REGRESSION MODELING IN COMPUTER-SUPPORTED LEARNING ENVIRONMENTS

Joachim Engel
University of Hannover, Germany
engel@math.uni-hannover.de

We consider the role of technology in learning concepts of modeling univariate functional dependencies. It is argued that simple scatter plot smoothers for univariate regression problems are intuitive concepts that- beyond their intended usefulness in providing a possible answer to more intricate regression problem - may serve as a paradigm for statistical thinking, detecting structure in noisy data. Simulation may play a decisive role in understanding the underlying concepts and acquiring insight into the relationship between structural and random variation.

INTRODUCTION

Dynamic and interactive software together with the computing power of today's accessible hardware illuminate key concepts of statistics and encourage an exploratory and activity-based working style in analyzing real data, including the creation of self-constructed methods of data representation and analysis and evaluating these by simulations. In our project *Statistical Thinking and Stochastic Modeling in Computer Supported Environments*, implemented in courses for teacher students in mathematics at the Universities of Ludwigsburg and Hannover, we investigated the didactical efficacy of various software environments including *Fathom*, *S-Plus* and *Lisp-Stat* on student learning of statistical concepts (Engel, 2002). In this paper we focus on one particular content of that project and its potential for enhancing statistical thinking: modeling functional dependencies in stochastic situations.

"Statistical thinking is concerned with learning and decision making under uncertainty. Much of that uncertainty stems from omnipresent variation. Statistical thinking emphasizes the importance of variation for the purpose of explanation, prediction and control" (Wild and Pfannkuch, 1999). Variation is the reason why sophisticated statistical methods were devised to filter out messages in data from the surrounding noise. A core concept of modeling statistical data is what Borovcnik (2004) calls the structural equation which represents data as decomposed into a signal to be recovered and noise.

Data	=	Signal Pattern Fit Model Explained Common	+	Noise Deviation Residual Residual unexplained variation specific causes

While the signal comprises controlled or explained causes why data vary, the noise contains the unexplained variation, usually modeled as a random quantity with expectation 0.

ON MODELING SCATTERPLOT DATA

Starting point for investigating functional relationships between two empirical variables is the collection of n pairs of measurements (x_1, y_1), …, (x_n, y_n) represented in a scatter plot. The objective of the modeling process is to derive a function f expressing the dependence of the two variables either through a functional term of the form $y = f(x)$ or as a function graph. A simple graph or functional equation $y = f(x)$ representing the data cloud is an efficient compression of the data which is easy to communicate to others and easier to interpret and compare with other graphs than the complete original data set. As for any type of mathematical model, the obtained representation may play a decisive role in understanding the dynamics driving the observed phenomena, predicting new data and, possibly, forming the basis for effective intervention. Many

excellent instructive examples for modeling functional dependencies, to be worked out with *Fathom*, can be found in Erickson (2005).

In many situations, the simplest way to derive a graph from a scatter plot (x_i, y_i), $i = 1, \ldots, n$ is through interpolation. Connecting the data with straight lines, interpolating polynomials or cubic spline functions results in curves that may well help at discerning possible trends in observations. Any type of interpolation is certainly appropriate when the observations represent error free measurements of the variables of interest, i.e., if $y_i = f(x_i)$ holds exactly. However, in all empirical situations the observations are subject to measurements errors, sampling errors or other disturbances ("noise"), attributed to lurking variables that cannot be measured, controlled or – if for no other reason than model simplicity– are not explicitly included. Recognizing that variables in empirical studies are usually disturbed by measurement and sampling errors leads to include random components into the model. The most common approach– in line with Borovcnik's representation – is to decompose the observations additively into $y_i = f(x_i) + \varepsilon_i$, where the model function *f* represents the trend or signal and ε_i is noise.

