In some cases, we may know certain statistical properties (e.g., variances, correlation) of the data error. That information can be sometimes used to obtain better estimates of the true values.

CORE OF THE METHOD

Combine all residuals into a sum function as

$$L_d = \sqrt[d]{\sum_{n=1}^{N} |r_n|^d}$$

where $d = 1$ is called L_1 norm, $d = 2$ gives the L_2 norm (Euclidean norm) used to determine geometric length of vectors, and $d \to \infty$ gives L_∞ norm that corresponds to largest residual value $\max_{1 \le n \le N} \|r_n\|$.

Example:

$$x = \begin{bmatrix} 2 \\ 5 \\ -3 \end{bmatrix}$$

$$\|x\|_1 = 10$$

$$\|x\|_2 = \sqrt{4+25+9} \approx 6.1644$$

$$\|x\|_\infty = 5$$

$$\|x\|_p = \sqrt[p]{2^p + 5^p + 3^p}$$

DEF: A norm is a function $\|\cdot\| : R^n \to R$ that satisfies

(1) $\|x\| \ge 0$, and $\|x\| = 0$ only if $x = 0$,

(2) $\|x + y\| \le \|x\| + \|y\|$,

(3) $\|\alpha x\| = |\alpha| \, \|x\|$.

Least-square fitting starts by combining all the residuals into a single value typically called χ^2

$$\chi^2 = \sum_{n=1}^{N} (\tilde{y}_n - \hat{y}_n)^2 = \sum_{n=1}^{N} \varepsilon_n^2 \sim N\varepsilon^2$$

in which the final order of magnitude estimate is accurate if all data errors are of similar magnitude ε.

A more complete representation uses weights (w) that could be used to emphasize the effect of certain residuals

$$\chi^2 = \sum_{n=1}^{N} w_n (\tilde{y}_n - \hat{y}_n)^2 = \sum_{n=1}^{N} \frac{(\tilde{y}_n - \hat{y}_n)^2}{\sigma_n^2} = \sum \left(\frac{observed - expected}{uncertainty} \right)^2$$

σ: measurement uncertainty associated to the statistical dispersion of the values attributed to a measured quantity. For a "good" model, we should expect that the difference between *expected* (model) and *observed* (data) values should be on the order of the uncertainty. This means that non-zero residuals are only due to the measurement uncertainty! This would produce chi-squared values $\sim N$. One can also compute the *reduced chi-squared* defined as $\chi_N^2 = \chi^2 / N$.

Rule of thumb (practical guide) to determine goodness-of-fit

Assuming that the measured values are Gaussian random variables, the fitting may not be so good (χ^2 seems too large) because:

- The model may be wrong, e.g., rather than $f(x) = a_0 + a_1 x$, one should have picked $f(x) = a_0 + a_1 x + a_2 x^2$.
- Whoever produced the uncertainties σ underestimated them, i.e., they are actually much larger which would improve the fitting.
- If $\chi^2 \approx 0$, you may see your model going through the data points, but this may not be necessarily good; you may be using too many parameters, thereby "overfitting".
- Whoever produced the uncertainties σ overestimated them, then many fitting attempts may look "good".

Typically, $\chi^2 \approx N - p$ implies a reasonably good fit in which N is the number of data points and p is the number of parameters. $N - p$ is also known as the number of *degrees of freedom*. Therefore, one can say that for a reasonably good fit (not too good/overfitting or not too bad/underfitting) should expect *chi-squared per degree of freedom to be roughly one.*

Straight-Line Fit and Linear Regression

$$y = f(x|a_0, a_1) = a_0 + a_1 x$$

Given a set of N measured points (y_i, x_i) with $i = 1, \ldots, N$, our goal is to find the values of the slope a_1 and the ordinate at the origin (intercept) a_0 that minimize the chi-square function.

$$\chi^2 = \sum_{i=1}^{N} \frac{\left(y_i - f(x_i|a_0, a_1)\right)^2}{\sigma_i^2} = \sum_{i=1}^{N} \frac{(y_i - a_0 - a_1 x)^2}{\sigma_i^2}$$

To find the minimum, we find values of the two parameters (slope and intercept) for which derivatives with respect to a_0 and a_1 are null simultaneously:

$$\frac{\partial \chi^2(a_0, a_1)}{\partial a_0} = 0$$

$$\frac{\partial \chi^2(a_0, a_1)}{\partial a_1} = 0$$

$$a_0 \sum_i \frac{1}{\sigma_i^2} + a_1 \sum_i \frac{x_i}{\sigma_i^2} = \sum_i \frac{y_i}{\sigma_i^2}$$

$$a_0 \sum_i \frac{x_i}{\sigma_i^2} + a_1 \sum_i \frac{x_i^2}{\sigma_i^2} = \sum_i \frac{x_i y_i}{\sigma_i^2}$$

