A Picture Can be Worth a Thousand Formulae - An Adventure in Model Fitting

Nicole A. Lazar
Department of Statistics
Carnegie Mellon University
Pittsburgh, PA 15213
nlazar@stat.cmu.edu

Summary

We often teach in data analysis courses that one should look at the data before beginning any serious modeling endeavor. Indeed, visualization and modeling should go hand in hand. Each can inform the other and each process is enriched by the use of the other. Judicious use of graphics can save the practitioner from trying to fit overly–complicated models, while at the same time opening a window on interpretation. The power of simple plots as a tool for learning about moderately complex data structures is demonstrated via example.

Key words: Changepoint model, Curve fitting, Permutation test, Visualization

1 Introduction

There is no doubt that statistics has made serious contributions to science, via a principled and mathematically grounded approach to model fitting. In recent years, the growth of computer processing speed and memory have made possible models and methods that could only have been imagined by previous generations of statisticians. With these advances, however, may come the temptation to fit ever more complicated models, even when a relatively simple, straightforward analysis, might suffice. It is important to distinguish between a "simple" analysis and a "simplistic" one. The former aims at providing an interpretable representation of the data, while remaining true to the characteristics and nature of the problem. The latter forces a model on the data, even if it is not particularly useful, in the name of interpretability (an example would be fitting a linear regression model to time series data).

Looking at the data, as a first step in understanding what sorts of simple models might be suitable, and whether a more complex analysis is needed, is something that is often advocated in theory, but much neglected in practice, with the interesting exception, perhaps, of the data mining community, which relies heavily on visualization. Appropriate graphics can always contribute to the modeling process, a fact which is overlooked by many in statistics and ignored by computer scientists, who take an algorithmic, black box, approach to model fitting (Breiman, 2001). We advocate the use of exploratory data analysis (Tukey, 1977) to our students, and rightfully so. Yet, the power of graphics as a tool for model fitting is not, perhaps, duly emphasized. The purpose of this paper is to show, using a real example, the insight that can be achieved from the simple approach to modeling, combined with effective use of graphics (see also the recent article by Gelman, Pasarica and Dodhia, 2002, for an argument for the use of graphical presentations instead of tables). In the next section, the data example is described. Section 3 gives details of the graphical approach, a comparison of different methods, and the conclusions that were drawn. Finally, in Section 4, are closing comments.

2 Description of the Data

Data were collected on the performance of two groups of subjects – patients diagnosed with autism or Asperger's syndrome, and healthy controls – on an eye movement task. Autism is a complex developmental disorder, defined by particular abnormalities in social behavior, language, imaginative play, as well as a narrow and repetitive pattern of behavior. People with autism may be of normal or above normal intelligence. Both autism and Asperger's syndrome are classified in the *Diagnostic and Statistical Manual IV* (*DSM-IV*; American Psychiatric Association, 1994) as *Pervasive Developmental Disorders* (PDD); individuals with PDD may exhibit various of the symptoms described above for autism, in different combinations and with differing degrees of severity. Thus Asperger's syndrome, for example, is characterized by impairments in social interaction, but generally there is no delay in language development, and intelligence tests in the average or above average range.

The eye movement task proceeded as follows. Subjects were required to fixate on a central stimulus for 3 to 5 seconds, after which a peripheral stimulus appeared for 1.5 seconds at one of three locations, 8, 16, or 24 degrees, to the left or right of center fixation. They were not to look at the light, but instead were to move their eyes immediately in the opposite direction to a point equidistant from center fixation; 1.5 seconds after the peripheral stimulus was presented, it was extinguished and feedback was provided by a light appearing at the location where subjects should have been fixating. If a subject made two consecutive errors during testing (i.e., she looked at the peripheral target instead of to the opposite side), the tester re–alerted the subject to the task instructions to ensure that poor performance was not a result of having forgotten task instructions. Thirty–six trials were presented. The percent of trials in which the subject looked toward the peripheral targets (response suppression errors) was recorded.

Subjects in this study ranged in age from 8 to 53, Data were cross–sectional, meaning that each subject was measured at a single age, as opposed to longitudinal, which would have involved following subjects over time. In all, there were 135 controls, and 85 patients. Young children and people with neurological disease respond reflexively to stimuli, so they should tend to look to the light. Healthy adults respond on the basis of plans and intentions, so they should be able to suppress the instinct to look in the direction of the stimulus. The percentage of *prosaccade errors*, that is, cases in which the subject looked to the light instead of away, should thus tend to be higher in patients overall. Similarly, for healthy controls, the error rate should be higher for children, decreasing through the teen years until reaching a plateau. The developmental pattern for the patient group is not *a priori* clear.

