

Relation between Two Continuous Variables

So far we have been dealing with problems where only one variable is measured. Expressions or functions which only depend on one variable are sometimes called *univariate*. If more than one variable is involved, we are dealing with *multivariate* problems. In the simplest case we have two variables involved, and we need a *bivariate* data analysis.

For two related variables, the *correlation* measures the association between the two variables. In contrast, a *linear regression* is used for the prediction of the value of one variable from another. For the correlation between many variables, you should look into “correlation tables” (nicely implemented in *seaborn*). And an extension of linear regression to more than two variables brings you into the realm of statistical modeling.

Correlation

Correlation Coefficient

The *correlation coefficient* between two variables answers the question: “Are the two variables related? I.e., if the two variables are normally distributed, the standard measure of determining the *correlation coefficient*, often ascribed to *Pearson*, is

$$r = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^n (X_i - \bar{X})^2} \sqrt{\sum_{i=1}^n (Y_i - \bar{Y})^2}}$$

With

$$s_{xy} = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{n - 1}$$

and s_x, s_y the sample standard deviations of the x and y values, respectively, this can also be written as

$$r = \frac{s_{xy}}{s_x \cdot s_y}.$$

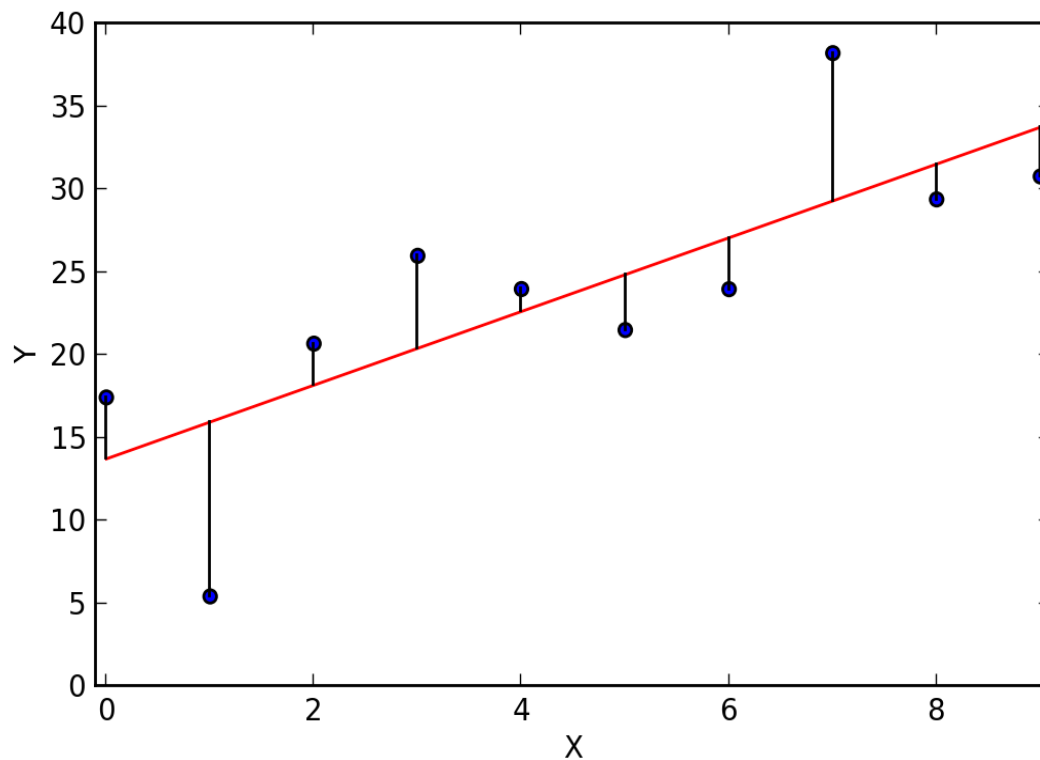
Pearson's correlation coefficient, sometimes also referred to as *population correlation coefficient* or *sample correlation*, can take any value from -1 to $+1$. Examples are given in the Figure below. Note that the formula for the correlation coefficient is symmetrical between x and y .

