Simple Linear Regression

CS 530
Chapman

Spring 2021

1

Take message for rest of course: Simple Linear Regression

- We will see the first instance of a predictive model
  - Linear model fit for continuous data
- We will see how to test
  - Whether the mode is appropriate for the data
  - Whether the mode fits the data at all
  - Whether the data trends according to the model
- We will see how to measure how well the model fits the data

2

Simple Linear Regression—Table of Contents

- What is (simple) linear regression?
- Constructing the linear regression model
- Evaluating the linear regression model

3

Simple Linear Regression

- Simple linear regression attempts to model relation between two variables.
  - Increase in one variable corresponds with proportional increase or decrease in another variable:

[Variable 2] = [Coefficient] × [Variable 1] + [“baseline”]

- Linear regression most appropriate for continuous variables (e.g., temperature, speed), or at least ordinal variables (e.g., education level, letter grading), categorical data (e.g., species, job title).

4

Simple Linear Regression

- Linear relationship represented by curve on graph that "comes closest" to data points.
- Resulting linear model can make predictions about data points where x value *known* & y value *unknown*
- Used when seeking a trend in the data

5

WHAT IS A LINEAR RELATION?

6

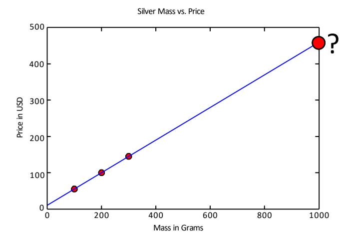
Linear relation: Silver Mass vs. Price

- We start with example of data with perfectly linear relationship.
- Helps define terms
- We see following advertisement online for buying quantities of silver, including shipping & handling.
- We want to buy 1 kg of silver. How much would it cost?

Mass	Price
100 g	\$55.00
200 g	\$100.00
300 g	\$145.00
⋮	⋮
1 kg	?

7

Linear relation: Silver Mass vs. Price



Plot of silver mass (or weight) vs. price: all points line up perfectly along straight line.

8

Linear relation: Silver Mass vs. Price

We can model this with the equation:

$$y = \beta_0 + \beta_1 x,$$

where

x ← independent (explanatory, predictor) variable, representing mass
y ← dependent (response) variable, representing price
β_0 ← y–intercept, representing the shipping and handling charge
β_1 ← slope, representing increase in price per additional gram

9

Linear relation: Silver Mass vs. Price

β1 is rate of change of price per unit mass (change in y over change in x):

$$\beta_1 = \frac{\Delta y}{\Delta x}$$

Using first two data points:

$$\beta_1 = \frac{\Delta y}{\Delta x} = \frac{\$100 - \$55}{200g - 100g}$$

$$\beta_1 = 0.45 \frac{\text{dollars}}{\text{gram}}$$

So, each additional gram of silver costs 45 cents.

10

Linear relation: Silver Mass vs. Price

β_0 is the base cost of shipping & handling, applied equally to any purchase of silver, regardless of size. To find β_0, we plug our value for β_1 & example data point into our equation:

$$y = \beta_0 + \beta_1 x$$

$$\$55 = \beta_0 + \left(0.45 \frac{\text{dollars}}{\text{gram}}\right)(100g)$$

$$\$55 = \beta_0 + \$45$$

$$\beta_0 = \$10$$

Cost of shipping & handling: \$10

11

Linear relation: Silver Mass vs. Price

Price for any bar of silver from online vendor can be modeled by equation

$$y = 10 + 0.45x$$

where

x ← grams of silver
y ← price in dollars

So, as 1 kg = 1000 g, $y = \$10 + \$0.45 \cdot 1000 = \$460$.

12

LINEAR REGRESSION: INTRODUCTION VIA SIMPLE EXAMPLE

13

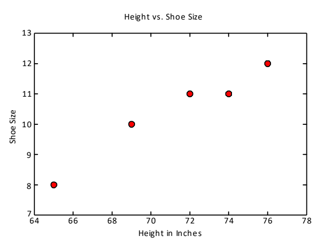
Linear Regression: Height vs. Shoe Size

Simple example: group of men with following heights & shoe sizes. What shoe size would we predict for man of average American height (70 inches)?