Using a simpler notation…

$$S \equiv \sum_{i=1}^{N} \frac{1}{\sigma_i^2} \qquad S_x \equiv \sum_{i=1}^{N} \frac{x_i}{\sigma_i^2} \qquad S_y \equiv \sum_{i=1}^{N} \frac{y_i}{\sigma_i^2}$$

The calculation of these 'S' terms relies exclusively on the data points!

$$S_{xx} \equiv \sum_{i=1}^{N} \frac{x_i^2}{\sigma_i^2} \qquad S_{xy} \equiv \sum_{i=1}^{N} \frac{x_i y_i}{\sigma_i^2} \qquad \Delta = S_{xx} S - S_x^2$$

The system of equations above can be re-written as

$$a_0 S + a_1 S_x = S_y$$

$$a_0 S_x + a_1 S_{xx} = S_{xy}$$

or equivalently in matrix form

$$\alpha\vec{a} = \vec{b}$$

$$\alpha = \begin{pmatrix} S & S_x \\ S_x & S_{xy} \end{pmatrix} \qquad \vec{a} = \begin{pmatrix} a_0 \\ a_1 \end{pmatrix} \qquad \vec{b} = \begin{pmatrix} S_y \\ S_{xy} \end{pmatrix}$$

To solve for $\vec{a}$, we do $\vec{a} = \alpha^{-1}\vec{b}$ in which

$$\alpha^{-1} = \frac{1}{\Delta}\begin{pmatrix} S_{xx} & -S_x \\ -S_x & S \end{pmatrix}$$

This is the *symmetric error matrix* or *covariance matrix* that carries the variances of the measured parameters (diagonal terms) and the covariance between the parameters in the two off-diagonal terms.

The estimators $\hat{a}_0$ and $\hat{a}_1$, which minimize the χ^2 function are therefore

$$\hat{a}_0 = \frac{1}{\Delta}\left(S_y S_{xx} - S_x S_{xy}\right)$$

$$\hat{a}_1 = \frac{1}{\Delta}\left(S S_{xy} - S_x S_y\right)$$

or equivalently in matrix form

$$\alpha\vec{a} = \vec{b}$$

$$\alpha = \begin{pmatrix} S & S_x \\ S_x & S_{xy} \end{pmatrix} \qquad \vec{a} = \begin{pmatrix} a_0 \\ a_1 \end{pmatrix} \qquad \vec{b} = \begin{pmatrix} S_y \\ S_{xy} \end{pmatrix}$$

To solve for $\vec{a}$, we do $\vec{a} = \alpha^{-1}\vec{b}$ in which

$$\alpha^{-1} = \frac{1}{\Delta}\begin{pmatrix} \sigma_{a_0}^2 & \sigma_{a_0 a_1}^2 \\ \sigma_{a_0 a_1}^2 & \sigma_{a_1}^2 \end{pmatrix}$$

This is the *symmetric error matrix* or *covariance matrix* that carries the variances of the measured parameters (diagonal terms) and the covariance between the parameters in the two off-diagonal terms.

The estimators $\hat{a}_0$ and $\hat{a}_1$, which minimize the χ^2 function are therefore

$$\hat{a}_0 = \frac{1}{\Delta}\left(S_y S_{xx} - S_x S_{xy}\right)$$

$$\hat{a}_1 = \frac{1}{\Delta}\left(S S_{xy} - S_x S_y\right)$$

The estimators $\hat{a}_0$ and $\hat{a}_1$, which minimize the χ^2 function are therefore

