## Loess (Locally Weighted Least Squares Regression)

<http://www.itl.nist.gov/div898/handbook/pmd/section1/pmd144.htm>

LOESS is one of many "modern" modeling methods that build on "classical" methods, such as linear and nonlinear least squares regression. Modern regression methods are designed to address situations in which the classical procedures do not perform well or cannot be effectively applied without undue labor. LOESS combines much of the simplicity of linear least squares regression with the flexibility of nonlinear regression. It does this by fitting simple models to localized subsets of the data to build up a function that describes the [deterministic part of the variation in the data](http://www.itl.nist.gov/div898/handbook/pmd/section1/pmd11.htm#def), point by point. In fact, one of the chief attractions of this method is that the data analyst is not required to specify a global function of any form to fit a model to the data, only to fit segments of the data.

<http://princeofslides.blogspot.de/2011/05/sab-r-metrics-basics-of-loess.html>

Using smoothers, there is no restriction on the functional form between X and Y with respect to intensity of the relationship, or direction (positive or negative). Of course, this means our fits are a bit more computationally intensive. And if not careful, it is very easy to overfit the data by trying to include every wiggle we see. But if done properly, one may be able to glean some extra information from the data by using a smoother instead of a restrictive linear model.

**Local Polynomial Regression Fitting**

Fitting is done locally. That is, for the fit at point x, the fit is made using points in a neighbourhood of x, **weighted by their distance from x** (with differences in ‘parametric’ variables being ignored when computing the distance). The **size of the neighbourhood is controlled by α** (set by span or enp.target). For α < 1, the neighbourhood includes proportion α of the points, and these have tricubic weighting (proportional to (1 - (dist/maxdist)^3)^3). For α > 1, all points are used, with the ‘maximum distance’ assumed to be α^(1/p) times the actual maximum distance for p explanatory variables.

<https://sites.google.com/site/davidsstatistics/using-r/smoothing-curves>

<http://stats.stackexchange.com/questions/82603/understanding-the-confidence-band-from-a-polynomial-regression>

<http://stats.stackexchange.com/questions/24993/how-to-get-an-r-squared-for-a-loess-fit>

Indeed, using Loess to achieve accuracy in a predictive model would be foolhardy. I think referring to Loess as a "model" conveys a possible misunderstanding about how it works and how it is intended to be used: it is really a *graphical, exploratory* tool to help see patterns and trends. Because it is really just a moving-window smoother, it acts as a fairly complicated spatial neighborhood model in which the fitted value at a point depends on which neighboring points exist in the dataset and on the values there.

**Spline**

**Bootstrap**

In [statistics](https://en.wikipedia.org/wiki/Statistics), bootstrapping can refer to any test or metric that relies on [random sampling with replacement](https://en.wikipedia.org/wiki/Random_sampling_with_replacement). Bootstrapping allows assigning measures of accuracy (defined in terms of bias, variance, confidence intervals, prediction error or some other such measure) to sample estimates. This technique allows estimation of the sampling distribution of almost any statistic using very simple methods.[[3]](https://en.wikipedia.org/wiki/Bootstrapping_%28statistics%29#cite_note-Varian-3)[[4]](https://en.wikipedia.org/wiki/Bootstrapping_%28statistics%29#cite_note-4) Generally, it falls in the broader class of [resampling](https://en.wikipedia.org/wiki/Resampling_%28statistics%29) methods.

The basic idea of bootstrapping is that inference about a population from sample data (sample → population) can be modeled by *resampling* the sample data and performing inference on (resample → sample). As the population is unknown, the true error in a sample statistic against its population value is unknowable. In bootstrap-resamples, the 'population' is in fact the sample, and this is known; hence the quality of inference from resample data → 'true' sample is measurable.

The simplest bootstrap method involves taking the original data set of N heights, and, using a computer, sampling from it to form a new sample (called a 'resample' or bootstrap sample) that is also of size N. The bootstrap sample is taken from the original using [sampling with replacement](https://en.wikipedia.org/wiki/Simple_random_sample) so, assuming N is sufficiently large, for all practical purposes there is virtually zero probability that it will be identical to the original "real" sample. This process is repeated a large number of times (typically 1,000 or 10,000 times), and for each of these bootstrap samples we compute its mean (each of these are called bootstrap estimates). We now have a histogram of bootstrap means. This provides an estimate of the shape of the distribution of the mean from which we can answer questions about how much the mean varies. (The method here, described for the mean, can be applied to almost any other [statistic](https://en.wikipedia.org/wiki/Statistic) or [estimator](https://en.wikipedia.org/wiki/Estimator).)

