Error Estimation

Source: *Richter “Estimating Errors in Least Square Fitting”* [*http://ipnpr.jpl.nasa.gov/progress\_report/42-122/122E.pdf*](http://ipnpr.jpl.nasa.gov/progress_report/42-122/122E.pdf)

The problem:

Data of the form is fitted by the function .

To determine the coefficients , **it is sought to minimize**

where is the standard deviation of the random errors of (which are assumed to be normally distributed).

is a coefficient vector that minimizes .

The variances of the elements are given by : the diagonal elements of the covariance matrix which is itself the inverse of the curvature or Hessian matrix .

The *weighted mean value of the variance of the fit is given by:* so that for constant data errors, **the mean standard error of the fit is**:

The **error in the value of the fitted function** is a function of even when the are all the same and independent of . Hence variance of the value of the fitted function (due to random data errors) is:

Where

For *linear* fitting where , and where is a column vector with elements .

**Standard Error of Fit:**

What is , the standard error of the fit?

Figure : Illustration of standard error of fit .

As is apparent from Figure 1, the standard error of fit is a very useful term! It is minimum at the centroid of the data points and maximum at the edges.

**Least Squares Fitting:**

can be obtained from knowledge of experimental errors or from analysis of the data itself. It is assumed that the errors are normally distributed.

**Linear Least Squares Fitting:**

are arbitrary basis functions for the independent variable .

In terms of linear algebra, the fitting function is an -element column vector, the coefficients are elements of an M element column vector and the basis functions are elements of an matrix . Hence:

i.e.

Lets now define the column vector and the matrix . Hence:

Extremum condition is: i.e.[[1]](#footnote-1)

Hence:

Here and is the covariance matrix: a symmetric matrix.

Hence:

**Suitability of Basis Functions:**

The value of should be of the order : the number of **degrees of freedom** of this system. Hence

This is the condition for the fit to be meaningful. If , is normally distributed with having mean 1 and standard deviation of .

**Standard Errors**

In matrix form:

The covariance of and is given by:

Since (errors are uncorrelated), hence: **.** Hence[[2]](#footnote-2),

Hence, the **errors in the coeffients are correlated**!

The variance of the coefficients are given by the **diagonal** elements:

Hence:

Hence the covariance:

This is independent of parameter values . For the special case where :

Here is a column vector whose elements are .

Hence, the **errors in**  **at two different values of are correlated!**

The **weighted mean value of over all**  is given by:

But . Hence:

Hence:

Hence if ,

This is analogous to the variance of the mean for the case of one value of where M=1. Hence, as a result of fitting to the function , **the variance of y is reduced by .**  Hence the higher the value of required to fit the data to reduce to a value close to 1, the larger will be the errors in the values of the fitted function.

**Very Important: For the fit to be meaningful:**

1. Data errors ( have to be correctly chosen otherwise fit is not meaningful even with .

Only if the fit is meaningful can one estimate the fitting errors meaningfully.

**Non-Linear Least Squares Problem**

No explicit solution for exists and an iterative procedure must be followed.

Where

And similarly:

Here:

If we define:

Then: **,** just like for linear least squares. The derivation then[[3]](#footnote-3), on similar lines, leads to:

Hence also,

i.e.

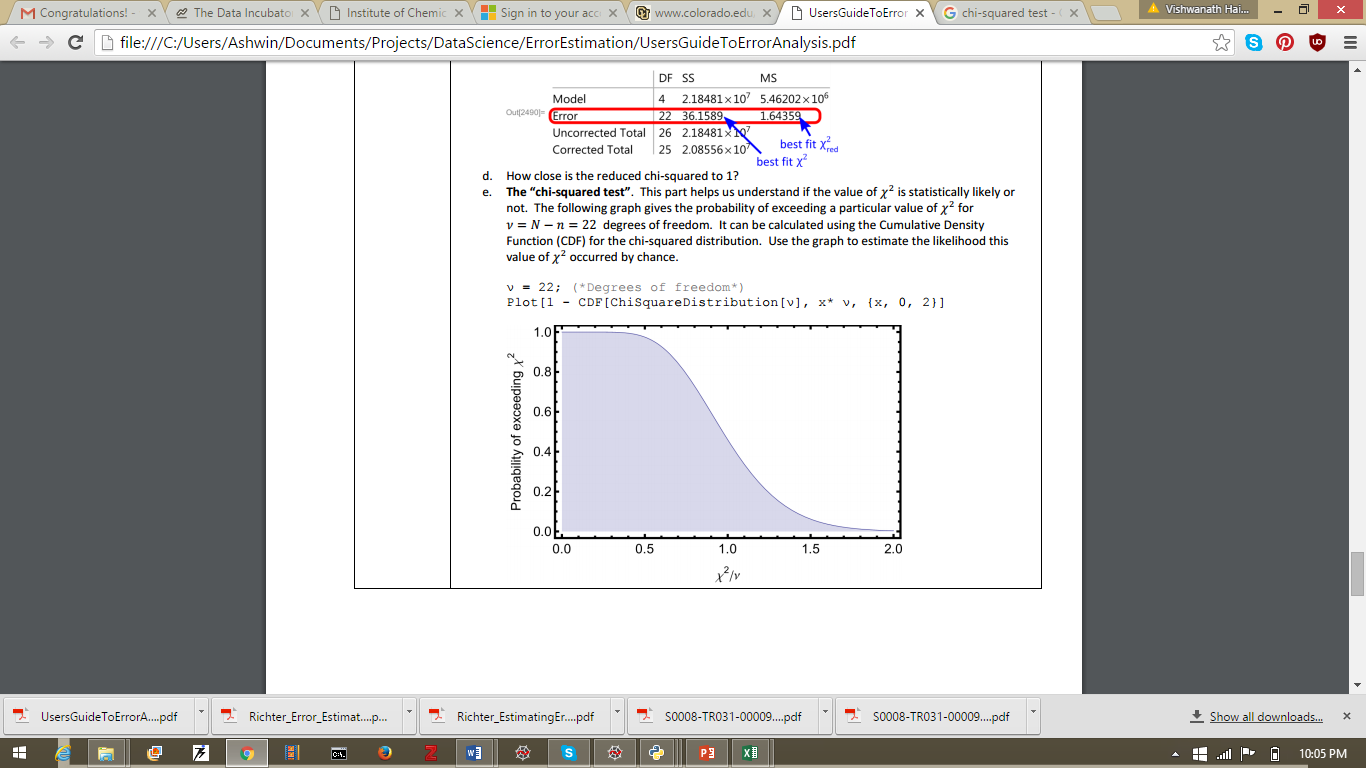
**The standard error of regression is:**

is the 68.3% confidence limit

is the 95.4% confidence limit

is the 99.73% confidence limit

**Note:** Notice that and are not functions of This is because the former are only meaningful if i.e. the dependence is already baked in.



