UNIVERSITÀ
DEGLI STUDI
DI TRIESTE

Statistics

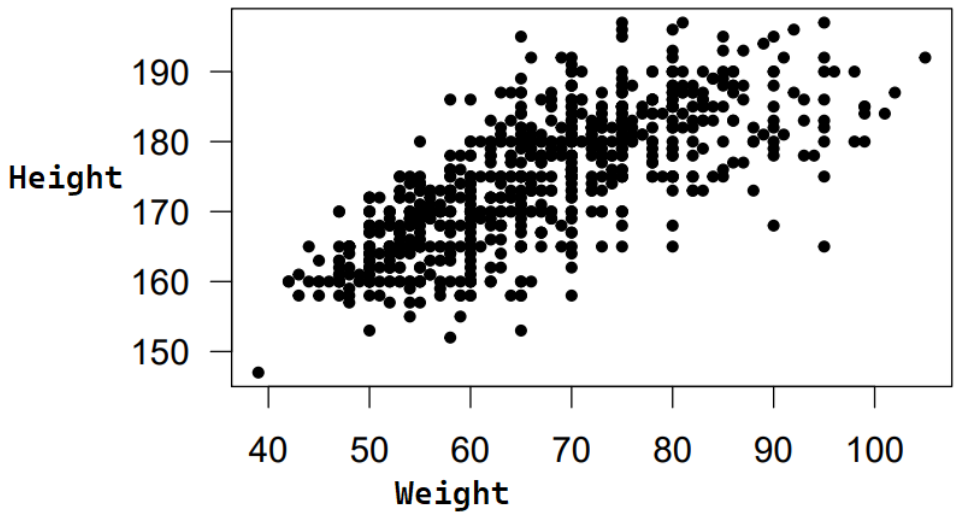
Regression

Sara Geremia

May 9th, 2024

Let's start from here:

Height and weight (in a set of units) are positively associated



A first model

Let's adopt the assumption of a linear relationship.

We can think of a model like:

$$\text{height} = \beta_0 + \beta_1 \text{weight} + \text{error}$$

Where the error represents the portion of height fluctuations not related to the weight (or, better said, that a linear function of the weight fails to explain)

A model of this type is called a simple linear regression model.

Simple Linear Regression Models

In the general case, we aim to explain a variable, let's say y, using another variable, let's say x, through a model of the form

$$y = \beta_0 + \beta_1 x + \text{error}$$

- y response or dependent variable, while
- x explanatory or independent variable, also called regressor, or predictor.
- β_0 intercept
- β_1 slope coefficient

β_0 and β_1 are the parameters of the model. The problem is how to "determine" β_0 and β_1.

Least Squares: Concept I

If we can calculate a "reasonable" value for these two parameters, let's say $\hat{\beta}_0$ and $\hat{\beta}_1$, we can then think of "predicting" the height of students using

$$height = \beta_0 + \beta_1 weight + \text{error}$$

It seems reasonable to choose two values for the parameters, $\hat{\beta}_0$ and $\hat{\beta}_1$, so that the regression line "reproduces" our data well, so that

$$\begin{aligned} y_1 &\approx \hat{\beta}_0 + \hat{\beta}_1 x_1 \\ y_2 &\approx \hat{\beta}_0 + \hat{\beta}_1 x_2 \\ &\vdots \\ y_N &\approx \hat{\beta}_0 + \hat{\beta}_1 x_N \end{aligned}$$

To make the idea "operational" we need to decide:

Least Squares: Concept II

- how we interpret the $\approx$ we wrote and
- how we combine the various approximations.

The most commonly used solution consist in choosing the two parameters by minimizing

$$s^2(\beta_0, \beta_1) = \sum_{i=1}^{N} (y_i - \beta_0 - \beta_1 x_i)^2$$

that is, by choosing $\hat{\beta}_0$ and $\hat{\beta}_1$ such that

$$s^2(\hat{\beta}_0, \hat{\beta}_1) \leq s^2(\beta_0, \beta_1)$$

for any $\beta_0 \in \mathbb{R}$ and $\beta_1 \in \mathbb{R}$.

In this case, it is said that the parameters are calculated using the least squares method.

