slope (continued)

$$
\begin{aligned}
b_1 \sum_i x_i^2 &= \sum_i x_i y_i - \left[\left(\frac{\sum_i y_i}{N} \right) - b_1 \left(\frac{\sum_i x_i}{N} \right) \right] \sum_i x_i \\
b_1 \sum_i x_i^2 &= \sum_i x_i y_i - \left(\frac{\sum_i y_i \sum_i x_i}{N} \right) - b_1 \left(\frac{(\sum_i x_i)^2}{N} \right) \\
b_1 \sum_i x_i^2 + b_1 \left(\frac{(\sum_i x_i)^2}{N} \right) &= \sum_i x_i y_i - \left(\frac{\sum_i y_i \sum_i x_i}{N} \right) \\
b_1 \left[\sum_i x_i^2 + \left(\frac{(\sum_i x_i)^2}{N} \right) \right] &= \sum_i x_i y_i - \left(\frac{\sum_i y_i \sum_i x_i}{N} \right) \\
b_1 &= \frac{\sum_i x_i y_i - \left(\frac{\sum_i y_i \sum_i x_i}{N} \right)}{\sum_i x_i^2 + \left(\frac{(\sum_i x_i)^2}{N} \right)}
\end{aligned}
$$

Solve for Slope (continued)

$$b_1 = \frac{\sum_i x_i y_i - \left(\frac{\sum_i y_i \sum_i x_i}{N}\right)}{\sum_i x_i^2 + \left(\frac{(\sum_i x_i)^2}{N}\right)}$$

Ratio of the sum of the crossproducts of x and y over the sum of squares for x