40% probability of exceeding 1.0 by random chance

Very low probability of exceeding 2.0 by random chance

Good chance of exceeding by random fluctuations

**Null Hypothesis:** The fit explains the data.

indicates a poor model fit: probability that the value of occurred by random chance is very low.

indicates that fit has not fully captured the data but null hypothesis cannot be rejected.

indicates that the match between observation and estimate is in accord with the error variance. Null hypothesis cannot be rejected.

indicates *over-fitting* of data i.e. model is (improperly) fitting noise or the error variance has been over-estimated. Null hypothesis cannot be rejected.

**Shapiro-Wilk** test can be used to see if the residuals are normally distributed. If they are, then the student-t distribution can be used to verify that the mean is statistically zero.

# Getting p-value for coefficient .

The statistic to test if a given coefficient is different from zero is:

. Get the value of and if this is less the 0.05, then the value of is significantly different from zero.

User’s Guide to Error Analysis

<http://www.colorado.edu/physics/EducationIssues/zwickl/Resources/Error%20Analysis%20Activity%202.pdf>

What is least squares fitting:

Where there are data points and the fit function is given by where is a vector of the fit parameters.

**Assumptions:**

1. **Gaussian Distribution** of the random fluctuations in each .
2. **Uncorrelated:** the random fluctuations in any one data point are uncorrelated with another data point.

If these two assumptions hold, minimizing gives the *most likely* function that reproduces the observed data.

Uncertainty in the data is calculated using the **residuals of the best fit**:

This is analogous to finding the variance of a repeated measurement.

**Note:** This was not calculated in Richter’s analysis.

The residuals must be randomly distributed about zero (no systematic variation).

**Note:** It is **not necessary** for a good fit line to pass through each set of error bars: though it should pass through most. If we have shown the error bars for standard error, then 68% of data points are expected to lie within their error bar.

**Applying the chi-squared test:**

The graph of the probability of exceeding the observed value for degrees of freedom is available (cumulative density function CDF of the chi-squared distribution). This gives us the likelihood that the observed value of occurred by chance.

The **null-hypothesis** for a chi-squared test is that the data are **independent** and **normally-distributed.**  The **chi-squared** test for goodness-of-fit is used to *reject* the null hypothesis that the data are independent.

**Note:** It is possible to get a good looking by overestimating experimental errors (.

Interpreting Sum of Squares

<http://facweb.cs.depaul.edu/sjost/csc423/documents/f-test-reg.htm>

<http://reliawiki.org/index.php/Simple_Linear_Regression_Analysis>

For observations and regression parameters:

1. Corrected Sum of Squares also called Sum of Squares of Regression :

This is a measure of **explained variation**

1. Sum of Squares for Error:

This is a meaure of **unexplained variation**

1. Corrected Sum of Squares (Total):
2. For multiple regression models:
3. Corrected Degrees of Freedom for Model:
4. Degrees of Freedom for Error:
5. Corrected Degrees of Freedom Total:
6. Mean of Squares of Model:
7. Mean of Squares for Error:
8. Mean of Squares Total:

**It is desirable to have MSM be large wrt MSE**.

# F-test

We want to test the following null hypothesis:

Steps of the F-test.

1. State the null and alternative hypothesis.
2. Compute test statistics assuming null-hypothesis is true:
3. Find a (1-a)\*100% confidence interval I for (DFM, DFE) degrees of freedom using an F-table or statistical software.
4. Accept the null-hypothesis if ; else reject.
5. Use statistical software to determine the value.

Implementation in Python

To implement this in Python, lets open a new file called *errorestimation.py*.