NONLINEAR DATA

The approach of curve fitting is based on the assumption that the unknown function $f(x) = f(x, \theta)$ belongs to a pre-specified or known class of functions characterized by a finite dimensional parameter θ (e.g. linear, exponential, logistic function). Then the objective is to determine that value of the unknown parameter such that the model function fits the data best. In the linear case the answer is standard, leading under the objective of minimizing the sum of squared errors to the formulae for OLS linear regression. But obviously, life is not always linear. Restricting yourself to linear regression then leads to a narrow mindset that not only cuts the learner off from many interesting problems, but also may impede appreciation and interpretation of linear regression itself.

One standard approach to data sets with nonlinear structure is to linearize the data by a suitable transformation. As an example we consider data obtained by heating different amounts of water for 30 seconds in a microwave oven observing the temperature increase (Erickson, 2005). A log transform of both *x* and *y* variable renders a linear structure leading after a back transform of the OLS fit to a model for the original data (see Figure 1). Transforming data with the purpose of creating linear structures is quite an instructive task that can easily implemented in most statistical packages. Dynamic software like Fathom is here particularly advantageous in situations where a suitable transformation depends on a yet unknown parameter. The software then allows choosing this parameter interactively while the effects of this choice can be observed dynamically in a scatter plot of the transformed data. By eyeball analysis, while the slider representing the unknown parameter is changed, a satisfactory linear structure may be obtained.

Most non-linear models used in intro stats classes can be transformed to linear and fit by ordinary least squares with the notable exception of periodic functions. Tidal patterns, sunrise times, data depending highly on the time of the day or the season of the year may not be modeled this way, because periodic structures cannot monotonically be transformed into linear. Then instead of the transform-back-transform approach one may try to minimize an outside criterion like OLS directly. Just as in the case of linear regression, the objective now is to find a value for the parameter θ minimizing $\sum \left[y_i - f(x_i, \theta) \right]^2 = \min_\theta$. For this general set-up Numerical Analysis offers sophisticated algorithms like Gauss-Newton. Insightful use of these algorithms asks for expertise (numerical stability, choice of pilot estimators, ...), that students barely won't have before attending graduate level classes in mathematics or statistics. Implementing these ideas is in the domain of the professional statistician's software like *R*, *S-Plus* or *Lisp-Stat* etc.

Nevertheless, in problems with a low-dimensional parameter θ we may approximate optimal parameters by trial and error, using software for illustration. However, feasibility is one important aspect, appropriateness a different issue. In stochastic modeling we always model both, signal AND noise, structure AND random deviation. Following the common (but by no means

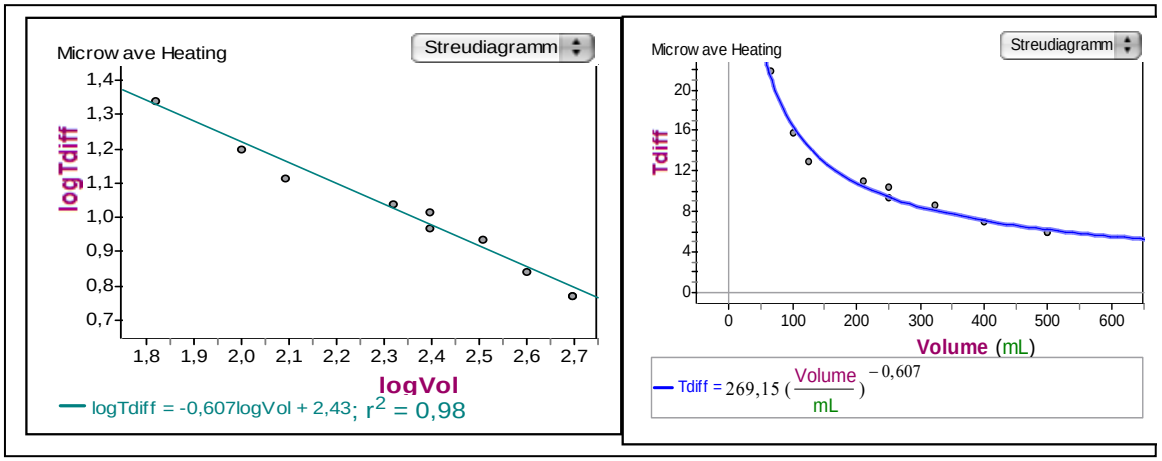

Figure 1: Microwave oven data. Linear fit after a log transformation (left), original data with a fitted curve obtained through back-transforming the linear function (right)

