

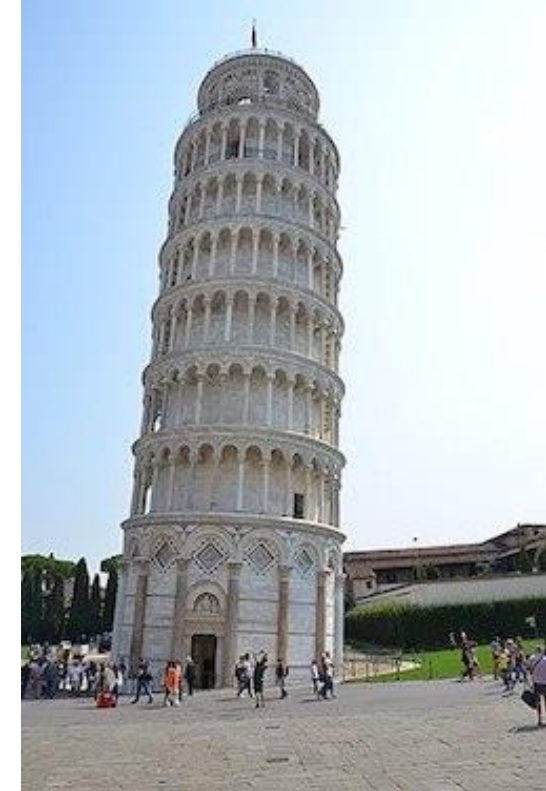
# Lecture 12: Linear Regression

---

Lecturer: Jie Fu

# A motivating example

- The leaning tower of Pisa continuously tilts over time.
- Measurements between 1975 and 1987 of the “lean” of a fixed point on the tower (the distance in measures of the actual position of the point, and its position if the tower was straight).
- We are interested in predicting the lean of the tower for year 1988 based on the data.
- How to do that?



| year | 1975   | 1976   | 1977   | 1978   | 1979   | 1980   | 1981   | 1982   | 1983   | 1984   | 1985   | 1986   | 1987   |
|------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|
| lean | 2.9642 | 2.9644 | 2.9656 | 2.9667 | 2.9673 | 2.9688 | 2.9696 | 2.9698 | 2.9713 | 2.9717 | 2.9725 | 2.9742 | 2.9757 |

