Introduction to Statistical Learning and Application CC1: Introduction to Statistical Learning and Linear Regression

Razan MHANNA
STATIFY team, Inria centre at the University Grenoble Alpes
LEMASSON/CHRISTEN team, Grenoble Institute of Neurosciences GIN

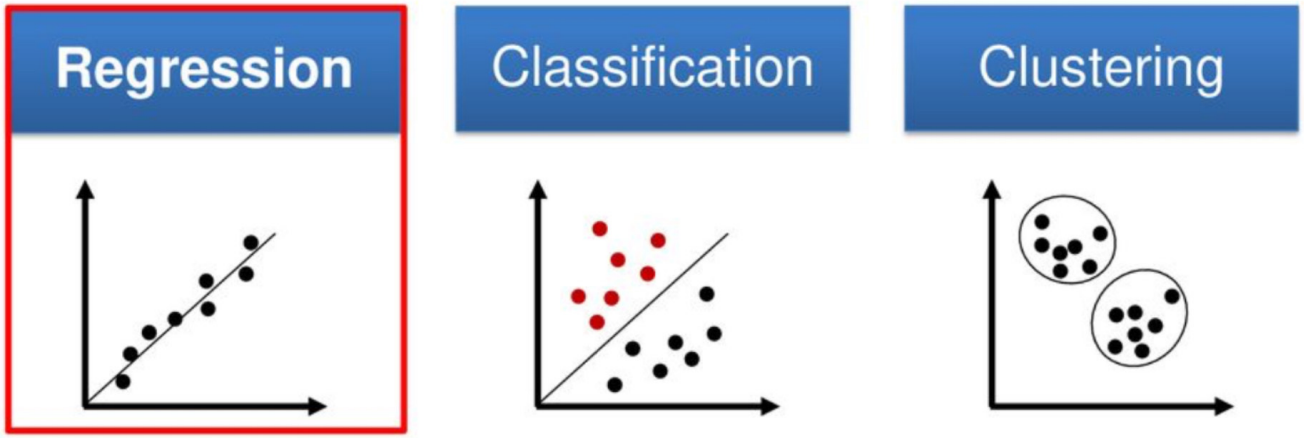
4 February 2025

Contents

Current Section

Course Information

- This is the complementary course of "Introduction to Statistical Learning and Applications" given to students from ENSIMAG and UGA by Professor Pedro Rodrigues.
- The classes will be given on Tuesdays from 11h30 to 13h at IM2AG, room F118.
- The materials used will be made available in this page https://github.com/ISLA-Grenoble/2025-complementary.



Course Content

- **Introduction to Statistical Learning**
- **Linear Regression**
  - Simple Linear Regression
  - Multiple Linear Regression
  - Extensions of the Linear Model
- **Classification**
  - Logistic Regression
  - Generative Models for Classification
- **Resampling Methods**
  - Cross-Validation
- **Dimension Reduction Methods**
  - Principal Component Analysis (PCA)
  - Partial Least Squares (PLS)

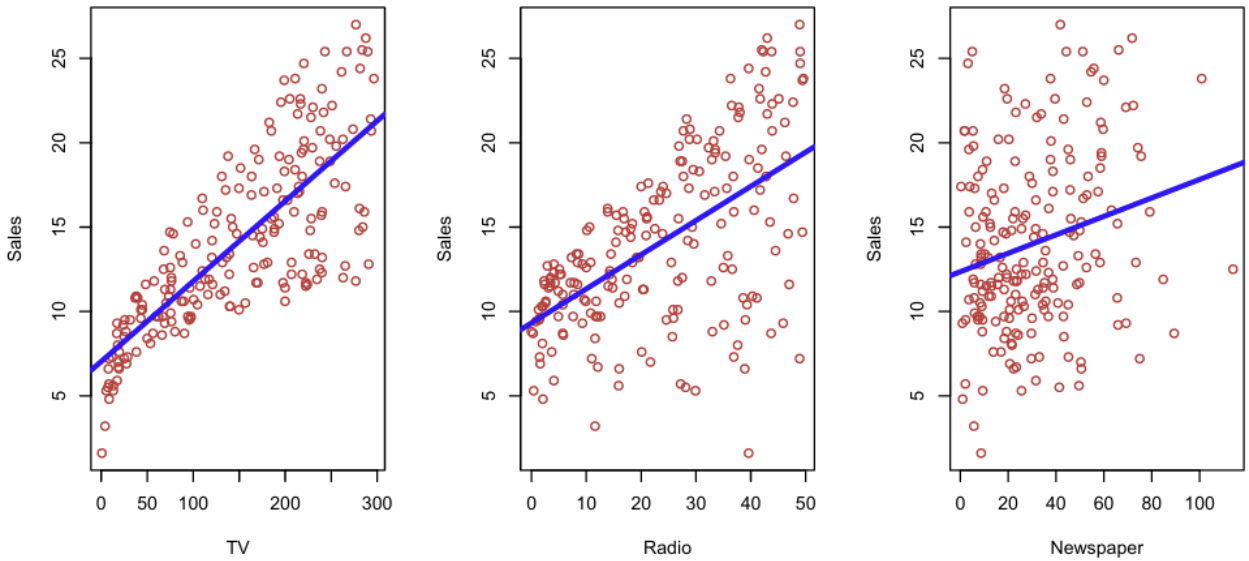
Course Content

- **Tree-Based Methods**
  - Regression Trees
  - Classification Trees
  - Bagging
  - Boosting
  - Random Forest
- **Support Vector Machines (SVM)**
- **Deep Learning**
  - Convolutional Neural Networks (CNN)

Current Section

Motivation for Statistical Learning

Suppose that we are statistical consultants hired by a client to investigate the association between advertising and sales of a particular product. The Advertising data set consists of the sales of that product in 200 different markets, along with advertising budgets for the product in each of those markets for three different media: TV, radio, and newspaper.



