THE SIGNIFICANCE OF RESIDUALS FOR MODELING DATA

Markus Vogel
University of Education, Heidelberg, Germany
vogel@ph-heidelberg.de

When modeling data one important rule to consider states: the model should fit the data and not vice versa. There is one problem well known by teachers and researchers: Students often do not realize the gap between data and model, they mistakenly consider the model as reality. In this situation the residuals defined as the difference between data and model become important: they remind us of modeling the trend in data and not the data itself. Whereas the model stands for the explained variation, the residuals represent the unexplained variation. This is at the core of statistical thinking. In this paper, the significance of the residuals for modeling data is examined from different perspectives.

MODELING DATA WITH FUNCTIONS

At all grade levels students in school learn about natural, technical and social phenomena of their daily experience. Often functional dependencies between different factors are of particular interest, like speed of sound or growth of plants. To explore these relationships the students have to analyze data, extract the trend in the data and describe it by a mathematical function. Especially in the natural sciences often elementary functions are appropriate to fit the data. Thus, analyzing data is one important approach to connect the ideas of functional thinking and modeling data in a genuine way as a special part of the overarching idea *Change and Relationships* (s. OECD, 2003).

Starting point for investigating functional relationships between two empirical variables are 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 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 (Engel, 2005). 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. This is at the core of statistical thinking: “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 & Pfannkuch, 1999)

What is this kind of analyzing data in terms of modeling? Data are numbers with context. According to Eichler and Vogel (2009) the data are the foundation of what Blum (2002) calls the *real model* of the situation. The process of mathematising can be described by the signal-noise metaphor of looking for signals in noisy processes (Konold & Pollatsek, 2002). In this context Borovcnik (2004) is talking about the *structural equation* which represents data as decomposed into a signal to be recovered and noise.

$$\text{data} = \text{signal} + \text{noise}$$

In the context of bivariate data, this leads to modeling the covariation with a mathematical function. This is the *mathematical model.* Accordingly, the structural equation becomes:

$$\text{data} = \text{function} + \text{residuals}$$

Figure 1 illustrates the structural equation in a concrete example by means of modeling the linear relationship between temperature and pressure of a gas within a pressure-vessel with constant volume.

The linear function represents the model as “deterministic concentrate of the data” the explained variation, in contrast the residuals (= data - function) stand for the unexplained variation of the data. As difference between function and data the residuals contain information about the goodness of fit: The less structured they appear to be scattered, the better the function fits the trend of the data. The residuals can be modeled stochastically. In the simplest case this noisy factor, usually denoted by ε, is modeled as independent random variable with expected value 0 and a constant variance.

In C. Reading (Ed.), *Data and context in statistics education: Towards an evidence-based society. Proceedings of the Eighth International Conference on Teaching Statistics (ICOTS8, July, 2010), Ljubljana, Slovenia*. Voorburg, The Netherlands: International Statistical Institute. www.stat.auckland.ac.nz/~iase/publications.php [© 2010 ISI/IASE]

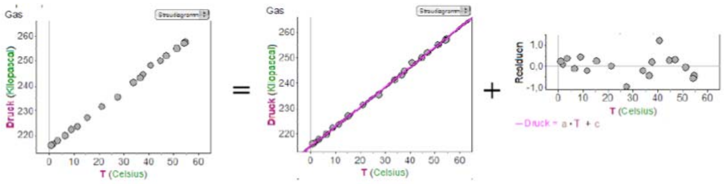

Figure 1. Example for the structural equation data = function + residuals

Students often mistake the model for reality (Hummenberger & Reichel, 1995). This problem often occurs for example when students recover laws of science like Newton’s law of dynamic F = m · a (with F force, m mass and a acceleration) or the relationship between path p and time t in case of free fall p = 0.5 · g · t^2 (without accounting for aerodynamic resistance) by modeling data. At school experiments in the natural sciences are usually carried out to rediscover such laws. Then their validity can become obvious to the students. With regard to the modeling process there is one entrapment in this situation: When students know which outcome of the data analysis process to be expected, they might be misled and see the data not fitting the model as something “wrong”. The residuals might be helpful in this situation: they result from the modeling process and remind us what has been first there - the data but not the function. They witness: it is a process of modeling and there is a gap between data being part of the real world and the mathematical model being part of the mathematical world. The residuals represent an important source of information: they contain that information being disregarded in the (first) cycle of the modeling. Thus, they have to tell us something.

WHAT RESIDUALS CAN TELL US

The residual plot can be explained to the students as being something like a magnifier of the modeling process. Whereas the scatter plot with the modeling function (e.g., middle of Figure 1) shows the modeling situation as a whole, the residual plot magnifies what has been left over. Looking in detail there can be found some interesting information.

