

# **STA 35C: Statistical Data Science III**

## **Lecture 5: Simple Linear Regression**

Dogyoon Song

Spring 2026, UC Davis

# Agenda

---

## Statistical learning:

- *Definition:* A set of tools for understanding data and making informed predictions
- *Goal:* Estimate a function  $f : X \rightarrow Y$  that
  - (1) minimizes the reducible error  $\hat{f}(X) - f(X)$ , and
  - (2) is interpretable
- Methodologies typically fall under parametric or nonparametric approaches

**Today:** We begin exploring concrete methods. Specifically, we'll discuss:

- Categorizing statistical learning problems
- (Linear) Regression
- Simple linear regression

# Supervised vs. unsupervised learning

---

Most statistical learning problems fall into two categories: supervised or unsupervised

In **supervised learning**:

- Each predictor observation  $x_i$  is accompanied by a response  $y_i$
- “Supervised” because the responses guide (supervise) the analysis
- Many classical statistical learning methods operate in the supervised learning domain
  - *Example*: linear regression, logistic regression, support vector machine, etc.

In **unsupervised learning**:

- We have observations  $x_i$  but no response  $y_i$
- “Unsupervised” because there is no response to guide the analysis
- Often used to explore relationships among observations or variables
  - *Example*: Cluster analysis, dimension reduction, etc.

Sometimes, whether an analysis is supervised or unsupervised is less clear-cut

# Illustration: Supervised vs. unsupervised learning

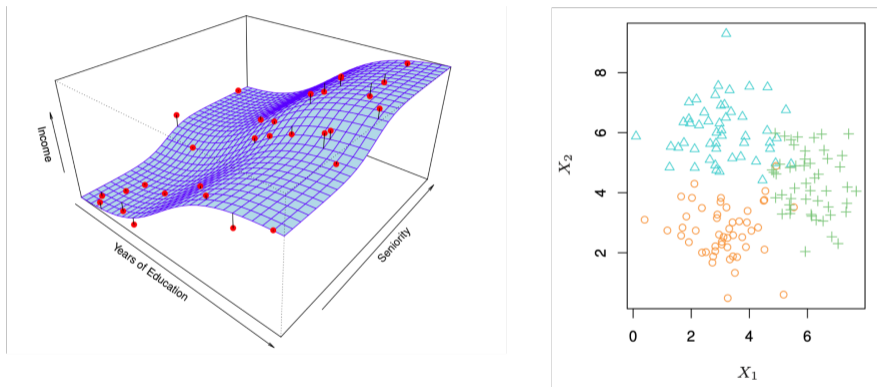


Figure: Supervised vs. unsupervised learning

For now, we focus on supervised learning

# Regression vs. classification

---

Variables can be quantitative or qualitative (categorical):

- Quantitative variables take numeric values
- Qualitative variables belong to one of  $K$  different *classes*

Depending on whether the **response** is quantitative or qualitative:

- Problems with a *quantitative* response are called **regression** problems
- Problems with a *qualitative* response are called **classification** problems

However, this distinction is not always crisp (e.g., linear vs. logistic regression)

Whether **predictors** are qualitative or quantitative is generally considered less important

# Illustration: Regression vs. classification

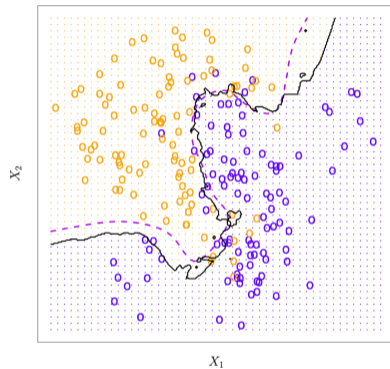
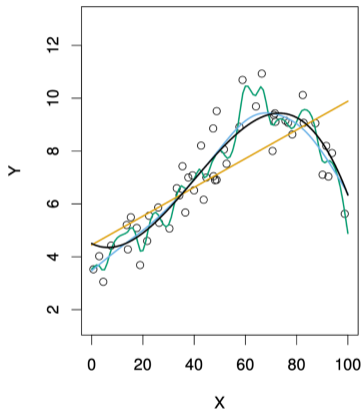


Figure: Regression vs. classification

## Regression: What do we want to do with this?

---

Regression problems are supervised learning problems with a *quantitative* response

Typical questions we want to address via regression include:

- Is there any relationship between  $X$  and  $Y$ ?
- How strong is it? (How much of  $Y$  is explained by  $X$ ?)
- How large is the association? (How does  $Y$  change per unit change in  $X$ ?)
- How accurately can we predict  $Y$  given  $X$ ?
- Is the relationship linear?

