

Linear regression

Purpose: model linear dependence of two quantitative variables & predict one variable from the value of the other.

ex. Old Faithful, MLB, Hanford/Columbia River cancer study, etc.

Basic idea: start with a normal distribution for Y ; allow the mean of Y to depend linearly on the value of a fixed, known **covariate**, x .

Model:

$$Y \sim \text{normal}(\beta_0 + \beta_1 x, \sigma^2)$$

In this model, $E(Y) = \mu = \beta_0 + \beta_1 x$, and $V(Y) = \sigma^2$.
Another (equivalent) way of writing the model:

$$Y = \beta_0 + \beta_1 x + \epsilon$$

where $\epsilon \sim \text{normal}(0, \sigma^2)$.

Data: each observation is an ordered pair, n observations in all, written as $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$.
Commonly depicted in a **scatterplot**.

Statistical inferences:

- Point estimates of unknown parameters β_0, β_1 , and σ^2
- Confidence intervals
- Hypothesis tests
- Estimate & CI for $\mu = \beta_0 + \beta_1 x$ at a particular value of x
- Prediction of a new value of Y at x
- Model evaluation
- Matrix representation

Pdf for Y is that of a normal distribution:

$$f(y) = \frac{1}{\sqrt{\sigma^2 2\pi}} e^{-\frac{[y - (\beta_0 + \beta_1 x)]^2}{2\sigma^2}}$$

Likelihood: product of normal pdf's, evaluated at (x_1, y_1) , (x_2, y_2) , ..., (x_n, y_n) :

$$L = \left(\frac{1}{\sqrt{\sigma^2 2\pi}} \right)^n e^{-\frac{1}{2\sigma^2} [(y_1 - \beta_0 - \beta_1 x_1)^2 + \dots + (y_n - \beta_0 - \beta_1 x_n)^2]}$$

ML estimates of β_0, β_1 jointly minimize a sum of squares:

$$(y_1 - \beta_0 - \beta_1 x_1)^2 + (y_2 - \beta_0 - \beta_1 x_2)^2 + \dots + (y_n - \beta_0 - \beta_1 x_n)^2$$

The ML estimates, $\hat{\beta}_0$ and $\hat{\beta}_1$ are the **least-squares** estimates. Formulas:

$$S_{xx} = \sum_{i=1}^n (x_i - \bar{x})^2 \quad S_{xy} = \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})$$

$$\hat{\beta}_1 = \frac{S_{xy}}{S_x}$$

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

The ML estimate of σ^2 is the average squared departure of the y_i 's from their estimated means:

$$\hat{\sigma}_{\text{ML}}^2 = \frac{1}{n} \sum_{i=1}^n \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i \right)^2$$

Usually the unbiased estimate of σ^2 is used:

$$\hat{\sigma}^2 = \frac{1}{n-2} \sum_{i=1}^n \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i \right)^2$$

Note: the term

$$\text{SS}(\text{residual}) = \sum_{i=1}^n \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i \right)^2$$

is the “residual sum of squares” (squared errors left over after model is fitted). The “regression sum of squares” is

$$\text{SS}(\text{regression}) = \sum_{i=1}^n \left(\hat{\beta}_0 + \hat{\beta}_1 x_i - \bar{y} \right)^2$$

or the squared departures of the values predicted by the regression model from the grand mean of the y_i 's. As in AOV, these sum to the “total sum of squares” in the y_i 's:

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

$$\text{SS}(\text{total}) = \text{SS}(\text{regression}) + \text{SS}(\text{residual})$$

Confidence intervals

Under repeated sampling (at the same values of the x_i 's), the parameter estimates $\hat{\beta}_0$ and $\hat{\beta}_1$ have normal distributions, with the following means and variances:

$$E(\hat{\beta}_0) = \beta_0 \text{ (unbiased)}$$

$$V(\hat{\beta}_0) = \sigma_{\hat{\beta}_0}^2 = \sigma^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}} \right)$$

estimate with $\hat{\sigma}^2$

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

$$V(\hat{\beta}_1) = \sigma_{\hat{\beta}_1}^2 = \frac{\sigma^2}{S_{xx}}$$

estimate with $\hat{\sigma}^2$

100(1 - α)% CIs based on Student's t distribution with $n - 2$ df:

$$\beta_0: \quad \hat{\beta}_0 \pm t_{\alpha/2} \sqrt{\hat{\sigma}^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}} \right)}$$

$$\beta_1: \quad \hat{\beta}_1 \pm t_{\alpha/2} \sqrt{\frac{\hat{\sigma}^2}{S_{xx}}}$$

Hypothesis test for β_0 :

$$H_0: \beta_0 = c \text{ (known constant)}$$

$$H_a: \beta_0 \left\{ \begin{array}{l} > \\ < \\ \neq \end{array} \right\} c$$

$$t = \frac{\hat{\beta}_0 - c}{\sqrt{\hat{\sigma}^2 \left(\frac{1}{n} + \frac{\bar{x}}{S_{xx}} \right)}}$$

$$\text{Reject } H_0 \text{ if } \left\{ \begin{array}{l} t \geq t_\alpha \\ t \leq -t_\alpha \\ |t| \geq t_{\alpha/2} \end{array} \right\} \text{ where } df = n - 2$$

Hypothesis test for β_1 :

$$H_0: \beta_1 = c \text{ (known constant)}$$

$$H_a: \beta_1 \left\{ \begin{array}{l} > \\ < \\ \neq \end{array} \right\} c$$

$$t = \frac{\hat{\beta}_1 - c}{\sqrt{\frac{\hat{\sigma}^2}{S_{xx}}}}$$

$$\text{Reject } H_0 \text{ if } \left\{ \begin{array}{l} t \geq t_\alpha \\ t \leq -t_\alpha \\ |t| \geq t_{\alpha/2} \end{array} \right\} \text{ where } df = n - 2$$

Note: for testing $H_0: \beta_1 = 0$ (model has no predictive value) vs. $H_a: \beta_1 \neq 0$ (model has some predictive value), one can show that

$$t^2 = f = \frac{SS(\text{regression})/1}{SS(\text{residual})/(n-2)}$$

and that this test statistic has an $F(1, n - 2)$ distribution under the null hypothesis. One would reject H_0 if $f \geq f_\alpha$. This test statistic is usually printed by computer regression packages.