Create account Not logged in Talk Contributions Log in

Article Talk Read Edit View history Search Go

WIKIPEDIA
The Free Encyclopedia

Main page
Contents
Featured content
Current events
Random article
Donate to Wikipedia
Wikipedia store

Interaction
Help
About Wikipedia
Community portal
Recent changes
Contact page

Tools
What links here
Related changes
Upload file
Special pages
Permanent link
Page information
Wikidata item
Cite this page

Print/export
Create a book
Download as PDF
Printable version

Languages
Add links

Lack-of-fit sum of squares

From Wikipedia, the free encyclopedia

In statistics, a **sum of squares due to lack of fit**, or more tersely a **lack-of-fit sum of squares**, is one of the components of a partition of the sum of squares in an analysis of variance, used in the numerator in an F-test of the null hypothesis that says that a proposed model fits well.

Contents [hide]

Sketch of the idea [edit]

In order for the lack-of-fit sum of squares to differ from the sum of squares of residuals, there must be more than one value of the response variable for at least one of the values of the set of predictor variables. For example, consider fitting a line

$$y = \alpha x + \beta$$

by the method of least squares. One takes as estimates of α and β the values that minimize the sum of squares of residuals, i.e., the sum of squares of the differences between the observed y-value and the fitted y-value. To have a lack-of-fit sum of squares that differs from the residual sum of squares, one must observe more than one y-value for each of one or more of the x-values. One then partitions the "sum of squares due to error", i.e., the sum of squares of residuals, into two components:

sum of squares due to error = (sum of squares due to "pure" error) + (sum of squares due to lack of fit).

The sum of squares due to "pure" error is the sum of squares of the differences between each observed y-value and the average of all y-values corresponding to the same x-value.

The sum of squares due to lack of fit is the *weighted* sum of squares of differences between each average of y-values corresponding to the same x-value and the corresponding fitted y-value, the weight in each case being simply the number of observed y-values for that x-value.[1][2] Because it is a property of least squares regression that the vector whose components are "pure errors" and the vector of lack-of-fit components are orthogonal to each other, the following equality holds:

$$\begin{aligned}
&\sum(\text{observed value} - \text{fitted value})^2 && \text{(error)} \\
&\quad = \sum(\text{observed value} - \text{local average})^2 && \text{(pure error)} \\
&\qquad + \sum \text{weight} \times (\text{local average} - \text{fitted value})^2 && \text{(lack of fit)}
\end{aligned}$$

Hence the residual sum of squares has been completely decomposed into two components.

Mathematical details [edit]

Consider fitting a line with one predictor variable. Define i as an index of each of the n distinct x values, j as an index of the response variable observations for a given x value, and n_i as the number of y values associated with the i th x value. The value of each response variable observation can be represented by

$$Y_{ij} = \alpha x_i + \beta + \varepsilon_{ij}, \qquad i = 1, \dots, n, \quad j = 1, \dots, n_i.$$

Let

$$\widehat\alpha, \widehat\beta$$

be the least squares estimates of the unobservable parameters α and β based on the observed values of x_i and Y_{ij}.

Let

$$\widehat Y_i = \widehat\alpha x_i + \widehat\beta$$

be the fitted values of the response variable. Then

$$\widehat\varepsilon_{ij} = Y_{ij} - \widehat Y_i$$

are the residuals , which are observable estimates of the unobservable values of the error term ε_{ij}. Because of the nature of the method of least squares, the whole vector of residuals, with

$$N = \sum_{i=1}^n n_i$$

scalar components, necessarily satisfies the two constraints

$$\sum_{i=1}^n \sum_{j=1}^{n_i} \widehat\varepsilon_{ij} = 0$$

$$\sum_{i=1}^n \left(x_i \sum_{j=1}^{n_i} \widehat\varepsilon_{ij} \right) = 0.$$

It is thus constrained to lie in an ($N - 2$)-dimensional subspace of $\mathbf{R}^N$, i.e. there are $N - 2$ "degrees of freedom for error".

Now let

$$\overline Y_{i\bullet} = \frac{1}{n_i} \sum_{j=1}^{n_i} Y_{ij}$$

be the average of all Y-values associated with the i th x-value.

