CSCI 3022 Intro to Data Science

Notebook day: 18-19 (maybe 20)

SLR: Simple linear regression
(1 x-value, 1 y-value)

OLS: ordinary least-squares

Prove that the OLS estimators: estimated

1. $\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$ intercept

2. $\hat{\beta}_1 = \frac{Cov[X,Y]}{\sum_{i=1}^n \left(X_i - \bar{X}\right)^2} = \frac{\sum_{i=1}^n \left(X_i - \bar{X}\right)\left(Y_i - \bar{Y}\right)}{\sum_{i=1}^n \left(X_i - \bar{X}\right)^2}$ slope

$(2-y)^2 + (3-y)^2 + (\pi - y)^2 = ay^2 + by + c$

satisfy

line

$$\hat{\beta}_0, \hat{\beta}_1 = \underset{\beta_0, \beta_1}{\text{argmin}} \sum_{i=1}^n \left(Y_i - [\beta_0 - \beta_1 X_i]\right)^2$$

data

calculus/optimization

Hint: what are $\frac{d}{dy}$ of

+ β_0 β_1

$$f(y) = \sum_{i=1}^n (a - y - bc)^2; \quad g(y) = \sum_{i=1}^n (a - b - yc)^2?$$

SLR Overview

Problem: use *predictor* x to describe *response* y, using a line. We're subject to *error*, *noise*, or *unexplained variability* ε.

1. Model: $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$
2. Terms: β_0 : the intercept of the goal line; β_1 : its slope
3. Assumptions on ε: **Independence, Homoskedasticity, Normality**

Goal:

Given sample data, which consists of n observed pairs, $(x_1, y_1), (x_2, y_2), \dots (x_n, y_n)$, construct an estimated "line of best fit":

$$y = \hat{\beta}_0 + \hat{\beta}_1 x$$

This line can then be used to make predictions or provide explanations for unobserved phenomena.

Estimating SLR Parameters

A line provides the **best fit** to the data if the sum of the squared vertical distances (deviations) from the observed points to that line is as small as it can be.

The sum of *squared vertical deviations* from the data points to the line $y = \beta_0 + \beta_1 x$ is then

$$\sum_{i=1}^{n} \left(\overbrace{Y_i}^{\text{Data}} - \overbrace{\beta_0 - \beta_1 X_i}^{\text{Line}} \right)^2$$

The point estimates of β_0 and β_1, denoted $\underline{\ \hat{\beta}_0; \hat{\beta}_1\ }$ are called the *least squares estimates*. They are those values that minimize SSE or sum of squared errors.

$$\hat{\beta}_0, \hat{\beta}_1 = \underset{\beta_0, \beta_1}{\operatorname{argmin}} \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_i)^2$$

(x_i, y_i)

Estimating SLR Parameters: Pen and Paper

Goal: find the minimizers of the function $f(\beta_0, \beta_1) = \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_i)^2$. Sounds like a Calculus problem!

y

$\frac{d}{d\beta_0} \sum$

$\frac{d}{d\beta_0} (f(y) + g(y)) = \frac{d f(y)}{d\beta_0} + \frac{d g(y)}{d\beta_0}$

$y_i - \beta x_i$

$\frac{d}{dy} (\# - y)^2 = \sum \boxed{(\# - y)} \, (0-1)$

d/dy

$= \sum_{i=1}^{n} 2(\# - y)$

d/dy

$\frac{d}{d\beta_1} \approx \frac{d}{dy} \sum (\# - X_i y)^2 = \sum \boxed{(\# - X_i y)} \, (0 - X_i)$

Estimating SLR Parameters: Pen and Paper

Goal: find the minimizers of the function $f(\beta_0, \beta_1) = \sum_{i=1}^n (Y_i - \beta_0 - \beta_1 X_i)^2$. Sounds like a Calculus problem!

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

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

Estimating SLR Parameters: Pen and Paper

Goal: find the minimizers of the function $f(\beta_0, \beta_1) = \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_i)^2$. Sounds like a Calculus problem!

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

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

For finding the joint maximum/minimum of multiple inputs, we end up with a system of equations: set both equal to zero and find the values that make both equal to zero.

