

Simple Linear Regression

Data

x	x_1	x_2	x_3	\cdots	x_n
y	y_1	y_2	y_3	\cdots	y_n

x: independent, predictor, or explanatory variable

y: dependent, response variable

x is deterministic and y is random

Scatter Plot: linear trend

Linear regression model

$$Y = \beta_0 + \beta_1 x + \epsilon, \quad \epsilon \sim N(0, \sigma^2)$$

The true regression line: $y = \beta_0 + \beta_1 x$

For fixed x : $Y \sim N(\beta_0 + \beta_1 x, \sigma^2)$

Source of variation for Y

Estimating β_0 and β_1

The measure of the goodness of fit

Vertical deviation: $y_i - (b_0 + b_1 x_i)$

Sum of Square Deviations:

$$f(b_0, b_1) = \sum_{i=1}^n [y_i - (b_0 + b_1 x_i)]^2$$

Principle of Least Square:

$$f(\hat{\beta}_0, \hat{\beta}_1) = \min f(b_0, b_1)$$

$\hat{\beta}_0$ and $\hat{\beta}_1$: the least square estimates of β_0 and β_1 .

Estimated regression line (least square line)

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

Formulas for $\hat{\beta}_0$ and $\hat{\beta}_1$

$$\frac{\partial f(b_0, b_1)}{\partial b_0} = \sum 2(y_i - b_0 - b_1 x_i)(-1) = 0$$

$$\frac{\partial f(b_0, b_1)}{\partial b_1} = \sum 2(y_i - b_0 - b_1 x_i)(-x_i) = 0$$

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

where

$$s_{xy} = \sum x_i y_i - (\sum x_i)(\sum y_i)/n$$

$$s_{xx} = \sum x_i^2 - (\sum x_i)^2/n$$

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

Example

Estimating σ^2

Fitted values:

$$\hat{y}_1, \hat{y}_2, \dots, \hat{y}_n$$

Residuals:

$$y_1 - \hat{y}_1, y_2 - \hat{y}_2, \dots, y_n - \hat{y}_n$$

Residual Sum of Square:

$$\text{SSE} = f(\hat{\beta}_0, \hat{\beta}_1) = \sum [y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i]^2$$

The estimate of σ^2

$$\hat{\sigma}^2 = s^2 = \frac{\text{SSE}}{n - 2}$$

The Coefficient of Determination

Total Sum of Squares:

$$\text{SST} = \sum (y_i - \bar{y})^2 = \sum y_i^2 - (\sum y_i)^2 / n$$

Error Sum of Square:

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

Regression Sum of Square

$$\text{SSR} = \text{SST} - \text{SSE}$$

Coefficient of Determination:

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

Example

Inferences about β_1 and β_0

Point estimators:

$$\hat{\beta}_1 = \frac{\sum (x_i - \bar{x})(Y_i - \bar{Y})}{\sum (x_i - \bar{x})^2} = \sum c_i Y_i$$

where $c_i = (x_i - \bar{x})/s_{xx}$, $s_{xx} = \Sigma(x_i - \bar{x})^2$.

$$\hat{\beta}_0 = \bar{Y} - \bar{x} = \Sigma(1/n - \bar{x}c_i)Y_i$$

$$\hat{\sigma}^2 = \frac{\Sigma(Y_i - \hat{\beta}_0 - \hat{\beta}_1x_i)^2}{n - 2}$$

$$E(\hat{\beta}_1) = \beta_1, E(\hat{\beta}_0) = \beta_0$$

$$Var(\hat{\beta}_1) = \Sigma c_i^2 Var(Y_i) = \frac{\sigma^2}{s_{xx}}$$

$$Var(\hat{\beta}_0) = \Sigma(1/n - \bar{x}c_i)^2 Var(Y_i) = \sigma^2(1/n + \frac{\bar{x}^2}{s_{xx}})$$

The distributions of $\hat{\beta}_1$ and $\hat{\beta}_0$:

t distributions:

$$\frac{\hat{\beta}_1 - \beta_1}{\hat{\sigma}/\sqrt{s_{xx}}} \sim t(n - 2)$$

$$\frac{\hat{\beta}_0 - \beta_0}{\hat{\sigma}\sqrt{1/n + \frac{\bar{x}^2}{s_{xx}}}} \sim t(n - 2)$$

100(1 - α)% **confidence intervals**

1. β_1

2. β_0

Hypothesis-testing for β_1 and β_0

1. $H_0 : \beta_1 = \beta_{10}$

Test statistic:

2. $H_0 : \beta_0 = \beta_{00}$

Test statistic:

Linear relationship test (model utility test)

1. t test

2. ANOVA

Inferences concerning $\mu_{Y\dot{x}^*}$

$$\mu_{Y\dot{x}^*} = \beta_0 + \beta_1 x^*$$

$$\hat{\mu}_{Y\dot{x}^*} = \hat{\beta}_0 + \hat{\beta}_1 x^*$$

$$\hat{\beta}_0 + \hat{\beta}_1 x^* = \sum \left[\frac{1}{n} + \frac{(x^* - \bar{x})(x_i - \bar{x})}{\sum (x_i - \bar{x})^2} \right] Y_i = \sum d_i Y_i$$

Let $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 x^*$

$$E(\hat{Y}) =$$

$$Var(\hat{Y}) =$$

Theorem

$$T = \frac{\hat{\beta}_0 + \hat{\beta}_1 x^* - (\beta_0 + \beta_1 x^*)}{S_{\hat{Y}}} = \frac{\hat{Y} - (\beta_0 + \beta_1 x^*)}{S_{\hat{Y}}} \sim t(n - 2)$$

100(1 - α) CI for $\mu_{Y_{x^*}}$

A prediction interval for a future value of Y

$$E(\hat{\beta}_0 + \hat{\beta}_1 x^* - Y) = 0$$

$$Var(\hat{\beta}_0 + \hat{\beta}_1 x^* - Y) = \left[1 + \frac{1}{n} + \frac{(x^* - \bar{x})^2}{S_{xx}} \right]$$

Theorem

$$T = \frac{\hat{\beta}_0 + \hat{\beta}_1 - Y}{S \sqrt{1 + \frac{1}{n} + \frac{(x^* - \bar{x})^2}{S_{xx}}}} \sim t(n - 2)$$

100(1 - α) CI for a future Y observation

Exmample

x	1	2	3	4	5	6	7	8	9	10
y	5.86	7.21	6.33	11.58	10.55	11.09	15.83	16.76	20.11	20.03
\hat{y}	4.80	6.52	8.24	9.96	11.67	13.39	15.11	16.82	18.54	20.26
$\hat{\epsilon}$	1.05	0.68	-1.91	1.62	-1.12	-2.30	0.72	-0.06	1.56	-0.23