Anderson-Darling Regression with two examples from biofilm engineering
by Don Simone Daly

A thesis submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy in Statistics
Montana State University

Abstract:
This dissertation explores the use of an optimization criterion based on the Anderson-Darling statistic (AD), a goodness-of-fit measure, to estimate the mean response in a variety of regression settings. This approach is best suited to the regression model where the distribution of the random component, and the linkage between this component and the mean response are known. In this situation, the AD model-fitting technique can outperform other techniques which do not use directly the available information about the distribution and linkage. This work, Anderson-Darling Regression (ADR), is an extension of Minimum Distance Estimation (MDE), pioneered by statisticians such as Parr (1981) and Boos (1981).

A terse history of MDE is presented, with an emphasis on its potential role in parametric and nonparametric regression. An ADR approach is described that accommodates many regression models: parametric and nonparametric, normal and non-normal, linear and nonlinear, natural and transformed. The ADR method can be applied easily to nonstandard regession models. ADR's ease of implementation is illustrated with two examples from biofilm engineering, and using conventional statistical software.

The ADR method does have limitations. Specifically, it may be seriously handicapped when the model is mis-identified, or when the estimator is biased. Therefore, a rigorous modeling approach is required that stresses model validation and diagnostics. On the plus side, the well-fit ADR model has residuals fitting the assumed random distribution — a definite benefit when assessing modeling assumptions, estimating standard errors and performing hypothesis tests.

ANDERSON-DARLING REGRESSION WITH TWO EXAMPLES FROM BIOFILM ENGINEERING

by

Don Simone Daly

A thesis submitted in partial fulfillment
of the requirements for the degree

of

Doctor of Philosophy

in

Statistics

MONTANA STATE UNIVERSITY
Bozeman, Montana

July 1997

APPROVAL

of a thesis submitted by

Don Simone Daly

This thesis has been read by each member of the thesis committee and has been found to be satisfactory regarding content, English usage, format, citations, bibliographic style, and consistency, and is ready for submission to the College of Graduate Studies.

July 23, 1997
Date

Martin Q. Hamilton
Martin Hamilton
Chairperson, Graduate Committee

Approved for the Major Department

7/25/97
Date

John Lund
John Lund
Head, Mathematical Sciences

Approved for the College of Graduate Studies

7/31/97
Date

Robert Brown
Robert Brown
Graduate Dean

STATEMENT OF PERMISSION TO USE

In presenting this thesis in partial fulfillment for a doctoral degree at Montana State University, I agree that the Library shall make it available to borrowers under rules of the Library. I further agree that copying of this thesis is allowable only for scholarly purposes, consistent with "fair use" as prescribed in the U. S. Copyright Law. Requests for extensive copying or reproduction of this thesis should be referred to University Microfilms International, 300 North Zeeb Road, Ann Arbor, Michigan 48106, to whom I have granted "the exclusive right to reproduce and distribute my dissertation for sale in and from microform or electronic format, along with the right to reproduce and distribute my abstract in any format in whole or in part."

Signature Don Simone Daly

Date July 21 1997

ACKNOWLEDGEMENTS

In completing my dissertation, I have reached a great milestone in my lifetime of learning. Reaching this point has been great fun because of the folks who have accompanied me along the way.

Marty Hamilton, my committee chair, and the other committee members have made learning a true joy. Robert "snatch the theorem from my palm" Boik made mathematical statistics one great adventure. Gary Bogar taught me the value of tenacity and rigor while John Lund taught me more napkin math than I thought possible, and much about the love of mathematics. Steve Cherry was the best classmate; his company alone made the trip worthwhile.

I wish to thank my parents, Audrey and John, my brother John, and my sisters, Jacquie, Judy, Susan, and Celinda. God knows the role they've played and this recognition, though minor, is the best I can offer. I also acknowledge Pat, my wife, whose prodding got me to the end. I applaud Tim Cashin, and Mike and Kathy Doughty for their encouragement over the years; I could not stand to hear "aren't you finished yet" one more time.

My work was supported in part by the Center for Biofilm Engineering at Montana State University, a National Science Foundation supported Engineering Research Center (cooperative agreement EEC-8907039), by the CBE's industrial associates, and, in particular, by Recep Acvi, Zibigniew Lewandowski, Jyostna Pendyala, and Frank Roe. Also, I could not have completed my degree without the technical and financial support of Battelle Memorial Institute and the Pacific Northwest National Laboratory; especially, Linda Wyrick and Brent Pulsipher, and Frank Ryan who brought clarity to my writing.

Finally, and, perhaps most importantly, I offer a special thanks to Leon Wagner, a man of quiet wisdom, my mentor and my friend. For all that I have learned during my formal education, I have learned so much more from the experiences I shared with Leon. I will judge my life a success if I can pass on a small part of what Dr. Leon taught me.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABSTRACT

This dissertation explores the use of an optimization criterion based on the Anderson-Darling statistic (AD), a goodness-of-fit measure, to estimate the mean response in a variety of regression settings. This approach is best suited to the regression model where the *distribution of the random component, and the linkage between this component and the mean response are known.* In this situation, the AD model-fitting technique can outperform other techniques which do not use directly the available information about the distribution and linkage. This work, Anderson-Darling Regression (ADR), is an extension of Minimum Distance Estimation (MDE), pioneered by statisticians such as Parr (1981) and Boos (1981).

A terse history of MDE is presented, with an emphasis on its potential role in parametric and nonparametric regression. An ADR approach is described that accommodates many regression models: parametric and nonparametric, normal and non-normal, linear and nonlinear, natural and transformed. The ADR method can be applied easily to nonstandard regession models. ADR's ease of implementation is illustrated with two examples from biofilm engineering, and using conventional statistical software.

The ADR method does have limitations. Specifically, it may be seriously handicapped when the model is mis-identified, or when the estimator is biased. Therefore, a rigorous modeling approach is required that stresses model validation and diagnostics. On the plus side, the well-fit ADR model has residuals fitting the assumed random distribution — a definite benefit when assessing modeling assumptions, estimating standard errors and performing hypothesis tests.

CHAPTER 1

INTRODUCTION

It is often desirable to describe a quantitative response as a function of the available structural or contextual information which shapes that response. This relationship between the response variable, and the explanatory variables may be described nicely by a regression model: for a given value of the explanatory variable $\mathbf{X} = (x_1, \ldots, x_p)$, the observed response Y is modeled as a deterministic function (systematic effect) of $\mathbf{X}$ perturbed by a random fluctuation. Knowledge of the functional relationship between $\mathbf{X}$ and the expected value of Y is usually of the greatest interest. This research explores one method of learning more about this functional relationship under a diverse set of circumstances.

Three Regression Models

With n data points $\{(\mathbf{X}_i, Y_i)\}_{i=1}^n$, the relationship between $\mathbf{X}$ and Y, the explanatory and response variables, can often be modeled using regression. A regression model features three elements:

- a *systematic effect*: a mean response function $m(\mathbf{X}) = \mathrm{E}[Y|\mathbf{X}]$ defining the expected value of the response Y in terms of $\mathbf{X}$, the explanatory variable.
- a *random effect*: a random perturbation or random error function $\epsilon(\mathbf{X})$ to accommodate stochastic behavior.
- a functional *linkage* that joins the mean response and the random perturbation into an expression describing the observed random variable: $Y = Y(m(\mathbf{X}), \epsilon(\mathbf{X}))$.

The term "linkage" in this context should not be confused with term "link" as used in the Generalized Linear Models context (McCullagh and Nelder, 1989).

A diverse set of models can be constructed by varying these three elements. Three regression models (Table 1) featuring these elements will be important to this work. Each model is associated with a certain regression method and is named accordingly. Two of the three models, LSR and NPR, represent two ends of a spectrum of regression models. These models will be used in comparisons to the third model, ADR, which lies in between the LSR and NPR with respect to the restrictiveness of the modeling assumptions.

All three models include a realized but unobserved contribution from the random element. Common to these models is the concept of a *natural residual*, a function of the observed response and the true (or estimated) response. Let $r(\mathbf{X})$ denote this natural residual and let us define $\epsilon(\mathbf{X})$ to be equivalent to $r(\mathbf{X})$ when $m(\mathbf{X})$ is known: $\epsilon(\mathbf{X}) \equiv r(Y(\mathbf{X}), m(\mathbf{X}))$. If $Y = Y(m(\mathbf{X}), \epsilon(\mathbf{X}))$, then the natural residual is

$$\begin{aligned} r(\mathbf{X}) &= r(Y(\mathbf{X}), \widehat{Y}(\mathbf{X})) \\ &= r(Y(\mathbf{X}), \widehat{m}(\mathbf{X})) \\ &= \epsilon(\mathbf{X}) \quad \text{when } m(\mathbf{X}) \text{ is known .} \end{aligned}$$

Table 1 lists the elements of these three regression models, ordered in terms of the assumptions underlying each method (top to bottom, from more to less restrictive).

Table 1: Three Regression Models

Model	$m(\mathbf{X})$	$\epsilon(\mathbf{X})$	Y
LSR	$m \in \mathrm{R}(\mathbf{X})$ $\mathbf{X} \in \mathrm{R}^{n \times p}$	$\epsilon \sim N(0, \sigma^2 \mathbf{I})$	$\mathbf{X}\beta + \epsilon$
ADR	$m(\mathbf{X}) \in \mathit{Smooth}$ $\mathrm{R}^{n \times p} \subset \mathit{Smooth}$ $\mathbf{X} \in \Re^p$	$\epsilon \sim F_\epsilon(\mu, \sigma^2)(\mathbf{X})$ F_ϵ known	$m(\mathbf{X}) \circ \epsilon$ $\circ \in (\times, +)$
NPR	$m(X) \in \mathit{Smooth}$ $X \in \Re$	$\mathrm{E}[\epsilon^2] < \infty$ F_ϵ unknown	$m(X) + \epsilon$

The first model listed in Table 1 is the classic Least Squares Regression model (LSR), the model most often adopted (Draper and Smith, 1981). The LSR mean response function $\mathbf{X}\beta$ is an

identified function, and linear in its unknown coefficients. The random component follows a Normal distribution, a member of a location-scale family, as does the conditional distribution of Y, $F_{Y|\mathbf{x}}$ (Lehmann, 1991). Though LSR may proceed without the Normality assumption, this assumption simplifies making probabilistic inferences concerning model estimates, predictions and hypothesis tests.

The Anderson-Darling Regression model (ADR) is more flexible than the LSR model. The mean response can be linear as in the LSR model, completely identified but nonlinear, or simply a smooth (p-differentiable) function. Other formulations of the mean response (and the linkage) are also possible. The random term in the ADR model can follow one of a variety of distributions upon which straightforward probabilistic inferences can be made concerning model estimates, predictions and hypothesis tests. Nonetheless, the distribution of the ADR random term, and the conditional distribution of Y, must belong to a known location, scale, or location-scale family.

T A diverse set of models can be constructed by varying these three elements. Three regression models (Table 1) featuring these elements will be important to this work. Two of the three models, LSR and NPR, were chosen due to their wide applicationprobabilistic inferences can be made concerning model estimates, predictions and hypothesis tests. Nonetheless, the distribution of the ADR random term, and the conditional distribution of Y, must belong to a known location, scale, or location-scale family.

The final model corresponds to Nonparametric Regression (NPR), often considered an exploratory data analysis technique, for which little is assumed about the mean response or the distribution of the error. The mean response is assumed to be a smooth function, while the random term is required to have finite first and second moments. An additive linkage between the deterministic and random terms is also assumed (Härdle, 1990). NPR also requires that F_ϵ be a member of a location, scale or location-scale family.

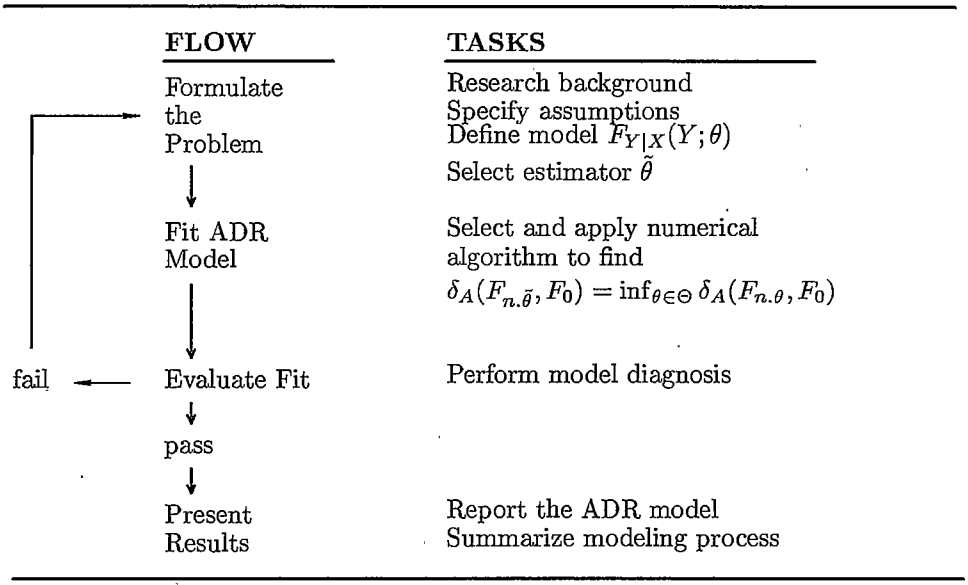
The Modeling Process

Stochastic modeling is an iterative process. The modeler must learn enough about the problem to specify appropriate assumptions and propose a reasonable model. The modeler must

choose an estimator; calculate estimates; and evaluate the results using diagnostic tools. If necessary, The modeler must make adjustments to the model or estimator, and recalculate until a satisfactory solution is found (Chatterjee and Hadi, 1988). Finally, he must apply the model and interpret the results in the context of the model's strengths and weaknesses.

Anderson-Darling Regression (ADR), proposed herein, is an estimation method amenable to this process as illustrated in Figure 1. This dissertation develops ADR, the ADR modeling process, and illustrates the use of ADR with a variety of models.

Figure 1: The ADR Regression Process



Both the LSR and NPR models have solution techniques that produce parameter estimates without requiring knowledge of the distribution of F_ϵ or, for that matter, of $F_{Y|X}$. Distributional concerns are usually handled "after the fact", using regression diagnostics and goodness-of-fit statistics. Fitting the assumed distribution is more of an issue in LSR, where calculating standard errors and confidence intervals is of greater interest.

The ADR method, however, requires that the distribution of $F_{Y|X}$ be faced head on. Measuring how well the observed residuals fit $F_{\hat{\epsilon}|X}$ is the heart of the ADR method. Parameter estimates that produce the best AD goodness-of-fit are the ADR estimates. With ADR, there is no avoiding the complete specification of the generator F_0 of the location, scale or location-scale family which includes the random component's distribution as a member.

In this dissertation, the LSR and NPR models, and estimators provide useful comparisons to ADR. It is assumed that the reader is familiar with LSR and NPR methods. Throughout the remaining document, 'LSR', 'NPR' and 'ADR' will refer to either the model (Table 1) or the related estimation procedure. The reference will be clear from the context.

The ADR method described herein is but one chapter in an extensive literature about Minimum Distance Estimation (MDE). For an introduction to MDE and the MDE literature, the reader is referred to a bibliography by Parr (1981).

Goal and Objectives

The goal of this dissertation is to present an alternative regression technique that is useful when extensive information is available about the distribution of the stochastic component and the linkage. My first objective is to show that the ADR estimation technique is a useful alternative to least squares linear regression or nonparametric regression for an often-encountered set of circumstances. My second objective is to illustrate the use of ADR modeling in a variety of regression settings. To that end, I derive some ADR diagnostics, and show how these may be applied.

A Motivating Example

A company wishes to evaluate the reliability of a vehicle guidance system (Hahn and Shapiro, 1967). Guidance system reliability is characterized by a system's expected time-to-failure and the variability in the time-to-failure from system to system. Characterization reduces to estimating the expected time-to-failure, and the time-to-failure distribution. Posing this problem in a regression setting (i.e., the mean response is constant) provides for an insightful introduction to estimation based on AD goodness-of-fit and to the ADR model-fitting process.

Problem Formulation.

Time-to-failure was observed for 20 guidance systems (Table 2). The sample distribution (Figure 2) suggests that time-to-failure has an asymmetric distribution. The symmetry of the sample distribution improves with a natural log transformation, however. Experience suggests that guidance system time-to-failure is a logNormal(μ,σ^2) random variable (Boos, 1982). Let $m(x) = \mu$, $\epsilon \sim N(0, \sigma^2)$ and $\ln(Y) = \mu + \epsilon$. Either the LSR or ADR model (Table 1) is appropriate for this problem.

Table 2: Time-to-Failure (hours) of 20 Vehicle Guidance Systems

1	4	5	6	15	20	40	40	60	93
95	106	125	151	200	268	459	827	840	1089

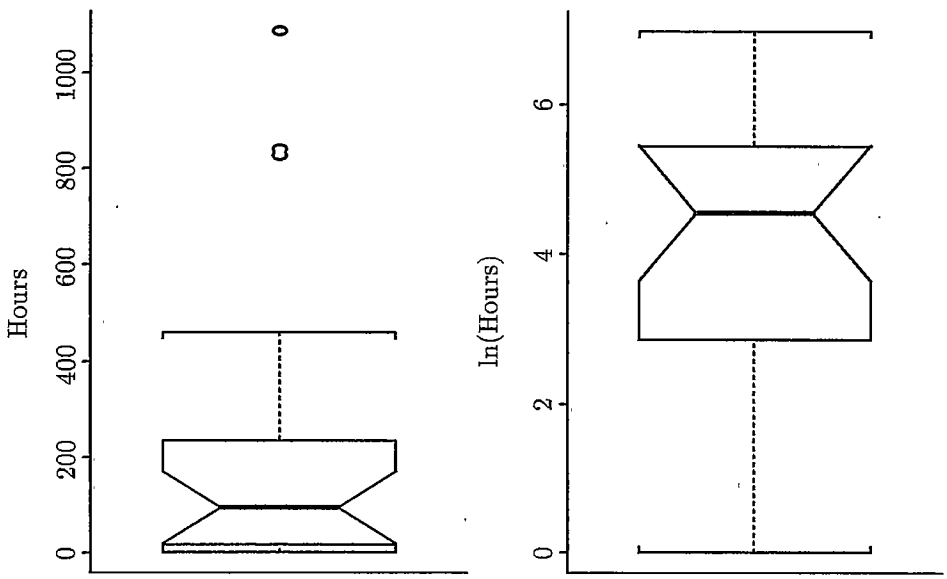


Figure 2: Time-to-Failure Sample Distribution for 20 Vehicle Guidance Systems. Left panel shows the boxplot of measured times-to-failure. The right panel shows the distribution of the log-transformed times.

ADR Modeling.

Though a logNormal model has been adopted, there is some doubt that this distribution applies to all observations. Therefore, the chosen estimator of (μ,σ^2) should be robust. Boos (1982) demonstrates that Anderson-Darling Estimation (ADE), the simplest form of ADR, is a robust estimation technique well suited to location-scale models such as this.

Anderson-Darling estimation (ADE) is one member of the family of Minimum Distance estimation (MDE) techniques. A minimum distance estimate $(\tilde{\mu}, \tilde{\sigma}^2)$ of (μ,σ^2) is the minimizer of the distance δ_{MD} between the empirical distribution F_n, and the cumulative distribution F_θ:

$$\delta_{MD}(F_n, F_{\tilde{\theta}}) = \inf_{\theta \in \Omega} \delta_{MD}(F_n, F_\theta) \; .$$

The Anderson-Darling estimate $(\tilde{\mu}, \tilde{\sigma}^2)$ of (μ,σ^2) is the minimizer of the Anderson-Darling statistic:

$$\delta_A(F_n, F_\theta) \;\; = \;\; \int_{-\infty}^{\infty} \frac{[F_n(y) - F_\theta(y)]^2}{F_\theta(y)[1 - F_\theta(y)]} dF_\theta(y) \qquad (1)$$

The AD statistic is the integral of the weighted and squared difference between the empirical and assumed cumulative distribution functions. The weighting is a function of the cumulative distribution function; the tails of this distribution resulting in significantly larger weights than its center. The AD statistic (1) can be rewritten in a closed form that simplifies computations (see Appendix A for my derivation):

$$\delta_A(F_n, F_\theta) \;\; = \;\; -1 - \sum_{i=1}^{n} \frac{2i-1}{n^2} \ln[F_\theta(y_{(i)})] + \frac{2(n-i)+1}{n^2} \ln[1 - F_\theta(x_{(i)})] \; . \qquad (2)$$

This closed form expression can be minimized using common optimizing routines. For examples in this dissertation, minimization is accomplished using a variant of the Newton-Marquardt algorithm found in *S-PLUS* © (StatSci Division, 1993).

The sample median and the interquartile range are reasonable candidates for initial parameter estimates to seed this iterative algorithm. For the guidance system problem, these initial estimates were $(\mu_0, \sigma_0^2) = (4.54, 3.29)$ (Table 3).

Table 3 summarizes the initial, LSR and ADR fits of the guidance system time-to-failure model. With regard to AD goodness-of-fit criterion, ADR produced the best estimates, though only slightly better than LSR. The ADR residuals are slightly more "logNormal like" than the LSR residuals; and have a slightly better distributional fit. Both fits satisfy the normality assumption so that inferences based on the fit of either method would be acceptable.

The empirical distribution functions for the three sets of estimators are shown against the estimated cumulative distribution functions in Figure 3. Visually, the LSR and ADR empirical

distribution functions (EDFs) appear to fit the hypothesized distribution equally well. Figure 4 provides a closer look at the residuals from each fit in terms of their contribution to the AD statistic. The larger contributions are related to the larger residuals, which is consistent with the definition of the AD statistic (observations in the distribution tails are weighted significantly higher than observations in the center).

Table 3: Initial, LSR and ADR Parameter Estimates for Guidance System Time-to-Failure Model.

Method	Estimator Estimator	Estimates $\hat{\mu}$	$\hat{\sigma}^2$	AD Score	AD GoF p-value
Initial	θ_0	4.54	3.29	0.031	0.048
LSR	$\widehat{\theta}$	4.15	3.75	0.013	0.606
ADR	$\tilde{\theta}$	4.22	4.02	0.012	0.686

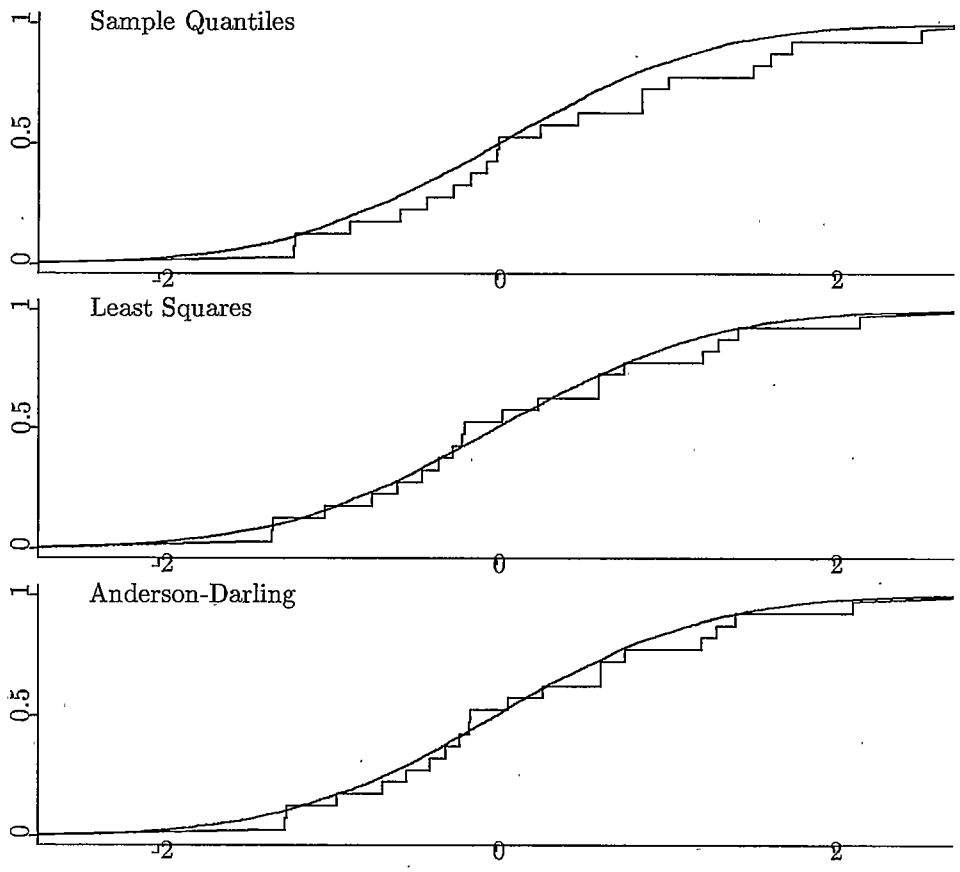


Figure 3: Empirical distributions of normalized residuals from Quantile (median and interquartile range), Least Squares and Anderson-Darling fits of the Time-to-failure logNormal model. The N(0,1) distribution function is marked by the smooth curve.

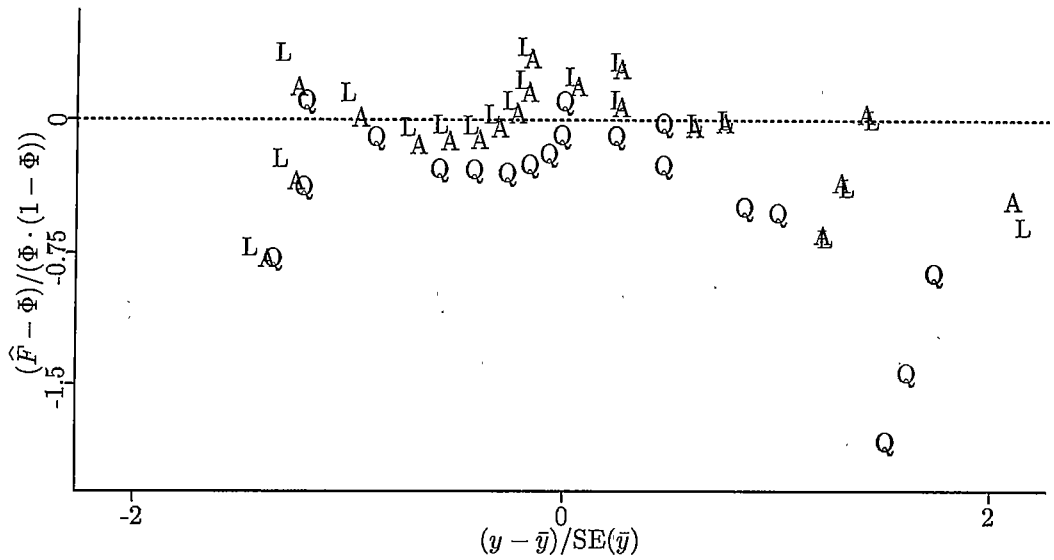


Figure 4: Normalized residuals from the **Q**uantile (median and interquartile range), **L**east Squares and **A**nderson-Darling fits of the Time-to-Failure logNormal model plotted against the contributions of the residuals to the AD statistic (i.e., $(F_n - \Phi)/(\Phi \cdot (1 - \Phi))$).

ADR Fit Evaluation.

The Anderson-Darling statistic $\delta_A(F_n, F_{\tilde{\theta}})$ is a measure of the goodness-of-fit of the empirical distribution F_n to the hypothesized distribution $F_{\tilde{\theta}}$. The distribution of $n\delta_A(F_n, F_{\tilde{\theta}})$ (see Figure 7, Chapter 2) is approximately distribution-free; i.e., minimally influenced by the assumed distribution of the random term (Boos, 1981). Boos conjectures that distribution of $n\delta_A(F_n, F_{\tilde{\theta}})$ is a weighted sum of squared standard normal random variables, adjusted for the number p of estimated parameters:

$$n\delta_A(F_n, F_{\tilde{\theta}}) \approx \sum_{i=p+1}^{\infty} \frac{Z_i^2}{i(i+1)}.$$

For the fitted ADR guidance system model, the approximate goodness-of-fit p-value, $P[\delta_A(F_n, F_\theta) \geq \delta_A(F_n, F_{\tilde{\theta}})]$, is 0.69. This value indicates there is no evidence to suggest that the ADR-fitted LogNormal distribution $F_{\tilde{\theta}}$ is not the source of the time-to-failure observations.

The robustness of the ADR method can be illustrated using repeated perturbations of one observation in the time-to-failure sample. Table 4 shows the effect of a series of perturbations on the ADR and LSR estimates. A smaller change is observed in the ADR parameter estimates compared to the change in the LSR estimates when one observation in the example dataset is perturbed.

Table 4: Initial, LSR and ADR Estimates for Guidance System Perturbed Time-to-Failure Datasets.

		LSR		ADR	
Case	11th Observation	$\widehat{\mu}$	$\widehat{\sigma}^2$	$\tilde{\mu}$	$\tilde{\sigma}^2$
1	.00095	3.58	9.88	3.94	6.57
2	.0095	3.69	7.60	3.95	6.09
3	.095	3.81	5.84	3.97	5.66
4	95	4.15	3.75	4.22	4.02
5	9500	4.38	5.00	4.39	5.05
6	95000	4.50	6.42	4.40	5.49
7	950000	4.61	8.38	4.41	5.91

Though the ADR method is robust, the Anderson-Darling statistic is very sensitive to outliers (Figure 5). With the approximate distribution of $n\delta_A(F_n, F_{\tilde{\theta}})$ and the perturbed datasets, we can evaluate this sensitivity in terms of change in p-value. When the scaled observation is far from the center of the data, the AD distance increases, and the p-value decreases. The AD statistic may be used as a measure of influence.

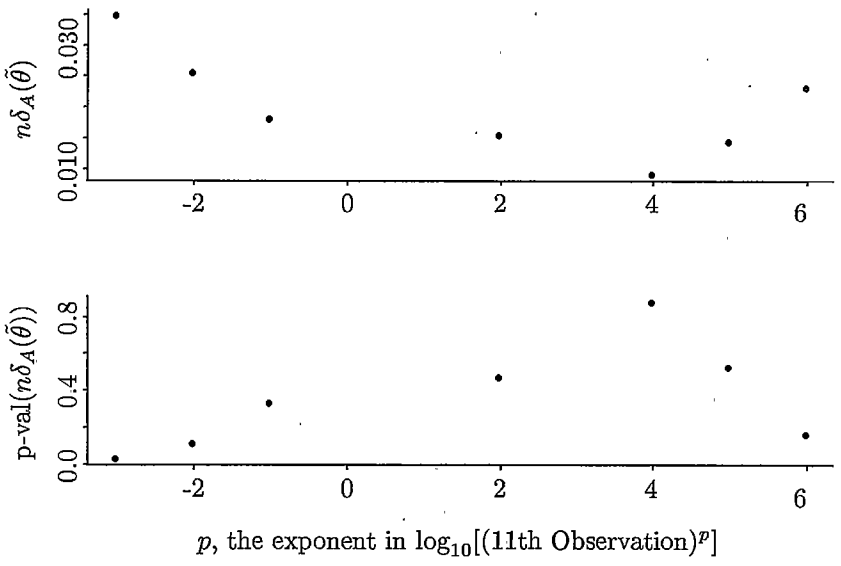


Figure 5: Anderson-Darling Scores with P-values When One Observation is Perturbed

The sensitivity of the Anderson-Darling statistic to sample characteristics is better understood using boxplots of the perturbed samples (Figure 6). Compare the changes in boxplots with the changes in the Anderson-Darling statistic. The AD statistic appears to be sensitive both to outliers and to lack of symmetry. The two boxplots on the left show the effects of both outliers and asymmetry; the AD statistics increase and their p-values decrease. The third and fourth

boxplots show no outliers; they are asymmetric, however. The asymmetry is reflected in larger AD statistics and smaller p-values. The AD statistic of the fifth boxplot has the largest p-value. In terms of Anderson-Darling goodness-of-fit, the residuals summarized in this boxplot are the closest of the seven sets of residuals to a Normal $N(0,1)$ distribution. The sixth and seventh boxplots show, once more, the sensitivity of the AD statistic to outliers and slight asymmetry; the smaller p-values result from the decreases in goodness-of-fit.

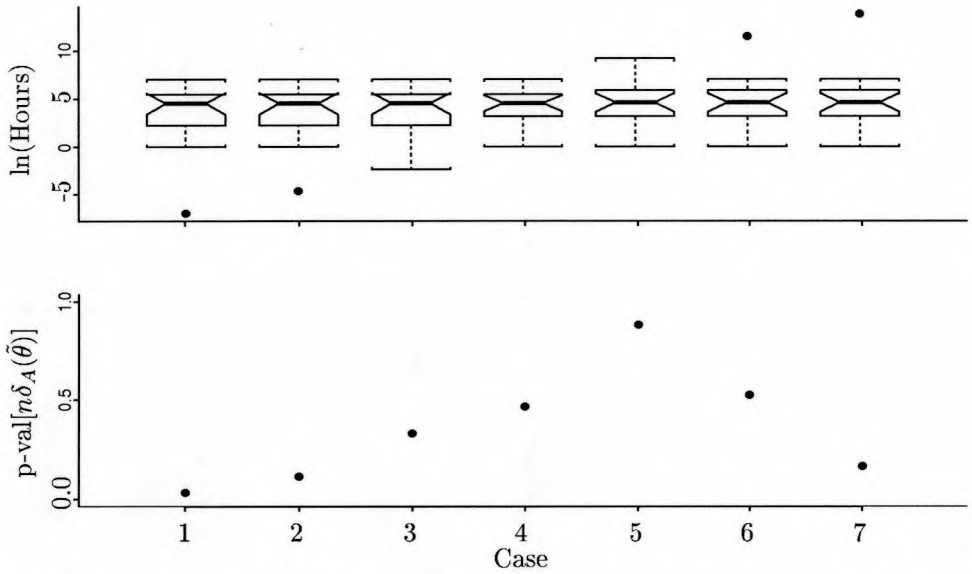


Figure 6: Boxplots of the Perturbed Time-to-Failure Datasets and Anderson-Darling p-values for the cases listed in Table 4.

Organization of Dissertation

Five chapters follow this introduction. Chapter 2, **Anderson-Darling Regression (ADR)**, begins with a detailed overview of Minimum Distance Estimation (MDE), a family of estimation methods based on minimizing a goodness-of-fit comparison between an assumed cumulative distribution function and the empirical distribution function. The bulk of the chapter is devoted to the theory and mathematics of Anderson-Darling Regression (ADR). ADR is an extension of Anderson-Darling Estimation (ADE), one member of the MDE family. The desirable features of ADR are discussed. The case is made that ADR is a good choice when fitting the ADR model and, in particular, when fitting a non-additive, non-normal ADR model.

Non-parametric ADR (NADR) and Empirical Anderson-Darling Estimation (EADE), two variants of ADR which may have the greatest potential as estimation methods, are illustrated Chapter 3, **Two Sensor Applications from Biofilm Engineering**. The first example presents the use of NADR to estimate a Direct Auger spectrum whose features reflect the elemental composition of an examined surface. The basis of this example is a first principles model that describes an Auger spectrum as an ordered set of realizations from an indexed family of Poisson distributions.

The second example illustrates the use of EADE to estimate the mean difference between electromagnetic measurements taken above cleaned and biofilm-covered surfaces. In this case, the physical properties of the sensor, a vibrating micro-probe, are not well defined, so do not suggest a distribution to describe the measurements. Nonetheless, it is possible to sample extensively with the micro-probe so that the cumulative distribution functions of electromagnetic strength measurements above both surfaces can be characterized sufficiently by their respective empirical distributions. The mean difference between these two empirical distributions is estimated using EADE.

A more detailed examination of **Nonparametric ADR** is presented in Chapter 4. The chapter begins with an introduction to nonparametric regression (NPR). Then NADR is presented as a variant of NPR which is useful when the random component of the NPR model is well-defined, but the mean response is not. In NADR, as presented here, the mean response is estimated using a kernel smoother; the optimal NPR smoothing parameter is chosen by minimizing the Anderson-Darling goodness-of-fit statistic.

ADR diagnostics are explored in Chapter 6. To start the chapter, classical regression diagnostics are reviewed. The adoption or adaption of these diagnostics for ADR is then discussed. The chapter closes with an illustration of the methods and a comparison to LSR diagnostics.

The final chapter presents a summary of my findings. Also, future work is discussed briefly.

CHAPTER 2

ANDERSON-DARLING REGRESSION

Anderson-Darling Regression (ADR), a procedure that capitalizes upon the specific information assumed about the distribution of the random component and the linkage in the ADR model (Table 1), is developed in this chapter. Anderson-Darling regression is one of many estimation methods available in the Minimum Distance Estimation (MDE) family. The chapter begins with an overview of MDE, then flows to the specifics of ADR, and explains why ADR is often a good choice when fitting the ADR model.

Minimum Distance Estimation

Minimum Distance Estimation is not one parameter estimation method but a collection of estimation methods. Many methods in the MDE collection can be applied easily to estimate consistently unknown parameters. Minimum distance estimation is designed to reflect the scientific modeler's desire to construct a model reproducing the probablistic structure of the real-world phenomenon under study (Kotz and Johnson, 1958). The theoretical foundation of MDE was presented by Wolfowitz (1957) in a series of papers. Wolfowitz desired to provide consistent parameter estimates in cases where other methods were not successful.

Minimum distance estimation is best explained by considering one of the simplest cases. Let $\mathbf{y} = \{y_i\}_{i=1}^n$ be a simple random sample from the distribution F_θ, a member of a parameterized family of probability distributions, $\mathcal{F} = \{F_\theta : \theta = (\theta_1, \ldots, \theta_k), \theta \in \Theta\}$. Let $\{y_{(i)}\}_{i=1}^n$ be the ordered realizations where $y_{(1)}$ and $y_{(n)}$ are the minimum and maximum, respectively. Let F_n be the empirical distribution function from the sample $\mathbf{y}$:

$$F_n(y) = \frac{1}{n} \sum_{i=1}^{n} I_{[y_{(i)}, \infty)}(y) \; .$$

Suppose δ is a measure of the distance between the empirical distribution function F_n and the functions $F_\theta \in \mathcal{F}$. The minimum distance estimate of θ will be any value $\tilde{\theta} \in \Theta$ such that $\tilde{\theta}$ is a minimizer of the distance between F_n and F_θ:

$$\delta(F_n, F_{\tilde{\theta}}) = \inf_{\theta \in \Theta} \delta(F_n, F_\theta) \, .$$

One must choose from many distance measures and techniques when estimating θ using MDE. MDE methods have been based on the distance between empirical and theoretical cumulative distribution functions, characteristic functions, density functions, and quantile functions, among others. MDE produces estimates by minimizing the difference between empirical and theoretical probabilistic structures; not the difference between observed and predicted values, as in least squares estimation.

Parr (1981) has published an extensive bibliography covering MDE research performed prior to 1980. In his bibliography, references are classified by subject matter including distance measure, MDE philosophy, regression applications, categorical or count data applications, and large sample theory. This literature, and the literature that has followed, has focused upon the theoretical aspects of MDE. The application of MDE, however, has received little attention. Interest in MDE waned in the mid '80's as witnessed by the decline in publications. Excluding a few papers such as Boik (1996), a literature search uncovered no significant articles concerning EDF-based MDE after 1986. The study of probability density functions and Hellinger distance in MDE has flourished, however. These studies are closely related to the study of estimating functions in Generalized Linear Models (McCullagh and Nelder, 1989).

In the classical MDE literature concerned with cumulative distribution functions, the distance measures most thoroughly studied belong to one of two families: Kolmogorov-Smirnov (*supremum*) and Cramér-Von Mises (*quadratic*). Anderson-Darling regression is based on the Anderson-Darling statistic and is a member of the Cramér-Von Mises family.

The Kolmogorov-Smirnov Family. Measures in this family are functions of the maximum difference between two distribution functions. Most *supremum* measures $\delta_{KS}(F_n, F_\theta)$

are defined in terms of the following. Let

$$D^+ = \sup_{-\infty<y<\infty} [F_n(y) - F_\theta(y)]$$

and

$$D^- = \sup_{-\infty<y<\infty} [F_\theta(y) - F_n(y)] .$$

The *Kolmogorov-Smirnov distance* is the maximum absolute distance between the theoretical and empirical cumulative distributions:

$$\delta_K(F_n, F_\theta) = \max(D^+, D^-) .$$

The *Kuiper distance* is the sum of the magnitudes of the largest positive and negative differences between the two cumulative distributions:

$$\delta_{Ku}(F_n, F_\theta) = D^+ + D^- .$$

Cramér-Von Mises (CVM) Family. Distance measures in the Cramér-Von Mises family are the integrated, weighted-and-squared difference of two distribution functions. Members of this family have the following form:

$$\delta_{CVM}(F_n, F_\theta) = \int_{-\infty}^{\infty} [F_n(y) - F_\theta(y)]^2 \psi(y) dF_\theta(y) .$$

The *Cramér-Von Mises distance* features a uniform weight: $\psi(y) = 1$. This distance measure reduces to a sum over the sample $\mathbf{Y}$ (D'Agostino and Stephens, 1986):

$$\begin{aligned} \delta_C(F_n, F_\theta) &= \int_{-\infty}^{\infty} [F_n(y) - F_\theta(y)]^2 dF_\theta(y) \\ &= \frac{1}{12n} + \sum_{i=1}^{n} [\frac{2i-1}{2n} - F_\theta(y_{(i)})]^2 . \end{aligned}$$

With the latter expression, estimating parameters is straightforward using a nonlinear least squares routine. Here, the CDF $F_\theta(\cdot)$ is the nonlinear component of the function $\delta_C(F_n, F_\theta)$ being fit.

