Linear Regression – Coefficient of Determination

EAS 199A Notes

Gerald Recktenwald
Portland State University
Department of Mechanical Engineering
gerry@me.pdx.edu

Overview

- Continuation of least squares curve fitting
- R^2 as a measure of goodness of fit

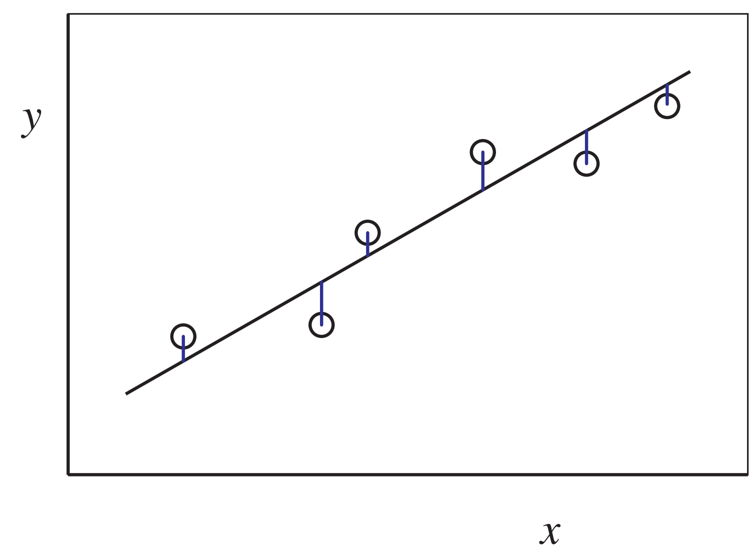
The Residual

The difference between the given y_i value and the fit function evaluated at x_i is

$$\begin{aligned} r_i &= y_i - \hat{y}_i \\ &= y_i - (mx_i + b) \end{aligned}$$

r_i is the *residual* for the data pair (x_i, y_i).

r_i is the vertical distance between the known data and the fit function.



Minimizing the Residual

Two criteria for choosing the “best” fit

$$\text{minimize} \sum |r_i| \qquad \text{or} \qquad \text{minimize} \sum r_i^2$$

For statistical and computational reasons choose minimization of $\rho = \sum r_i^2$

$$\rho = \sum_{i=1}^{m} \left[y_i - (mx_i + b) \right]^2$$

The best fit is obtained by the values of m and b that minimize ρ.

Coefficients of a Line Fit

Given the data

$$(x_i, y_i), \qquad i = 1, \ldots, n$$

finding the minimum of ρ (the minimum of the sum of squares) yields

$$m = \frac{n \sum x_i y_i - \left(\sum x_i\right)\left(\sum y_i\right)}{n \sum x_i^2 - \left(\sum x_i\right)^2} \tag{1}$$

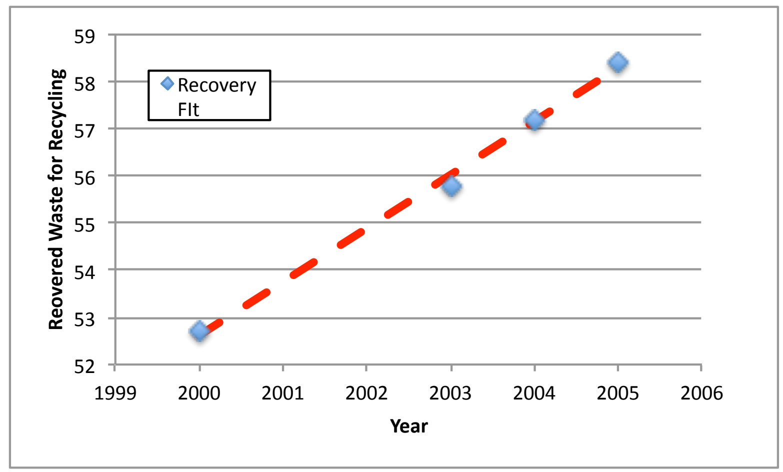
$$b = \frac{\sum y_i - m \sum x_i}{n} \tag{2}$$

$$= \frac{\left(\sum x_i^2\right)\left(\sum y_i\right) - \left(\sum x_i\right)\left(\sum x_i y_i\right)}{n \sum x_i^2 - \left(\sum x_i\right)^2} \tag{3}$$

How do we assess the quality of the fit?

Using $y = mx + b$ implies that the data has a linear relationship between the input x and the output y. That is not always the case.

For example, factors other than time are likely to influence the rate of recycling between 2000 and 2004.

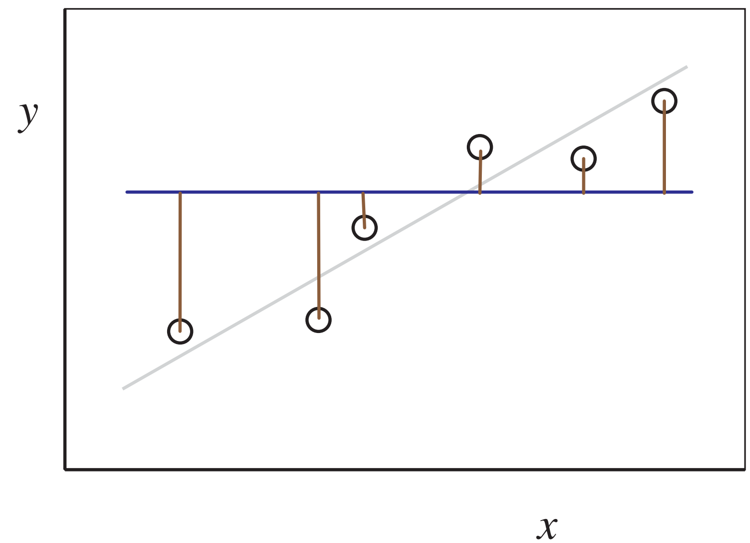


The mean of the y_i data

If $\bar{y}$ is a good model of the data, there there is no meaningful relationship between y and x. In that case, x is said to *not explain the data*.

The mean of the dependent variable is

$$\bar{y} = \frac{1}{n}\sum y_i$$



The R^2 Statistic

R^2 is a measure of how well the fit function follows the trend in the data. $0 \leq R^2 \leq 1$.

Define:

$\hat{y}$ is the value of the fit function at the known data points.

For a line fit $\quad \hat{y}_i = c_1 x_i + c_2$

Then:

$$R^2 = \frac{\sum (\hat{y}_i - \bar{y})^2}{\sum (y_i - \bar{y})^2} = 1 - \frac{\sum (y_i - \hat{y})^2}{\sum (y_i - \bar{y})^2}$$

When $R^2 \approx 1$ the fit function follows the trend of the data.
When $R^2 \approx 0$ the fit is not significantly better than approximating the data by its mean.

Alternative single pass formula for R^2

The value of R^2 produced by the preceding equations is equivalent to

$$R^2 = \left[\frac{n \sum x_i y_i - \left(\sum x_i\right)\left(\sum y_i\right)}{\sqrt{n \sum x_i^2 - \left(\sum x_i\right)^2}\,\sqrt{n \sum y_i^2 - \left(\sum y_i\right)^2}} \right]^2$$