# Graphical Presentation of a Nonparametric Regression with Bootstrapped Confidence Intervals

Mark Nicolich & Gail Jorgensen
Exxon Biomedical Science, Inc., East Millstone, NJ

## INTRODUCTION

Parametric regression (least-squares) techniques are used to estimate a statistical model that attempts to predict a variable based on one, or more, other variables. The model is required to have a specified algebraic form such as a straight line, a parabola, or an exponential curve. An example would be predicting a persons annual income based on their age, years of schooling and gender using a linear model of the form:

income = A + B*age + C*years_of_school + D*gender

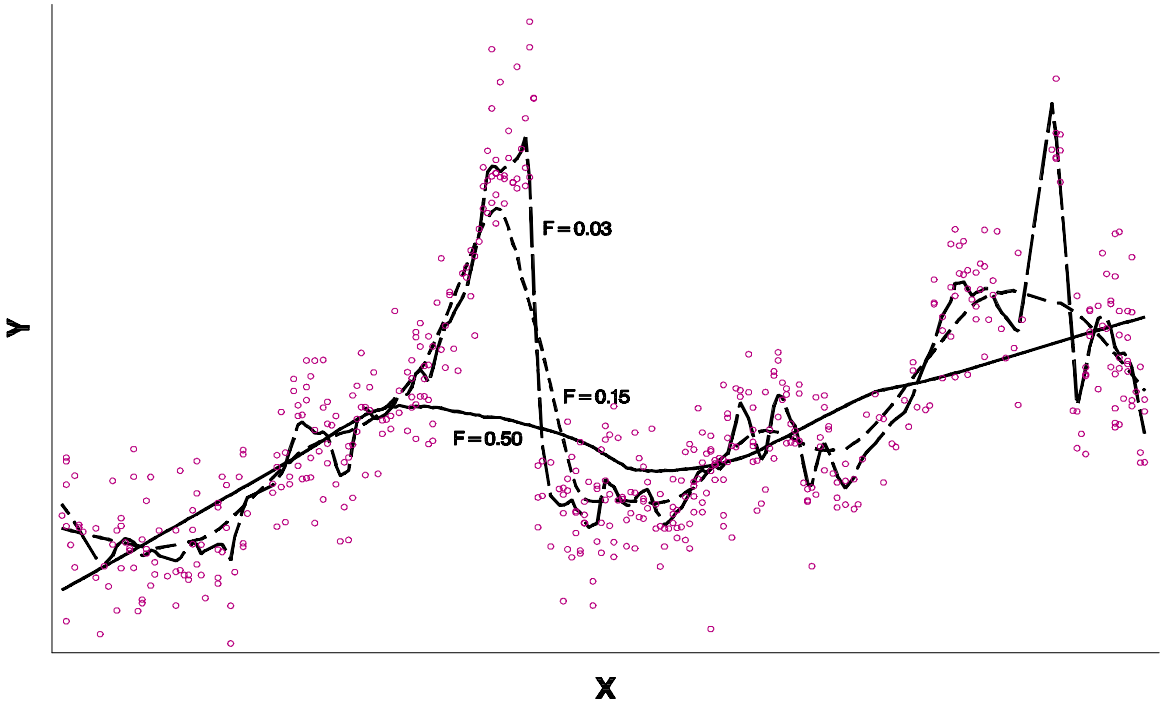
These models are restrictive because it is not always possible to find a simple mathematical model form to describe the relationship that exists between the variables.

An alternative approach is to use nonparametric models which do not require exact specification of the model form. The approach is known under several names such as nonparametric regression and smoothing. Both the parametric and nonparametric methods have their advantages and disadvantages. The parametric approach has the problem of inexact results if the model is not correctly specified, while the nonparametric approach has the problem of not making full use of the data when the model is known exactly. The appropriate method to use depends on the data at hand and the questions to be answered by the data.

## NONPARAMETRIC REGRESSION

There are several nonparametric methods (Härdle, 1989) which can be used to form the regression line. The methods include kernel smoothing, spline fitting or smoothing, L-Smoothing, R-smoothing, M-smoothing, and LOWESS techniques. The techniques are mathematically related to each other, but have different properties which are advantageous in different situations. The technique discussed in this paper is a useful method proposed by Cleveland (1979) which has the advantage of not being sensitive to outliers (it is robust) and allowing the user to easily adjust the degree of smoothing without the curve having an excessively wiggly shape. The technique is usually called LOcally WEighted Scatterplot Smoothing (LOWESS or LOESS). It fits a simple straight line (or higher order polynomial function) in successive regions of the data, then iteratively reforms the results to create a smooth, continuous curve. The result of a LOWESS regression is a line which best fits the data locally, but is not constrained to be of a particular mathematical form. Visual inspection and judgment can then be used to determine the nature of the relationship between the variables. The resulting curve is not sensitive to missing observations and can be used as a useful tool for finding spurious or outlying observations.