Statistical Learning Setup

- Goal: Understand the relationship between advertising and sales.
- Data: Sales in 200 markets, with advertising budgets for TV, radio, and newspaper.
- Objective: Develop an accurate model that can be used to predict sales on the basis of the three media budgets.
- Inputs(Predictors): Advertising budgets (TV, radio, newspaper), denoted as X_1, X_2, X_3.
- Quantative Response: Sales (denoted as Y).

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

where is some fixed but unknown function of X1, ..., Xp, and ε is a random error term, which is independent of X and has mean zero. In essence, statistical learning refers to a set of approaches for estimating f.

Reasons for Estimating f

Two main reasons to estimate f: **Prediction and Inference.**

Prediction:

- Inputs X are readily available, but output Y is difficult to obtain.
- We estimate f to predict Y using $\hat{Y} = \hat{f}(X)$.
- The accuracy of $\hat{(}Y)$ as a prediction for Y depends on two quantities: the reducible error and the irreducible error.
- Why is the irreducible error larger than zero? Due to unmeasurable factors that are useful in predicting Y, or simply the quantity ε may contain unmeasurable variations.

$$E(Y - \hat{Y})^2 = [f(X) - \hat{f}(X)]^2 + \text{Var}(\varepsilon)$$

- The focus of this course is on techniques for estimating f with the aim of minimizing the reducible error.
- The irreducible error sets an upper bound on prediction accuracy. This bound is unkown in pratice.

Inference

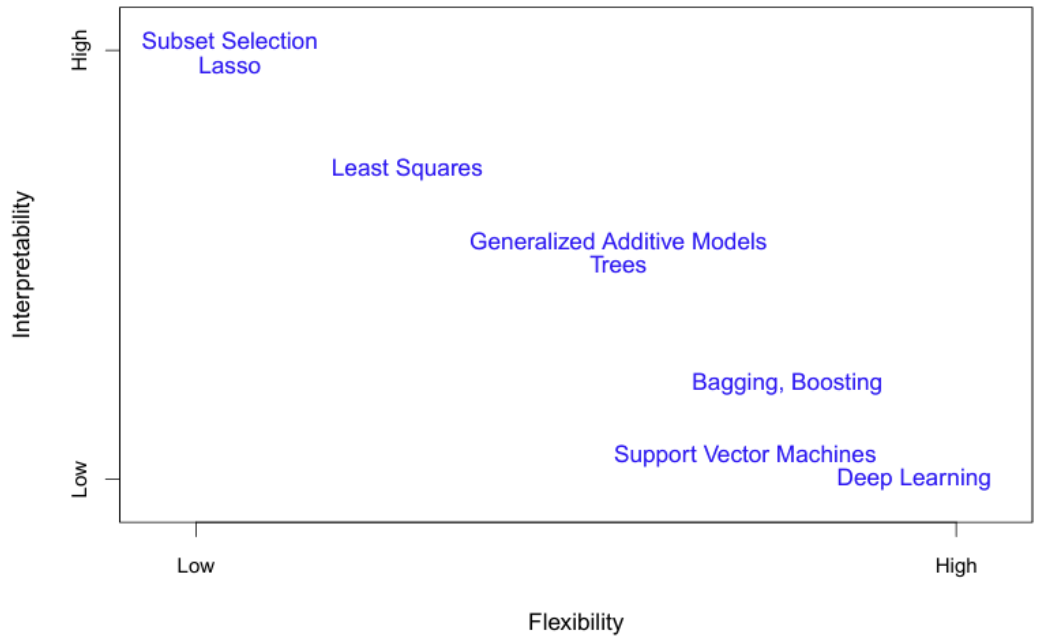
Inference: Understand the association between Y and $X_1, X_2, \ldots, X_p$. Unlike prediction, fcan't be trated as a blackbox; we need the exact form of $\hat{f}$.

- Key questions:
  1. Which predictors are associated with the response?
  2. What is the relationship between Y and each predictor? Some predictors may have a positive relationship with Y. Other predictors may have the opposite relationship.
  3. Can the relationship be summarized with a linear model, or is it more complex?

Depending on whether our ultimate goal is prediction, inference, or a combination of the two, different methods for estimating f may be appropriate. For example:

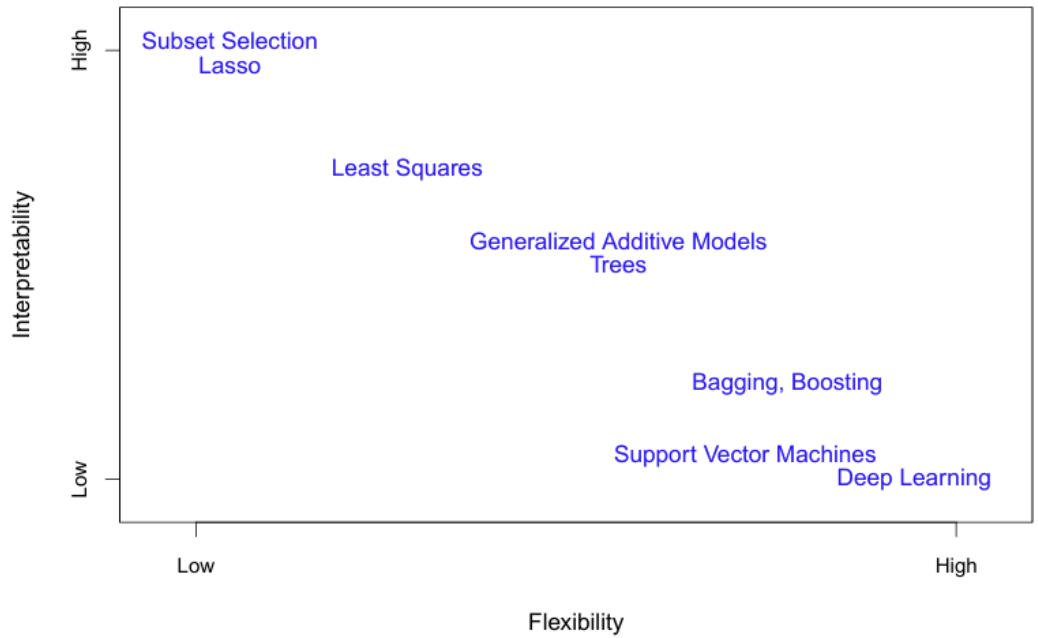
- *Linear Models: Easy to interpret but may lack prediction accuracy.*
- *Non-linear Models: More accurate predictions but harder to interpret.*

The Trade-Off Between Prediction Accuracy and Model Interpretability



Why Use Restrictive Models? When inference is the goal, restrictive models are more interpretable. For instance, the linear model may be a good choice for inference since it will be quite easy to understand the relationship between Y and X1,X2,...,Xp. In contrast, very flexible approaches, such as the splines and boosting.