Questions of interest centered on comparing the developmental curves for the two groups. For the control group, as described above, theory predicts that the rate of prosaccade errors will decline from age 8 until some point in the teen years, and will remain steady after that. One question then, was at what age does this stabilization occur? Second, what happens with the patient group? Do the patients follow the same general trend as that of the healthy controls? Finally, at what ages, if any, are the two curves significantly different?

The data are plotted in Figure 1. A number of immediate observations are possible:

1. The patient data look quite flat, although the paucity of subjects older than 25 in this

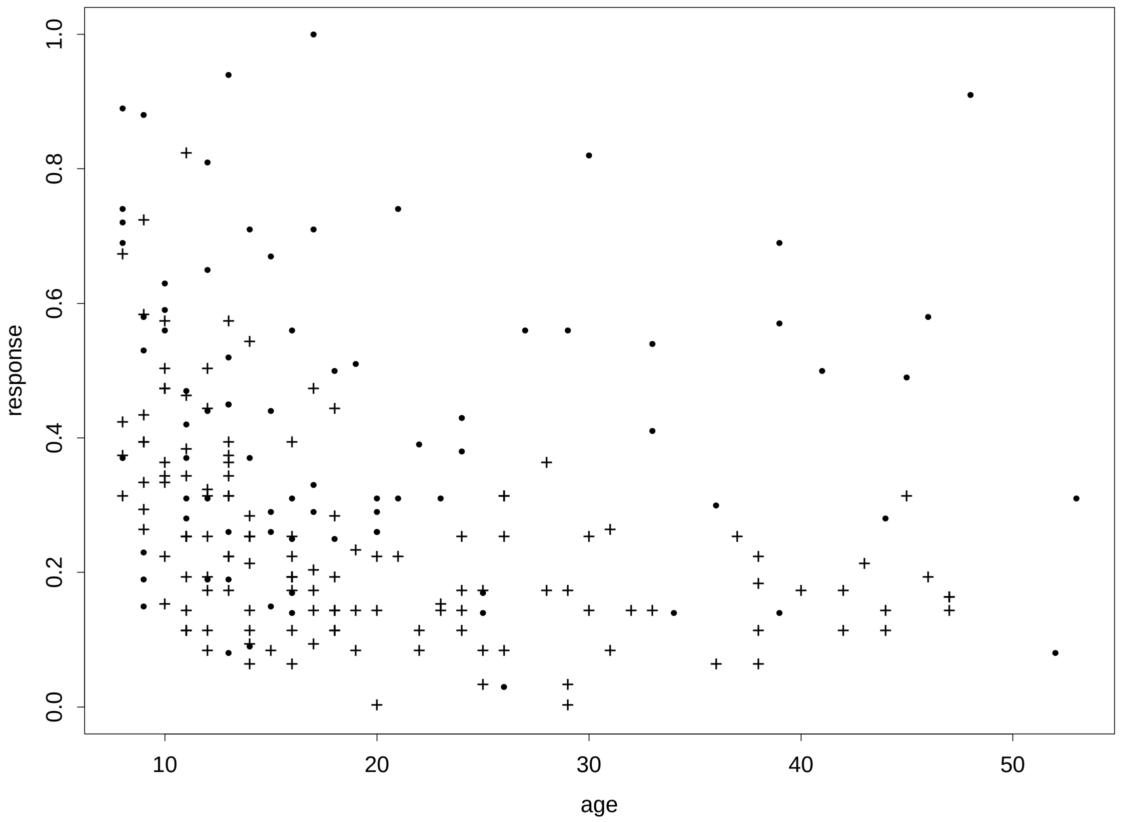


Figure 1: **Data for both groups of subjects, plotted on the same axes. Dots represent patients and plus signs represent controls.**

group makes it hard to draw definite conclusions.

2. The control data display a trend, with a decline in the error rate, indicating improved task performance (fewer prosaccade errors), through the late teen years, leveling off by around age 20.

3. The patient group overall seems to be located at a higher level than the control group, indicating poorer performance on the task (more prosaccade errors).

3 Visualization and Model Fitting

Simple inspection of the raw data did not provide a clear suitable model, although there was at least an initial indication that the age–task curve for the patients might be flat. As a first attempt at model fitting, therefore, I tried a series of nonparametric curves, using various smoothers and values of their tuning parameters. A representative example of the results, using the `lowess` function in Splus (Venables and Ripley, 1997), is displayed in Figure 2. Examination of the Figure reveals that the general trend in the nonparametric fit accords well with the first impressions from looking at the data, although the argument for describing the patient group by the mean line is weakened.