If P_y is the proportion of observations whose value is less than or equal to y, then nP_y is a binomial random variable with true proportion $F_\theta(y)$; i.e., $nP_y \sim Bi(n, F_\theta(y))$. The variance of P_y is $F_\theta(y)[1 - F_\theta(y)]$. The *Anderson-Darling distance* uses this variance in the weight $\psi(y)$, thus increasing the influence of extreme observations:

$$\delta_A(F_n, F_\theta) = \int_{-\infty}^{\infty} \frac{[F_n(y) - F_\theta(y)]^2}{F_\theta(y)[1 - F_\theta(y)]} dF_\theta(y) .$$

We can rewrite δ_A as a sum over the sample $\mathbf{Y}$ (see my derivation in Appendix A). Let $F_{\theta.i} = F_\theta(y_{(i)})$, then

$$\begin{aligned} \delta_A(F_n, F_\theta) &= -1 - \sum_{i=1}^{n} \frac{(2i-1)}{n^2} [\ln(F_{\theta.i}) + \ln(1 - F_{\theta.n+1-i})] \\ &= -1 - \sum_{i=1}^{n} \left[\frac{2i-1}{n^2} \ln(F_{\theta.i}) + \frac{2(n-i)+1}{n^2} \ln(1 - F_{\theta.i}) \right] . \end{aligned} \tag{3}$$

As with all CVM estimators, the necessary computations to evaluate the measure δ_A, and to determine the estimate $\tilde{\theta}$ can be accomplished using common numerical optimization routines.

The advantage of using one CVM distance measure over another, or over a Kolmogorov-Smirnov-type distance measure, is not clear. The literature provides little guidance in this area although Boos (1981) suggests that MDE using the Anderson-Darling distance strikes a nice balance between robustness and efficiency.

Anderson-Darling Regression

Suppose that $\mathbf{Y} = \{(\mathbf{X}_i, Y_i)\}_{i=1}^n$ is a random sample from a process under study, wherein Y_i is the *ith* response and $\mathbf{X}_i$ is a vector of values of the independent variables. Suppose that preliminary research indicates that the ADR model (Table 1) is appropriate because

- the systematic effect is a smooth function $m(\mathbf{X})$
- the error distribution F_ϵ is a member of a location-scale family with specified generator F_0
- $m(\mathbf{X})$ and F_ϵ are linked via an additive or multiplicative model: $Y_i = m(\mathbf{X}_i) \circ \epsilon_i$.

Let $\tilde{m}(\cdot)$ be an estimate of the systematic component. Then, a residual, denoted r_i, which is an estimate of the unobserved realization of the random error ϵ_i, can be calculated for each observation. For instance, if the linkage is additive,

$$r_i = Y_i - \tilde{m}(\mathbf{X}_i)$$

and if the linkage is multiplicative, then

$$r_i = \frac{Y_i}{\tilde{m}(\mathbf{X}_i)} .$$

The set of residuals $\mathbf{r} = \{r_i\}_{i=1}^n$ is an estimator of $\{\epsilon_i\}_{i=1}^n$, the unobserved sample of size n from the random error distribution F_ϵ. Without loss of generality, let us assume that $\mathbf{r}$ is an estimator of the unobserved sample from the random error distribution F_0, the generator of the location-scale family to which F_ϵ belongs.

Let $F_{n.\theta}$ be the empirical distribution of the residuals $\mathbf{r}$. The subscript $n.\theta$ emphasizes the dependence, in this case, of the empirical distribution upon both the sample size and the parameters (through the residuals). Suppose that the estimate $\tilde{\theta}$ is the minimizer of the Anderson-Darling distance between the empirical and hypothesized distribution functions:

$$\delta_A(F_{n.\tilde{\theta}}, F_0) = \inf_{\theta \in \Theta} \delta_A(F_{n.\theta}, F_0)$$

Let us call $\tilde{\theta}$ "the Anderson-Darling estimator of θ", and this estimation procedure "Anderson-Darling Regression (ADR)".

A Point of Emphasis. By definition, an empirical distribution is the function of independent and identically distributed random variables. The residuals $\{r_i = r(y_i, m(\mathbf{X}_i)\}$ are the iid random variables in the regression setting, not the response variables $\{y_i\}$ whose distributions depend upon $m(\mathbf{X})$, and therefore, are not identically distributed. ADR must be defined in terms of these residuals. This is possible if F_ϵ and $F_{Y|\mathbf{X}}$ belong to a location/scale family.

Careful reading reveals that the definition of ADR provided here is not identical to the MDE definition provided above. In the earlier description, the empirical distribution F_n remained

fixed, while the parameters of the cumulative distribution F_θ varied as we searched Θ for a θ to minimize δ_A. In the ADR method just outlined, the cumulative distribution F_0, the generator of the specified location-scale family, is held constant. The empirical distribution $F_{n.\theta}$, and the featured values of F_0 vary, however, because the residuals, being functions of θ, vary as Θ is searched for a θ to minimize δ_A.

Desirable Features of MDE and ADR Estimators

Cramér Von Mises estimators and AD estimators, in particular, are often consistent, asymptotically normal, and robust or efficient for a wide variety of distributions. Table 5 lists these and other desirable features common to many MDE estimators including ADR. Almost all of these will be discussed, in order of their appearance in this table.

Table 5: Desirable Features of MDE and ADR Estimators.

Existence
Strong consistency
Fisher consistency
Asymptotic normality
Asymptotic full efficiency
Robustness
MLE-like invariance: $\tilde{g}(\theta) = g(\tilde{\theta})$
Goodness-of-fit measure
Concrete interpretation
Ease of application
Simplicity of computation

Existence of $\tilde{\theta}$. If Θ is compact and F_θ is continuous in θ, then $\delta_A(\theta)$ will also be continuous in θ. It follows that at least one minimizing value $\tilde{\theta}$ exists, because a continuous function on a compact set attains its minimum (Royden, 1988).

Consistency. Under nominal regularity conditions, MDE estimators are strongly consistent (Parr, 1981; Parr and Schucany, 1980; Boos, 1981, 1982). Boos (1982) also identifies conditions which ensure that all CVM $\tilde{\theta}_n$ converge, with probability one, to the true parameter.

Consider the following assumptions. Suppose that

A1: $\mathbf{Y} = \{Y_i\}_{i=1}^n$ is a set of independent, identically-distributed random variables with distribution $F(y) = F_{\theta_0}(y)$ where $\theta_0 \in \Theta, \Theta \subset \Re^k$

A2: $w_\theta(y)$ is a nonnegative weight function which is positive on at least part of the support of F

A3: δ is a Cramér-Von Mises type distance measure:

$$\delta_{F_n}(\theta) = \int_{-\infty}^{\infty} [F_n(y) - F_\theta(y)]^2 w_\theta dy$$

A4: $\tilde{\theta}_n$, when it exists, is the minimizer of $\delta_{F_n}(\theta)$, so that

$$\delta_{F_n}(\tilde{\theta}_n) = \inf_{\theta \in \Theta} \delta_{F_n}(\theta)$$

A5: There are constants C_1 and C_2 such that

$$\sup_{\theta \in \Theta} \int_{-\infty}^{\infty} [F(y)(1 - F(y))]^{2(1-\epsilon)} w_\theta(y) dy \leq C_1$$

$$\sup_{\theta \in \Theta} \int_{-\infty}^{\infty} [F(y)(1 - F(y))]^{1-\epsilon} |F(y) - F_\theta(y)| w_\theta(y) dy \leq C_2$$

for some $\epsilon \in (0, 1)$.

Theorem (Boos). Under assumptions **A1-A5**, if Θ is compact, $\delta_F(\theta)$ is continuous in θ, and $\tilde{\theta}_n$ exists, then $\tilde{\theta}_n \xrightarrow{wp1} \theta_0 \quad \text{as} \quad n \to \infty$.

Boos' Theorem can be restated to apply directly to Anderson-Darling Regression. First, with regard to assumptions **A1-A4:**

A1: $r = r(y|\theta_0)$; and $\mathbf{r} = \{r_i\}_{i=1}^n$ is a set of independent, identically-distributed random variables with distribution $F_0(r) = F_0(r_{\theta_0})$, where F_0 is the generator of a location-scale family which includes F_{θ_0} as a member, where $\theta_0 \in \Theta, \Theta \subset \Re^k$

A2: $w_0(r) = [F_0(r)(1 - F_0(r))]^{-1}$ is a nonnegative weight function which is positive on at least part of the support of F_0

A3: δ_A is a Cramér-Von Mises type distance measure:

$$\delta_A(\theta|F_{n.\theta}) = \int_{-\infty}^{\infty} [F_{n.\theta}(r) - F_0(r)]^2 w_0 dF_0(r)$$

A4: $\tilde{\theta}_n$, when it exists, is the minimizer of $\delta_A(\theta|F_{n.\theta})$ so that

$$\delta_A(\tilde{\theta}_n|F_{n.\tilde{\theta}_n}) = \inf_{\theta \in \Theta} \delta_A(\theta|F_{n.\theta})$$

When posed for ADR, Assumption **A5** becomes

A5: There are constants C_1 and C_2 such that

$$\sup_{\theta \in \Theta} \int_{-\infty}^{\infty} [F_0(r_\theta)(1 - F_0(r_\theta))]^{2(1-\epsilon)} w_0(r) dF_0(r) \leq C_1$$

$$\sup_{\theta \in \Theta} \int_{-\infty}^{\infty} [F_0(r_\theta)(1 - F_0(r_\theta))]^{1-\epsilon} |F_0(r) - F_0(r_\theta)| w_0(r_\theta) dF_0(r) \leq C_2$$

for some $\epsilon \in (0, 1)$.

If these assumptions are met, then $\tilde{\theta}_n \xrightarrow{wp1} \theta_0$ as $n \to \infty$. Checking these assumptions may be a challenge. Boos (1981) provides an example based on the extreme-value, location-scale distribution, where he checks these assumptions, and invokes the theorem. Neither work presented here, nor Boos' work has shown that Assumption **A5** is met by a general class of distributions such as the exponential family. Nonetheless, for many location-scale families, Boos conjectures that the ADR estimators are strongly consistent.

Fisher Consistency. Recall that $\tilde{\theta} \in \Theta$ is the Anderson-Darling estimator of θ if $\tilde{\theta}$ is the minimizer of the Anderson-Darling distance:

$$\delta_A(F_\theta, F_n) = \int_{-\infty}^{\infty} \frac{(F_n - F_\theta)^2}{F_\theta(1-F_\theta)}\, dF_\theta \ .$$

Suppose that $\Theta \subset \Re^1$. Then $\tilde{\theta}(\cdot)$ can be expressed as a statistical functional; $\tilde{\theta}(\cdot)$ maps a distribution F_n (or F_{θ_0}) to an element $\tilde{\theta}$ in the parameter space Θ:

$$\tilde{\theta} = \tilde{\theta}(F_n) \ .$$

The estimator $\tilde{\theta}$ is *Fisher consistent* if $\tilde{\theta}(F_{\theta_0}) = \theta_0$ (Cox and Hinkley, 1974). The Anderson-Darling estimator $\tilde{\theta}(\cdot)$ is Fisher consistent because $\tilde{\theta}(F_{\theta_0}) = \theta_0$:

$$\begin{aligned}
\delta_A(F_{\theta_0}, F_{\theta_0}) &= \int_{-\infty}^{\infty} \frac{(F_{\theta_0} - F_{\theta_0})^2}{F_{\theta_0}(1-F_{\theta_0})}\, dF_{\theta_0} \\
&= 0 \\
&= \inf_{\theta \in \Theta} \int_{-\infty}^{\infty} \frac{(F_{\theta_0} - F_\theta)^2}{F_\theta(1-F_\theta)}\, dF_\theta \\
&= \delta_A(F_{\tilde{\theta}}, F_{\theta_0}) \ .
\end{aligned}$$

Asymptotic Normality. The δ_C estimator for the location parameter μ will be asymptotically normal if F_θ has a uniformly continuous density. Boos (1981), Parr and DeWet (1981), and others, give less stringent conditions for asymptotic normality of other estimators from the Cramér-Von Mises family. Anderson-Darling estimators are usually asymptotically normal. The diverse family of MDE estimators, however, includes many which are not asymptotically normal. For instance, *sup*-norm estimators are often not asymptotically normal (Rao et al., 1975).

Further work by Boos (1982) outlines conditions for which Anderson-Darling estimators are asymptotically normal. He also determines the form of these asymptotic distributions. To report this result, let us first simplify notation.

Define

$$\delta_n(\theta) \quad = \quad \delta_{F_n}(\theta)$$

and

$$\delta_n'(\theta) \quad = \quad (\frac{\partial}{\partial\theta_1}\delta_n(\theta), \cdots, \frac{\partial}{\partial\theta_k}\delta_n(\theta))^T \; .$$

In addition, let $\delta_n''(\theta)$ be the $k \times k$ matrix of second partial derivatives:

$$\delta_n''(\theta) = \left[\frac{\partial^2}{\partial\theta_i\partial\theta_j}\delta_n(\theta)\right]$$

Define C to be the covariance matrix of $\Delta \cdot IC_{\theta_0}(Y)$ with $Y \sim F_{\theta_0}$ and where $IC_{\theta_0}(c)$ is the influence curve of $\tilde{\theta}$,

$$IC_{\theta_0}(c) = -\Delta^{-1}\int_{-\infty}^{\infty}[I(c \le y) - F_{\theta_0}(y)][-F_{\theta_0}^1(y), \cdots, -F_{\theta_0}^k(y)]^T w_{\theta_0}(y)dy \; ,$$

with

$$F_\theta^i(y) = \frac{\partial}{\partial\theta_i}F_\theta$$

and Δ is the $k \times k$ matrix of probability limits of second derivatives,

$$\frac{1}{2}\delta_n''(\tilde{\theta}) \xrightarrow{p} \Delta \; .$$

Theorem (Boos). With the previous assumptions, if $\tilde{\theta}_n$ exists and $\tilde{\theta}_n \xrightarrow{p} \theta_0$ where $\theta_0 \in \Theta^\circ$ (the interior of Θ) and

(a) all derivatives $\delta_n^i(\theta), \delta_n^{ij}(\theta), i, j = 1, \cdots, k$ exist
in some open neighborhood of θ_0;

(b) $n^{\frac{1}{2}}[\frac{1}{2}\delta_n'(\theta_0)] \xrightarrow{d}$ multivariate normal $MVN(0, C)$;

(c) $\frac{1}{2}\delta_n''(\theta^*) \xrightarrow{p} \Delta$, where $(\theta^* - \theta_0)^T(\theta^* - \theta_0) \le (\tilde{\theta}_n - \theta_0)^T(\tilde{\theta}_n - \theta_0)$
and Δ has rank k ;

then

$$\tilde{\theta}_n - \theta_0 \quad = \quad \frac{1}{n}\sum_{i=1}^{n} IC_{\theta_0}(Y_i) + o_p(n^{-\frac{1}{2}})$$

and

$$n^{\frac{1}{2}}(\tilde{\theta}_n - \theta_0) \quad \xrightarrow{d} \quad MVN(0, \Delta^{-1}C\Delta^{-1}) \, .$$

Again, these results will apply directly to the ADR estimators defined previously.

Efficiency. Boos (1982) has also shown that AD estimators are asymptotically fully efficient estimators for location, scale and location-scale models. Others have demonstrated the efficiency of CVM estimators with properly chosen weights $\psi(\theta)$ (Parr and DeWet, 1981; Parr and Schucany, 1980). The choice of CVM weight usually involves a trade-off between efficiency and robustness.

Robustness. Millar (1981) has shown CVM estimators are very robust against local deviations from the model. The influence curve of a CVM estimator is bounded and continuous if the model is correct, and if

$$\int_{-\infty}^{\infty} \left| \frac{\partial F_\theta(y)}{\partial \theta} \right| w_\theta(y) dF_\theta(y) < \infty \, .$$

For a location model, Kotz and Johnson (1958) show that the influence curve is

$$IC_{F_\theta}(c) = \frac{\int_{-\infty}^{\infty} [F_\theta(y) - I(c \leq y)] w_\theta(y) f_\theta^2(y) dy}{\int_{-\infty}^{\infty} w_\theta(y) f_\theta^3(y) dy} \, .$$

The CVM estimator of the location parameter of a normal distribution has an asymptotic variance of $1.095\sigma^2$. A small 9.5% price must be paid to gain a high degree of robustness. Boos (1981) states that Anderson-Darling distance "provides a nice balance between robustness and efficiency in a variety of models because of its weight function $[F_\theta(1 - F_\theta)]^{-1}$". He then provides several examples with particular emphasis on location-scale families.

MLE-Like Invariance. MDE estimators possess a certain MLE-like invariance. Namely,

$$\tilde{\gamma}(\theta) = \gamma(\tilde{\theta})$$

The proof of MLE-like invariance of ADR estimators follows directly from a similar invariance proof for Maximum Likelihood Estimators (Mood et al., 1974).

Theorem (Daly): Let $\tilde{\theta} = (\tilde{\theta}_1, \ldots, \tilde{\theta}_k)$ be the ADR estimator of $\theta = (\theta_1, \ldots, \theta_k)$ in F_θ. Let $\gamma(\theta) = (\gamma_1(\theta), \ldots, \gamma_p(\theta))$ for $1 \leq p \leq k$ be a transformation on the parameter space Θ. If Γ is the range of $\gamma(\cdot)$, then

$$\gamma : \Theta \overset{onto}{\longmapsto} \Gamma$$

and the ADR estimator of $\gamma(\theta)$, $\tilde{\gamma}(\theta)$, is $\gamma(\tilde{\theta})$.

Proof. Note that the mapping $\gamma(\cdot)$ partitions Θ. If $\gamma \in \Gamma$ and $\Theta_\gamma = \{\theta : \gamma(\theta) = \gamma\}$ then

$$\Theta_\gamma \cap \Theta_\lambda = \emptyset \quad \text{for} \quad \gamma \neq \lambda$$

and

$$\Theta = \bigcup_{\gamma \in \Gamma} \Theta_\gamma .$$

Let us define

$$\Delta_\gamma(F_n, F_\theta) = \inf_{\theta \in \Theta_\gamma} \delta(F_n, F_\theta) .$$

We can view Δ_γ as the distance induced by $\gamma(\cdot)$. To find the ADR estimator $\tilde{\gamma}(\theta)$, let us minimize the distance induced by $\gamma(\cdot)$. Hence, the ADR estimate of γ is any value $\tilde{\gamma}$ such that $\Delta_{\tilde{\gamma}}(F_n, F_\theta) \leq \Delta_\gamma(F_n, F_\theta)$ for all $\gamma \in \Gamma$.

We see that $\Delta_{\gamma(\tilde{\theta})} \leq \Delta_\gamma$ for any $\gamma \in \Gamma$ because

$$\begin{aligned}
\Delta_\gamma(F_n, F_\theta) &= \inf_{\theta \in \Theta_\gamma} \delta(F_n, F_\theta) \\
&\geq \inf_{\theta \in \Theta} \delta(F_n, F_\theta) \\
&= \delta(F_n, F_{\tilde{\theta}}) \\
&= \inf_{\theta \in \{\theta : \gamma(\theta) = \gamma(\tilde{\theta})\}} \delta(F_n, F_\theta) \\
&= \Delta(F_n, F_{\tilde{\theta}}) \, .
\end{aligned}$$

Hence,

$$\inf_{\gamma \in \Gamma} \Delta_\gamma(F_n, F_\theta) \geq \Delta(F_n, F_{\tilde{\theta}})$$

and

$$\tilde{\gamma}(\theta) = \gamma(\tilde{\theta}) \, .$$

Goodness-of-Fit. Minimum distance estimation provides not only a method of estimating parameters but also a natural statistic for judging model validity. The minimum Anderson-Darling distance $\delta_A(F_n, F_\theta)$ has been studied for testing goodness-of-fit in the composite null case (Boos, 1981). Boos found distribution-free approximations to the asymptotic null distribution of $n\delta_A$, even though the distributions of these statistics depend upon the family $\mathcal{F}$ and, possibly, upon the specific point $\tilde{\theta}$ in the parameter space Θ. Figure 7 shows the distribution of $n\delta_A(F_n, F_{\tilde{\theta}})$ when two parameters are estimated. The approximate distribution of $n\delta_A$ is a weighted sum of squared standard normal random variables, adjusted for the number p of estimated parameters:

$$n\delta_A(F_n, F_{\tilde{\theta}}) \sim \sum_{i=p+1}^{\infty} \frac{Z_i^2}{i(i+1)}$$

An MDE estimate is the result of an effort to minimize directly the lack of fit of the model to the data. An MDE estimate makes the EDF of the residuals as much like the assumed

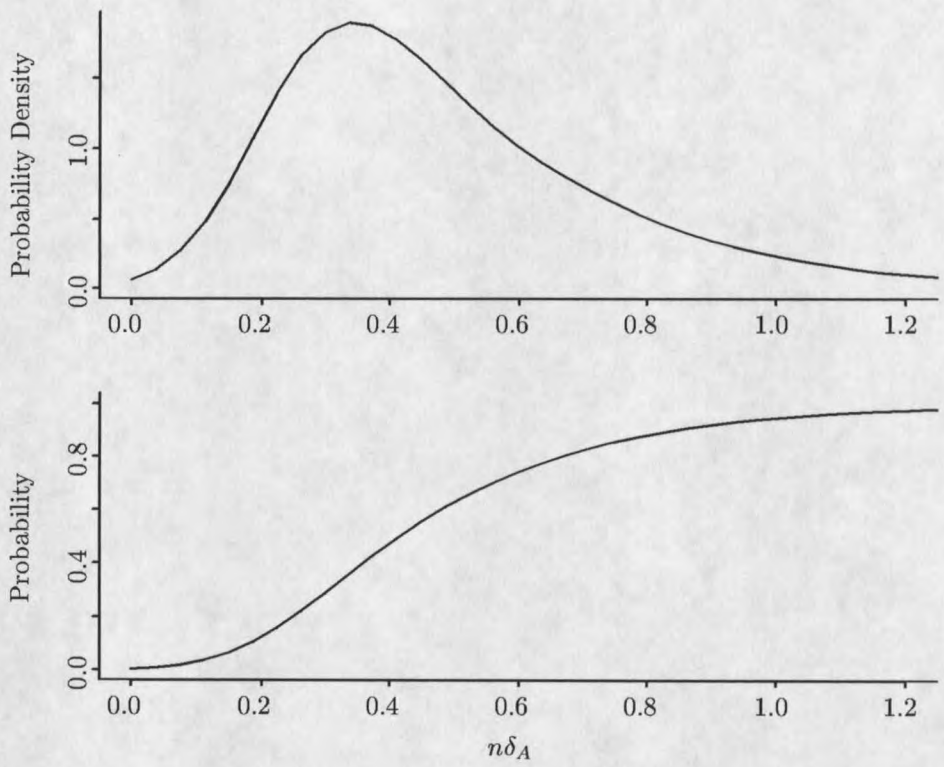


Figure 7: Approximate asymptotic probability density and cumulative distribution of $n \cdot \delta_A$ when two parameters are estimated.

distribution as possible. This is illustrated by the series of boxplots presented in Figure 6. The MDE distance is useful in selecting among competing models, or in testing a null hypothesis where the selection criterion, minimum distance (MD) goodness-of-fit, has a well-understood statistical interpretation (Boos, 1982). As with all omnibus goodness-of-fit tests, MD goodness-of-fit tells you only that the model has passed or failed, and, by itself, does not provide great insight into which model assumption(s) may have been violated.

Concrete Interpretation. Unlike the case for most other estimators, a concrete and useful interpretation of an MDE estimator is possible even when the model is incorrect (i.e., when the parent distribution F is not in $\mathcal{F}$) (Parr and Schucany, 1980). In this case, the MDE estimator $F_{\tilde{\theta}}$ can be interpreted as an approximation to F. For instance, the Kolmogorov-Smirnov estimator is the L^∞ projection of F_n into $\mathcal{F}$. If the CVM weight is chosen as $\psi = 1/f_\theta$, then $F_{\tilde{\theta}}$ is an L^2 projection of F_n into $\mathcal{F}$. Thus, in a probablistic sense, the MDE estimator $\tilde{\theta}$ is a point in Θ giving the best possible approximation $F_{\tilde{\theta}}$ to F.

The concrete interpretation of an ADR estimator as the identifier of a distribution $F_{\tilde{\theta}}$ in $\mathcal{F}$ that most closely approximates the true parent distribution, given the data and measure of distance, follows directly from ADR being an MDE technique. With the ADR weight $\psi = [F_\theta(1 - F_\theta)]^{-1}$, $F_{\tilde{\theta}}$ can be thought of as a weighted L^2 projection of F_n into $\mathcal{F}$. The ADR estimator $\tilde{\theta}$ is a point in Θ giving the best possible approximation $F_{\tilde{\theta}}$ to F in a probablistic and weighted L^2 sense.

Ease of Application. MDE is easy to apply. Given a set of data, the modeler must select an appropriate parametric model, cumulative distribution function and goodness-of-fit measure. Usually, transformations are not required to apply MDE, as is often the case when applying the LSR model and Normal theory. The diversity of the available goodness-of-fit measures allows for flexibility. The ease of application and simplicity of computation are apparent when the closed form of the Anderson-Darling distance is examined:

$$\delta_A(F_n, F_\theta) = -1 - \sum_{i=1}^{n} \left[\frac{2i-1}{n^2} \ln(F_i) + \frac{2(n-i)+1}{n^2} \ln(1 - F_i) \right]$$

Simplicity of Computation. Finally, computation is generally simple. All that is required is an all-encompassing optimization routine to compute the estimate. These routines are readily available in most analysis software such as *SPLUS*©, and *MATLAB*©. The distribution functions and goodness-of-fit measures, particularly for CVM-type estimators, require modest effort to encode.

CHAPTER 3

TWO APPLICATIONS FROM BIOFILM ENGINEERING

In this chapter, Anderson Darling regression is illustrated using two examples. Both examples were small parts of solutions to biofilm engineering problems. In the first example, ADR is used to estimate the direct Auger spectrum; i.e., the expected energy intensity or electron frequency as function of electron velocity. The frequency at any given velocity and, hence, the spectrum depend upon the elemental composition of the examined surface. The relevant particle physics and the electron monitoring system are well-understood. The *distribution underlying the intensity measurements is known* with a fair degree of certainty so that an ADR model (Table 1) provides a fair description.

The second example illustrates the use of ADR to estimate the difference between two electromagnetic field strength distributions; one measured above a biofilm-covered surface, and the other measured above a cleaned surface. The distribution characterizing the uncertainties in electromagnetic strength measurements is not known in this case. First principles provide little guidance. Nonetheless, it is possible to sample the two distributions extensively using automated techniques. The cumulative distribution functions of the electromagnetic strength measurements for both surfaces can be well-characterized by empirical distributions. Differences in field strength above the two surfaces then can be assessed by comparing the two empirical distributions.

Auger Spectroscopy and Relative Elemental Abundance

The relative elemental compositions of microbially colonized surfaces are of great interest to biofilm researchers. Micro-scale relative elemental composition can be evaluated using Auger spectroscopy (pronounced Ō-zhay). Auger spectroscopy is a particle-beam micro-analytic

technique used to identify chemical elements on a surface and to estimate the relative abundances of those elements (Fuchs et al., 1990) . At Montana State University, the surface under inspection by Auger spectroscopy is probed with a $3000eV$ electron beam. In response to the electron bombardment, element-specific Auger electrons are emitted by atoms on or just beneath the inspected surface. Numerous background electrons are generated by other means, such as scattering, across a range of kinetic energies up to $3000eV$. The resulting electron shower is filtered, and electrons in a specified energy band are passed to a detector.

The detected electrons are counted and the count is sometimes recorded. More often, though, the electron intensity, a function of the electron count, is recorded. Electron intensity is defined as

$$I(\epsilon) = \frac{\epsilon \cdot R(\epsilon) \cdot N(\epsilon)}{\Delta t} \tag{4}$$

where ϵ is the signature kinetic energy of the monitored energy band, $R(\epsilon)$ is the resolving power of the Auger system (i.e., the detector bandwidth at energy ϵ), $N(\epsilon)$ is the count of passed electrons detected in the band, and Δt is the length of the probing period.

A common approach in Auger analysis is to successively sample the same location for several short periods. In this case, the average electron intensity

$$\bar{I}(\epsilon) = \frac{1}{n} \sum_{i=1}^{n} I_i(\epsilon)$$

is reported.

The set of electron counts indexed by signature kinetic energy, , $N(\epsilon)$, is known as the electron distribution (Figure 8). The electron intensity $I(\epsilon)$ or average electron intensity $\bar{I}(\epsilon)$ across signature kinetic energies is called the direct Auger spectrum (Figure 9).

A graph of the estimated electron distribution from an Auger scan of a 304 stainless steel sample displays characteristics common to all Auger electron distributions (Figure 8). Electrons with low kinetic energies, say less than $50eV$, are abundant. The number of electrons counted quickly drops for kinetic energies between $50eV$ and $100eV$. The electron distribution varies smoothly and, for the most part, is locally constant for kinetic energies between $100eV$ and

$3000eV$. The smoothness of the electron distribution is broken by sharp increases in counts at the Auger energies of elements present on the surface. These increases appear as distinct dimples on the distribution graph.

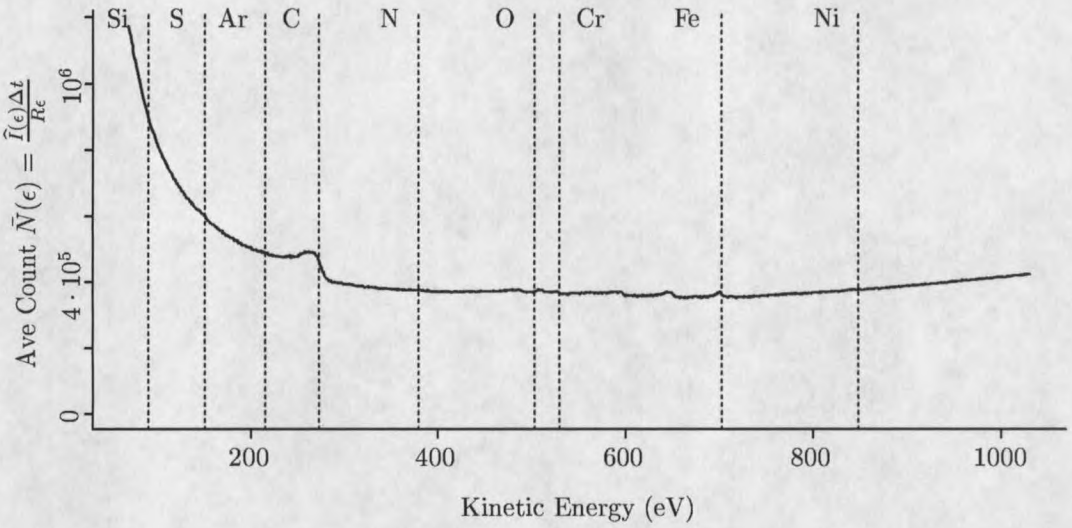


Figure 8: A scaled estimate of an Auger electron distribution from a direct spectrum of a 304 stainless steel sample. The horizontal positions of the vertical dotted lines are at the Auger energies of silicon, sulfur, argon, carbon, nitrogen, oxygen, chromium, iron, and nickel.

The usefulness of Auger spectroscopy resides in the response of surface or near-surface atoms of an element to the exciting effect of the electron beam. An excited atom often emits an electron with a unique or characteristic kinetic energy called its Auger energy. This electron with the characteristic kinetic energy is called the Auger electron. First principles suggest that, over a probing period, Auger electrons are produced for each element in proportion to the abundance of that element on the probed surface (Briggs and Seah, 1983).

Ideally, only Auger electrons would contribute to the electron distribution. This is not the case, however. Auger electrons are a very small proportion, about 10^{-6}, of all the electrons counted. The detected Auger electrons result in a count of electrons greater than the normal background count at the Auger energy and at nearby kinetic energies. The normal background count includes non-Auger electrons such as those generated by beam scattering. The background count varies smoothly over the monitored kinetic engeries (Briggs and Seah, 1983). The element-specific Auger counts create a small positive dimple centered at the element's Auger energy on the relatively smooth graph of the direct Auger spectrum (Figure 9).

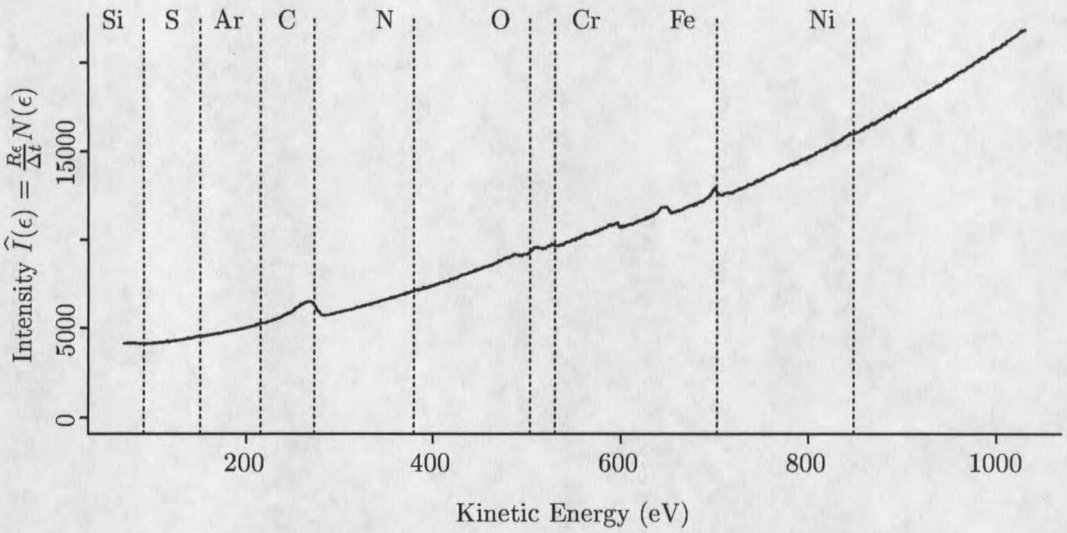


Figure 9: A direct Auger spectrum from a scan of a 304 stainless steel sample. The horizontal positions of the vertical dotted lines are at the Auger energies of silicon, sulfur, argon, carbon, nitrogen, oxygen, chromium, iron, and nickel.

An element often generates secondary Auger-like peaks on the graph, each peak being almost smooth, nearly symmetric and unimodal. These secondary dimples are clearly visible preceding the iron Auger energy (Figure 9). Though, these secondary dimples may embody important information, Auger spectroscopists do not analyze these dimples at this time.

Using the established algorithm of the Auger spectroscopists, the relative abundance of an element is estimated from a numerical derivative of the direct Auger spectrum, or from a measured derivative Auger spectrum provided by the Auger spectrograph (Davis et al., 1976; Burden et al., 1978). On a graph of a derivative spectrum (Figure 10), one tentatively identifies the elements residing on the probed surface by noting significant deviations from zero or from an independent knowledge of potential surface elements. In the example presented here, elements *C, Cr, Fe* and *Ni* are expected because they are the key elements in the composition of 304 stainless steel. The elements *Ar, O* and *N* might also be present because *O* and *N* are common elements in the atmosphere, and *Ar* gas is used in the Auger spectrometer. The spectroscopist also expects that the elements *Si* and *S* also reside on the surface of this speciman.

For each potential surface element, one locates a tell-tale positive spike followed by a negative spike in a neighborhood of the element's Auger energy, and measures the vertical distance

between the tips of the two spikes. This distance is called the "peak-to-peak distance" by Auger spectroscopists.

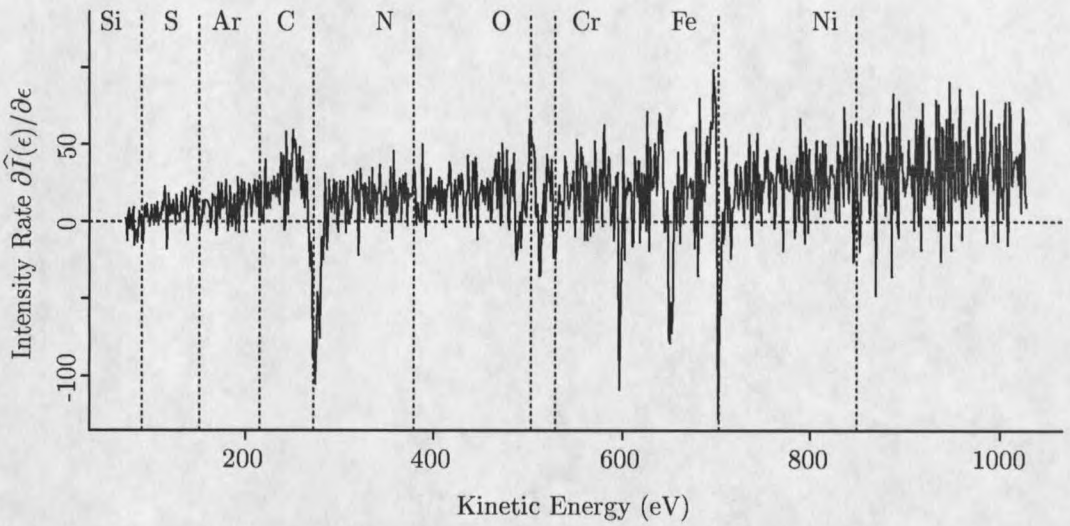


Figure 10: A (numerical) derivative Auger spectrum from a scan of a 304 stainless steel sample. The horizontal positions of the vertical dotted lines are at the Auger energies of silicon, sulfur, argon, carbon, nitrogen, oxygen, chromium, iron, and nickel.

Each peak-to-peak distance D_i is multiplied by a catalogued, element-specific sensitivity factor S_i which reflects the varying detectabilities of elements by Auger spectroscopy. The standard sensitivity factor for each element is determined by an empirical calibration in which the peak-to-peak distance of a pure sample of the element is compared to the distance of a silver standard. Finally, the relative abundance of an element is determined by calculating the ratio R_i of its corrected peak-to-peak distance to the sum of the corrected peak-to-peak distances of all elements of interest:

$$R_i = \frac{S_i D_i}{\sum_j S_j D_j} \,. \tag{5}$$

An Auger Analysis Example. A stochastic model of Auger spectroscopy is developed in Appendix B. The foundation of this model is the particle physics underlying Auger spectroscopy. Appendix B includes an examination of the standard Auger estimators with regards to this stochastic model. The results of this work are incorporated into this example.

As shown in Appendix B, if we assume that the count of electrons at a given kinetic energy $C(\epsilon)$ is a Poisson process

$$C(\epsilon) \sim Poisson(\lambda_A(\epsilon) + \lambda_B(\epsilon))$$

where $\lambda_A(\epsilon)$ and $\lambda_B(\epsilon)$ are the rates of the underlying, independent Auger and background Poisson processes, then the Auger intensity is distributed

$$I(\epsilon) \sim \frac{\epsilon R}{\Delta t} C(\epsilon) \, .$$

With this stochastic description of a direct Auger spectrum, we can proceed to estimate the relative elemental abundances within a sampled region of the example specimen, using a nonparametric version of Anderson-Darling regression (NADR; see Chapter 4).

An Overview of Auger NADR. In this application of NADR, the optimal bandwith for an NPR kernel smoother is estimated using Anderson-Darling distance. The bandwidth resulting in the smallest AD score between the empirical distribution of the normalized Poisson-related residuals and the N(0,1) distribution is selected. This bandwidth provides the best fit, in an Anderson-Darling sense, of a kernel smooth to the measured spectrum.