The Trade-Off Between Prediction Accuracy and Model Interpretability



Why Use Restrictive Models? When inference is the goal, restrictive models are more interpretable. For instance, the linear model may be a good choice for inference since it will be quite easy to understand the relationship between Y and X1,X2,...,Xp. In contrast, very flexible approaches, such as the splines and boosting.

The Trade-Off Between Prediction Accuracy and Model Interpretability

- Least Squares Regression: A simple and interpretable method that fits a linear relationship between input and output but may lack flexibility for complex patterns.
- Lasso: A linear regression variant that adds sparsity by shrinking some coefficients to zero, improving interpretability while maintaining predictive power.
- Generalized Additive Models (GAMs): Extend linear models by allowing non-linear relationships, increasing flexibility but making interpretation harder.
- Non-Linear Methods: Techniques like bagging, boosting, support vector machines, and neural networks offer high predictive accuracy but at the cost of interpretability and complex inference.

How Do We Estimate f?

Given n data points, the training data is used to estimate f by finding a function $\hat{f}$ such that $Y \approx \hat{f}(X)$. These methods are classified into parametric and non-parametric approaches.

Parametric methods: assume a specific functional form for f, simplifying the estimation process, then use training data to estimate the model's parameters:

- Assume a functional form, e.g., linear model

$$f(X) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$$

Although efficient, parametric methods may suffer if the true function f deviates significantly from the assumed model, leading to poor estimates.

How Do We Estimate f?

Given n data points, the training data is used to estimate f by finding a function $\hat{f}$ such that $Y \approx \hat{f}(X)$. These methods are classified into parametric and non-parametric approaches.

Parametric methods: assume a specific functional form for f, simplifying the estimation process, then use training data to estimate the model's parameters:

- Assume a functional form, e.g., linear model

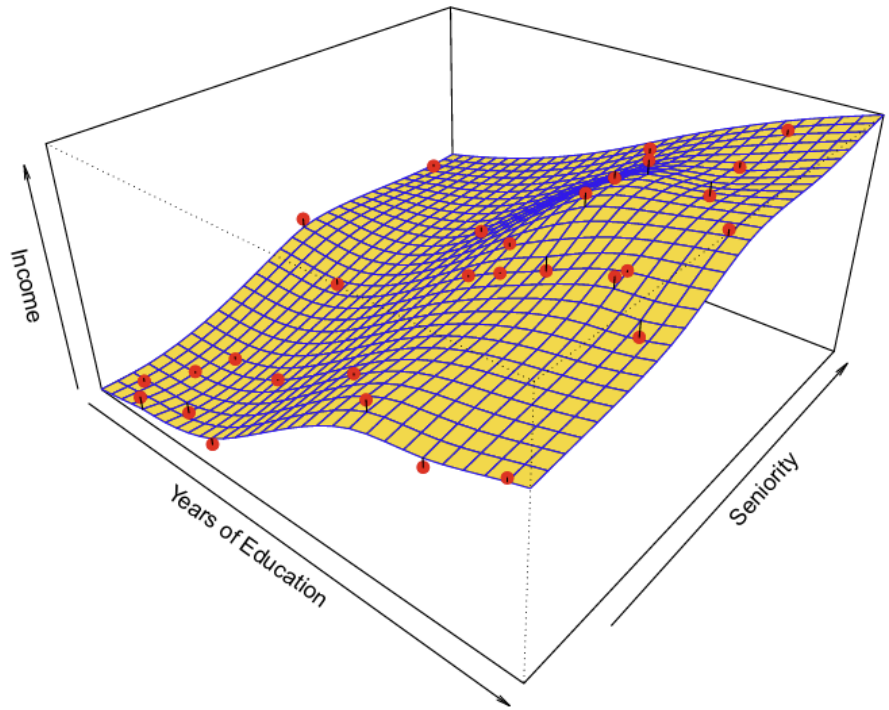
$$f(X) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$$

Although efficient, parametric methods may suffer if the true function f deviates significantly from the assumed model, leading to poor estimates.

Non-Parametric Methods

Non-parametric methods do not assume a functional form for f, aiming for a flexible fit:

- These methods can more accurately capture complex relationships but require a larger number of observations.
- Example: Thin-plate splines fit the data closely. Overfitting!



Supervised vs Unsupervised Learning

- Supervised learning involves predictor measurements $x_1, \ldots, x_n$ with associated response measurements y_i.
  - Linear regression, logistic regression, Generalized Additive Models (GAM), boosting, support vector machines.
- In unsupervised learning, only predictors $x_1, \ldots, x_n$ are available, without corresponding responses y_i.
  - Clustering analysisâgrouping similar observations based on predictor variables.
- Semi-supervised learning aims to incorporate both labeled and unlabeled data. Some observations have both predictors and responses, while others only have predictors.
  - Scenario arises when predictors are cheap to collect, but responses are expensive.

Current Section

What is Simple Linear Regression?