Coefficient of determination

In order to interpret r , let me first define a few common terms.

Residuals

Differences between observed values and predicted values.



Best-fit linear regression line (red) and residuals (black).

A data set has values y_i , each of which has an associated modelled value f_i (also sometimes referred to as \hat{y}_i). Here, the values y_i are called the *observed values*, and the modelled values f_i are sometimes called the *predicted values*.

In the following \bar{y} is the mean of the observed data:

$$\bar{y} = \frac{1}{n} \sum_{i=1}^n y_i$$

where n is the number of observations.

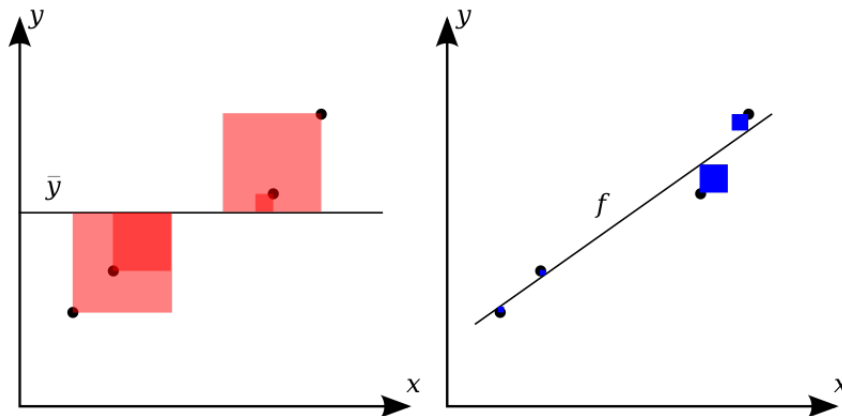
The “variability” of the data set is measured through different sums of squares:

$SS_{\text{tot}} = \sum_i (y_i - \bar{y})^2$, the total sum of squares (proportional to the sample variance);

$SS_{\text{mod}} = \sum_i (f_i - \bar{y})^2$, the sum of squares of the model values, also called the explained sum of squares;

$SS_{\text{res}} = \sum_i (y_i - f_i)^2$, the sum of squares of residuals, also called the residual sum of squares.

The notations SS_R and SS_E should be avoided, since in some texts their meaning is reversed to “Residual sum of squares” and “Explained sum of squares”, respectively.



The better the linear regression (on the right) fits the data in comparison to the simple average (on the left graph), the closer the value of R^2 is to one. The areas of the blue squares represent the squared residuals with respect to the linear regression. The areas of the red squares represent the squared residuals with respect to the average value (from Wikipedia)

With these expressions, the most general definition of the *coefficient of determination*, R^2 , is

$$R^2 \equiv 1 - \frac{SS_{\text{res}}}{SS_{\text{tot}}}.$$

Since

$$SS_{\text{tot}} = SS_{\text{mod}} + SS_{\text{res}}$$

Therefore the equation above is equivalent to

$$R^2 = \frac{SS_{\text{mod}}}{SS_{\text{tot}}}$$

For simple linear regression (i.e. line-fits), the *coefficient of determination* or R^2 is the square of the correlation coefficient r . It is easier to interpret than the correlation coefficient r : values of R^2 close to 1 are good, values close to 0 are poor. Note that for general models it is common to write R^2 , whereas for simple linear regression r^2 is used.