The N(0,1) distribution is the distribution for comparison in this instance. Recall that if N_i is an independent Poisson random variable with parameter λ, then $N = \sum_{i=1}^{n} N_i$ is also Poisson distributed with parameter $n\lambda$. Further, the normalized count Y where

$$Y = \frac{N - n\lambda}{\sqrt{n\lambda}}$$

is approximately distributed N(0,1) if the expected electron count $n\lambda$ is large (McCullagh and Nelder, 1989). The normalized intensity or normalized direct spectrum at kinetic energy ϵ

$$Y = \sqrt{\frac{n\Delta t}{\epsilon R}} \frac{\bar{I} - \mathrm{E}[\bar{I}]}{\sqrt{\mathrm{E}[\bar{I}]}}$$

is also approximately distributed N(0,1) because

$$
\begin{aligned}
Y &= \frac{X - n\lambda}{\sqrt{n\lambda}} \\
&= \frac{\frac{1}{n}\sum N_i - \lambda}{\frac{1}{n}\sqrt{n\lambda}} \\
&= \frac{\bar{N} - \lambda}{\frac{1}{\sqrt{n}}\sqrt{\lambda}} \\
&= \frac{\sqrt{n}(\bar{N} - \lambda)}{\sqrt{\lambda}} \\
&= \sqrt{n}\frac{\frac{\epsilon R}{\Delta t}\bar{N} - \frac{\epsilon R}{\Delta t}\lambda}{\sqrt{\frac{\epsilon R}{\Delta t}}\sqrt{\frac{\epsilon R}{\Delta t}\lambda}} \\
&= \sqrt{n}\frac{\bar{I} - \mathrm{E}[\bar{I}]}{\sqrt{\frac{\epsilon R}{\Delta t}}\sqrt{\mathrm{E}[\bar{I}]}} .
\end{aligned}
$$

It follows that

$$
Y = \sqrt{\frac{n\Delta t}{\epsilon R}}\frac{\bar{I} - \widehat{I}}{\sqrt{\widehat{I}}}
$$

is approximately distributed N(0,1) because the Anderson-Darling estimator $\widehat{I}$ is a consistent estimator of $\mathrm{E}[\bar{I}]$ (see Chapter 2).

Example Results. Figure 11 shows a graph of the AD scores for a range of bandwidths from kernel smooths of the direct Auger spectrum featured in Figure 9. The optimal AD bandwidth was 1.75 in this application. The features of the graph of AD score versus bandwidth is typical for an NADR fit of an Auger spectrum. The AD scores first drop off rapidly with increasing bandwidth. The scores then reach a minimum and begin to climb slowly. Work presented in the next chapter suggests that bandwidths at, or greater than, the bandwidth associated with the minimum AD score, will result in over-smoothing. Nonetheless, I am not aware of an objective criteria for selecting a bandwidth from the neighborhood bounded above by this minimal bandwidth.

Figure 12 displays the raw and normalized residuals from the Gaussian kernel smooth of the same direct Auger spectrum (Figure 9). Note the increasing spread of the intensity raw residuals with increasing kinetic energy, which parallels the increase in the direct spectrum with increasing

kinetic energy. This is expected, given that the direct spectrum is a realization of a Poisson process. The dashed lines are located at two standard deviations above and below 0. The decrease in spread of the residuals (relative to the spread of the dashed lines) with the increase in kinetic energy indicates that the present model may be inadequate. The smaller-than-expected spread suggests under-smoothing; the estimated mean response follows the observations too closely given the assumed N(0,1) distribution. The Auger system settings (R, Δt, and Number-of-Replicates), which have prominent roles in all estimates, may be incorrect. The spread also suggests that a more sophisticated under-dispersed Poisson model may be required.

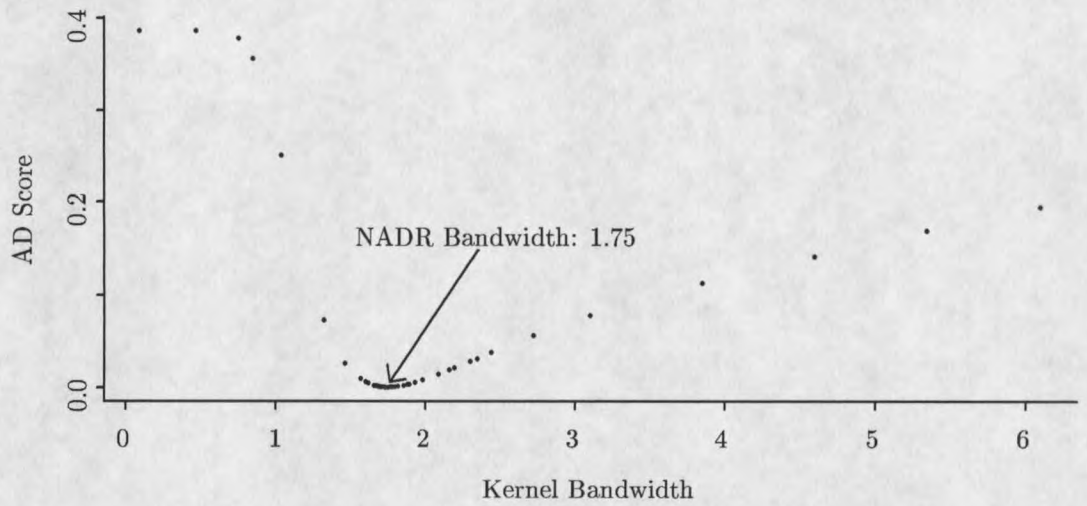


Figure 11: Anderson-Darling score as a function of kernel smoother bandwidth. A kernel bandwidth of 1.75 results in a nearly minimum AD score.

Figure 13 presents a nonparametric estimate of the density of the normalized residuals with a graph of the N(0,1) density. The disconcerting behavior of the varying spread of the normalized residuals notwithstanding, the agreement between the two densities is notable. Anderson-Darling regression works as advertised — it produces residuals in close agreement with the underlying parent distribution!

Figure 14 presents standard (top row) and NADR (bottom row) smooths (left column) of the carbon "dimple" on a direct Auger spectrum of a 304 stainless steel sample. Standard (top row) and NADR (bottom row) derivative estimates form the right column. The smoothness of the NADR-based estimates relative to the standard estimates is the most striking difference between the standard and NADR sets.

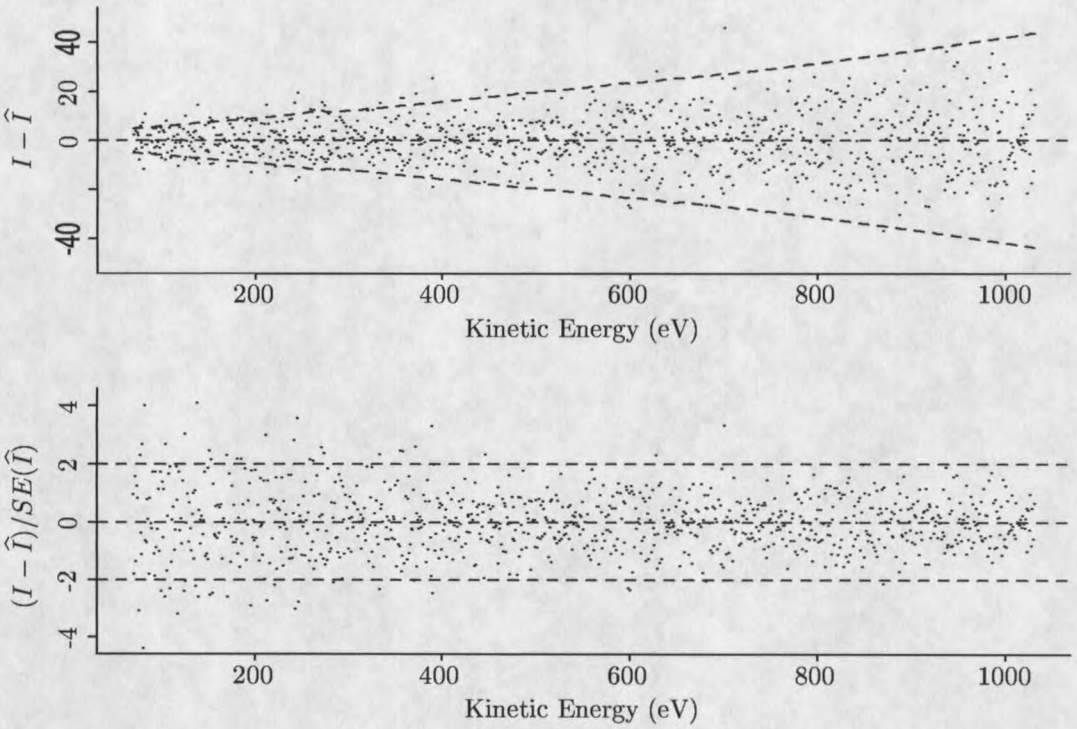


Figure 12: Raw and normalized residuals from the NADR kernel smoother fit to the observed direct Auger Spectrum.

Under-dispersion. Initial fits suggested that an under-dispersed Poisson model may be more appropriate. That is, if the number of electrons N is Poisson distributed with parameter λ, then the appropriate mean and variance for an under-dispersed Poisson model are

$$\begin{aligned} \mathrm{E}[N] &= \lambda \\ \mathrm{Var}[N] &= \tau^2 \lambda \end{aligned}$$

where τ^2 is the dispersion parameter. An under-dispersed Poisson model was fit by adjusting the normalized residuals accordingly:

$$Y = \frac{1}{\tau} \frac{\sqrt{n\Delta t}}{\sqrt{\epsilon R}} \frac{\bar{I} - \hat{I}}{\sqrt{\hat{I}}} .$$

A dispersion parameter value of .85 was selected and used to generate the results presented

(Figure 13). This value resulted in residuals with the lowest AD score with a more homogeneous appearanceover the kinetic energy domain. Heterogeniety was not entirely cured with this model refinement, however, which produced a spread of residuals similar to the original model (Figure 12). This result suggests that a more careful investigation of the physics of Auger electron production and detection is warranted.

Relative Abundance Estimates. As shown in appendix B, the stochastic model of the Auger spectrum coupled with the standard estimation method leads to an alternative estimator of peak-to-peak distance:

$$\widehat{D}(\epsilon_A) = 2\frac{\epsilon_A R}{\Delta t}\widehat{\lambda}'_A(\epsilon_A - \widehat{K})$$

where the Auger derivative estimate $\widehat{\lambda}'_A$ is determined from the direct Auger estimate $\widehat{\lambda}$ by numerical differentiation and $\epsilon_A - \widehat{K}$, the location of the maximum peak preceding ϵ_A, is determined by inspection of the Derivative spectrum (Figure 14).

The standard error of this estimator is easily derived because the numerical derivative of the direct Auger spectrum in this application is a linear combination of independent Auger Intensities:

$$\mathrm{Var}[\widehat{D}(\epsilon_A)] = \mathrm{Var}\left[\widehat{\frac{dI_{\epsilon_U}}{d\epsilon_A}}\right] + \mathrm{Var}\left[\widehat{\frac{dI_{\epsilon_L}}{d\epsilon_A}}\right]$$

where

$$\mathrm{Var}\left[\widehat{\frac{dI_i}{d\epsilon_i}}\right] = \frac{1}{576}\Big[\mathrm{Var}[I_{i-3}] + 36\mathrm{Var}[I_{i-2}] + 441\mathrm{Var}[I_{i-1}] + 441\mathrm{Var}[I_{i+1}] + 36\mathrm{Var}[I_{i+2}] + \mathrm{Var}[I_{i+3}]\Big]$$

and

$$\begin{aligned}\mathrm{Var}[\widehat{I}] &= \mathrm{Var}\left[\frac{1}{n}\sum_{i=1}^{n} I_i(\epsilon)\right] \\ &= \frac{\epsilon^2 R^2}{n(\Delta t)^2}[\lambda_A(\epsilon) + \lambda_B(\epsilon)]\,.\end{aligned}$$

An estimate of the relative abundance of an element is found by combining these NADR-based peak-to-peak estimates using the standard ratio formula.

Table 6 presents approximate 95% confidence intervals derived from NADR estimates of peak-to-peak distances for a set of elements suspected to be residing on the surface of the example 304 steel specimen. The set includes elements commonly found in stainless steel and elements common to the earth's atmosphere. The confidence intervals of nitrogen and nickel include 0. Based on a back-of-the-envelope hypothesis test, there is no strong evidence to support that nitrogen and nickel reside on the surface (There is the question of power, however).

Table 6: NADR point estimates and 95% confidence intervals for peak-to-peak distances of elements expected on the surface of the example 304 steel specimen.

Element	95% Lower Bound	Point	95% Upper Bound
Si	7.42	12.75	18.08
S	9.44	18.72	27.99
Ar	4.54	18.91	33.27
C	129.30	147.90	166.60
N	-2.90	26.42	55.74
O	42.03	87.00	131.90
Cr	11.22	58.12	105.00
Fe	112.69	184.70	256.60
Ni	-43.50	54.53	152.50

Table 7 contains relative abundance estimates obtained by the standard Auger method and by NADR. Under NADR, two sets are listed: one set with nitrogen and nickel included, and one without. The standard set, and the NADR set with with nitrogen and nickel included are in close agreement.

Table 7: Relative elemental abundance estimates of elements expected on the surface of the example 304 steel specimen.

	Relative Elemental Abundance (%)		
	Std. Method	NADR Method	
Element		All	No N nor Ni
C	34.6	32.3	36.7
N	2.8	3.6	NA
O	7.3	7.4	8.5
Cr	7.5	7.9	9.0
Fe	38.9	40.3	45.8
Ni	6.9	8.5	NA

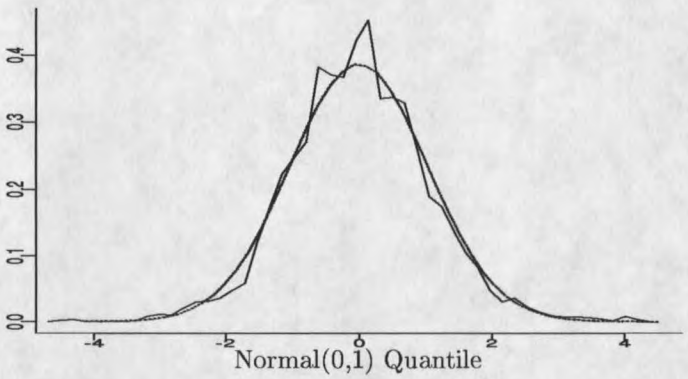


Figure 13: Density estimate from the Auger normalized NADR residuals (jagged line), compared with the N(0,1) density (smooth line).

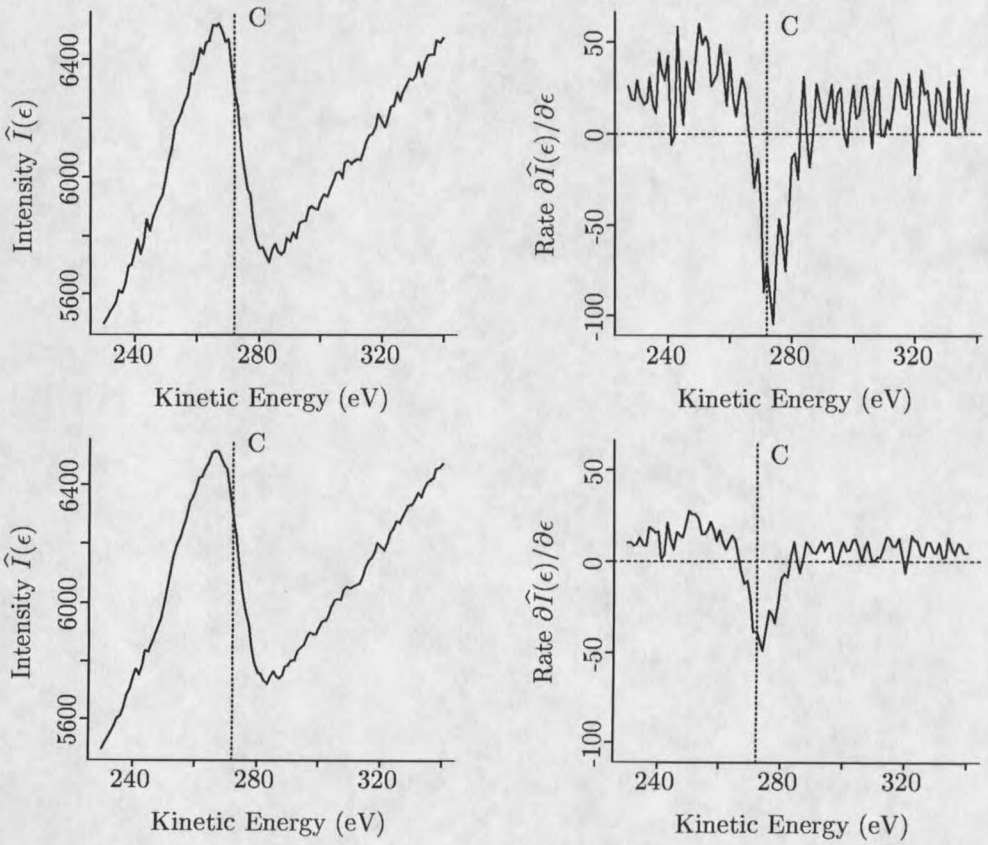


Figure 14: Standard (top) and NADR (bottom) smooths of a direct Auger carbon "Dimple" with estimates of the derivative spectrum for a scan of a 304 stainless steel sample. The horizontal position of the vertical dotted line is at the Auger energy of carbon.

A Vibrating Microprobe and Microbially-Influenced Corrosion.

To evaluate microbially-influenced corrosion, scientists and engineers from MSU's Center for Biofilm Engineering have developed miniature probes to monitor surface chemistry at the micron scale. One such device is the vibrating microprobe. This miniature probe measures electromagnetic field strength, and the flow of electron currents at the micron scale. The recorded response, however, is often a filtered, and coarsely digitized translation of the field strength sensed at the probe tip. Because the response undergoes many unknown transformations before it is recorded, and is often poorly resolved due to digitization, characterizing the probe's uncertainty by extensive sampling under controlled conditions may be more reasonable than attempting an involved derivation of its measurement error distribution.

To illustrate this empirical approach, a uniformly thick biofilm was grown on a polished steel coupon. The biofilm was then removed from one area of the coupon. The electromagnetic fields above the 300-micron-by-300-micron cleaned area and above an adjacent 300-micron-by-300-micron covered area were then sampled repeatedly using a vibrating micro-probe. A total of 400 measurements were collected above each area (Figure 15).

The difference in electromagnetic field strength distributions between the two areas was expected to be a shift in the mean alone. The shapes of the distributions were expected to be the same. Therefore, if Y_c represents a measurement from above the cleaned surface, and Y_b a measurement from above the biofilm-covered surface, then

$$\begin{aligned} Y_c &\sim F_\xi \\ Y_b &\sim \mu + Y_c \end{aligned}$$

where F_ξ is the random distribution for measurements above the cleaned surface and μ represents the difference in means between the two distributions.

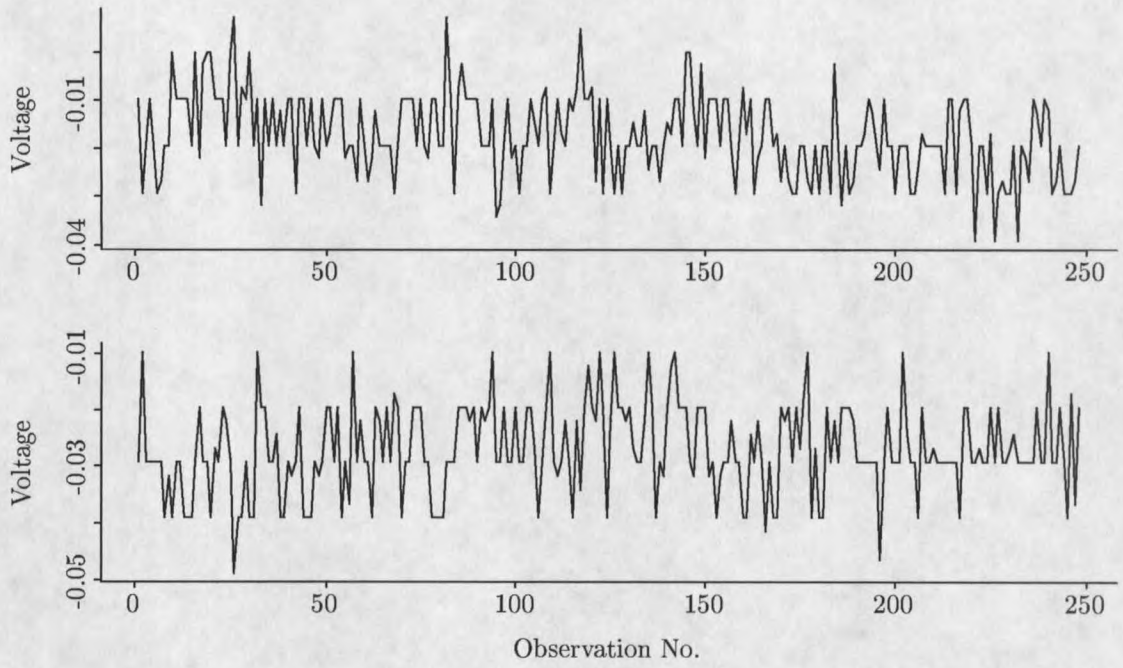


Figure 15: Vibrating probe measurements of electromagnetic field strength above cleaned (top) and biofilm-covered (bottom) surfaces. The measurements are presented in order recorded from left to right. The discrete changes or steps in voltage result from digitization of the measured voltage prior to recording.

We can estimate μ using Anderson-Darling regression. In this application, however, the empirical distribution EDF_ξ of the 400 measurements collected above the cleaned surface fills the role of the true distribution F_ξ in the Anderson-Darling distance measure. This variation to the ADR method does affect the definition of δ_A significantly. The statistic δ_A becomes a weighted difference between two step functions – a new closed form expression for δ_A must be found.

Figure 16 shows the empirical distribution of the cleaned-surface measurements, and the empirical distribution of the biofilm-covered-surface measurements. The description of the difference between the two CDFs as a shift of the mean is appropriate (the striking similarity between the two EDFs raises suspicions ; nonetheless, the two measurement sets are independent).

The ADR estimate of the mean shift is $-1.17 \cdot 10^{-2}$ volts, while the difference in arithmetic means of the two series of measurements is $-1.01 \cdot 10^{-2}$ volts. The relative difference between these two estimates (with respect to the ADR estimate) is 13.7%.

An estimate of the standard error for the difference in means is $1.70 \cdot 10^{-5}$ volts. Treating the Empirical Anderson-Darling Estimate (EADE; see Chapter 6) as the gold standard and

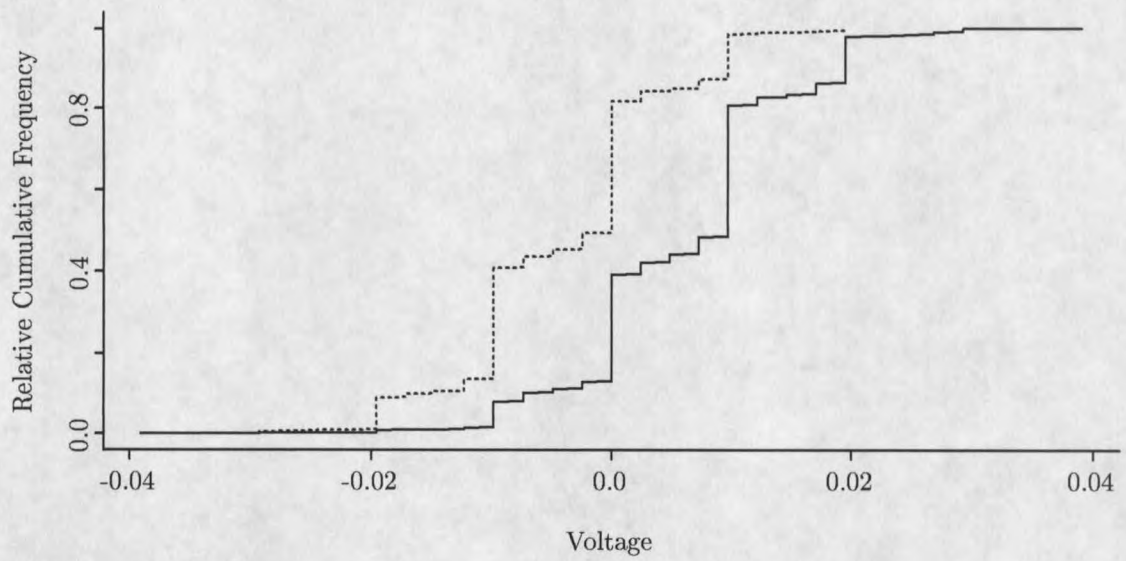


Figure 16: Vibrating probe EDFs for measurements above a cleaned and biofilm-covered surface. The solid line denotes the distribution of the cleaned-surface electromagnetic field strength measurements, while the dotted line denotes the distribution of measurments above the biofilm-covered surface.

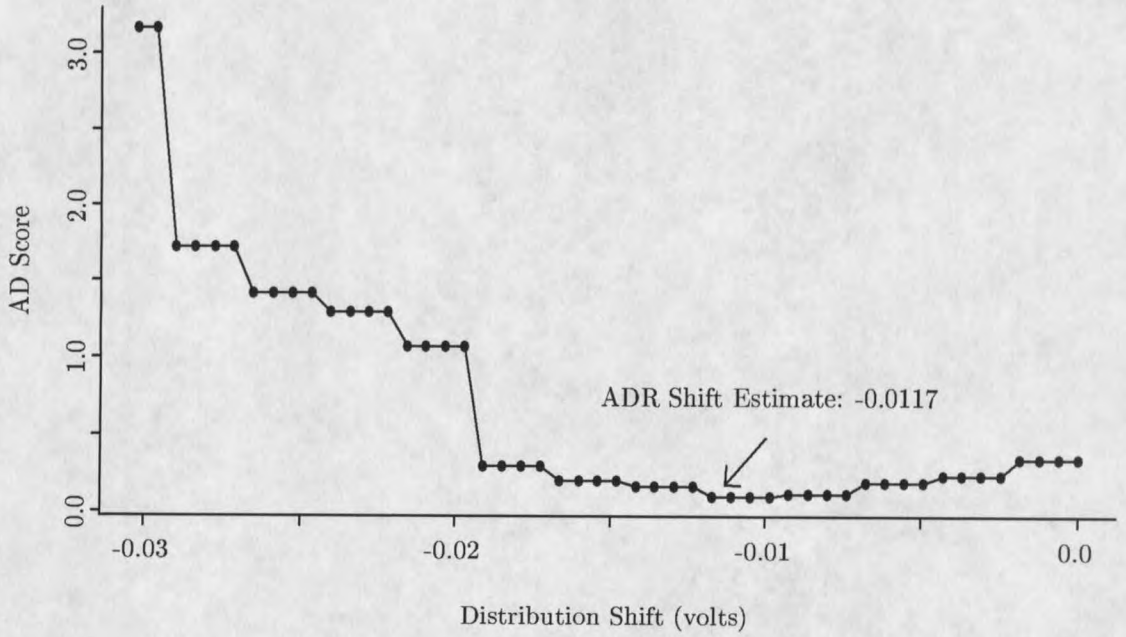


Figure 17: Anderson-Darling score as a function of the shift of the biofilm-covered field strength distribution.

invoking the central limit theorem, an 95% confidence interval for the difference between the ADR value and the classic difference estimate is

$$[\mu_{ADR} - (\bar{X}_{ctrl} - \bar{X}_{trt})] \pm 2 * SE[\bar{X}_{ctrl} - \bar{X}_{trt}] = (1.56 \cdot 10^{-3}, 1.63 \cdot 10^{-3})$$

The difference is statistically significantly different from 0. The significance from a practical standpoint has not been determined.

In this case, applying classical methods may have served us well in estimating the simple expression for the mean shift response function, $m(x) = K$. A more complex expression for the mean response, however, might lead us to a different conclusion. Figure 18 shows the two EDFs adjusted for EADE-estimated mean shift. The two curves are visually almost indistinguishable. This example provides a simple illustration of EADE.

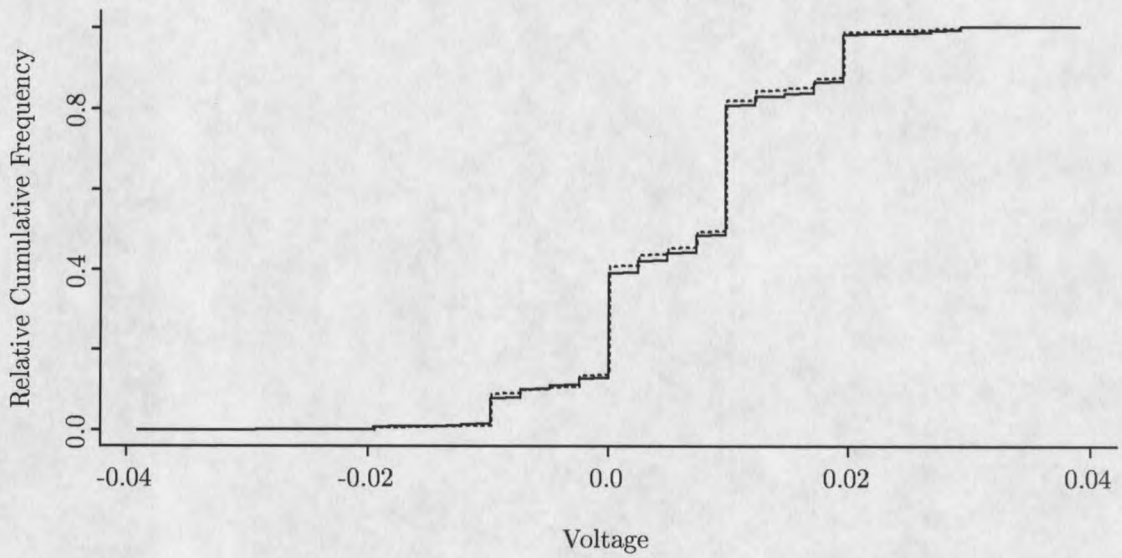


Figure 18: The cleaned field strength measurement EDF and biofilm-covered field strength measurement EDF adjusted for the NADR-estimated mean shift. The solid line denotes the distribution of clean surface electromagnetic field strength measurements while the dotted line denotes the biofilm-covered surface distribution.

CHAPTER 4

NONPARAMETRIC REGRESSION METHODS

Often, too little is known to postulate the Least Squares Regression model (Table 1). For instance, minimal information may be available about the functional form of the mean response. The random component may not be Normally distributed; the linkage may be non-additive. Often, classical regression is too inflexible.

Knowledge of the model may be limited to assumptions about the degree of smoothness of the mean response, to bounds on the first and second moments of the distribution of the random effect, and to the form of the linkage connecting the two. If the linkage between the mean response and the random effect is additive, a nonparametric regression model (Table 1) can be formulated. Nonparametric regression (NPR) techniques to estimate the mean response function of this model are available (Härdle, 1990).

Suppose that the distribution of the random effect is known, as is the structure linking the random effect to an unidentified mean response. Often, this is the case when sampling with routinely-used sensors; the sensing mechanism is described very well by widely accepted physical laws, so the distribution of the random component and the linkage are well understood. The mean response function remains unidentified, however, because an accurate description of the sensed phenomenon in terms of physical laws is not available. Classical regression cannot be applied directly here, and may not be appropriate. NPR is a viable option if the linkage is additive. NPR, however, cannot use directly the available distributional information.

I propose a variation of standard NPR that uses the identified distribution of the random effect directly in the estimation of the mean response. This variation also allows for a non-additive linkage between an unidentified mean response and the random effect. I call this variation "Nonparametric Anderson-Darling Regression (NADR)".

NADR is one method in a family of nonparametric minimum-distance regression methods (NMDR). NMDR uses a minimum distance measure to select an appropriate value for an NPR smoothing parameter. The purpose of this chapter is to illustrate the use of one minimum distance measure, the Anderson-Darling distance measure, to select this value. My intent is to illuminate the potential of NMDR through NADR examples.

To this end, NADR will be compared to and contrasted with three NPR techniques: cross-validated NPR (CV), an adaptive NPR smoothing procedure (ANPR), and a hybrid combination of CV and NADR denoted by CVAD. The comparisons and contrasts will be based on the performance of these methods on several models featuring a variety of known mean response functions, random distributions, and linkages. The models have been chosen to explore the breadth of applicability of NADR.

Performance will be evaluated using four measures: average squared error (ASE), maximum deviation (SUP), Anderson-Darling error (ADE), and mean squared error (MSE). The first two measure the deviation of the estimated mean from the true mean response. The third measures the deviation of the residuals from the true distribution. The fourth is a classical measure of regression performance and is included here for completeness. Discussions of performance will focus on the sampling distributions of these performance measures generated from 1000 simulations of each model.

The chapter begins with a brief introduction to NPR, and the use of cross-validation to select a nonparametric smoothing parameter. Next, the ADR method is adapted for use in a nonparametric setting. The body of the chapter follows this discussion, and is devoted to comparison of the nonparametric methods.

Nonparametric Regression

Standard nonparametric regression is a method of estimating an unidentified mean response function for an additive model whose random error distribution has bounded first and second moments. The name notwithstanding, nonparametric regression estimators do have parameters. The term "nonparametric" refers to the lack of a prespecified functional form for the mean response. A more flexible, parameterized approximation of the unidentified function is used

instead. In NPR parlance, the approximating function is called a "smooth", and the NPR algorithm a "smoother".

Härdle (1990) provides a nice introduction to nonparametric regression. He notes that smoothing can often be thought of as a local averaging procedure. If

$$Y_i = m(x_i) + \epsilon_i$$

$$\mathrm{E}[\epsilon_i^2] < \infty$$

and $m(\cdot)$ is a smooth function, then

$$\widehat{m}(x) = \sum_{j=1}^{n} w_j(x) Y_j$$

with the set of weights $\{w_j(x)\}_{j=1}^n$ being a function of the chosen smoothing method. Further, if the weights $\{w_j(x)\}_{j=1}^n$ are positive and sum to 1, then $\widehat{m}(x)$ is a least squares estimate at the point x. That is, $\widehat{m}(x)$ is a solution to the minimization problem

$$\min_{\theta} \sum_{j=1}^{n} w_j(x)(Y_j - \theta)^2 = \sum_{j=1}^{n} w_j(x)(Y_j - \widehat{m}(x))^2 \, .$$

Common NPR smoothers are the 'kernel', 'spline' and 'orthogonal series'. Each features one or more parameters that affect the resulting smooth. One parameter affecting a kernel smooth is the 'bandwidth'. An important parameter in spline smoothing is the 'span'. Parameters of an orthogonal series smoother include the number and identity of the orthogonal terms. Even after an NPR technique is chosen, the choice of values for the smoothing parameters may still greatly influence the estimate of the mean response. Härdle (1990) describes methods for selecting the best values for these parameters.

This discussion will focus upon kernel smoothing, wherein the sequence of weights $\{w_j(x)\}_{j=1}^n$ are determined by a scalable, density-like function. This function, commonly referred to as the *kernel* K, is a continuous, bounded, and symmmetric real function. A Gaussian kernel

will be used for examples in this chapter. If

$$f(x,\mu) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left[-\frac{(x-\mu)^2}{\sigma^2}\right]$$

then

$$w_j(x) = \frac{f(x_j,x)}{\sum_{j=1}^{n} f(x_j,x)}$$

It follows that

$$w_j(x) \geq 0 \quad \text{for} \quad j=1,\ldots,n\,,$$

and

$$\sum_{j=1}^{n} w_j(x) = 1\,.$$

Kernel smoothers have a simple structure. Indeed, Härdle (1990) notes that kernel smoothers are local polynomial fits . Härdle (1990) also notes that most k-nearest neighbor smoothers, spline smoothers and orthogonal-series smoothers are asymptotically equivalent to kernel smoothers. Focusing on the kernel smoother, therefore, does not seriously slight the others. Furthermore, kernel smoothers are readily available tools in statistical software such as *SPLUS* (StatSci Division, 1993), or are easy to implement.

The accuracy of a kernel-smoothed estimate of the mean response $m(\cdot)$ is a function of the choice of the kernel and the bandwidth. Härdle argues that this accuracy depends mainly on the choice of the value of the smoothing parameter. He notes that the basic idea behind smoothing parameter selection techniques such as cross-validation, generalized cross-validation, and the plug-in method is to estimate the average squared error, or an equivalent measure. The hope is that the smoothing parameter value minimizing this measure results in an optimal approximation to the unknown mean response function. I will concentrate on selecting the optimal bandwidth of a Gaussian kernel smoother using cross-validation. This approach is intuitive, easy to implement, and should give results similar to the other two methods (generalized cross-validation and plug-in).

Simply stated, cross-validation selects the smoothing parameter value that minimizes the predicted sum of squares. For a given value of the smoothing parameter, say θ, the predicted sum of squares is

$$\text{PRESS}(\theta) = \sum_{i=1}^{n} (y_i - m_{\theta.(i)}(x_i))^2 \ .$$

The subscript notation $\theta.(i)$ underscores the dependency of the smooth $m_\theta(\cdot)$ on the smoothing parameter θ and denotes that the *ith* observation was removed when the smooth was computed.

Cross-validation often proceeds by selecting a dense set of values, say of size m, for the smoothing parameter, and then methodically calculating PRESS(θ) for each value. Modern optimization routines may allow an iterative search for the minimizing value. Using either approach, cross-validation could take a long time. With few exceptions, the smooth must be computed n times for each value of the smoothing parameter considered.

A work-around fix, the Weighted Averaging over Rounded Points (WARPing) technique, has been proposed (Härdle, 1990). The fix is to reduce the number of computations by effectively reducing the number of observations. To WARP, bin the x values, find the bin averages for both x and y, and then choose the smoothing parameter using cross-validation on this smaller set. The ability of this approach to produce an accurate estimate is not well understood. Though he outlined the method and promoted its use, Härdle did not discuss the ramifications of WARPing. Other schemes to minimize the computations exist, or can be easily proposed. Exploring these, however, is beyond the scope of this paper.

If the distribution of the random term is identified, then an analysis of the NPR residuals could be used to adjust the smoothing parameter. It is this thought that leads us to a more formal, direct use of NPR residuals to select the smoothing parameter.

Nonparametric Anderson-Darling Regression

In nonparametric Anderson-Darling regression (NADR), an optimal value for a smoothing parameter is selected using the Anderson-Darling goodness-of-fit statistic. The smoothing parameter value associated with the smallest Anderson-Darling goodness-of-fit statistic determines the NADR estimate of the mean response.

On the surface, a problem can often and easily be formulated so that a nonparametric regression algorithm coupled with a goodness-of-fit measure could provide a seemingly reasonable estimate. Consider the model

$$Y(x) = f(m(x), \epsilon)$$

where the linkage function $f(\cdot)$ is known, the distribution of ϵ has been identified, and $m(\cdot)$ is an unknown, smooth function. Suppose that the model can be rewritten in a 'natural residual' form. I define the **natural residual** to be

$$\epsilon(x) = f^{-1}(m(x), Y) \; .$$

To find the NADR estimate of $m(x_i)$, we proceed in a fashion similar to cross validation. Select a dense set of values, say, of size m, for the smoothing parameter. Then, for each parameter value, find the NPR estimate $\widehat{m}(x)$ and, subsequently, estimate set of unobserved errors. For each x_i:

$$\begin{aligned} \widehat{\epsilon}(x_i) &= r_i \\ &= f^{-1}(\widehat{m}(x_i), Y_i) \; . \end{aligned} \tag{6}$$

Next, compute the Anderson-Darling distance between the empirical distribution function $\mathrm{EDF}(r_1, \ldots, r_n)$ and the true cumulative distribution function. Finally, choose the smoothing parameter that produces the smallest Anderson-Darling statistic to determine the NADR estimate of $m(\cdot)$. In contrast to the $n \cdot m$ smooths possibly required for cross validation, only m smooths are required for NADR.

