DS 3010

Spring 2026

Wenting Xu

Iowa State University

Module 1: Introduction to Multiple Linear Regression

Wenting Xu

Iowa State University

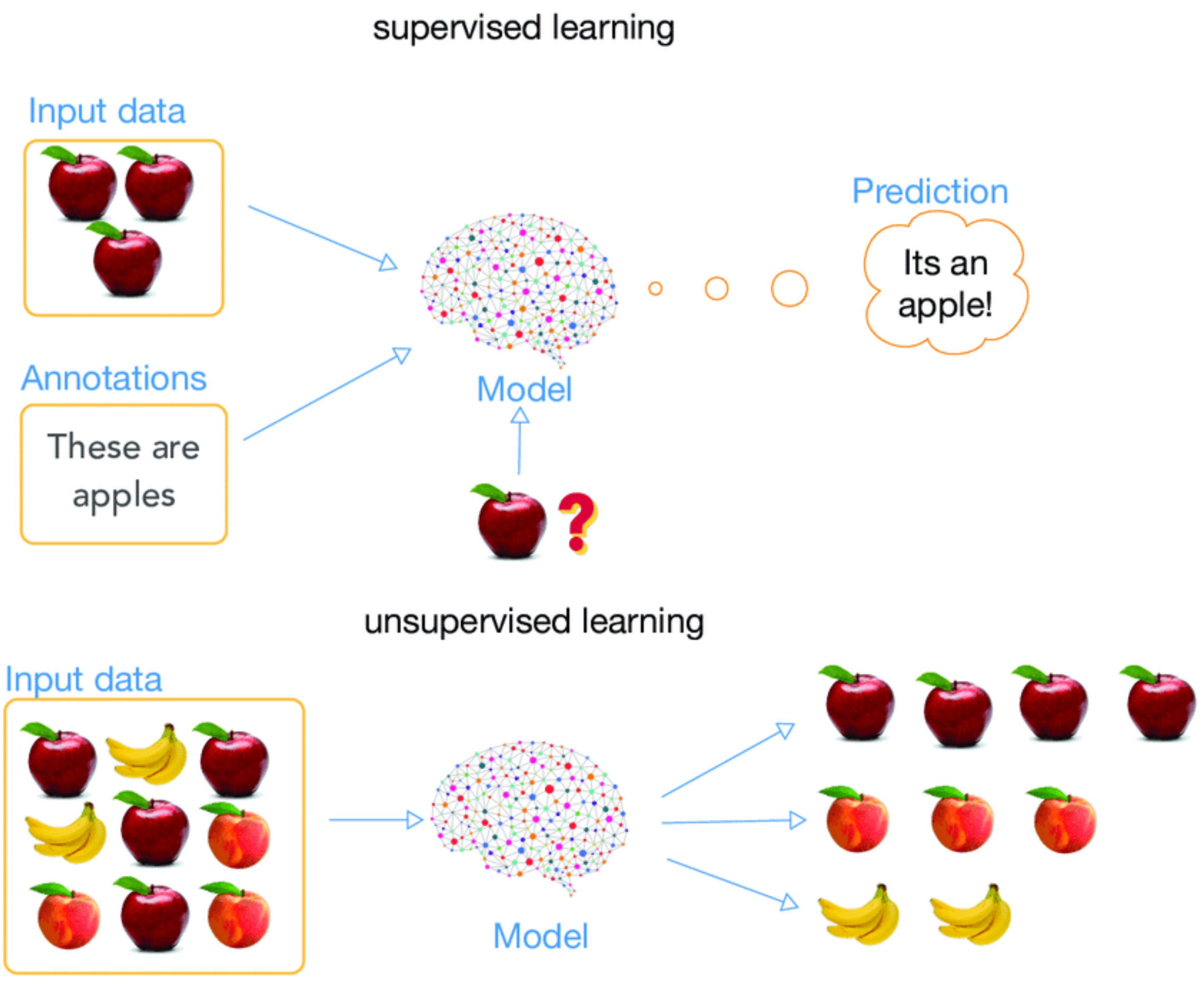
Copyright & Attribution

Some lecture slides and instructional materials in this course are adapted from the following sources:

•*An Introduction to Statistical Learning: With Applications in R (Second Edition)*
Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani
Springer, 2021

•Online course materials developed by Trevor Hastie, Robert Tibshirani, and collaborators.