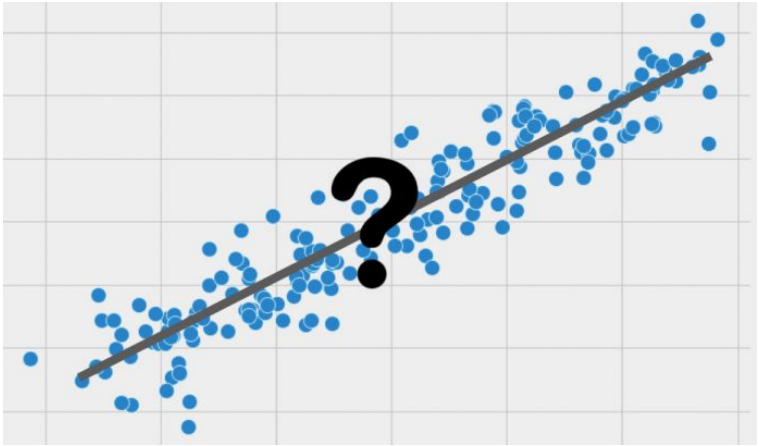
Definition

Simple Linear Regression is a statistical method used to model the relationship between two variables by fitting a straight line to the data.

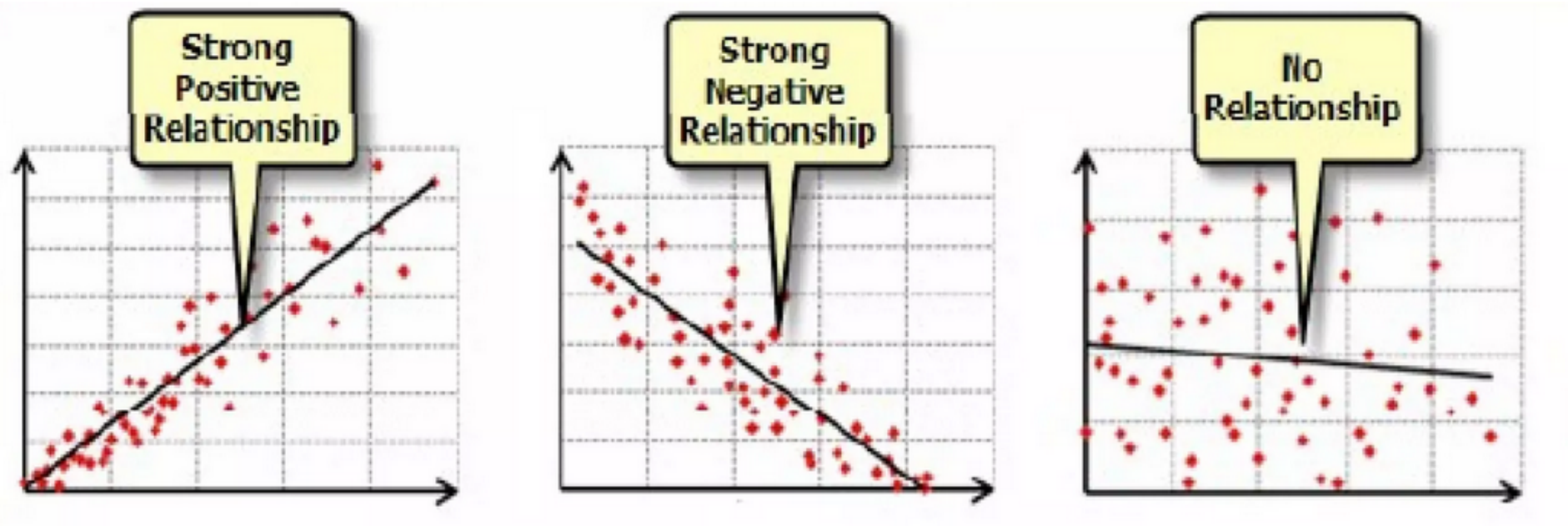
- **Regression** is a supervised learning technique that helps in finding the model function f between variables.
- It is used to fit the best straight line between a set of data points.
- After a graph is properly scaled, the data points must "look" like they fit a straight line, not a parabola or any other shape.
- The line is used as a model to predict a variable y from another variable x.
- Finding the **"best-fit"** line is the goal of simple linear regression.

What is Simple Linear Regression?

Input, predictive or Independent varible x: this the variable whose value that is believed to influence the value of another variable,
Output, Response or Dependent variable y: this is the varaible whose value that is believed to be influenced by the value of another variable,
Best-fit Line: represents our model. It is the line that best-fits our data points. The line represents the best estimate of the y value for every given input of x,