$$\hat{a}_0 = \frac{1}{\Delta}\left(S_y S_{xx} - S_x S_{xy}\right)$$

$$\hat{a}_1 = \frac{1}{\Delta}\left(S S_{xy} - S_x S_y\right)$$

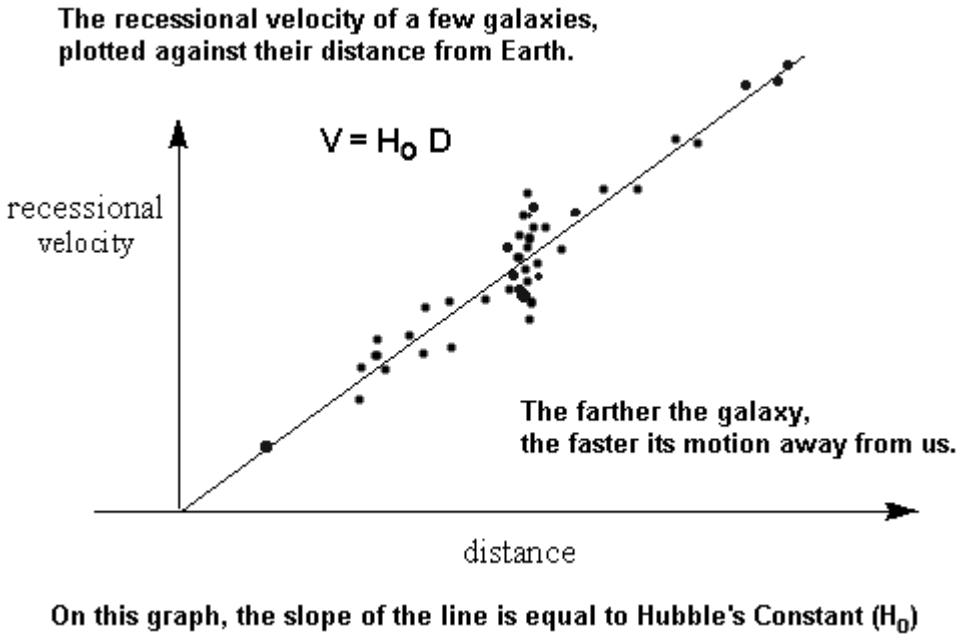
These equations yield estimators for the straight-line parameters that best fit the measured data points.

Note that if the values y_i are given without estimates of their standard deviations, σ_i, it suffices to set all $\sigma_i = 1$.

According to the textbook of C. A. Pruneau, the term **fit** or **curve-fitting** is usually reserved for problems where a model is used to infer a linear relationship between the dependent and the independent variables (known a priori).

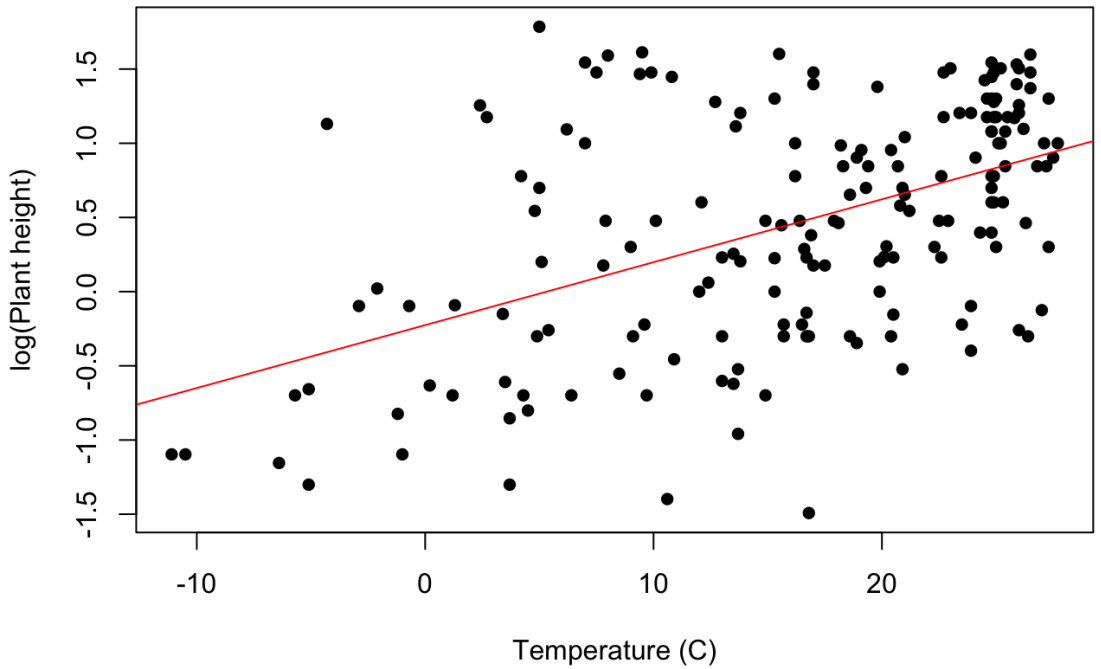
Linear regression is typically used for cases where no model is known a priori, or whenever large variances characterize both variables. The procedure thus yields an estimate of the trend between the variables, akin to an estimate of correlation.