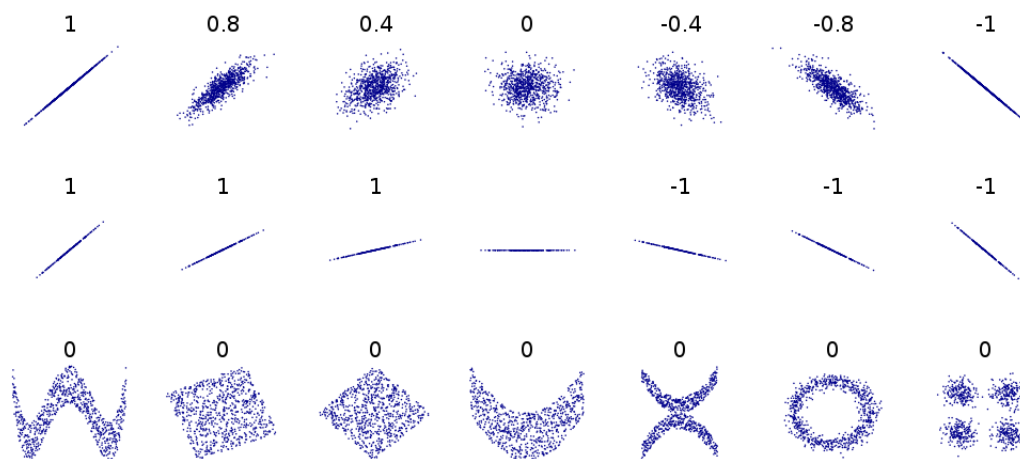
Relation to unexplained variance

In a general form, R^2 can be seen to be related to the unexplained variance, since the second term compares the unexplained variance (variance of the model's errors) with the total variance (of the data).

Examples

How large R^2 or \bar{R}^2 must be to be considered good depends on the discipline. They are usually expected to be larger in the physical sciences than it is in biology or the social sciences. In finance or marketing, it also depends on what is being modeled.

Caution: the sample correlation and R^2 are misleading if there is a nonlinear relationship between the independent and dependent variables!



Several sets of (x, y) points, with the correlation coefficient of x and y for each set. Note that the correlation reflects the non-linearity and direction of a linear relationship (top row), but not the slope of that relationship (middle), nor many aspects of nonlinear relationships (bottom). N.B.: the Figure in the center has a slope of 0 but in that case the correlation coefficient is undefined because the variance of Y is zero. (From: Wikipedia)

Rank correlation

If the data distribution is not normal, a different approach is necessary. In that case one can rank the set of subjects for each variable and compare the orderings. There are two commonly used methods of calculating the rank correlation.