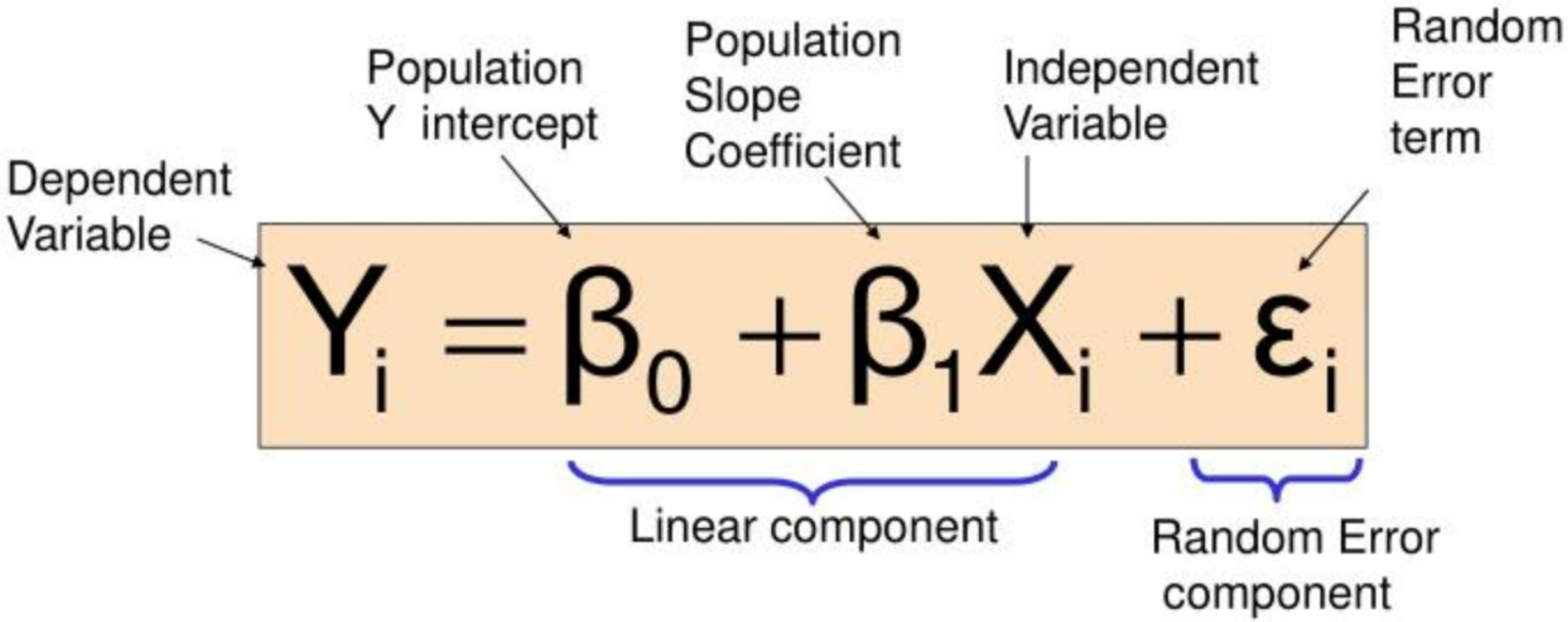


Types of Linear Regression Models



Model equation

The simple linear regression equation provides an estimate of the population regression line.



The individual random error e_i have a mean of zero.

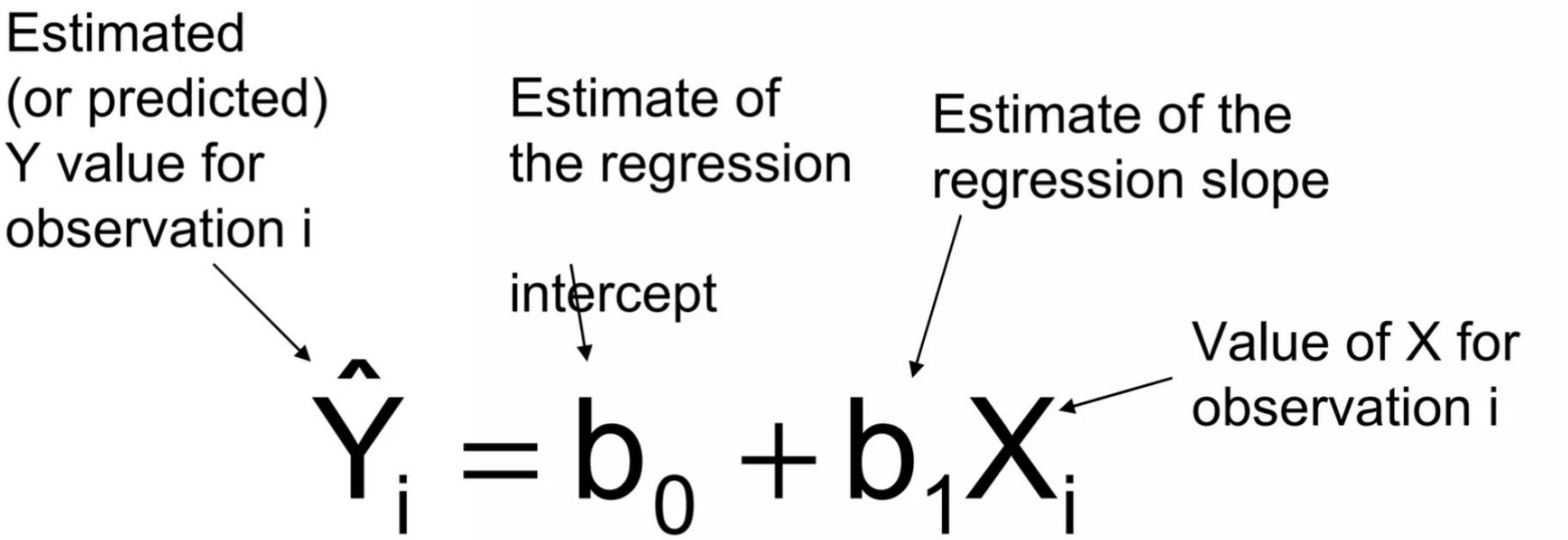
Outline

Least Squarres Method

b_0 and b_1 are obtained by finding the values of b_0 and b_1 that minimize the sum of the squarred differences between y and $\hat{y}$

$\min \sum (y_i - \hat{y}_i)^2 = \min \sum (y_i - (b_0 + b_1 x_i))^2$

- b_0 is the estimated average value of Y when the value of X is zero,
- b_1 is the estimated change in the average value of Y as a result of a one-unit change in X.

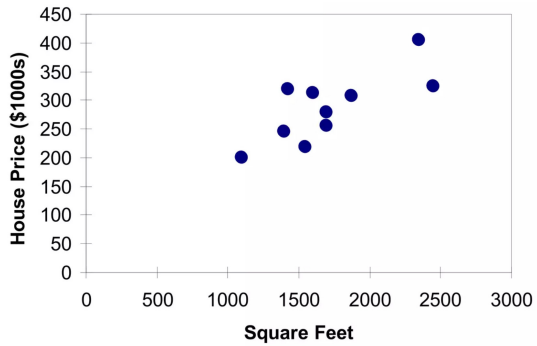


Method’s details

Example

A real estate agent wishes to examine the relationship between the selling price of a home and its size measured in square feet. Y is the house price in $1000s.

House Pice	Squarre Feet
245	1400
312	1600
279	1700
308	1875
199	1100
219	1550
405	2350
324	2450
319	1425
255	1700



$\hat{house price}$=98.24833 + 0.10977 square feet

Predict the price for a house with 2000 square feet: 98.25 + 0.1098(2000)=317.85.

The predicted price for a house with 2000 square feet is 317,850 dollars.