We need to import several things. See Figure 2

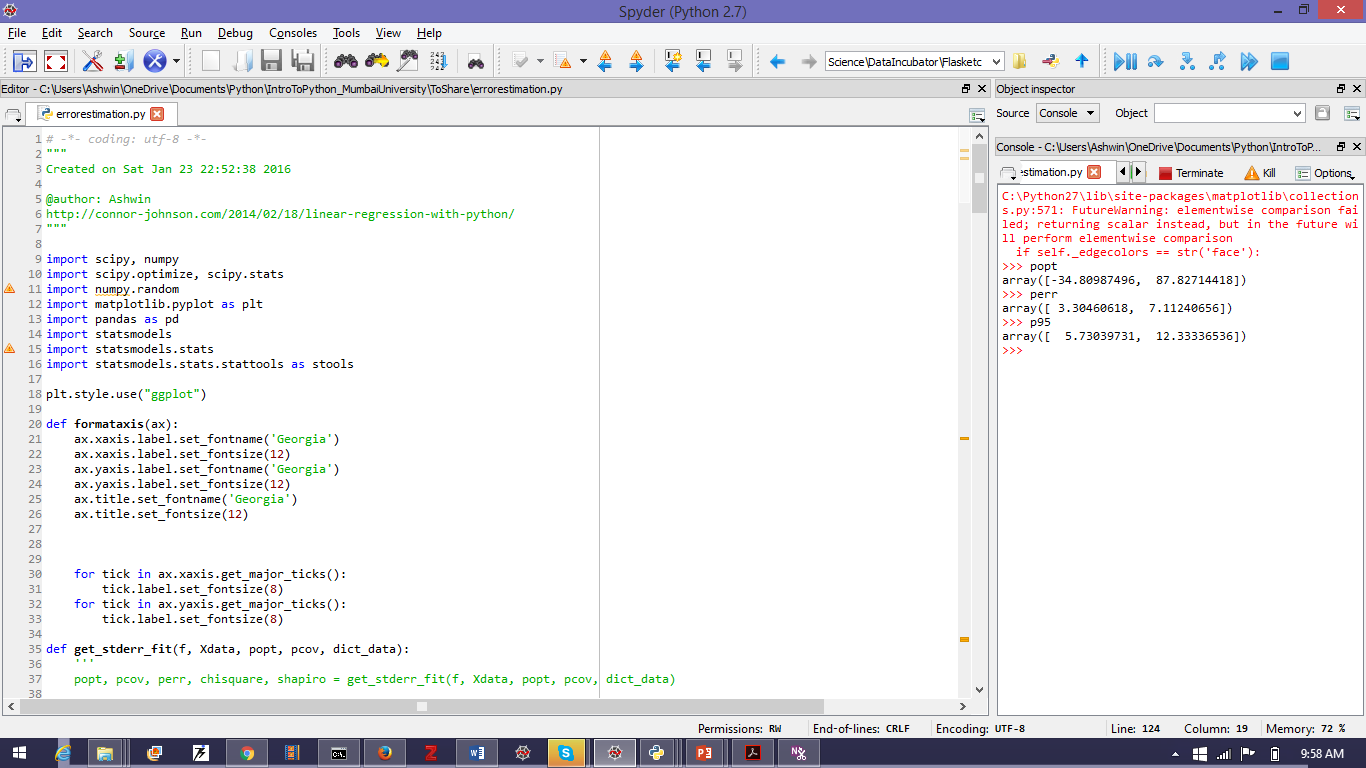


Figure : Various libraries that we need to import for analysis of a curve-fitting.

We define a function *fitdata* as shown in Figure 3. The inputs of the function are explained in the docstring (the green text in the triple quotes).

We will use the Levenberg-Marquadt algorithm to optimize the **p** (see discussion at the beginning of this chapter). This is a *non-linear* least square fitting.

The code corresponding to this is *within* the *fitdata* function’s block. It appears in Figure 4. Sums of squares of various deviations appear in Figure 5 and analysis of goodness of fit appears in Figure 6.

The code for the plotting appears in Figure 7. The calculation of the standard error of the fit appears in Figure 8 and the code for formatting the axes in Figure 9.

See Figure 10, Figure 11 and Figure 12 for an illustration of importing, preparing and fitting data from an Excel sheet.

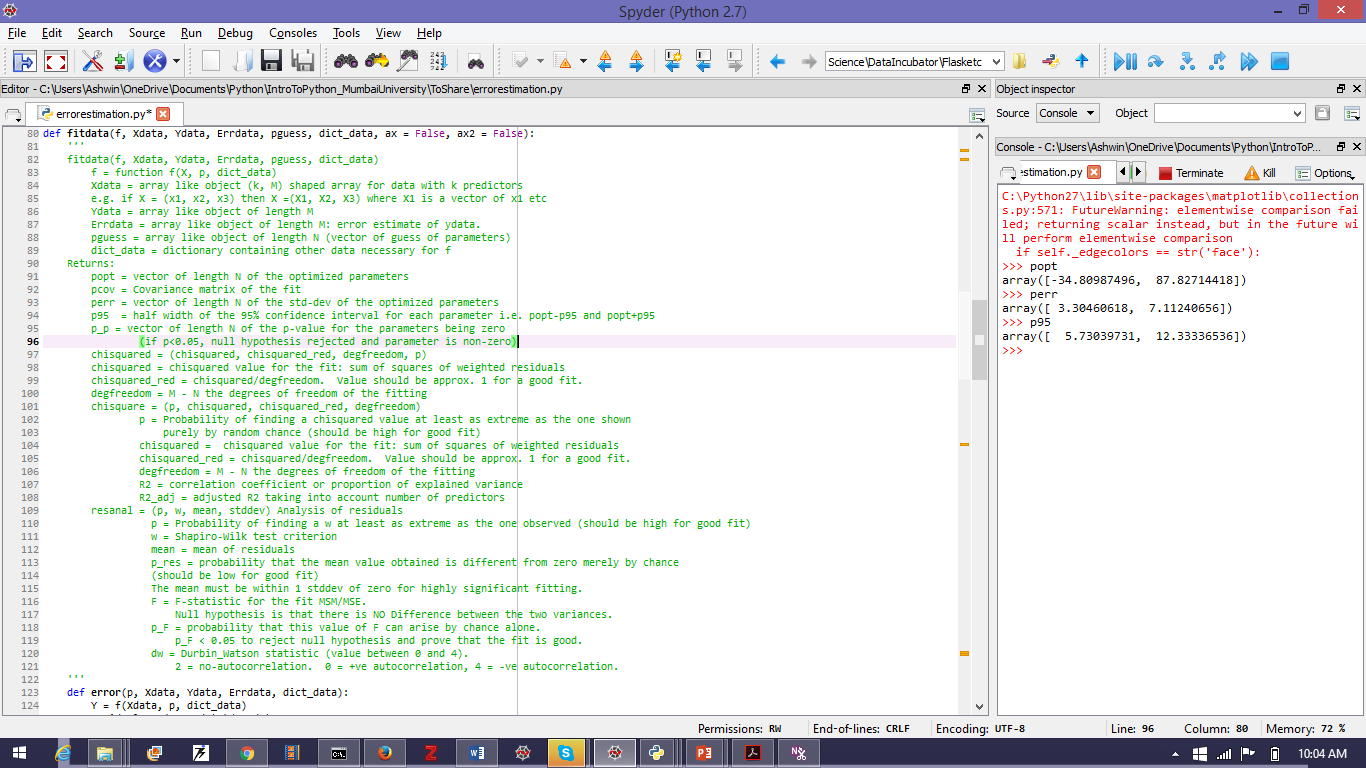


Figure : Function fitdata and its docstring explaining its inputs and outputs.

