|  |  |
| --- | --- |
| Methods for Confidence interval calculations  With numeric-oriented approach | Abstract  TODO: list of methods, approach, ext…  Daniel Abutbul  Statistics Exercise |

Contents

[Motivation 2](#_Toc106371425)

[Working example 2](#_Toc106371426)

[Extrapolation 3](#_Toc106371427)

# Motivation

For many applications, we encounter the problem of creating predictions based on observed data. While a very general case can be formulated, if one wishes to practice, understand, implement and finally use such predictions, it is best first to solve the simplest imaginable problem, and try and solve it in as many methods as possible, while pointing out their advantages and disadvantages.

In this report, we will review and implement various methods for extrapolating the value of a straight line calibrated from observed data.

We will assume certain knowledge in probability theory and linear regression.

## Working example

Throughout this report, we will use the same simple example of the straight line:

The straight line formula tells us exactly the value of at any value of . For many real life situation, though, the formula and its parameters are unknown to the researcher. Instead, the researcher may have a sample of points which lie near the straight line, and originated from the straight line, but with added noise. This can represent for example a scientific experiment of measuring the value of at different ’s, with added measurement noise.

The sample is represented by a set of points , at each one we have the value . For our simplest working example, we assume the noise comes from a normal distribution, and we model it using a random variable , such that is sampled normally around the true value :

It is very common, and still useful, to restrict ourselves to the case that are i.i.d. variables:

Finally, our very specific working example will be of the data set:

The straight line, a possible sample points with noise from the straight line, and theirs best fit ( minimized) are given in Figure 1:

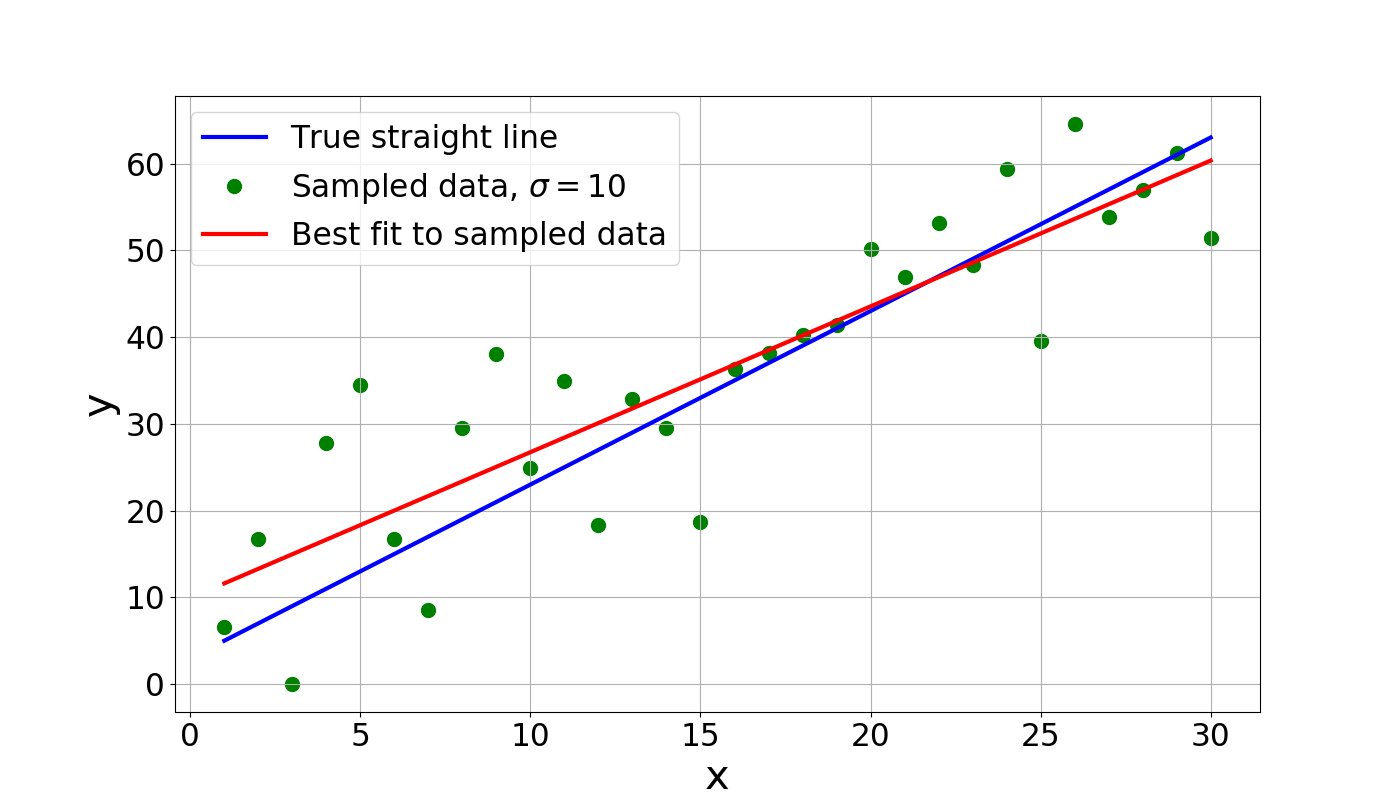


Figure 1: The straight line , sampled with normal noise ( i.i.d. random variables) and fitted with a straight line.

## Extrapolation and Confidence Intervals

Our task is to construct an extrapolation toward a value that is not in our data. In this report we will try to extrapolate the value of the straight line toward:

We will not only predict the value of at (that would most likely simply be the best fit to the observed data). We will decide on some range around the extrapolated value, such that in most noise realization the true value of the straight line will be within our predicted range, as demonstrated in Figure 2:

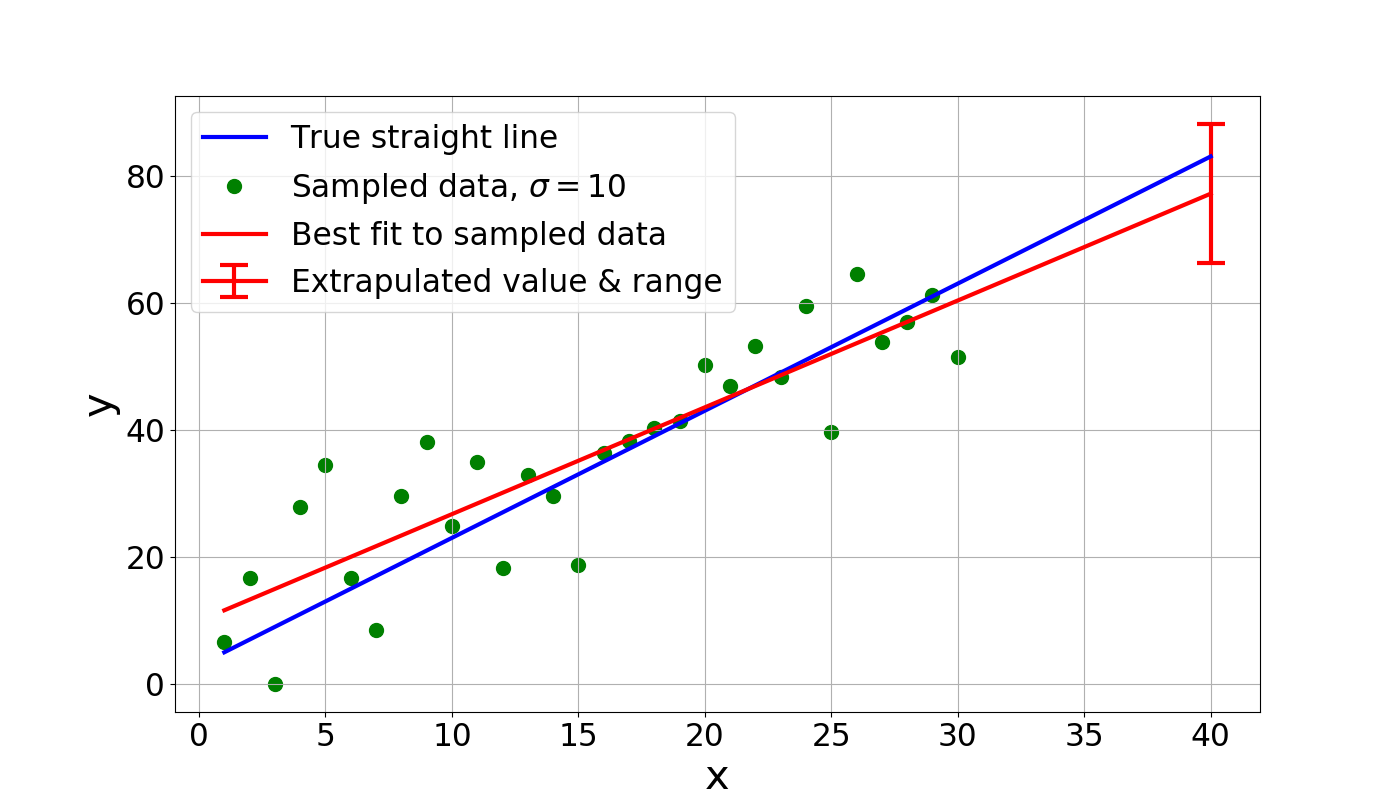


Figure 2: Extrapolation from sample x=1…30 toward and some range around it. We choose the range such that for most noise realizations, the true value of the straight line will fall within our range.

The Confidence Interval (CI) is a range around that extrapolated point, **constructed based on the sampled data**. We will say a range is a 95%CI if the true value falls within the range with probability of 0.95. Notice a 95%CI changes every noise realization; for each realization, we construct a range based on the observed sampled data. For each realization, the true value will either be inside or outside that range.

When we know the true line, which rarely happens in real life but does happen in this report, we can test our method for constructing 95%CI by running it for different noise realization. For each 95%CI construction method, we test if indeed in 95% of noise realizations the true line falls within that range.

In general, though, since we have only the sampled data and we construct some range around it, we can never know for sure if the true line indeed falls within the constructed range. When we call a range a 95%CI in such a case, we actually **model** the noise, and then construct a range such that we believe the probability the true line is inside that range is 0.95.

## Other literature

There is a lot of literature on the subject. The goal of this report is also to implement and practice the different methods, and so it does not claim novelty over existing literature (and definitely does not put as a goal to resemble a scientific peer-reviewed paper). Some main sources used in the preparation of this report are ‎[1]-[TODO].

# Exact solution

Conveniently, the problem of calculating the 95%CI for the case of a straight line has an exact solution. To see it, we will use the linear regression formulation, for which a general model that is linear in its parameters is given by:



( is an vector, and is vector. The result is a scalar)

Where represent the sample point at which we calculate (notice for a straight line: ), and are the parameters of the linear model (for a straight line we have and so ). As mentioned above, we wish to investigate the case when we sample from a linear model with added noise, modeled by the i.i.d. normal random variables .

Our samples are , , and we can rearrange them into one vector:

is called the design matrix. For samples of a straight line at points we have . The best fit as function of the noise is well known ‎[2] to be given by:

( is an matrix, and is an square matrix which has inverse. is matrix, and is matrix, and so the result is an matrix)

And so, our best fit prediction to any point is given using equation ‎(6) with the replacement of the true parameters by the fitted ones :

We create a prediction toward a new , and so it will be natural to calculate the variance of at any point . Notice ‎[3], and so:

Where is just defined as (notice it’s a number ). Since is normally distributed, so is : (proof can be found in ‎[3]).

In normal distribution a range of around the mean covers 95.4% of the area under the distribution, and so it will be most convenient to calculate the 95.4%CI throughout the report, using . Formula ‎(10) gives us then the 95.4%CI around any prediction constructed symmetrically around , since the probability for to fall within a range of from the true is 0.954. We can plot the 95.4%CI range as function of (for the same sample data from Figure 2), as shown in Figure 3:

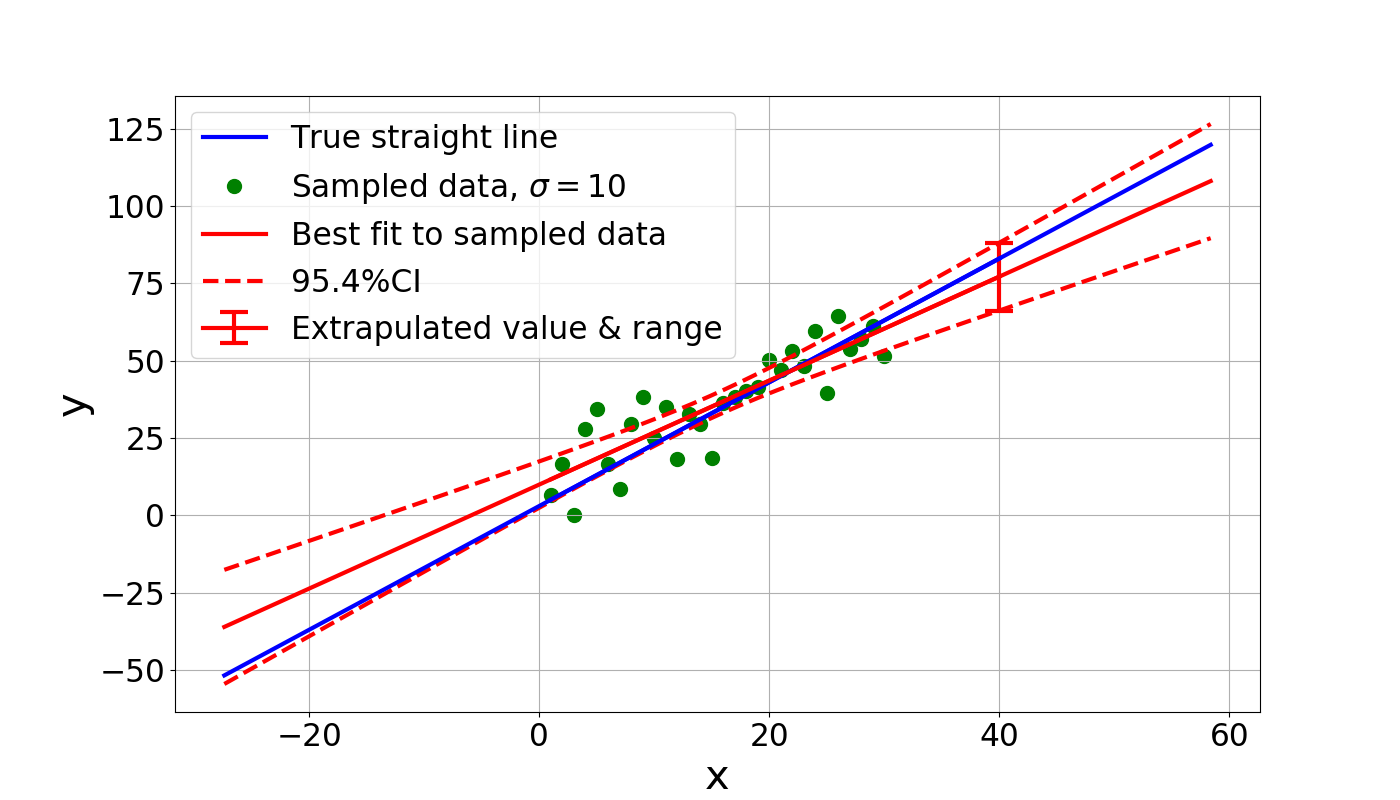


Figure 3: CI (95.4%CI) around the best fit. It can be seen that the prediction is more accurate around the middle of the sample, but as the extrapolation deviate from the sample the uncertainty increase.

The error bar at is of size 21.9, that is . It is about half the size of our original noise .

For any CI calculation method for which the range magnitude is independent of the noise realization, like in the case of the exact solution, we should then get the same CI that was calculated exactly here, that is a range of size 21.9.

# Sources

1. “Numerical Recipes, the Art of Scientific Computing”, Third Edition, <https://e-maxx.ru/bookz/files/numerical_recipes.pdf> . William H. Press, Saul A. Teukolsky, William T. Vetterling & Brian P. Flannery, Cambridge university press
2. Wikipedia linear regression, <https://en.wikipedia.org/wiki/Linear_regression>
3. <https://www.stat.purdue.edu/~boli/stat512/lectures/topic3.pdf>