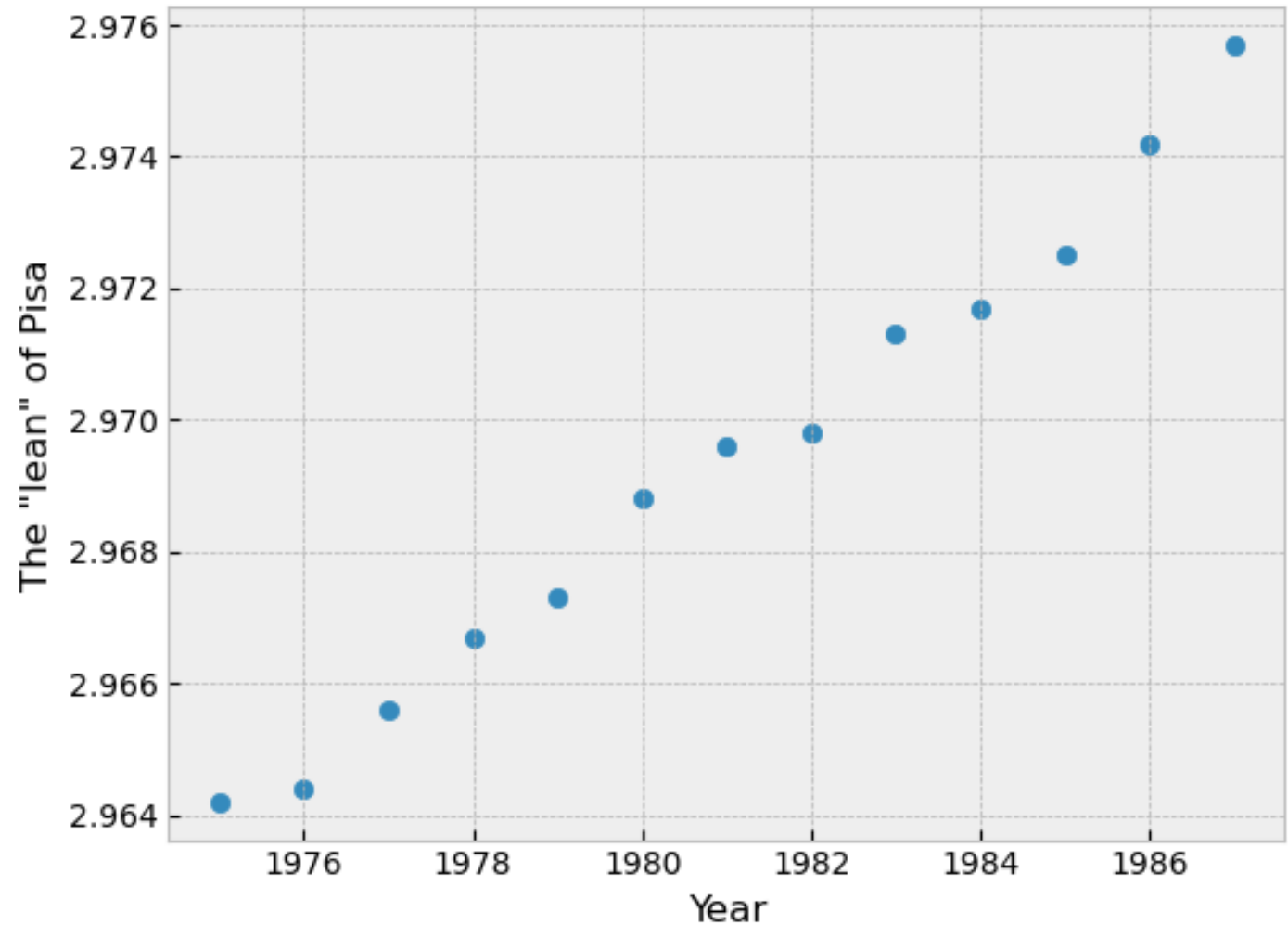
# A general problem statement

---

- We consider the case of only **two variables** for illustration.
- We wish to **model the relation** between two variables of interest,  $x$  and  $y$ 
  - e.g., year and the tilt of the Pisa tower.
- based on a collection of data pairs  $(x_i, y_i)$ ,  $i = 1, \dots, n$ .
  - e.g.  $x_i$  is the year, and  $y_i$  the measured “lean” of the tower in that year.

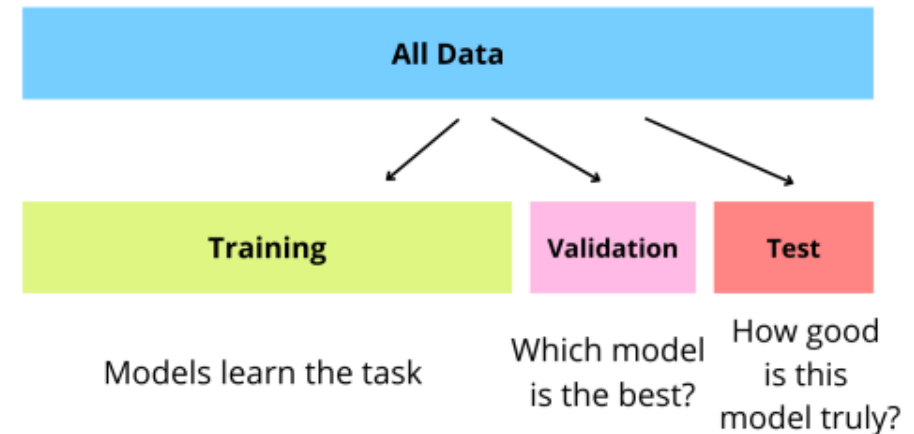
# Visualize the data in python

```
x = range(1975, 1988)
print(len(x))
y = np.array([2.9642, 2.9644, 2.9656,
2.9667, 2.9673, 2.9688, 2.9696,
2.9698, 2.9713, 2.9717, 2.9725,
2.9742, 2.9757])
print(len(y))
plt.scatter(x,y, label='Data')
plt.xlabel('Year')
plt.ylabel('The "lean" of Pisa')
```



# Data

- Training data:
  - Used to train the model by adjusting its parameters. The model learns patterns from this data.
- Validation data:
  - Used to evaluate model performance **during training**.
- Test data:
  - Used to assess the final model's performance on **unseen** data. Provides an unbiased evaluation of its generalization ability.
  - Not available during any part of the learning process