exclusive) approach of an additive decomposition of the data into structure and noise, there is still a lot of freedom in how to model noise. The simplest situation is based on the assumption that $y_i = f(x_i, \theta) + \varepsilon_i$, where the noise ε is modeled as independent random variables with expected value 0 and a constant variance of σ^2 independent from *x*, while in more complex situations, for example, the error terms may be stochastically dependent or the variance of ε varies with the value of *x*. Also, additivity in linking structure and noise is most common if only for the sake of simplicity. Specifying how structure and noise are connected is just part of the model, and the choice of a suitable model depends on the context of the data. In standard regression, the assumption is additive homogeneous noise. However, when transforming, we change this assumption. As an example, consider again the Microwave data. The model function for the original data in Figure 1 has been obtained after a linear regression of the log transformed data. Implicitly, we assumed $\log(y_i) = a\log(x_i) + b + \varepsilon_i$. This model corresponds to a multiplicative error model for the original data specified as $y_i = \beta \cdot x_i^{\,a} \cdot \exp(\varepsilon_i) \approx \beta \cdot x_i^{\,a}(1+\varepsilon_i)$ with $\beta = \exp(b)$. This means that we now have indeed specified a heteroscedastic model with non-constant error variance of $\left[\beta \cdot x_i^{\,a}\right]^2 \mathrm{var}(\varepsilon_i)$. From a statistical point of view, it is important to make explicit that OLS is good only under assumptions. And when we transform models, we not only change assumptions about the trend, but also about the random deviations from the trend.

SIMULATION

Simulation may serve here as a very helpful teaching strategy to illustrate the differences of the two underlying concepts. We generated data according to $y_i = b \cdot x_i^{\,a} + \varepsilon_i$ and $z_i = b \cdot x_i^{\,a+\varepsilon_i}$. While the *y*-Data are custom tailored for direct OLS minimization, the *z*-Data are in fact heteroscedastic and have constant error variance exactly after a log-transform. Therefore, pursuing the linearization with the *y*-data is just as inappropriate as applying a direct OLS minimization via Gauss-Newton for the *z*-Data, see Figure 2.

BASIC CONCEPTS OF DATA SMOOTHING

The problem with fitting curves from a parametric family is that their derivation may be guided by intuition and experience from the field of application, but it often lacks an objective justification. Quite often one may have no idea at all which type of functions might be suitable to model the data at hand. Using the wrong parametric model then leads to misspecifications. This drawback calls for methods with more flexibility because the assumption of a parametric functional class becomes a “straight jacket” imposing a given structure on the data or excluding possibly existing data structure by assumptions – a contradiction to the principles of exploratory data analysis and a discovery-oriented approach to learning.

