Finding Outliers in Models of Spatial Data

David W. Scott
Dept of Statistics, MS-138
Rice University
Houston, TX 77005-1892
scottdw@stat.rice.edu
www.stat.rice.edu/∼scottdw

J. Blair Christian
Dept of Statistics, MS-138
Rice University
Houston, TX 77005-1892
blairc@stat.rice.edu
www.stat.rice.edu/∼blairc

Abstract

Statistical models fit to data often require extensive and challenging re-estimation before achieving final form. For example, outliers can adversely affect fits. In other cases involving spatial data, a cluster may exist for which the model is incorrect, also adversely affecting the fit to the "good" data. In both cases, estimate residuals must be checked and rechecked until the data are cleaned and the appropriate model found. In this article, we demonstrate an algorithm that fits models to the largest subset of the data that is appropriate. Specifically, if a hypothesized linear regression model fits ninety percent of the data, our algorithm can not only find an excellent fit as if only that "good" data were presented, but will also highlight the ten percent of the "bad" data that is not fit. Our work in digital government has focused on mapping data. Thus we illustrate how models fit to census track data work, and how the data in the "bad" set can be viewed spatially through ArcView or other tools. This approach greatly simplifies the task of modeling spatial data, and makes us of advanced map visualization tools to understand the nature of subsets of the data for which the model is not appropriate.

1 Introduction

The study of spatial data has an important role in the mission of governmental entities. The topic has drawn the interest of statisticians (Cressie, 1991) and geographers (MacEachren, 1995), among others. The main presentation tool is the choropleth map, which displays small geographical units by means of color shading (Tobler, 1978; Tufte, 1990). Two outstanding collections of such maps may be found in Pickle et al. (1996) and Brewer and Suchan (2001). Color choice is critical to successful presentation, and Brewer (1999) has described a theory of successful color mapping. Choropleth maps are inherently discontinuous. Smoothing methodologies such as kriging and nonparametric regression can aid visual understanding (Scott, 1992, 2000), but at the cost of some statistical bias. Whittaker and Scott (1994, 1999) have argued that the needs of many decision makers are aided by such a big picture, without insisting on complete fidelity to the raw data.

One of the many challenges in federal mapping is a better understanding of how many variables are related spatially to variables of interest. One classical dataset much studied is the Boston housing data (Harrison et al, 1978); the goal is to predict median housing prices, but much of the data does not fit a single multivariate model. Robust methods are often applied to such data, but are iterative and sometimes difficult to interpret.

We have introduced an alternative multivariate regression fitting algorithm, called L2E (Scott, 2001). In place of a maximum likelihood or a least-squares criterion, L2E attempts to estimate parameters so the *shape* of the residuals is as close to Normal as possible. Why does this alternative criterion help? If the data contain a cluster of "bad" data fitted by least-squares, the residuals will not be Normal and may be skewed, for example. Regression diagnostic plots must be carefully screened and interpreted in order to identify and then correct these practical problems. Making the residuals look Normal will avoid such difficult diagnostic steps. Instead, 90% of the residuals will look Normal (with a mean of 0), and the 10% of residuals corresponding to the bad data will be clearly shown.

In the case of spatial data, another diagnostic is available. Namely, the values of the residuals may be plotted spatially in order to better understand the nature of the "bad" data. Of course, the bad data may in fact be the most interesting data. A simple estimation example is presented.

2 Fitting Multivariate Regressions by L2E

Given a variable of interest, y, and a number of covariates, $(x_1, x_2, \ldots, x_p)$, collected over a number of spatial units (census tracks, for example), we seek to model y as a simple function of the predictors, $\{x_k\}$:

$$\hat{y}(x_1, x_2, \ldots, x_p) = \beta_0 + \sum_{j=1}^{p} \beta_j \cdot x_j \, .$$

Given estimates of the $\beta_j' s$, a residual for the i-th case would be computed according to the formula:

$$\epsilon_i = y_i - \beta_0 - \sum_{j=1}^{p} \beta_{ij} \cdot x_{ij} \, .$$

If there was a cluster of bad data, then the residuals would not look Normal and be centered at 0.

With the L2E, least-squares is replaced by another criterion, which can be optimized using standard software, for example, *nlmin* in the Splus package. The criterion to minimize is simply

$$\frac{1}{2\sqrt{\pi}\sigma_\epsilon} - \frac{2}{n} \sum_{i=1}^{n} \phi(\epsilon_i | 0, \sigma_\epsilon),$$

where σ_ϵ^2 is the variance of the residuals and $\phi(x|\mu, \sigma^2)$ is the density of a Normal random variable.

The new exciting extension which we will illustrate below was introduced by Scott and Szewcyzk (2003). Specifically, instead of modeling the residuals as Normal, $N(0, \sigma_\epsilon^2)$, we add an additional weight parameter, w, and use the residual model $w \cdot N(0, \sigma_\epsilon^2)$.