Measuring comparable objects of one sort or collecting data of a physical process like free fall, one can observe: with increasing values usually the residuals of the modeling function often also increase. This phenomenon is known as heteroscedasticity. The left-hand scatter plot in Figure 2 illustrates the measurement width vs. length (measured in centimeters) of a sample of shells of the native butter clam (data are available at http://seattlecentral.edu/qelp/sets/001/001.html). The linear function stands for the averaged ratio of width and length. The corresponding residual plot clearly shows residuals of increasing magnitude. The same effect can also be observed when measuring leafs of a tree. Heteroscedasticity can also be found in non-linear situations. The scatter plot on the right of Figure 2 illustrates modeling data of free fall by a quadratic function. With increasing time the residuals also increase.

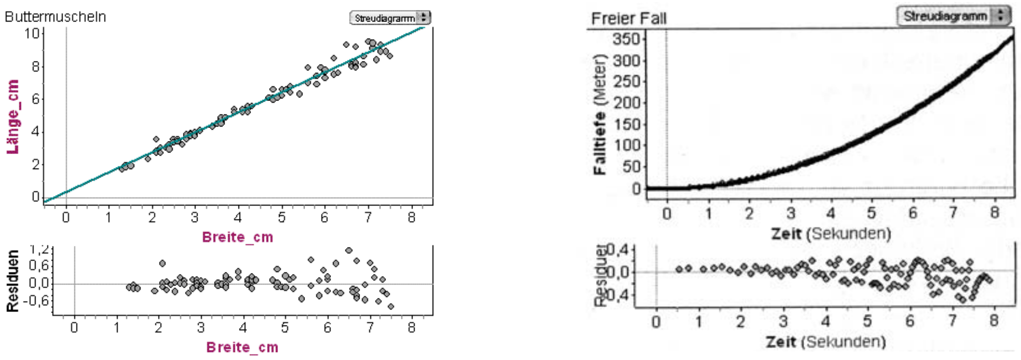

Figure 2. Increasing residuals with increasing values

Whereas in case of the clams the major variability of major shells seem to be a good explanation for this effect, in case of free-fall-data the increasing imprecision of the measurements explains his phenomena of heteroscedasticity plausibly.

From the structural equation “data = function + residuals” directly results: If the residuals increasingly differ from the modeling function either systematically upwards or systematically downwards the function increasingly worsens in being an adequate model. A data-set of radioactive decay of Barium-137m can illustrate what is meant (s. Figure 3 left-hand). The data-fit through an exponential function increasingly worsens. Although there are good reasons to take an exponential function as model, that is the solution of the differential equation $\frac{df(x)}{dx} = c \cdot f(x)$, it does not work for the whole process. What happens is that the curve modeling the exponential decrease does not contain the base rate of natural radiation.

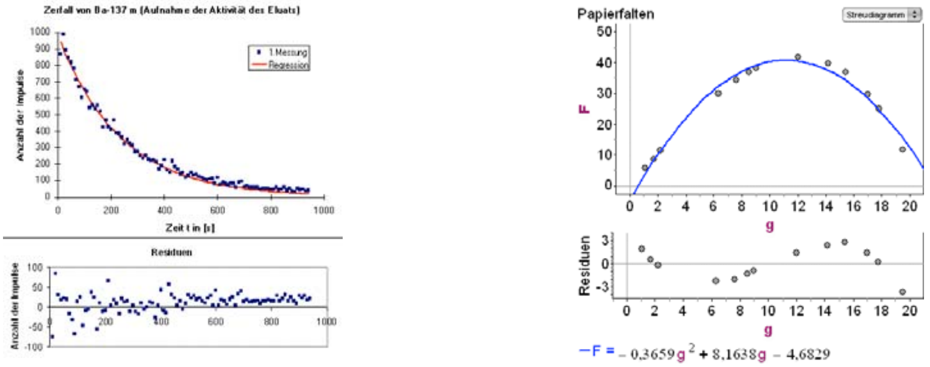

Figure 3. Residuals focus on goodness of fit

The potential of a careful residual analysis on improving the original model is illustrated in Figure 3 (right-hand): The data showing the functional dependency between baseline and area of triangle result from folding a piece of paper in a special way (for details see Biehler et al., 2007). Just by looking at the scatter plot, a quadratic function results in a seemingly satisfactory model. But when inspecting the residual plot, it becomes apparent that a cubic function may be more appropriate.

These examples illustrate that residuals contain an important message. They are much more than a waste of modeling. Erickson (2005) formulates: “How close is close enough? A real function never goes through all the points. It only comes close. There is no firm rule, but we will learn about a tool here - the residual plot - that may be the most important piece of data analysis machinery since the slide rule.”