We partition the sum of squares due to error into two components:

$$\sum_{i=1}^n \sum_{j=1}^{n_i} \widehat\varepsilon_{ij}^{\,2} = \sum_{i=1}^n \sum_{j=1}^{n_i} \left(Y_{ij} - \widehat Y_i \right)^2$$

$$= \underbrace{\sum_{i=1}^n \sum_{j=1}^{n_i} \left(Y_{ij} - \overline Y_{i\bullet}\right)^2}_{\text{(sum of squares due to pure error)}} + \underbrace{\sum_{i=1}^n n_i \left(\overline Y_{i\bullet} - \widehat Y_i\right)^2}_{\text{(sum of squares due to lack of fit)}}.$$

Probability distributions [edit]

Sums of squares [edit]

Suppose the error terms ε_{ij} are independent and normally distributed with expected value 0 and variance σ^2. We treat x_i as constant rather than random. Then the response variables Y_{ij} are random only because the errors ε_{ij} are random.

It can be shown to follow that if the straight-line model is correct, then the **sum of squares due to error** divided by the error variance,

$$\frac{1}{\sigma^2}\sum_{i=1}^{n}\sum_{j=1}^{n_i}\widehat\varepsilon_{ij}^{\,2}$$

has a chi-squared distribution with $N - 2$ degrees of freedom.

Moreover, given the total number of observations N, the number of levels of the independent variable n, and the number of parameters in the model p:

- The sum of squares due to pure error, divided by the error variance σ^2, has a chi-squared distribution with $N - n$ degrees of freedom;
- The sum of squares due to lack of fit, divided by the error variance σ^2, has a chi-squared distribution with $n - p$ degrees of freedom (here $p = 2$ as there are two parameters in the straight-line model);
- The two sums of squares are probabilistically independent.

The test statistic [edit]

It then follows that the statistic

$$F = \frac{\text{lack-of-fit sum of squares/degrees of freedom}}{\text{pure-error sum of squares/degrees of freedom}}$$

$$= \frac{\left.\sum_{i=1}^{n} n_i\left(\overline{Y}_{i\bullet} - \widehat{Y}_i\right)^2\right/(n-p)}{\left.\sum_{i=1}^{n}\sum_{j=1}^{n_i}\left(Y_{ij} - \overline{Y}_{i\bullet}\right)^2\right/(N-n)}$$

has an F-distribution with the corresponding number of degrees of freedom in the numerator and the denominator, provided that the model is correct. If the model is wrong, then the probability distribution of the denominator is still as stated above, and the numerator and denominator are still independent. But the numerator then has a noncentral chi-squared distribution, and consequently the quotient as a whole has a non-central F-distribution.

One uses this F-statistic to test the null hypothesis that there is no lack of linear fit. Since the non-central F-distribution is stochastically larger than the (central) F-distribution, one rejects the null hypothesis if the F-statistic is larger than the critical F value. The critical value corresponds to the cumulative distribution function of the F distribution with x equal to the desired confidence level, and degrees of freedom $d_1 = (n - p)$ and $d_2 = (N - n)$. This critical value can be calculated using online tools[3] or found in tables of statistical values.[4]

The assumptions of normal distribution of errors and independence can be shown to entail that this lack-of-fit test is the likelihood-ratio test of this null hypothesis.

See also [edit]

- Linear regression

Notes [edit]

1. ^ Brook, Richard J.; Arnold, Gregory C. (1985). *Applied Regression Analysis and Experimental Design*. CRC Press. pp. 48–49. ISBN 0824772520.
2. ^ Neter, John; Kutner, Michael H.; Nachstheim, Christopher J.; Wasserman, William (1996). *Applied Linear Statistical Models* (Fourth ed.). Chicago: Irwin. pp. 121–122. ISBN 0256117365.
3. ^ Soper, D.S. "Critical F-value Calculator (Online Software)". *Statistics Calculators*. Retrieved 19 April 2012.
4. ^ Lowry, Richard. "VassarStats". *Concepts and Applications of Inferential Statistics*. Retrieved 19 April

2012.

Categories: Analysis of variance | Regression analysis | Design of experiments | Least squares

This page was last modified on 28 January 2015, at 18:52.