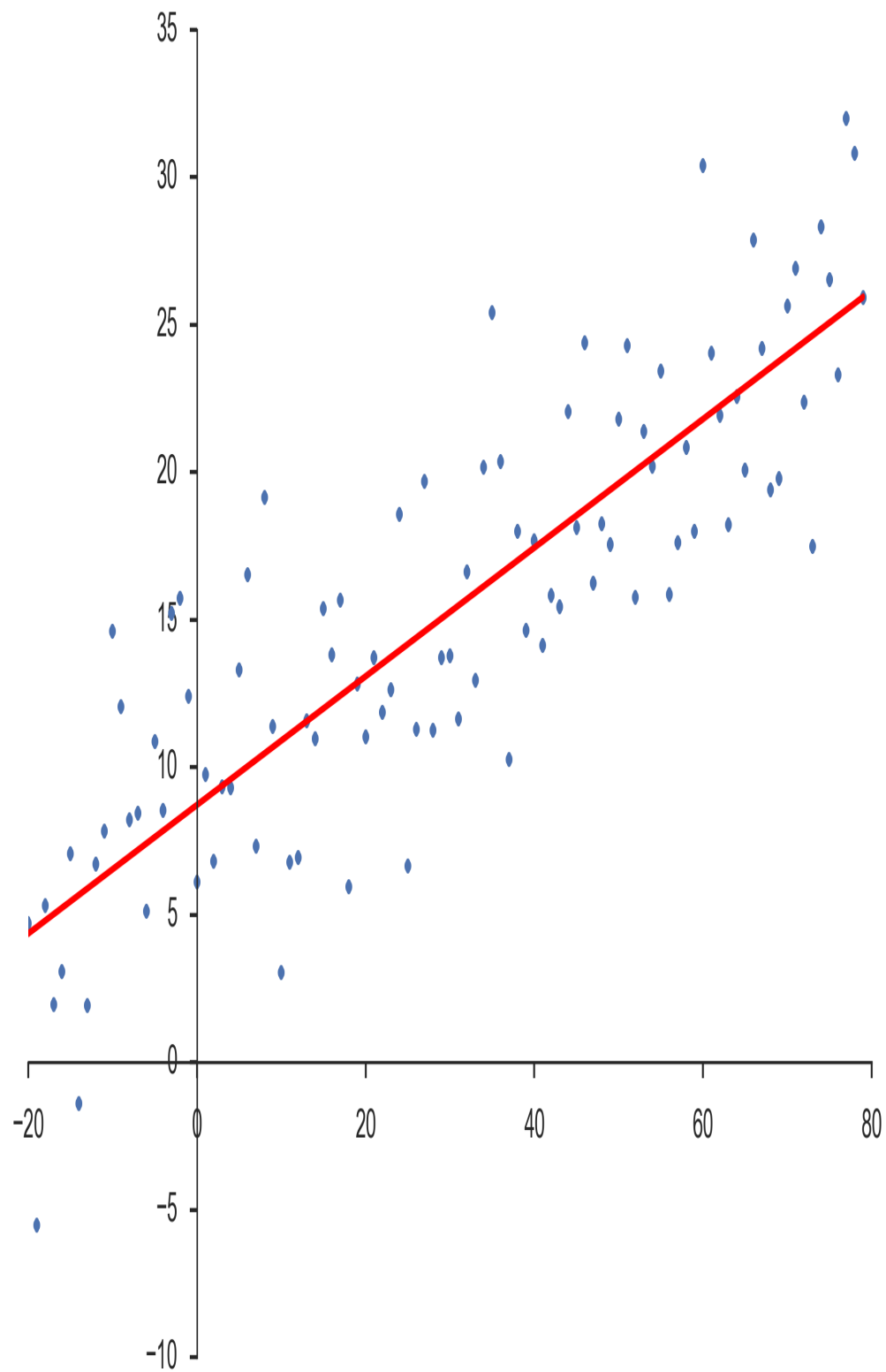
- *Spearman's ρ* , which is exactly the same as the Pearson correlation coefficient r calculated on the ranks of the observations.
- *Kendall's τ* . is also a rank correlation coefficient, measuring the association between two measured quantities. It is harder to calculate than Spearman's rho, but it has been argued that confidence intervals for Spearman's rho are less

reliable and less interpretable than confidence intervals for Kendall's tau-parameters.

Regression

General linear regression model

We can use the method of *regression* when we want to predict the value of one variable from the other.



Linear regression. (From Wikipedia)

When we search for the best-fit line to a given (x_i, y_i) dataset, we are looking for the parameters (k, d) which minimize the sum of the squared *residuals* ϵ_i in

$$y_i = k * x_i + d + \epsilon_i$$

where k is the *slope* or *inclination* of the line, and d the *intercept*. This is in fact just the one-dimensional example of the more general technique, which is described in the next section. Note that in contrast to the correlation, this relationship between x and y is no more symmetrical: it is assumed that the x -values are known exactly, and that all the variability lies in the residuals.

Simple Regression

Example of *simple linear regression* with 7 observations. Suppose there are 7 data points $\{y_i, x_i\}$, where $i = 1, 2, \dots, 7$. The simple linear regression model is

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i,$$

where β_0 is the y-intercept and β_1 is the slope of the regression line. This model can be represented in matrix form as

$$\begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \\ y_5 \\ y_6 \\ y_7 \end{bmatrix} = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ 1 & x_3 \\ 1 & x_4 \\ 1 & x_5 \\ 1 & x_6 \\ 1 & x_7 \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \epsilon_3 \\ \epsilon_4 \\ \epsilon_5 \\ \epsilon_6 \\ \epsilon_7 \end{bmatrix}$$

where the first column of ones in the design matrix represents the y-intercept term while the second column is the x-values associated with the y-value.