Taylor Series Approximations. NADR should begin with an assessment of the bias of the natural residual (Equation 6) which often may be an exceedingly difficult task to perform rigorously. Therefore, let us consider approximating the mean and variance of the natural residual using a Taylor series approximation (Mood et al., 1974), though wary that the difference between the true values and the Taylor series approximations may be very large. For functions of random variables in general, and quotients in particular, we have

$$\begin{aligned}
\mathrm{E}[g(X,Y)] &\approx g(\mu_X,\mu_Y) + \frac{1}{2}\mathrm{Var}[X]\frac{\partial^2}{\partial x^2}g(x,y)\Big|_{\mu_X,\mu_Y} + \frac{1}{2}\mathrm{Var}[Y]\frac{\partial^2}{\partial y^2}g(x,y)\Big|_{\mu_X,\mu_Y} \\
&\quad + \mathrm{Cov}[X,Y]\left[\frac{\partial^2}{\partial x \partial y}g(x,y)\Big|_{\mu_X,\mu_Y}\right] \\
\mathrm{Var}[g(X,Y)] &\approx \mathrm{Var}[X]\left[\frac{\partial}{\partial x}g(x,y)\Big|_{\mu_X,\mu_Y}\right]^2 + \mathrm{Var}[Y]\left[\frac{\partial}{\partial y}g(x,y)\Big|_{\mu_X,\mu_Y}\right]^2 \\
&\quad + 2\,\mathrm{Cov}[X,Y]\left[\frac{\partial}{\partial x}g(x,y)\Big|_{\mu_X,\mu_Y} \cdot \frac{\partial}{\partial y}g(x,y)\Big|_{\mu_X,\mu_Y}\right]
\end{aligned}$$

and

$$\begin{aligned}
\mathrm{E}\left[\frac{Y}{X}\right] &\approx \frac{\mu_Y}{\mu_X} - \frac{1}{\mu_X^2}\mathrm{Cov}[Y,X] + \frac{\mu_Y}{\mu_X^3}\mathrm{Var}[X] \\
\mathrm{Var}\left[\frac{Y}{X}\right] &\approx \left(\frac{\mu_Y}{\mu_X}\right)^2\left(\frac{\mathrm{Var}[Y]}{\mu_Y^2} + \frac{\mathrm{Var}[X]}{\mu_X^2} - 2\,\frac{\mathrm{Cov}[Y,X]}{\mu_Y\mu_X}\right) .
\end{aligned}$$

Here, I use "$\approx$" to mean the expected value and variance of the function is "approximately equal to" the expected value and variance of its truncated Taylor series expansion.

Assessing Bias Using the Taylor Series Approximation. To illustrate the use of the Taylor series approximation to assess bias, consider a multiplicative model with a $\chi^2(\nu)$ error term:

$$Y_i = m(x_i)\cdot\epsilon_i\ .$$

In this case, the natural residual function is

$$r_i = \frac{Y_i}{\widehat{m}(x_i)}\ . \tag{7}$$

To find the NADR estimate $\widehat{m}(\cdot)$, I would assume that the natural residuals are distributed as independent $\chi^2(\nu)$ random variables. I would then identify the kernel smoothing parameter that produces a residual empirical distribution most closely matching, in an Anderson-Darling sense, the $\chi^2(\nu)$ cumulative distribution.

The assumption that the natural residuals are distributed as independent $\chi^2(\nu)$ random variables is key to making the NADR estimate. The natural residual is the ratio of a χ^2 random variable and a weighted sum of χ^2 random variables; a ratio whose numerator and denominator are not independent. Therefore, let us look closely at the expectation and variance of the natural residual function. We would like the moments of the natural residual to match the $\chi^2(\nu)$ moments. That is, we would like

$$\mathrm{E}[r_i] = \nu \qquad \text{and} \qquad \mathrm{Var}[r_i] = 2\nu \; . \tag{8}$$

Evaluation of the mean and variance of the natural residual, a ratio, is most difficult. So, let us consider the Taylor series approximations of these two functions. The approximate expected value of the natural residual is

$$\begin{aligned} \mathrm{E}[r_i] &= \mathrm{E}\left[\frac{Y_i}{\widehat{m}(x_i)}\right] \\ &\approx \frac{\mu_{Y_i}}{\mu_{\widehat{m}(x_i)}} - \frac{1}{\mu^2_{\widehat{m}(x_i)}}\mathrm{Cov}[Y_i, \widehat{m}(x_i)] + \frac{\mu_{Y_i}}{\mu^3_{\widehat{m}(x_i)}}\mathrm{Var}[\widehat{m}(x_i)] \; . \end{aligned} \tag{9}$$

Let us evaluate the terms of this approximation separately and then recombine the results to approximate $\mathrm{E}[r_i]$. First, let us consider μ_{Y_i} and $\mu_{\widehat{m}(x_i)}$:

$$\begin{aligned} \mu_{Y_i} &= \mathrm{E}[m(x_i) \cdot \epsilon_i] \\ &= m(x_i) \cdot \nu \end{aligned}$$

$$\begin{aligned} \mu_{\widehat{m}(x_i)} &= \mathrm{E}\left[\sum_j w_j(x_i) Y_j\right] \\ &= \sum_j w_j(x_i) m(x_j) \cdot \nu \\ &= \left[m(x_i) + \mathrm{Bias}[x_i, m(x), K, h]\right] \cdot \nu \; . \end{aligned}$$

The bias $\mathrm{Bias}[x_i, m, K, h]$ varies in sign and magnitude depending upon x_i, the mean response m and its derivatives, the kernel K, and the kernel bandwidth h. Nonetheless, Härdle (1990) shows that, under mild conditions, the smooth $\sum_j w_j(x) m(x_j)$ converges in probability to the true mean response $m(x)$ at points of continuity of $m(\cdot)$. I believe that this bias affects cross-validation NPR

and NADR equally. I will ignore (though, not forget) this bias to simplify this evaluation and, therefore, I set

$$\mu_{\widehat{m}(x_i)} = m(x_i) \cdot \nu \, .$$

Next, let us evaluate the variance and covariance terms of the Taylor series expansion:

$$\begin{aligned}
\mathrm{Var}[Y_i] &= \mathrm{Var}[m(x_i) \cdot \epsilon_i] \\
&= 2\, m^2(x_i) \cdot \nu
\end{aligned}$$

$$\begin{aligned}
\mathrm{Var}[\widehat{m}(x_i)] &= \mathrm{Var}\left[\sum_j w_j(x_i) Y_j\right] \\
&= \sum_j w_j^2(x_i) \mathrm{Var}[m(x_j)\epsilon_j] \\
&= 2 \sum_j w_j^2(x_i) m^2(x_j) \cdot \nu
\end{aligned}$$

$$\begin{aligned}
\mathrm{Cov}[Y_i, \widehat{m}(x_i)] &= \mathrm{E}\left[(Y_i - m(x_i) \cdot \nu)\left(\sum_j w_j(x_i) Y_j - \sum_j w_j(x_i) m(x_j) \cdot \nu\right)\right] \\
&= \mathrm{E}\left[\sum_j w_j(x_i)[(Y_i - m(x_i) \cdot \nu)(Y_j - m(x_j) \cdot \nu)]\right] \\
&= \sum_j w_j(x_i) \mathrm{E}[(Y_i - m(x_i) \cdot \nu)(Y_j - m(x_j) \cdot \nu)] \\
&= \sum_j w_j(x_i) \mathrm{Cov}(Y_i, Y_j) \\
&= w_i(x_i) \mathrm{Var}(Y_i) \\
&= 2\, w_i(x_i) m^2(x_i) \cdot \nu \, .
\end{aligned}$$

Substituting these expressions into Equation 9,

$$\begin{aligned}
\mathrm{E}[r_i] &\approx \frac{m(x_i) \cdot \nu}{m(x_i) \cdot \nu} - \frac{2\, w_i(x_i) m^2(x_i) \cdot \nu}{m^2(x_i) \cdot \nu^2} + \qquad (10) \\
&\quad \frac{m(x_i) \cdot \nu}{m^3(x_i) \cdot \nu^3}\left[2 \sum_j w_j^2(x_i) m^2(x_j) \cdot \nu\right] \\
&= 1 - \frac{2}{\nu} w_i(x_i) + \frac{2 \sum_j w_j^2(x_i) m^2(x_j)}{m^2(x_i) \cdot \nu} \, . \qquad (11)
\end{aligned}$$

We can approximate the variance of r_i in a similar fashion:

$$\begin{aligned}
\mathrm{Var}[r_i] &= \mathrm{Var}\left[\frac{Y_i}{\widehat{m}(x_i)}\right] && (12)\\
&\approx \left[\frac{\mu_{Y_i}}{\mu_{\widehat{m}(x_i)}}\right]^2\left[\frac{\sigma^2_{Y_i}}{\mu^2_{Y_i}} + \frac{\sigma^2_{\widehat{m}(x_i)}}{\mu^2_{\widehat{m}(x_i)}} - \frac{2\mathrm{Cov}[Y_i,\widehat{m}(x_i)]}{\mu_{Y_i}\mu_{\widehat{m}(x_i)}}\right]\\
&= \left[\frac{m(x_i)\cdot\nu}{m(x_i)\cdot\nu}\right]^2\left[\frac{2\,m^2(x_i)\cdot\nu}{m^2(x_i)\cdot\nu^2} + \frac{2\cdot\nu\sum_j w_j^2(x_i)m^2(x_j)}{m^2(x_i)\cdot\nu^2} - \right. && (13)\\
&\quad \left.\frac{4\,w_i(x_i)m^2(x_i)\cdot\nu}{m^2(x_i)\cdot\nu^2}\right]\\
&= \frac{2}{\nu} - \frac{4\,w_i(x_i)}{\nu} + \frac{2\sum_j w_j^2(x_i)m^2(x_j)}{m^2(x_i)\cdot\nu}\,. && (14)
\end{aligned}$$

Let us attempt to bound the mean and variance of r_i. From Equation 11, the expected value of the natural residual is bounded by

$$1 - \frac{2}{\nu}w_i(x_i) \le \mathrm{E}[r_i] \le 1 + \frac{2\sum_j w_j^2(x_i)m^2(x_j)}{m^2(x_i)\cdot\nu}\,.$$

The lower bound can be simplified by noting that $w_i(x_i)$ is less than or equal to one. I will conjecture that $\frac{\sum_j w_j^2(x_i)m^2(x_j)}{m^2(x_i)}$ is nearly equal to one. Nonetheless, if $K_i = \frac{\sum_j w_j^2(x_i)m^2(x_j)}{m^2(x_i)}$ then

$$1 - \frac{2}{\nu} \le \mathrm{E}[r_i] \le 1 + \frac{2}{\nu}\cdot K_i\,.$$

Note that the interval bounding the expectation shrinks in width as the degrees of freedom ν increases; The expected value of the natural residual is very near one when ν is large as is expected given the relationship between the numerator and denominator of the ratio.

In a similar fashion, bounds on the variance of the natural residual can be developed from Equation 14:

$$\frac{2}{\nu} - \frac{4\,w_i(x_i)}{\nu} \le \mathrm{Var}[r_i] \le \frac{2}{\nu} + \frac{2\sum_j w_j^2(x_i)m^2(x_j)}{m^2(x_i)\cdot\nu}\,.$$

and,

$$0 \le \mathrm{Var}[r_i] \le \frac{2}{\nu}[1 + K_i]\,.$$

These derivations show that NADR should not be applied directly to this χ^2 model if the

sample size is small because the natural residuals do not follow the χ^2 distribution. In the exploratory spirit of NPR, however, NADR with a corrected residual may provide adequate results. The expected value of the natural residual (Equation 7) suggests a corrected residual whose moments more closely match χ^2 moments (Equations 8):

$$\begin{aligned} r_i^* &= \nu \cdot r_i \\ &= \nu \cdot \frac{Y_i}{\widehat{m}(x_i)} \end{aligned}$$

with

$$\nu - 2 \leq \mathrm{E}[r_i^*] \leq \nu + 2K_i$$

and

$$0 \leq \mathrm{Var}[r_i^*] \leq 2\nu \cdot (1 + K_i) \, .$$

Comparisons of Nonparametric Methods

In this section, NADR is compared to three nonparametric regression techniques: cross-validated nonparametric regression (CV), CV using Anderson-Darling distance (CVAD) and adaptive bandwidth smoothing (ANPR). NADR and CV have been discussed earlier in this section. An introduction to adaptive bandwidth smoothing is provided by Härdle (1990). The adaptive bandwidth smoothing algorithm used in this paper is available as an *SPLUS* function called *supsmu* and is described in the *SPLUS* documentation (StatSci Division (1993)). The hybrid routine, CVAD, was constructed by replacing the *PRESS* statistic in CV with a "predicted Anderson-Darling distance" wherein the usual set of residuals $\{r_i\}$ are replaced by leave-one-out set of residuals $\{r_{(i)}\}$.

The performance of each method is measured by averaged squared error (ASE), maximum deviation (SUP), Anderson-Darling error (ADE) and mean squared error (MSE):

$$\begin{aligned} ASE(\widehat{m}) &= \frac{1}{n}\sum_{i=1}^{n}[m(x_i) - \widehat{m}(x_i)]^2 \\ SUP(\widehat{m}) &= \max[\{m(x_i) - \widehat{m}(x_i)\}_{i=1}^{n}] \end{aligned}$$

$$ADE(\widehat{m}) = -1 - \sum_{i=1}^{n} \frac{2i-1}{n^2} \ln[F_\theta(r_{(i)})] + \frac{2(n-i)+1}{n^2} \ln[1 - F_\theta(r_{(i)})]$$

$$MSE(\widehat{m}) = \frac{1}{n} \sum_{i=1}^{n} [y_i - \widehat{m}(x_i)]^2 .$$

Performance is measured on a variety of models (Table 8). Each model was simulated 1000 times for each nonparametric regression technique to generate sampling distributions of the performance measures. The discussion of performance focuses on these performance distributions, and graphs of example fits or residuals.

Table 8: Example Models for Regression Performance Comparisons.

	Model	Linkage	Mean Response	Random Distribution
	Normal	Additive	$A[\sin(2\pi x^3)]^3$	N(0,1)
1			$A = 1.7$, $x \in [0,1]$	
2			$A = 3.4$	
3			$A = 10$	
4	Student's t	Additive	$3.4[\sin(2\pi x^3)]^3$ $x \in [0,1]$	Student's t(4)
5	Spline	Additive	$3.4[\sin(2\pi x^3)]^3$ $x \in [0,1]$	Asymmetric
6	χ^2	Multiplicative	x^2 $x \in [10,100]$	$\chi^2(4)$
	Poisson	NA	$B(x-45)^2 + 100$	Poisson(m(x))
7			$B = 4$, $x \in [10,100]$	
8			$B = 8$	
9			$B = 13.3$	

Additive Model Examples

In these examples, the four smoothing methods were applied to simulated data from five models (Table 8). The datasets for the first three models are based on the additive model

$$Y_i = m(x_i) + \epsilon_i$$

with

$$m(x) = A[\sin(2\pi x^3)]^3$$

$$x \in [0,1] \qquad A \in \{1.7, 3.4, 10\} .$$

This function has very smooth to not-so-smooth features, which will allow us to compare the bias/fidelity balance achieved by the competing smoothing parameter selection techniques.

Three random error distributions were chosen for study (models 1-5, Table 8). The first distribution, N(0,1), allows for evaluation of the regression techniques in a familiar setting (models 1-3). The second distribution, Student's t(4), provides a heavy-tailed version of the first (model 4). The third is an asymmetric distribution (model 5). Both the Student's t(4) and the asymmetric distribution were scaled to have mean 0 and variance 1. The three error distributions are displayed in Figure 19.

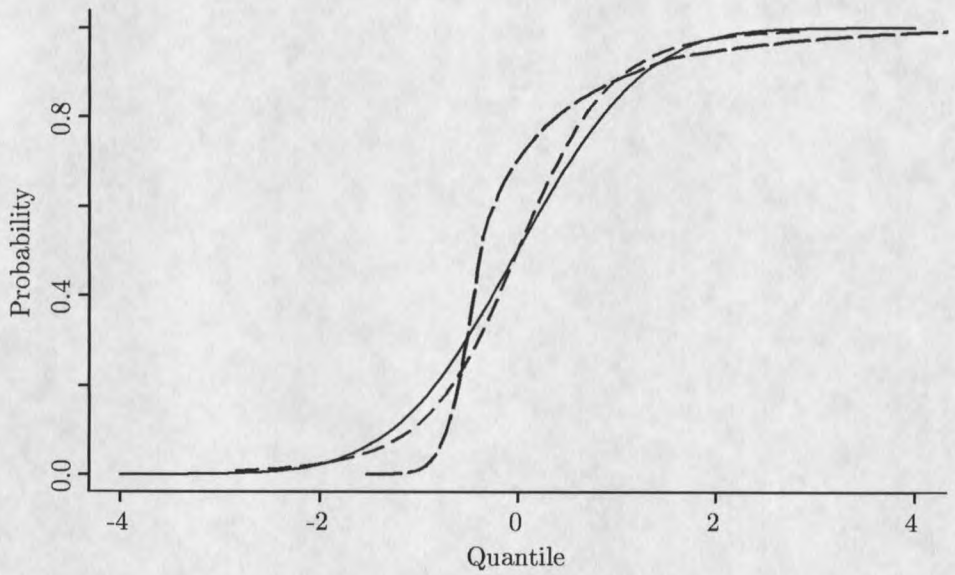


Figure 19: Three error distibutions featured in the additive model examples: N(0,1) — solid line; Student's t(4) — short-dashed line; custom asymmetric distribution — long-dashed line.

To evaluate the influence of the magnitude of the errors on smoother performance, the mean responses of models 1, 2, and 3 were scaled so that the variance of the N(0,1) error distribution was 5%, 15%, and 30% of the range of the mean response, respectively. To evaluate the influence of distribution type, the mean responses of the Student's t and asymmetric models (models 4 and 5, respectively) were scaled so that their variances roughly correspond to 15% of the ranges of their mean responses.

To generate a dataset (one of the 1000 from each model), the mean response function was evaluated at 100 x-values that were evenly spaced over the interval $[0, 1]$. A random sample of size

100 was drawn from the appropriate error distribution, and then added to the mean values to create the y's. Scatterplots of example datasets from the five models, with the true mean responses, are shown in Figure 20.

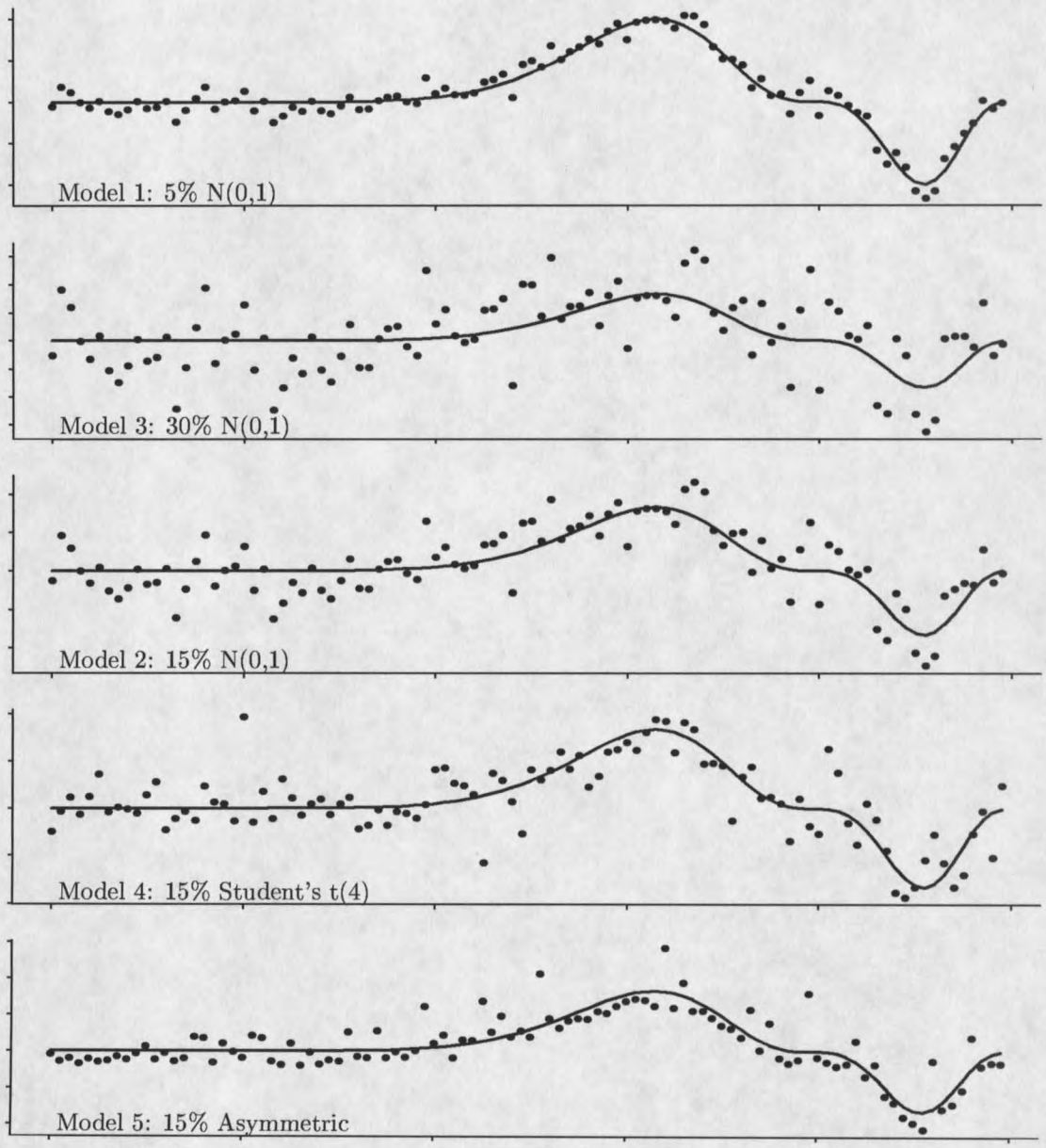


Figure 20: Examples of datasets from models 1 through 5. Each plot shows the scatter of 100 observations from one model about the true mean response.

Mean and variance of the additive-model natural residual. The natural residual of the additive model is

$$r_i = [Y_i - \widehat{m}(x_i)]/\sigma$$

where σ is known (i.e., $\sigma = 1$). The mean and variance of this residual can be approximated and then bounded in a straightforward fashion:

$$\begin{aligned} \mathrm{E}[r_i] &= \mathrm{E}[Y_i - \widehat{m}(x_i)] \\ &= m(x_i) - \sum_j w_j(x_i) m(x_j) \\ &\approx 0 \end{aligned}$$

because of Equation 9, and

$$\begin{aligned} \mathrm{Var}[r_i] &= \mathrm{Var}[Y_i - \widehat{m}(x_i)] \\ &= \mathrm{Var}[Y_i] + \mathrm{Var}[\widehat{m}(x_i)] - 2\,\mathrm{Cov}[Y_i, \widehat{m}(x_i)] \\ &= 1 + \sum_j w_j^2(x_i) - 2w_i(x_i)\,. \end{aligned}$$

Earlier, we saw that $1/n \leq 2\,w_i(x_i) - \sum_j w_j^2(x_i) \leq 1$. Hence, the variance is bounded below by 0 and above by 1:

$$0 \leq \mathrm{Var}[r_i] \leq \frac{n-1}{n}\,.$$

We expect that estimating the mean response function using NADR and the natural residuals will result in over-smoothing. Because the variance of the natural residuals may be less than the model's true variance, NADR will inflate the residuals by over-smoothing to match this variance. Nonetheless, the mean and variance of the natural residual do not provide enough guidance to construct a corrected residual. So, let us proceed with NADR, but be aware that the NADR smooth may be too smooth.

Results and Discussion (Models 1-5) Example NADR, CVAD, CV, and Adaptive NPR smooths based on the medians of the simulated sampling distributions of the smoothing

parameter are shown in Figures 21 and 22. Visually, the example smooths from all methods appear to be equivalent in their estimation of the true mean response for a given error distribution. The NADR and CVAD smooths are nearly identical in all cases. These two appear smoother than either the CV smooth or the ANPR smooth. The CV smooth exhibited less bias than the other methods at the peak and trough of the mean response function. The NADR and CVAD smooths, however, exhibit less bias than the other methods over a region where the true response was constant. The ANPR smooth exhibits as much bias as, or more bias than, the NADR and CVAD smooths in the constant region, and more bias than the CV smooth in the peak and trough.

Smoother performance in terms of the four performance measures described previously is summarized by the boxplots in Figures 23, 24 and 25. Each boxplot summarizes the results for one performance measure applied to a set of 1000 simulations. These sample distributions summarize the performance measures from smooths of datasets wherein $\mathrm{Var}(\epsilon_i)$ was 15% of the range of the true mean response (models 2, 4, and 5). Performance results for the 5% and 30% N(0,1) models (1 and 3) are comparable to the results for the 15% N(0,1) model, and will not be presented here.

For the measures ASE, MSE, and SUP, the relationships among the four performance measure sampling distributions appear to be the same across all the additive error distributions. That is, the performance measure distributions for ANPR have the smallest mean value and narrowest spread. Most often, the performance measure distributions for CV are very similar to ANPR performance measure distributions, while CVAD performance measure distributions are very similar to those of NADR. The mean and spread generally increase from ANPR to CV and CVAD and, finally, to NADR.

For the ADE performance measure, these relationships are reversed. Performance measure distributions for NADR have the smallest mean value and narrowest spread. The mean and spread generally increase from NADR to CVAD and CV and, finally, to ANPR. Nonetheless, the distributions overlap substantially in all cases; no method appears to perform substantially better or worse than the others. Hence, other considerations such as a good fit to the assumed error distribution for construction of confidence intervals may determine the appropriate method to use.

Differences in processing time do exist. As implemented, the ANPR method is 5 to 15 times faster than NADR. In turn, NADR is 10 to 30 times faster than CV or CVAD. Nonetheless, given

the present speed of computing, these processing time differences may be insignificant (outside of a simulation study, that is).

From an objective viewpoint, the performance measures favor the adaptive smoothing approach (ANPR) of the additive model, across a breadth of asymmetric and symmetric error distributions and error scales. Subjectively, the fixed bandwidth methods (NADR, CV and CVAD) provide a more visually appealing smooth. From a practical standpoint, any of the methods would perform adequately.

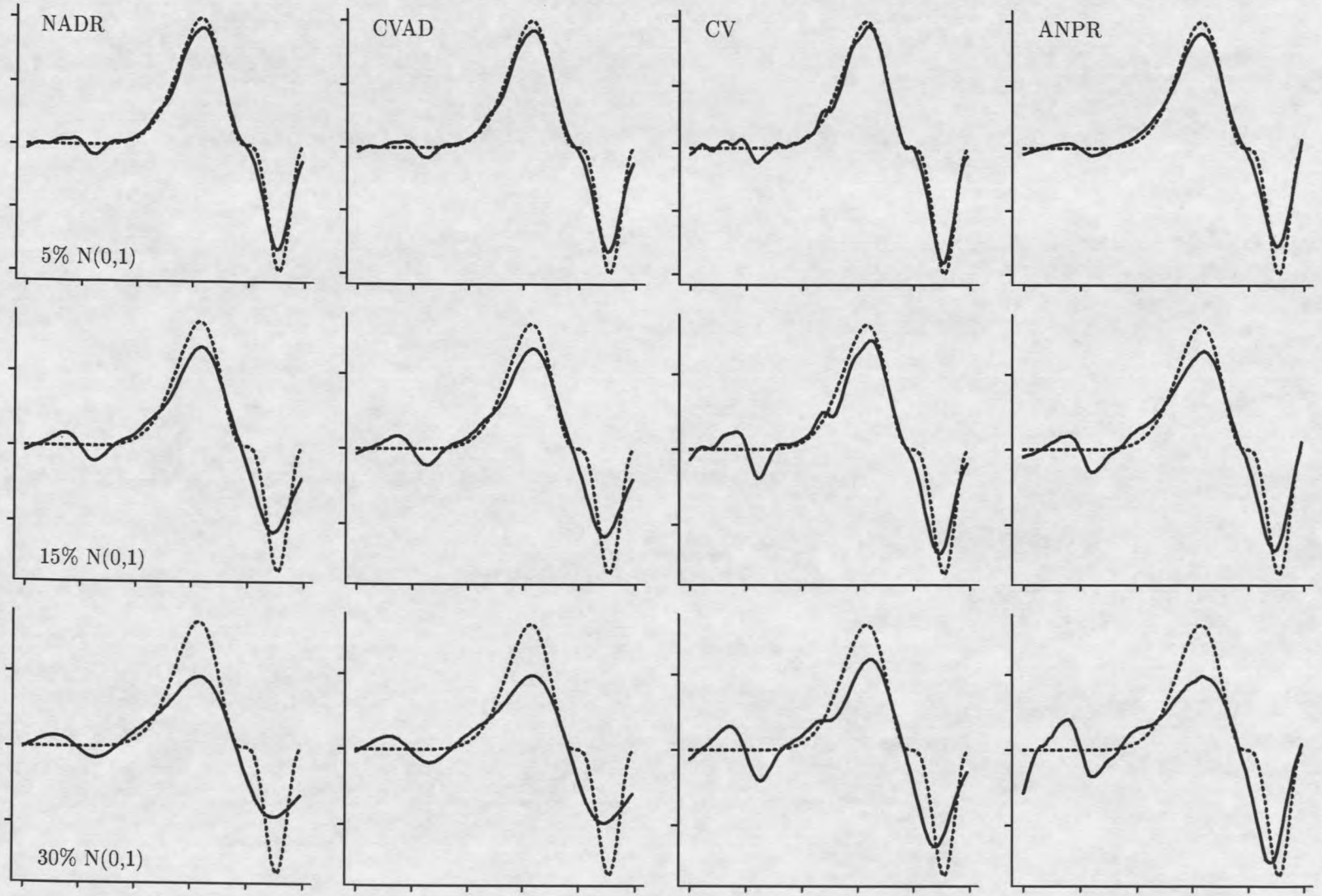


Figure 21: NADR, Cross-validated AD NPR, Cross-validated NPR,and Adaptive-bandwith Smooths of the three N(0,1) Datasets. Rows, from top to bottom, correspond to Var(ϵ) of 5%, 15% and 30% of the range of the true mean response, respectively. Each column corresponds to one smoothing technique. In each panel, the solid line denotes the smooth and the dashed line traces the true mean response.

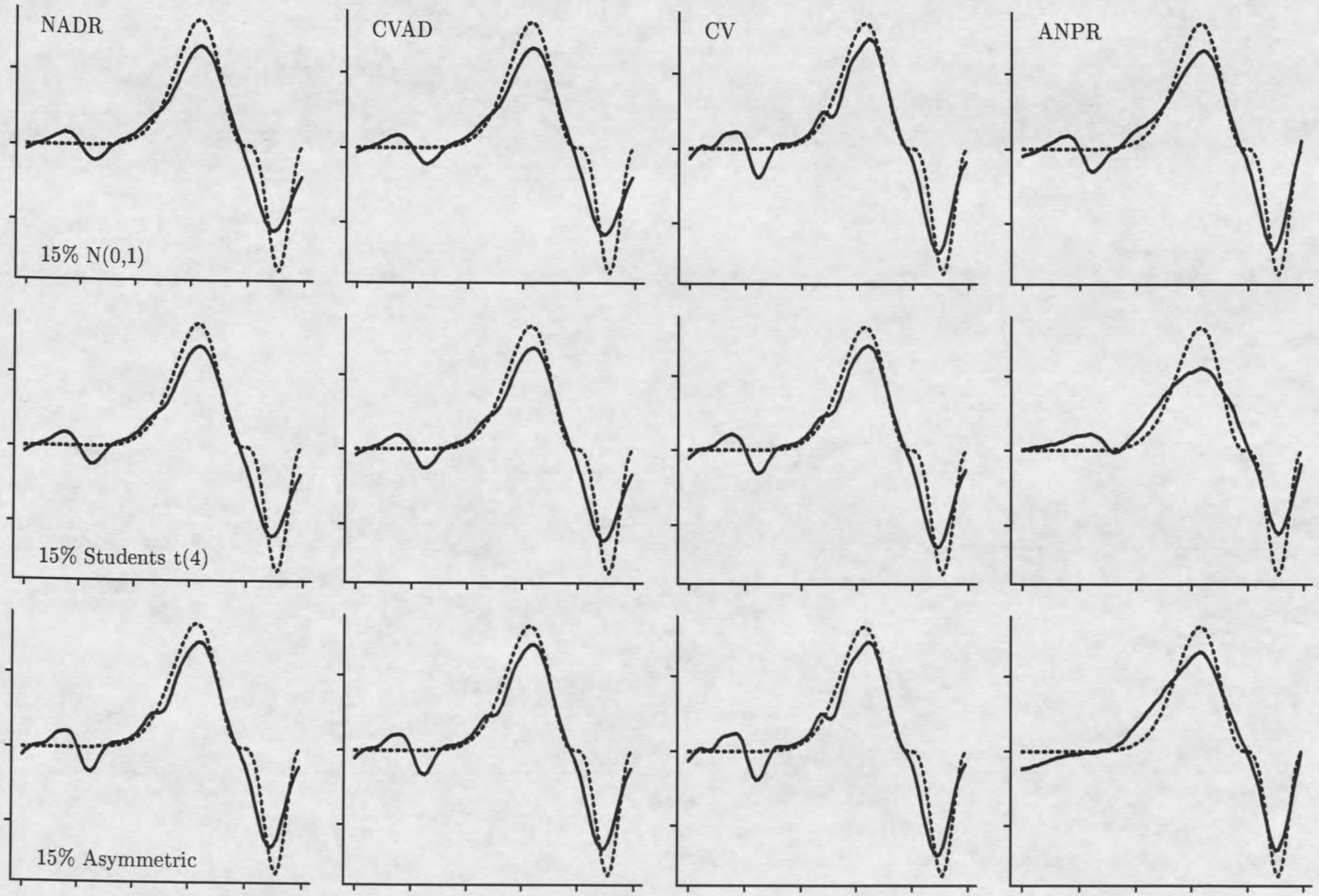


Figure 22: NADR, Cross-validated AD NPR, Cross-validated NPR, and Adaptive-bandwith Smooths of Datasets with N(0,1), Student's t(4) and Asymmetric Distributions where Var(ϵ) is 15% of the range of the true mean response. Rows, from top to bottom, correspond to N(0,1), Student's t(4) and Asymmetric models, respectively. Each column corresponds to one smoothing technique. In each panel, the solid line denotes the smooth and the dashed line traces the true mean response.

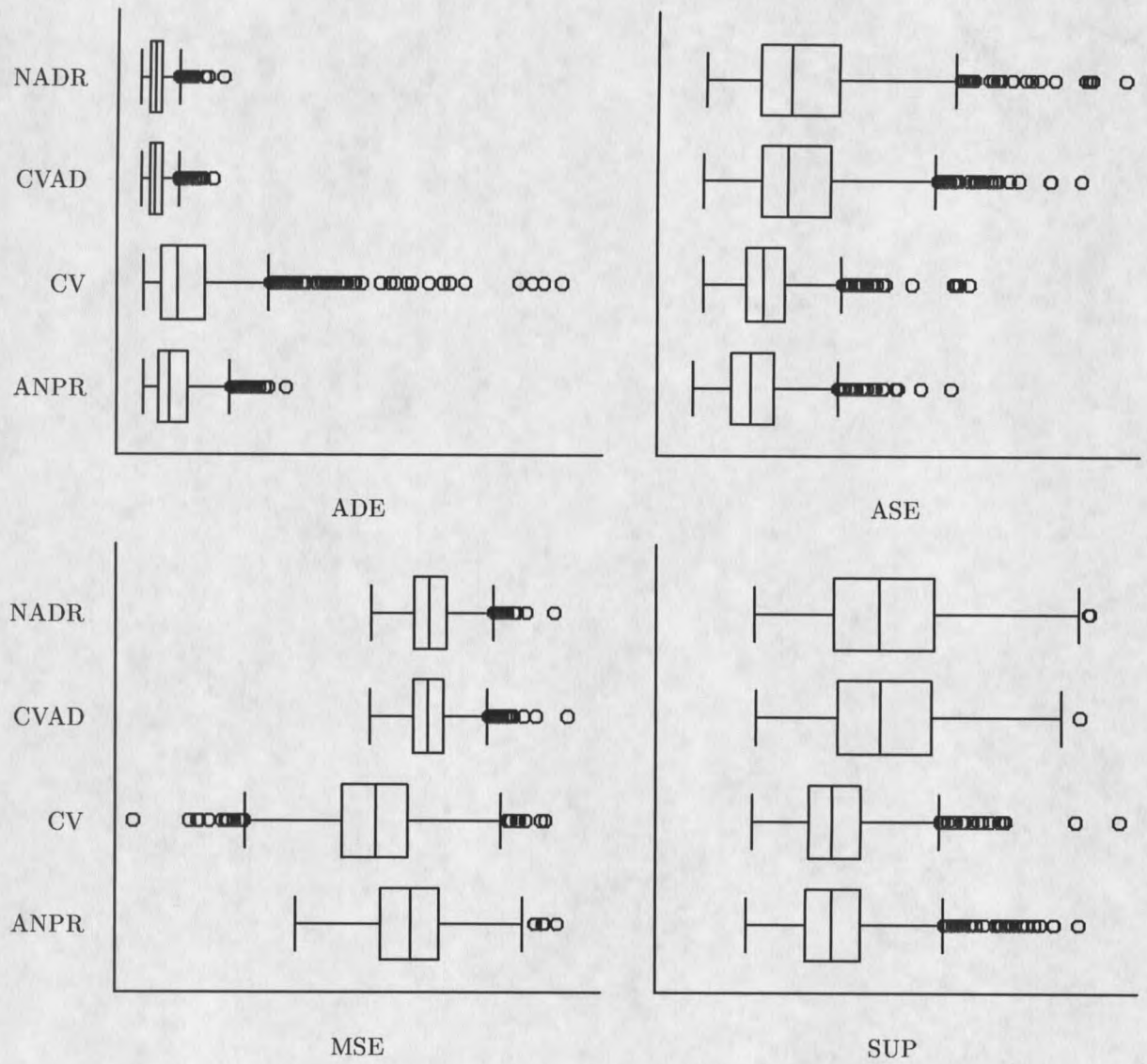


Figure 23: N(0,1) Performance Results for 15% Datasets (model 2).

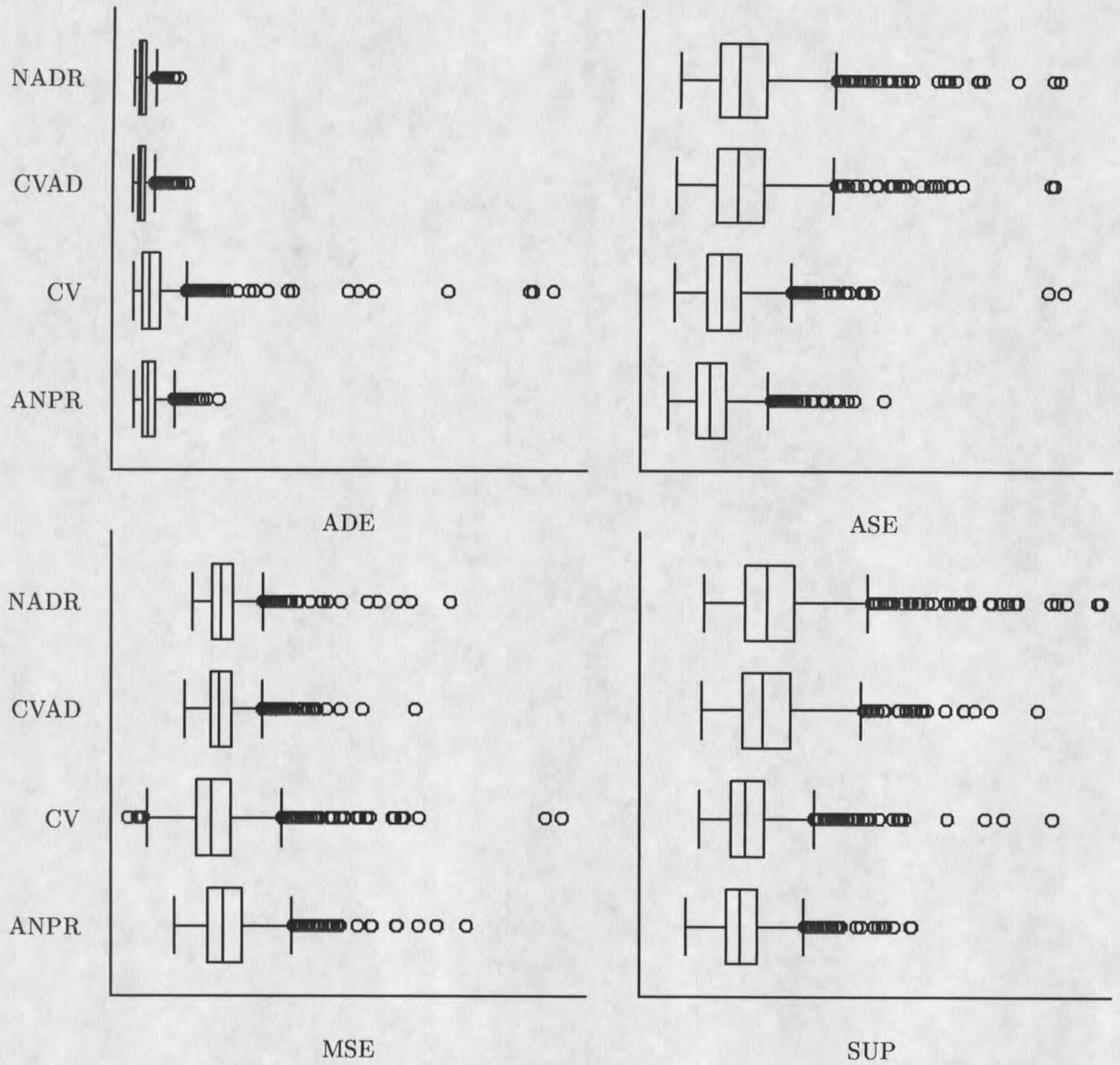


Figure 24: Student's t(4) Performance Results (model 4).