Additional interesting features that were not apparent from the initial inspection of the data are evident in Figure 2. First of all, the patient responses appear to follow those of the controls through the teen years, although there is evidence of an increased difference after age 20 or so. Second, the lines are roughly parallel over much of the range of the data. Even though there is considerable overlap of the responses of the two groups, especially in the early teen years, the lines nowhere cross – patients always make more mistakes than controls. Third, the plateaus reached by the two groups when performance levels out, are different, around 0.2 for the controls and 0.4 for the patients. In other words, healthy adults make prosaccade errors in roughly twenty percent of the trials; autistic adults make such errors in almost half of the trials.

I now had a plausible model in hand, one that I could explain to the scientist, one that would permit us to reach conclusions about the phenomenon in question. I was curious, though, to see how much these conclusions depended on the particular fitting technique. Plotted in Figure

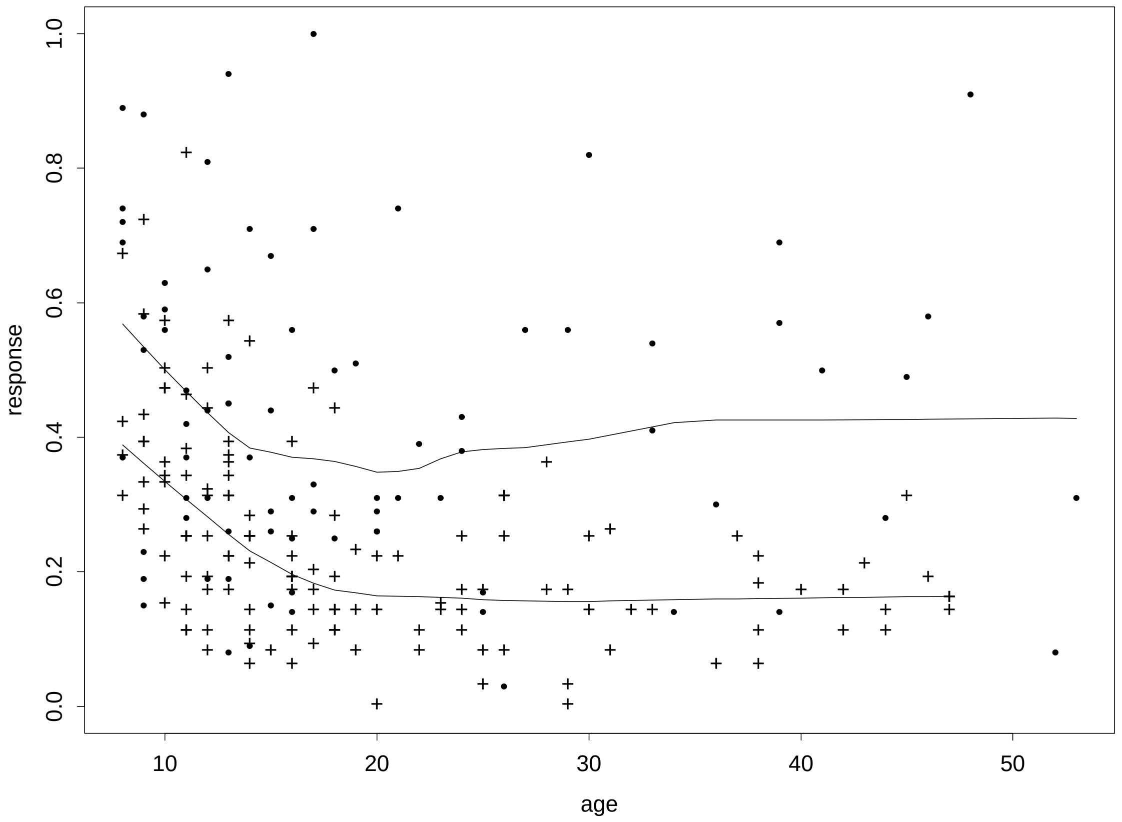
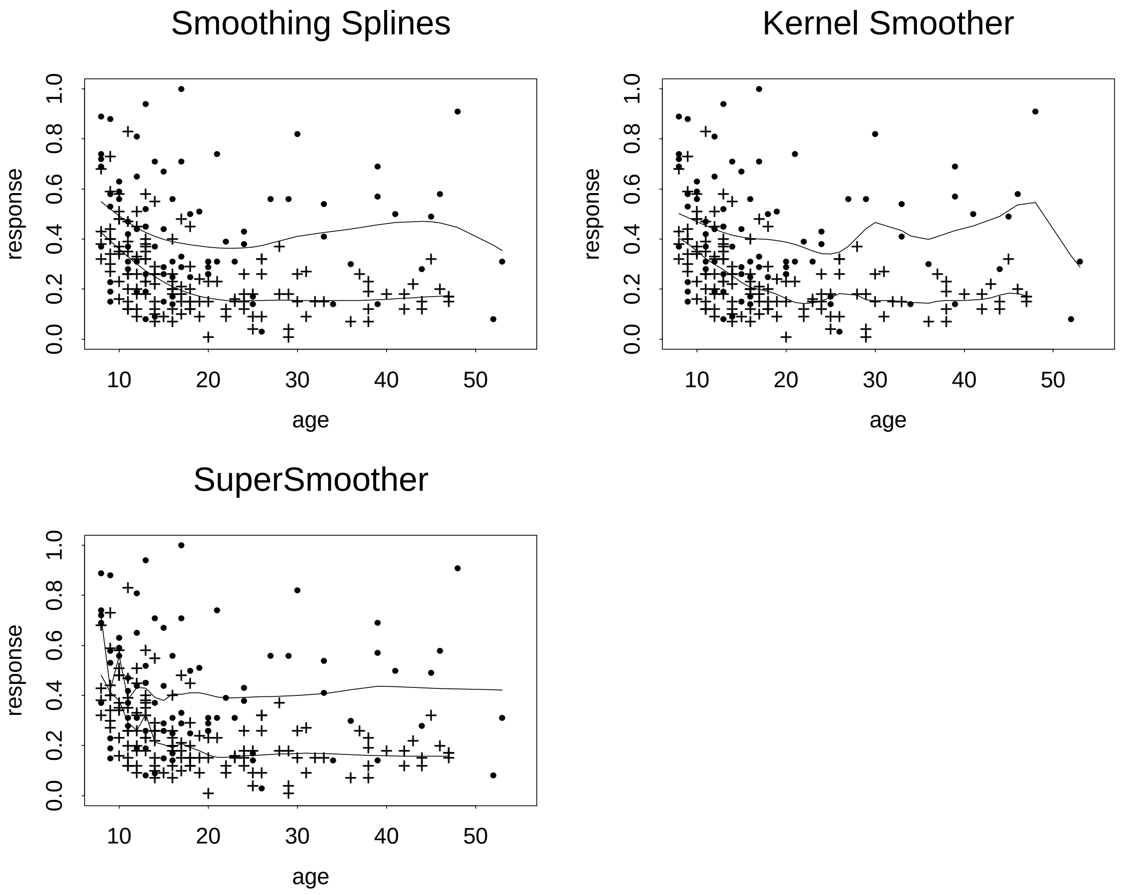


