HCMC University of Technology

Dung Nguyen

Probability and Statistics

Linear regression Model

Outline I

Example I

Hooke's law states that the force applied to a spring is proportional to the distance that the spring is stretched. Thus, if F is the force applied and x is the distance that the spring has been stretched, then $F = kx$. The proportionality constant k is called the spring constant. Some physics students want to determine the spring constant for a given spring. They apply forces of 2, 5, and 7 pounds, which have the effect of stretching the spring 5, 8, and 10 inches, respectively.

$$5k = 2 \implies k = 0.4$$
$$8k = 5 \implies k = 0.625$$
$$10k = 7 \implies k = 0.7.$$

First guess:

$$k = \frac{0.4 + 0.625 + 0.7}{3} = \frac{1}{3} \cdot 0.4 + \frac{1}{3} \cdot 0.625 + \frac{1}{3} \cdot 0.7.$$

Example II

First guess:

$$k = \frac{0.4 + 0.625 + 0.7}{3} = \frac{1}{3} \cdot 0.4 + \frac{1}{3} \cdot 0.625 + \frac{1}{3} \cdot 0.7.$$

Optimal value:

$$\begin{aligned} k &= \frac{5 \cdot 2 + 8 \cdot 5 + 10 \cdot 7}{5^2 + 8^2 + 10^2} \\ &= \frac{5^2 \cdot 0.4 + 8^2 \cdot 0.625 + 10^2 \cdot 0.7}{5^2 + 8^2 + 10^2} \\ &= \frac{25}{189} \cdot 0.4 + \frac{64}{189} \cdot 0.625 + \frac{100}{189} \cdot 0.7. \end{aligned}$$

Also

$$\begin{bmatrix} 5 & 8 & 10 \end{bmatrix} \begin{bmatrix} 5 \\ 8 \\ 10 \end{bmatrix} k = \begin{bmatrix} 5 & 8 & 10 \end{bmatrix} \begin{bmatrix} 2 \\ 5 \\ 7 \end{bmatrix}$$

Thus $189k = 120$.

Empirical Models

Many problems in engineering and science involve exploring the relationships between two or more variables -> Regression analysis.

- In a chemical process: the yield of the product is related to the process-operating temperature.
- Regression analysis can be used to build a model to predict yield at a given temperature level.

- The simple linear regression considers a single regressor or predictor x and a dependent or response variable Y.
- The expected value of Y at each level of x is a random variable.

$$E(Y|x) = \alpha + \beta x$$

- We assume that each observation, Y, can be described by the model.

$$Y = \alpha + \beta x + \epsilon$$

That is

$$\begin{aligned} Y_1 &= \alpha + \beta X_1 + \epsilon_1 \\ Y_2 &= \alpha + \beta X_2 + \epsilon_2 \\ &\dots \\ Y_n &= \alpha + \beta X_n + \epsilon_n \end{aligned}$$

The least-squares estimates of the intercept and slope in the simple linear regression model are

$$b = \frac{S_{xy}}{S_{xx}}$$

and

$$a = \overline{y} - b\overline{x}.$$

with $S_{xx} = \sum_{i=1}^{n}(x_i - \overline{x})^2$ and $S_{xy} = \sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})$, where

$$S_{xx} = \sum_{i=1}^{n}(x_i - \overline{x})^2 = \sum_{i=1}^{n} x_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)^2$$

$$S_{xy} = \sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y}) = \sum_{i=1}^{n} x_i y_i - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)$$

$$S_{yy} = \sum_{i=1}^{n}(y_i - \overline{y})^2 = \sum_{i=1}^{n} y_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} y_i\right)^2.$$

The fitted or estimated regression line is therefore

$$\hat{y}_i = a + bx_i$$

Note that each pair of observations satisfies the relationship

$$y_i = a + bx_i + \epsilon_i, \quad i = 1, \ldots, n$$

where $e_i = y_i - \hat{y}_i$ is called the residual. The residual describes the error in the fit of the model to the i-th observation y_i.