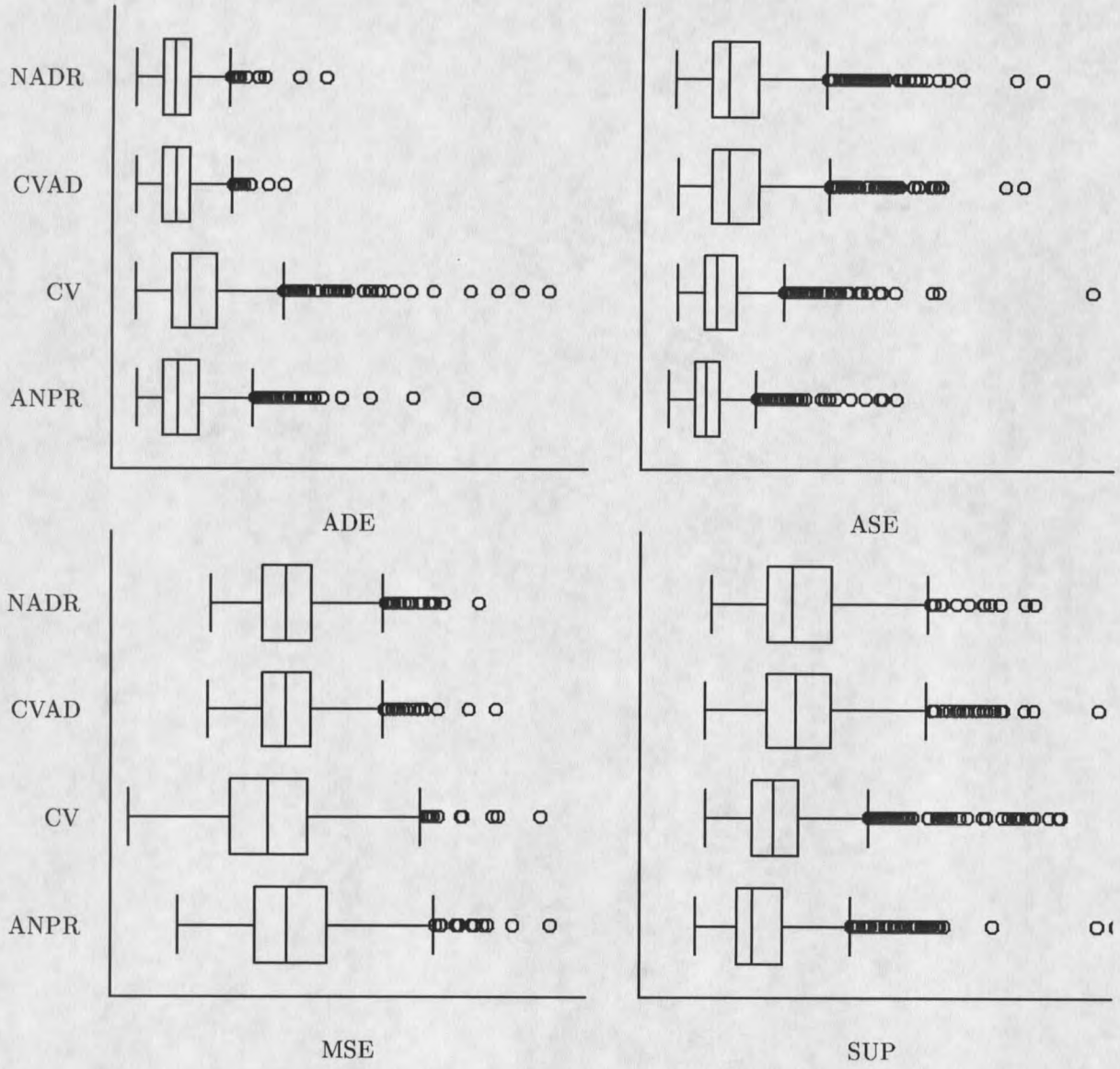


Figure 25: Spline Performance Results (model 5).

A $\chi^2(4)$ Multiplicative Model Example

Let us now examine an example featuring a multiplicative model (model 6 in Table 8):

$$Y_i = m(x_i) \cdot \epsilon_i \, .$$

Earlier, I noted that it was not appropriate to apply NADR directly to this model using the natural residual. Instead, I suggested a corrected residual whose moments more closely match χ^2 moments (Equations 8):

$$r_i^* = \nu \cdot \frac{Y_i}{\widehat{m}(x_i)} \, .$$

For certain smoothers, I will use the log transformation to adjust the multiplicative model so that the new model conforms to the the standard NPR assumption of additive linkage (Table 1). We will fit the log-transformed model when applying CV, CVAD and ANPR. I will then back-transform $\widehat{\log}[m(x)]$ to obtain an estimate of $m(x)$. If possible, the back-transformation must account for bias in the estimate. No transformation will be required for NADR.

For the log-transformed model,

$$\log[Y_i] = \log[m(x_i)] + \log[\epsilon_i]$$

the expected value of the nonparametric estimator of the mean response can be approximated:

$$\begin{aligned}
\mathrm{E}[\widehat{\log}[m(x)]] &= \mathrm{E}\left[\sum_{j=1}^{n} w_j \log(Y_j)\right] \\
&= \sum_{j=1}^{n} w_j \mathrm{E}[\log(Y_j)] \\
&= \sum_{j=1}^{n} w_j \mathrm{E}\left[\log(m(x_j)) + \log(\epsilon_j)\right] \\
&\approx \sum_{j=1}^{n} w_j \left[\log(m(x_j)) + \log(\nu) - \frac{1}{\nu}\right] \\
&\approx \sum_{j=1}^{n} w_j \log(m(x_j)) + \log(\nu) - \frac{1}{\nu} \, .
\end{aligned}$$

It is common practice to back transform LogNormal estimates. Hence, we see that on the original scale, the expected value of the standard nonparametric kernel estimator is

$$\begin{aligned} \mathrm{E}[\widehat{m}(x)] &= \mathrm{E}[\exp(\widehat{\log[m(x)]})] \\ &\approx \nu \exp(-\nu) \cdot \prod_{j=1}^{n} m(x_i)^{w_j} \\ &\neq m(x) \, . \end{aligned}$$

The expected value of the standard NPR estimator using the log-transform approach is not what we desire.

Results and Discussion (model 6). A scatterplot of an example multiplicative $\chi^2(4)$ dataset with the true mean response is shown in Figure 26.

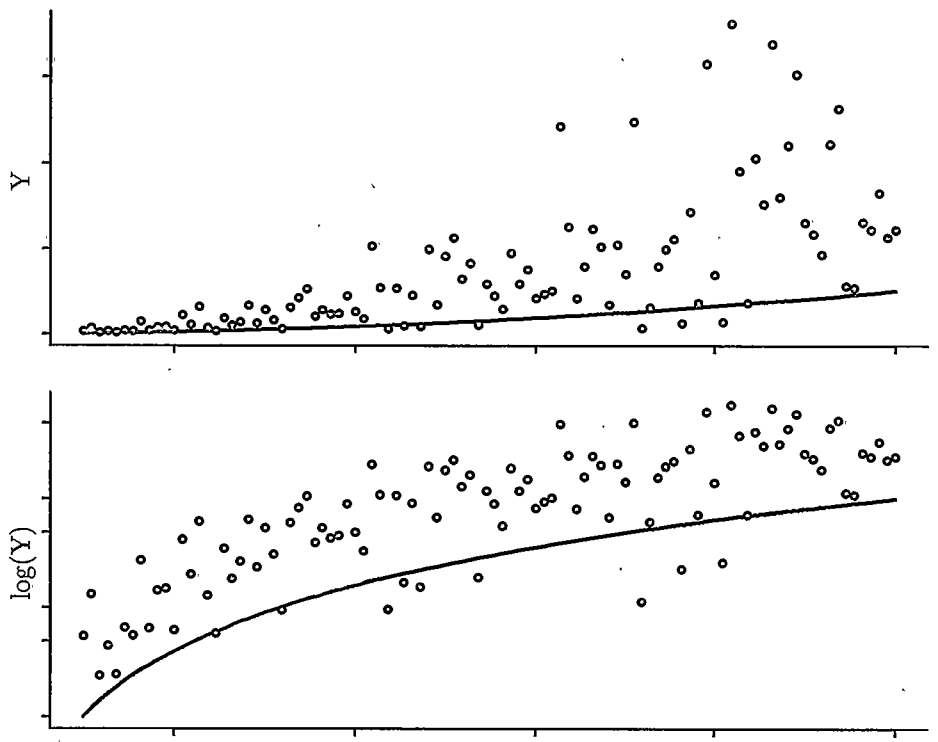


Figure 26: Observations of the true mean response x^2 (solid line) perturbed by $\chi^2(4)$ errors. The top panel shows the dataset in the original scale. The bottom panel presents the log-transformed dataset.

NADR, CV, CVAD and ANPR smooths of an example $\chi^2(4)$ dataset using the appropriate median bandwidth, when applicable, are shown in Figure 27. Boxplots of the performance distributions are presented in Figure 28.

Visually, the CVAD smooth appears to be the best estimate of the true mean response. The NADR and CV estimates are similar to each other, and much less smooth than the CVAD estimate. The ANPR smooth is very similar to the CVAD smooth, though it displays a certain jitter as seen in the additive ANPR model smooths.

Quantitatively, the characteristics of the four distributions for a given performance measure are more alike than not. In particular, the middle quartiles appear almost identical, for the most part.

The boxplots of the performance measures suggest that the four methods would provide similar results. The graphs of the example smooths suggest otherwise. Nonetheless, this visual difference may be due to the unique qualities of the example dataset which is one realization of model 6.

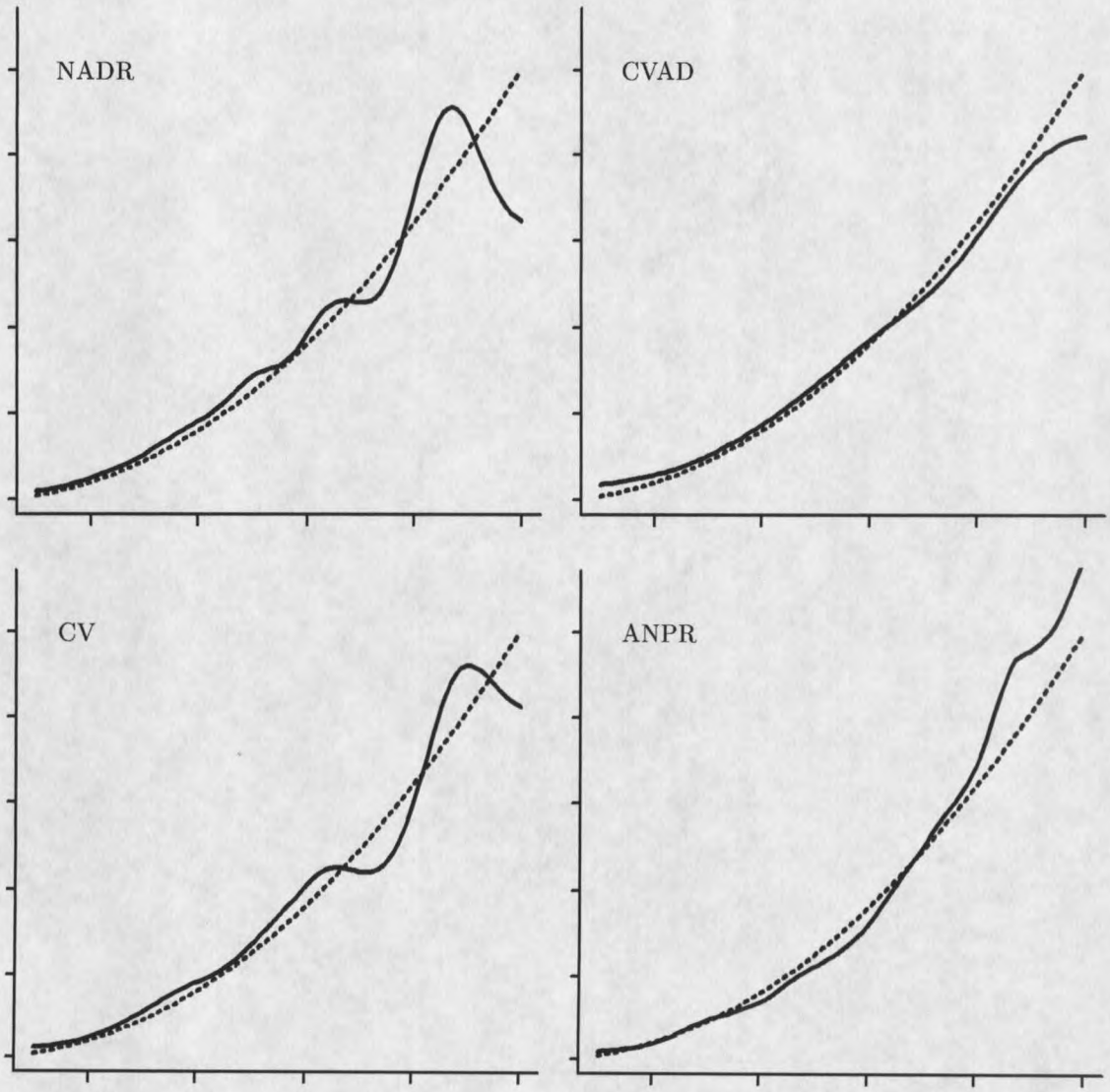


Figure 27: NADR, CV, CVAD and ANPR smooths of an example $\chi^2(4)$ multiplicative model dataset. The solid line in each panel denotes the smooth while the dashed line marks the true mean response.

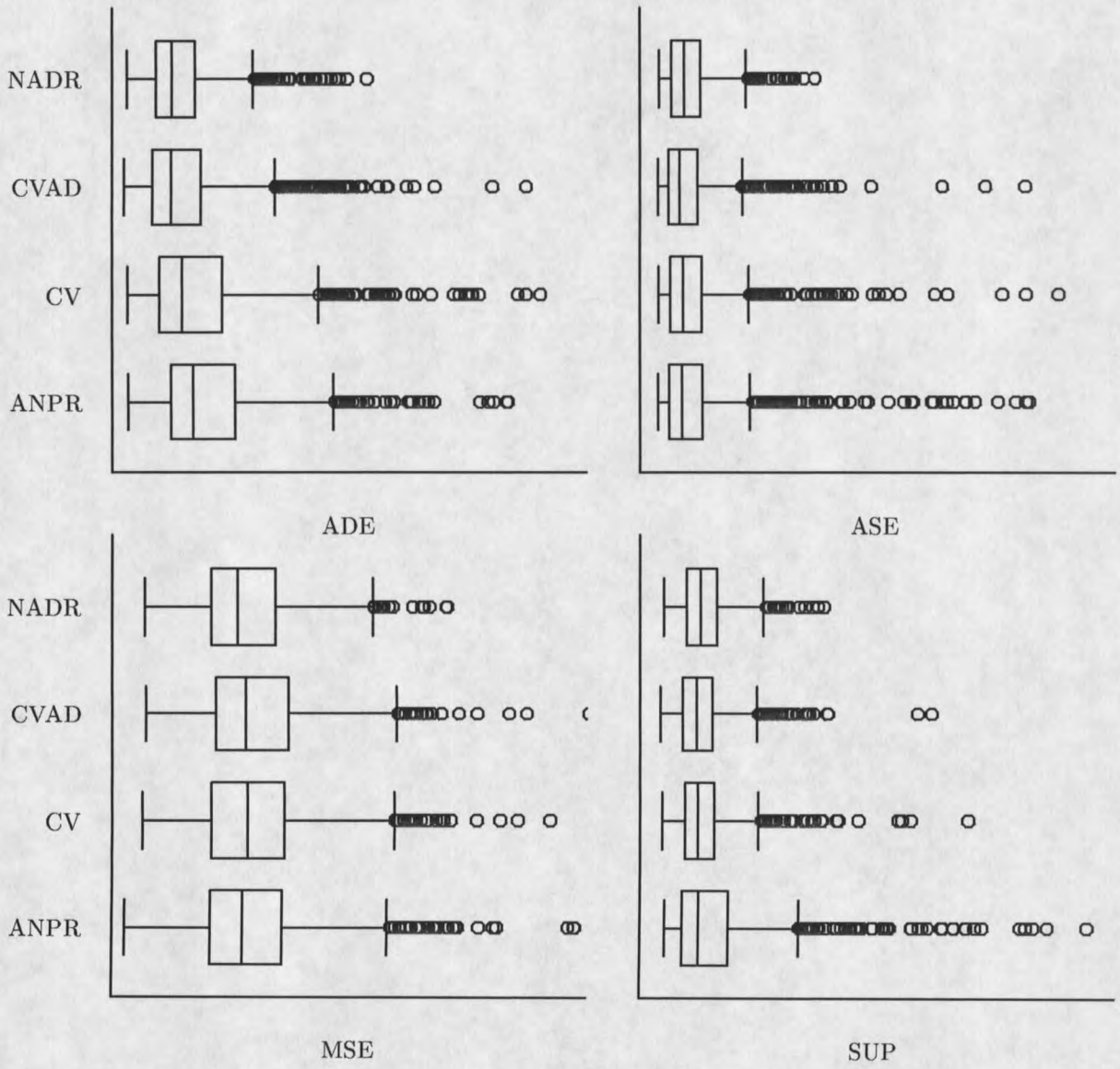


Figure 28: $\chi^2(4)$ Performance Results (model 6).

A Poisson($\lambda(m(x))$) Model Example.

Suppose that $Y(x)$ is distributed as a Poisson random variable with parameter $\lambda(x)$. Let us assume that $\lambda(x)$ is large so that the standardized random variable is approximately normally distributed (McCullagh and Nelder, 1989):

$$\begin{aligned} r(x) &= \frac{Y(x) - \lambda(x)}{\sqrt{\lambda(x)}} \\ &\sim N(0,1)\,. \end{aligned}$$

As in the previous examples, we could obtain an estimate of $\lambda(x)$ using NADR and the natural residuals. To estimate $\lambda(x)$, we would select a value for the smoothing parameter, calculate the standardized residuals (which serve as the natural residuals, in this case), and compare their empirical distribution to the N(0,1) cumulative distribution. We would label the smooth with the smallest AD goodness-of-fit statistic as the best NADR estimate of $\lambda(x)$.

Let us begin with an examination of the mean and variance of the standardized residual. First,

$$\begin{aligned} \mathrm{E}[r_i] &= \mathrm{E}\left[\frac{Y_i - \widehat{\lambda}(x_i)}{\sqrt{\widehat{\lambda}(x_i)}}\right] \\ &\approx 0\,. \end{aligned}$$

Because

$$\mathrm{E}\left[\frac{Y_i - \widehat{\lambda}(x_i)}{\sqrt{\widehat{\lambda}(x_i)}}\right] = \mathrm{E}\left[\left(Y_i - \sum_k w_k Y_k\right)\left(\sum_k w_k Y_k\right)^{-1/2}\right]$$

it follows from a Taylor series approximation that

$$\mathrm{E}[r_i] \approx \frac{1}{2}\frac{1}{\sqrt{\lambda_i}}\left[\sum_j w_j^2 - w_i\right]$$

and, hence,

$$\mathrm{E}[r_i] \to 0 \quad \text{as} \quad \lambda(x_i) \to \infty\,.$$

For large values of $\lambda(x_i)$, the expected value of the natural residual is close to zero.

A Taylor series approximation of the variance of r_i is easily produced:

$$\begin{aligned}
\mathrm{Var}[r_i] &= \mathrm{Var}\left[(Y_i - \sum_k w_k Y_k)(\sum_k w_k Y_k)^{-1/2}\right] \\
&\approx \sum_j \mathrm{Var}[Y_j]\frac{\partial}{\partial Y_j}\left[(Y_i - \sum_k w_k Y_k)(\sum_k w_k Y_k)^{-1/2}\Big|_{\mu_{\mathbf{Y}}}\right]^2 \\
&\quad +2\sum_{i<j}\mathrm{Cov}[Y_i, Y_j]\left[\frac{\partial^2}{\partial Y_i \partial Y_j}(Y_i - \sum_k w_k Y_k)(\sum_k w_k Y_k)^{-1/2}\Big|_{\mu_{\mathbf{Y}}}\right] \\
&= 1 - 2w_i + \sum_j w_j^2 \,.
\end{aligned}$$

The variance $\mathrm{Var}[r_i]$ can be bounded by considering the case when $w_i = 1$ and the case $w_j = 1/n$ for all $j = 1, \ldots, n$. Then,

$$0 \leq \mathrm{Var}[r_i] \leq \frac{n-1}{n} \,.$$

The approximate mean and variance of r_i, and the bounds on these, suggest no correction to the natural residual for use in NADR. The variance upper bound (i.e., $\mathrm{Var}[r_i] < 1$) suggests that NADR may over-smooth in order to produce residuals with a larger variance more closely matching the N(0,1) distribution.

Poisson Datasets. The Poisson models for the three examples considered in this section featured a quadratic mean response function (models 7, 8, and 9 of Table 8),

$$\lambda(x) = b_1(x - 45)^2 + 100$$

with

$$x \in [10, 100]$$

and

$$b_1 \in \{80, 13.337, 4\} \,.$$

As in previous examples, datasets were generated wherein the random variation among x-neighboring random variables is a slight to significant proportion of the overall variability. The

parameter, b_1, was chosen so that the minimum variance of the random variable is roughly 5%, 30% and 100% of the range of the mean response. Further, the mean y-intercept (100) was chosen so that λ could always be considered "large" and the normal approximation would be acceptable. Scatterplots of example Poisson datasets and the true mean responses are shown in Figure 29.

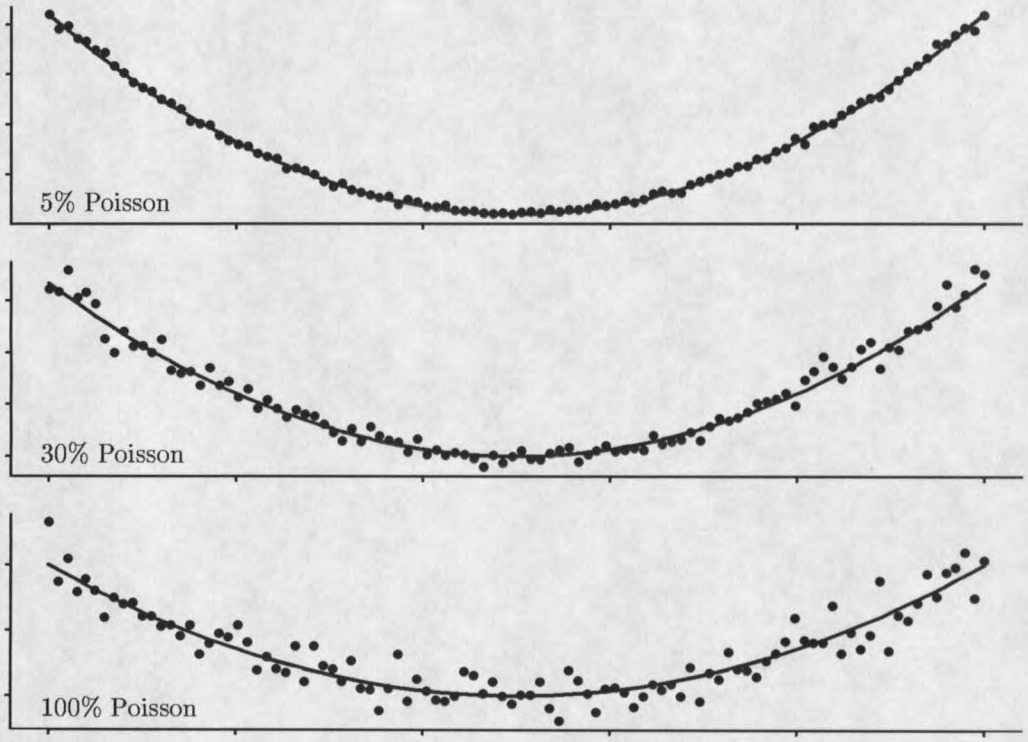


Figure 29: From top to bottom, example Poisson datasets (points) with the true mean response (solid line), where the $Var(\epsilon)$ is 5%, 30% and 100% of the range of the true mean response.

Results and Discussion (models 7-9). The example Poisson dataset was smoothed using the median bandwidths from the NADR, CV, and CVAD simulations and then graphed (Figure 30). The graphs show that the median smooths by all methods and for all levels of variability closely approximated the true mean response. The interquartile ranges of standardized differences $(\widehat{m}(x_i) - m(x_i))/m(x_i)$ ranged from -1% to 2%. Visually, the smoothers performed equally (and surprisingly) well across the range of variability.

Boxplots of the performance measures from the simulations wherein the minimum variance was 100% of the range of the true mean response are presented in Figure 31. The boxplots for performance results from the 5% and 30% simulations were similar to the 100% results and will not be discussed here.

The natural residual in this case is equivalent to the natural residual featured in the additive model. Therefore, we could expect the smoothing methods to perform under this scenario much the same as in the additive N(0,1) model case. The plots of smooths and the boxplots bear this out. With respect to ASE and SUP, ANPR had the smallest average error with approximately the same spread as CV and CVAD. The sample mean and variance of the ASE increased from ANPR to CV and CVAD and, finally, to NADR.

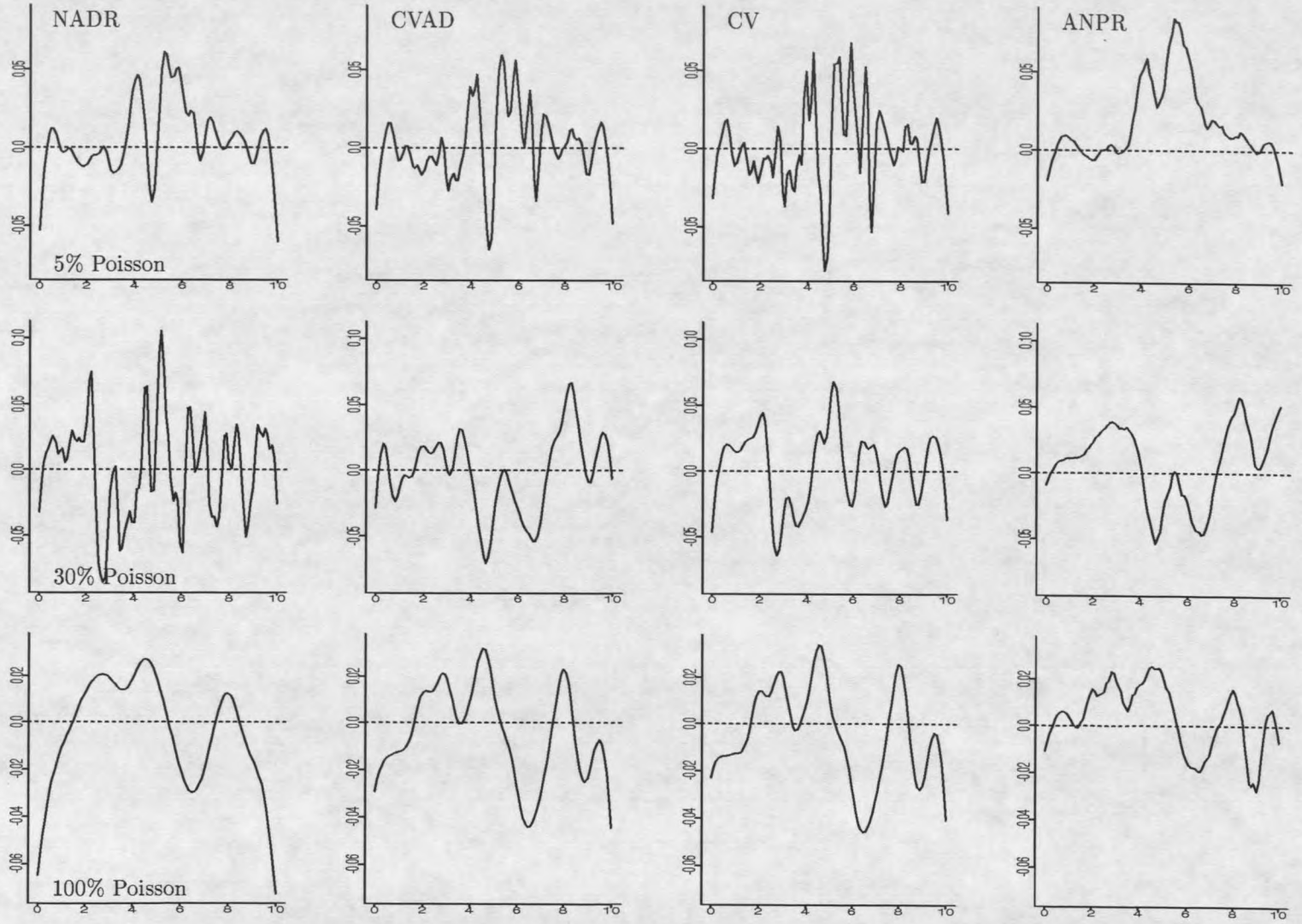


Figure 30: Differences between the true mean response and the NADR, Cross-validated NPR, Cross-validated AD and Adaptive-bandwith smooths of the three Poisson Datasets. Rows, from top to bottom, correspond to models with Var(ϵ) of 5%, 30%, and 100% of the range of the true mean response, respectively. Each column corresponds to one smoothing technique.

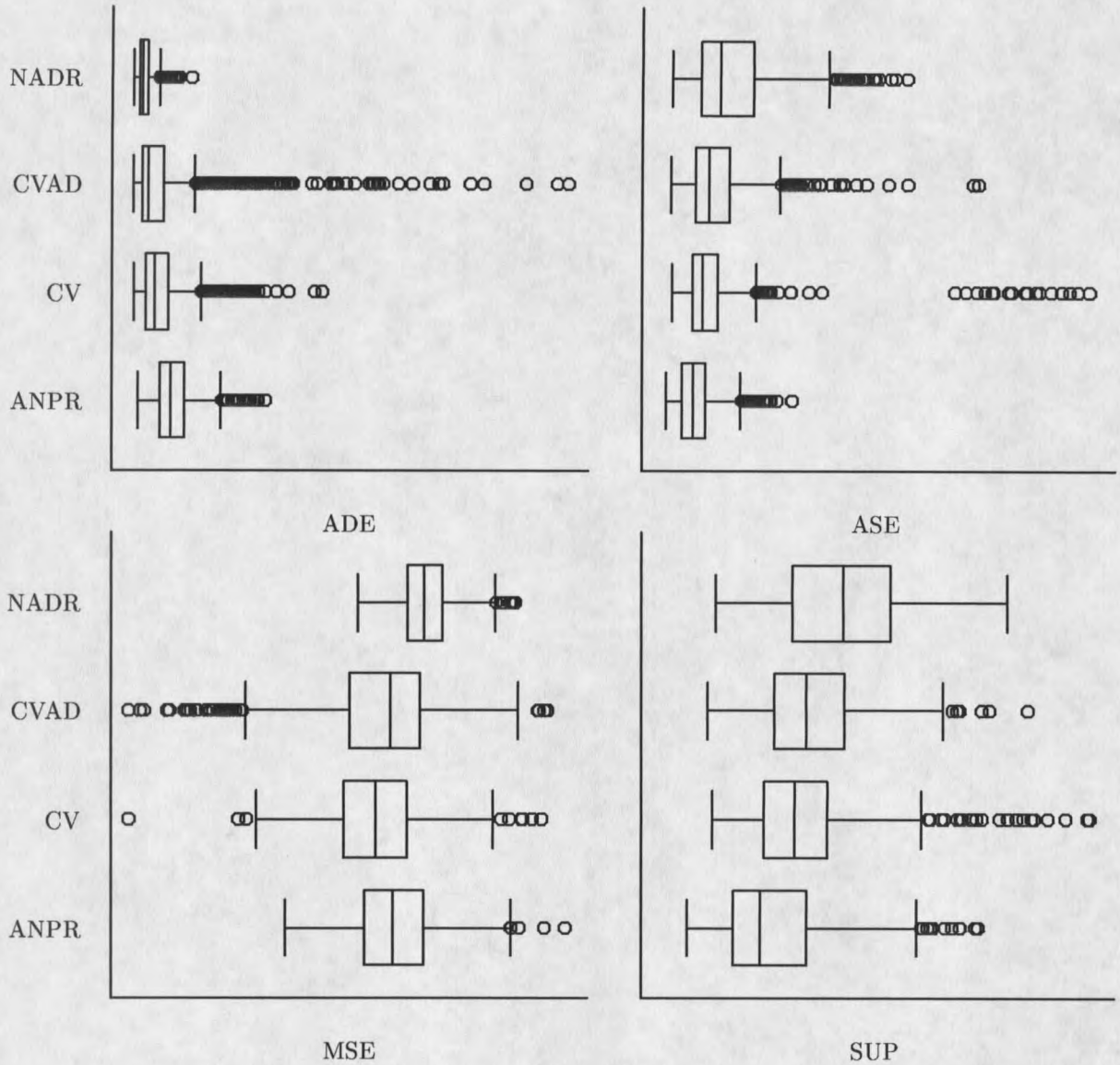


Figure 31: Poisson Performance Results for the 100% Datasets.

Summary of Simulation Results

In almost all cases, the four smoothing methods (NADR, CVAD, CV and ANPR) performed equally well. Quantitatively, NADR marginally out-performed the other methods in the mulitiplicative model example. ANPR and the others marginally outperformed NADR in the additive model example, as well a Poisson model example which had very similar characteristics to the additive models.

Visually, the NADR procedure most often produced the smoothest estimates. CVAD and CV estimates appeared less smooth, in that order. In regions where the true mean response changed quickly, increased smoothness also resulted in increased bias. The bias was reduced, however, in regions where the true mean response changed slowly. ANPR produced estimates which, though often as smooth as NADR estimates in overall appearance, had a disconcerting local roughness not characteristic of the true mean response functions.

With respect to implementation, the commercially-available ANPR method is the best choice. This adaptive method requires no user-selected parameter settings to provide reasonable results. It runs very quickly in comparison with the other methods. Usually, it was a factor of 10 times faster than the NADR method and a factor of 100 to 1000 times faster than the CV or CVAD methods. The variability in processing speed of the latter two results from the ability to optimize the cross-validation algorithm in some cases.

This work makes it clear that one should evaluate the bias of nonparametric estimators, especially when an additive model technique such as cross validated nonparametric regression is applied to a non-additive model that has been transformed to an additive model. The mean response of the transformed model most likely will not be the mean response of the original model. Nonetheless, the bias induced by the transformation may be corrected or reduced by taking actions suggested by the examination of the bias of the estimator with respect to the assumed model.

CHAPTER 5

ADR DIAGNOSTICS

Modeling would be uncomplicated if the design set spanned the design space at a sufficient density, and if all data followed the assumed stochastic model. A model could be fit and exercised without regard to evaluating the model assumptions, experimental design, or sampling plan. Alas, this is not often the case.

Some design points may have unintentionally high leverage. Some design variables may be highly collinear. Simplifying stochastic assumptions rarely apply to real phenomena or measurements. The stochastic model may not apply for a subset of the data. One or more observations may exert inordinate influence on parameter estimates, or may be outliers with regard to the mean response. Consequently, model fitting would be incomplete if the modeler failed to critique the model fit.

The need to critique model fit has lead to the development of model diagnostic methods. To assess ADR fit, an ADR diagnostic method must address three questions:

- Do I have the correct random distribution, mean response, and stochastic model?
- Do some observations depart from the ADR model? Are there outliers, or other data anomalies?
- How are ADR parameter estimates influenced by each datum?

Methods to appraise LSR fits have been developed by many authors, and these diagnostic methods add substantial value to the LSR modeling process. Developing methods for model diagnosis is more effective if one follows a few guiding principles. Cook and Weisberg (1986) identify three desirable properties of a diagnostic method. A good diagnostic

- identifies aspects of a problem that do not conform to the hypothesized model
- suggests appropriate remedial action
- reveals important phenomena

Diagnostics possessing these features have been developed for LSR, but diagnostic methods for ADR are not available. Because diagnostics are a necessary part of the ADR modeling process (Table 1), I have derived some ADR diagnostic methods, and my results to date are presented in this chapter. These results show how to adapt a subset of LSR diagnostics for use in ADR. In addition, this chapter also examines some diagnostics specific to ADR.

Regression Diagnostics

A fitted model is determined by the variables, the observations, and the modeling assumptions that tie these together. To evaluate fit, the effect of each of these factors must be investigated. Methods to investigate these factors have been proposed by many authors. Almost all these methods share two features: a numerical measure of the effect, and a display of this measure to aid in assessing the effect. For example, a plot of the residual empirical distribution provides a visual assessment of the effect of assuming a certain error distribution (though, this plot is not limited to this use). Commonly used diagnostic measures are functions of the raw and transformed residuals, as well as the points in the design space. Often-used graphical displays include

- residual-based graphs such as histograms, boxplots, and stem-and-leaf plots
- scatter plots of residuals against sampling sequence, functions of the explanatory variables, and fitted values
- probability, Q-Q and other goodness-of-fit plots
- leverage and influence plots

The collection of diagnostic methods relating to variables, observations and LSR modeling assumptions are extensive. It appears that many of these methods can be applied directly, and

almost all can be applied, in spirit, to assess ADR fitted models, because the driving philosophy behind each LSR diagnostic method is relevant in the ADR situation. The embodiment of this philosophy in an ADR diagnostic tool, however, may take a much different form.

For instance, useful LSR measures to evaluate the impact of a point in the design space (a row vector in the design matrix $\mathbf{X}$) have been developed from the projection matrix $\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$. The projection matrix has no direct role in ADR, however. Indeed, if the model is nonlinear, the projection matrix may be only tenuously related to the solution. Nonetheless, the need to assess the impact of points in the design space remains, and this need must be met with a measure appropriate to the model, and the fitting technique.

Quantitative goodness-of-fit diagnostics such as R^2, though important in LSR, are often lost among the other LSR diagnostics. Goodness-of-fit is at the forefront of ADR, however. Indeed, emphasis on goodness-of-fit and the AD goodness-of-fit statistic is the defining difference between the LSR and ADR methods. Therefore, interpretation of diagnostics in the ADR setting almost always can be couched in terms of goodness-of-fit.

Observations and Regression

An observation that significantly affects LSR model fit has been labelled, by convention, an "outlier", an "influential observation", or a "high leverage point" (Chatterjee and Hadi, 1988):

- **Outlier.** An observation for which the as yet-to-be-defined residual is large in magnitude compared to residuals of other observations in the data set.
- **Influential Observation.** An observation that excessively influences parameter estimation as compared to other observations in the data set.
- **High-Leverage Point.** A point "for which the input vector $\mathbf{x}_i$ is, in some sense, far from the rest of the data" (Hocking and Pendleton, 1983). An observation which is extreme or isolated in the design space (the column space of $\mathbf{X}$) will have high leverage. Leverage is a property of the predictor variables and the mean response function, and not of the reponse variable.

Determining the role of each observation in the fit begins with the definition and discussion of model residuals.

Residuals and Outliers.

The influence of an observation on a model is assessed by examining the difference between the observation and its fitted value. In practice, five related definitions of this difference, or "residual", have proven useful when evaluating the influence of an observation (assuming an additive model).

The first definition establishes the ordinary (natural) residuals. In matrix notation,

$$\mathbf{e} = \mathbf{Y} - \widehat{\mathbf{Y}}$$

or, element by element,

$$e_i = Y_i - \widehat{Y}_i \, .$$

Caution is required in evaluating ordinary residuals, because their magnitudes reflect both the random effects, and the structure in the contextual information (i.e., the design $\mathbf{X}$). The scale of the residuals, in physical units, may also be unfamiliar, further clouding the evaluation. Therefore, a second and third definition are offered. These two address the issue of an unfamiliar scale, and are called the normalized (**a**) and standardized (**b**) residuals, respectively:

$$\mathbf{a} = \frac{\mathbf{e}}{\sqrt{\mathbf{e}'\mathbf{e}}}$$

or, alternatively,

$$a_i = \frac{e_i}{\sqrt{\mathbf{e}'\mathbf{e}}}$$

and

$$\mathbf{b} = \frac{\mathbf{e}}{\widehat{\sigma}}$$

$$\text{where} \quad \widehat{\sigma} \;=\; \sqrt{\frac{\mathbf{e}'\mathbf{e}}{n-p}}$$

or, element by element,

$$b_i \;=\; \frac{e_i}{\widehat{\sigma}}\,.$$

The remaining two definitions attempt to adjust for the influence of the projection matrix $\mathbf{X}(\mathbf{X}'\mathbf{X})^{-}\mathbf{X}'$ as well as for the unfamiliar scale. The second of these includes an estimate of the variance σ^2 calculated without, and hence, independent of, the *ith* observation. These residuals are called, respectively, the internally and externally studentized residuals:

$$\begin{aligned} r_i &= \frac{e_i}{\widehat{\sigma}\sqrt{1-p_{ii}}} \\ r_i^{\star} &= \frac{e_i}{\widehat{\sigma}_{(i)}\sqrt{1-p_{ii}}} \end{aligned}$$

where p_{ii} is the ith diagonal element of the projection matrix and

$$\sigma_{(i)} = \sqrt{\frac{\sum_{j\neq i}(Y_j - \widehat{Y}_j)^2}{(n-p-1)}}\,.$$

The definitions for ordinary, normalized and standardized residuals apply directly to Anderson-Darling regression. Directly applying the definitions for internally and externally studentized residuals in ADR is not possible, however, because the estimates, $\tilde{\mathbf{Y}}$, arise in practice from an iterative procedure, and do not result from the application of a projection matrix. Nonetheless, these residuals may be informative in the ADR setting when a projection matrix can be constructed because the design will affect the ADR fit if the response is a function of the design variables.