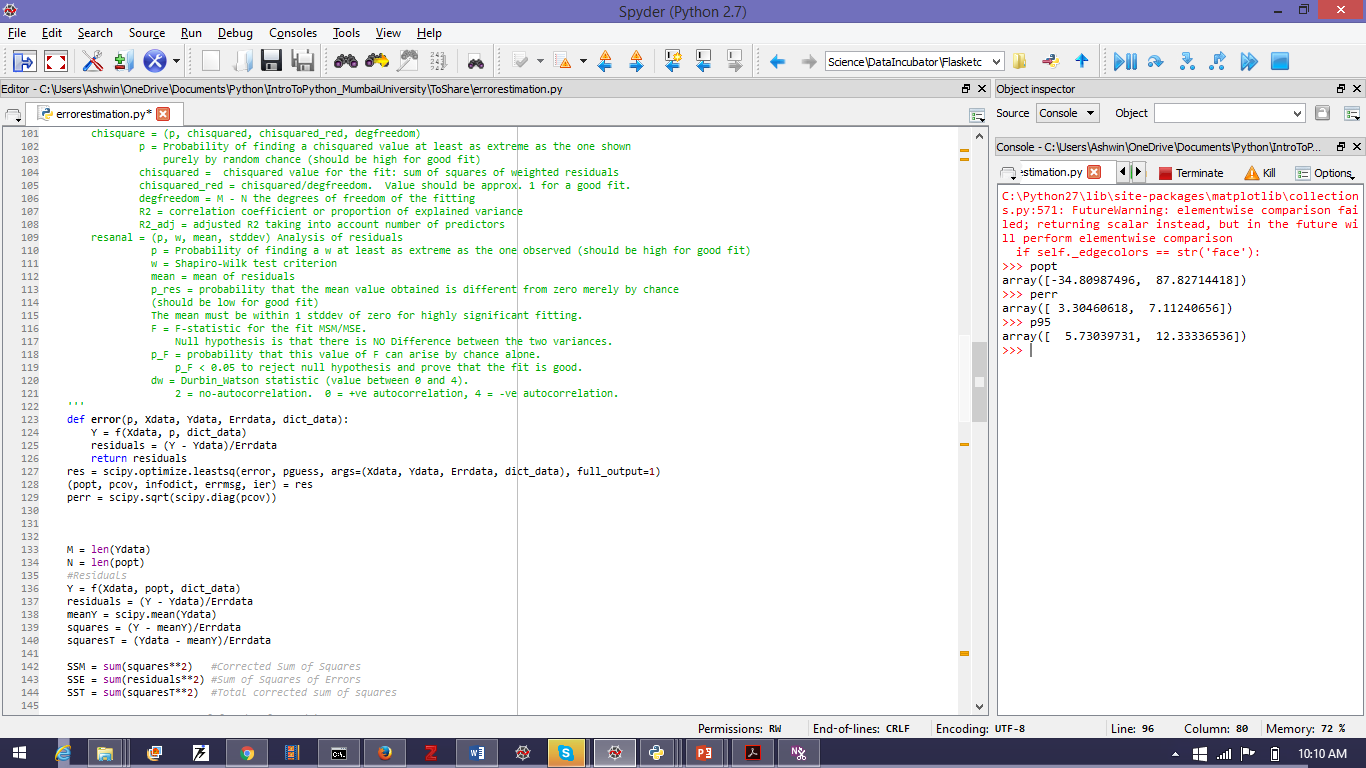


Figure : Code to optimize the parameters **p** using the scipy.optimize.leastsq module with full\_output = 1. Also shown is **perr:** the vector of the standard-deviation for each of the **p** values.

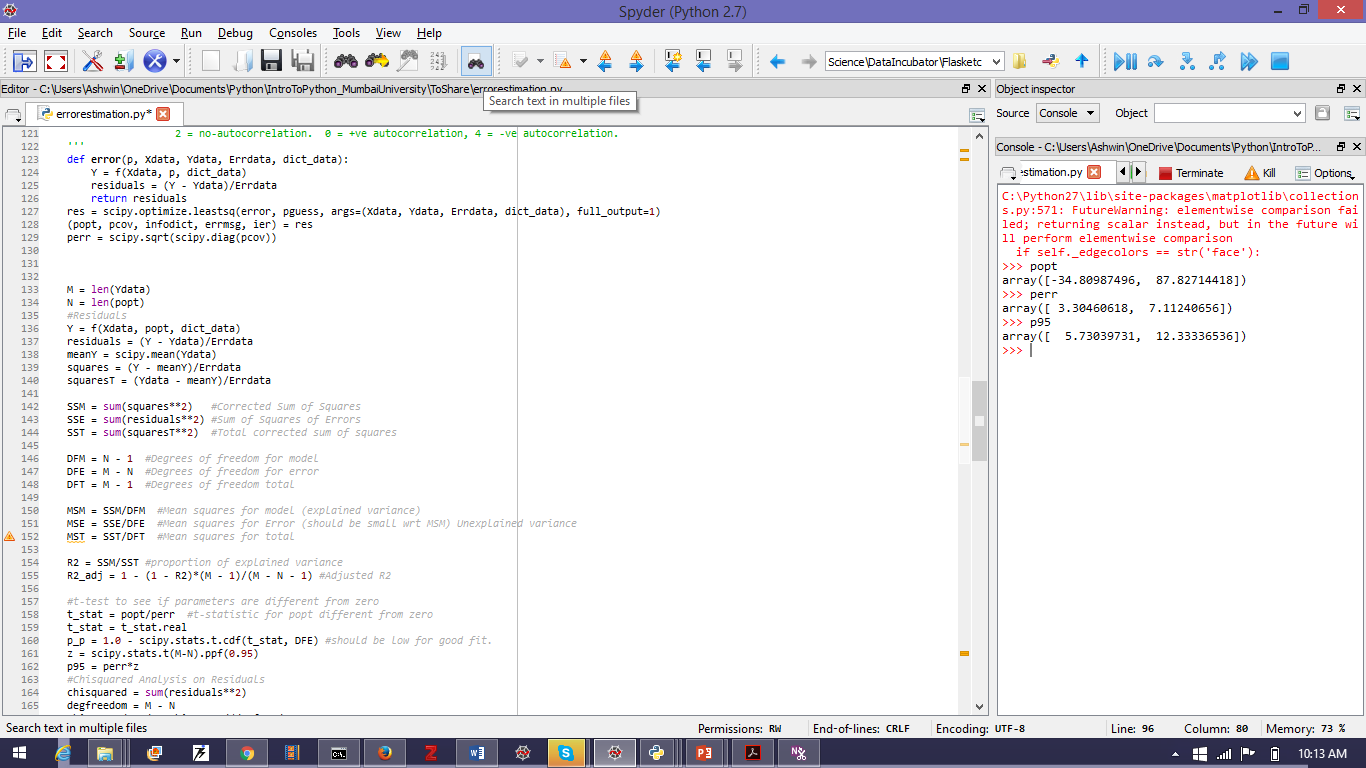


Figure : Code to analyse the sum of squares of various deviations of the fit. See the part about the F-test above.