Figure 2: **Nonparametric curve fits. The upper line is the patients, the lower line is the controls. The fits use the default smoothing parameters in the Splus function `lowess`.**

3 are the results from several different procedures. Inspection of the Figure shows that the control results especially are robust to a variety of smoothing procedures. Smoothing splines and a kernel smoother with a normal kernel give essentially the same picture as the lowess fit. The supersmoother (`supsmu` function in Splus) agrees with the others in substance, but is slightly more wiggly in the early years of the sample. In all cases, different values of the smoothing parameters have little effect on the substantive conclusions. On the other hand, the results for the patients are more ambiguous. Smoothing splines and kernel smoother techniques point to an increase in the percentage of errors made by older patients over those in their twenties and thirties. This is not apparent with the supersmoother. Furthermore, for some values of the smoothing parameters, there is an indication that just fitting the mean line to the data would suffice, although the evidence is by no means conclusive.

Figure 4 shows the fit from a piecewise linear model, for each group separately. One aim of this exercise being simplicity, the changepoints were found in a somewhat *ad hoc* way. Placing the changepoint at each age, I fit the piecewise linear models and looked for the one that minimized the residual sum of squares. For controls, this method worked as I had expected, with the residual sum of squares decreasing up to a certain age, and then starting to increase again. This gave an estimated changepoint at age 15, matching with previous observations and one of the research hypotheses. For patients, the method was not as clearly successful, since the residual sum of squares did not monotonically decrease to a minimum and then monotonically increase with age. Rather, especially at the younger ages, the residual sum of squares oscillated, and then began monotonically increasing at age 14. The absolute minimum was found by placing the changepoint at age 9. This does not fit any scientific theory, but does lend further credence to the supposition that the mean model is adequate for the patient group. Other observations,