Influence.

Measures of influence also can be defined. Two common diagnostics in linear regression, Cook Distance and Welsch-Kuh Distance, make sense in the Anderson-Darling regression setting. Welsch Distance and Modified Cook Distance, two measures closely related to Welsch-Kuh Distance, are also directly applicable in ADR diagnosis.

The Cook Distance is a global measure of the influence of the *ith* observation. Welsch-Kuh Distance measures the local influence of the *ith* observation on the fit at the *ith* point. Welsch Distance and Modified Cook Distance are weighted versions of Welsch-Kuh Distance. Let k be the dimension of the design space $\mathbf{X}$ and p_{ii} the ith diagonal element of the projection matrix. Then,

Cook Distance

$$\begin{aligned} C_i &= \frac{(\widehat{\mathbf{y}} - \widehat{\mathbf{y}}_{(i)})^T(\widehat{\mathbf{y}} - \widehat{\mathbf{y}}_{(i)})}{k\widehat{\sigma}^2} \\ &= \frac{p_{ii}}{p(1-p_{ii})} r_i^2 \end{aligned} \tag{15}$$

Welsch-Kuh Distance

$$\begin{aligned} WK_i &= \frac{\widehat{y}_i - \widehat{y}_{i(i)}}{\mathrm{SE}(\widehat{y}_i)} \\ &= \frac{n-1}{1-p_{ii}} r_i^* \end{aligned} \tag{16}$$

Welsch Distance

$$W_i = WKi\sqrt{\frac{n-1}{1-p_{ii}}} \tag{17}$$

Modified Cook Distance

$$C_i^* = WK_i\sqrt{\frac{n-k}{k}}. \tag{18}$$

These diagnostics include estimates $\widehat{y}_{i(i)}$ from datasets with the *ith* observation excluded. In classical regression, the closed form estimator $\widehat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$ can be manipulated so that re-estimation is not required in order to determine these measures of influence. ADR, however, is based on an iterative optimization algorithm. Most likely, no closed form estimator exists. No algebraic manipulation is available to foreshorten "leave one out" calculations. Hence, computation of an ADR influence measure such as Cook distance or Welsch-Kuh distance requires n "leave one out" fits for a sample of n observations.

Welsch-Kuh distance also requires an estimate of $\mathrm{SE}(\widehat{y}_i)$. Without a closed form ADR estimator $\widehat{y}_i$, a closed form estimator of standard error most likely does not exist. Nonetheless, we can calculate Welsch and Welsch-Kuh distances by following a suggestion by Tukey (1958) and Efron (1982); estimate the variance of an estimator by jackknifing:

$$\mathrm{Var}(\tilde{\theta}) = \frac{n-1}{n}\sum_{i=1}^{n}[\tilde{\theta}_{(i)} - \tilde{\theta}_{(.)}]^2$$

where $\tilde{\theta}_{(i)}$ is the parameter estimate with the ith observation removed, and

$$\tilde{\theta}_{(.)} = \frac{\sum \tilde{\theta}_{(i)}}{n}.$$

If the ADR model is linear in its parameters (i.e., $m(\mathbf{X}) = \mathbf{X}\beta$), we can estimate the variance of $\mathrm{Var}(\tilde{y}_i)$ by

$$\begin{aligned}\mathrm{Var}(\tilde{y}_i) &= \mathrm{Var}(\mathbf{x}_i^T\tilde{\beta}) \\ &= \mathbf{x}_i^T\mathrm{Var}(\tilde{\beta})\mathbf{x}_i\,.\end{aligned}$$

Leverage.

As defined above, a high leverage point is an observation that is somehow remote in the design space $\mathbf{X}$. Measuring leverage in LSR models reduces to considering the role of the projection matrix alone in the estimation of the parameters.

Directly applying LSR leverage diagnostics to ADR is not possible because the LSR projection matrix $\mathbf{P}$ has no counterpart in ADR. Nonetheless, an isolated or extreme point in the design space will affect the ADR fit. Therefore, for ADR cases when the projection matrix can be defined, projector-based diagnostics should provide a measure of how extreme a point is in the design space.

ADR Diagnostics Based on $\delta_A(F_n, F_\theta)$

The ADR method produces estimates by directly seeking the best goodness-of-fit of a model to the data by minimizing the AD distance $\delta_A(F_{n.\theta}, F_0)$. Let $\tilde{\delta}_A$ be the minimum Anderson-Darling distance:

$$\begin{aligned} \tilde{\delta}_A &= \delta_A(F_{n.\tilde{\theta}}, F_0) \\ &= \inf_{\theta \in \Theta} \delta_A(F_{n.\theta}, F_0) \ . \end{aligned}$$

In general, if the model is under-fit, the residuals, and $\tilde{\delta}_A$, are larger than expected. If the model is over-fit, so that the residuals are much smaller than expected, then $\tilde{\delta}_A$ is smaller than expected.

What more can be learned from $\tilde{\delta}_A$ about a model's goodness-of-fit? We know that $\tilde{\delta}_A$ is an overall or omnibus measure of the fit of $F_{\tilde{\theta}}$ to F_n (or $F_{n.\tilde{\theta}}$ to F_0). As a goodness-of-fit statistic, $\tilde{\delta}_A$ is useful for testing the null hypothesis F_θ equals $F_{\tilde{\theta}}$. If $\tilde{\delta}_A$ is too large, we would reject the fitted model, attempt to determine its shortcomings or shorcomings in the data, make appropriate adjustments, and then refit.

The value of $\tilde{\delta}_A$ may also tell us about the influence of individual observations. Boos (1981) suggests that even though Anderson-Darling estimation is a robust technique, the Anderson-Darling statistic is sensitive to outliers, and therefore, observing the differences in $\tilde{\theta}$ may be informative as each datum is left out in turn and $\tilde{\delta}_A$ is recalculated. The closed-form expression of δ_A, a sum of the individual contributions of the order statistics, also suggests a measure of the influence of an individual observation. These, and the role of $\tilde{\delta}_A$ as a goodness-of-fit measure, will be explored in the following sections.

Goodness-of-Fit: Diagnosing with $\tilde{\delta}_A$.

The approximate null distribution of $\tilde{\delta}_A$ when two or more parameters are estimated is very similar to approximate null distribution when one parameter is estimated (Figure 7). Comparing $\tilde{\delta}_A$ resulting from fitting a two or more parameter model using ADR to the appropriate approximate null distribution is helpful when diagnosing the goodness of fit. If $\tilde{\delta}_A$ is in the right tail, the evidence indicates that $\tilde{\delta}_A$ is larger than expected and the model is under-fit. If the statistic $\tilde{\delta}_A$

falls in the left tail, the evidence suggests that the model is over-fit. A more insightful examination of over- or under-fitting can be assessed by reviewing a QQ plot

The Distribution of $\tilde{\delta}_A$. The question becomes "When is $\tilde{\delta}_A$ to be considered too large or too small?". This question can be answered only by refering to the distribution of $\tilde{\delta}_A$. An unfortunate complication arises. For each parametric family, $\tilde{\delta}_A$ has a unique null distribution.

All is not lost, however. From a study of a range of symmetric, two-parameter, location-scale distributions, Boos (1981) conjectured that the null limiting distribution of $n\tilde{\delta}_A$ could be reasonably approximated by the distribution of

$$A_2^2 = \sum_{i=3}^{\infty} \frac{Z_i^2}{i(i+1)}$$

where the Z_i are independent standard normal random variables. He suggests that this approximation is valid for more than symmetric location-scale models, based on work using Monte Carlo simulation (Boos, 1981). He offered an analogy, based on the χ^2 goodness-of-fit statistic whose degrees of freedom decrease as the number of estimated parameters increased. He suggests that

$$A_p^2 = \sum_{i=p+1}^{\infty} \frac{Z_i^2}{i(i+1)}$$

be used for evaluating the case of p estimated parameters. The approximate limiting null distribution of $n\tilde{\delta}_A$ with no parameters estimated is A_0^2. The approximate limiting null distributions with the estimation of 1, 2 or more parameters, and the resulting loss in degrees of freedom, would be $A_1^2, A_2^2, \ldots$. Boos' results are based on work by Pollard (1980) who presented conditions for the convergence in distribution of a Cramér-Von Mises MD statistic to an random variable that was shown later to be a weighted sum of χ^2 random variables:

$$\inf_{\theta \in \Theta} n \int_{-\infty}^{\infty} [F_n(y) - F_\theta(y)]^2 w_\theta(y) dy \xrightarrow{d} \min_{x=(x_1,\cdots,x_p)} \int_0^1 [W(t) - x \cdot h(t)]^2 v(t) dt \, .$$

Here, $W(t)$ is the Brownian bridge on $C[0,1]$, $h(t) = (F_0^1(F_0^{-1}(t)), \ldots, F_0^p(F_0^{-1}(t)))'$ and $v(t) = \frac{w_0(F_0^{-1}(t))}{f_0(F_0^{-1}(t))}$ where $w_0(F_0^{-1}(t))$ is $w_\theta(y)$ with $\theta = \theta_0$ and $y = F_0^{-1}(t)$. Letting

$B(W) = \int W(t)h(t)v(t)dt$, Boos showed that

$$\begin{aligned}\min_{x=(x_1,\cdots,x_p)} \int_0^1 [W(t) - x \cdot h(t)]^2 v(t)dt &= \int W^2(t)v(t)dt + \min_x [x'\delta x - 2x'B(W)] \\ &= \int W^2(t)vtdt - B(W)'\Delta^{-1}B(W)\end{aligned}$$

where Δ is the matrix of probability limits of second derivatives (i.e., $\frac{1}{2}\frac{\partial}{\partial_i \partial_j} F_n \xrightarrow{P} \Delta_{ij}$). Boos noted that $\Delta^{-1}B(W)$ is an alternative representation of the limiting normal distribution of $n^{\frac{1}{2}}(\tilde{\theta} - \theta)$.

Anderson and Darling have shown that

$$\int W^2(t)v(t)dt = \sum_{i=1}^{\infty} \frac{Z_i^2}{i(i+1)},$$

where the Z_i are iid standard normal random variables defined in terms of the Brownian bridge $W(t)$. Boos found that, when estimating the location parameter of a logistic distribution, $B(W)'\Delta^{-1}B(W) = Z_1^2/2$. Hence, $n\tilde{\delta}_A \xrightarrow{d} A_1^2$. Boos felt that it was reasonable to expect that $B(W)'\Delta^{-1}B(W) \approx Z_1^2/2$ for other one-parameter parent distributions. He further conjectured that a similar result would be true for two, three, or more parameter parent distributions so that A_1^2 (or A_2^2, A_3^2, ...) is an approximation to the limiting distribution of $n\tilde{\delta}_A$ when one parameter (or 2, 3, ... parameters) is estimated.

ADR Residuals: Diagnosing with $\tilde{\delta}_i$.

Let us now focus upon the contribution of each observation to δ_A. Recall the simpler expression for the Anderson-Darling distance (Equation 3):

$$\begin{aligned}\delta_A(F_\theta, F_n) &= -1 + \sum_{i=1}^{n} \frac{2i-1}{n^2} |\ln[F_\theta(r_{(i)})]| + \\ &\quad \frac{2(n-i)+1}{n^2} |\ln[1 - F_\theta(r_{(i)})]| .\end{aligned}$$

The closed form representation of δ_A suggests a natural estimator for the contribution of

one observation. Let us define δ_i to be the contribution of the ith order statistic to δ_A:

$$\delta_i = \frac{2i-1}{n^2}|\ln(F_i)| + \frac{2(n-i)+1}{n^2}|\ln(1-F_i)| . \tag{19}$$

Then, the distance δ_A can be written as a sum of the contributions from each observation:

$$\delta_A = -1 + \sum_{i=1}^{n} \delta_i .$$

We see that the sum of the contributions must be greater than or equal to one, because δ_A is nonnegative by definition. Further, the expected value of this sum must exceed one because each term in the definition of δ_i is positive.

In some sense, δ_i is a measure of the goodness of fit due to the $(i)th$ observation. It is reasonable then to ask what δ_i tells us about the fit of the model.

Two pieces of information accompany each δ_i: the magnitude of the weighted lack-of-fit of one residual, and the index i of the $(i)th$ order statistic. Both must be considered in interpreting δ_i. For instance, a large δ_i, and a relatively small or large i (and hence, a large magnitude residual $|y_i - \tilde{y}|$) may indicate that the $(i)th$ observation is an outlier in the classic sense. A small δ_i and middle-valued i (with a residual $y_i - \tilde{y}$ near zero) may indicate that the $(i)th$ observation is a high influence point. Therefore, it is beneficial to look for expressions based on δ_i that provide insight about the outlier or influence potential of the $(i)th$ observation.

We know that the expected value of δ_i is greater than zero from its definition (Equation 19). Therefore, judging individual fit on the magnitude of δ_i alone may be misleading. So, let us examine δ_i with respect to its expected value and variance. Let us consider a more easily interpretable, standardized diagnostic based on δ_i:

$$\delta_i^* = \frac{\delta_i - \mathrm{E}[\delta_i]}{\sqrt{\mathrm{Var}[\delta_i]}} .$$

The Expected Value of δ_i. Let us find the expected value of δ_i:

$$\begin{aligned} \mathrm{E}[\delta_i] &= \mathrm{E}\left[-\frac{2i-1}{n^2}\ln(U_i) - \frac{2(n-i)+1}{n^2}\ln(1-U_i)\right] \\ &= -\frac{2i-1}{n^2}\mathrm{E}[\ln(U_i)] - \frac{2(n-i)+1}{n^2}\mathrm{E}[\ln(1-U_i)] \,. \end{aligned} \tag{20}$$

To evaluate $\mathrm{E}[\delta_i]$ we must determine the expectations of the two logarithms: $\mathrm{E}[\ln(U_i)]$ and $\mathrm{E}[\ln(1-U_i)]$. If $\{Y_i\}_{i=1}^n$ is a simple random sample from F_θ, then $\{F_\theta(Y_i)\}_{i=1}^n$ is a simple random sample from a uniform $U(0,1)$ distribution. Let $U_i = F_\theta(Y_{(i)})$. The statistic U_i is the *ith* order statistic from a $U(0,1)$ because U_i is a function of the *ith* order statistic $Y_{(i)}$, and the transformation F_θ preserves order. The random variable U_i has a known density function Mood et al. (1974);

$$\begin{aligned} f_{U_i}(u) &= \frac{n!}{(i-1)!(n-i)!}[F_U(u)]^{i-1}[1-F_U(u)]^{n-i} f_U(u) \\ &= \frac{n!}{(i-1)!(n-i)!}u^{i-1}(1-u)^{n-i} \\ &= \frac{\Gamma(n+1)}{\Gamma(i)\Gamma(n-i+1)}u^{i-1}(1-u)^{(n-i+1)-1} \,. \end{aligned}$$

The function $f_{U_i}(u)$ is the the probability density function of a Beta random variable: $U_i \sim \mathrm{Beta(i, n-i+1))}$.

To find $\mathrm{E}[\ln(U_i)]$ and $\mathrm{E}[\ln(1-U_i)]$, let us apply an elegant approach provided by Boik (1997). Suppose $U \sim \mathrm{Beta}(\alpha, \beta)$. Then,

$$\int_0^1 \frac{u^{\alpha-1}(1-u)^{\beta-1}}{\mathrm{B}(\alpha,\beta)}\,\mathrm{d}u = 1$$

where

$$\mathrm{B}(\alpha,\beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)} \,.$$

Now, let us take the derivative of both sides with respect to the parameter α. In this case, we can take the derivative under the integral because the necessary regularity conditions are met; namely, α is a continuous over its parameter space.

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}\alpha}\left(\int_0^1 \frac{u^{\alpha-1}(1-u)^{\beta-1}}{\mathrm{B}(\alpha,\beta)}\,\mathrm{d}u\right) &= -\frac{1}{\mathrm{B}(\alpha,\beta)}\frac{\mathrm{d}}{\mathrm{d}\alpha}(\mathrm{B}(\alpha,\beta)) + \frac{1}{\mathrm{B}(\alpha,\beta)}\int_0^1 \frac{\ln(u)u^{\alpha-1}(1-u)^{\beta-1}}{\mathrm{B}(\alpha,\beta)}\,\mathrm{d}u \\ &= -\frac{1}{\mathrm{B}(\alpha,\beta)}\frac{\mathrm{d}}{\mathrm{d}\alpha}(\mathrm{B}(\alpha,\beta)) + \mathrm{E}[\ln(U)] \\ &= 0\,.\end{aligned}$$

It follows that the expected value of $\ln(U)$ is

$$\begin{aligned}\mathrm{E}[\ln(U)] &= \frac{1}{\mathrm{B}(\alpha,\beta)}\frac{\mathrm{d}}{\mathrm{d}\alpha}(\mathrm{B}(\alpha,\beta)) \\ &= \frac{1}{\mathrm{B}(\alpha,\beta)}\frac{\mathrm{d}}{\mathrm{d}\alpha}\left(\frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}\right) \\ &= \frac{1}{\mathrm{B}(\alpha,\beta)}\left[\frac{\Gamma'(\alpha)}{\Gamma(\alpha)}\frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)} - \frac{\Gamma'(\alpha+\beta)}{\Gamma(\alpha+\beta)}\frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}\right] \\ &= \frac{1}{\mathrm{B}(\alpha,\beta)}[\psi(\alpha)\mathrm{B}(\alpha,\beta) - \psi(\alpha+\beta)\mathrm{B}(\alpha,\beta)] \\ &= \psi(\alpha) - \psi(\alpha+\beta)\end{aligned}$$

where $\psi(\cdot)$ denotes the DiGamma function (Abramowitz and Stegun, 1964).

Applying this result, we can express the expected values of the two functions of the order statistic U_i in terms of the DiGamma function by substituting i for α and $n-i+1$ for β:

$$\begin{aligned}\mathrm{E}[\ln(U_i)] &= \psi(i) - \psi(n+1) \\ \mathrm{E}[\ln(1-U_i)] &= \psi(n-i) - \psi(n+1)\,.\end{aligned}$$

Substituting these expected values into Equation 20, we see that the expected value of δ_i is

$$\begin{aligned}\mathrm{E}[\delta_i] &= -\frac{2i-1}{n^2}[\psi(i) - \psi(n+1)] - \frac{2(n-i)+1}{n^2}[\psi(n-i+1) - \psi(n+1)] \\ &= \frac{-1}{n^2}[(2i-1)\psi(i) + (2(n-i)+1)\psi(n-i+1) - 2n\,\psi(n+1)]\,.\end{aligned}$$

The expected value $E[\delta_i]$ depends upon the position i of the *ith* order statistic and the sample size n; it does not depend on the parent distribution F_θ. In this sense, $\mathrm{E}[\delta_i]$ is distribution-free. Routines to compute $\psi(\cdot)$ are available (though, not standard elements of most

software packages) so that computing $\mathrm{E}[\delta_i]$ with this formula requires some additional effort (Press et al., 1992).

The Variance of δ_i. To determine the variance of δ_i, let us recall that

$$\mathrm{Var}(\delta_i) = \mathrm{E}[\delta_i^2] - \mathrm{E}^2[\delta_i]$$

. Therefore, let us find $\mathrm{E}[\delta_i^2]$ and then combine with the previous result for $\mathrm{E}[\delta_i]$. Again, let $U_i = F_\theta(Y_{(i)})$. Then

$$\begin{aligned}
\mathrm{E}[\delta_i^2] &= \mathrm{E}\left[\left(-\frac{2i-1}{n^2}\right)^2 \ln^2(U_i) + 2\cdot\frac{2i-1}{n^2}\cdot\frac{2(n-i)+1}{n^2}\ln(U_i)\ln(1-U_i) + \right. \\
&\quad \left.\left(\frac{2(n-i)+1}{n^2}\right)^2 \ln^2(1-U_i)\right] \\
&= \left(\frac{2i-1}{n^2}\right)^2 \mathrm{E}[\ln^2(U_i)] + 2\cdot\frac{2i-1}{n^2}\cdot\frac{2(n-i)+1}{n^2}\mathrm{E}[\ln(U_i)\ln(1-U_i)] + \\
&\quad \left(\frac{2(n-i)+1}{n^2}\right)^2 \mathrm{E}[\ln^2(1-U_i)] \,.
\end{aligned}$$

Once more, we must calculate the expectations of functions of the order statistic U_i: $\mathrm{E}[\ln^2(U_i)]$, $\mathrm{E}[\ln(U_i)\ln(1-U_i)]$, and $\mathrm{E}[\ln^2(1-U_i)]$. And, once more, we can apply the elegant "derivative" approach. Let $U \sim \mathrm{Beta}(\alpha,\beta)$. Then

$$\begin{aligned}
\frac{\mathrm{d}^2}{\mathrm{d}\alpha^2}\int_0^1 \frac{u^{\alpha-1}(1-u)^{\beta-1}}{\mathrm{B}(\alpha,\beta)}\,\mathrm{d}u &= \psi'(\alpha+\beta) - \psi'(\alpha) + \mathrm{E}[\ln^2(U)] \\
\frac{\mathrm{d}^2}{\mathrm{d}\alpha\mathrm{d}\beta}\int_0^1 \frac{u^{\alpha-1}(1-u)^{\beta-1}}{\mathrm{B}(\alpha,\beta)}\,\mathrm{d}u &= \psi'(\alpha+\beta) + \mathrm{E}[\ln(U)\ln(1-U)] \\
\frac{\mathrm{d}^2}{\mathrm{d}\beta^2}\int_0^1 \frac{u^{\alpha-1}(1-u)^{\beta-1}}{\mathrm{B}(\alpha,\beta)}\,\mathrm{d}u &= \psi'(\alpha+\beta) - \psi'(\beta) + \mathrm{E}[\ln^2(1-U)] \,.
\end{aligned}$$

Because these derivatives also equal 0 as in the previous case, it follows that

$$\begin{aligned}
\mathrm{E}[\ln^2(U)] &= \psi'(\alpha) - \psi'(\alpha+\beta) \\
\mathrm{E}[\ln(U)\ln(1-U)] &= -\psi'(\alpha+\beta) \\
\mathrm{E}[\ln^2(1-U)] &= \psi'(\beta) - \psi'(\alpha+\beta)
\end{aligned}$$

where $\psi'(\cdot)$ denotes the TriGamma function (Abramowitz and Stegun, 1964). Applying these results with $\alpha = i$ and $\beta = n - i + 1$, the expected value of δ_i^2 is

$$\begin{aligned} E[\delta_i^2] &= \left(\frac{2i-1}{n^2}\right)^2 [\psi'(i) - \psi'(n+1)] - 2 \cdot \frac{2i-1}{n^2} \cdot \frac{2(n-i)+1}{n^2} \psi'(n+1) \\ &\quad \left(\frac{2(n-i)+1}{n^2}\right)^2 [\psi'(n-i+1) - \psi'(n+1)] \\ &= \left(\frac{2i-1}{n^2}\right)^2 \psi'(i) - \frac{4}{n^2}\psi'(n+1) + \left(\frac{2(n-i)+1}{n^2}\right)^2 \psi'(n-i+1)\,. \end{aligned}$$

Collecting results and substituting, we find that the variance of δ_i is

$$\begin{aligned} \mathrm{Var}(\delta_i) &= \mathrm{E}[\delta_i^2] - \mathrm{E}^2[\delta_i] \\ &= \left(\frac{2i-1}{n^2}\right)^2 \psi'(i) - \frac{4}{n^2}\psi'(n+1) + \left(\frac{2(n-i)+1}{n^2}\right)^2 \psi'(n-i+1) - \\ &= \left[\frac{-1}{n^2}[(2i-1)\psi(i) + (2(n-i)+1)\psi(n-i+1) - 2n\psi(n+1)]\right]^2 . \end{aligned}$$

Routines to compute $\psi'(\cdot)$ are available (Press et al., 1992). Computing $\mathrm{Var}[\delta_i]$ with this formula also can be accomplished with moderate effort.

An Illustration of ADR Diagnostics.

Chatterjee and Hadi (1988) illustrate the use of several regression diagnostics by modeling data extracted from the records of 30 employees who attended a company's health club on a regular basis. In their illustration, an employee's time in a one-mile run (Y_i) was modeled as a linear combination of the employee's weight ($X1_i$), resting pulse rate ($X2_i$), arm and leg strength($X3_i$), and time in a quarter-mile run ($X4_i$) (Table 9. Chatterjee and Hadi (1988) employee health model is defined in Equation 21). The health club dataset is listed in Table 10.

$$\begin{aligned} Y_i &= \beta_0 + \beta_1 X1_i + \beta_2 X2_i + \beta_3 X3_i + \beta_4 X4_i + \epsilon_i \qquad (21) \\ \epsilon_i &\overset{iid}{\sim} N(0, \sigma^2) \end{aligned}$$

Table 9: Health Club Variables

X1	:	weight in pounds
X2	:	resting pulse rate per minute
X3	:	arm and leg strength (number of pounds lifted)
X4	:	time in 1/4-mile run (seconds)
Y	:	time in 1-mile run (seconds)

Table 10: Health Club Dataset

Person	X1	X2	X3	X4	Y
1	217	67	260	91	481
2	141	52	190	66	292
3	152	58	203	68	338
4	153	56	183	70	357
5	180	66	170	77	396
6	193	71	178	82	429
7	162	65	160	74	345
8	180	80	170	84	469
9	205	77	188	83	425
10	168	74	170	79	358
11	232	65	220	72	393
12	146	68	158	68	346
13	173	51	243	56	279
14	155	64	198	59	311
15	212	66	220	77	401
16	138	70	180	62	267
17	147	54	150	75	404
18	197	76	228	88	442
19	165	59	188	70	368
20	125	58	160	66	295
21	161	52	190	69	391
22	132	62	163	59	264
23	257	64	313	96	487
24	236	72	225	84	481
25	149	57	173	68	374
26	161	57	173	65	309
27	198	59	220	62	367
28	245	70	218	69	469
29	141	63	193	60	252
30	177	53	183	75	338

Table 11 lists the LSR and ADR estimates, with standard errors, for the six parameters $(\beta_0, \ldots, \beta_4, \sigma^2)$ of the health model (Equation 21). Though the estimates from the two methods appear strikingly different, the differences are not significant in light of the standard errors. From these results, I would draw the same conclusions in both cases regarding the quality of the fit of the model, and the significance of each regressor. On first glance at the estimates,

the model fits well ($R^2 > 0.8$); the intercept and $X2$ terms, however, are not significant. For a more in-depth evaluation of model fit, and an introduction to ADR diagnostics, let us proceed with a diagnosis of the fit, following the lead of Chatterjee and Hadi (1988).

Table 11: Health Model Coefficient Estimates with Standard Errors.

	LSR		ADR	
Parameter	estimate	SE	estimate	SE
Intercept	-3.62	56.10	-8.72	55.50
X1	1.27	0.29	1.74	0.64
X2	-0.53	0.86	0.073	1.20
X3	-0.51	0.25	-0.77	0.31
X4	3.90	0.75	3.01	1.41
$\widehat{\sigma}$	28.7	—	30.3	—
R^2	0.853	—	0.827	—

Chatterjee and Hadi (1988) identified three observations (numbered 23, 28 and 30 in Table 10) as problematic. Their evidence is succinctly summarized by a Leverage-Residual (L-R) plot (Figure 32). The two panels in this figure display measures of leverage (the p_{ii}s or the diagonal elements of the projection (hat) matrix) against the normalized residuals for both the LSR (top panel) and ADR (bottom panel) fits to the full dataset. The LSR L-R panel shows that observation number 23 has high leverage but is not an outlier nor an influential point. Number 28 is an influential observation but may not be an outlier nor a high leverage point. Number 30 is an outlier but neither high leveraging nor influential.

The ADR L-R panel suggests that both points 23 and 28 are high-leverage points in the LSR sense, while observations 11, 21, and, possibly, 25 and 30 are outliers. No observation appears to be both high leverage and an outlier in the ADR fit.

To explore further the influence of observations 23, 28, and 30, the health model (Equation. 21) was refit to reduced datasets using the LSR and ADR methods, where each of these suspect observations was discarded in turn. One measure of the influence of these observations is the *Student's t* statistic (Table 12). After comparing the t statistics of the coefficient estimates to '2', a rough threshold of *Student's t* significance, I conclude that the three observations have a similar influence on both the LSR and ADR sets of the estimates; the corresponding LSR and ADR t statistics for the reduced datasets generally move in the same direction relative to the t statistics for the full dataset.

Table 12: *Student's t* statistics for Health model coefficient estimates from the full and reduced datasets (All Obs., and minus Obs. 23, 28 and 30, respectively).

	All Obs.		- Obs. 23		- Obs. 28		- Obs. 30	
Parameter	LSR	ADR	LSR	ADR	LSR	ADR	LSR	ADR
Intercept	0.06	-0.16	-0.50	-0.15	-0.28	-0.38	0.59	-1.12
$X1$	4.42	2.70	4.11	2.06	2.87	1.61	5.14	2.66
$X2$	-0.61	0.06	-0.73	-0.19	-0.78	-0.06	-1.52	0.05
$X3$	-2.05	-2.36	-1.29	-1.60	-1.55	-1.77	-2.74	-1.98
$X4$	5.22	2.24	5.14	2.50	5.70	2.13	6.01	2.30

A closer comparison of outliers, leverage and influence can be made using side-by-side boxplots of LSR and ADR diagnostic scores (Figure 33). The diagnostics presented in the six panels of this figure cover leverage (diagonal elements of the hat matrix), outliers (internally-studentized residuals), global influence (Cook distance), and local influence (Welsch-Kuh, Welsch, and modified Cook distances). The ADR diagnostic scores were computed using the standard formulas (Equations 15, 16, 17 and 18, respectively).

It is worth noting one outcome of ADR: the symmetry of the internally-studentized residuals, as assumed in the model (Equation 21). It is also worth noting that, in the LSR case, observations with significant leverage (23 and 28) also have significant global influence (large Cook distances); which is not the case for ADR. In general, each observation has a larger global influence in ADR; the ADR Cook distances are relatively larger than their LSR counterparts. The comparison of the local influence diagnostics, however, suggest that the local influence of each observation in ADR is reduced relative to the local influence of each observation in LSR. This speaks to the robustness of ADR.

Figure 34 shows the values of three potential ADR-specific diagnostics. The first panel features the $\delta_A(i)'s$, the AD scores from the ADR fits to the n leave-one-out health datasets. The second panel presents the approximate p-values for the n leave-one-out AD scores featured in the first panel. The third panel displays the values of the contribution of the *ith* observation to the AD score (i.e., $\delta_i = \frac{[F_n(y)-F_\theta(y)]^2}{F_\theta(y)[1-F_\theta(y)]}$).

Coefficient estimates do not appear to vary significantly (Table 11) as observations are included or excluded; again, a testament to the robustness of ADR. Nonetheless, there are extreme $\delta_A(i)$ scores with extreme $\delta_A(i)$ p-values. The goodness-of-fit hangs on whether or not these observations are included in the estimation dataset. Therefore, it appears that an extreme

leave-one-out AD score and its p-value flag an outlier or an inlierwith respect to the assumed distribution.

The boxplot of individual contributions δ_i to the AD score is included for completeness. Observations having relatively large contributions would bear investigation. In this case, however, no observation makes such a contribution. As noted earlier, the normalized contribution δ_i^* may be a better measure of a single observation's contribution to δ_A because it accounts for the differences in expected values and standard errors among the raw contributions.

Figure 35 displays the ratios of the ADR diagnostic scores to the LSR scores against the observation number. Three diagnostics are featured: internally-studentized residuals, Cook distance and Welsch-Kuh distance. In all three panels, observations with nearly-equivalent diagnostic scores are located near the dotted horizontal line (where the ratio equals 1). In this case, ratios of the internally-studentized residuals show that the LSR and ADR residuals almost always agree in sign, though they vary in magnitude. With regard to Cook distance, ADR Cook distance is often significantly larger by a factor of 5 or more than LSR Cook distance. Welsch-Kuh ratios (and equivalently, Welsch and modified Cook ratios) do not indicate any agreement between the two methods in measure of local influence .

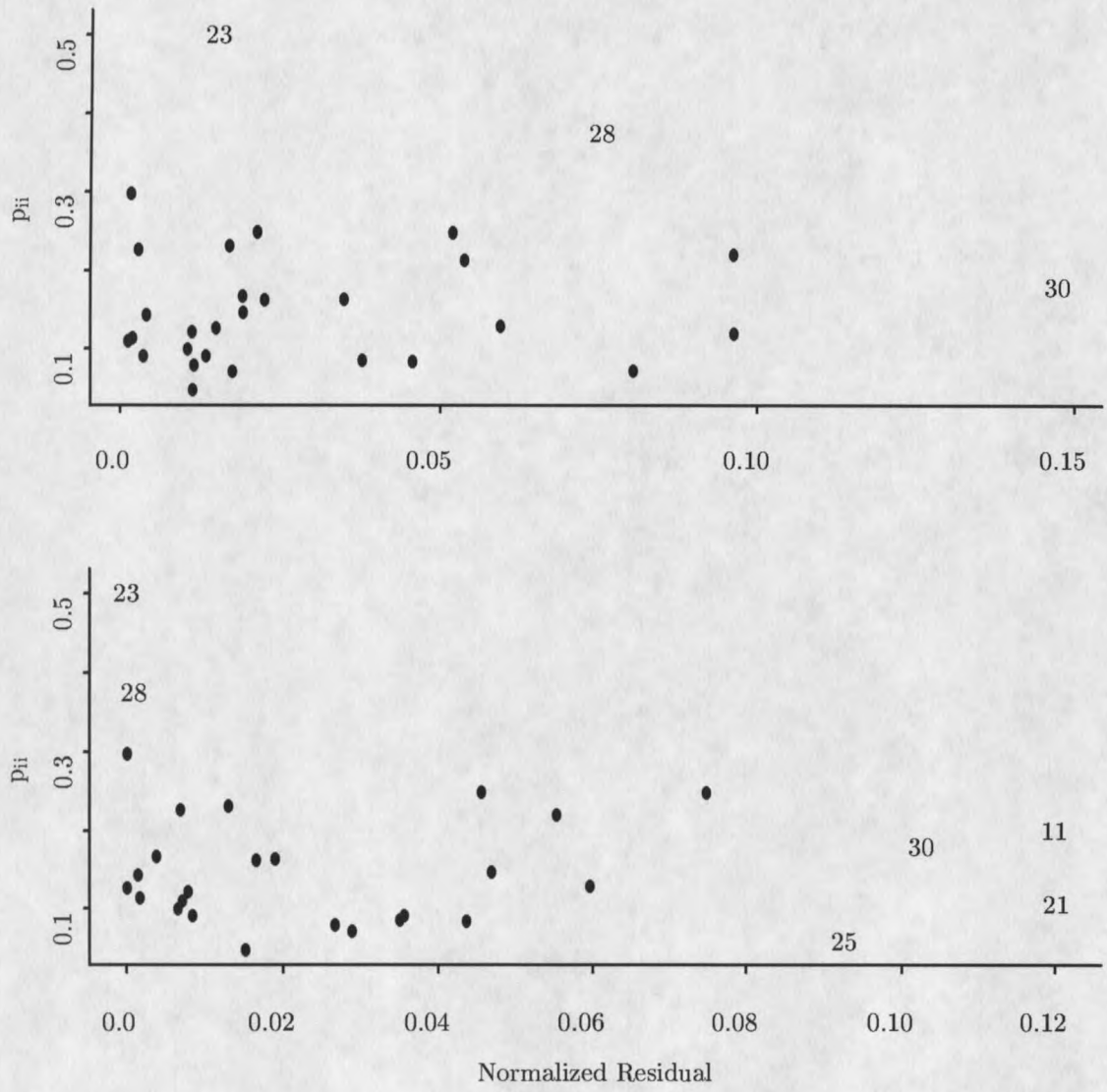


Figure 32: LSR (top) and ADR L-R (bottom) Plots. The LSR panel suggests that observation 23 is a high leverage point, observation 30 is an outlier, and observation 28 is a high leverage point and, possibly, an outlier. The ADR panel suggests that both points 23 and 28 are high leverage points while observations 11, 21, and, possibly, 25 and 30 are outliers.

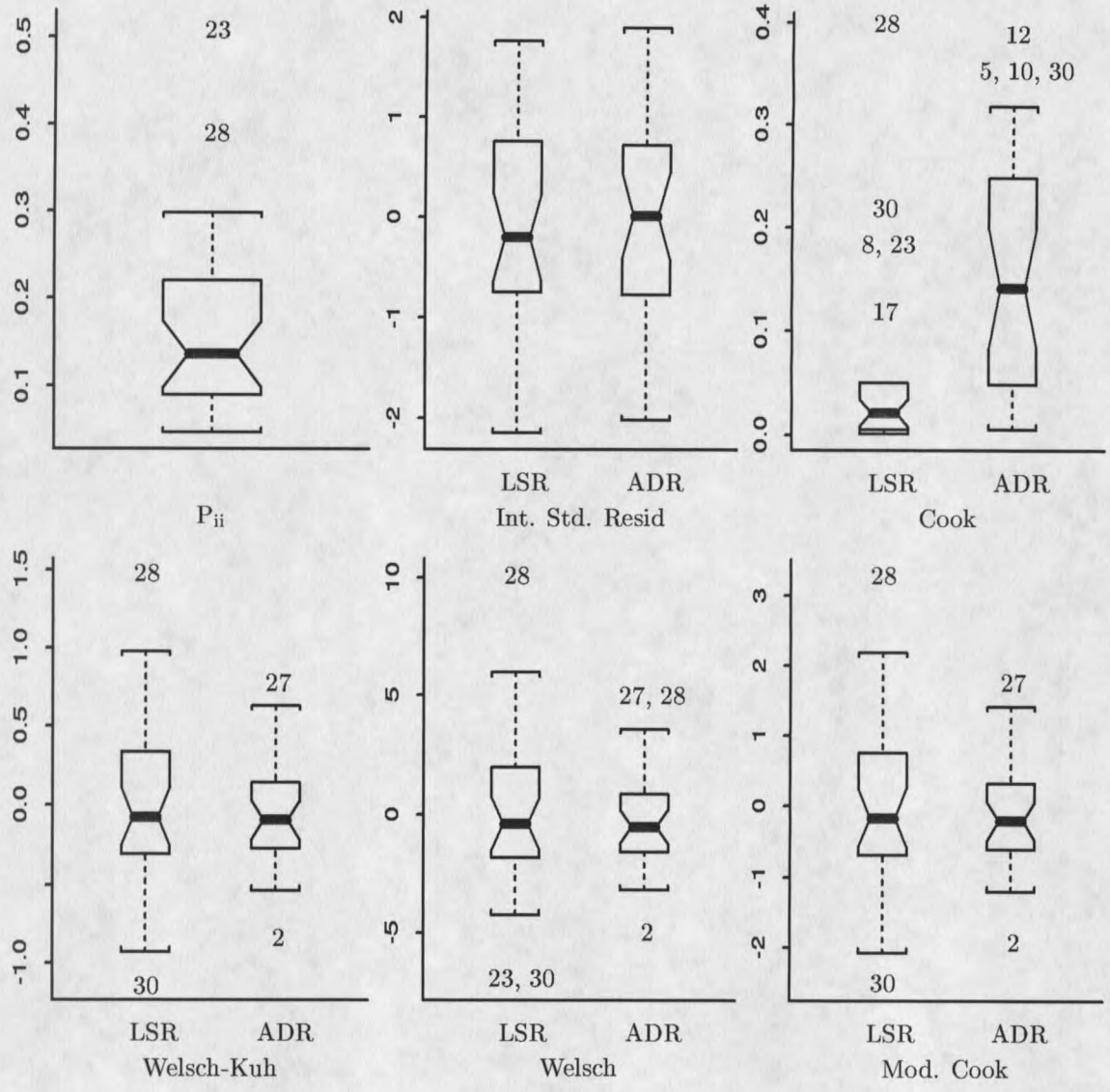


Figure 33: Boxplots of LSR and ADR scores on standard LSR diagnostic measures. Panels from left to right and top to bottom display boxplots of the hat-matrix diagonal values, internally-studentized residuals, Cook distance, Welsch-Kuh distance, Welsch distance, and modified Cook distance.