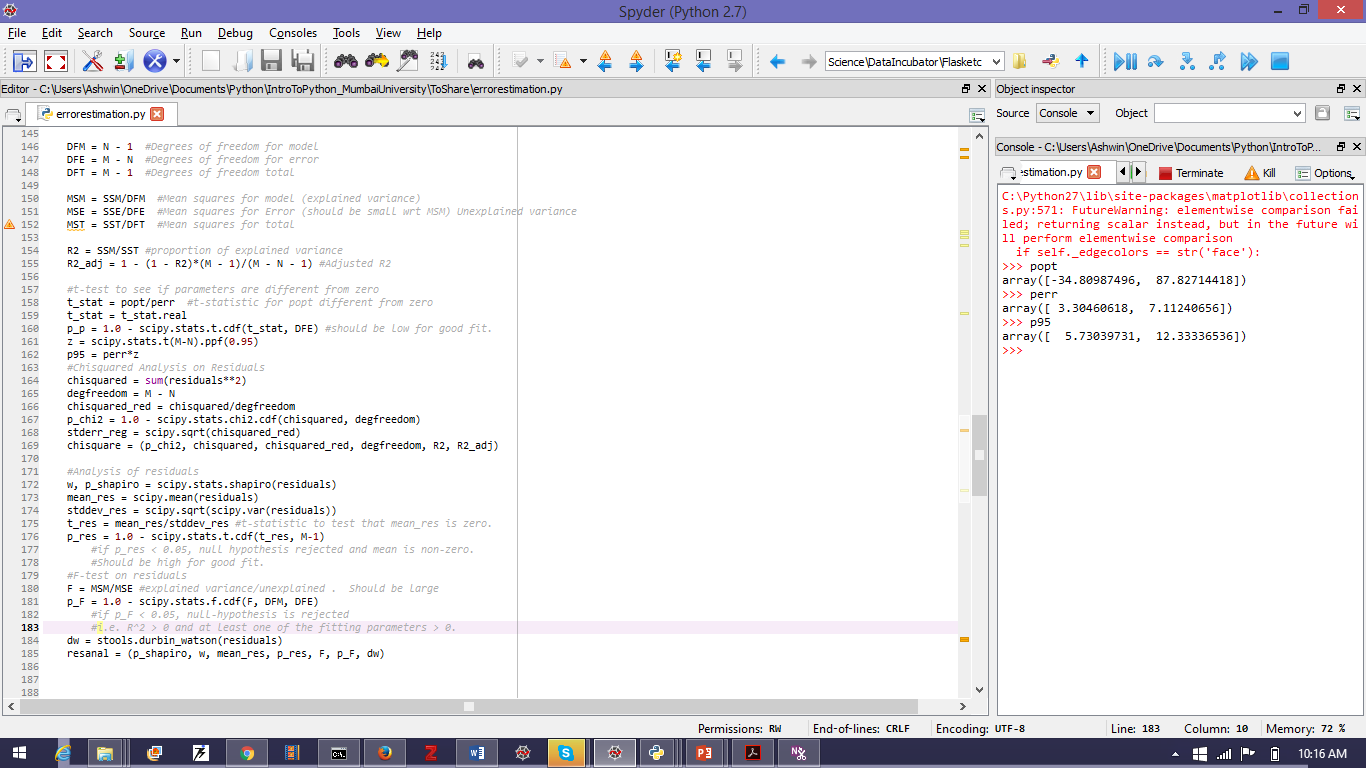


Figure : Test to see goodness of fit. is the vector of the 95% confidence range of p i.e. p = p with 95% confidence. Shapiro test is to check if the residuals are normally distributed and the Durbin-Watson test is to check if they are correlated. p\_shapiro should be > 0.05 and dw should be near 2 and away from 0 and 4.