Smoothing Splines

Kernel Smoother

SuperSmoother

Figure 3: **Fits from different smoothing techniques:** `smooth.spline`, `ksmooth` **and** `supsmu` **in Splus. In all cases, the lower line is the controls and the upper line the patients. Splus default smoothing parameters are not always used, as they may result in very noisy fits.**

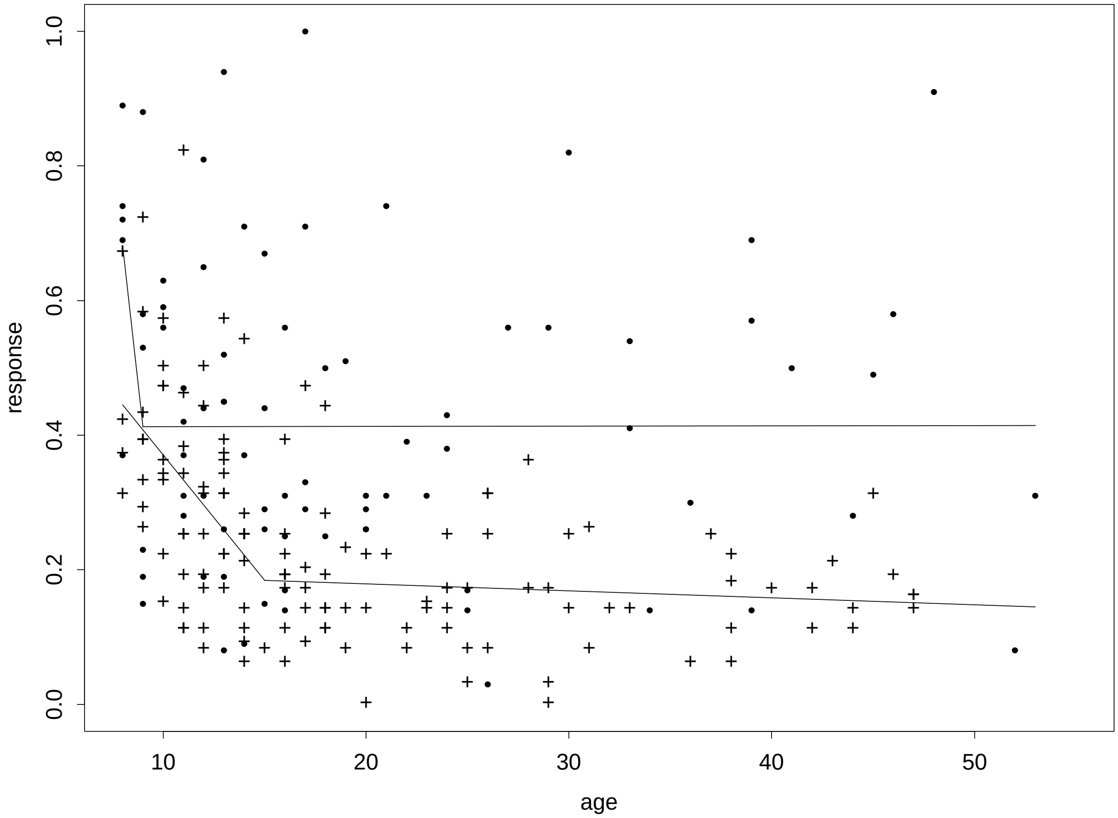


Figure 4: **Piecewise linear fits. The upper line is the patients, the lower line is the controls. For the controls, there is a clear changepoint at age 15, as explained in the text. For the patients, the changepoint is put at age 9, so that the mean model might be adequate.**

namely that the curves are approximately parallel through the adult years, and that there is no overlap of the patient and control curves, are also borne out by this analysis. These general conclusions are not substantially modified by placing the changepoint at age 14 for the patient group, as might be warranted by the observation that this is the age at which the residual sum of squares starts to increase monotonically (see Figure 5). One change is that with the shift of the changepoint, the difference between the two groups seems to increase with age. This also coincides with the nonparametric analysis.