Height (inches)	Shoe Size
72	11
74	11
76	12
69	10
65	8
⋮	⋮
70	?

14

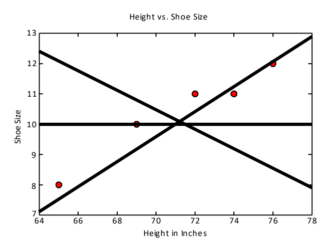
Linear Regression: Height vs. Shoe Size



Scatter plot of height vs. shoe size.
This time it is not a perfect linear relation.

15

Linear Regression: Height vs. Shoe Size



There is infinite number of lines—infinite models—to potentially fit to this data. What is "best fit" line—the "best" model?

16

Linear Regression: Height vs. Shoe Size

Data points don't fit perfectly onto line, but relationship still appears "line-like" in nature. We can model this with equation:

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

where

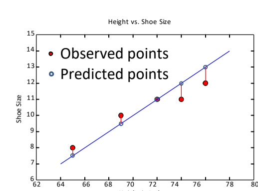
x ← man's height in inches
y ← man's shoe size
β_0 ← theoretical shoe size for a man who is 0" tall
β_1 ← increase in shoe size for each additional 1" in height
ϵ ← error term (a.k.a. noise)

17

Ordinary Least Squares (OLS): Cost Function of Regression

No line can pass through every point in this data set.
Vertical distances between the observed data points & our model line termed <u>residuals</u>—represented by red vertical lines in the graph.

Model (line) with best fit defined as one with "least overall distance" between observed points & predicted points—*best-fit model minimizes sum of residuals.*



18

Ordinary Least Squares (OLS): Cost Function of Regression

Each of these residuals e_i can be written as

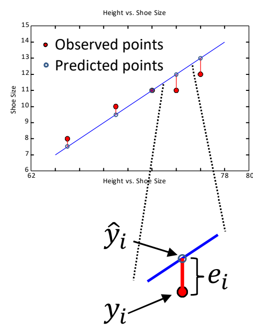
$$e_i = y_i - \hat{y}_i$$

where

e_i ←distance from data point to regression line

y_i ←actual y value for data point at x_i (red/black markers)

$\hat{y}_i$ ←predicted y value at x_i (blue markers on the line)



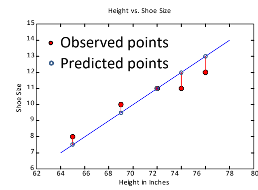
19

Ordinary Least Squares (OLS): Cost Function of Regression

We want to minimize the sum of the squares of these residuals, called the Residual Sum of Squares (RSS):

$$\text{RSS} = \sum_{i=1}^{n} (e_i)^2$$
$$= \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \ ,$$

so that our regression line comes as close as possible—in above sense—to original data points.



Note: above formulation minimizes RSS <u>in y-direction only</u>. It does not minimize distance perpendicular to regression line.

20

ORDINARY LEAST SQUARES (OLS)

21

Ordinary Least Squares (OLS): Deriving the Formulation

OLS assumes that data's true relationship is

$$y = \beta_0 + \beta_1 x + \epsilon.$$

But we have just a sample of data (rather than height & shoe size of every person on earth). So, we look for coefficient estimates $\hat{\beta}_0$ & $\hat{\beta}_1$:

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

The ϵ term designates variability we cannot control (no model can fully explain randomness of real world). Plugging into RSS:

$$\text{RSS} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$
$$= \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2$$

22

Ordinary Least Squares (OLS): Deriving the Formulation

We seek values of $\hat{\beta}_0$ and $\hat{\beta}_1$ that represent closest possible line to our data points (in y direction). So, we wish to minimize RSS.

Viewing RSS as a function of $\hat{\beta}_0$ & $\hat{\beta}_1$ values, we wish to find its minimum. In other words, we want to find lowest point on blue surface on the right (designated as red dot), because that will give us $\hat{\beta}_0$ & $\hat{\beta}_1$ associated with smallest RSS.