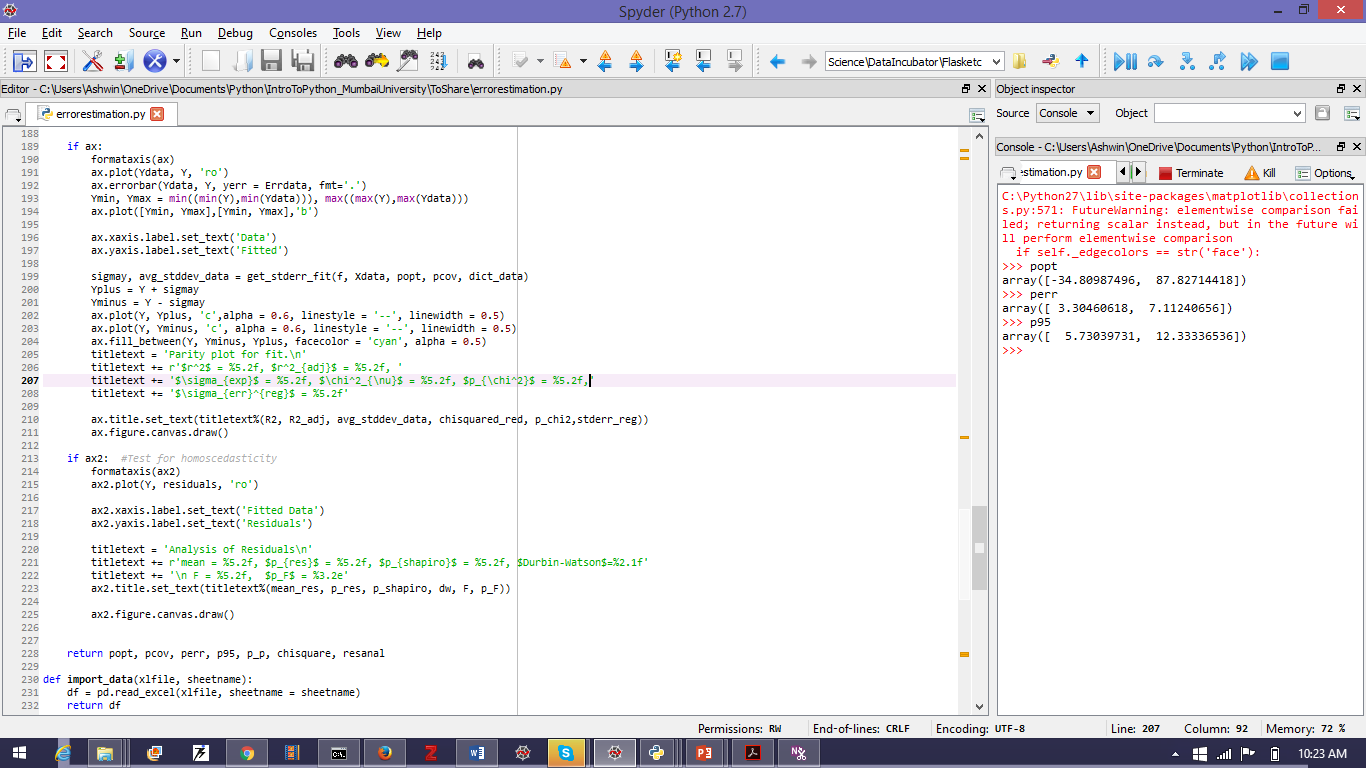


Figure : Plotting and return

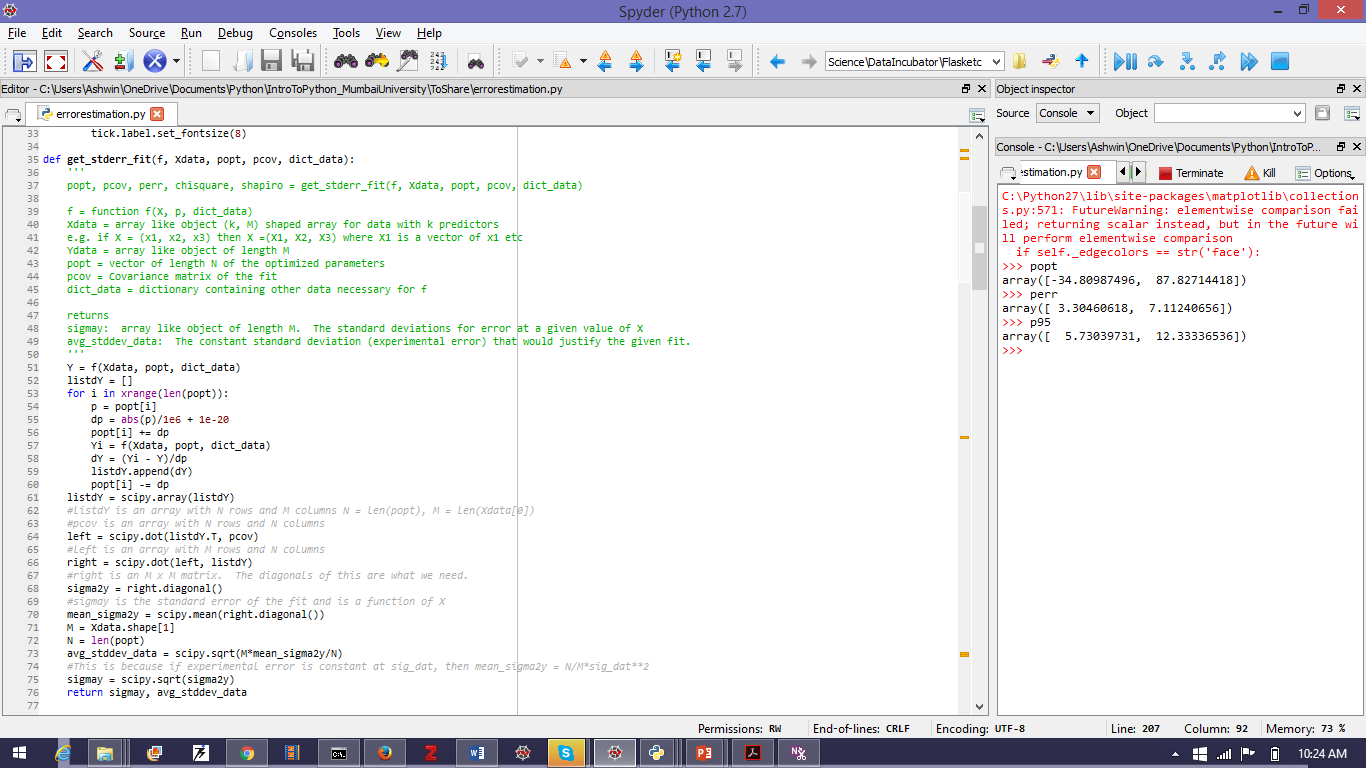


Figure : Code to calculate standard error of the fit.