Assumptions of Regression

Use the acronym **LINE**

- Linearity: the underlying relationship between X and Y is linear,
- Independence of errors: error values are statistically independent,
- Normality of errors: error values are normally distributed for any given value of X,
- Equal Variance (homoscedasticity)! The probability distribution of the errors has constant variance.

Current Section

Measures of variation

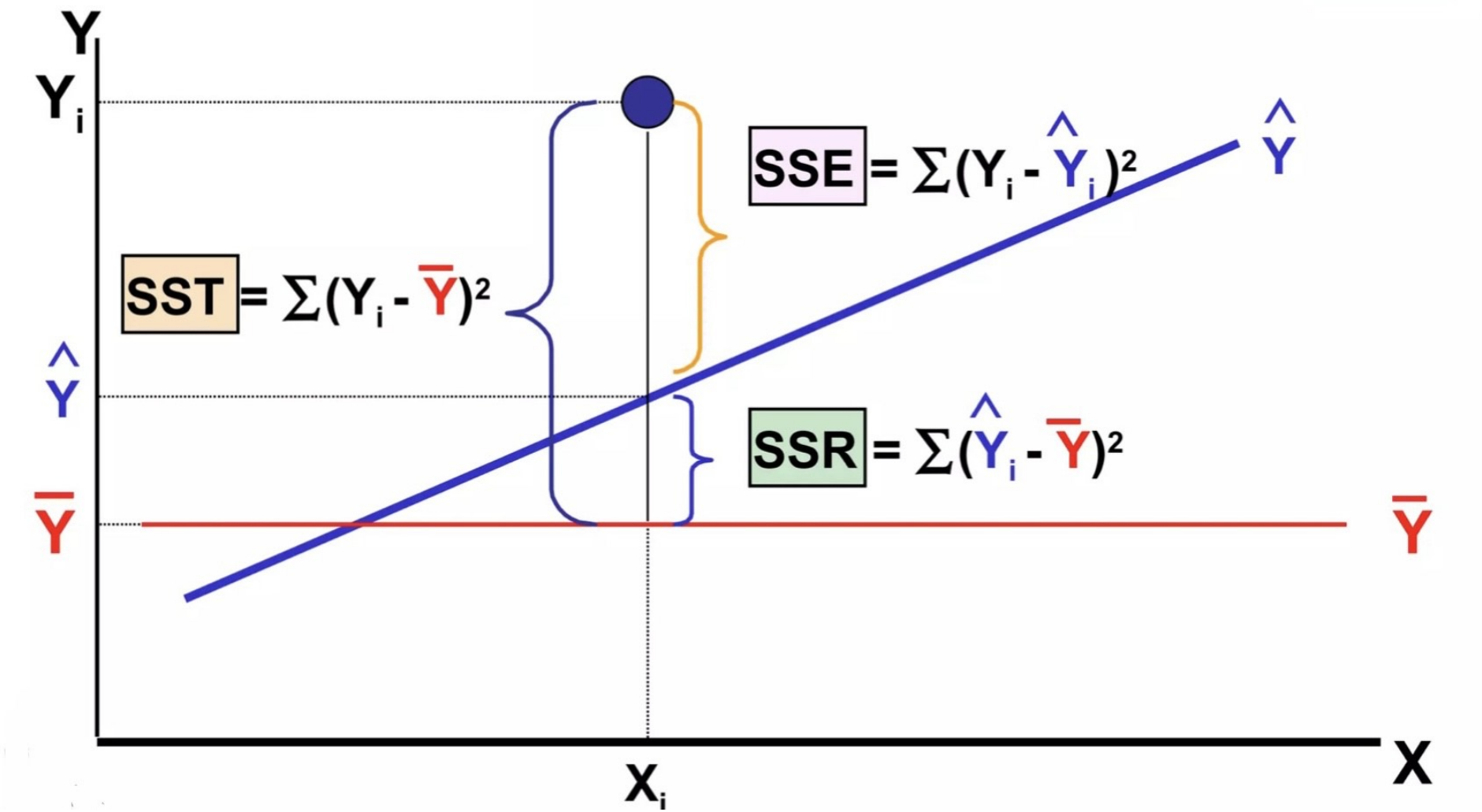
- SST= total sum of squarres
  Measures the variation of the Y_i values around their mean.
- SSR= regression sum of squares
  Explained variation attributable to the relationshiop between X and Y.
- SSE= error sum of squarres
  Variation attributable to factors other than the relationship between X and Y.

- Total variation is made up of two parts:

$$SST = SSR + SSE$$

Total Sum of Squares	Regression Sum of Squares	Error Sum of Squares
$SST = \sum (Y_i - \bar{Y})^2$	$SSR = \sum (\hat{Y}_i - \bar{Y})^2$	$SSE = \sum (Y_i - \hat{Y}_i)^2$

Measures of variation

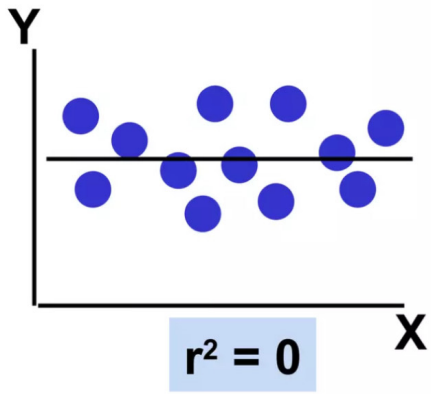


Current Section

Coefficient of Determination r^2

The coefficient of determination is the portion of total variation in the dependent variable that is explained by variation in the independent variable. It is also called r-squared and is denoted as r^2

$$r^2 = \frac{SSR}{SST} = \frac{regressionsumofsquarres}{totalsumofsquares}$$



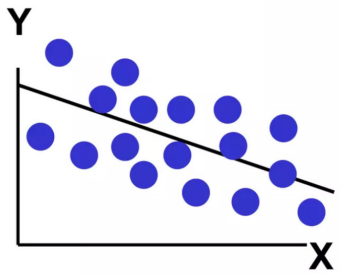
$r^2 = 0$

No linear relationship between X and Y:

The value of Y does not depend on X. (None of the variation in Y is explained by variation in X)

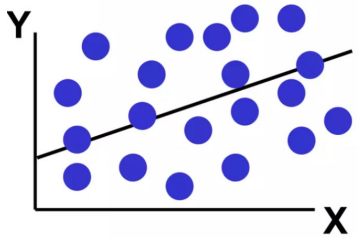
Note that r^2 values range b/w 0 and 1.