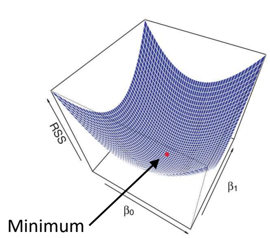
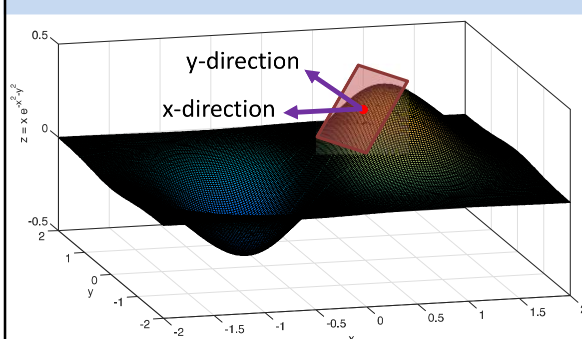


Image Source

23

Reminder: Partial derivatives



24

Reminder: Partial derivatives

How does the derivation work technically?

$$z = f(x,y) = x \cdot e^{-x^2-y^2}$$

$\Rightarrow$

$$\frac{\partial z}{\partial x} = \frac{f(x,y)}{\partial x} = \frac{\partial\left(x \cdot e^{-x^2-y^2}\right)}{\partial x} = e^{-x^2-y^2} - 2x^2 \cdot e^{-x^2-y^2}$$

y kept constant

$$\frac{\partial z}{\partial y} = \frac{f(x,y)}{\partial y} = \frac{\partial\left(x \cdot e^{-x^2-y^2}\right)}{\partial y} = -2xy \cdot e^{-x^2-y^2}$$

x kept constant

25

Ordinary Least Squares (OLS): Deriving the Formulation

The minimum ($\hat{\beta}_0$,$\hat{\beta}_1$) point is where the tangent plane to the surface is horizontal.
This occurs where the function's partial derivatives are all 0.
So, we can find minimum value of RSS—a.k.a. minimize RSS—by setting partial derivatives of RSS, with respect to $\hat{\beta}_0$ & $\hat{\beta}_1$,to 0.

$$\frac{\partial z}{\partial x} = 0; \frac{\partial z}{\partial y} = 0$$

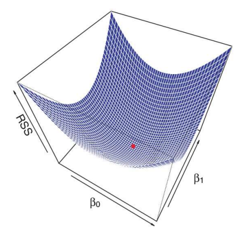


Image Source

26

Ordinary Least Squares (OLS): Deriving the Formulation

Let's start by taking the partial derivative of RSS with respect to $\hat{\beta}_0$, set derivatives equal to 0, and solve for $\hat{\beta}_0$:

$$0 = \frac{\partial(\text{RSS})}{\partial\hat{\beta}_0} = \frac{\partial\left(\sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2\right)}{\partial\hat{\beta}_0} = -2\left[\sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)\right]$$

$$0 = \sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)$$

$$n\hat{\beta}_0 = \sum_{i=1}^{n}(y_i - \hat{\beta}_1 x_i)$$

$$\hat{\beta}_0 = \frac{\sum_{i=1}^{n} y_i}{n} - \hat{\beta}_1 \frac{\sum_{i=1}^{n} x_i}{n}$$

$$\boxed{\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}}$$

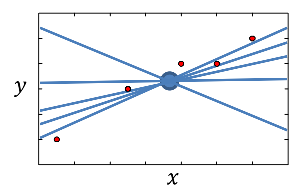
27

Ordinary Least Squares (OLS): Deriving the Formulation

- This yielded the equation

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \quad \text{or} \quad \bar{y} = \hat{\beta}_0 + \hat{\beta}_1 \bar{x}$$

- Where $\bar{x}$ & $\bar{y}$ are sample means for x & y, respectively.
- This equation makes some intuitive sense, as $\hat{\beta}_0$ is the y-intercept of a line passing through the average x- and y-values of our data. Hence, regression line is guaranteed to go through "the middle" of our data points.



28