Opening Sol'n

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

Partial

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

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

Opening Sol'n

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

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

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

Opening Sol'n

$$\sum_{i=1}^{n} Y_i = \bar{Y} \cdot n$$

$$\frac{\sum_{i=1}^{n} Y_i}{n} = \bar{Y} \Rightarrow$$

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

$$\frac{\partial f}{\partial \beta_0} = \frac{\partial}{\partial \beta_0} \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_i)^2 = \sum_{i=1}^{n} -2 (Y_i - \beta_0 - \beta_1 X_i) \overset{set}{=} 0$$

n

$$\frac{\partial f}{\partial \beta_1} = \frac{d}{d\beta_1} \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_i)^2 = \sum_{i=1}^{n} -2 X_i (Y_i - \beta_0 - \beta_1 X_i) \overset{set}{=} 0$$

Opening Sol'n

From the β_0 row:

drop -2

$$0 = \sum_{i=1}^{n} Y_i - \beta_0 - \beta_1 X_i \implies 0 = n\bar{Y} - n\beta_0 - n\beta_1 \bar{X} \implies \boxed{\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}}$$

intercept: use averages,

$\bar{Y} - \hat{\beta}_1 \bar{X}$

estimated slope.

Opening Sol'n

From the β_0 row:

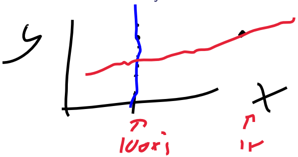
$$0 = \sum Y_i - \beta_0 - \beta_1 X_i \implies 0 = n\bar{Y} - n\beta_0 - n\beta_1 \bar{X} \implies \beta_0 = \bar{Y} - \beta_1 \bar{X}$$

Plugging that into the other row gives

$$0 = \sum X_i (Y_i - (\bar{Y} - \beta_1 \bar{X}) - \beta_1 X_i) = \sum X_i \left(Y_i - \bar{Y} + \beta_1 (\bar{X} - X_i)\right)$$

$$0 = \underbrace{\sum X_i \left(Y_i - \bar{Y}\right)}_{\text{without } \beta_1} + \beta_1 \underbrace{\sum X_i \left(\bar{X} - X_i\right)}_{\text{with } \beta_1} \implies \beta_1 = \frac{\sum X_i \left(Y_i - \bar{Y}\right)}{\sum X_i \left(X_i - \bar{X}\right)}$$

Opening Sol'n



(2,3)

$(2-2.5)+(3-2.5)=0$

$2(2-2.5)+3(3-2.5)=.5$

From the β_0 row:

$$0 = \sum Y_i - \beta_0 - \beta_1 X_i \implies 0 = n\bar{Y} - n\beta_0 - n\beta_1\bar{X} \implies \beta_0 = \bar{Y} - \beta_1\bar{X}$$

Plugging that into the other row gives

$$0 = \sum X_i(Y_i - (\bar{Y} - \beta_1\bar{X}) - \beta_1 X_i) = \sum X_i \left(Y_i - \bar{Y} + \beta_1(\bar{X} - X_i)\right)$$

$$0 = \sum X_i \left(Y_i - \bar{Y}\right) + \beta_1 \sum X_i \left(\bar{X} - X_i\right) \implies \beta_1 = \frac{\sum X_i \left(Y_i - \bar{Y}\right)}{\sum X_i \left(X_i - \bar{X}\right)}$$

... are we done?
Note: $\sum(X_i - \bar{X}) = 0$, so $\bar{X}\sum(X_i - \bar{X}) = 0$. That's the difference between our current solution and the prior slide.

Estimating SLR Parameters: Results

For a model of the form $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i;\ \varepsilon \sim N(0, \sigma^2)$

1. $\hat{\beta}_0 =$

2. $\hat{\beta}_1 =$

What happens if $\beta_0 \approx 0$? If $\beta_1 \approx 0$?

Estimating SLR Parameters: Results

For a model of the form $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i;\ \varepsilon \sim N(0, \sigma^2)$

1. $\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$

intercept uses averages

2. $\hat{\beta}_1 = \frac{Cov[X,Y]}{\sum_{i=1}^{n}(X_i - \bar{X})^2} = \frac{\sum_{i=1}^{n}(X_i-\bar{X})(Y_i-\bar{Y})}{\sum_{i=1}^{n}(X_i-\bar{X})^2}$

slope: $\frac{\text{X \& Y relationship}}{\text{spread of X}}$

$s^2 = \frac{\sum (x_i - \bar{x})^2}{n-1}$

What happens if $\beta_0 \approx 0$? If $\beta_1 \approx 0$?

Estimating SLR Parameters: Results

For a model of the form $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$; $\varepsilon \sim N(0, \sigma^2)$

1. $\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$

2. $\hat{\beta}_1 = \frac{Cov[X,Y]}{\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2} = \frac{\sum_{i=1}^{n}\left(X_i - \bar{X}\right)\left(Y_i - \bar{Y}\right)}{\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2}$

What happens if $\beta_0 \approx 0$? If $\beta_1 \approx 0$?
One result: the regression line goes through $(0, \beta_0)$. It also goes through $(\bar{X}, \bar{Y})$!

Daily Recap

Today we learned

1. Regression!

Moving forward:

- nb day Friday

Next time in lecture:

- More Regression!