Data fitting

1 Least-squares fit.

We often wish to determine one characteristic y of an experiment as a function of some other quantity x. That is,instead of making a number of measurements of a single quantity x, we make a series of N measurements of the pair (x_i, y_i), one for each of several values of the index i, which runs from 1 to N. Our object is to find a function $y = y(x)$ that describes the relation between these two measured variables. In this chapter we consider the problem of pairs of variables (x_i, y_i) that are linearly related to one another.

1.1 Correlated variables

Correlation is a description of a relationship between variables. A challenge in measuring correlation is that the variables we want to compare might not be expressed in the same units. For example, height might be in centimeters and weight in kilograms. And even if they are in the same units, they come from different distributions. There are two common solutions to these problems:

1. Transform all values to standard scores. This leads to the *Pearson* coefficient of correlation.

2. Transform all values to their percentile ranks. This leads to the *Spearman* coefficient.

If X is a series of values, x_i , we can convert to standard scores by subtracting the mean and dividing by the standard deviation: $z_i = (x_i - \mu)/\sigma$.

The numerator is a deviation: the distance from the mean. Dividing by σ normalizes the deviation, so the values of Z are dimensionless (no units) and their distribution has mean 0 and variance 1. If X is normally-distributed, so is Z; but if X is skewed or has outliers, so does Z. In those cases it is more robust to use percentile ranks. If R contains the percentile ranks of the values in X, the distribution of R is uniform between 0 and 100, regardless of the distribution of X.

1.2 Covariance

Covariance is a measure of the tendency of two variables to vary together. If we have two series, X and Y, their deviations from the mean are $dx_i = x_i - \mu_x$, $dy_i = y_i - \mu_y$ where μ_x and μ_y are the mean of X and Y. if X and Y vary together, their deviations tend to have the same sing.

If we multiply them together, the product is positive when the deviations have the same sign and negative when they have the opposite sign. So adding up the products gives a measure of the tendency to vary together. Covariance is the mean of these products:

$$Cov(X, Y) = \frac{1}{n} \sum dx_i dy_i \tag{1}$$

where n is the length of the two series.

Covariance is useful in some computations, but it is seldom reported as a summary statistic because it is hard to interpret. Among other problems, its units are the product of the units of X and Y. So the covariance of weight and height might be in units of kilogram-meters, which does not mean much.

Covariance of two variables

```
# Compute the covariance of the two variables in each cov1.dat, cov2.dat, and cov3.dat files.

reset  # This clear all variables in memory
import numpy as np

c1=np.loadtxt('cov1.dat')
dx=c1[0,:]-np.mean(c1[0,:])
dy=c1[1,:]-np.mean(c1[1,:])
cov=1/len(dx)*np.sum(dx*dy)

# Check that the covariance cov(x,x) is the variance of x
print(cov=1/len(dx)*np.sum(dx*dx))
print(np.var(c1[0,:]))
```

1.3 Correlation

One solution to this problem is to divide the deviations by σ, which yields standard scores, and compute the product of standard scores:

$$p_i = \frac{(x_i - \mu_x)}{\sigma_x}\frac{(y_i - \mu_y)}{\sigma_y} \tag{2}$$

The mean of these products is $\rho = \frac{1}{n}\sum p_i$, which can be rewrite as:

$$\rho = \frac{Cov(X,Y)}{\sigma_x \sigma_y} \tag{3}$$

This value is called *Pearson‘s correlation* after Karl Pearson, an influential early statistician. It is easy to compute and easy to interpret. Because standard scores are dimensionless, so is ρ.

Pearson‘s correlation is always between -1 and +1 (including both). The magnitude indicates the strength of the correlation. If $\rho = 1$ the variables are perfectly correlated, which means that if you know one, you can make a perfect prediction about the other. The same is true if $\rho = 1$. It means that the variables are negatively correlated, but for purposes of prediction, a negative correlation is just as good as a positive one. Pearson‘s correlation is a measure of how much better is a correlation.

So if $\rho = 0$, does that mean there is no relationship between the variables? Unfortunately, no. Pearson‘s correlation only measures linear relationships. If there is a nonlinear relationship, ρ understates the strength of the dependence. Figure 1 shows scatter plots and correlations for several carefully-constructed datasets.