Examples of Approximate values

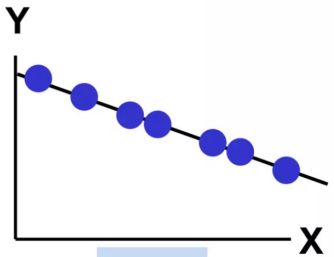


$0 < r^2 < 1$

Weaker linear relationships between X and Y:



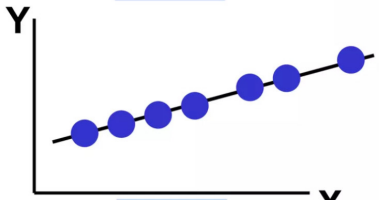
Some but not all of the variation in Y is explained by variation in X



$r^2 = 1$

Perfect linear relationship between X and Y:

$r^2 = 1$



100% of the variation in Y is explained by variation in X

Coefficient of Determination r^2

- Measures the goodness "fit" of the model,
- Assesses the usefulness or predictive value of the model,
- Is interpreted as the proportion of variability of the observed values of Y explained by the regression of Y on X.
- E.g. R^2=71.9%, almost 72 % of the variation in house prices is explained by the regression on their surface.

[]

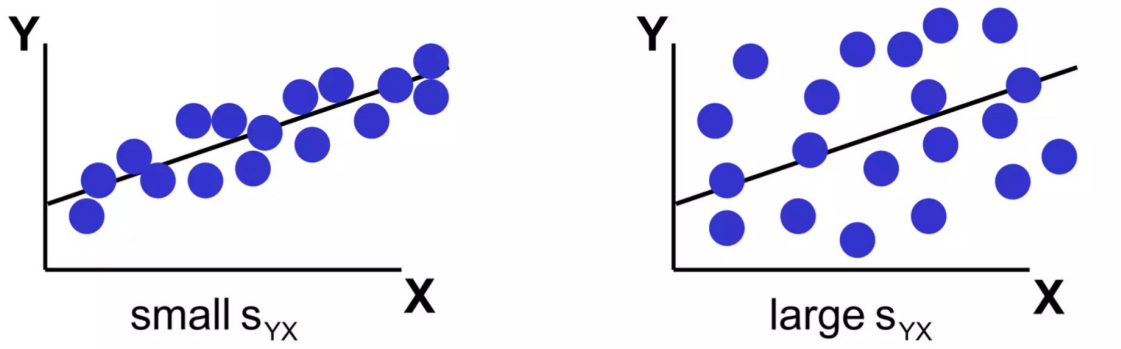
Standard Error of Estimate

- The standard deviation of the variation of observations around the regression line is estimated by

$$S_{YX} = \sqrt{\frac{SSE}{n-2}} = \sqrt{\frac{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}{n-2}}$$

where SSE= error sum of squares and n=sample size.

- The magnitude of S_{YX} should always be judged relative to the size of the Y values in the sample data, i.e. S_{YX}= \$41.33k is moderately small relative to house prices in the \$200 - \$300k range.



Current Section

Multiple Linear regression

$$Y = \beta_0 + \beta_1 X1 + \beta_2 X2 + ... + \beta_p X_p + \varepsilon$$

$$X = \begin{bmatrix} 1 & x_{11} & x_{12} & \dots & x_{1p} \\ 1 & x_{21} & x_{22} & \dots & x_{2p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \dots & x_{np} \end{bmatrix}$$