Supervised learning vs Unsupervised learning



[1] https://devopedia.org/supervised-vs-unsupervised-learning

Most statistical learning problems fall broadly into one of two categories:

1. Supervised learning → data is labeled
2. Unsupervised learning → no labels (just predictors)

Supervised learning

This is the setting where you have **labelled** data:

$$(Y, X_1, X_2, \ldots, X_p)$$

what we want to predict

- Y is our response (outcome of interest), X's are our predictors.
- We sometimes refer to X as our input and Y as our output.
- Usually, we are interested in learning the relationship between a set of **inputs** (X's) and **output** (Y).
- Majority of machine learning problems/techniques fall into this category.

Y is continuous Y is categorical

- We refer to this setting as prediction or classification.

Unsupervised learning

This is the setting where you only have **unlabelled** data:

$$(X_1, X_2, \ldots, X_p)$$

- We no longer have an associated response Y.
- Prediction and classification models are no longer appropriate here.
- In some sense, we are working blind: we are *unsupervised* because we lack a response variable Y that can supervise our analysis.
- This setting is considered much more challenging.
- Applications that fall under unsupervised learning?

Group users w/ similar behaviors

As the course title suggests...

This class will largely focus on supervised learning.

Examples of supervised learning techniques you may already know:

↳ Regression $\begin{cases} \text{Linear regression (prediction)} \rightarrow \text{cont. var} \\ \text{Logistic regression (classification)} \rightarrow \text{cat. var} \end{cases}$

↳ KNN (nearest neighbors)