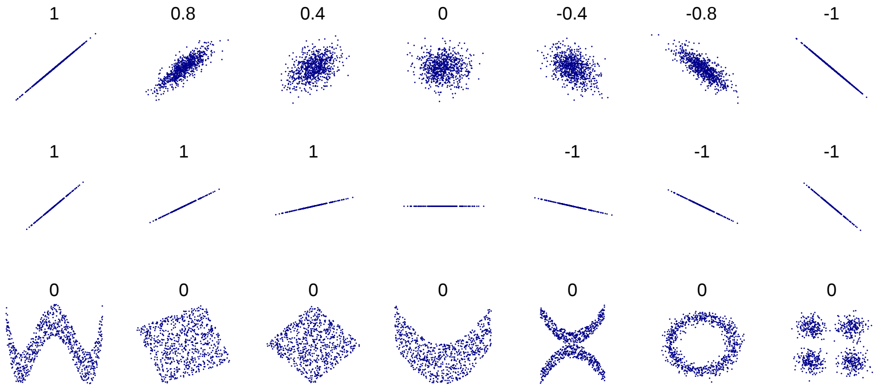

Figure 1: Scatter plots and correlations for several data sets.

The top row shows linear relationships with a range of correlations; you can use this row to get a sense of what different values of ρ look like. The second row shows perfect correlations with a range of slopes, which demonstrates that correlation is unrelated to slope. The third row shows variables that are clearly related, but because the relationship is non-linear, the correlation coefficient is 0. The moral of this story is that you should always look at a scatter plot of your data before blindly computing a correlation coefficient.

Correlation of two variables

```
# Make a plot and compute the correlation of the two variables in each cov1.dat, cov2.dat, and
    cov3.dat files.

reset  # This clear all variables in memory
import numpy as np
import matplotlib.pyplot as plt
plt.ion()
c1=np.loadtxt('cov1.dat')
plt.plot(c1[0,:],c1[1,:],'.b',linewidth=2)

dx=c1[0,:]-np.mean(c1[0,:])
dy=c1[1,:]-np.mean(c1[1,:])
cov=1/len(dx)*np.sum(dx*dy)
cor1=cov/(np.std(c1[0,:])*np.std(c1[1,:]) )
print(cor1)

#cor1=-0.051
#cor2=-0.995
#cor3=-1.0
#cor4=-0.069
```

1.4 Least-squares fit to a straight line.

Correlation coefficients measure the strength and sign of a relationship, but not the slope. There are several ways to estimate the slope; the most common is a linear least squares fit. A "linear fit"is a line intended to model the relationship between variables. A "least squares"fit is one that minimizes the mean squared error (MSE) between the line and the data.

Suppose we have a sequence of points, Y, that we want to express as a function of another sequence X. If there is a linear relationship between X and Y with intercept α and slope β, we expect each y_i to be roughly $\alpha + \beta x_i$. But unless the correlation is perfect, this prediction is only approximate. The deviation, or *residual*, is:

$$\epsilon_i = (\alpha + \beta x_i) - y_i.$$

The residual might be due to random factors like measurement error, or non-random factors that are unknown. If we get the parameters α and β wrong, the residuals get bigger, so it makes intuitive sense that the parameters we want are the ones that minimize the residuals. As usual, we could minimize the absolute value of the residuals, or their squares, or their cubes, etc. The most common choice is to minimize the sum of squared residuals:

$$min(\sum \epsilon_i^2).$$

There are good reasons to do this: 1) Squaring has the obvious feature of treating positive and negative residuals the same, which is usually what we want, 2) Squaring gives more weight to large residuals, but not so much weight that the largest residual always dominate, and 3) If the residuals are independent of x, random, and normally distributed with $\mu = 0$ and constant (but unknown) σ, then the least squares fit is also the maximum likelihood estimator of α and β.