# Model Flexibility

*As model flexibility increases, training error will decrease.*

*The model can fit more and more of the variance in the training set.*

*The test error may or may not decrease.*

*If training error is much larger than test error, the model is **overfitting**.*

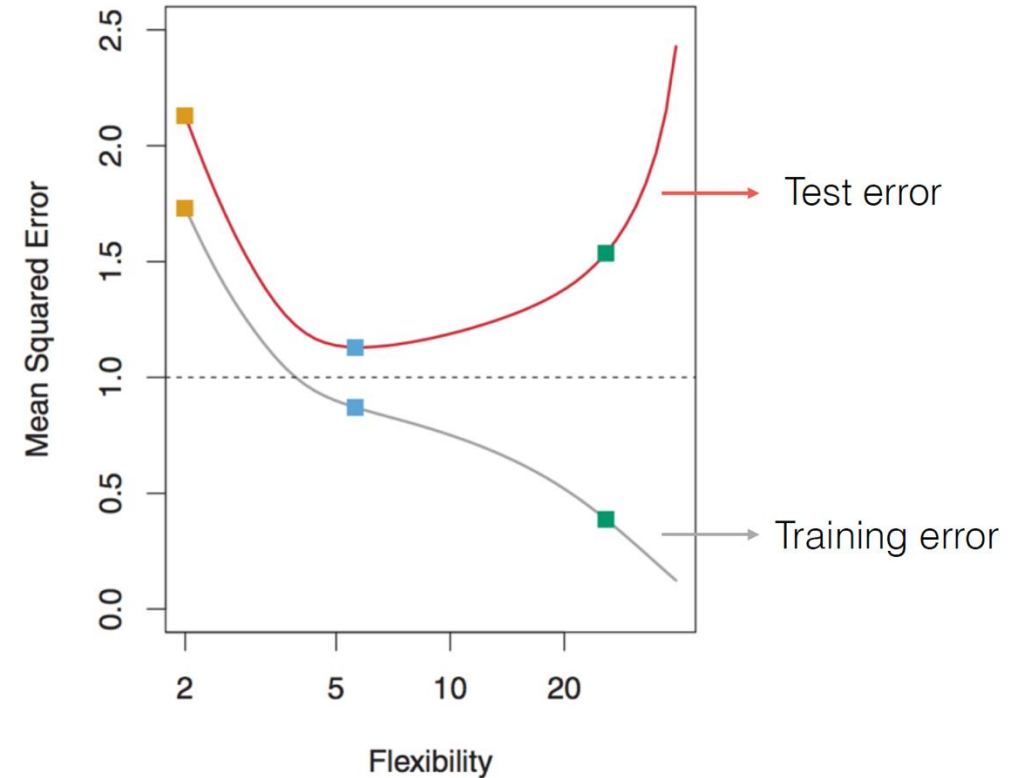
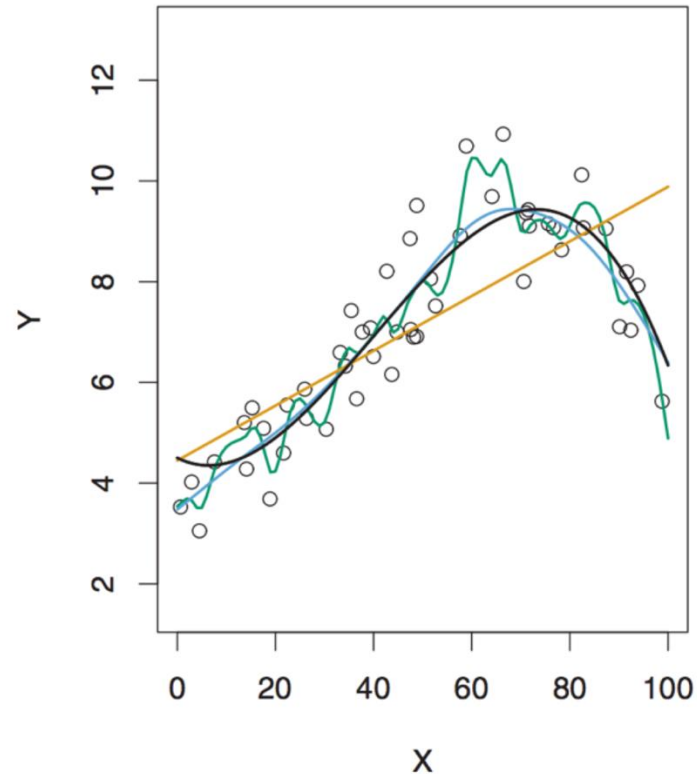