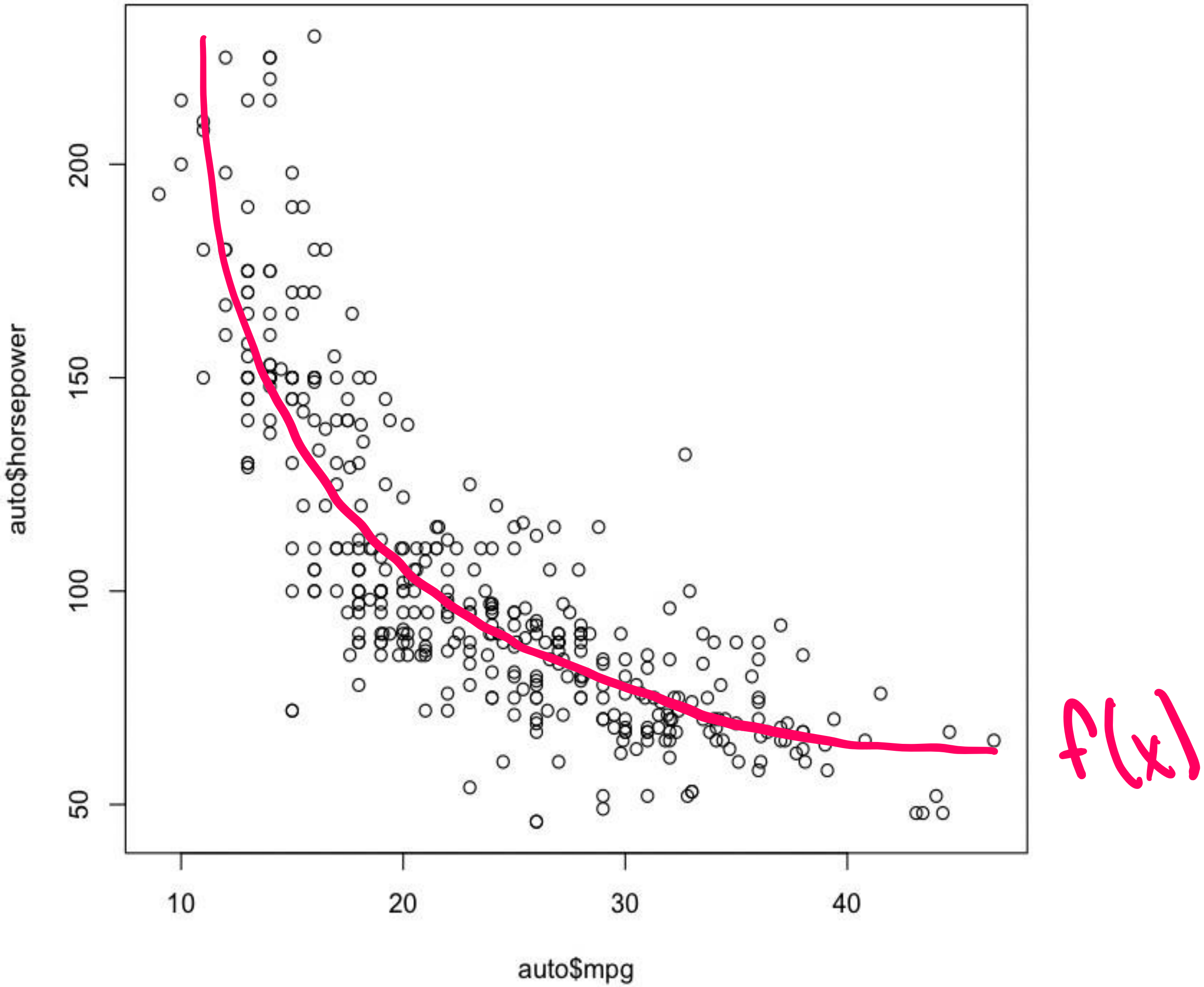
↳ Decision tree

Supervised learning setup

$$Y = f(X) + \epsilon$$

- The function f captures the systematic relationship between X and Y.
- f is fixed and unknown.
- ϵ represents — ? random error
- Our goal: to estimate (learn) the function f, using a dataset. This will allow us **model** the relationship between X and Y.

Example



How do we estimate $f(X)$? Ideas?

Multiple Linear Regression

Supervised Learning

- Linear regression is a key building block of predictive modeling and an important tool to have in your tool kit.

- Multiple linear regression (more than one predictor). $X_1, X_2, \ldots, X_p$

- Simple linear regression (only one predictor). X_1

Multiple Linear Regression

Motivation: $Y = \beta_0 + \beta_1 \cdot X_1 + \varepsilon$

1. Can provide an exact and interpretable description of the relationship between Y and X.
2. Widely used. Simple.
3. In terms of prediction, can often outperform more complicated models.
4. Inference is well-studied in this setting.
5. The fundamentals covered here are the building blocks for more complicated models.

Predict salary upon graduation

- Y = income upon graduation.
- X_1 = gpa.
- X_2 = number of internship hours.
- X_3 = major

Suppose I hand you a dataset with this information for 1,000 students who graduated college last year. My goal is to be able to predict a current student's future salary (Y) given their gpa, number of internship hours, and major.

How would we formulate this problem?

How might we use this data?

Multiple Linear Regression Preliminaries

Regression setup:

\# of observations

$$Y_i = f(X_i) + \epsilon_i, \qquad i = 1\ldots, n.$$

If we are willing to make a *key* assumption that the relationship between X and Y is *approximately* linear, then

$$f(X_i) = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots + \beta_p X_{ip}.$$

\# of predictors

Population regression line:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots + \beta_p X_{ip} + \epsilon_i, \qquad i = 1\ldots, n.$$

Assumptions

1. Relationship between Y and $X = (X_1, X_2, \ldots, X_p)$ is approximately linear.
2. $E(\epsilon) = 0$.
3. $\mathrm{Var}(\epsilon) = \sigma^2$.
4. ϵ's are uncorrelated.

