STA 35C: Statistical Data Science III

Lecture 4: Simple Linear Regression

Dogyoon Song

Spring 2025, UC Davis

Agenda

Statistical learning:

- *Definition:* A set of tools for understanding data and making informed predictions
- *Goal:* Estimate a function $f : X \to Y$ that
  (1) minimizes the reducible error $\hat{f}(X) - f(X)$, and
  (2) is interpretable
- Various methodologies, largely categorized into parametric vs. nonparametric

Today, we will begin to learn some concrete methods

Specifically, let's discuss:

- Categorization of 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**:

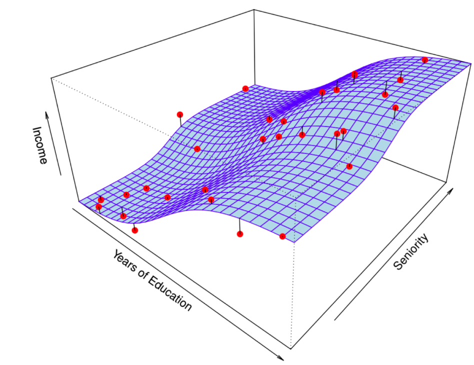
- Each predictor observation x_i is accompanied by a response y_i
- "Supervised" because the responses guide (supervise) the analysis
- Many classical statistical learning methods operate in the supervised learning domain
  - *Example:* linear regression, logistic regression, support vector machine,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



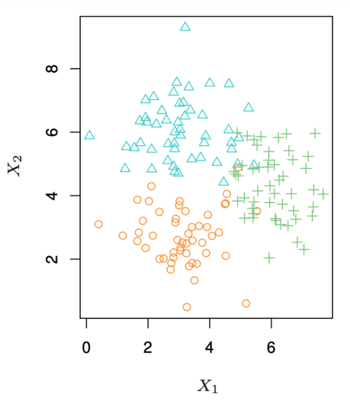


Figure: Supervised vs. unsupervised learning

Regression vs. classification

Variables can be quantitative or qualitative (categorical):

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

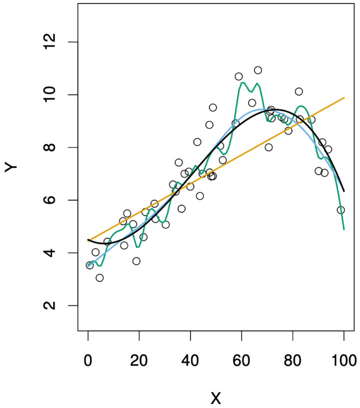
Depending on whether the **response** is quantitative or qualitative:

- 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



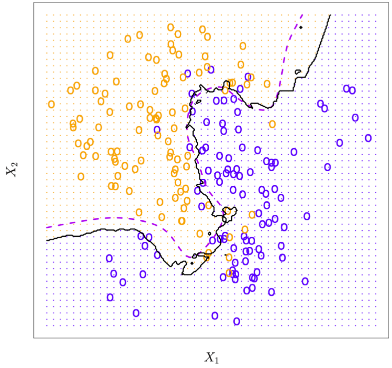


Figure: Regression vs. classification

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 regerssion 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 additionally ask:

- 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*: β_0 (intercept), β_1 (slope) are fixed, unknown constants
- ϵ 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