With multiple predictors, we can also ask (in future lectures):

- Which  $X$  are associated with  $Y$ ?
- Are there interactions among  $X$ ?

# Simple linear regression

---

Simple linear regression predicts  $Y$  from a single variable  $X$ , assuming an approximately linear relationship between  $X$  and  $Y$

Mathematically, we assume

$$Y = \beta_0 + \beta_1 X + \epsilon,$$

- *Model parameters*:  $\beta_0$  (intercept),  $\beta_1$  (slope) are fixed, unknown constants
- $\epsilon$  is an error term

We often say we *regress*  $Y$  on  $X$

Once we have *estimated*  $\hat{\beta}_0$  and  $\hat{\beta}_1$  from *training data*, we can predict

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

# Example

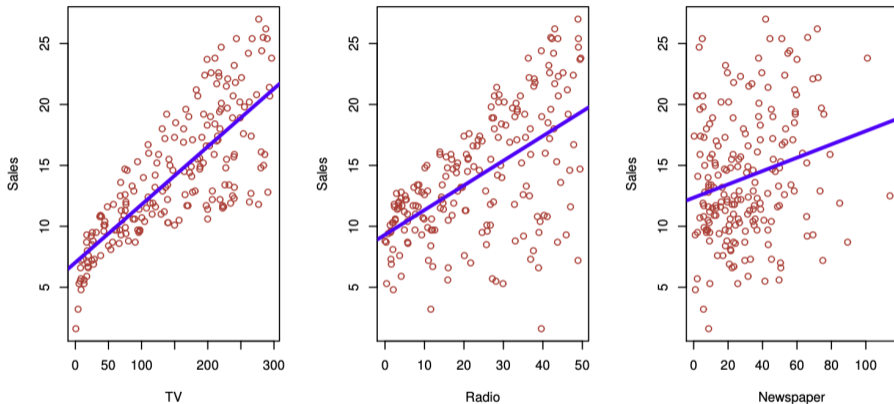


Figure: The Advertising data set shows Sales of a product in 200 different markets against advertising budgets for three media: TV, Radio, and Newspaper [JWHT21, Figure 2.1].

## Estimating the coefficients: Least squares

---

In practice,  $\beta_0$  and  $\beta_1$  are unknown and must be estimated from data

$$(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n).$$

We want the fitted line  $\hat{\beta}_0 + \hat{\beta}_1 x$  to be close to the true line  $\beta_0 + \beta_1 x$

The most common approach involves the **least squares** criterion:

- The Residual sum of squares (RSS) is defined as

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

- The *least squares* approach chooses  $\hat{\beta}_0$  and  $\hat{\beta}_1$  to minimize the RSS
- The solutions are

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2} \quad \text{and} \quad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

where  $\bar{y} := \frac{1}{n} \sum_{i=1}^n y_i$  and  $\bar{x} := \frac{1}{n} \sum_{i=1}^n x_i$

## Pop-up quiz: Least squares coefficients

---

**Dataset:** Three observations:

$$(x_1, y_1) = (1, 2), \quad (x_2, y_2) = (2, 3), \quad (x_3, y_3) = (3, 6).$$

**Question:** What are the estimated slope  $\hat{\beta}_1$  and intercept  $\hat{\beta}_0$ ?

*Hints (partial sums to speed up):*

$$\bar{x} = 2, \quad \bar{y} = \frac{2 + 3 + 6}{3} = 3.667, \quad \sum (x_i - \bar{x})^2 = 2, \quad \sum (x_i - \bar{x})(y_i - \bar{y}) = 4.$$

**Multiple-choice answers:**

- a)  $\hat{\beta}_1 = 2, \quad \hat{\beta}_0 = -0.333$
- b)  $\hat{\beta}_1 = 1.5, \quad \hat{\beta}_0 = 1$
- c)  $\hat{\beta}_1 = 2, \quad \hat{\beta}_0 = 0$
- d)  $\hat{\beta}_1 = 3, \quad \hat{\beta}_0 = -2$

# Properties of the least squares estimator

$(\hat{\beta}_0, \hat{\beta}_1)$  estimate  $(\beta_0, \beta_1)$  using data, so they need not be the same