MLR

Population regression line:

$$Y_i = \beta_0 + \beta_1 * X_{i1} + \beta_2 * X_{i2} + ... + \beta_p * X_{ip} + \epsilon_i,\ i = 1.....n$$

How to obtain data-driven estimates for $\hat{\beta}$ from our dataset?

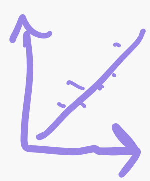
^ : est. value (for an unknown parameter)

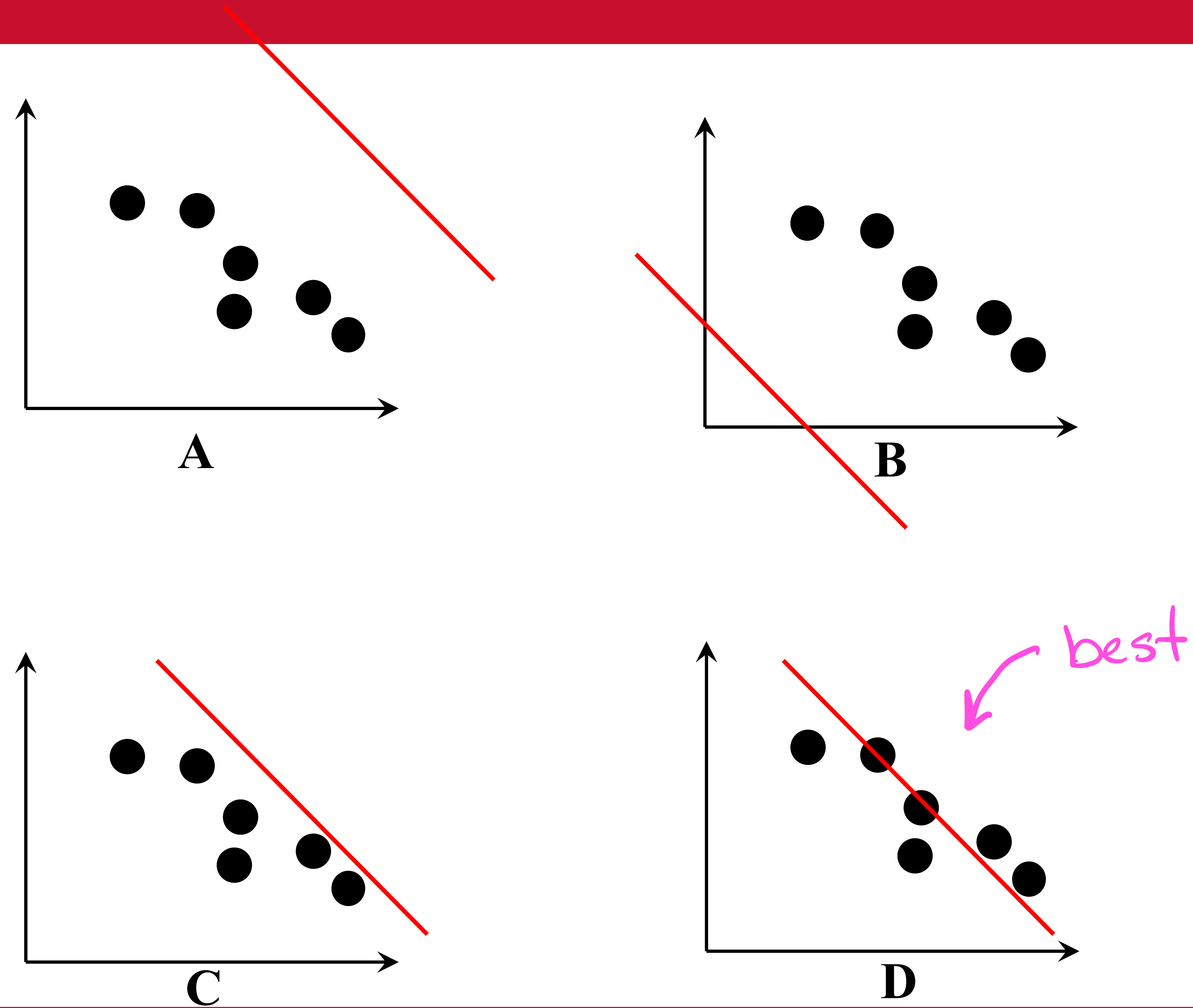
$\hat{B}_0 \to B_0$

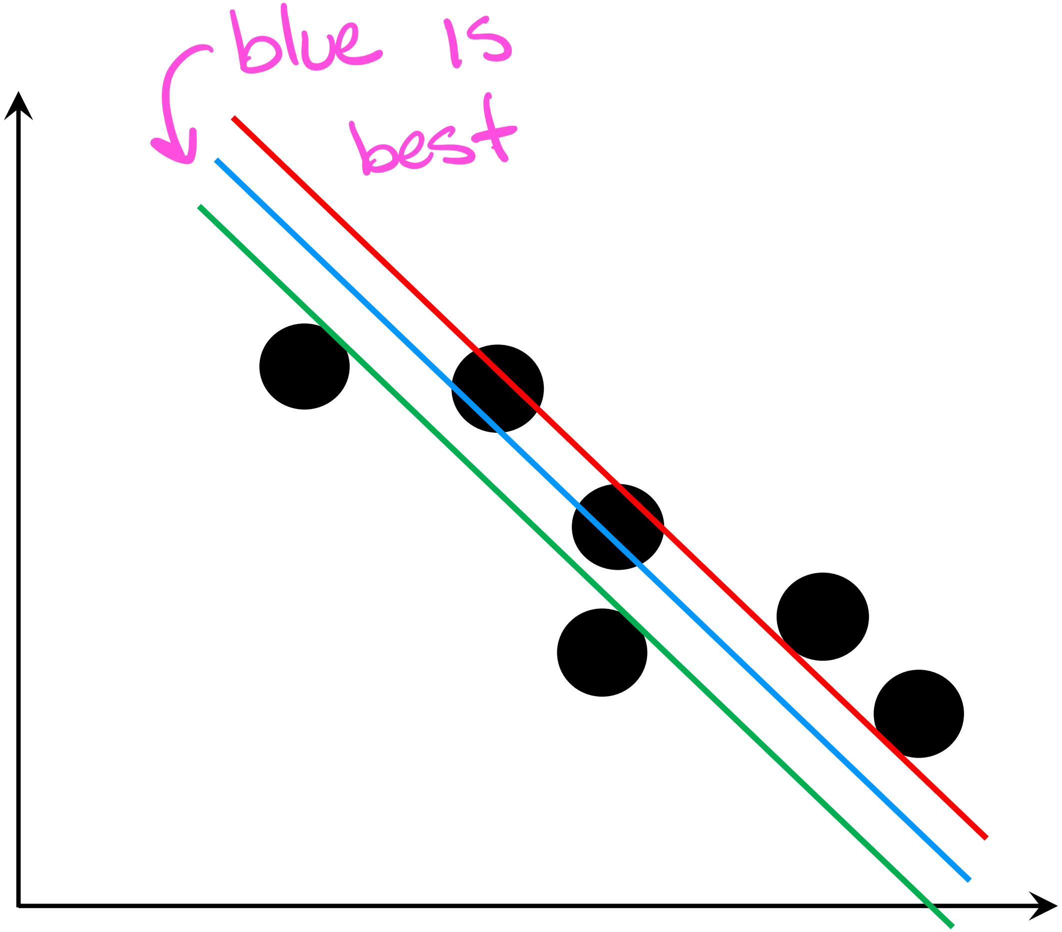
unknown

$Y = B_0 + B_1 \cdot X_1 + \varepsilon$

$\hat{Y} = \hat{B}_0 + \hat{B}_1 \cdot X_1$




A
B
C
D
best



Which line is the best line ?

How to obtain estimates for β ?