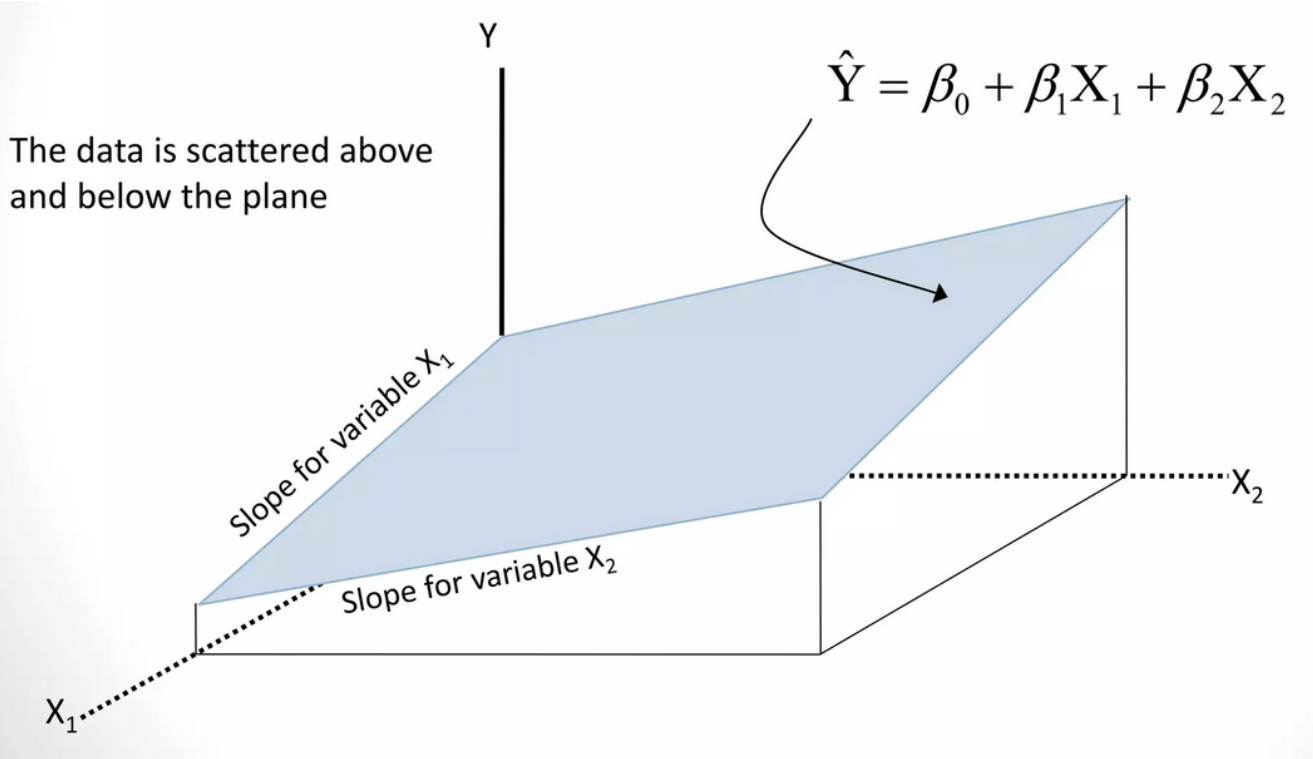
$Y=(y_1 y_2 \dots \dots y_n)^T$

Matrix notation

$$Y=X\beta + \varepsilon$$
$$Y \in \mathbb{R}^n; X \in \mathbb{R}^{n\times(p+1)}; \varepsilon \in \mathbb{R}^n; \beta \in \mathbb{R}^{p+1}$$

n is the sample size (number of observations) and p is the number of predictors.

Multiple Linear regression



Multiple Linear regression

- We want to predict $\hat{Y} = X\hat{\beta}$, by following the same method as in simple linear regression, our estimator will minimize $\mathbb{E}_{\mathbb{X}\mathbb{Y}}\left[(Y - f(X))^2\right]$.
- This means finding the plane or hyper-plane that minimizes the error between the y values in the observations set and the y values that the plane or hyper-plane passes through.
- In other words, we want the plane or hyper-plane that "best fits" the training samples.

$$\hat{\beta} = (X^T X)^{-1} X^T y$$
If $\varepsilon_i \sim \mathcal{N}(0, \sigma_\varepsilon)$ *then* $\hat{\beta} \sim \mathcal{N}(\beta, \sigma^2 (X^T X)^{-1})$

Multicollinearity: IVs shouldn't be overly correlated (e.g >.7), if so-consider removing one.

Current Section

Exercise: Multiple Linear Regression with a Categorical Variable

A company wants to predict employees' salaries in 1K dollars based on:

- Years of Experience (X_1) - Numerical variable.
- Education Level (X_2) - Categorical variable:
  - Bachelorâs degree Encoded as $X_2 = 0$
  - Masterâs degree Encoded as $X_2 = 1$

Years of Experience (X_1)	**Education Level (X_2)**	**Salary (Y) in 1000s**
1	Bachelor (0)	30
2	Bachelor (0)	35
3	Master (1)	50
4	Bachelor (0)	50
5	Master (1)	60
6	Master (1)	65

Question: Find the Regression Model

Using the least squares estimation method, determine the multiple regression model:

$$Y = a + b_1 X_1 + b_2 X_2$$

where:

- a is the intercept.
- b_1 represents the impact of years of experience.
- b_2 represents the impact of having a master's degree (compared to a bachelor's).

Tasks:

1. Compute the regression coefficients b_1 and b_2.
2. Determine the intercept a.
3. Predict the salary for an employee with:
   - 4 years of experience and a Masterâs degree.
   - 6 years of experience and a Bachelorâs degree.

Current Section

Linear regression issues

- Sensitivity to outliers,
- Multicollinearity leads of high variance of the estimator,
- Prone to overfit if there is a lot of variables,
- Hard to interpret if the number of predictors is large(needs a smaller subset that exhibits strongest effects).

- *Predictive Accuracy*
- *Model Interpretability by removing irrelevent features(by setting the corresponding coefficient estimates to zero), we can obtain a model that is more easily interpreted.*

Current Section

Code for Linear Regression in Python

```
# Importing required libraries
import numpy as np
import pandas as pd
import statsmodels.api as sm
import matplotlib.pyplot as plt

# Sample Data: Years of Experience and Salary
data = {
    'Years of Experience': [1, 2, 3, 4, 5, 6, 7, 8, 9, 10],
    'Salary (in $1000)': [30, 35, 50, 50, 60, 65, 70, 75, 8
}

# Creating a DataFrame
df = pd.DataFrame(data)
```

Code for Linear Regression in Python

```
# Defining dependent and independent variables
X = df['Years of Experience']
y = df['Salary (in $1000)']

# Adding constant term (for intercept)
X = sm.add_constant(X)

# Fitting the model
model = sm.OLS(y, X).fit()

# Displaying the summary of the regression
print(model.summary())

# Plotting the regression line
plt.scatter(df['Years of Experience'], df['Salary (in $1000
plt.plot(df['Years of Experience'], model.fittedvalues, col
plt.xlabel('Years of Experience')
plt.ylabel('Salary (in $1000)')
```

Model Summary Analysis

The model summary output from ‘statsmodels‘ provides the following key results:

Regression Equation

$$\text{Salary} = 26.00 + 5.50 \times \text{Years of Experience}$$

- Intercept (a): 26.00 (Salary when years of experience is 0)
- Slope (b_1): 5.50 (Increase in salary for each additional year of experience)
- R-squared: 0.991 (Indicates a very good fit to the data)
- **p-values:** Both the intercept and slope are statistically significant (p-values < 0.05)

Dependent Variable:	Salary (in $1000)
R-squared: **Adj. R-squared:**	0.991 0.990
Intercept (const): **Years of Experience (X):**	$26.00\,(\text{std err} = 2.55, p < 0.001)$ $5.50\,(\text{std err} = 0.40, p < 0.001)$