Design Matrix

Quadratic Fit

The equation for a quadratic fit to the given data is

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \epsilon_i,$$

This can be rewritten in matrix form:

$$\begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \\ y_5 \\ y_6 \\ y_7 \end{bmatrix} = \begin{bmatrix} 1 & x_1 & x_1^2 \\ 1 & x_2 & x_2^2 \\ 1 & x_3 & x_3^2 \\ 1 & x_4 & x_4^2 \\ 1 & x_5 & x_5^2 \\ 1 & x_6 & x_6^2 \\ 1 & x_7 & x_7^2 \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \epsilon_3 \\ \epsilon_4 \\ \epsilon_5 \\ \epsilon_6 \\ \epsilon_7 \end{bmatrix}$$

General Formulation

In general, this can be rewritten in matrix form as:

$$y = X\beta + \epsilon,$$

the matrix X is the *design matrix*.

Y is a vector of dimension $(n \times 1)$ and is called the endogenous variable, X is a matrix of dimension $(n \times k)$ where each column is an explanatory variable and ϵ is the error term. β is the vector of dimension $(k \times 1)$ and contains the parameters we want to estimate.

Coding

If you have vectors x, y containing your data, you can use *statsmodels* to create a design matrix that also includes the 1 's for the offset:

```
import statsmodels.api as sm
Xmat = sm.add_constant(x)
```

The parameters are then easily found as

```
params = np.linalg.lstsq(Xmat, y)
```

However, you get a lot more information if you use the OLS-fit from *statmodels*:

```
import numpy as np
import statsmodels.api as sm

# Generate artificial data
nobs = 100
X = np.random.random(nobs)
X = sm.add_constant(X)
beta = [5, 3.5]
e = np.random.random(nobs)
y = np.dot(X, beta) + e

# Fit regression model
results = sm.OLS(y, X).fit()
```



```
# Inspect the results
print(results.summary())
```

yields the following results:

```

                                OLS Regression Results
=====
Dep. Variable:                  y      R-squared:
Model:                        OLS      Adj. R-squared:
Method:                    Least Squares  F-statistic:
Date:                Fri, 04 Jul 2014  Prob (F-statistic):
Time:                14:49:08      Log-Likelihood:
No. Observations:                100      AIC:
Df Residuals:                    98      BIC:
Df Model:                        1
=====
               coef      std err          t      P>|t|      [ 95.0%
-----
const           5.4410        0.059     92.685      0.000        5.32
x1              3.5718        0.104     34.250      0.000        3.36
=====
Omnibus:                 21.620    Durbin-Watson:
Prob(Omnibus):            0.000    Jarque-Bera (JB):
Skew:                    0.223    Prob(JB):
Kurtosis:                1.908    Cond. No.
=====

```

The meaning of many of these parameters is described in the chapter on “Statistical Models”.

From the *results*, you can extract e.g. the model parameters, standard errors, confidence intervals, and residuals:

```
params = results.params
std_err = results.bse
ConfInt = results.conf_int()
residuals = results.resid
```

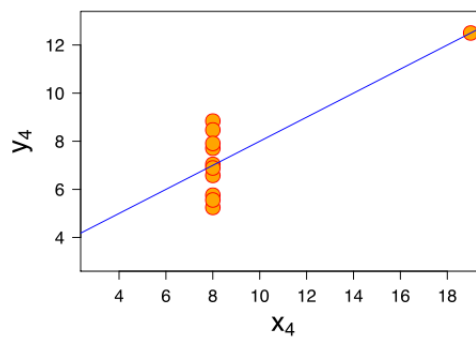
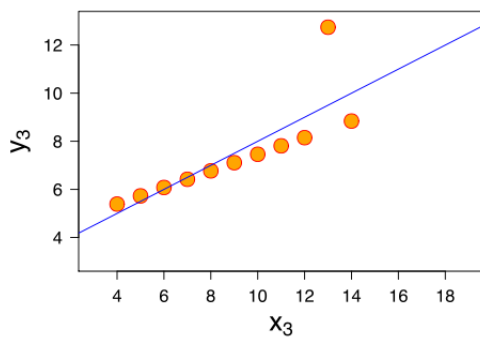
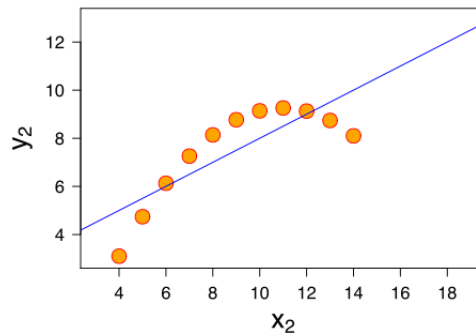
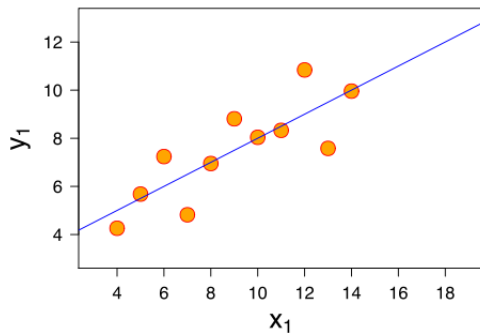
Assumptions