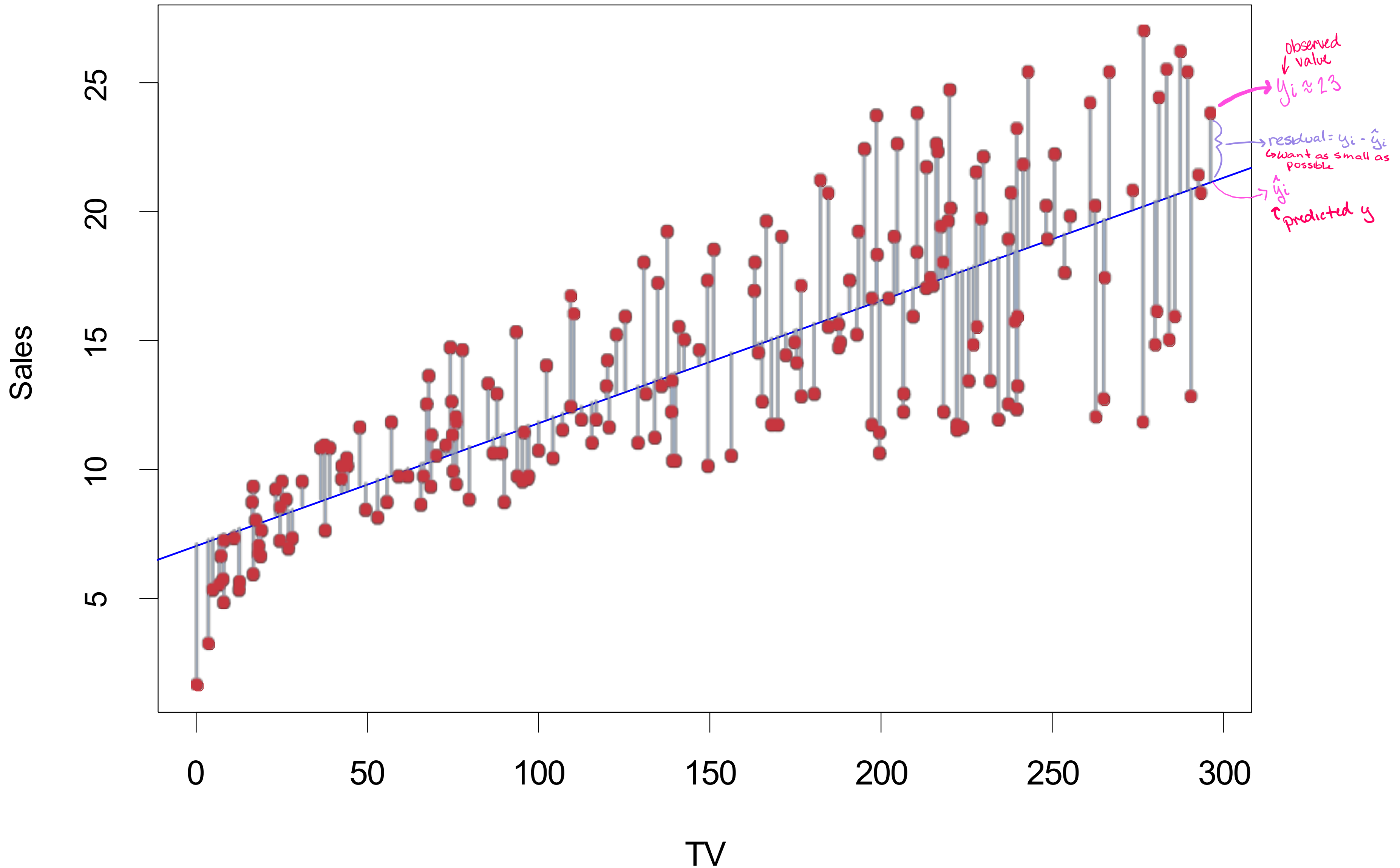
Let's start with the simple linear regression case (we only have 1 predictor X_1).

- Our goal is to find estimates for the coefficients $\hat{\beta}_0$ and $\hat{\beta}_1$.
- We have our data: $(y_i, x_i), i = 1, . . . , n.$
- We want to obtain coefficient estimates such that the linear model fits the available data well. In other words, we want:

$$y_i \approx \hat{\beta}_1 + \hat{\beta}_1 * x_i, i = 1,2, \ldots n$$

- We want to the line to be as **close** as possible to the data points. The problem boils down to: **how do we define closeness here**?