Estimate $\mathsf{E}(Y|X = x^*)$: $\quad \hat{y}^* = a + bx^*$

The sum of squares of the residuals

$$SSE = \sum (y_i - \hat{y}_i)^2.$$

The total sum of squares of the response variable

$$SST = \sum (y_i - \overline{y})^2.$$

The sum of squares for regression

$$SSR = \sum (\hat{y}_i - \overline{y})^2$$

Fundamental identity

$$SST = SSE + SSR$$

Computational formula

$$SST = S_{yy}$$

$$SSR = bS_{xy}$$

$$SSE = S_{yy} - bS_{xy}$$

It can be shown that

$$\mathsf{E}(SSE) = (n-2)\sigma^2.$$

An unbiased estimator of σ^2

$$S^2 = \frac{SSE}{n-2}$$

Coefficient of determination

$$r^2 = 1 - \frac{SSE}{SST}.$$

Properties of the Least Squares Estimators I

- Slope Properties

$$\mathsf{E}(b) = \beta, \quad \mathsf{V}(b) = \frac{\sigma^2}{S_{xx}} \equiv \sigma_b^2$$

The estimated standard error of the slope

$$S_b = \sqrt{\frac{S^2}{S_{xx}}}.$$

Moreover

$$T = \frac{b-\beta}{S_b} \sim \mathbf{t}(n-2).$$

Properties of the Least Squares Estimators II

- Intercept Properties

$$\mathsf{E}(a) = \alpha, \quad \mathsf{V}(a) = \frac{\sigma^2 \mu_{xx}}{S_{xx}} \equiv \sigma_a^2.$$

with $\mu_{xx} = \frac{1}{n}\sum x_i^2$. The estimated standard error of the slope

$$S_a = \sqrt{\frac{S^2 \mu_{xx}}{S_{xx}}} = \sqrt{S_b^2 \mu_{xx}}.$$

Moreover

$$T = \frac{a - \alpha}{S_a} \sim \mathbf{t}(n-2).$$

Confidence interval

Confidence interval for the slope

$$b \pm t_{v/2,n-2} \cdot S_b$$

Confidence interval for the intercept

$$a \pm t_{v/2,n-2} \cdot S_a$$

Hypothesis testing

- Slope: $T = \frac{b-\beta_0}{S_b}$ with

$$H_1 : \beta \neq \beta_0, \quad H_1 : \beta < \beta_0, \quad H_1 : \beta > \beta_0$$

- Intercept: $T = \frac{a-\alpha_0}{S_a}$ with

$$H_1 : \alpha \neq \alpha_0, \quad H_1 : \alpha < \alpha_0, \quad H_1 : \alpha > \alpha_0$$

Observation Number	Hydrocarbon Level x(%)	Purity y(%)
1	0.99	90.01
2	1.02	89.05
3	1.15	91.43
4	1.29	93.74
5	1.46	96.73

Michaelis-Menten Equation

The hydrolysis of carbobenzozyglycyl-L-tryptophan catalyzed by pancreatic carboxypeptidase occurs ac- cording to the reaction:

$$\text{carbobenzozyglycyl-L-tryptophan} + H_2O \rightarrow \text{carbobenzoxyglycine} + \text{L-tryptophan}$$

The following data on the rate of formation of L-tryptophan at 25 °C and pH = 7.5 was obtained:

Substrate Concentration (mM)	Rate
2.5	0.026
5.0	0.037
10.0	0.054
15.0	0.061
20.0	0.065

1. Plot the initial rate as a function of substrate concentration. What shape does the data follow? 95% confidence intervals for β_1 and β_2, and find the r^2 value for the fit. Are they what you expect?
2. Now linearize the model using the Lineweaver-Burk method and solve for V_{max} and K_M. Find the 95% confidence intervals for the slope and intercept of your Lineweaver-Burk plot and determine the r^2 value.
3. Make a residual plot to assess the fit from part d.