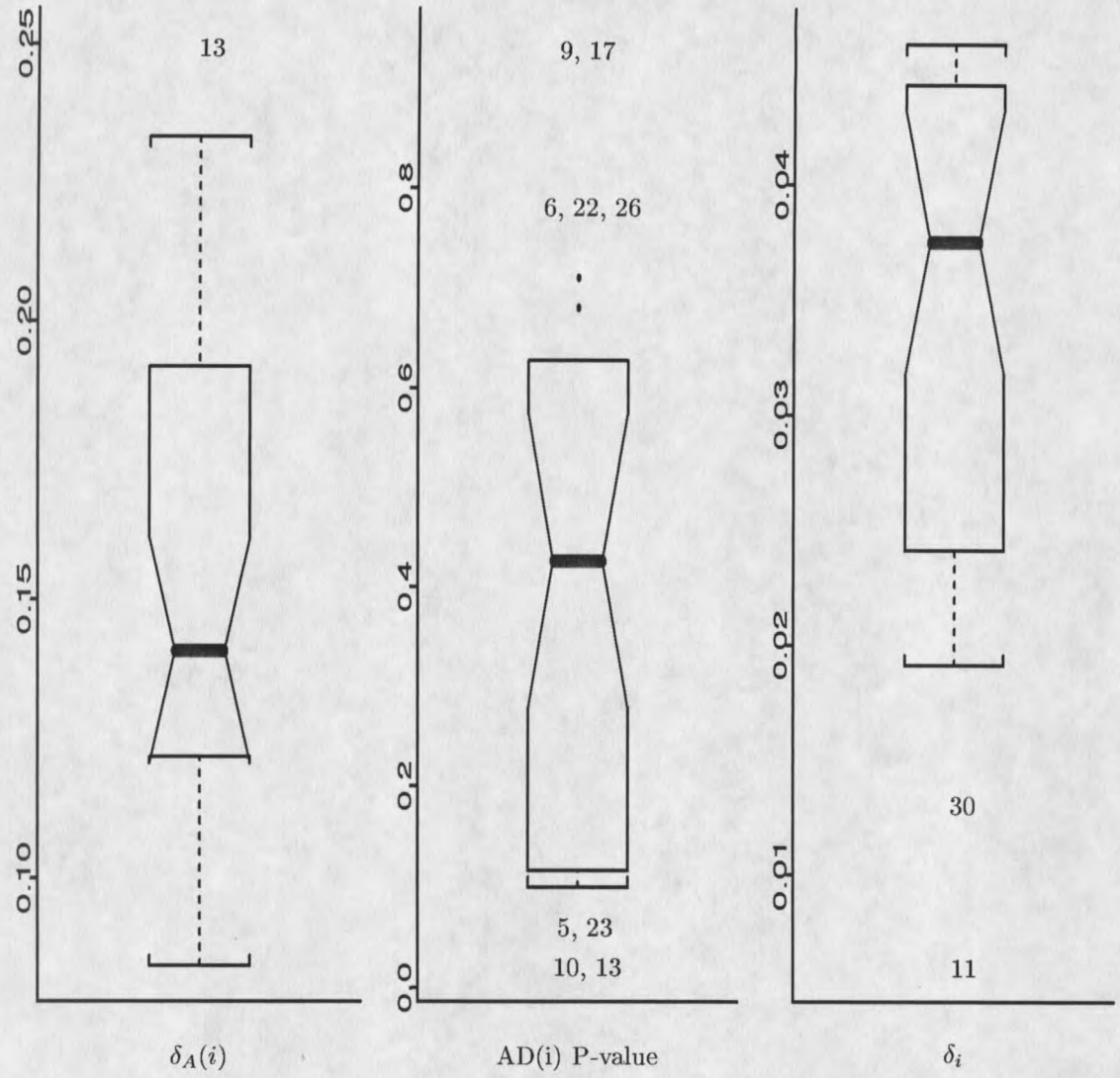


Figure 34: Boxplots of scores on diagnostic measures specific to ADR. The three panels from left to right display boxplots of Anderson-Darling score and p-value for the n leave-one-out datasets, and the contribution of the ith value to the Anderson-Darling score.

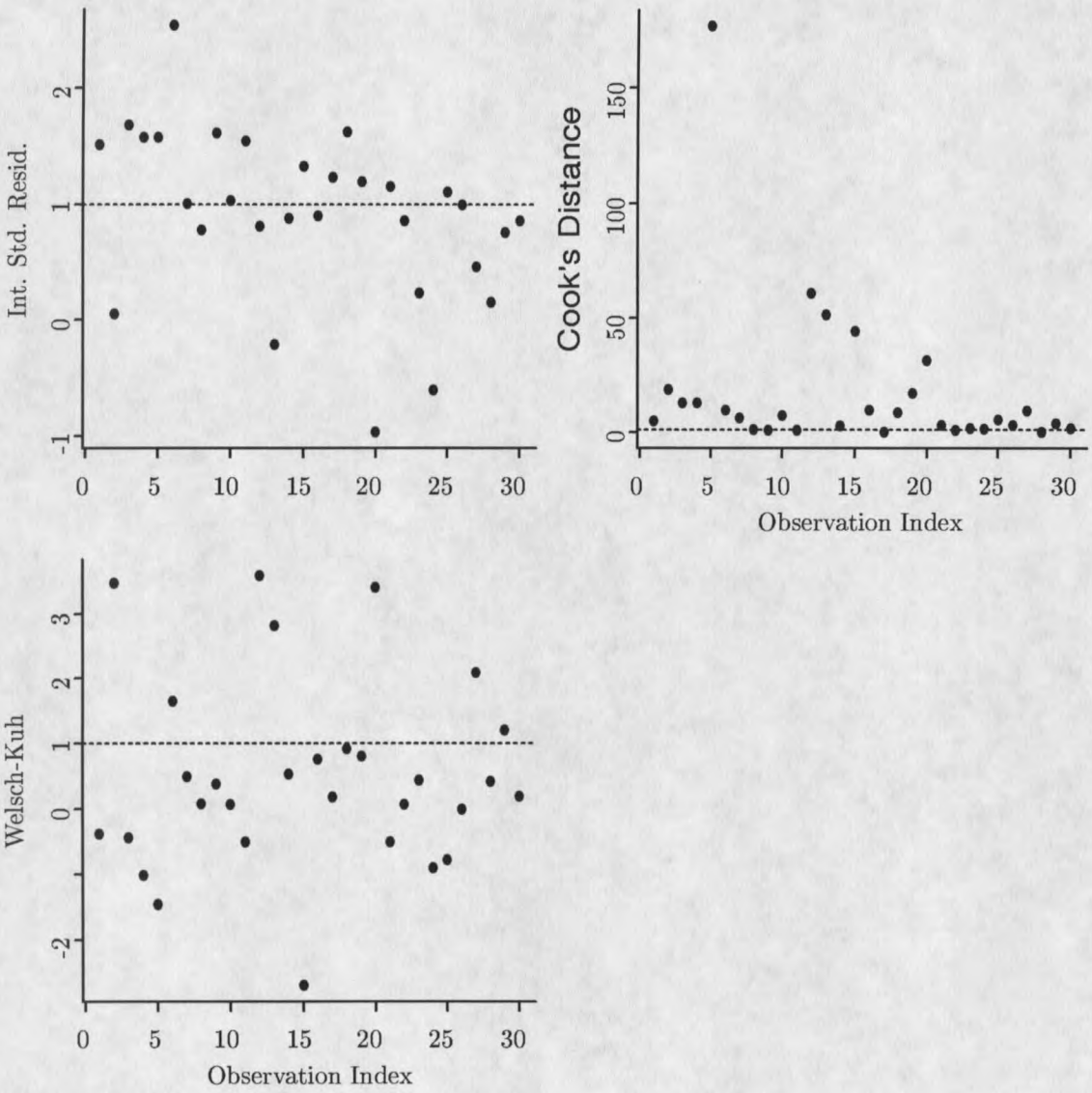


Figure 35: The Ratios of ADR to LSR Diagnostic Scores.

CHAPTER 6

SUMMARY AND FUTURE WORK

This dissertation has presented Anderson-Darling Regression as a useful and broadly-applicable alternative to classic parametric and nonparametric regression methods. In particular, the technique outlined here has been tailored to the ADR model described in Table 1. This model, where the distribution of the random component, and the linkage between the deterministic and random components are identified, is more flexible than the classic least squares model, but more structured than the classic nonparametric model. ADR was tailored to be flexible but powerful due to its use of all the available information in a natural way.

The ADR method is built upon the assumption that the observed random variable follows a location/scale distribution. Estimates are determined by comparing an empirical distibution based on the ADR estimate to the assumed location/scale cumulative distribution via a goodness-of-fit statistic; in this case, the Anderson-Darling distance measure. The actual comparison involves a residual EDF and the location/scale family generator CDF. The fundamental philosophy driving this work has two components:

- if we are presented with a significant amount of information about the random component, and the linkage between the deterministic and random components, then we should use all this information in a direct fashion to make our estimates,
- using all the information in a direct fashion should provide estimates as good as, if not better than, estimates from techniques that do not make use of all the available information directly.

ADR Theory

The work in chapter 2 establishes ADR's pedigree. The desirable properties of ADR estimators were summarized in Table 5. Practicing applied statisticians will find these theoretical niceities reassuring. It is the ease of application, straightforward computation and concrete interpretation, however, that are most appealing. The fact that the evaluation of the complicated integral at the heart of the Anderson-Darling distance measure reduces to a nonlinear regression problem that is easily tackled with common statistical software is most satisfying. Explaining ADR to clients is also easy — ADR chooses the model from a collection of candidate models that is closest, in a probabilistic sense, to the true model.

In this dissertation, I have often referred to the approximate limiting null distribution of $n\delta_A$ in examples featuring small or moderate sample sizes (Boos, 1982). The accuracy of this approximation in these cases is unknown. The accuracy of this approximation must be explored in a variety of small or moderate sample size settings in order to instill confidence in ADR inferences.

The role of ADR in hypothesis testing, and variable selection was not discussed earlier but does bear mentioning. The foundation of statistical inference with regard to hypothesis testing, and variable selection rests on an important but lightly regarded detail — the assumed distribution must fit the data. Because ADR is based on goodness-of-fit, one must confront the quality of fit directly. Using Boos' approximation to determine p-values, a feature that could be a central component of a canned ADR routine, one would know immediately if probabilistic interpretations of point estimates, confidence intervals and hypothesis tests are valid.

ADR may also have a further role in hypothesis testing, and variable selection. Hypothesis testing depends upon a good fit of the residuals to the hypothesized random distribution, and a very good estimate of standard errors. The focus of ADR is optimizing this fit. Through hypothesis testing, ADR can be used as variable selection technique.

Even with all these desirable properties and the excellent theoretical work conducted by Boos and others, there is still much to do. For instance, extending Boos' work with approximate limiting distributions would be worthwhile. Nonetheless, progress in MDE theory will require an individual with significant mathematical statistics skills and a strong interest in MDE. Given the dearth of recent articles, theoretical or applied, this individual apparently does not exist. It remains to be seen if the potential of MDE and ADR, as described herein or previously documented by others, will be fulfilled.

Non-Parametric ADR

To find the NADR estimate of $m(\cdot)$, we proceed in a fashion similar to cross-validated NPR:

- select a dense set of values, say of size m, for the smoothing parameter,
- for each parameter value, find the NPR estimate $\widehat{m}(x_i)$ and, then, an estimate the unobserved error,

$$\begin{aligned} \widehat{\epsilon}(x_i) &= r_i \\ &= f^{-1}(\widehat{m}(x_i), Y_i) \end{aligned} \tag{22}$$

- compute the Anderson-Darling distances between the residual EDFs (one for each parameter value) and the true family generator CDF,
- choose the smoothing parameter that produces the smallest Anderson-Darling distance to determine the best ADR estimate of $m(\cdot)$.

Intuitively, this approach seems reasonable. Nonetheless, simulations show that this approach sometimes works very well and, sometimes, not well at all. Depending on the model and the smoother, the residuals may be significantly biased in location, in scale, or in both. Choosing a smooth that is based on biased residuals using a goodness-of-fit measure that assumes unbiased residuals may cause significant error.

Ideally, NADR would be performed using the true distribution of the residuals; a distribution difficult to determine. My derivations show that NADR should not be directly applied to certain models because the natural residuals do not follow the assumed distribution. My derivations show that this is also the case when applying standard NPR. Using bias-corrected residuals may be a major improvement, but, the corrected residuals still may not have the proper distribution. NADR may identify a smoothing parameter that is too large or too small, leading to over- or under-smoothing as it strives to match the residuals to the assumed CDF. In applying NADR, the bias of the nonparametric estimator must be evaluated carefully.

There is also an effect due to the lack of independence among NPR residuals. It is not clear how this affects ADR, and this effect requires further study.

ADR Diagnostics

The use of diagnostics in ADR, and the use of goodness-of-fit as a general diagnostic to measure the effect of one observation or variable on model fit was tersely addressed in this paper. Defining and computing potentially informative ADR diagnostic measures is easy. Uncovering useful interpretations of these measures, however, is difficult. The main problem is that an EDF is fundamental to ADR, and an EDF is a function of the order statistics. Order statistics are not easy to work with. With regard to one diagnostic based on Anderson-Darling distance, this paper establishes a normalized statistic δ_i^* wherein the role of the order statistics has been simplified. More work with this statistic and ADR diagnostics is needed, however, if ADR is to become a useful method for solving real world problems.

Empirical Anderson-Darling Estimation

Consider the additive model $Y = m_\theta(\mathbf{X}) + \epsilon$. The body of this paper has explored the questions that arise when one estimates the mean response $m_\theta(\mathbf{X})$ by choosing the value $\tilde{\theta} \in \Theta$ that minimizes the Anderson-Darling distance

$$\delta_A(F_{n.\tilde{\theta}}(\mathbf{r}), F_0(\mathbf{r})) = \inf_{\theta \in \Theta} \delta_A(F_{n.\theta}(\mathbf{r}), F_0(\mathbf{r}))$$

where $\mathbf{r} = \mathbf{r}(\widehat{m}_\theta(\mathbf{x}), \mathbf{y})$ is the set of sample residuals, and $F_{n.\theta}(\mathbf{r})$ is the empirical distribution.

Suppose, however, that the distribution F_0 is not known, but that substantial empirical information about F_0 is available or is easily obtained, say in the form of an empirical distribution F_m. Then, one could estimate $m_\theta(\cdot)$ using Anderson-Darling regression by substituting F_m for F_0. One would find the ADR estimate of θ by selecting $\tilde{\theta} \in \Theta$ that minimizes the Anderson-Darling distance between F_m and $F_{n.\theta}$. If one denotes this minimal AD distance between the two EDFs by $\widehat{\delta}_A(F_{n.\tilde{\theta}}, F_0)$, then $\tilde{\theta}$ is that $\theta \in \Theta$ such that

$$\begin{aligned} \widehat{\delta}_A(F_{n.\tilde{\theta}}, F_0) &= \delta_A(F_{n.\tilde{\theta}}, \widehat{F}_0) \\ &= \delta_A(F_{n.\tilde{\theta}}, F_m) \\ &= \inf_{\theta \in \Theta} \delta_A(F_{n.\tilde{\theta}}, F_m) \ . \end{aligned}$$

I call this method "Empirical Anderson-Darling Estimation (EADE)".

EADE Motivation and Support

The use of EADE was illustrated in an example application in Chapter 3. That example featured a microprobe measuring system so complex that an analytical determination of its error distribution may be intractable. Nonetheless, it was very easy using this new sampling device under controlled conditions to generate an instrument-characterizing empirical distribution. This made EADE an appealing option for estimating the mean response in an experiment conducted under a different set of conditions.

One vote for the EADE approach is indirectly cast by Lehmann (1991) who proved that the EDF F_m is the uniformly miminal variance unbiased estimator of F_θ under mild conditions. The fact that the empirical distribution F_m is an unbiased estimator of F_θ follows directly from the definition of F_m as an arithmetic average of indicator functions:

$$\begin{aligned} E[\widehat{F}_\theta(y)] &= E[F_m(y)] \\ &= E[\frac{1}{m}\sum_{k=1}^{m} I_{[y_{(k)},\infty)}(y)] \\ &= \frac{1}{m}\sum_{k=1}^{m} E[I_{(-\infty,y)}(y_k)] \\ &= \frac{1}{m}\sum_{k=1}^{m} \int_{-\infty}^{y} dF_\theta(x) \\ &= \frac{1}{m}\sum_{k=1}^{m} F_\theta(y) \\ &= F_\theta(y)\ . \end{aligned} \tag{23}$$

Lehmann (1991) provides a clever proof showing that F_m is the UMVU estimator of F_θ. Let $\mathbf{Y} = \{Y_i\}_{i=1}^m$ be a set of independent and identically distributed random variables with distribution F_θ. Lehmann notes that the set of order statistics $\mathbf{Y}_{(\cdot)} = \{Y_{(i)}\}_{i=1}^m$ constitutes a complete sufficient statistic. Further, an estimator $\theta(\mathbf{Y})$ is a function of the order statistics $\mathbf{Y}_{(\cdot)}$ if and only if it is symmetric in its m arguments. For distribution families $\mathcal{F}$ for which the order statistics are complete, there can exist at most one symmetric, unbiased estimator of any estimand, and this is the uniformly minimal variance unbiased estimator (UMVUE). A proof that F_m is the UMVU estimator of F_θ follows from these facts.

Building upon Lehmann's result, let us replace F_0 with F_m in the Anderson-Darling regression method. The Anderson-Darling goodness-of-fit statistic becomes a measure of the distance between two empirical distributions, with F_m acting as the gold standard.

$$\begin{aligned} \widehat{\delta}_A(F_{n.\theta}, F_0) &= \delta_A(F_{n.\theta}, \widehat{F}_0) \\ &= \delta_A(F_{n.\theta}, F_m) \\ &= \int_{-\infty}^{\infty} \frac{(F_m - F_n)^2}{F_m(1-F_m)}\, dF_m\ . \end{aligned}$$

I conjecture that this integral expression reduces to a closed form expression for Anderson-Darling distance between two EDFs; this closed form likely will not be equal to the standard closed form (Equation 2), however. Given that it involves two step functions, it is easy to imagine evaluating the integral over one step of F_m. It is also easy to imagine that the necessary bookkeeping may be tedious over the domain of F_m.

Now, $F_m \to F_0$ as $m \to \infty$. I conjecture that $\delta_A(F_{n.\theta}, F_m) \to \delta_A(F_{n.\theta}, F_0)$ as $m \to \infty$, and that the resulting parameter estimates converge to the ADR parameter estimates:

$$\tilde{\theta}_m \to \tilde{\theta} \text{ as } m \to \infty .$$

If my conjectures are true, then $\tilde{\theta}_m$ is close to $\tilde{\theta}$ and, hence, to θ (for large values of m). Proving these conjectures is beyond the scope of this paper. Given the potential of EADE, however, these are good candidates for future work.

EADE Bootstrapping

Evaluating the quality of an ADR estimate is difficult when one has to compare an EDF to an assumed CDF. It is a significant leap in difficulty if the comparison is between two EDFs. Bootstrapping, however, may reduce this difficulty.

When the parameter estimate $\tilde{\theta}$ has been found by minimizing $\delta_A(F_{n.\theta}, F_0)$, the statistical significance of the goodness-of-fit can be ascertained by comparing $\delta_A(F_{n.\tilde{\theta}}, F_0)$ to the appropriate limiting null distribution described by Boos (1982). The statistical significance of the goodness-of-fit when $\tilde{\theta}$ has been found by minimizing $\delta_A(F_{n.\tilde{\theta}}, F_m)$ may be approximated by comparison to Boos limiting null distribution. Nonetheless, the precision of this approximation must be questioned.

Suppose that the parameter estimate $\tilde{\theta}$ has been found by minimizing $\delta_A(F_{n.\theta}, F_m)$. Then, $\delta_A(F_{n.\tilde{\theta}}, F_m)$ is a measure of the goodness-of-fit of the model, and the best available estimate of $\delta_A(F_{n.\tilde{\theta}}, F_0)$. The sampling distribution of this measure, reflecting the use of the estimate F_m for F_0, can be estimated by bootstrapping. In this case, $\tilde{\theta}$, the minimizer of $\delta_A(F_{n.\theta}, F_m)$, would be held fixed. Then the sampling distribution of $\delta_A(F_{n.\tilde{\theta}}, F_m)$ would be estimated by resampling,

recomputing F_m, and recomputing $\delta_A(F_{n.\tilde{\theta}}, F_m)$. Finally, the extreme quantiles, say the 5th and 95th quantiles of the sampling distribution, could be compared to the appropriate Boos limiting null distribution to establish an informal measure of confidence in the goodness-of-fit.

REFERENCES CITED

Abramowitz, M. and Stegun, I. (1964). *Handbook of Mathematical Functions.* National Bureau of Standards Applied Mathematics Series. U.S. Government Printing Office, Washington, D.C.

Anderson, K. (1995). Maximum Likelihood Estimation with Splus. Personal communication. Battelle, Richland, WA.

Boik, R. (1996). Transforming Linear Functions to Normality: Optimal Component Powers. *Communication in Statistics: Simulation*, 25(2):351–367.

Boik, R. (1997). Calculation of E[ln(U)] when $U \sim$ Beta(α,β). Personal communication. Montana State University. Bozeman, MT.

Boos, D. (1981). Minimum Distance Estimators for Location and Goodness of Fit. *Journal of the American Statistical Association: Theory and Methods*, 76:663–670.

Boos, D. (1982). Minimum Anderson-Darling Estimation. *Communication in Statistics: Theory and Methods*, 11(24):2747–2774.

Briggs, D. and Seah, M. (1983). *Practical Surface Analysis by Auger and X-ray Photoelectron Spectroscopy.* John Wiley and Sons, Inc, New York, New York.

Burden, R., Faires, J., and Reynolds, A. (1978). *Numerical Analysis.* Prindle, Weber and Schmidt, Inc, Boston, Massachusetts.

Chambers, J. and Hastie, T. (1992). *Statistical Models in S.* Wadsworth and Brooks/Cole, Inc., Pacific Grove, California. 608 pages.

Chatterjee, S. and Hadi, A. (1988). *Sensitivity Analysis In Linear Regression.* John Wiley and Sons, Inc., New York, New York. 315 pages.

Cook, R. and Weisberg, S. (1986). *Residuals and Influence In Regression.* Chapman and Hall, New York, New York. 230 pages.

Cox, D. and Hinkley, D. (1974). *Theoretical Statistics.* Chapman and Hall, Ltd., London, Great Britain.

D'Agostino, R. and Stephens, M. (1986). *Goodness-of-Fit Techniques.* Statistics: textbooks and monographs. Marcel Dekker, Inc., New York, New York.

Davis, L., MacDonald, N., Palmberg, P., Riach, G., and Weber, R. (1976). *Handbook of Auger Electron Spectroscopy.* Physical Electronics Industries, Inc., Eden Prairie, Minnesota.

Draper, N. and Smith, H. (1981). *Applied Regression Analysis.* John Wiley and Sons, Inc., New York, New York.

Efron, B. (1982). The Jackknife, the Bootstrap and Other Resampling Plans. *Society for Industrial and Applied Mathematics, CBMS-NSF Regional Conference Series in Applied Mathematics*, 1.

Fuchs, E., Oppolzer, H., and Rehme, H. (1990). *Particle Beam Microanalysis; Fundamentals, Methods and Applications.* VCH Verlagsgesellschaft mbH, New York, New York.

Hahn, G. and Shapiro, S. (1967). *Statistical Models in Engineering.* John Wiley and Sons, Inc., New York, New York.

Härdle, W. (1990). *Applied Nonparametric Regression.* Cambridge University Press, New York, New York.

Hocking, R. and Pendleton, O. (1983). Goodness of Fit and Parameter Estimation. *Communications in Statistics: Theory and Methods*, 12(5):497–527.

Kotz, S. and Johnson, N. (1958). *Encyclopedia of Statistical Sciences.* John Wiley and Sons, Inc., New York, New York.

Lehmann, E. (1991). *Theory of Point Estimation.* Wadsworth, Inc., Pacific Grove, California.

McCullagh, P. and Nelder, J. (1989). *Generalized Linear Models, 2nd Edition.* Chapman and Hall, Inc., New York, New York.

Millar, P. (1981). Unknown. *Zeit. WahrsheinlichKeitsth verw. Geb.*, 55:73–89.

Mood, A., Graybill, F., and Boes, D. (1974). *Introduction to the Theory of Statistics.* McGraw-Hill, Inc., New York, New York. 564 pages.

Parr, W. (1981). Minimum Distance Estimation: a Bibliography. *Communication in Statistics: Theory and Methods*, A10(12):1205–1224.

Parr, W. and DeWet, T. (1981). On Minimum Cramer-Von Mises-Norm Parameter Estimation. *Communication in Statistics: Theory and Methods*, A10(12):1149–1166.

Parr, W. and Schucany, W. (1980). Minimum Distance and Robust Estimation. *Journal of the American Statistical Association*, 75:616–624.

Pollard, D. (1980). The Minimum Distance Method of Testing. *Metrika*, 27:43–70.

Press, W., Teukolsky, S., Vetterling, W., and Flannery, B. (1992). *Numerical Recipes in C.* Cambridge University Press, New York, New York.

Rao, P., Schuster, E., and Littell, R. (1975). Unknown. *Annals of Statistics*, 3:862–873.

Royden, H. (1988). *Real Analysis.* Macmillan Publishing Company, New York, New York. 444 pages.

StatSci Division (1993). S-PLUS, version 3.2. (software package). Mathsoft, Inc. Seattle, Washington.

Swokowski, E. (1988). *Calculus With Analytic Geometry.* Mathematics. Prindle,Weber and Schmidt - Kent Co., Boston, Massachusetts.

Tukey, J. (1958). Bias and Confidence in Not Quite Large Samples. *Annals of Mathematical Statistics*, 29:614.

Wolfowitz, J. (1957). The Minimum Distance Method. *Annals of Mathematical Statistics*, 28:75–88.

APPENDICES

APPENDIX A

A Closed-Form Expression For $\delta_A(F_n, F_\theta)$

That the integral expression of the Anderson-Darling discrepancy reduces to a sum of log-transformed valuations of the underlying cumulative distribution function is almost unbelievable to any reasoning being. The equivalence is often quoted in the relevant literature, for example, Boos (1981). None that I have reviewed, however, provide a derivation of this almost miraculous relationship. Counting myself among the skeptical beings in this universe, I could not accept the often-quoted equivalence as fact. This appendix presents the details of my derivation of this equivalence. I have found that deriving this result, though not difficult, is long, tedious and involved. It is clear now why the derivation of this almost unbelievable result is not provided.

The Anderson-Darling measure of discrepancy is

$$\delta_A(F_n, F_\theta) = \int_{-\infty}^{\infty} \frac{[F_n(y) - F_\theta(y)]^2}{F_\theta(y)[1 - F_\theta(y)]}\, dF_\theta(y) \,. \tag{1}$$

We can simplify this integral and express δ_A in a compact closed form:

$$\delta_A(F_n, F_\theta) = -1 + \sum_{i=1}^{n} \frac{2i-1}{n^2} \left[|\ln[F_\theta(y_{(i)})]| + |\ln[1 - F_\theta(y_{(n+1-i)})]| \right] \,. \tag{2}$$

To see the simplifying details, let $F_n = F_n(y)$, $F_\theta = F_\theta(y)$ and $\{y_{(i)}\}_{i=1}^n$ be the set of ordered observations. Also, recall the definition of the empirical distribution function:

$$\begin{aligned} F_n(y) &= \frac{1}{n} \sum_{i=1}^{n} I_{(-\infty, y]}(y_{(i)}) \\ &= \frac{1}{n} \sum_{i=1}^{n} I_{[y_{(i)}, \infty)}(y) \,. \end{aligned}$$

The function $I_{[y_{(i)},\infty)}(y)$ is the indicator function:

$$I_{[y_{(i)},\infty)}(y) = \begin{cases} 0 & \text{if } y \in (-\infty, y_{(i)}) \\ 1 & \text{if } y \in [y_{(i)}, \infty) \end{cases} \,.$$

Integral Expansion. I begin by expanding the right side of Equation 1 by partitioning the domain of the integral and decomposing the integral into a sum of integrals over the

partitioned domain:

$$\begin{aligned}
\delta_A(F_n, F_\theta) &= \int_{-\infty}^{\infty} \frac{(F_n - F_\theta)^2}{F_\theta(1-F_\theta)}\, dF_\theta \\
&= \int_{-\infty}^{y_{(1)}} \frac{(F_n - F_\theta)^2}{F_\theta(1-F_\theta)}\, dF_\theta + \qquad (3) \\
&\quad \sum_{i=1}^{n-1} \int_{y_{(i)}}^{y_{(i+1)}} \frac{(F_n - F_\theta)^2}{F_\theta(1-F_\theta)}\, dF_\theta + \qquad (4) \\
&\quad \int_{y_{(n)}}^{\infty} \frac{(F_n - F_\theta)^2}{F_\theta(1-F_\theta)}\, dF_\theta\ . \qquad (5)
\end{aligned}$$

Simplifying the First Term. The first integral (Term 3) of this expression can be simplified because $F_n(y) = 0$ for $y \in (-\infty, y_{(1)})$: :

$$\begin{aligned}
\int_{-\infty}^{y_{(1)}} \frac{(F_n - F_\theta)^2}{F_\theta(1-F_\theta)}\, dF_\theta &= \int_{-\infty}^{y_{(1)}} \frac{F_\theta^2}{F_\theta(1-F_\theta)}\, dF_\theta \\
&= \int_{-\infty}^{y_{(1)}} \frac{F_\theta}{1-F_\theta}\, dF_\theta \\
&= \int_{-\infty}^{y_{(1)}} \frac{1}{1-F_\theta} - 1\ dF_\theta \\
&= |\ln[1 - F_\theta(y_{(1)})]| - F_\theta(y_{(1)})\ .
\end{aligned}$$

Thus, the first term of the original expansion (Term 3) reduces to a simple closed form:

$$\int_{-\infty}^{y_{(1)}} \frac{(F_n - F_\theta)^2}{F_\theta(1-F_\theta)}\, dF_\theta = |\ln[1 - F_\theta(y_{(1)})]| - F_\theta(y_{(1)})\ .$$

Simplifying the Second Term. For a summand in the second term (Term 4),

$$\begin{aligned}
\int_{y_{(i)}}^{y_{(i+1)}} \frac{(F_n - F_\theta)^2}{F_\theta(1-F_\theta)}\, dF_\theta &= \int_{y_{(i)}}^{y_{(i+1)}} \frac{F_\theta^2 - 2F_\theta F_n + F_n^2}{F_\theta(1-F_\theta)}\, dF_\theta \\
&= \int_{y_{(i)}}^{y_{(i+1)}} \frac{F_\theta^2}{F_\theta(1-F_\theta)}\, dF_\theta \qquad (6) \\
&\quad -2\int_{y_{(i)}}^{y_{(i+1)}} \frac{F_\theta F_n}{F_\theta(1-F_\theta)}\, dF_\theta \qquad (7) \\
&\quad +\int_{y_{(i)}}^{y_{(i+1)}} \frac{F_n^2}{F_\theta(1-F_\theta)}\, dF_\theta\ . \qquad (8)
\end{aligned}$$

The first integral (Term 6) in this sum can be expressed in a simple closed form (as above):

$$\begin{aligned}\int_{y_{(i)}}^{y_{(i+1)}} \frac{F_\theta^2}{F_\theta(1-F_\theta)}\,dF_\theta &= |\ln(1-F_\theta)| - F_\theta \Big|_{y_{(i)}}^{y_{(i+1)}} \\ &= |\ln[1-F_\theta(y_{(i+1)})]| - F_\theta(y_{(i+1)}) - \\ &\quad |\ln[1-F_\theta(y_{(i)})]| + F_\theta(y_{(i)})\,.\end{aligned}$$

The second component (Term 7) does not equal zero as above, however. On $[y_{(1)}, y_{(i+1)})$, $F_n(y) = i/n$ for $y \in [y_i, y_{i+1}), i = 1, \ldots, n-1$. Hence,

$$\begin{aligned}\int_{y_{(i)}}^{y_{(i+1)}} \frac{-2F_\theta F_n}{F_\theta(1-F_\theta)}\,dF_\theta &= \int_{y_{(i)}}^{y_{(i+1)}} \frac{-2F_n}{1-F_\theta}\,dF_\theta \\ &= -2\int_{y_{(i)}}^{y_{(i+1)}} \frac{i}{n}\frac{1}{1-F_\theta}\,dF_\theta \\ &= -\frac{2i}{n}|\ln(1-F_\theta)| \Big|_{y_{(i)}}^{y_{(i+1)}} \\ &= \frac{2i}{n}|\ln(1-F_\theta(y_{(i)})| - \frac{2i}{n}|\ln(1-F_\theta(y_{(i+1)})|\,.\end{aligned}$$

To reduce the last component (Term 8), let us complete the square in the denominator:

$$\begin{aligned}\int_{y_{(i)}}^{y_{(i+1)}} \frac{F_n^2}{F_\theta(1-F_\theta)}\,dF_\theta &= \int_{y_{(i)}}^{y_{(i+1)}} \left(\frac{i}{n}\right)^2 \frac{1}{F_\theta(1-F_\theta)}\,dF_\theta \\ &= \left(\frac{i}{n}\right)^2 \int_{y_{(i)}}^{y_{(i+1)}} \frac{1}{(\frac{1}{2})^2 - (F_\theta - \frac{1}{2})^2}\,dF_\theta\,.\end{aligned}$$

The resulting integral has a closed form (Swokowski, 1988):

$$\int \frac{1}{a^2-u^2}\,du = \frac{1}{2a}\ln\left|\frac{u+a}{u-a}\right|\,.$$

Let $u = F_\theta - \frac{1}{2}$ and $a = \frac{1}{2}$. Then, $du = dF_\theta$ and

$$\begin{aligned}\int \frac{1}{(\frac{1}{2})^2 - (F_\theta - \frac{1}{2})^2}\,dF_\theta &= \frac{1}{2(\frac{1}{2})}\ln\left|\frac{F_\theta - \frac{1}{2} + \frac{1}{2}}{F_\theta - \frac{1}{2} - \frac{1}{2}}\right| \\ &= \ln\left|\frac{F_\theta}{F_\theta - 1}\right| \\ &= \ln(F_\theta) + |\ln(1-F_\theta)|\,.\end{aligned}$$

Therefore, the last component (Term 8) reduces to a closed form:

$$\begin{aligned}\left(\frac{i}{n}\right)^2 \int_{y_{(i)}}^{y_{(i+1)}} \frac{1}{(\frac{1}{2})^2 - (F_\theta - \frac{1}{2})^2}\, dF_\theta &= \left(\frac{i}{n}\right)^2 \left[\ln(F_\theta) + |\ln(1-F_\theta)|\right] \Big|_{y_{(i)}}^{y_{(i+1)}} \\ &= \left(\frac{i}{n}\right)^2 \left[\ln[F_\theta(y_{(i+1)})] + |\ln[1-F_\theta(y_{(i+1)})]|\right] - \\ &\quad \left(\frac{i}{n}\right)^2 \left[\ln[F_\theta(y_{(i)})] + |\ln[1-F_\theta(y_{(i)})]|\right] .\end{aligned}$$

Collecting results, I can write a summand in the second term (Term 4) as a closed form:

$$\begin{aligned}\int_{y_{(i)}}^{y_{(i+1)}} \frac{(F_n - F_\theta)^2}{F_\theta(1-F_\theta)}\, dF_\theta &= -F_\theta(y_{(i+1)}) + |\ln[1-F_\theta(y_{(i+1)})]| + \\ &\quad F_\theta(y_{(i)}) - |\ln[1-F_\theta(y_{(i)})]| - \\ &\quad \frac{2i}{n}|\ln[1-F_\theta(y_{(i+1)})]| + \frac{2i}{n}|\ln[1-F_\theta(y_{(i)})]| + \\ &\quad \left(\frac{i}{n}\right)^2 \left[\ln[F_\theta(y_{(i+1)})] + |\ln[1-F_\theta(y_{(i+1)})]|\right] - \\ &\quad \left(\frac{i}{n}\right)^2 \left[\ln[F_\theta(y_{(i)})] + |\ln[1-F_\theta(y_{(i)})]|\right] .\end{aligned}$$

Simplifying the Third Term. Finally, we can evaluate the last term (Term 5) in the expansion of $\delta_A(F_n, F_\theta)$. This term reduces quickly to a simple closed form because $F_n(y) = 1$ for $y \in [y_n, \infty)$:

$$\begin{aligned}\int_{y_{(n)}}^{\infty} \frac{(F_n - F_\theta)^2}{F_\theta(1-F_\theta)}\, dF_\theta &= \int_{y_{(n)}}^{\infty} \frac{(1-F_\theta)^2}{F_\theta(1-F_\theta)}\, dF_\theta \\ &= \int_{y_{(n)}}^{\infty} \frac{1-F_\theta}{F_\theta}\, dF_\theta \\ &= \int_{y_{(n)}}^{\infty} \frac{1}{F_\theta} - 1\ dF_\theta \\ &= \ln(F_\theta) - F_\theta \Big|_{y_{(n)}}^{\infty} \\ &= |\ln[F_\theta(y_{(n)})]| + F_\theta(y_{(n)}) - 1 .\end{aligned}$$

Collecting the Results. Collecting the evaluations of the three terms (Terms 3, 4 and 5), we find a closed form expression for Equation 1:

$$\begin{aligned}
\delta_A(F, F_n) \;=\; & -F_\theta(y_{(1)}) + |\ln[1 - F_\theta(y_{(1)})]| + \\
& \Big[\sum_{i=1}^{n-1} -F_\theta(y_{(i+1)}) + |\ln[1 - F_\theta(y_{(i+1)})]| + \\
& F_\theta(y_{(i)}) - |\ln[1 - F_\theta(y_{(i)})]| - \\
& \frac{2i}{n}|\ln[1 - F_\theta(y_{(i+1)})]| + \frac{2i}{n}|\ln[1 - F_\theta(y_{(i)})]| + \\
& \left(\frac{i}{n}\right)^2 [\ln[F_\theta(y_{(i+1)})] + |\ln[1 - F_\theta(y_{(i+1)})]] - \\
& \left(\frac{i}{n}\right)^2 [\ln[F_\theta(y_{(i)})] + |\ln[1 - F_\theta(y_{(i)})]|] \Big] + \\
& |\ln[F_\theta(y_{(n)})]|1 + F_\theta(y_{(n)}) - 1 \,. \qquad (9)
\end{aligned}$$

Simplifying the Closed Form Expression. To simplify this expression (Equation 9), let us first simplify the notation. Let $F_i = F_\theta(y_{(i)})$. With this simplification, however, $F_n = F_\theta(y_{(n)})$ and F_n no longer denotes the empirical distribution function. This should not be a problem at this juncture. Rewriting Equation 9 in the simplified notation,

$$\begin{aligned}
\delta_A \;=\; & |\ln(1 - F_1)| - F_1 + \\
& \Big[\sum_{i=1}^{n-1} F_i - |\ln(1 - F_i)| - F_{i+1} + |\ln(1 - F_{i+1})| - \\
& \frac{2i}{n}|\ln(1 - F_{i+1})| + \frac{2i}{n}|\ln(1 - F_i)| - \\
& \frac{i^2}{n^2}|\ln(F_{i+1})| + \frac{i^2}{n^2}|\ln(1 - F_{i+1})| + \\
& \frac{i^2}{n^2}|\ln(F_i)| - \frac{i^2}{n^2}|\ln(1 - F_i)| \Big] + \\
& |\ln(F_n)| + F_n - 1 \,.
\end{aligned}$$

We proceed by grouping like factors:

$$\begin{aligned}
\delta_A &= -1 - F_1 + \sum_{i=1}^{n-1}(F_i - F_{i+1}) + F_n + && (10)\\
&\quad \sum_{i=1}^{n-1}\left[-\frac{i^2}{n^2}|\ln(F_{i+1})| + \frac{i^2}{n^2}|\ln(F_i)|\right] + |\ln(F_n)| + && (11)\\
&\quad |\ln(1-F_1)| + \sum_{i=1}^{n-1}\Big[\ (-1 + \frac{2i}{n} - \frac{i^2}{n^2})|\ln(1-F_i)| + \\
&\quad (1 - \frac{2i}{n} + \frac{i^2}{n^2})|\ln(1-F_{i+1})|\Big] \ . && (12)
\end{aligned}$$

Addressing this sum one group at a time, we see first that the first term (Term 10) with factors F_i is a telescoping series and sums to zero. The second term (Term 11) reduces to a simple sum following more expanding and grouping like factors:

$$\begin{aligned}
&\sum_{i=1}^{n-1}\left[-\frac{i^2}{n^2}|\ln(F_{i+1})| + \frac{i^2}{n^2}|\ln(F_i)|\right] + |\ln(F_n)| \\
&\quad = \sum_{i=1}^{n-1} -\frac{i^2}{n^2}|\ln(F_{i+1})| + \sum_{i=1}^{n-1}\frac{i^2}{n^2}|\ln(F_i)| + |\ln(F_n)| \\
&\quad = \sum_{i=2}^{n-1} -\frac{(i-1)^2}{n^2}|\ln(F_i)| - \frac{(n-1)^2}{n^2}|\ln(F_n)| + \\
&\quad \frac{1}{n^2}|\ln(F_1)| + \sum_{i=2}^{n-1}\frac{i^2}{n^2}|\ln(F_i)| + |\ln(F_n)| \\
&\quad = \frac{2\cdot 1 - 1}{n^2}|\ln(F_1)| + \sum_{i=2}^{n-1}\frac{i^2-(i-1)^2}{n^2}|\ln(F_i)| + \frac{n^2-(n-1)^2}{n^2}|\ln(F_n)| \\
&\quad = \sum_{i=1}^{n}\frac{2i-1}{n^2}|\ln(F_i)| \ .
\end{aligned}$$

A summand featuring factors like $|\ln(1-F_i)|$ (Term 12) also reduces to a simple form:

$$\begin{aligned}
&(-1 + \frac{2i}{n} - \frac{i^2}{n^2})|\ln(1-F_i)| + (1 - \frac{2(i-1)}{n} + \frac{(i-1)^2}{n^2})|\ln(1-F_{(i-1)+1})| \\
&\quad = \frac{n^2 - 2ni + 2n + i^2 - 2i + 1 - n^2 + 2ni - i^2}{n^2}|\ln(1-F_i)| \\
&\quad = \frac{2(n-i)+1}{n^2}|\ln(1-F_i)| \ .
\end{aligned}$$

The sum of these summands can now be collapsed to a nice form;

$$\sum_{i=1}^{n-1}(-1+\frac{2i}{n}-\frac{i^2}{n^2})|\ln(1-F_i)|+(1-\frac{2i}{n}+\frac{i^2}{n^2})|\ln(1-F_{i+1})|$$
$$= \sum_{i=1}^{n}\frac{2(n-i)+1}{n^2}|\ln(1-F_i)|$$

Combining All the Simplifications. Collecting these simplifications, the Anderson-Darling measure of discrepancy can be written in closed form. Returning to the original definitions of F_n as the empirical distribution and F_θ as the cumulative distribution, we find that

$$\delta_A(F_n,F_\theta) = -1+\sum_{i=1}^{n}\frac{2i-1}{n^2}|\ln(F_i)|+\frac{2(n-i)+1}{n^2}|\ln(1-F_i)|\,. \tag{13}$$

We can rewrite this equation (13) to discover an alternate expression for δ_A that is commonly found in the literature. First,

$$\delta_A(F_n,F_\theta) = -1+\sum_{i=1}^{n}\frac{2i-1}{n^2}|\ln(F_i)|+\sum_{i=1}^{n}\frac{2(n-i)+1}{n^2}|\ln(1-F_i)|\,.$$

In the second sum, let us modify the subscripts by setting $i=n+1-j$. Then,

$$\sum_{i=1}^{n}\frac{2(n-i)+1}{n^2}|\ln(1-F_i)| = \sum_{j=1}^{n}\frac{2j-1}{n^2}|\ln(1-F_{n+1-j})|\,.$$

Hence, we find that the Anderson-Darling measure of discrepancy (1) does have the simple closed form noted earlier (2):

$$\delta_A(F_n,F_\theta) = -1+\sum_{i=1}^{n}\frac{2i-1}{n^2}\left[\,|\ln(F_i)|+|\ln(1-F_{n+1-i})|\right]\,.$$

APPENDIX B

A Stochastic Model of an Auger Spectrum

This appendix provides a statistical foundation for the standard Auger methods. To begin, a stochastic model of an Auger spectrum is developed. The foundation of this model is the particle physics underlying Auger spectroscopy and the standard methods to monitor this particle phenomenon outlined in Chapter 3 and described in more detail by Fuchs et al. (1990). Then, estimators based on this model that complement the standard Auger methods are derived. Together, these provide a statistical approach to identify chemical elements on a surface and to estimate their relative abundances.