Least Squares: Parameter Determination I

Step 1 Fixing β_1 to any value, the problem becomes

$$\inf_{\beta_0 \in \mathbb{R}} \sum_{i=1}^{N} (y_i - \beta_0 - \beta_1 x_i)^2 = \inf_{\beta_0 \in \mathbb{R}} \sum_{i=1}^{N} (z_i - \beta_0)^2$$

with $z_i = y_i - \beta_1 x_i$.

We know the constant that minimizes the mean of the squares of deviations from a value is the arithmetic mean of the z_i.

So

$$\hat{\beta}_0 = \frac{1}{N} \sum_{i=1}^{N} z_i = \frac{1}{n} \sum_{i=1}^{N} (y_i - \beta_1 x_i) = \overline{y} - \beta_1 \overline{x}$$

where $\overline{y}$ and $\overline{x}$ denote respectively the mean of the y_i and that of the x_i.

Least Squares: Parameter Determination II

Step 2 The quantity to minimize then becomes

$$s^2(\hat{\beta}_0, \beta_1) = \sum_{i=1}^{N} [y_i - \overline{y} - \beta_1(x_i - \overline{x})]^2.$$

Deriving with respect to β_1 and setting the derivative to zero yields the equation (for β_1)

$$-2\sum_{i=1}^{N}(x_i - \overline{x})[(y_i - \overline{y}) - \beta_1(x_i - \overline{x})] = 0,$$

which can be rewritten as

$$\sum_{i=1}^{N}(x_i - \overline{x})(y_i - \overline{y}) = \beta_1 \sum_{i=1}^{N}(x_i - \overline{x})^2.$$

Least Squares: Parameter Determination III

If $\sum_{i=1}^{N}(x_i - \overline{x})^2 > 0$, the above equation has a unique solution

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{N}(x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{N}(x_i - \overline{x})^2}$$

So,

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

$$\hat{\beta}_1 = \frac{\sigma_{XY}}{\sigma_X^2}$$

where $\overline{y}$, $\overline{x}$, σ_X^2, and σ_{XY} are respectively the mean of the response variable, the mean and the variance of the explanatory variable, and the covariance between response and explanatory.

It must hold that $\sigma_X^2 > 0$. This is very reasonable: β_1 tells us how the response varies with the explanatory variable, but if $\sigma_X^2 = 0$, the explanatory variable did not vary at all in the available data.

Example: dataset `trees`

$$\sum y_i = 935{,}3 \qquad \sum x_i = 410{,}7$$
$$\sum x_i^2 = 5736{,}5 \qquad \sum x_i y_i = 13887{,}86.$$

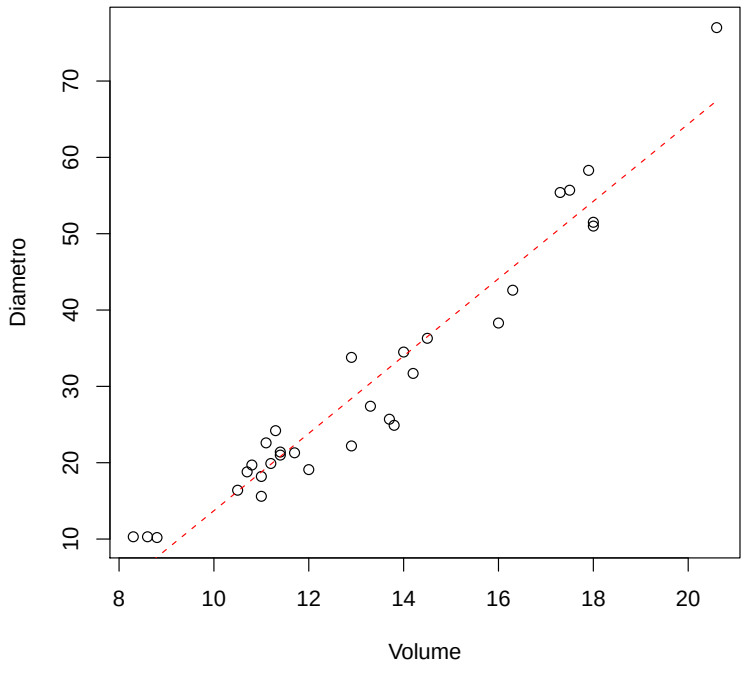
So

$$\overline{y} = 935{,}3/31 = 30{,}2$$
$$\overline{x} = 410{,}7/31 = 13{,}2$$
$$\sigma_X^2 = (5736{,}5/31) - 13{,}2^2 = 9{,}5$$
$$\sigma_{XY} = (13887{,}86/31) - 13{,}2 \times 30{,}2 = 48{,}3.$$