To use the technique of linear regression, the following assumptions should be fulfilled:

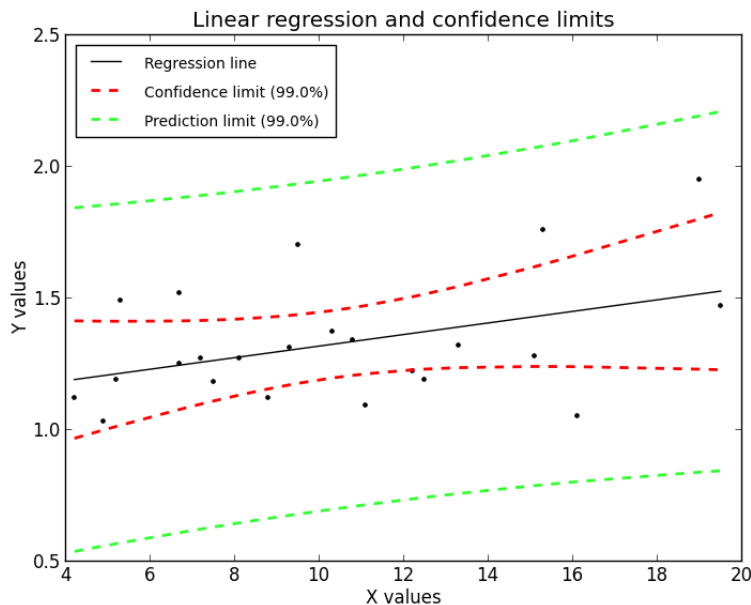
1. The *independent variables* (i.e. x) are exactly known.
2. Validity. Most importantly, the data you are analyzing should map to the research question you are trying to answer. This sounds obvious but is often overlooked or ignored because it can be inconvenient. For example, a linear regression does not properly describe a quadratic curve.
3. Additivity and linearity. The most important mathematical assumption of the regression model is that its deterministic component is a linear function of the separate predictors.

4. Independence of errors.
5. Equal variance of errors.
6. Normality of errors.

IP[y] [80_multivariate.ipynb](#)



The sets in the Anscombe's quartet have the same linear regression line but are themselves very different.



Regression, with confidence intervals for the mean, as well as for the predicted data. The red dotted line shows the confidence interval for the mean; and the green dotted line the confidence interval for predicted data. (This can be compared to the standard error and the standard deviation for a population.)

Since to my knowledge there exists no program in the Python standard library (or numpy, scipy) to calculate the confidence intervals for a regression line, I include my corresponding program *fitLine.py*. The output of this program is shown in the figure below. This program also shows how Python programs intended for distribution should be documented.

Exercises

1. Correlation

Read in the data for the average yearly temperature at the Sonnblick, from https://github.com/thomas-haslwanter/statsintro/blob/master/Data/data_others/AvgTemp.xls. Calculate the Pearson and Spearman correlation, and Kendall's tau, for the temperature vs. year.

2. Regression

For the same data, calculate the yearly increase in temperature, assuming a linear increase with time. Is this increase significant?

3. Normality Check

For the data from the regression model, check if the model is ok by testing if the residuals are normally distributed (e.g. by using the

Komogorov–Smirnov test)



[4] This section has been taken from Wikipedia