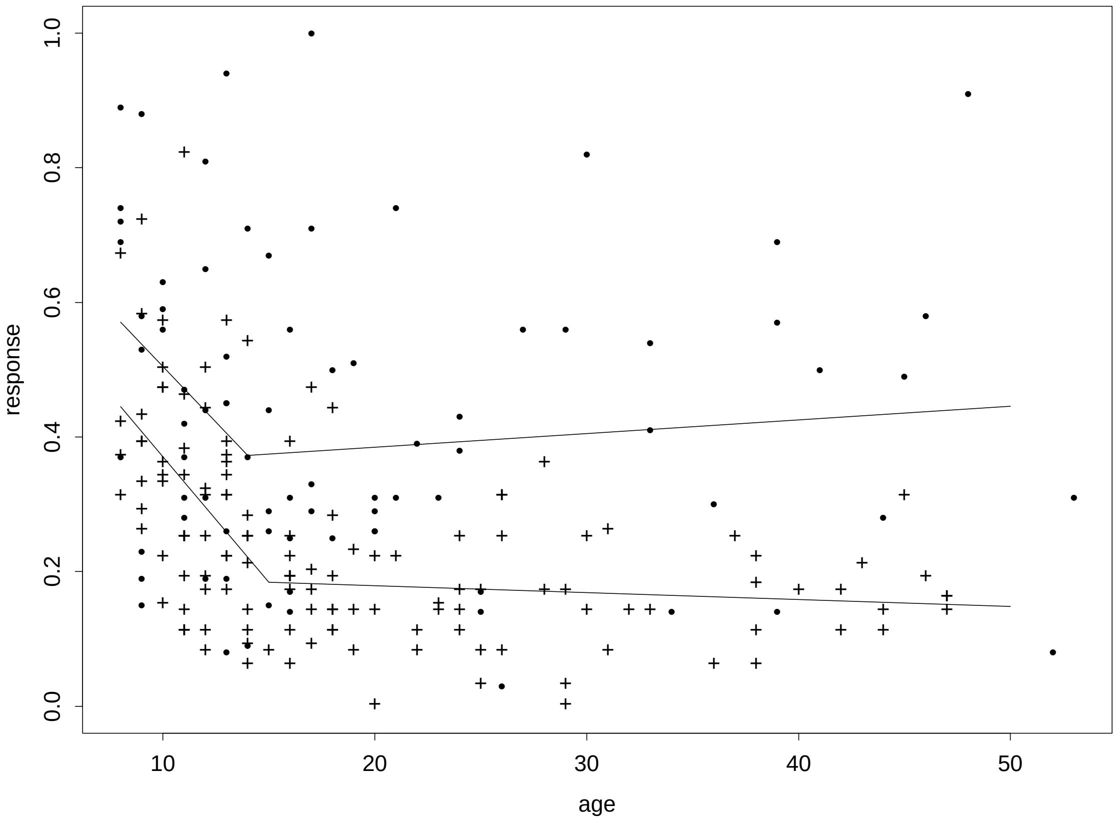


Figure 5: **Piecewise linear fits. The upper line is the patients, the lower line is the controls. For the controls, there is a clear changepoint at age 15, as explained in the text. For the patients, the changepoint is now put at age 14. As age increases, the difference between the two groups also seems to increase.**

Interestingly, even for the patient group, the model with a single changepoint at age 9 improves slightly over both the mean model and the simple linear regression model with no changepoint. Thus there is some evidence that the changepoint does capture a real feature of the data, although the scientific interpretation in this case is not clear.

Beyond having a model for the developmental patterns of the two groups, there was still an inferential question of interest to the researchers, namely: do the groups differ, and if so, at what ages? To answer this question, I decided to go back to the nonparametric `lowess` fits, as they provided an extra level of sensitivity to fluctuations in the data that the piecewise linear models didn't have. Staying within the simple, graphics driven approach, I found the difference between the fitted values of the two nonparametric curves, on an age by age basis (therefore, for ages at which there weren't subjects in both groups, no difference could be calculated). Then, I permuted the data into two groups, in mimicry of the original structure, found the nonparametric fits, and calculated the difference in the fitted values between the curves. I repeated this exercise 500 times, giving me 500 permutation difference curves.

Based on these 500 permutations of the data, I calculated 95% pointwise confidence intervals for the difference between the two nonparametric curves. The results are plotted in Figure 6, marked by + signs. The difference between the fits for the two groups in the study is the solid line. Nowhere does the line dip in to the confidence bands, indicating that the two groups are different everywhere. Note that the line itself is not strictly flat, that is, the lines aren't perfectly parallel. Indeed, the difference is slightly increasing with age. Also, it is worth noting that the confidence bands at the later ages are based on fewer points, as there are relatively few subjects older than 30 (36 out of 220 subjects; of these 22 are controls and 14 are patients). If there are not at least two subjects at a given age, that age cannot be used in the calculation