Michael Friendly of York University has developed a SAS® MACRO that will generate a LOWESS plot based on the LOWESS technique. It is easy to use and can produce hard copy plots via SAS/GRAPH®. There are several user controlling parameters such as the width of the region of interest, the degree of the fitted polynomial, and some characteristics of the resulting plot. The program is available on the SAS Institute Web Site as LOWESS2 from www.sas.com/samples/A56143. The degree of smoothing in the program is controlled by the width of the initial region or window used for the initial simple regression, and is expressed as a fraction of the data range. The choice of the window width is subjective and is based on the goals of the analysis. A wide window results in a smooth curve that shows 'long term' trends, while a narrow window will show more of the local variation. Experimentation with the width will often indicate the best choice for the problem being studied. The plot below is for some artificial data and shows three LOWESS curves based on a linear regression and with increasing degrees of smoothness. When the window width is set at 3% of the range (F=0.03) the curve is wiggly, as the width is increased to 0.15 the curve is smoother and misses the high peaks and drops. When the width is set at 50% the curve shows only general trends. The extreme case of 100% would be




a straight line; if the initial regressions were second order polynomials, the 100% curve would have a simple second order curve.

## BOOTSTRAPPING and CONFIDENCE LIMITS

A weakness of the nonparametric regression techniques is that, because the regression curve is not based on a specific mathematical model and distribution, it is not possible to directly calculate confidence intervals (CI). The CIs are useful in determining the precision of the predicted model and give the user an idea of how useful the model is in a particular region.

For the kernel and spline smoothers there are techniques for constructing pointwise confidence intervals and variability bands (Härdle, 1989). These methods are useful, but can be computationally intensive. However, confidence intervals can be generated for nonparametric regression lines through the bootstrap technique. The bootstrap technique (Efron and Tibshirani, 1993) is a data-based method of simulating an empirical distribution of results, in this case nonparametric regression curves, through a resampling process. The bootstrap has been applied to almost all the nonparametric regression methods, each with several variants in technique (Efron and Tibshirani, 1991,1993; Härdle, 1989; Shao and Tu, 1995). The fundamental idea of bootstrapping is to start with a sample (the *original* sample), then create a new sample (the *bootstrap* sample) by sampling with replacement from the original sample. The process is repeated a large number of times to generate a series of bootstrapped samples, and the statistic of interest (such as the mean) is computed for each sample. Because the sampling is with replacement, the bootstrap sample will likely contain several duplicate observations from the original sample, and nearly each bootstrap sample will be unique. Each bootstrap sample represents a plausible alternate sample from the original population, and the set of calculated statistics yields information about variation in the original sample. Since a distribution of different outcomes was generated from one sample, it seems as if you have lifted yourself by your own bootstraps.

In this paper we make a simple application of the bootstrap technique to the LOWESS regression. We create a set of (n=100, for example) bootstrap samples and their associated LOWESS regression lines. The set of nonparametric bootstrapped regression lines constitutes an empirical distribution of regression lines. At each independent value (X value), points on the y-axis which are at the 2.5th percentile and 97.5th percentile of the set of bootstrapped regression lines are selected to form the 95% CI of the nonparametric regression line (Efron and Tibshirani, 1993).

The nonparametric bootstrap CI has different properties from the usual parametric CI. Parametric CIs are smooth and have a parabola shape because of the underlying mathematical assumptions, and are narrowest at the mean of the predicted and predictor variables. Bootstrap CIs are based on actual data and are therefore jagged and without predetermined shape. The width of the CI at any given point on the x-axis will vary, depending on the placement of datapoints at or near that point.