Computing a least squares fit is quick, easy and often good enough. This is how it goes:

1. Compute the sample means, $\bar{x}$ and $\bar{y}$, the variance of X, $Var(X)$, and the covariance of X and Y, $Cov(X,Y)$.
2. The estimate of the slope is:
   $$\hat{\beta} = \frac{Cov(X,Y)}{Var(X)}$$
3. And the intercept is:
   $$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x}$$

1.5 Goodness of fit.

Having fit a linear model to the data, we might want to know how good it is. One way to evaluate a model is its predictive power. In the context of prediction, the quantity we are trying to guess is called a dependent variable and the quantity we are using to make the guess is called an explanatory or independent variable. To measure the predictive power of a model, we can compute the coefficient of determination, more commonly known as *R-squared*:

$$R^2 = 1 - \frac{Var(\epsilon)}{Var(Y)}$$

The term $Var(\epsilon)/Var(Y)$ is the ratio of mean squared error with and without the explanatory variable, which is the fraction of variability left unexplained by the model. The complement, R^2, is the fraction of variability explained by the model. If a model yields $R^2 = 0.64$, you could say that the model explains 64% of the variability, or it might be more precise to say that it reduces the MSE or your predictions by 64%.

In the context of linear least squares model, it turns out that there is a simple relationship between the coefficient of determination and Pearson‘s correlation coefficient, ρ: $R^2 = \rho^2$.

1.6 Least-squares fit to a polynomial.

Arrays can be used for least-squares fit, making very easy to extend the method to polynomials. Let’s use the least-squares method to fit a power-series polynomial, which can be expresed as:

$$y(x) = \sum_{k=1}^{m} a_k f_k(x) \tag{4}$$

where the functions $f_k(x)$ could be the powers of x ($f_1(x) = 1, f_2(x) = x, f_3(x) = x^2, \ldots$) or they could be other function of x as long as they *do not involve the parameters* $a_1, a_2, a_3, \ldots$. With this definition, χ^2 becomes:

$$\chi^2 = \sum \left[\frac{1}{\sigma_i} \left[y_i - \sum_{k=1}^{m} a_k f_k(x_i) \right] \right]^2 \tag{5}$$

where σ_i are the uncertainties in the observations y_i.

The method of least squares requires that we minimize χ^2, our measurement of the goodness of the fit to de data, with respect to the parameters $a_1, a_2, a_3, \ldots$. The minimum is determined by taking partial derivatives with respect to each parameter in the expression for χ^2 in Eq. 5, and setting them to zero. We obtain a set of m coupled linear equations for the m parameters a_l, with the index l running from 1 to m. this system that can be solved by using a matrix method. The expression in matrix form is: $\boldsymbol{\beta} = \boldsymbol{a\alpha}$, where the elements of the row matrix $\boldsymbol{\beta}$ are defined by:

$$\beta_k \equiv \sum \left[\frac{1}{\sigma_i^2} y_i f_k(x_i) \right] \tag{6}$$

those of the symetric matrix $\boldsymbol{\alpha}$ by:

$$\alpha_{lk} \equiv \sum \left[\frac{1}{\sigma_i^2} f_l(x_i) f_k(x_i) \right] \tag{7}$$

and the elements of the row matrix $\boldsymbol{a}$ are the parameters of the fit. To solve for the parameter matrix $\boldsymbol{a}$ we multiply by $\epsilon = \alpha^{-1}$, which gives:

$$\boldsymbol{a} = \boldsymbol{\beta\epsilon} \tag{8}$$

The solution of Eq. 8 requires that the matrix $\boldsymbol{\alpha}$ be inverted. The symmetric matrix $\boldsymbol{\alpha}$ is called *curvature matrix* because of its relation to the curvature of the χ^2 function in parameter space. The diagonal elements of the square matrix $\boldsymbol{\epsilon}$, called the error matrix ore the covariance matrix, are the variances of the parameters a_k and the off diagonal elements are the covariances: $\sigma^2_{a_l} = \alpha^{-1}_{ll}$, $\sigma^2_{a_{lk}} = \alpha^{-1}_{lk}$. The average uncertainty in the estimation of $y(x)$ is given by:

$$\boldsymbol{\sigma^2} = \frac{1}{N-m}\sum_{i=1}^{N}(y_i - y(x_i))^2 \tag{9}$$

where N is the number of data and m the number of free parameters.
We can also calculate uncertainty for a predicted value $y(x_i)$ using the error matrix. For example, for $m = 3$ the uncertainty for a predicted value $y_p = a_1 + a_2x_p + a_3x_p^2$ is given by:

$$s^2 = 1 \cdot \epsilon_{11} + x_p^2 \cdot \epsilon_{22} + x_p^4 \cdot \epsilon_{33} + 2(x_p \cdot \epsilon_{12} + x_p^2 \cdot \epsilon_{13} + x_p^3 \cdot \epsilon_{23}) \tag{10}$$

Where $\epsilon_{11}, \epsilon_{22}, \epsilon_{33}$ are de variances of a_1, a_2, a_3, and $\epsilon_{12}, \epsilon_{13}, \epsilon_{23}$ are the covariant terms in the symmetric error matrix.