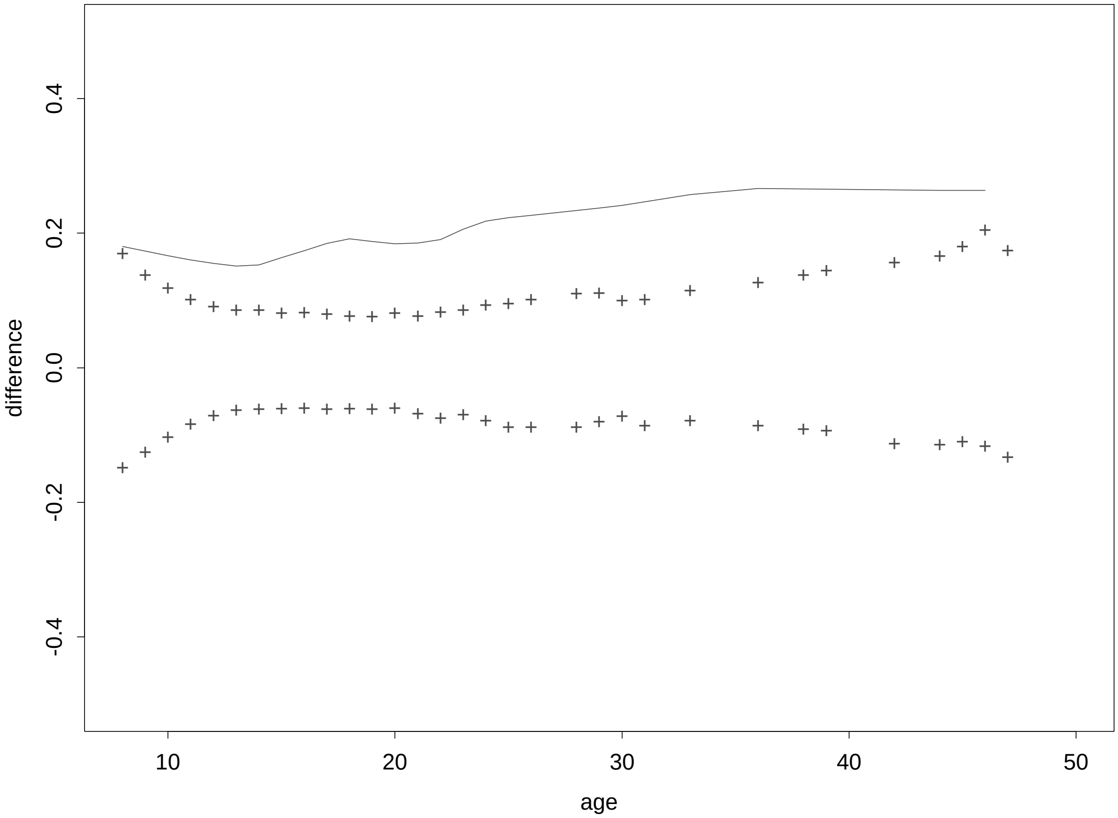


Figure 6: **The solid line is the difference between the loess fits for the patients and controls. Marked by the + signs are pointwise 95% permutation confidence intervals, based on 500 permutations of the data. Gaps in the confidence bands are ages with fewer than two subjects.**

of the permutation confidence bands.

For comparison, Figure 7 shows the difference curve and 95% pointwise permutation confidence bands for a similar study, based on a separate sample, aimed at exploring the differences in prosaccade error rates between males and females, as a function of age. As can be seen, the difference between the two groups is within, or very close to, the confidence bounds, with the exception of the teen years, when the boys outperform the girls.

While the permutation confidence intervals required a small amount of programming and