What this very unusual (and unexpected) model is really doing can be made clearer by the following observation. If we must deal with a "bad" data cluster, then the residual plot will have a separate component for that cluster. In other words, we believe the residual density is actually a mixture of two components, one centered at 0 (the good data), and a second elsewhere (and of unknown shape). The "magic" of the L2E algorithm, as described in Scott and Szewczyk (2001), is that we can use L2E to estimate only the major mixture component. The criterion given above is modified only slightly to:

$$\frac{w^2}{2\sqrt{\pi}\sigma_\epsilon} - \frac{2\,w}{n}\sum_{i=1}^{n}\phi(\epsilon_i|0,\sigma_\epsilon).$$

The optimization is over the parameters w and σ_ϵ, as well as the parameters $\{\beta_0, \beta_1, \ldots, \beta_p\}$, which are used implicitly to recompute the estimated residuals, $\{\epsilon_i\}$.

The estimated value of the weight, w, indicates the amount of the data that the L2E algorithm believes is being fitted by the multivariate regression model. Thus, when all works well, the investigator simultaneously obtains a model fit that is very good as though only the "good" data were fitted, as well as an explicit estimate of the fraction of "bad" data. By sorting on the magnitude of the fitted residuals, the "suspect" data can clearly be identified.

The output of such models can be used in our previous digital government work fitting smooth slices of regression functions (Whittaker and Scott, 1994; Scott and Whittaker, 1996; Whittaker and Scott, 1999; Scott and Wojciechowski, 2001; Christian and Scott, 2002).

3 Application to Boston Housing Data

The Boston housing data were originally modeled to determine if levels of air pollution appeared to be correlated with housing prices. Data were assembled for the 506 census tracts in greater Boston in the early 1970's. The 13 predictor variables included information on per capita crime rates, proportion of residential land zoned for large lots, proportion of non-retail business acres, average age and number of rooms per dwelling, full-value property-tax rates, and pupil-teacher ratios, in addition to the nitric oxides concentration (the pollution variable). Dummy variables were added to account for spatial correlation, for example, adjacency to the Charles River, distances to five Boston employment centers, and accessibility to radial highways.

We begin by examining the residuals of a least-squares fit to these data; see Figure 1a. We can easily see that the shape is far from the assumed shape $N(0, \sigma_\epsilon^2)$. The residuals seem skewed to the right, with a number of outliers on the right (and a couple on the left, too).

We next fit these data by the L2E method, including the weight parameter, w. The fitted value of w was 84.5%. Thus from the beginning we have an indication that the assumed multivariate regression model will not fit all of these data. In fact, 15.5% of the data are not well-fit by the model. (However, 84.5% are well-fit, which is not bad in many situations.) When we examine the residuals of the L2E fit to these data in Figure 1b, we can easily see that the majority of the data are well-fitted by the assumed shape $2 \cdot N(0, \sigma_\epsilon^2)$. The residuals seem skewed to the right, with a number of outliers on the right (and a couple on the left, too). Note that the Normal curve depicted in this figure has area 84.5%, rather than a full 100%, since we are fitting the partial mixture model.

Now that we have the standardized residuals from the L2E fit, we can examine the locations of the census tracts that are "different" from the vast majority of the data. We expect most Normal residuals to fall in the range, $(-3, 3)$. Here, of course, only 84.5% fall in that range.

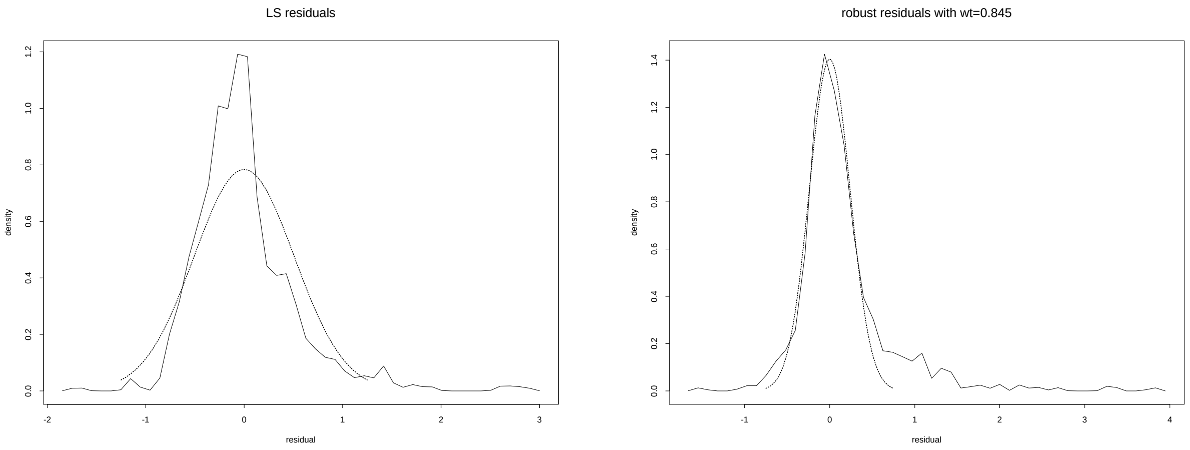

Figure 1: Kernel smooth of residual density (solid line), with assumed Normal fit (dotted line). The least-squares case is shown on the left, and the L2E case on the right.