As an example for the case of $m = 2$, we have the following equations:

$$\chi^2 = \sum\left[\frac{1}{\sigma_i}(y_i - a - bx_i)\right]^2 \tag{11}$$

$$a = \frac{1}{\Delta}\begin{vmatrix} \sum\frac{y_i}{\sigma_i^2} & \sum\frac{x_i}{\sigma_i^2} \\ \sum\frac{x_iy_i}{\sigma_i^2} & \sum\frac{x_i^2}{\sigma_i^2} \end{vmatrix} \tag{12}$$

$$b = \frac{1}{\Delta}\begin{vmatrix} \sum\frac{1}{\sigma_i^2} & \sum\frac{y_i}{\sigma_i^2} \\ \sum\frac{x_i}{\sigma_i^2} & \sum\frac{x_iy_i}{\sigma_i^2} \end{vmatrix} \tag{13}$$

$$\Delta = \begin{vmatrix} \sum \frac{1}{\sigma_i^2} & \sum \frac{x_i}{\sigma_i^2} \\ \sum \frac{x_i}{\sigma_i^2} & \sum \frac{x_i^2}{\sigma_i^2} \end{vmatrix} \tag{14}$$

The estimated uniform variance is given by:

$$\sigma^2 \simeq s^2 = \frac{1}{N-2} \sum (y_i - \bar{y})^2 \tag{15}$$

where $\bar{y} = a + bx$

The uncertainties in coefficients are given by:

$$\sigma_a^2 = \frac{1}{\Delta} \sum \frac{x_i^2}{\sigma_i^2}; \quad \sigma_b^2 = \frac{1}{\Delta} \sum \frac{1}{\sigma_i^2} \tag{16}$$

Also, for $m = 3$, we have:

$$\chi^2 = \sum \left[\frac{1}{\sigma_i} (y_i - a_1 - a_2 x - a_3 x^2) \right]^2 \tag{17}$$

$$a_1 = \frac{1}{\Delta} \begin{vmatrix} \sum y_i \frac{1}{\sigma_i^2} & \sum \frac{x_i}{\sigma_i^2} & \sum \frac{x_i^2}{\sigma_i^2} \\ \sum y_i \frac{x_i}{\sigma_i^2} & \sum \frac{x_i^2}{\sigma_i^2} & \sum \frac{x_i^3}{\sigma_i^2} \\ \sum y_i \frac{x_i^2}{\sigma_i^2} & \sum \frac{x_i^3}{\sigma_i^2} & \sum \frac{x_i^4}{\sigma_i^2} \end{vmatrix} \tag{18}$$

$$a_2 = \frac{1}{\Delta}\begin{vmatrix} \sum \frac{1}{\sigma_i^2} & \sum y_i \frac{1}{\sigma_i^2} & \sum \frac{x_i^2}{\sigma_i^2} \\ \sum \frac{x_i}{\sigma_i^2} & \sum y_i \frac{x_i}{\sigma_i^2} & \sum \frac{x_i^3}{\sigma_i^2} \\ \sum \frac{x_i^2}{\sigma_i^2} & \sum y_i \frac{x_i^2}{\sigma_i^2} & \sum \frac{x_i^4}{\sigma_i^2} \end{vmatrix} \tag{19}$$

$$a_3 = \frac{1}{\Delta}\begin{vmatrix} \sum \frac{1}{\sigma_i^2} & \sum \frac{x_i}{\sigma_i^2} & \sum y_i \frac{1}{\sigma_i^2} \\ \sum \frac{x_i}{\sigma_i^2} & \sum \frac{x_i^2}{\sigma_i^2} & \sum y_i \frac{x_i}{\sigma_i^2} \\ \sum \frac{x_i^2}{\sigma_i^2} & \sum \frac{x_i^3}{\sigma_i^2} & \sum y_i \frac{x_i^2}{\sigma_i^2} \end{vmatrix} \tag{20}$$

