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.

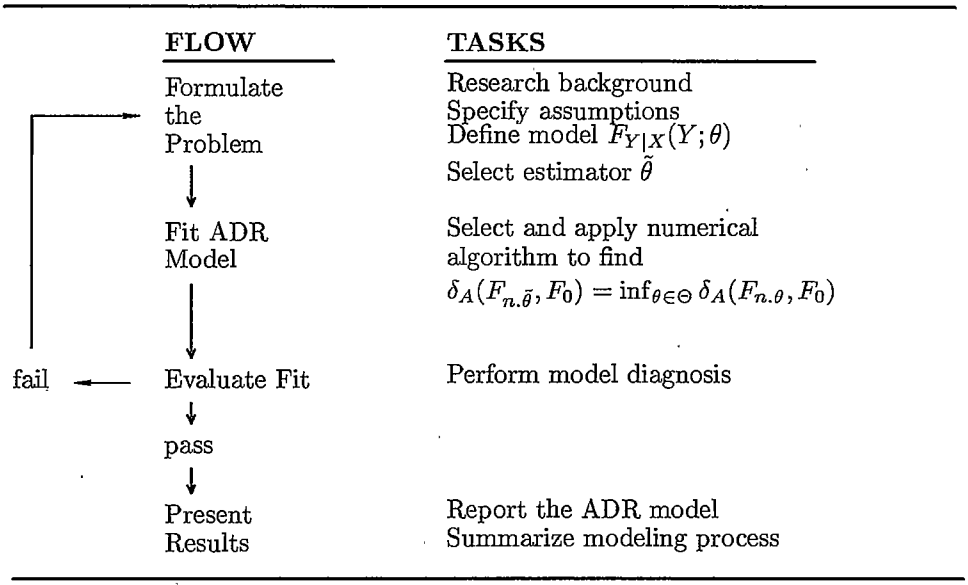
The Modeling Process

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

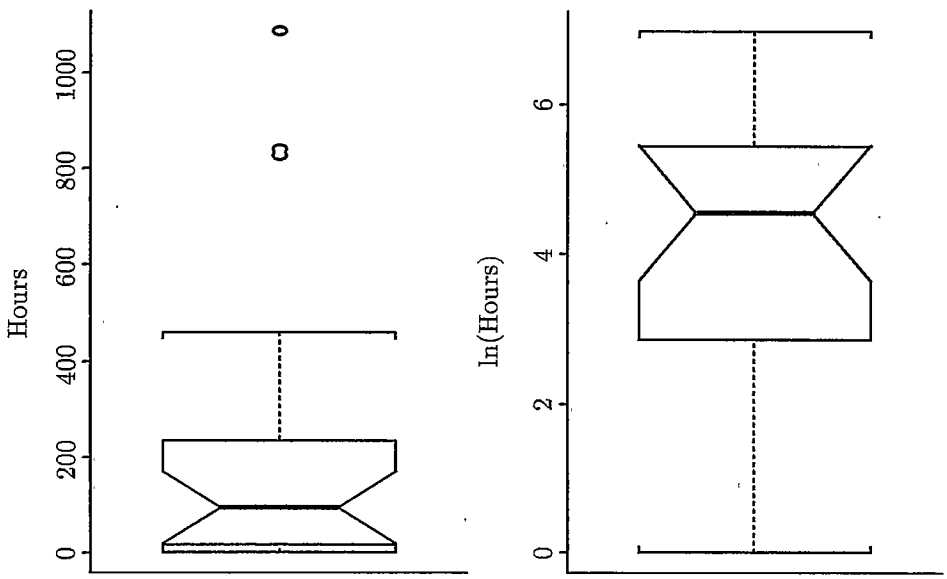


Figure 2: Time-to-Failure Sample Distribution for 20 Vehicle Guidance Systems. Left panel shows the boxplot of measured times-to-failure. The right panel shows the distribution of the log-transformed times.