In Figure 2, we show a choropleth map of the L2E residuals for all 506 census tracts. The dark blue residuals (corresponding to predictions that are much too low) are indeed clustered in the western region. Armed with this knowledge, the investigator could attempt to add an additional indicator variable so that the model would fit a larger fraction of the data.

The dark red residuals (corresponding to predictions that are too high) cannot be seen very well in Figure 2, since they are all located in the downtown region. In Figure 3, we zoom in on that region. We see that these are very special tracts, and probably should be analyzed separately from the rest of the data.

4 Discussion and Future Directions

The digital government work of our research team has focused on a number of statistical tools and techniques which can handle many of challenging real situations. In this paper, we have examined the all-too-common problem of statistical models which fit a large fraction of the data, but not all of the data. The common practice of iteratively trying to figure out which data points are the problematic ones is very difficult and, in our experience, often leads to unnecessary failure. This is not too surprising as the number of possible remedies with 13 predictor variables is almost unlimited, and both a novice and experienced investigator may not be able to find the correct combination of fixes.

In contrast, our new fitting technique goes straight to the heart of the issue, finding the "best" for the largest fraction of "good" data that can be identified. Diagnostics are much easier with this approach. The generality of our approach will become clearer, as regression is the fundamental model for many if not most data analyses.

Acknowledgments: Research was supported in part by the National Science Foundation grants NSF EIA-9983459 (digital government) and DMS 99-71797 (non-parametric methodology). The

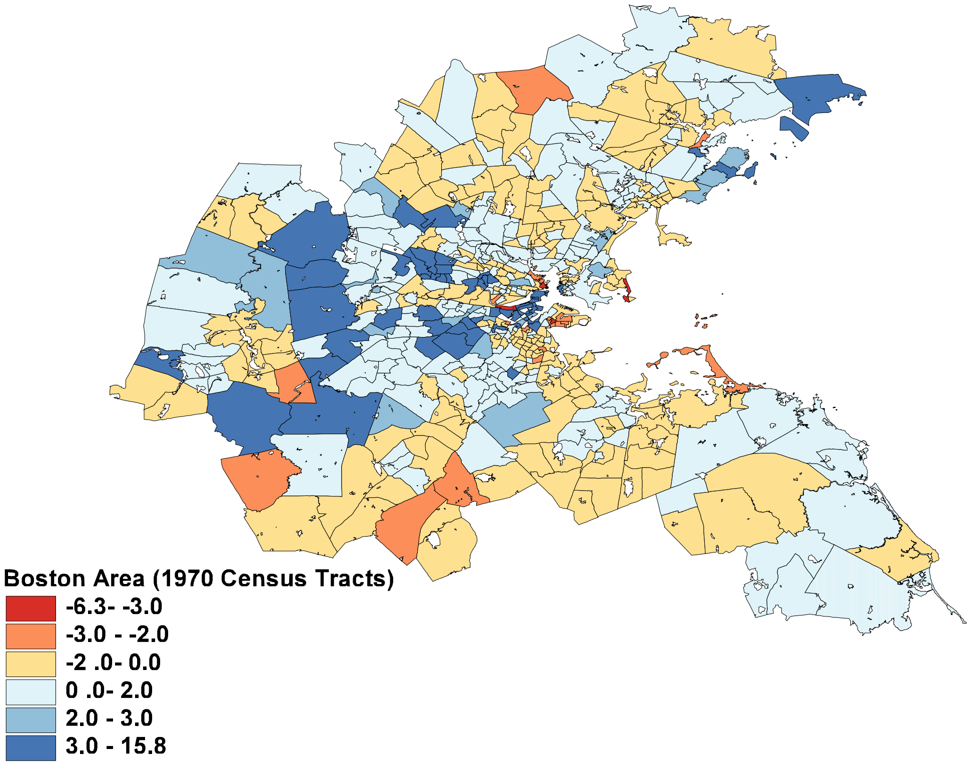

Figure 2: Color-coded standardized residuals by census tract.

authors would like to thank Linda Pickle and Sue Bell for their collaboration, as well as our DG collaborators Drs. Carr, MacEachren, and Brewer.