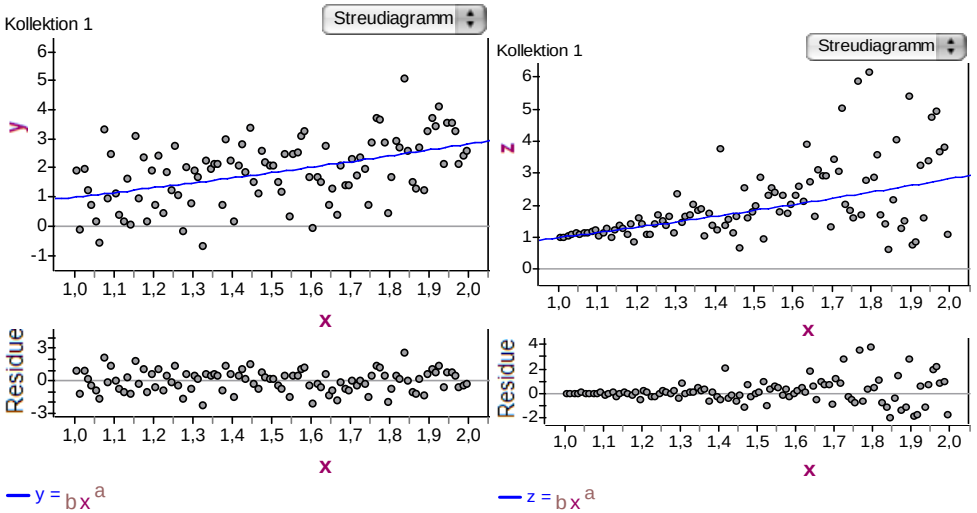

Figure 2: Simulated data with the signal $f(x) = x^{3/2}$. The added noise in the left frame is homoscedastic, rendering the data suitable for direct OLS, while the noise in the right is chosen such that the data are suitable for a log-transform. Notice their heteroscedasticity.

Computationally intensive smoothing methods that allow the derivation of a model curve with a minimum of a-priori specifications, have been around for several decades now and are implemented in most professional software packages including R, S-Plus and Lisp-Stat. Some go by fancy names like local polynomial approximation, kernel estimation, wavelet regression, smoothing splines etc., but their underlying idea is simple and intuitive (Weldon, 2002). For a comprehensive view, see e.g., Simonoff (1996). The most easily accessible approach may be the regressogram, where the sample space is sliced into cells - just like for a histogram in case of univariate data - and the average of the *y*-values is calculated within these cells. The next step is a "moving regressogram" or running mean, i.e., the cell or smoothing interval is centered round the point *x* of estimation. Then *x* is pulled over an input grid of the sample space. The result is a moving average, still a ragged curve. Introducing smoothly weighted moving averages (like Epanechnikov weights) leads immediately to smoother results which are mathematically and aesthetically more satisfying. Finally, while the moving average can be considered as a "locally constant" estimator (within the smoothing interval centered round *x* we fit the cell mean as functional value), we may also consider a local linear estimator (fitting a linear model within the smoothing interval). These remarks indicate that starting from the intuitive regressogram climbing upwards to gradually higher levels of sophistication is straightforward. However, the genuine purpose of teaching elementary concepts of smoothing should not originate in the strive to introduce recent concepts of statistical methodology into the introductory statistics classroom– a rather questionable curricular orientation – but its contribution to promoting statistical thinking.

I am grateful to Cliff Konold for drawing my attention to *TinkerPlots* (Test release of version 2.0) and its implementation of basic smoothing ideas. Besides the moving average (as well as a moving median and midrange) *TinkerPlots* allows a polygon-type variant of the regressogram by connecting the midpoints of the regressogram with straight lines, a direct analogue to the frequency polygon as a refinement of the histogram for univariate data. Figure 3 shows a scatter plot of the electricity usage data (Simonoff, 1996) in an all-electric home for 55 months. Average daily electricity usage (in Kilowatt Hours) is plotted against average daily temperature (Degrees Fahrenheit).

A delicate question for any smoothing method is the choice of a smoothing parameter, here the window size or bandwidth. Automatic bandwidth choices based on some optimality criteria and sophisticated plug-in or cross-validation procedures are possible, but barely accessible except for the very advanced students. Moreover, for the purpose of promoting statistical thinking, an automatic choice is not even desirable. For exploratory purposes it suffices

completely to choose the window size by hand through a slider. Exploring with several window widths instills a sense of balancing out between recovering signal and suppressing noise.