RESEARCH QUESTIONS

Addressing this enthusiastic statement our approach investigates students’ actions and thinking patterns when modeling data by elementary function according to the signal-noise metaphor. We concentrate on the question, which competencies and skills are required from students in their modeling process when they have to account for residuals by fitting a curve. In particular, in our research we focus on two research questions:

- How do students handle residuals in their modeling processes? How do they appreciate and interpret them with regard to the signal-noise metaphor? Do they accept respectively regard them at all?
- Can residuals actually be helpful for students to understand the modeling process and help them to realize the gap between model and reality?

First of all, we want to get information about how students look at different function models without having learned about the structural equation before.

SOME DETAILS OF A PRELIMINARY STUDY

In a first step of our preliminary study we focus on how students without being experienced in modeling data with functions will judge two different ready-made models given to them. For this, 26 middle-school students (9th grade, aged 14-15) were given two graphs modeling the functional dependency between height and weight of six adult persons (s. Figure 4). Without further information about residuals, they were asked which of the two models they consider to be more appropriate to describe the relationship between body size and weight. Besides asking for a choice between the two models, our main interest focused on their reasoning, why they preferred one model over the other.

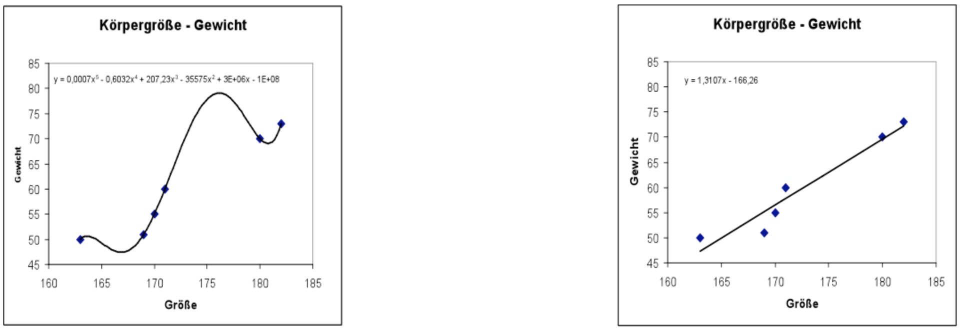

Figure 4. Two different models describing functional dependency between body height and weight of 6 people

There were different arguments for the left-hand model like e.g. “the left-hand model is better because showing the variability more exactly” or “the left-hand model can be read more easily by fitting all points, in contrast the right-hand model seems to be unclear, the model matches only two points.” First results of a qualitative analysis of all students’ comments reveals - besides many questions in detail - a tendency for two patterns lying beyond the arguments for the strongly deterministic model left-hand of Figure 4: the more exact the fit, the more precise and hence the better the model. Moreover, there was a tendency to neglect the data context and argue only on the basis of scatter plots given to the students. These preliminary results will be reviewed by interviewing students who preferred the full-fit model. Thus, more information will be available about students thinking. This information should be helpful to plan and design further studies addressing the research questions mentioned above.

CONCLUSIONS

From a mathematical perspective the residuals enrich the process of modeling bivariate data with functions: they contain information about the phenomenon modeled by function and about the goodness of fit. How students realize and handle residuals, and what has to be done, so they can benefit from them in their activities of modeling data, this is an interesting subject of further research in field of statistical thinking”.

REFERENCES

Biehler, R., Prömmel, A., & Hofmann, T. (2007). Optimales Papierfalten: Ein Beispiel zum Thema “Funktionen und Date”. *Der Mathematikunterricht*, *53*(3), 23-32.

Blum, W. (2002). ICMI-Study 14: Applications and modelling in mathematics education: Discussion document. *Educational studies in mathematics*, *51*, 149–171.

Borovcnik, M. (2005). *Probabilistic and statistical thinking.* URL: http://www.ethikkommission-kaernten.at/lesenswertes/Upload/CERME_Borovcnik_Thinking.pdf (14.01.2010).

Eichler, A., & Vogel, M. (2009). *Leitidee Daten und Zufall.* Wiesbaden: Vieweg+Teubner.

Konold, C., & Pollatsek, A. (2002). Data analysis as the search for signals in noisy processes. *Journal for Research in Mathematics Education,* 33(4), 259-289.

Engel, J. (2005). *Regression modelling in computer-supported learning environments.* http://www.stat.auckland.ac.nz/~iase/publications/17/7G3_ENGE.pdf. (14.01.2010).

Erickson, T. (2005). *The model shop. Using data to learn about elementary functions.* Oakland: eeps media.

Humenberger, J., & Reichel, H.-C. (1995). *Fundamentale Ideen der Angewandten Mathematik und ihre Umsetzung im Unterricht.* Mannheim: BI-Wissenschaftsverlag.

OECD. (2003). *The PISA 2003 assessment framework – mathematics, reading, science and problem solving knowledge and skills.* http://www.pisa.oecd.org/dataoecd/46/14/33694881.pdf. (14.01.2010)

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