References

Brewer, Cynthia A. (1999), "Color Use Guidelines for Data Representation," Proceedings of the Section on Statistical Graphics, American Statistical Association, Baltimore, pp. 55-60.

Brewer, Cynthia A., and Trudy A. Suchan (2001), *Mapping Census 2000: The Geography of U.S. Diversity,* CENSR/01-1, U.S. Government Printing Office, Washington DC.

Cressie, Noel A. C. (1991), *Statistics for spatial data,* Wiley-Interscience, New York.

Harrison, D. and Rubinfeld, D.L. (1978), "Hedonic Housing Prices and the Demand for Clean Air," Journal of Environmental Economics and Management, 5, 81-102.

MacEachren, Alan M. (1995), *How maps work,* Guilford Publications.

Pickle, L. W., Mungiole, M., Jones, G. K., and White, A. A. (1996), *Atlas of United States Mortality,* HHS-CDC, Hyattsville, MD.

Scott, D.W. (1992), *Multivariate Density Estimation: Theory, Practice, and Visualization,* John Wiley, New York.

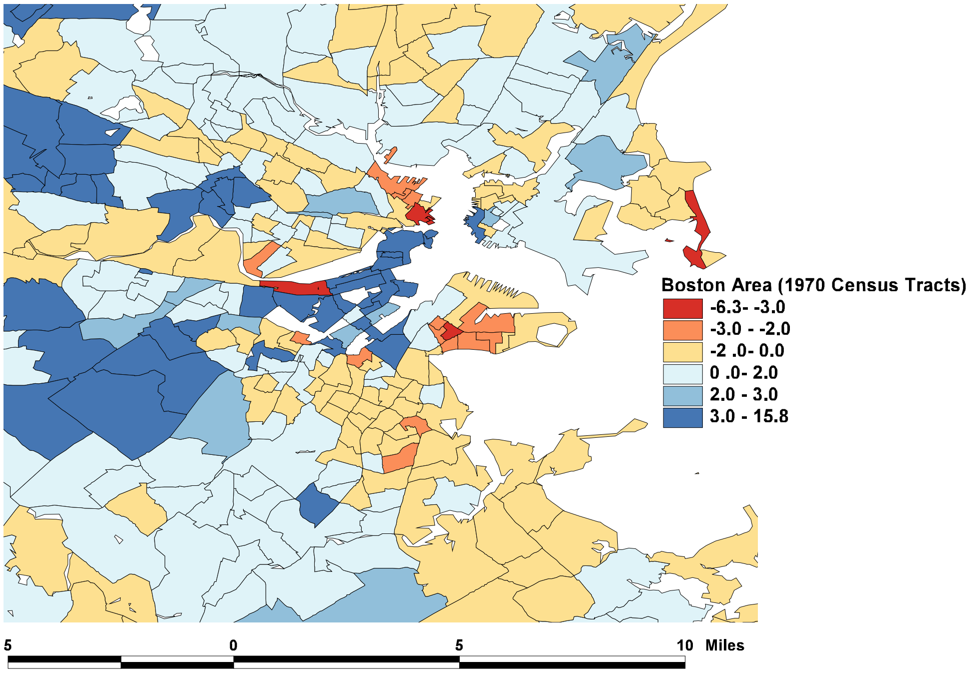

Figure 3: Color-coded standardized residuals by census tract.

Scott, D.W. (2000), "Multidimensional Smoothing and Visualization," In *Smoothing and Regression. Approaches, Computation and Application,* M. G. Schimek, Ed., John Wiley, New York, pp. 451-470 (with 5 color plates).

Scott, D.W. (2001), "Parametric Statistical Modeling by Minimum Integrated Square Error," Technometrics, 43, pp. 274–285.

Scott, D.W. and Whittaker, G. (1996), "Multivariate Applications of the ASH in Regression," Communications in Statistics, 25, pp. 2521-2530.

Tobler, W. R. (1978), "Data Structures for Cartographic Analysis and Display", Proceedings of the Computer Science and Statistics 11th Annual Symposium on the Interface, pp. 134-139.

Tufte, Edward R. (1990), *Envisioning information,* Graphics Press, Cheshire, CT.

Whittaker, G. and Scott, D.W. (1994), "Spatial Estimation and Presentation of Regression Surfaces in Several Variables Via the Averaged Shifted Histogram," Computing Science and Statistics, 26, pp. 8-17.

Whittaker, G. and Scott, D.W. (1999), "Nonparametric Regression for Analysis of Complex Surveys and Geographic Visualization," Sankhya, Series B, special issue on small-area sampling, P. Lahiri and M. Ghost, Eds., 61, pp. 202-227.

Scott, D.W. and Szewczyk, W.F. (2003), "The Stochastic Mode Tree and Clustering," J. Comp. Graph. Stat., to appear.