FIGURE 2.9, ISL (8th printing 2017)

# Linear regression

---

- Linear regression assumes that the relationship between the **input variables (features)** and the **output variable (target)** can be described by a straight line:

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

- $Y$ : the dependent variable (output)
- $X$  is the independent variable (input)
- $\beta_0$  is the intercept
- $\beta_1$  is the slope
- $\varepsilon$  is the error term

# Simple Linear Regression

---

- Estimate  $\beta$ s using training data:

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

The estimated value of  $\beta$ s are  $\hat{\beta}$ s

For a particular realization of  $X$ , aka  $X = x$ , the predicted output is denoted  $\hat{y}$ :

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

Goal: Pick  $\hat{\beta}_0, \hat{\beta}_1$ , such that the model is a good fit to the training data.

$$y_i \approx \hat{\beta}_0 + \hat{\beta}_1 x_i, \forall i = 1, \dots, n$$



## Example:

---

House sizes:

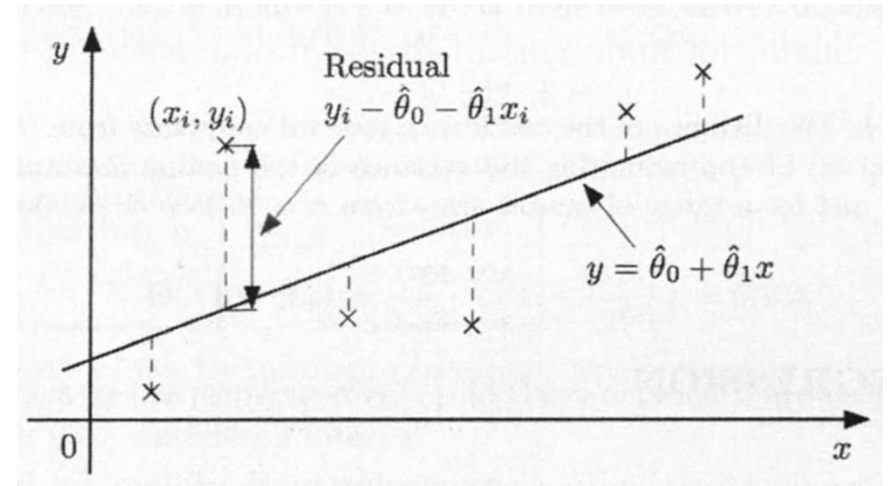
| <b>Size in<br/>feet<sup>2</sup> (x)</b> | <b>Price (\$) in<br/>1000's (y)</b> |
|---|-------------------------------------|
| 2104                                    | 460                                 |
| 1416                                    | 232                                 |
| 1534                                    | 315                                 |
| 852                                     | 178                                 |

Hypothesis:

$$h_{\theta}(x) = -40 + 0.25x$$

Use matrix-vector multiplication:

- Residual for the i-th sample:
- Residual sum of squares (RSS):



- The model fit using least squares finds the parameters, that minimize the RSS:

$$\hat{\beta}_0, \hat{\beta}_1 = \arg \min_{\beta_0, \beta_1} \left[ \sum_{i=1}^n \left( y^{(i)} - \left( \beta_0 + \beta_1 x^{(i)} \right) \right)^2 \right]$$

- It can be verified that:

$$\hat{\beta}_1 = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2}$$

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

# How good is the model fit?

---

- Residual standard error (RSE):

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

where:

- n is the number of observations,
- p is the number of predictors (independent variables). (features) used in a regression model to predict the dependent variable. In **simple linear regression**, there is only **one predictor** (p = 1)
- It represents the standard deviation of residuals and helps assess the goodness of fit for a regression model.