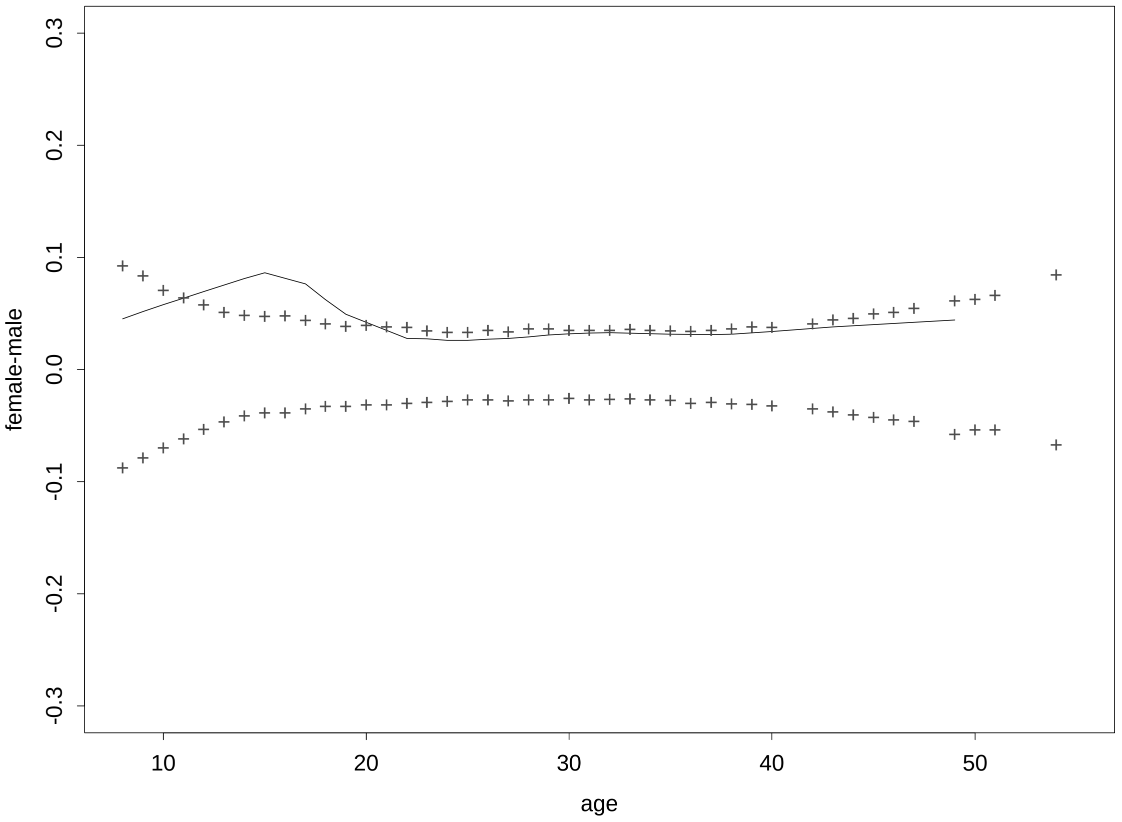


Figure 7: **The solid line is the difference between the loess fits for males and females. Marked by the + signs are pointwise 95% permutation confidence intervals, based on 1000 permutations of the data. Gaps in the confidence bands are ages with fewer than two subjects.**

tailoring to the specific data set at hand, making it more difficult to automate the procedure for use by my colleagues, again I was very easily able to explain to them what I had done and the logic behind it.

4 Conclusions

In the model fitting problem presented to me, I tried to adhere to a number of principles. First, the belief that visualization is a powerful tool in understanding the data, and that it can play a crucial role in modeling. Second, that simple models, although not simplistic ones, are usually "good enough" in the sense that they will give an approximation to the story that the data can tell. A sophisticated model will be able to tease out nuances that a simple one will overlook, and this ability should not be underestimated. However, in real world work, it is often necessary to have a quick and dirty method that gets at the essential point. This is where simple models, if done well, can shine.

Third, a method that results in a model that is easily explained and that can be implemented by the researchers themselves (once they understand what the model does and does not do for them) is preferable to a complicated procedure that looks like magic to those who need to use it on a daily basis. Especially when the analysis to be performed needs to be a routine part of the researcher's toolbox, it behooves us as methodologists to provide them with tools that they feel comfortable with and can explain to others. The setting that generated these data is an experimental paradigm frequently used by a local research group; the piecewise linear model therefore had the added bonus of being easily implemented by a non–statistician. Indeed, I sat down with a research assistant from the group, and was very easily able to explain to her how to find the changepoint, and plot the fitted models. More importantly, she understood the meaning of the model and hence felt comfortable both using it and explaining it to others.

Finally, it helps to look at the same data set from a variety of perspectives, trying out models of differing levels of complexity, to see if a consistent story emerges. To the extent that it doesn't, a more elaborate scheme might be required.

ACKNOWLEDGMENTS: The author thanks Dr. John Sweeney, Department of Psychiatry, University of Illinois, Chicago and Dr. Beatriz Luna, Department of Psychiatry, Western Psychiatric Institute and Clinic, University of Pittsburgh, for providing the problem setting and the data sets. This work was partially supported by NICHHD grant U19-HD35469.

REFERENCES

American Psychiatric Association (1994) *Diagnostic and Statistical Manual of Mental Disorders, Fourth Edition.*

Breiman, L. (2001) Statistical modeling: The two cultures (with discussion). *Statistical Science*, **16**, 199–231.

Gelman, A., Pasarica, C. and Dodhia, R. (2002) Let's practice what we preach: Turning tables into graphs. *The American Statistician*, **56**, 121–130.

Tukey, J.W. (1977) *Exploratory Data Analysis*, New York: Addison–Wesley.

Venables, W.N. and Ripley, B.D. (1997) *Modern Applied Statistics with S-PLUS*, Second Edition. New York: Springer.