A great advantage of bootstrap is its simplicity. It is a straightforward way to derive estimates of [standard errors](https://en.wikipedia.org/wiki/Standard_error_%28statistics%29) and [confidence intervals](https://en.wikipedia.org/wiki/Confidence_intervals) for complex estimators of complex parameters of the distribution, such as percentile points, proportions, odds ratio, and correlation coefficients. Bootstrap is also an appropriate way to control and check the stability of the results. Although for most problems it is impossible to know the true confidence interval, bootstrap is asymptotically more accurate than the standard intervals obtained using sample variance and assumptions of normality.[[15]](https://en.wikipedia.org/wiki/Bootstrapping_%28statistics%29#cite_note-DiCiccio1996-15)

**Do a resampling with replacement.**

**Parametric bootstrap**

In this case a parametric model is fitted to the data, often by [maximum likelihood](https://en.wikipedia.org/wiki/Maximum_likelihood), and samples of [random numbers](https://en.wikipedia.org/wiki/Random_number_generation) are drawn from this fitted model. Usually the sample drawn has the same sample size as the original data. Then the quantity, or estimate, of interest is calculated from these data.

**Cross-Validation**

Cross-validation, sometimes called rotation estimation,[[1]](https://en.wikipedia.org/wiki/Cross-validation_%28statistics%29#cite_note-1)[[2]](https://en.wikipedia.org/wiki/Cross-validation_%28statistics%29#cite_note-Kohavi95-2)[[3]](https://en.wikipedia.org/wiki/Cross-validation_%28statistics%29#cite_note-Devijver82-3) is a [model validation](https://en.wikipedia.org/wiki/Model_validation) technique for assessing how the results of a [statistical](https://en.wikipedia.org/wiki/Statistics) analysis will generalize to an independent data set. It is mainly used in settings where the goal is prediction, and one wants to estimate how [accurately](https://en.wikipedia.org/wiki/Accuracy) a predictive model will perform in practice. In a prediction problem, a model is usually given a dataset of known data on which training is run (training dataset), and a dataset of unknown data (or first seen data) against which the model is tested (testing dataset).[[4]](https://en.wikipedia.org/wiki/Cross-validation_%28statistics%29#cite_note-Newbie_question:_Confused_about_train.2C_validation_and_test_data.21-4) The goal of cross validation is to define a dataset to "test" the model in the training phase (i.e., the validation dataset), in order to limit problems like overfitting, give an insight on how the model will generalize to an independent data set (i.e., an unknown dataset, for instance from a real problem), etc.

One round of cross-validation involves [partitioning](https://en.wikipedia.org/wiki/Partition_of_a_set) a [sample](https://en.wikipedia.org/wiki/Statistical_sample) of [data](https://en.wikipedia.org/wiki/Data) into [complementary](https://en.wikipedia.org/wiki/Complement_%28set_theory%29) subsets, performing the analysis on one subset (called the training set), and validating the analysis on the other subset (called the validation set or testing set). To reduce [variability](https://en.wikipedia.org/wiki/Variance), multiple rounds of cross-validation are performed using different partitions, and the validation results are averaged over the rounds.

Cross-validation is important in guarding against [testing hypotheses suggested by the data](https://en.wikipedia.org/wiki/Testing_hypotheses_suggested_by_the_data) (called "[Type III errors](https://en.wikipedia.org/wiki/Type_III_error)"[[5]](https://en.wikipedia.org/wiki/Cross-validation_%28statistics%29#cite_note-Mosteller1948-5)), especially where further [samples](https://en.wikipedia.org/wiki/Statistical_sample) are hazardous, costly or impossible to collect.

**Sampling from distributions**

One way to sample from a known probability density function (pdf) is to use [**inverse transform sampling**](http://en.wikipedia.org/wiki/Inverse_transform_sampling). First, you integrate the pdf to get the cumulative distribution function (cdf). Next, you find the inverse of the cdf. Finally, apply this inverse cdf to each number in a sample of Uniform(0,1) observations.

<http://blog.quantitations.com/tutorial/2012/11/20/sampling-from-an-arbitrary-density/>

Another option for sampling that doesn’t require finding the inverse cdf is [**rejection sampling**](http://blog.quantitations.com/tutorial/2012/11/24/rejection-sampling/).

The **accepance-rejection method** (ARM) is among the modst widely applicable mechanisms for generating random samples.