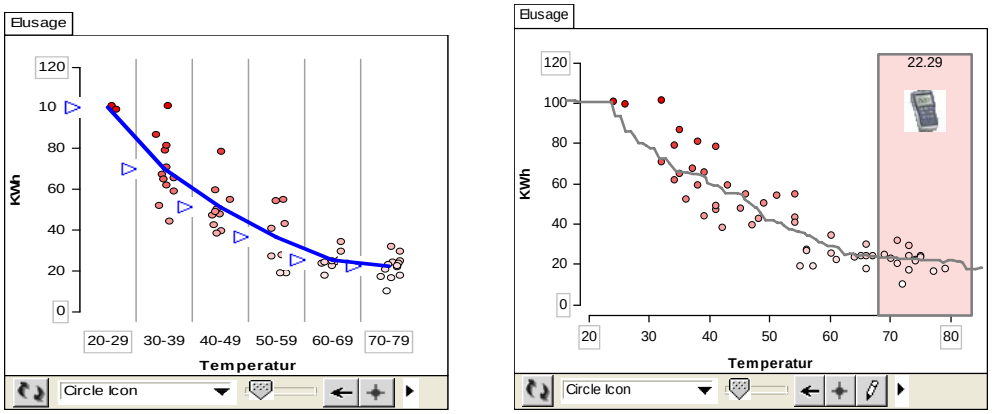

Figure 3: Regresso-Polygon (left) and moving average curve. Notice that the moving average curve at the center of the shaded window is computed by averaging over all ordinate values within the window.

For small bandwidths the resulting curve follows quite closely the ruggedness of the data while with larger bandwidths – just in the case of large bins of a histogram – the noise is averaged out and the result is a smooth curve. However, when the bandwidth is too large, structure in the data will be smeared out together with the noise. The bandwidth mediates between signal and random noise. Figure 4 shows different representations for the smoothed electrical usage data.

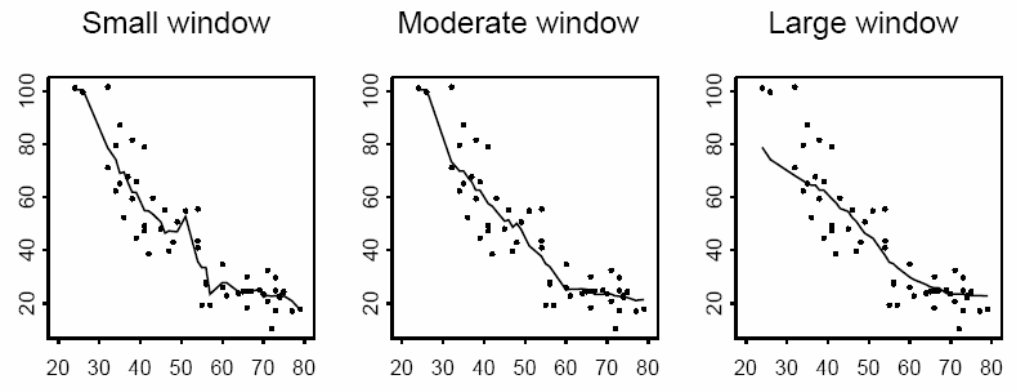

Figure 4: Smoothed representation of the electrical usage data based on moving averages with different window sizes or bandwidths

ROLE OF SIMULATION

The discovery and specification of trends in bivariate data – discerned first through visual inspection, then through numerical considerations and the use of modern technology – forms an important part of the data analysis curriculum. When teaching about modeling scatter plot data I let my students first draw free-hand graphs, based on eyeball inspection of the data before introducing scatter plot smoothers and curve fitting. Novices in probability and statistics tend to stick to a deterministic-mechanistic view of the world, which either doesn't allow room for chance or knows only trend free randomness. When considering noisy observations in empirical data, the random part has to be separated from the deterministic trend. Here computer simulations offer the opportunity to develop and deepen a sense for random fluctuation in real data in order to focus on the relationship between systemic structure and random noise in data. underlying curve and thus explore the relationship between filtered systematic trend and residual noise Figure 5 illustrates this effect of the bandwidth choice for a simulated data set, using a smoothly weighted moving average. Figure 5 illustrates the effect of the bandwidth for a simulated data set, using a smoothly weighted moving average. In the oversmoothed fit the structure of the main peak is

distorted while the small peak is almost lost. The undersmoothed fit shows too much variability and a number of random peaks.