Hence

$$\hat{\beta} = 48{,}3/9{,}5 = 5{,}1$$
$$\hat{\beta}_0 = 30{,}2 - 5{,}1 \times 13{,}2 = -37{,}1.$$

Scatterplot with Regression Line



The ability to describe the variation in volume seems fine, except perhaps for the more "outlying" observations.

Observed, Predicted, Residual Values

The following quantities are of interest.

- y_i observed value for Y on the i-th statistical unit
- $\hat{y}_i$ predicted value for Y on the i-th statistical unit: $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$
- r_i residual for the i-th statistical unit: $r_i = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i$

The predicted value lies on the estimated regression line; the residuals measure the distance between the observed value and the regression line.

Mean of Residuals

It is easy to verify that the mean of the residuals is zero.

$$\begin{aligned}\sum_{i=1}^{N} r_i &= \sum_{i=1}^{N} y_i - N\hat{\beta}_0 - \hat{\beta}\sum_{i=1}^{N} x_i \\ &= N\overline{y} - N(\overline{y} - \hat{\beta}\overline{x}) - N\hat{\beta}\overline{x} = 0\end{aligned}$$

Residual Variance

$$
\begin{aligned}
\mathrm{var}(r_1, \ldots, r_N) &= \sigma_R^2 = \frac{1}{N}\sum_{i=1}^{N} r_i^2 = \\
&= \frac{1}{N}\sum_{i=1}^{N}[(y_i - \overline{y}) - \hat{\beta}(x_i - \overline{x})]^2 = \\
&= \frac{1}{N}\sum_{i=1}^{N}(y_i - \overline{y})^2 + \frac{\hat{\beta}^2}{N}\sum_{i=1}^{N}(x_i - \overline{x})^2 - \\
&\quad -2\frac{\hat{\beta}}{N}\sum_{i=1}^{N}(x_i - \overline{x})(y_i - \overline{y}) = \\
&= \sigma_Y^2 + \hat{\beta}^2\sigma_X^2 - 2\hat{\beta}\sigma_{XY} = \\
&= \sigma_Y^2 + \sigma_{XY}^2/\sigma_X^2 - 2\sigma_{XY}^2/\sigma_X^2 = \\
&= \sigma_Y^2 - \sigma_{XY}^2/\sigma_X^2
\end{aligned}
$$

Residual Variance (continued)

- The residual variance, which coincides with the mean squared residuals, is always no larger than the variance of the response.
- It can be used to get a "numerical idea" of how well the model fits the data.
- The smaller σ_R^2 is, the more the regression line "explains" the variations in the response. When $\sigma_R^2 = 0$, all observations lie on the regression line.
- When $\sigma_{XY} = 0$, meaning in the absence of a linear relationship, $\sigma_R^2 = \sigma_Y^2$.

Coefficient of Determination

The fraction of the variance of the response (Y) explained by the simple linear regression model is given by

$$R^2 = 1 - \frac{\sigma_R^2}{\sigma_Y^2}$$

$$0 \leq R^2 \leq 1$$

$R^2 = 1 \longrightarrow \sigma_R^2 = 0$: the model perfectly explains the response.
$R^2 = 0 \longrightarrow \sigma_R^2 = \sigma_Y^2$: the model explains nothing.

Example: `trees` Dataset

$$\bar{y} = 935.3/31 = 30.2$$
$$\sigma_X^2 = (5736.5/31) - 13.2^2 = 9.5$$
$$\sigma_{XY} = (13887.86/31) - 13.2 \times 30.2 = 48.3.$$

Moreover

$$\sum y_i^2 = 36324.99$$

So

$$\sigma_Y^2 = 36324.99/31 - 30.2^2 = 261.5$$

and thus

$$\sigma_R^2 = 261.5 - 48.3^2/9.5 = 15.9.$$

The coefficient of determination is

$$R^2 = 1 - 15.9/261.5 = 0.94,$$

meaning the model explains just under 95% of the variance in the response.