# How good is the model fit?

---

- R<sup>2</sup> statistic:

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

- RSS: Residual sum of squares (RSS)

$$\text{RSS} = \epsilon^{(1)2} + \epsilon^{(2)2} + \dots + \epsilon^{(n)2} = \sum_{i=1}^n (y^{(i)} - \hat{y}^{(i)})^2$$

- TSS: Total sum of squares (TSS)

$$\text{TSS} = \sum_{i=1}^n (y^{(i)} - \bar{y})^2$$

# How good is the model fit?

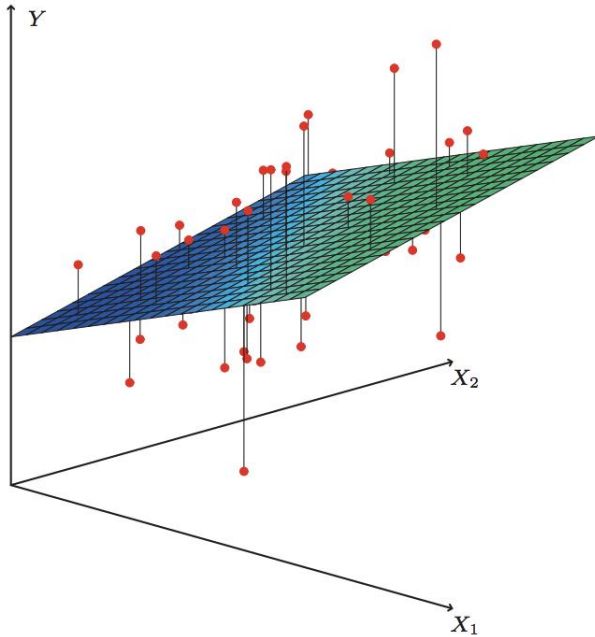
---

- If we know the data comes from a linear model, then  *$R^2$  close to 1 in order to call the fit good.*
- Otherwise, if the linear model is only a crude approximation, then *with  $R^2$  of 0.4 may be considered good.*

# Multiple Linear Regression

- Predict the response variable using more than one predictor variable.

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p + \epsilon \quad (1)$$





A company wants to predict the **price of a house** based on the following features:

- **Size** ( $x_1$ ) in square feet
- **Number of bedrooms** ( $x_2$ )
- **Age of the house** ( $x_3$ ) in years

| Size ( $x_1$ ) | Bedrooms ( $x_2$ ) | Age ( $x_3$ ) | Price ( $y$ ) (in 1000s) |
|----------------|--------------------|---------------|--------------------------|
| 1500           | 3                  | 10            | 300                      |
| 2000           | 4                  | 5             | 450                      |
| 2500           | 4                  | 2             | 500                      |
| 1800           | 3                  | 20            | 280                      |
| 2200           | 5                  | 7             | 480                      |

Table 1: House Price Prediction Data

# Multiple Linear Regression

---

- Using matrix notation

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p + \epsilon \quad (1)$$

# Least squares

---

- RSS: Residual sum of squares (RSS)

$$\text{RSS} = \epsilon^{(1)2} + \epsilon^{(2)2} + \dots + \epsilon^{(n)2} = \sum_{i=1}^n (y^{(i)} - \hat{y}^{(i)})^2$$

- The model fit using least squares finds that minimize the RSS.

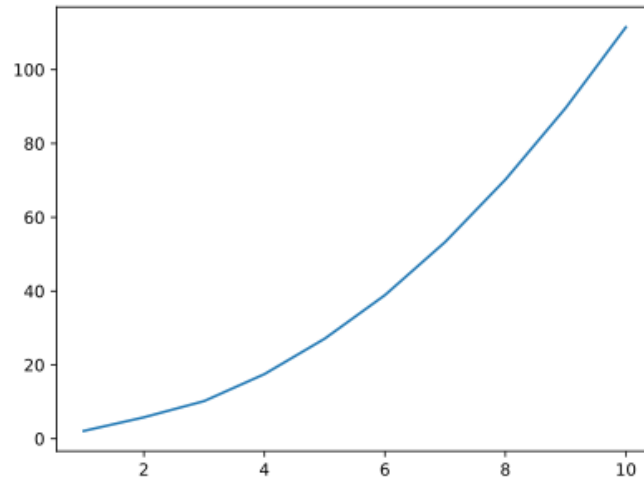
- Analytical solution to least squares

$$\mathbf{X}^\top \mathbf{X} \hat{\beta} = \mathbf{X}^\top \mathbf{Y} \quad \Rightarrow \quad \hat{\beta} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{Y}$$

# Non-linearity of Data

- A simple way to extend linear regression to model non-linear relationships is via polynomial regression.

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \cdots + \beta_n x^n + \epsilon \quad (1)$$



$x = [1, 2, 3, 4, 5, 6, 7, 8, 9, 10]$

$y = [2.1, 5.8, 10.2, 17.5, 27.1, 38.9, 53.3, 70.2, 89.6, 111.5]$

- Which degree of polynomials to select?
- Can we derive the analytical solution?