$$\Delta = \begin{vmatrix} \sum \frac{1}{\sigma_i^2} & \sum \frac{x_i}{\sigma_i^2} & \sum \frac{x_i^2}{\sigma_i^2} \\ \sum \frac{x_i}{\sigma_i^2} & \sum \frac{x_i^2}{\sigma_i^2} & \sum \frac{x_i^3}{\sigma_i^2} \\ \sum \frac{x_i^2}{\sigma_i^2} & \sum \frac{x_i^3}{\sigma_i^2} & \sum \frac{x_i^4}{\sigma_i^2} \end{vmatrix} \tag{21}$$

2 Nonlinear functions

In all the procedures present here we have assumed that the fitting function was linear in the coefficients. By that we mean that the function can be expressed as a sum of separate terms each multiplied by a single coefficient. How can we fit data with a function that is not linear in the coefficients? For example, how can we solve a function like $P(t) = (1/\tau)e^{-t/\tau}$ using the least-squares method?. The method of least-squares does not yield a straightforward analytical solution for such functions. Here we consider approximate solutions to such problems using linear-regression techniques.

2.1 Linearization

It is possible to transform some functions into linear functions. For example, if we were to fit an exponential decay problem of the form:

$$y = ae^{-bx} \tag{22}$$

where a and b are the unknown parameters, it would seem reasonable to take logarithms of both sides and to fit the resulting straight line equation

$$\ln y = \ln a - bx \tag{23}$$

The method of least squares minimizes the value of χ^2 with respect to each of the coefficients $\ln a$ and $\ln b$ where χ^2 is given by

$$\chi^2 = \sum \left(\frac{1}{\sigma_i'^2} (\ln y_i + \ln a - bx_i)^2 \right) \tag{24}$$

where we must use weighted uncertainties σ_i' instead of σ_i to account for the transformation of the dependent variable:

$$\sigma_i' = \frac{d(\ln y_i)}{dy} \sigma_i = \frac{1}{y_i} \sigma_i \tag{25}$$

In general, if we fit the function $f(y)$ rather than y, the uncertainties σ_i in the measured quantities must be modified by

$$\sigma_i' = \frac{df(y)}{dy_i} \sigma_i \tag{26}$$

2.2 Errors in the parameters

If we modify the fitting function so that instead of fitting the data points y_i with the coefficient $a, b, \ldots$, we fit modified data points $y'_i = f(y_i)$ with coefficients $a', b', \ldots$, then our estimates of the errors in the coefficients will pertain to the uncertainties in the modified coefficients $a', b', \ldots$, rather than to the desired coefficients $a, b, \ldots$. If the relationship between the two sets of coefficients is defined to be $a' = f_a(a)$ and $b' = f_b(b)$, then the correspondence between the uncertainties $\sigma'_a, \sigma'_b, \ldots$ in the modified coefficients and the uncertainties $\sigma_a, \sigma_b, \ldots$ in the desired coefficients is obtained in a manner similar to that for σ'_i and σ_i in Eq. 26

$$\sigma'_a = \frac{df_a(a)}{da}\sigma_a \; ; \quad \sigma'_b = \frac{df_b(b)}{db}\sigma_b \tag{27}$$

Thus, if the modified coefficient is $a' = \ln a$, the estimated error in a is determined form the estimated error in a', acording to Eq. 27 with $f_a = \ln a$:

$$\sigma'_a = \frac{d(\ln a)}{da}\sigma_a = \frac{\sigma_a}{a} \tag{28}$$

Values of χ^2 for testing the goodness of fit should be determined from the original uncertainties of the data σ_i and from the unmodified equation, although Eq. 24 should give approximately equivalent results when weighted with the modified uncertainties σ'_i.

3 Exercises.

- Download the files in folder "Data for exercises chapter #3". Plot the data and find the correlation coefficients for those dataset for it makes sense. Using the least-squares method, find the parameter for a linear o polynomial correlation. Find the average uncertainties of each fit.