Least squares estimation

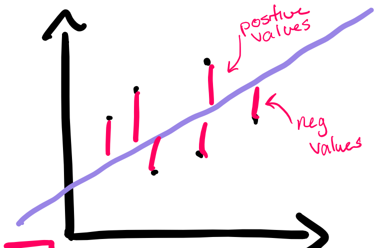


Details

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

$$= \sum_{i=1}^{n} \left[y_i - (\beta_0 + \beta_1 x_1) \right]$$

↳ residual sum of squares (RSS)



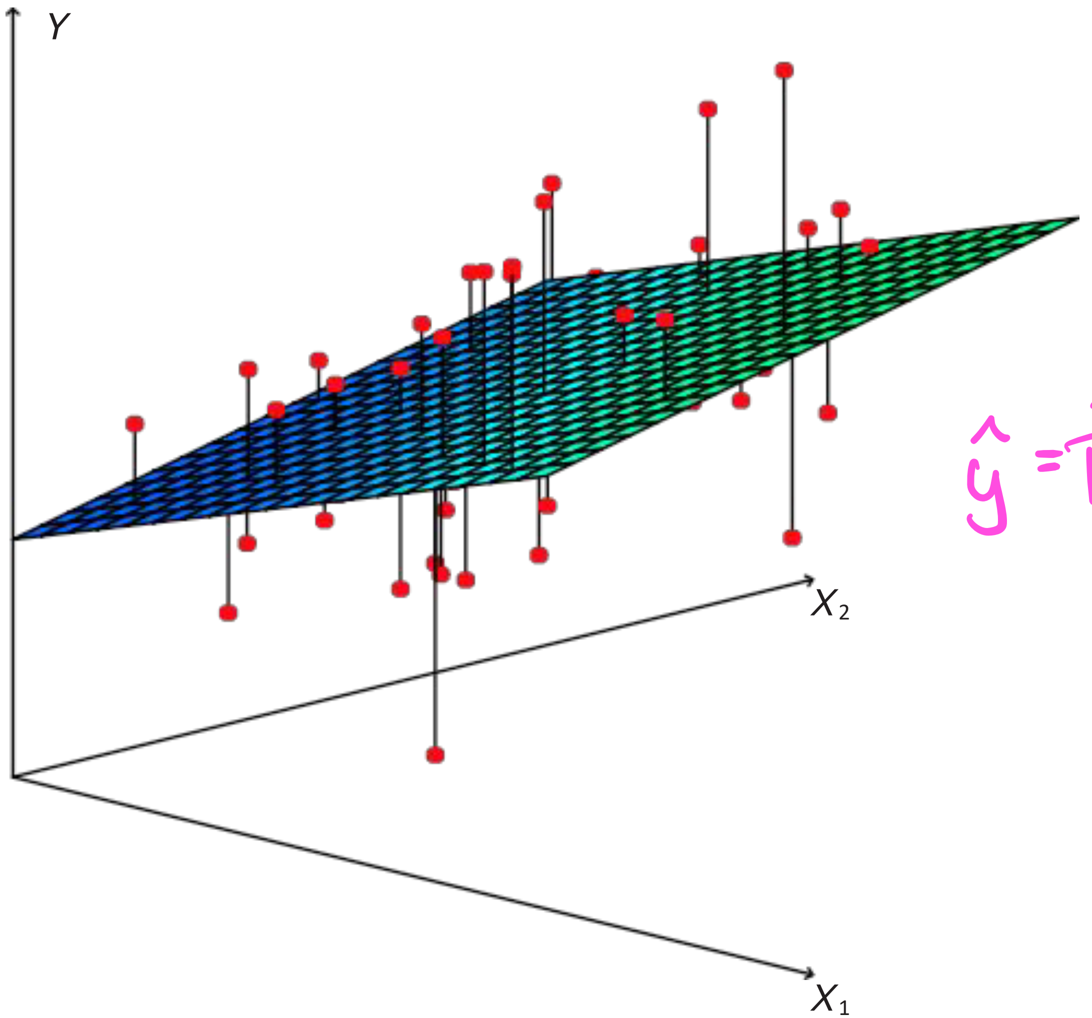
$$\sum_{i=1}^{n} (y_i - \hat{y}_i) \rightarrow \text{cancellation}$$