CURVE FIT



Hubble's law, also known as the Hubble–Lemaître law, is the observation in physical cosmology that galaxies are moving away from Earth at speeds proportional to their distance.

LINEAR REGRESSION



Linear regression often deals with linear fitting in which the **data points do not have uncertainties.**

The fact that the measurements y_i each carry an error σ_i implies the model parameters a_0 and a_1 are known with limited precision only. We can estimate their respective errors using error propagation technique (assuming a_0 and a_1 to be independent):

$$\sigma_{a_0}^2 = \sum_{j=1}^{N} \left(\frac{\partial a_0}{\partial y_j} \right)^2 \sigma_j^2 = \frac{S_{xx}}{\Delta}$$

$$\sigma_{a_1}^2 = \sum_{j=1}^{N} \left(\frac{\partial a_1}{\partial y_j} \right)^2 \sigma_j^2 = \frac{S}{\Delta}$$

The variances above correspond respectively to the diagonal elements $(\alpha^{-1})_{11}$ and $(\alpha^{-1})_{22}$.

Generalizations

Nonlinear χ^2 minimization schemes in which one minimizes the generalized function

$$\chi^2 = \sum_{j=1}^{N} \left(\frac{\tilde{y}_j - f\left(x_j | \vec{a}\right)}{\sigma_j} \right)^2$$

where f can be a nonlinear function (e.g., exponential, logarithm, trigonometric, etc.) carrying multiple parameters. These schemes use advanced multidimensional minimization approaches (some resembles multidimensional root-finding), e.g., the Gauss-Newton and Levenberg-Marquardt methods.

Generalizations

The least-squares method here introduced for linear fits can be readily extended to polynomials in which the model function can be generalized as

$$f(x) = a_0 + a_1 x + \cdots + a_m x^m = \sum_{j=1}^{m} a_j x^j$$

and a matrix form system of equations of the kind $\vec{a} = \alpha^{-1}\vec{b}$ can also be obtained. The same scheme works if a subset of the polynomial basis function is not linear, e.g., $f(x) = \sum_{j=0}^{m} a_j \phi_j(x)$ which could be $f(x) = a_0 + a_1 e^{-x^2}$ (still linear in the a's).

It is also possible to show that α^{-1} is the *error* or *covariance matrix* as

$$\alpha^{-1} = \begin{bmatrix} \sigma_{a_1}^2 & \sigma_{a_1 a_2}^2 \cdots & \sigma_{a_1 a_m}^2 \\ \sigma_{a_2 a_1}^2 & \ddots & \vdots \\ \vdots & & \\ \sigma_{a_m a_1}^2 & \cdots & \sigma_{a_m}^2 \end{bmatrix}$$

Generalizations

The matrix above contains the estimates of the parameter variances along the diagonal, and the covariances in the off-diagonal and through the error propagation definitions (assuming independence between measurements), we can write:

$$\hat{\sigma}_{a_k}^2 = \sum_{j=1}^{N} \left(\frac{\partial a_k}{\partial y_j} \right)^2 \sigma_j^2$$

$$\hat{\sigma}_{a_k a_l}^2 = \sum_{j=1}^{N} \frac{\partial a_k}{\partial y_j} \frac{\partial a_l}{\partial y_j} \sigma_j^2$$

But note that for polynomials high-order m, the matrix α is prone to become **ill conditioned**, and its inversion may become numerically unstable. This can be fixed using methods that rely on **orthogonal polynomials** transformations (Chebyshev, Legendre, Jacobi, …).

Multi-variable Linear Regression

Linear regression can be expanded to account for the case in which a dataset with N measurements of $m+1$ independent variables such as

$$\left(y_i \middle| x_i^{[1]}, \ldots, x_i^{[m]}\right) \text{ for } i = 1, \ldots, N$$

The model to fit is then generalized as

$$y(\vec{x}) = a_0 + a_1 x^{[1]} + a_2 x^{[2]} + \cdots + a_m x^{[m]} = a_0 + \sum_{k=1}^{m} a_k x^{[k]}$$

The χ^2 minimization scheme is generalized as

$$\chi^2 = \sum_{j=1}^{N} \left(\frac{\tilde{y}_j - f\left(x_j | \vec{a}\right)}{\sigma_j} \right)^2 = \sum_{j=1}^{N} \left(\frac{\tilde{y}_j - a_0 - \sum_{k=1}^{m} a_k x_j^{[k]}}{\sigma_j} \right)^2$$