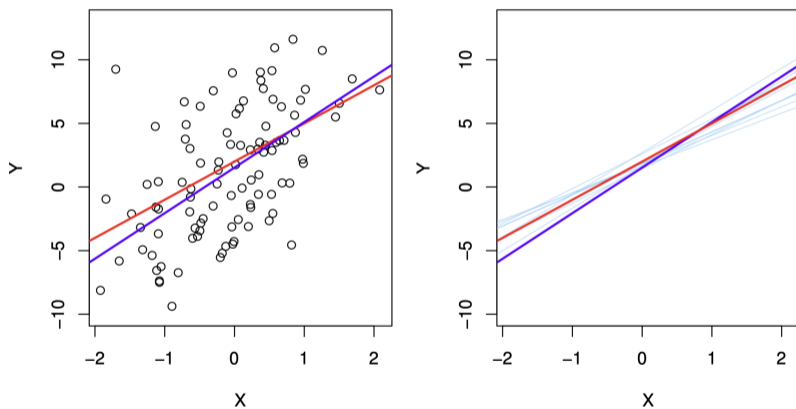


Figure: Least squares coefficient estimates from 100 data points. Red: population regression line, Blue: least squares line, Light blue: ten separate least squares lines [JWHT21, Figure 3.3].

## Properties of the least squares estimator (cont'd)

---

**Unbiasedness:**  $\hat{\beta}_0$  and  $\hat{\beta}_1$  are *unbiased* estimators of  $\beta_0$  and  $\beta_1$

- $\mathbb{E}[\hat{\beta}_0] = \beta_0$  and  $\mathbb{E}[\hat{\beta}_1] = \beta_1$
- If we *repeat the least squares regression using new samples* from the same population, then their average converges to the population regression line

**Sampling distribution:** Nevertheless, we only have one dataset!

- We care about how far  $\hat{\beta}_0, \hat{\beta}_1$  can deviate from their expected values
- Under certain assumptions<sup>1</sup>, the standard errors of  $\hat{\beta}_0$  and  $\hat{\beta}_1$  are:

$$\text{SE}(\hat{\beta}_0)^2 = \sigma^2 \left[ \frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^n (x_i - \bar{x})^2} \right] \quad \text{and} \quad \text{SE}(\hat{\beta}_1)^2 = \frac{\sigma^2}{\sum_{i=1}^n (x_i - \bar{x})^2}$$

where  $\sigma^2 = \text{Var}(\epsilon)$ , which is typically unknown and has to be estimated

---

<sup>1</sup>For these to be strictly valid, we must assume  $\epsilon_i$  for all  $i$  have variance  $\sigma^2$  and are uncorrelated

## Inference about the model parameters

---

Usually, we estimate  $\sigma^2 = \text{Var}(\epsilon)$  using the *residual standard error* (RSE):

$$\hat{\sigma} = \text{RSE} = \sqrt{\frac{\text{RSS}}{n-2}}$$

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

**Confidence Intervals:** Standard errors can be used to compute confidence intervals

- A 95% confidence interval of  $\beta_i$  is approximately  $\hat{\beta}_i \pm 1.96 \cdot \text{SE}(\hat{\beta}_i)$
- There is approximately a 95% chance that the (random) interval

$$\left[ \hat{\beta}_i - 1.96 \cdot \text{SE}(\hat{\beta}_i), \hat{\beta}_i + 1.96 \cdot \text{SE}(\hat{\beta}_i) \right]$$

contains the true value of  $\beta_i$

## Inference about the model parameters (cont'd)

---

**Hypothesis Testing:** Standard errors can also be used for hypothesis tests on  $\beta_0$  or  $\beta_1$

- A common test:

$$H_0 : \beta_1 = 0 \quad (\text{no relationship}) \quad \text{vs.} \quad H_1 : \beta_1 \neq 0 \quad (\text{some relationship}).$$

- Compute a *t*-statistic and compare to a *t*-distribution w/  $(n - 2)$  degrees of freedom:

$$t = \frac{\hat{\beta}_1 - 0}{\text{SE}(\hat{\beta}_1)}$$

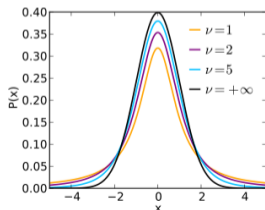


Figure: *t*-distribution (image from [Wikipedia](#)).