↳ want this to be as close to zero, so just taking the sum does <u>not</u> work (since there are both pos. + neg. values)

Least Squares = the best fitting line is the one that minimizes RSS

↳ the method chooses $\hat{\beta}_0, \hat{\beta}_1$ by minimizing RSS

Multiple linear regression



$\hat{y} = \hat{\beta}_0 + \beta_1 x_1 + \beta_2 x_2$

Extending to multiple linear regression

Least squares (minimize RSS)

$$\sum_{i=1}^{n}(y_i-\hat{y}_i)^2 = \sum_{i=1}^{n}\left[y_i-(\hat{\beta}_0+\hat{\beta}_1 X_1+\beta_2 X_2+\ldots+\beta_p X_p\right]^2$$

Interpretation of least squares coefficients

$\hat{\beta}_j$ can be interpreted as the average change in Y associated with a 1 unit change in X_j, *holding all other predictors constant*.

In a study investigating the relationship between the number of years of work experience (x) and the monthly salary earned (y) for employees in a particular industry, a linear regression analysis was conducted. The resulting regression equation is

$$\hat{y} = 800 \times \ x + 3000 \quad b_0$$

Interpret the meaning of b_0 and b_1 in this regression line

b_0:

A. The expected monthly salary for an employee with zero years of work experience.

B. The increase in monthly salary for each additional year of work experience.

C. The average monthly salary of all employees in the industry.

D. The maximum monthly salary an employee can earn in this industry.

In a study investigating the relationship between the number of years of work experience (x) and the monthly salary earned (y) for employees in a particular industry, a linear regression analysis was conducted. The resulting regression equation is

$$\hat{y} = \overset{b_1}{800} \times \ x + 3000$$

Interpret the meaning of b_0 and b_1 in this regression line

$b_{1:}$

A. The expected monthly salary for an employee with zero years of work experience.

B. The expected increase in monthly salary for each additional year of work experience.

C. The average monthly salary of employees in this industry.

D. The minimum possible monthly salary an employee can earn.

In a study investigating the relationship between the number of years of work experience (x) and the monthly salary earned (y) for employees in a particular industry, a linear regression analysis was conducted. The resulting regression equation is

$$\hat{y} = 800 \times x + 3000$$

Interpret the meaning of b_0 and b_1 in this regression line

$b_{1:}$

What is the predicted monthly salary for an employee with 3 years of experience ?

$\hat{y} = 3 \cdot 800 + 3000 = 2400 + 3000 = 5400$

$\hat{y} = 5400$

In a study investigating the relationship between the number of years of work experience (x_1), the number of internship hours(x_2) and the monthly salary earned (y) for employees in a particular industry, a linear regression analysis was conducted. The resulting regression equation is

$$\hat{y} = 800 \times\ x_1 + 100 \times\ x_2 + 3000$$

Interpret the meaning of b_1 (800) in this regression line

b_0

A. The expected monthly salary for an employee with zero years of work experience and zero internship hours.

b_1

B. The expected increase in monthly salary for each additional year of work experience, holding internship hours constant.

b_2

C. The average increase in monthly salary for each additional internship hour completed.

D. The total increase in monthly salary due to both work experience and internship hours.

$$\hat{y} = \underset{\beta_0}{\underset{\uparrow}{50}} + \underset{\beta_1}{\underset{\uparrow}{20}}(GPA) + \underset{\beta_2}{\underset{\uparrow}{0.07}}(IQ) + \underset{\beta_3}{\underset{\uparrow}{35}}(L)$$