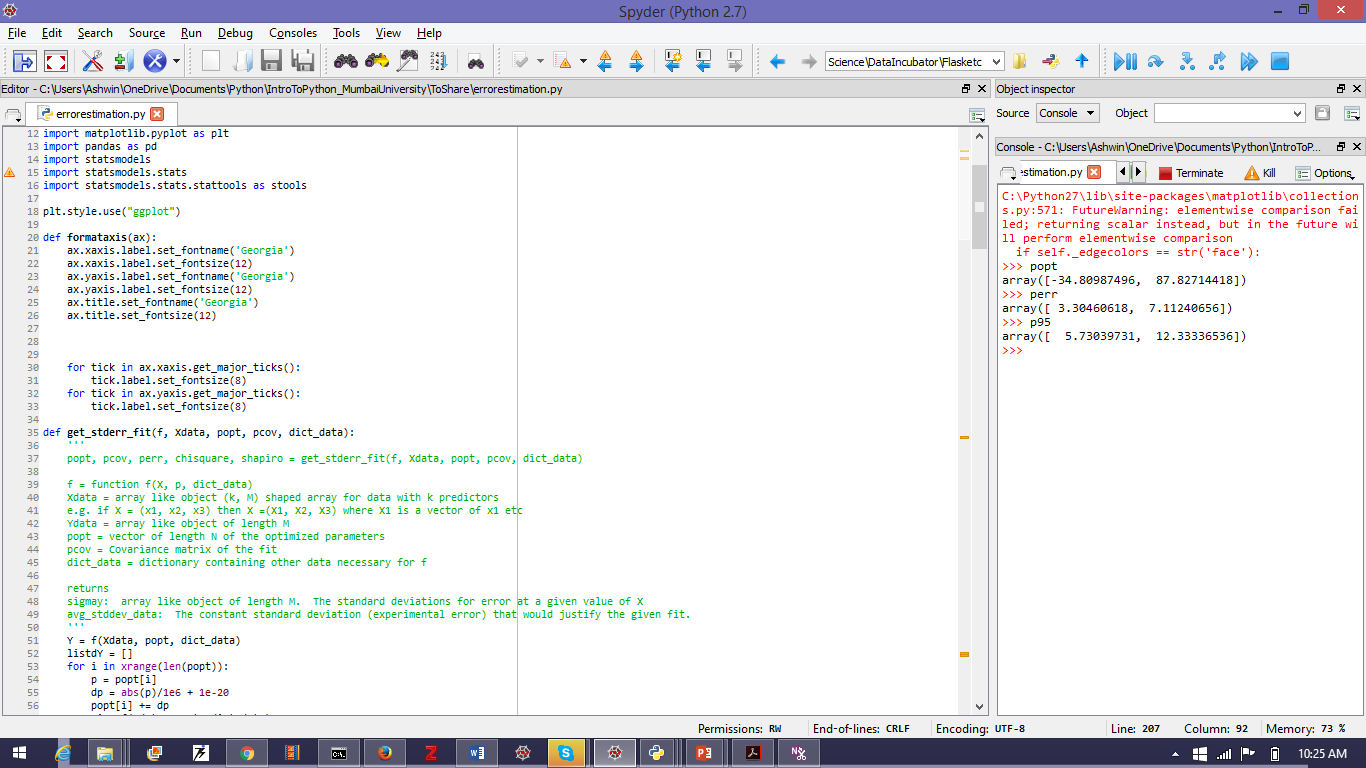


Figure : Code to format axis.

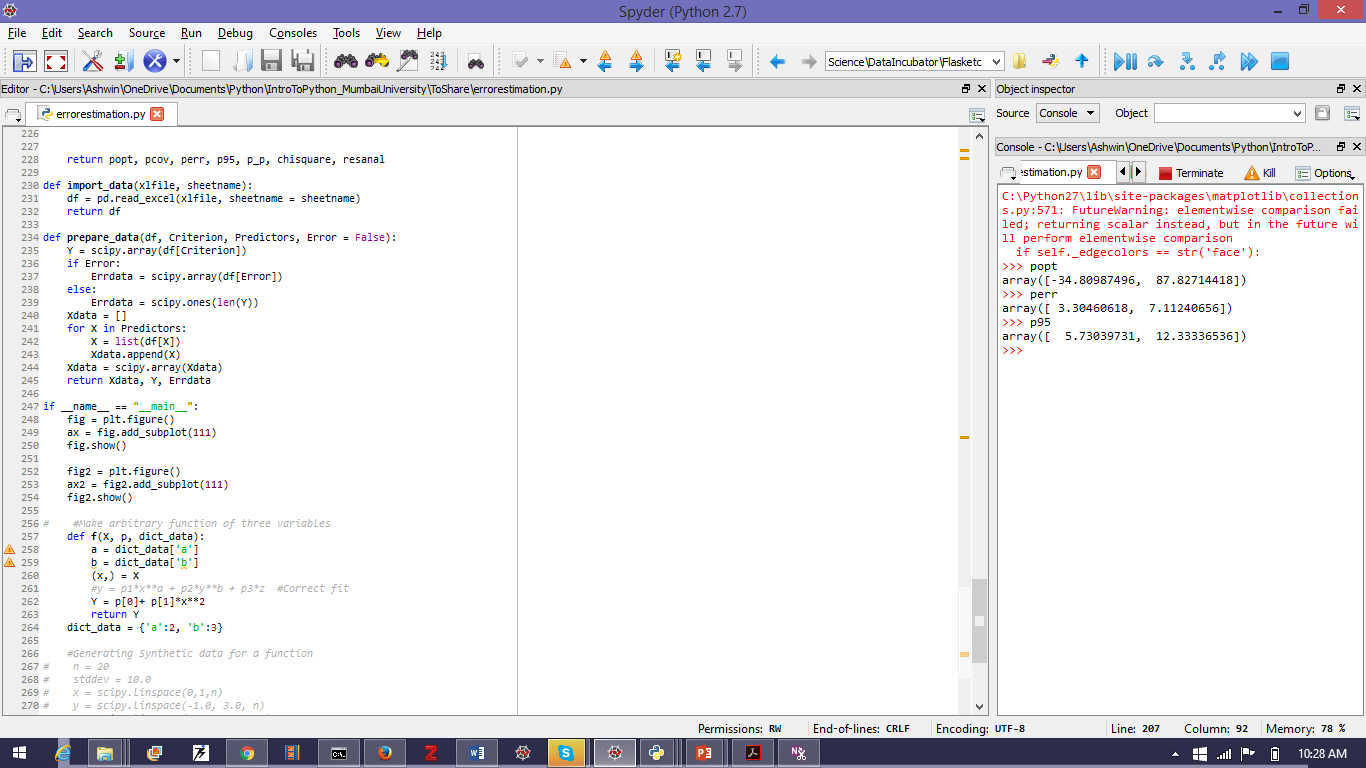


Figure : Code for importing data from an excel file and preparing it for fitting. See Figure 11 for illustration.