## Pop-up quiz: Confidence interval / hypothesis test

---

Let's use the *same dataset* as before:

$$(1, 2), (2, 3), (3, 6),$$

We found  $\hat{\beta}_1 = 2$  and  $\hat{\beta}_0 = -0.333$ ; also observe that  $\sum (x_i - \bar{x})^2 = 2$

1) **Residual standard error:** We have  $n = 3$ , so  $n - 2 = 1$ . The RSS is

$$\text{RSS} = \sum_{i=1}^3 (y_i - \hat{y}_i)^2 = 0.666 \dots \quad (\approx \frac{2}{3}) \quad \implies \quad \hat{\sigma} = \sqrt{\frac{\text{RSS}}{n-2}} = ?$$

2) **Standard error of  $\hat{\beta}_1$ :**  $\text{SE}(\hat{\beta}_1) = \sqrt{\frac{\hat{\sigma}^2}{\sum (x_i - \bar{x})^2}} = ?$

3) **Hypothesis test:**  $H_0 : \beta_1 = 0$  vs.  $H_1 : \beta_1 \neq 0$ . What is the  $t$ -statistic  $t = \frac{\hat{\beta}_1 - 0}{\text{SE}(\hat{\beta}_1)} = ?$

**Multiple-choice answers:**

A)  $\hat{\sigma} \approx 0.82$ ,  $\text{SE}(\hat{\beta}_1) \approx 0.58$ ,  $t \approx 3.46$ .

B)  $\hat{\sigma} \approx 1.0$ ,  $\text{SE}(\hat{\beta}_1) \approx 0.5$ ,  $t \approx 2.0$ .

C)  $\hat{\sigma} \approx 0.58$ ,  $\text{SE}(\hat{\beta}_1) \approx 0.82$ ,  $t \approx 0.71$ .

D)  $\hat{\sigma} \approx 0.50$ ,  $\text{SE}(\hat{\beta}_1) \approx 0.25$ ,  $t \approx 8.0$ .

## Assessing the accuracy of the model

---

The quality of a linear fit is typically assessed using RSE or the  $R^2$  statistic

- **Residual standard error (RSE):** “average deviation of  $Y$  from the regression line”

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

which is an estimate of the standard deviation of  $\epsilon$

- **The  $R^2$ :** “the proportion of variance in  $Y$  explained by  $X$ ”

$$R^2 = \frac{\text{TSS} - \text{RSS}}{\text{TSS}} = 1 - \frac{\text{RSS}}{\text{TSS}}$$

where  $\text{TSS} = \sum_{i=1}^n (y_i - \bar{y})^2$  is the total sum of squares (TSS)

## The $R^2$ statistic

---

The  $R^2$  ranges from 0 to 1, and is independent of the scale of  $Y$

- $R^2$  near 1 indicates most variability in  $Y$  is explained by the linear regression model
- $R^2$  near 0 indicates little variability is explained
  - e.g., the linear model assumption is wrong or the error variance  $\sigma^2$  is large

Recall the (sample) correlation:

$$\widehat{\text{Cor}}(X, Y) = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2} \sqrt{\sum_{i=1}^n (y_i - \bar{y})^2}}$$

- In simple linear regression,  $R^2 = r^2$ , where  $r = \text{Cor}(X, Y)$
- This equality does *not* hold in multiple linear regression

## Wrap-up

---

Simple linear regression assumes a model:

$$Y = \beta_0 + \beta_1 X + \epsilon,$$

with the model parameters typically estimated by least squares

We can address the five key questions:

- Is there any relationship between  $X$  and  $Y$ ?  $\Rightarrow$  Test  $H_0 : \beta_1 = 0$
- How much of  $Y$  is explained by  $X$ ?  $\Rightarrow R^2$
- How does  $Y$  change per unit change in  $X$ ?  $\Rightarrow \hat{\beta}_1$
- How accurately can we predict  $Y$  given  $X$ ?  $\Rightarrow \text{Var}(Y|X)$ , related to  $R^2$
- Is the relationship linear?  $\Rightarrow \hat{\beta}_0, \hat{\beta}_1, R^2$

**Next lecture:** multiple linear regression

# References

---



Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani.

*An Introduction to Statistical Learning: with Applications in R*, volume 112 of *Springer Texts in Statistics*.

Springer, New York, NY, 2nd edition, 2021.