Conceptually, nonparametric CIs have the same interpretation as the parametric CIs. For any exposure value on the x-axis the 95% CI will enclose the true response in approximately 95% of the instances in which it is calculated. The non-parametric 95% confidence intervals are based on the actual data rather than on an assumption about the distribution of the residuals as in parametric least-squares analyses. Therefore, in the regions where there are only a few datapoints one should be cautious about interpreting the intervals. If there are only a few datapoints relatively close together the effective variation of the bootstrap estimate at that interval will be narrow or small. Conversely, a few widely spaced points in an interval will yield a large estimate of the variation. While there are other methods to estimate the CI for nonparametric analyses, none perform well in the region of sparse data.

## SAS MACRO and EXAMPLE

We have developed a SAS MACRO to generate a series of bootstrap samples from a data sample, then determine the bootstrapped CI, and plot the data, LOWESS curve and CIs. The MACRO can be altered to set the number of bootstrap samples and the width of the CI.

As an example of how to use the MACRO we have a set of simulated data. The data show the relation-ship between the change over the work day of the forced expiratory volume during the first second of a pulmonary function test ($FEV_1$) and the amount of sunlight the worker was exposed to that day. The $FEV_1$ is a measure of lung function and is sensitive to upper airway restriction.

To generate 95% CI for the LOWESS plots, one hundred bootstrap samples were used. The highest and lowest 2.5% points were used to form a 95% confidence interval (Efron and Tibshirani, 1993). Since there were 100 points the 2.5% point was the median of the 2 and 3% points and the 97.5% point was the median of the 97 and 98% point. The resulting plot is at the end of the paper. All the LOWESS regressions were based on a window width of 0.4.

As can be seen in the plot, the LOWESS regression procedure is highly influenced by few datapoints at “high exposures”. Although there are no large reductions in lung function at higher exposures, there are so few datapoints that definitive conclusions about exposure-response are difficult where data are sparse.

## OTHER REGRESSIONS

The fundamentals of the LOWESS technique are easy to understand, and relatively straight forward to implement. Many analyses need more than one independent variable to explain variation in the dependent variable, such as the example in the introduction of this paper. There are techniques which can be used when there are multiple independent variables. Two methods that have received attention are the *projection pursuit* method and the *grand tour* method. A paper by Cook, et al, (Cook, et al, 1995) presents an overview and comparison of the two techniques. The paper by Schluter, et al, (Schluter and Nychka, 1994) discusses the projection pursuit method, presents an extensive example, and provides a FORTRAN program to estimate models.

## CONCLUSIONS

This MACRO and the LOWESS technique with CI are useful tools for data analysis. They are an improvement over parametric regressions when the form of the regression model is not known or is too complicated to specify a model. They can be combined with parametric regression as another test of the model adequacy.

On the negative side it may be thought that the choice of window width can lead to different conclusions, but this is not different from choosing different parametric models in ordinary regression. On the positive side, the choice of window width allows consideration of immediate vs long term changes in the data. There are applications in exposure response studies where the effects of short term vs long term exposures could be usefully modeled with these techniques.

A few possible weaknesses of the LOWESS technique are that it can be used for interpolation but not extrapolation, it is unfamiliar to many researchers, and it can be computationally time consuming. None of these possible problems should keep researchers from using this technique in their data exploration and analysis, and promoting its acceptance in the published literature.

## REFERENCES

Cleveland, WS, "Robust Locally Weighted Regression and Smoothing Plots", *Journal of the American Statistical Association*, 74, 1979, pp 829-836.

Cook, D., Buja, A., Cabrera, J., and Hurley, C., "Grand Tour and Projection Pursuit", *Journal of Computational and Graphical Statistics*, 4(3), 1995, pp 155-172.

Efron, B and Tibshirani, R, "Statistical Data Analysis in the Computer Age", *Science*, 253, July 26, 1991, pp 390-395.

Efron, Bradley and Tibshirani, R, *An Introduction to the Bootstrap*, Chapman and Hall, New York, 1993.

Härdle, Wolfgang, *Applied Nonparametric Regression*, Cambridge University Press, NY, 1989.

Schluter, D. and Nychka, D., "Exploring Fitness Surfaces", *The American Naturalist*, 143(4), 1994, pp 597-616.

Shau, J. and Tu, D., *The Jackknife and Bootstrap*, Springer , New York, 1995.



Mark Nicolich & Gail Jorgensen
732.873.6241 & 732.873.6124
Exxon Biomedical Science, Inc.
Mettlers Road, CN 2350
East Millstone, NJ 08875,2350

# Simulated Data Depicting Nonparametric Regression of △FEV1% vs. Sunlight Exposure with Bootstrap 95% CI