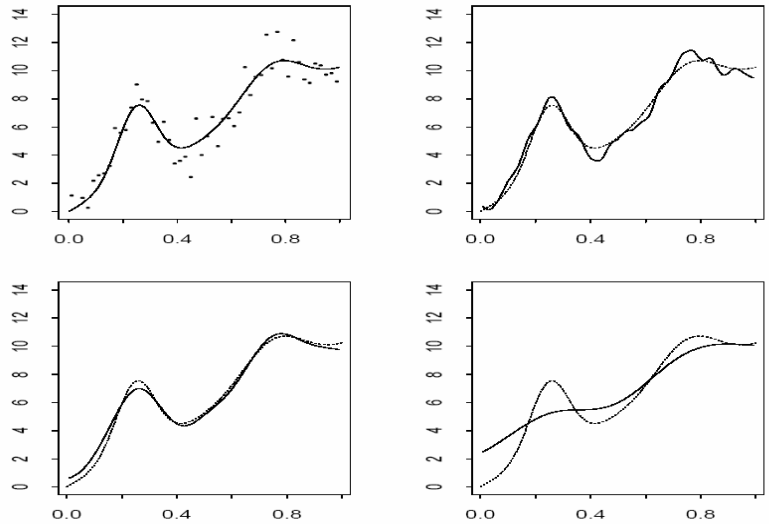

Figure 5: Simulated data with true regression function (left above), an undersmoothed fit to the data (right above), an oversmoothed fit (right below) and an adequately smoothed fit (left below).

CONCLUSION

When modeling scatter plot data students often find it difficult to give the resulting curve a proper interpretation. With some time series data students may be tempted to interpolate while with other data the imposed functional structure may be questionable. A quick introduction of technical concepts like least-squares regression coefficients followed by a model check afterwards through residual analysis draws attention away from the fact that scatter plot modeling–like any statistical activity–represents an attempt to detect structural information from data corrupted by random noise. In contrast, a teaching approach that emphasizes exploring how much of the variation in the data is random and how much is due to structure focuses on concepts, interpretation and understanding rather than on mathematical formulae. Technology supported elementary smoothing – in contrast to classical regression teaching - offers here the opportunity to address directly the issue of detecting trends in the presence of random variation. This approach is not only exploratory but also addresses explicitly the quest for systematic structure in noisy data. It challenges students to express and discuss their ideas, whose initial concept may range from data interpolation to a global average.

REFERENCES

Borovcnik, M. (2005). Probabilistic and Statistical Thinking. CERME, in press.

Engel, J. (2002). Activity-based statistics, computer simulation and formal mathematics. In B. Phillips (Ed.), *Proceedings of the Sixth International Conference on Teaching of Statistics*, Cape Town. Voorburg, The Netherlands: International Statistical Institute.

Erickson, T. (2005). *The Model Shop. Using Data to Learn about Elementary Functions.* Oakland, CA: EEPS media.

Konold, C. (2002). Alternatives to scatterplots. In B. Phillips (Ed.), *Proceedings of the Sixth International Conference on Teaching of Statistics*, Cape Town. Voorburg, The Netherlands: International Statistical Institute.

Simonoff, J. S.(1996). *Smoothing Methods in Statistics*. New York: Springer

Weldon, K. L. (2002). Advance topics for a first service course in statistics. In B. Phillips (Ed.), *Proceedings of the Sixth International Conference on Teaching of Statistics*, Cape Town. Voorburg, The Netherlands: International Statistical Institute.

Wild, C. and M. Pfannkuch (1999). Statistical thinking in empirical enquiry. *International Statistical Review*, 3, 223-266.