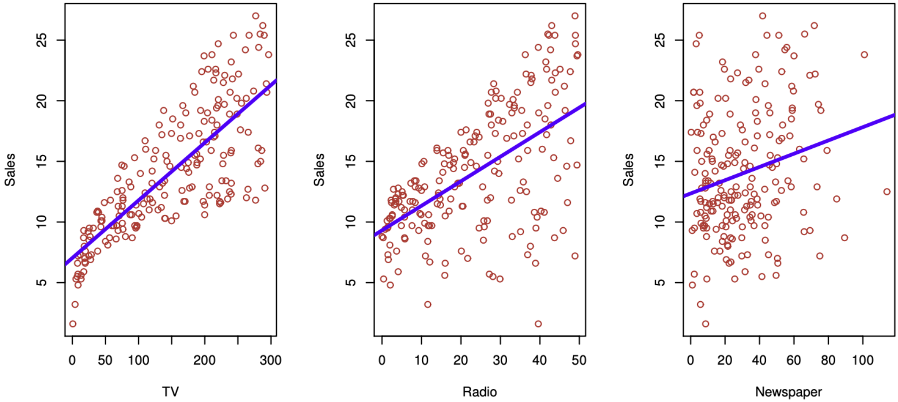


Figure: The Advertising data set shows Sales of a product in 200 different markets against advertising budgets for three media: TV, Radio, and Newspaper [JWHT21, Figure 2.1].

Estimating the coefficients: Least squares

In practice, β_0 and β_1 are unknown and must be estimated from data

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

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

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

- 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 $\bar{x} := \frac{1}{n} \sum_{i=1}^{n} x_i$

Properties of the least squares estimator

$(\hat{\beta}_0, \hat{\beta}_1)$ estimate (β_0, β_1) using data, so they need not be the same

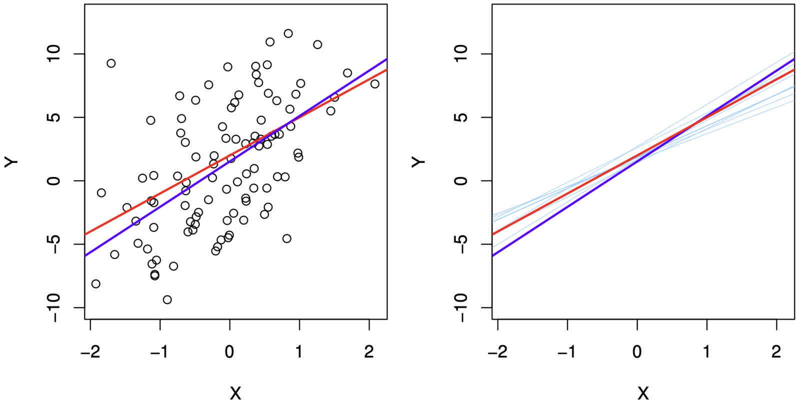


Figure: Least squares coefficient estimates from 100 data points. Red: population regression line, Blue: least squares line, Light blue: ten separate least squares lines [JWHT21, Figure 3.3].

Properties of the least squares estimator (cont'd)

$\hat{\beta}_0$ and $\hat{\beta}_1$ are *unbiased* estimators of β_0 and β_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, then their average converges to the population regression line

Nevertheless, we only have one dataset!

- We care about how far estimates can deviate from the expected value in average
- The standard errors of $\hat{\beta}_0$ and $\hat{\beta}_1$ can be computed[1] using:

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

where $\sigma^2 = \mathrm{Var}(\epsilon)$

[1] For these to be strictly valid, we must assume ϵ_i for all i have variance σ^2 and are uncorrelated

Inference about the model parameters

Usually, $\sigma^2 = \mathrm{Var}(\epsilon)$ is unknown, but can be estimated using the *residual standard error*:

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

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

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

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

  will contain the true value of β_i

Inference about the model parameters (cont'd)

Hypothesis Testing: Standard errors can also be used for hypothesis testing on β_0, β_1

- We often test

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

- In practice, we compute a *t-statistic*:

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

Assessing the accuracy of the model

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

The **Residual standard error** (RSE) is an estimate of the standard deviation of ϵ

The $\mathbf{R^2}$ represents the proportion of variance in Y explained by X:

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

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

The R^2 takes on a value between 0 and 1, and is independent of the scale of Y

- R^2 near 1 indicates most variability in Y is explained by the regression
- R^2 near 0 indicates little variability is explained
    - This can happen when the linear model is wrong or the error variance σ^2 is high

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.