Implicit Assumptions in the Standard Methods.

With the standard Auger estimate of relative abundance described in Chapter 3, the Auger spectroscopist assumes that the area of the Auger dimple on the electron distribution is directly proportional to the abundance of the associated element on the surface. One further assumes that the peak-to-peak distance is directly proportional to the area of the dimple. Hence, the proportion of peak-to-peak distance corrected for sensitivity is equivalent to the proportion of the element on the surface. Also, one assumes that the proportionality constants are equal across all elements. In essence, we are assuming a linear relationship between peak-to-peak distance and absolute abundance.

Let us trace the effect of these assumptions from the actual elemental abundance at the probed surface to the standard abundance estimate, using a deterministic line of reasoning. Let A_i be the abundance of element i on the surface, N_i be the number of Auger electrons emitted for element i, and C_i be the count of Auger electrons in the neighborhood of the Auger energy ϵ_i that form the *ith* Auger dimple. Let B_i be the area of this dimple and D_i the related peak-to-peak distance. Assuming that the next term in the chain is directly related to the previous term,

$$\begin{aligned} A_i &= h_i N_i \\ N_i &= k_i C_i \\ C_i &= l_i B_i \\ B_i &= q_i D_i \,, \end{aligned}$$

where h_i, k_i, l_i, and q_i are proportionality factors. Ideally, if R_i is the relative elemental

abundance of element i, then

$$\begin{aligned} R_i &= \frac{A_i}{\sum_{j=1}^{p} A_j} \\ &= \frac{h_i N_i}{\sum_{j=1}^{p} h_j N_j} \\ &= \frac{h_i k_i C_i}{\sum_{j=1}^{p} h_j k_j C_j} \\ &= \frac{h_i k_i l_i B_i}{\sum_{j=1}^{p} h_j k_j l_j B_j} \\ &= \frac{h_i k_i l_i q_i D_i}{\sum_{j=1}^{p} h_j k_j l_j q_j D_j} . \end{aligned}$$

Let $s_i = h_i k_i l_i q_j$. Upon substitution, this formula is similar to the standard formula for the calculation of an element's relative abundance (Equation 5):

$$R_i = \frac{s_i D_i}{\sum_{j=1}^{p} s_j D_j} .$$

The factor s_i and the empirically-determined Auger sensitivity factor for element i appear to be the same. The Auger sensitivity factor s_i apparently summarizes the accumulated effect of the series of linearity assumptions and proportionality constants.

The standard formula produces an estimate of relative abundance. The literature, however, does not provide a standard formula to estimate the uncertainty in the standard relative abundance estimate nor the uncertainty in any of the other estimates. I will now show that it is possible to identify a surface element with more confidence, to provide a better estimate of relative elemental abundance, and to estimate the standard errors by modeling the process with a stochastic model.

A Stochastic Model For Auger Spectra.

Let us begin at the surface and work to an estimate of relative abundance, taking stock of the randomness in the process. At this time, let us not stray too far from the present method, which has proven to be a reliable, though not comprehensive, approach. Instead, let us consider the randomness in the generation of the electrons and propagate this variability through the standard method.

A General Poisson Model. As a start, let us assume that the probed area is fixed for a given examination. Hence, the true abundance A_i and relative abundance R_i of element i in this area are constant. The number of detectable electrons in the system from one probing of this area to the next, however, is a random variable.

Let $N = N(\epsilon)$ be the total number of detectable electrons with kinetic energy ϵ in the system, during a time period of length Δt. Note that N is a random variable. Suppose that N is a sum of the number of generated Auger electrons, N_A, and the number of other, say background, electrons, N_B, in the system:

$$N(\epsilon) = N_A(\epsilon) + N_B(\epsilon) .$$

These random variables (N, N_A, N_B) cannot be measured directly, but it is plausible that they are Poisson distributed.

Let $C = C(\epsilon)$ be the total count of electrons detected in a given time period. This count is also a random variable, and a fraction of N. It follows that C is the sum of the count of Auger electrons, C_A, and the count of background electrons, C_B, fractions of N_A and N_B, respectively:

$$C(\epsilon) = C_A(\epsilon) + C_B(\epsilon) .$$

Though C can be measured, the other two variables (C_A and C_B) cannot.

Let us assume that at each kinetic energy, the arrivals of Auger electrons at the detector are independent of the arrivals of background electrons. Let's also assume that, for a given kinetic energy, the counts of Auger and background electrons are Poisson random variables. Further, let us suppose that the dependence among the Auger random variables across kinetic energies is uniquely determined by an identified function of the Poisson parameter, say $\lambda_A(\epsilon)$. Let us assume a similar relationship among the background random variables, say $\lambda_B(\epsilon)$. These assumptions are reasonable given the operational characteristics of a typical Auger spectroscopy system and the earlier assumption that N_A, N_B and N are Poisson distributed. We will refer to the collections of random variables, Auger and background, as Poisson processes. Namely,

$$\begin{aligned} C_A(\epsilon) &\sim Poisson(\lambda_A(\epsilon)) \\ C_B(\epsilon) &\sim Poisson(\lambda_B(\epsilon)) . \end{aligned}$$

It follows directly that $C(\epsilon)$, the sum of independent Poisson random variables, is a Poisson random variable and the collection over all kinetic energies is a Poisson process, say the cumulative process:

$$C(\epsilon) \sim Poisson(\lambda_A(\epsilon) + \lambda_B(\epsilon)) .$$

Observations about the Auger Process. The Auger and background processes differ markedly. The graphs (Figures 8, 9 and 10) suggest that the background process is the dominating process and produces almost all the electrons counted. The background process appears to peak early, near a kinetic energy of 100eV. From 100eV to 3000eV, the background process changes very smoothly. It decreases almost monotonically from 100eV to 700eV and then rises very slowly thereafter.

These graphs suggest that the Auger process produces a small but perceptible contribution to the cumulative count of electrons. Graphically, the Auger process appears as an almost smooth curve with a nearly symmetric unimodal peak centered at its Auger energy (Figure 36). The detected Auger electrons create a dimple on the otherwise smooth background electron distribution. Comparison of several direct spectra indicates that with a known increase in surface elemental abundance or an increase in instrument sensitivity, the height and width of the dimple increases. When spectroscopists speak of the shape of the dimple, they use the term "second-order". I will show that they may be thinking "Gaussian", as reflected in their choice of peak-to-peak distance as the single dimple-based measure leading to an estimate of relative elemental abundance.

For any particular element, the highest number of electrons will be counted at its Auger energy. Nonetheless, because of scattering, filtering and other inherent variabilities, some Auger electrons are counted at lower energies, and some at higher energies. The number of Auger electrons counted decreases monotonically, symmetrically, and sharply as their kinetic energy deviates more and more from the element's Auger energy. These observations, together with the spectroscopists' view of dimple shape, suggest a model for the dimple function:

$$\lambda_A(\epsilon) = a \cdot \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{1}{2}\frac{(\epsilon - \epsilon_A)^2}{\sigma_A^2}\right] .$$

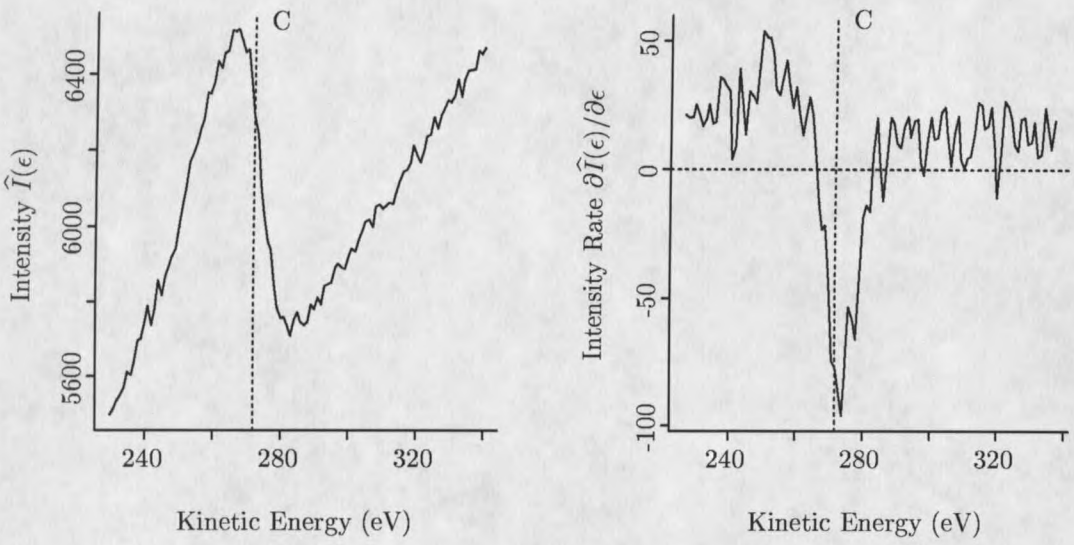


Figure 36: An almost smooth, nearly symmetric carbon "dimple" on a direct Auger spectrum with an estimate of its derivative from a scan of a 304 stainless steel sample. The horizontal position of the vertical dotted line is at the Auger energy of carbon.

where ϵ_A is an element's Auger energy, σ_A is a shape parameter affecting the spread of the function, and a is a parameter affecting its height.

The background process is assumed to be constant in the neighborhood of ϵ_A over which $\lambda_A(\epsilon)$ varies significantly. That is,

$$\lambda_B(\epsilon) = \lambda$$

for $\epsilon \in (\epsilon_A - C\sigma_A, \epsilon_A + C\sigma_A)$ where λ and C are constants.

The electron distribution $\{C(\epsilon)\}_{\epsilon=0eV}^{3000eV}$ is one realization of the cumulative process. Let us look at the expected relative elemental abundances based on the averaging of replicate realizations.

Expectations about the Auger Process.

Adopting the Poisson model described earlier, statistical properties of of the Auger process can be examined. This examination can be extended to the standard estimators such as the direct and derivative spectra, the peak-to-peak distance, and relative elemental abundance.

Expected Values of Auger Estimators. Assuming the stochastic model of the electron distribution follows the Poisson process described earlier, the expected number of detected electrons for the cumulative process is

$$\begin{aligned} \mathrm{E}[C(\epsilon)] &= \lambda(\epsilon) \\ &= \lambda_A(\epsilon) + \lambda_B(\epsilon) \, . \end{aligned}$$

Recall that the direct Auger spectrum is the product of $C(\epsilon)$ and the linear function $R(\epsilon)/\Delta t$. In practice, spectroscopists consider $R(\epsilon)$ a constant, $R(\epsilon) = 0.005$. Therefore, the notation will be shortened to R (Equation 4). If $I = I(\epsilon)$ denotes the direct Auger spectrum, then the expected value of the direct Auger spectrum is

$$\begin{aligned} \mathrm{E}[I] &= \mathrm{E}\left[C(\epsilon) \cdot \frac{\epsilon R}{\Delta t}\right] \\ &= \frac{\epsilon R}{\Delta t} \cdot \mathrm{E}[C(\epsilon)] \\ &= \frac{\epsilon R}{\Delta t} \cdot \lambda(\epsilon) \\ &= \frac{\epsilon R}{\Delta t}[\lambda_A(\epsilon) + \lambda_B(\epsilon)] \, . \end{aligned}$$

Recalling that expectation is an integration, the expected value of the derivative Auger spectrum can be determined by differentiating under the integral (within the expectation) with respect to kinetic energy ϵ. This is possible because the integrand is continuous in the location parameter ϵ.

$$\begin{aligned} \mathrm{E}[I'] &= \mathrm{E}\left[\frac{\partial}{\partial \epsilon}(C(\epsilon) \cdot \frac{\epsilon R}{\Delta t}\right] \\ &= \frac{\partial}{\partial \epsilon}\mathrm{E}\left[C(\epsilon) \cdot \frac{\epsilon R}{\Delta t}\right] \\ &= \frac{R}{\Delta t} \cdot [\lambda_A(\epsilon) + \lambda_B(\epsilon)] + \frac{\epsilon R}{\Delta t}[\lambda'_A(\epsilon) + \lambda'_B(\epsilon)] \, . \end{aligned}$$

Now, suppose that the Auger energy for an element is ϵ_A. Let us also assume that $\lambda_A(\cdot)$ is a smooth, symmetric, and unimodal function centered on ϵ_A. Then, the expected mode of the distribution of Auger electrons across the range of kinetic energies will be $\lambda_A(\epsilon_A)$. Furthermore, the largest and smallest values of the derivative λ' will occur at, say, $\epsilon_A - K$ and $\epsilon_A + K$; the

magnitude of K increasing with an increase in Auger electrons. It follows from the assumptions that

$$\begin{aligned} \lambda_A(\epsilon) &\geq 0 \quad \text{for all} \quad \epsilon \\ \lambda_A(\epsilon_A - K) &= \lambda_A(\epsilon_A + K) \\ \lambda'_A(\epsilon_A - K) &= -\lambda'_A(\epsilon_A + K)\,. \end{aligned}$$

For the expected background process, we have

$$\begin{aligned} \lambda_B(\epsilon) &> 0 \quad \text{for all} \quad \epsilon \\ \lambda_B(\epsilon - K) &\geq \lambda_B(\epsilon + K) \quad \text{for all} \quad \epsilon > 100eV \\ \lambda'_B(\epsilon) &< 0 \quad \text{for all} \quad \epsilon > 100eV \\ \lambda'_B(\epsilon - K) &< \lambda'_B(\epsilon + K) \quad \text{for all} \quad \epsilon - K > 100eV \\ \lambda''_B(\epsilon) &> 0 \quad \text{for all} \quad \epsilon - K > 100eV\,. \end{aligned}$$

The expected peak-to-peak distance D for an element with Auger energy ϵ can now be calculated:

$$\begin{aligned} \mathrm{E}[D(\epsilon_A)] &= \mathrm{E}[I'(\epsilon_A - K) - I'(\epsilon_A + K)] \\ &= \frac{R}{\Delta t}[\lambda_A(\epsilon_A - K) + \lambda_B(\epsilon_A - K)] + \frac{R}{\Delta t}(\epsilon_A - K)[\lambda'_A(\epsilon_A - K) + \lambda'_B(\epsilon_A - K)] - \\ &\quad \frac{R}{\Delta t}[\lambda_A(\epsilon_A + K) + \lambda_B(\epsilon_A + K)] - \frac{R}{\Delta t}(\epsilon_A + K)[\lambda'_A(\epsilon_A + K) + \lambda'_B(\epsilon_A + K)] \\ &= \frac{R}{\Delta t}[2\epsilon_A\lambda'_A(\epsilon_A - K) + [\lambda_B(\epsilon_A - K) - \lambda_B(\epsilon_A + K)] + \\ &\quad \epsilon_A[\lambda'_B(\epsilon_A - K) - \lambda'_B(\epsilon_A + K)] - K[\lambda'_B(\epsilon_A - K) + \lambda'_B(\epsilon_A + K)]]\,. \end{aligned}$$

Suppose that the background process is very smooth. Then, for small K, $\lambda'_B(\epsilon - K) = \lambda'_B(\epsilon + K)$ or, both $\lambda_B(\epsilon - K) = \lambda_B(\epsilon + K)$ and $\lambda'_B(\epsilon - K) = \lambda'_B(\epsilon + K) = 0$. Then the expected peak-to-peak distance is

$$\mathrm{E}[D(\epsilon_A)] = \frac{R}{\Delta t}[2\epsilon_A\lambda'_A(\epsilon_A - K) + [\lambda_B(\epsilon_A - K) - \lambda_B(\epsilon_A + K)]$$

or

$$\mathrm{E}[D(\epsilon_A)] \;=\; 2\epsilon_A \frac{R}{\Delta t}\lambda'_A(\epsilon_A - K)\,.$$

Though the effect of the background process on D is reduced under the supposition of almost no change in background, the influence of the Auger energy, a multiplicative factor, is not.

Let us now examine the relative elemental abundance for one of two elements, under the assumption of a very smooth background process. We see that if R_1 denotes relative elemental abundance of one element, then

$$\begin{aligned}
\mathrm{E}[R_1] &= \mathrm{E}\left[\frac{S_1 D_1}{S_1 D_1 + S_2 D_2}\right] \\
&\approx \frac{2\epsilon_1 S_1 \lambda'_A(\epsilon_1 - K_1)}{2\epsilon_1 S_1 \lambda'_A(\epsilon_1 - K_1) + 2\epsilon_2 S_2 \lambda'_A(\epsilon_2 - K_2)} \\
&= \frac{\epsilon_1 S_1 \lambda'_A(\epsilon_1 - K_1)}{\epsilon_1 S_1 \lambda'_A(\epsilon_1 - K_1) + \epsilon_2 S_2 \lambda'_A(\epsilon_2 - K_2)}\,.
\end{aligned}$$

where the detectability factors S_1 and S_2 are included to account for the varying sensitivity of the Auger spectrometer to the two elements and the necessary proportionality constants alluded to above. It's clear that the empirically-determined sensitivity factors also account for the multiplicative effects of the Auger energies.

Using a set of assumptions that are widely accepted in practice, we have derived a statistical result which validates the present method of estimating relative elemental abundances. This validation also points the way to estimation of uncertainties.

Auger variance estimators. Variances for many of these estimators can be derived in a straightforward fashion. In particular,

$$\begin{aligned}
\mathrm{Var}[\bar{I}] &= \mathrm{Var}\left[\frac{1}{n}\sum_{i=1}^{n} I_i(\epsilon)\right] \\
&= \mathrm{Var}\left[\frac{1}{n}\sum_{i=1}^{n} \frac{\epsilon \cdot R \cdot N_i(\epsilon)}{\Delta t}\right]
\end{aligned}$$

$$
\begin{aligned}
&= \frac{\epsilon^2 R^2}{n^2(\Delta t)^2}\mathrm{Var}\left[\sum_{i=1}^{n} N(\epsilon)\right] \\
&= \frac{\epsilon^2 R^2}{(n\Delta t)^2}\lambda(\epsilon) \\
&= \frac{\epsilon^2 R^2}{n(\Delta t)^2}[\lambda_A(\epsilon) + \lambda_B(\epsilon)] \,.
\end{aligned}
$$

If the derivative spectrum is found by numerical differentiation, then deriving a variance for the derivative of $I(\epsilon)$ may be straightforward, because a numerical derivative is often a linear combination of independent random variables. For instance, consider the following numerical derivative suggested by Burden (Burden et al. (1978)):

$$
\begin{aligned}
\frac{dy_i}{dx_i} &= (-y_i[x_i - 3h] + 6y_i[x_i - 2h] - 21y_i[x_i - 1h] + \\
&\quad 21y_i[x_i + 1h] - 6y_i[x_i + 2h] + y_i[x_i + 3h])/24h \,.
\end{aligned}
$$

To apply this formula in this application, let $y_i = I_i$, $x_i = i$ and $h = 1$. Then, the variance of this derivative estimator is

$$
\begin{aligned}
\mathrm{Var}\left[\frac{\widehat{\partial I_i}}{\partial \epsilon_i}\right] &= \frac{1}{576}\Big[\mathrm{Var}[I_{i-3}] + 36\mathrm{Var}[I_{i-2}] + 441\mathrm{Var}[I_{i-1}] + \\
&\quad 441\mathrm{Var}[I_{i+1}] + 36\mathrm{Var}[I_{i+2}] + \mathrm{Var}[I_{i+3}]\Big] \,.
\end{aligned}
$$

The variance of the peak-to-peak distance estimator follows directly:

$$
\mathrm{Var}[\widehat{D}(\epsilon_A)] = \mathrm{Var}\left[\frac{\widehat{\partial I_{\epsilon_U}}}{\partial \epsilon_A}\right] + \mathrm{Var}\left[\frac{\widehat{\partial I_{\epsilon_L}}}{\partial \epsilon_A}\right] \,.
$$

The subscripts ϵ_U and ϵ_L denote kinetic energies of the maximum (upper) and minimum (lower) peaks for an element with Auger energy ϵ_A.

An assessment of the presence of an element on a surface can be made using an approximate 95% confidence interval for the element's peak-to-peak distance. An approximate interval would be

$$
\widehat{D}(\epsilon_A) \pm 2\mathrm{SE}[\widehat{D}(\epsilon_A)] \,.
$$

It is unlikely that an element is present or that its relative abundance can be estimated reasonably well if this interval contains 0 because it would be difficult to distinguish its peaks from those generated by the background process.

APPENDIX C

Splus Functions

Splus © functions that I have composed for ADR are listed and described in this appendix. Because MLE and ADR estimation share similar estimation protocols, I have also included a brief discussion of similar functions for MLE composed by Anderson (1995).

Approximating the Limiting Null Distribution of $n\delta_A$

Boos (1981) conjectured that the null limiting distribution of $n\tilde{\delta}_A$ could be approximated by the distribution of

$$A_p^2 = \sum_{i=p+1}^{\infty} \frac{Z_i^2}{i(i+1)},$$

where the Z_i are standard normal random variables and p is the number of estimated parameters. Accepting his conjecture, I approximate null limiting distribution of $n\tilde{\delta}_A$ using simulation. My approximate distribution for any p is the empirical distribution of a large sample (*n.samp*) from the finite sum of *p.trunc* weighted standard normal deviates. The finite sum includes the $p+1$ to $p.trunc+1$ terms of Boos' approximating series.

The function **AD.null.fun** generates an *n.samp* by *p.max* matrix **AD.null.sample** of simulated random samples. The *jth* column of **AD.null.sample** corresponds to a random sample from A_{j-1}. The matrix contains approximating distributions for 1 to $p.max - 1$ estimated parameters.

```
AD.null.fun <- function(n.samp=5000,p.trunc=1000,p.max=9)
 {
  AD.null.sample <- matrix(NA,nrow=n.samp,ncol=p.max+1)
  for(i in 1:n.samp)
   {
    for(p in 0:p.max)
     {
      z <- rnorm(p.trunc,mean=0,sd=1)
      i.p <- seq(1+p,n.trunc+p)
      AD.null.sample[i,k+1] <- sum(z^2/(i.k*(i.k+1)))
     }
   }
  return(AD.null.sample)
 }
```

To estimate a p-value for any ADR fit, I select the column from **AD.null.sample** corresponding to the number of estimated parameters (i.e., $n.param + 1$). Then, I determine the proportion of values in the column greater or equal to $n.AD.fit$, the sample-size-corrected Anderson-Darling score ($n\delta_A$) from the final ADR fit. This algorithm is embodied in the function

AD.p.val.fun.

```
AD.p.val.fun <- function(n.AD.fit,AD.null.sample,n.est.param)
 {
  n.approx <- dim(AD.null.sample)[[1]]
  p.val <- sum(AD.null.sample[,n.param+1] >= n.AD.fit)/n.approx
  return(p.val)
 }
```

Splus ADR and MLE Estimation Functions

Estimating parameters using ADR requires that the Anderson-Darling goodness-of-fit statistic be minimized. This task is simplified because the Anderson-Darling integral can be expressed in closed form as a sum over the sample $\mathbf{Y}$. If $F_{\theta.i} = F_\theta(y_{(i)})$, then

$$\begin{aligned}\delta_A(F_n, F_\theta) &= -1 - \sum_{i=1}^{n} \frac{(2i-1)}{n^2}[\ln(F_{\theta.i}) + \ln(1 - F_{\theta.n+1-i})] \\ &= -1 - \sum_{i=1}^{n} \left[\frac{2i-1}{n^2}\ln(F_{\theta.i}) + \frac{2(n-i)+1}{n^2}\ln(1-F_{\theta.i})\right] .\end{aligned}$$

Finding parameter values to minimize this Anderson-Darling statistic can be done in a variety of ways. I have approached this problem from two directions: brute force and with a touch-of-elegance. I resorted to the brute force approach when the elegant approach faltered.

In the brute force approach, I generated a set of points spanning the parameter space and then identified the parameter value resulting in smallest Anderson-Darling statistic. I then generated a denser set of points spanning a neighborhood of the first candidate point, evaluated the sum and, again found the point resulting in the smallest score. I continued in this fashion until all points in a neighborhood gave nearly identical AD scores (scores agreed to five digits or more). Though not an elegant solution, it was easy to monitor the progress towards an estimate and develop an appreciation/understanding of the final estimates.

The touch-of-elegance approach made use of optimization routines built into Splus, particularly, the functions *ms* and *nlmin*. I developed my first AD optimization routine using *ms*, an older Splus function. I found this function to be unreliable; frought with convergence problems and having a tendency to "lock-up" Splus. I have since developed a new routine based on *nlmin*. I

have not encountered the level of problems I faced with the *ms*-based function, though I have not used this new routine as extensively.

The function **fit.adr**, based on *nlmin*, is listed below. I wrote a generic version in an attempt to handle all models. After several trials, I found that it was better to have a function framework that I could easily customize to the model of interest. The necessary custom components in this model are the mean response function, the residual function and the random distribution. The function listed below is 'customized' to find the ADR bandwidth estimate for a kernel smooth of data following a multiplicative χ^2 model. Customization may include updating the argument list; in this case, by adding the kernel type and initial bandwidth estimate. Note that the residual function includes a correction for bias.

```
fit.adr <- function(x,y,init.bandwidth,kernel=''gaussian'',
                max.it=100,max.fun.cal=100,prt.lvl=0)
 {
# Assign the data to Frame 1.  Make sure the attributes of the data
#  agree with the expections of the mean response function.

  assign(''x.var'',as.vector(x),frame=1)
  assign(''y.var'',as.vector(y),frame=1)

# Define the ADR Score function for this application. Not that ''x'' in
# this function is the parameter or parameter vector to be compatible
# with ''nlmin'' which returns the parameter estimates as a scalar or
# vector labelled ''x''.

  chisq.df <- 4

ADR.fun <- function(x)
 {
# Fit the mean response.

  fit <- ksmooth(x.var,y.var,kernel=kernel,bandwidth=x)$y

# Calculate the residuals and sort.

  resid <- sort((y.var/fit)*chisq.df)

# Calculate the cumulative probability for each residual.

  theo.p <- pchisq(resid,chisq.df)
  theo.p <- ifelse(theo.p == 0,1e-16,theo.p)

# Calculate the Anderson-Darling Score.

  K1 <- (2*1:n-1)/n^2
  K2 <- (2*(n-1:n)+1)/n^2
```

```
  EDF1 <- log(theo.p)
  EDF2 <- log(1-theo.p)
  AD.score <-  -1 - sum(K1*EDF1 + K2*EDF2)

  return(AD.score)
 }

  nlmin(ADR.fun,init.beta,max.iter=max.it,max.fcal=max.fun.cal,
        print.level=prt.lvl)
 }
```

The ADR estimation function **adr.fit** is a modification of a simlar function for maximum likelihood estimation. The function **FIT.MLE** was written by Kevin Anderson to solve Maximum Likelihood problems. **FIT.MLE** must contain a user-supplied function that computes the actual negative-log-likelihood, "nlogbl.fun". This function is then passed to "nlmin" to find the MLE's. The data is passed to "nlogbl.fun" by assigning it to FRAME 1 following the approach of Chambers Chambers and Hastie (1992). This method of passing data is akin to the use of common blocks in FORTRAN

```
FIT.MLE <- function(x.data, y.data, init.guess,mi=60,mf=100,pl=0)
 {
  assign("x.dat", as.vector(x.data), frame = 1)
  assign("y.dat", as.vector(y.data), frame = 1)

# Define the negative loglikelihood function in terms of the data
#   (x.dat,y.dat) and the parameters (x):

  nlogbl.fun <- function(x)
   {

   }

# Set up the minimization function.

   nlmin(nlogbl.fun,init.guess,max.iter=mi,max.fcal=mf,print.level=pl)
 }
```

The function **COV.MLE** estimates the covariance matrix of the MLEs using numerical derivatives. This function requires the same negative-log-likelihood computing function "nlogbl.fun". The user supplies the delta used in the numerical derivatives by giving COV.MLE the maximum likelihood estimates from FIT.MLE,say MLE.1 and another point estimate, say MLE.0 which should be close to MLE.1. A good choice for MLE.0 is the MLE estimates from one of the later iterations of "nlmin". To obtain MLE.0, run FIT.MLE with pl=1 to see how many

iterations it takes "nlmin" to converge. Then run FIT.MLE one less iteration by setting "mi".

```
COV.MLE <- function(x.data, y.data, MLE.0, MLE.1) {

  assign("x.dat", as.vector(x.data), frame = 1)
  assign("y.dat", as.vector(y.data), frame = 1)

# Define the negative loglikelihood function in terms of the data
#   (x.dat,y.dat) and the parameters (x):

  nlogl.fun <- function(x)
   {

   }

# Estimate the Covariance by approximating the Hessian matrix.

  assign("nlogl.fun", nlogl.fun, frame = 1)
  delta <- abs(MLE.1-MLE.0)
  np <- length(delta)
  Hess <- matrix(0,np,np)
  nlogl.hat <- nlogl.fun(MLE)

  for(i in 1:np)
   {
    di <- rep(0,np) ; di[i] <- delta[i]
    Hess[i,i] <- (nlogl.fun(MLE.1 + di) - 2*nlogl.hat +
                  nlogl.fun(MLE.1 - di))/ delta[i]^2
   }

  hess.calc <- function(x,i1,i2,delta){
      d1 <- rep(0,length(x)) ; d2 <- d1
      d1[i1] <- delta[i1] ; d2[i2] <- delta[i2]
      (nlogl.fun(x + d1 + d2) - nlogl.fun(x - d1 + d2) -
       nlogl.fun(x + d1 - d2) + nlogl.fun(x - d1 - d2))/
      (4*delta[i1]*delta[i2])
  }

  for(i in 1:(np-1)) {
     for(j in (i+1):np) {
       Hess[i,j] <- hess.calc(MLE.1,i,j,delta)
       Hess[j,i] <- Hess[i,j]
     }
   }
  solve(Hess)
 }
```

General Splus functions

Along the way, I accumulated a diverse collection of simple and complex Splus functions. This section lists a sample of these that the reader may find useful.

MEAN RESPONSE FUNCTIONS mu.fun(*)

classical linear response functions.

```
constant.fun <- function(x,beta)
  {
    b0 <- beta[1]
    lin.x <- b0
    return(m=lin.x)
  }

line.fun <- function(x,beta)
  {
    intercept <- beta[1]
    slope <- beta[2]
    return(intercept + slope * x)
  }

lin.fun <- function(x,beta)
  {
    b0 <- beta[1]
    b1 <- beta[2]
    lin.x <- b0 + b1*x
    return(m=lin.x)
  }

quad.fun <- function(x,beta)
  {
    b0 <- beta[1]
    b1 <- beta[2]
    b2 <- beta[3]
    quad.x <- b0 + b1*x +b2*x^2
    return(m=quad.x)
  }

cube.fun <- function(x,beta)
  {
    b0 <- beta[1]
    b1 <- beta[2]
    b2 <- beta[3]
    b3 <- beta[4]
    quad.x <- b0 + b1*x + b2*x^2 + b3*x^3
    return(m=cube.x)
  }
```

nonparametric response functions.

```
lowes <- function(x,y,beta)
  {
    approx(lowess(x,y,f=beta),x)
  }
```

RESIDUAL FUNCTIONS *.resid.fun(*)

Multiplicative Error Structure.

```
mult.resid.fun <- function(y,fit)
  {
    residuals <- y/fit
    return(residuals=residuals)
  }
```

Additive Error Stucture

```
add.resid.fun <- function(y,fit)
  {
    residuals <- y-fit
    return(residuals=residuals)
  }
```

Standardized Additive Error Stucture

```
std.add.resid.fun <- function(y,fit,beta)
  {
    residuals <- y-fit
    sd.resid <- beta[2]
    std.residuals <- residuals/sd.resid
    return(residuals=std.residuals)
  }
```

RESIDUAL DISTRIBUTION FUNCTIONS resid.distn(*)

```
resid.distn <- function(x,y,beta,mu.fun,resid.fun=mult.resid.fun,
                        family="chisq",...)
  {
    n <- length(x)
    fit <- mu.fun(x, beta)
    fit <- ifelse(fit == 0, 1e-06, fit)
    residuals <- resid.fun(y,fit)
    resid.index <- sort.list(residuals)
    empirical.q <- residuals[resid.index]
    empirical.p <- seq(1/(n - 0.5), (n - 0.5)/n, length = n)
```

```
      theoretical.p <- eval(as.name(paste("p", family, sep = "")))(
                empirical.q, ...)
      theoretical.q <- eval(as.name(paste("q", family, sep = "")))(
                empirical.p, ...)
      pdiff <- theoretical.p - empirical.p
      qdiff <- theoretical.q - empirical.q
      return(list(pdiff=pdiff, qdiff = qdiff, theo.p = theoretical.p,
               emp.p = empirical.p, theo.q = theoretical.q,
               emp.q = empirical.q,resid.index = resid.index ))
    }

  resid.distn2 <- function(x,y,beta,mu.fun=constant.fun,
                        resid.fun=std.add.resid.fun,family="norm",...)
    {
      n <- length(x)
      fit <- mu.fun(x, beta)
      fit <- ifelse(fit == 0, 1e-06, fit)
      residuals <- resid.fun(y,fit,beta)
      resid.index <- sort.list(residuals)
      empirical.q <- residuals[resid.index]
      theoretical.p <- eval(as.name(paste("p", family, sep = "")))(
                empirical.q, ...)
      return(list(theo.p = theoretical.p))
    }
```

DISTRIBUTION METRICS metric(*)

KOLMOGOROV Metric.

```
  kolmgorov <- function(x,y,beta,resid.distn,mu.fun,
                     resid.fun,family,...)
    {
      results <- resid.distn(x,y,beta,mu.fun,resid.fun,family,...)
      return(max(abs(results$pdiff))
    }
```

KUIPER Metric.

```
  kuiper <- function(x, y, beta, resid.distn, mu.fun, resid.fun, family, ...)
    {
     results <- resid.distn(x, y, beta, mu.fun, resid.fun, family, ...)
     return(score = max(results$pdiff) + abs(min(results$pdiff)))
    }
```

LEVY Metric.

```
  levy <- function(x, y, beta, resid.distn, mu.fun, resid.fun, family, ...)
    {
      tmp<- resid.distn(x, y, beta, mu.fun, resid.fun, family,...)
```

```
    theo.theta <- atan(tmp$theoretical.p/tmp$theoretical.q)
    theo.phi <- (theo.theta - 45*(pi/180))
    theo.c <- sqrt(tmp$theoretical.p^2 + tmp$theoretical.q^2)
    theo.a.new <- theo.c*cos(theo.phi)
    theo.b.new <- theo.c*sin(theo.phi)

    emp.theta <- atan(tmp$empirical.p/tmp$empirical.q)
    emp.phi <- (emp.theta - 45*(pi/180))
    emp.c <- sqrt(tmp$empirical.p^2 + tmp$empirical.q^2)
    emp.a.new <- emp.c*cos(emp.phi)
    emp.b.new <- emp.c*sin(emp.phi)

    new.emp.b <- approx(theo.a.new,theo.b.new,emp.a.new)
    levy.score <- max(abs(new.emp.b$y-emp.b.new),na.rm=T)

    return(score=levy.score)
  }
```

CRAMER-VON MISES Metric.

```
  cramer <- function(x,y,beta,resid.distn,mu.fun,resid.fun,family, ...)
    {
      results <- resid.distn(x,y,beta,mu.fun,resid.fun,family,...)
      n <- length(x)
      Wn <- 1/(12*n) + sum((results$theo.p - ((2*(1:n)-1)/2*n)^2)
      return(Wn=Wn)
    }
```

ANDERSON-DARLING Metric (a weighted C-VM metric).

```
  anderson <- function(x,y,beta,resid.distn,mu.fun,resid.fun,family, ...)
    {
      results <- resid.distn(x,y,beta,mu.fun,resid.fun,family,...)
      n <- length(x)
      An <- -sum((2*(1:n)-1)*(log(results$theo.p) +
               log(1-rev(sort(results$theo.p))))/n - n
      return(An=An)
    }

  anderson2 <- function(x,y,beta,resid.distn2,mu.fun,resid.fun,family, ...)
    {
      results <- resid.distn2(x,y,beta,mu.fun,resid.fun,family,...)
      n <- length(x)
      An <- -n^(-2)*sum((2*(1:n)-1)*(log(results$theo.p) +
               log(1-rev(sort(results$theo.p)))) - 1
      return(An=An)
    }
```