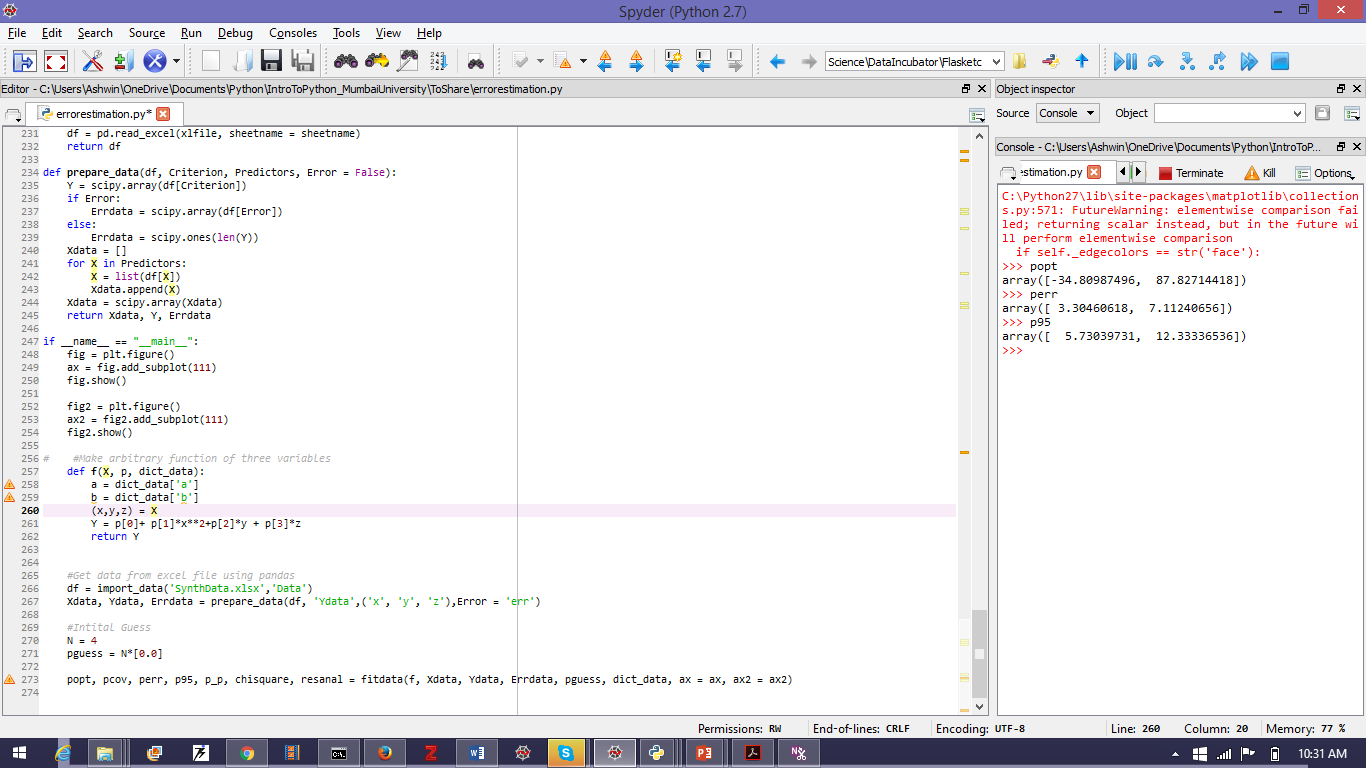


Figure : Illustration of use of importing, preparing and fitting data. The Excel file is named ‘Synthdata.xlsx’ and it contains a sheet called ‘Data’ formatted as shown in Figure 12

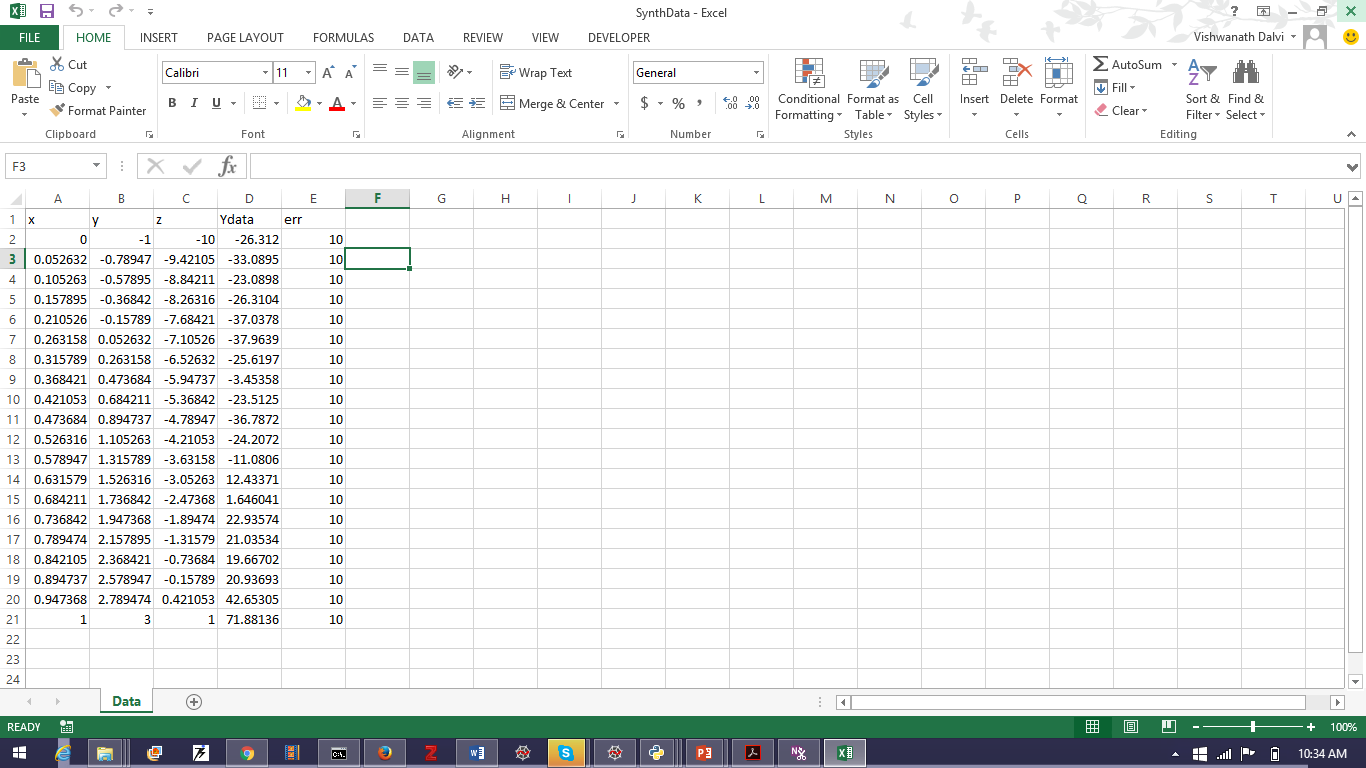


Figure : The Excel sheet used for the above problem. Note that the **first** row is only the headings and the headings are referenced in the code of Figure 11.

1. . Hence [↑](#footnote-ref-1)
2. and [↑](#footnote-ref-2)
3. Here replaces of linear least squares. [↑](#footnote-ref-3)