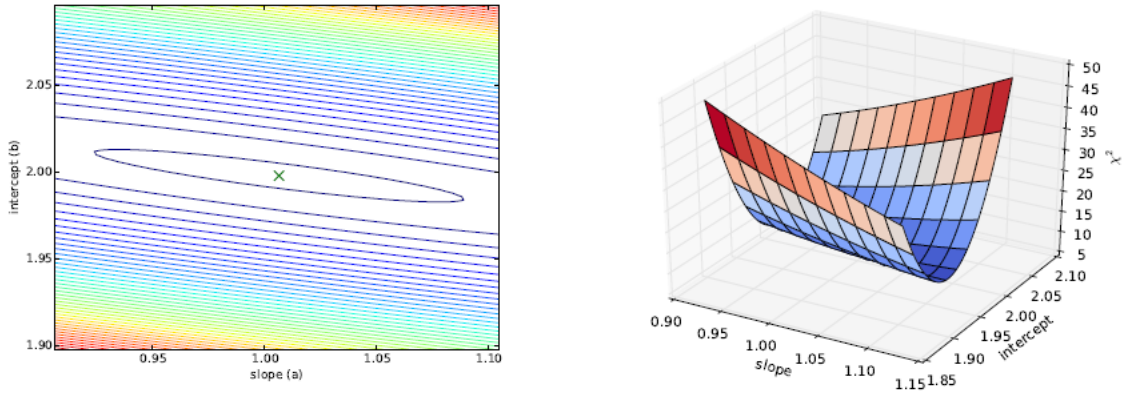
Note that such fitting procedure requires significance tests to determine the quality of the fit! Examples are t-test (for the model components), F-test for the significance of the parameters, and coefficient of determination.

IMPORTANT:

The linear regression discussed here in analytical form assumed that the model is linear in the fitting parameters. This 'linearity' is not a requirement in which χ^2 minimization shown above can be applied for nonlinear models provided that each variable y_i is Gaussian distributed.

The main complication for nonlinear functions is that an analytic solution for the best-fit values and the errors is in general no longer available. This limitation can be overcome with the use of sophisticated numerical methods to minimize the χ^2 statistic. A "brute force" scheme to achieve that is to construct an m-dimensional grid of all possible parameter values, evaluate χ^2 at each point, and then find the global minimum; those would be the parameters regarded as the best estimate for the model.

This direct grid-search method can become unfeasible as the number of free parameters increases. Moreover, to find the parameter uncertainties using grid-search requires a knowledge of the expected variable of χ^2 around the minimum. Among the methods that can be used to estimate the parameters and their covariance matrix is the *Markov Chain Monte Carlo Technique.*



(a) contours of constant χ^2. (b) surface

Figure 2.5: χ^2 landscape for straight line fit to 100 data points.

Gradient Descent

Have function $\chi^2(\vec{a})$ and we want to

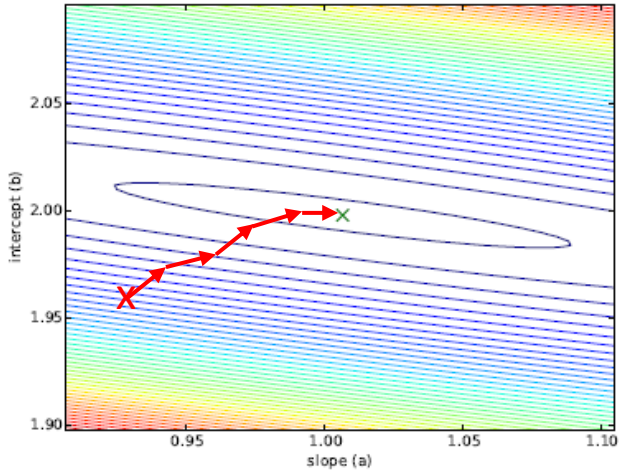
$$\min_{\vec{a}} \chi^2(\vec{a})$$

- Give an initial condition to the set of parameters $\{\vec{a}\}$
- Keep changing $\{\vec{a}\}$ to minimize χ^2 until reach the minimum (up to a certain tolerance)

$$a_0 \coloneqq a_0 + \gamma \frac{\partial \chi^2(a_0, a_1)}{\partial a_0}$$

$$a_1 \coloneqq a_1 + \gamma \frac{\partial \chi^2(a_0, a_1)}{\partial a_1}$$

γ: learning rate
Normally a sufficiently small number.



(a) contours of constant χ^2.

More advanced Gradient Descent methods such as **'Adam' (Adaptive Moment Estimation)** use a varying (adaptive) learning rate which helps in the minimization of more complicated functions.