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

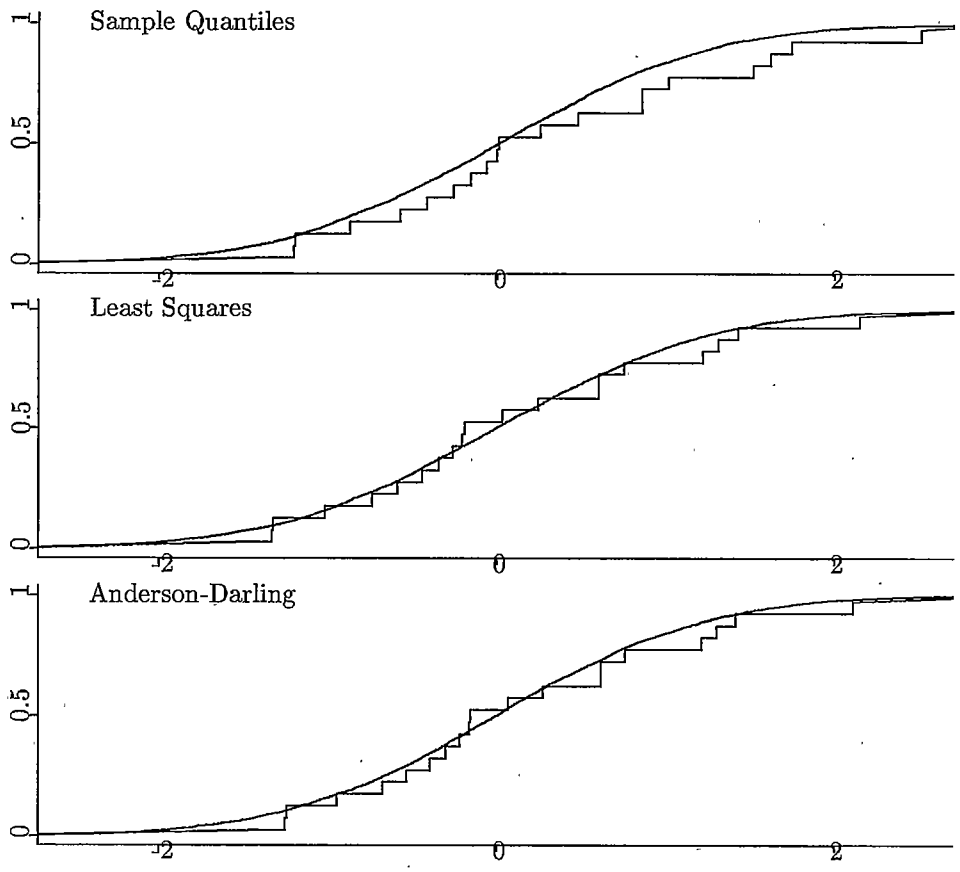


Figure 3: Empirical distributions of normalized residuals from Quantile (median and interquartile range), Least Squares and Anderson-Darling fits of the Time-to-failure logNormal model. The N(0,1) distribution function is marked by the smooth curve.

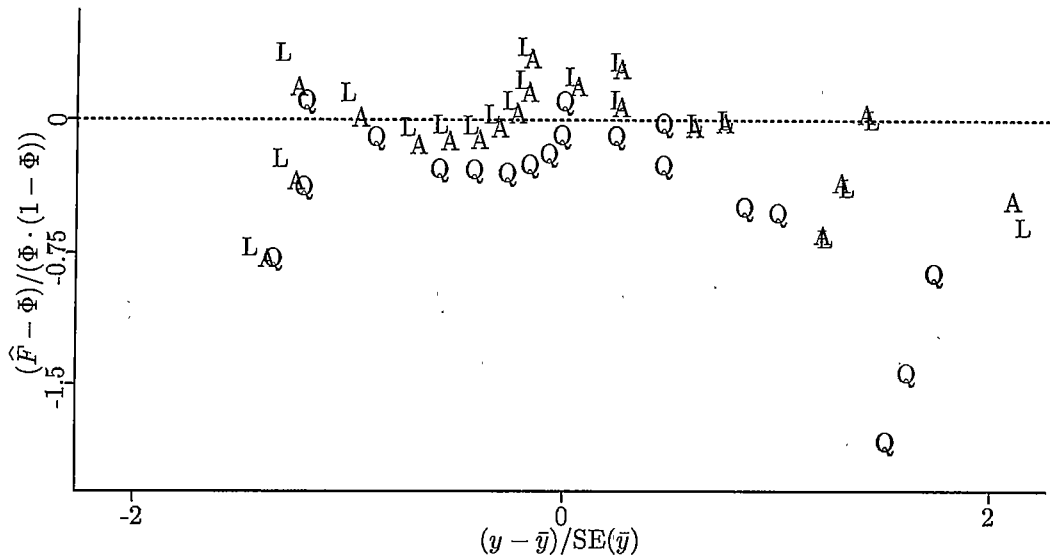


Figure 4: Normalized residuals from the **Q**uantile (median and interquartile range), **L**east Squares and **A**nderson-Darling fits of the Time-to-Failure logNormal model plotted against the contributions of the residuals to the AD statistic (i.e., $(F_n - \Phi)/(\Phi \cdot (1 - \Phi))$).

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.

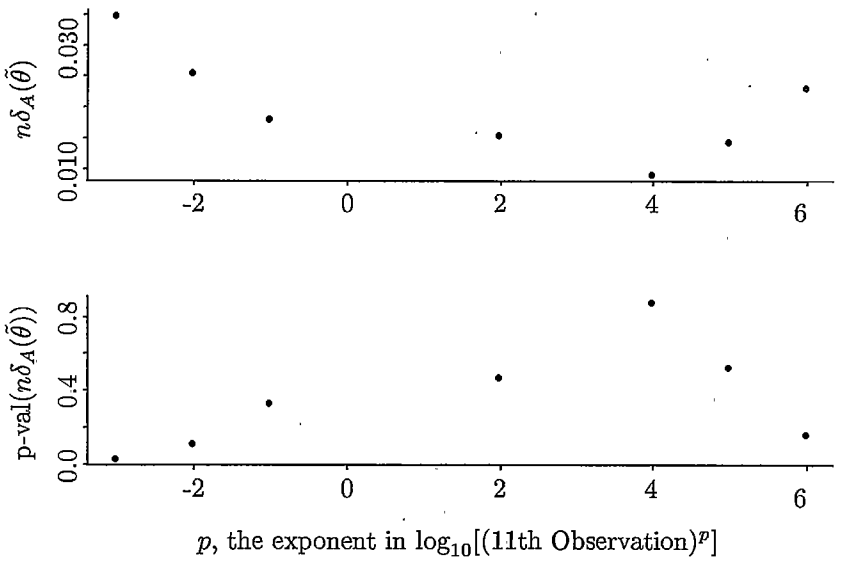


Figure 5: Anderson-Darling Scores with P-values When One Observation is Perturbed

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.

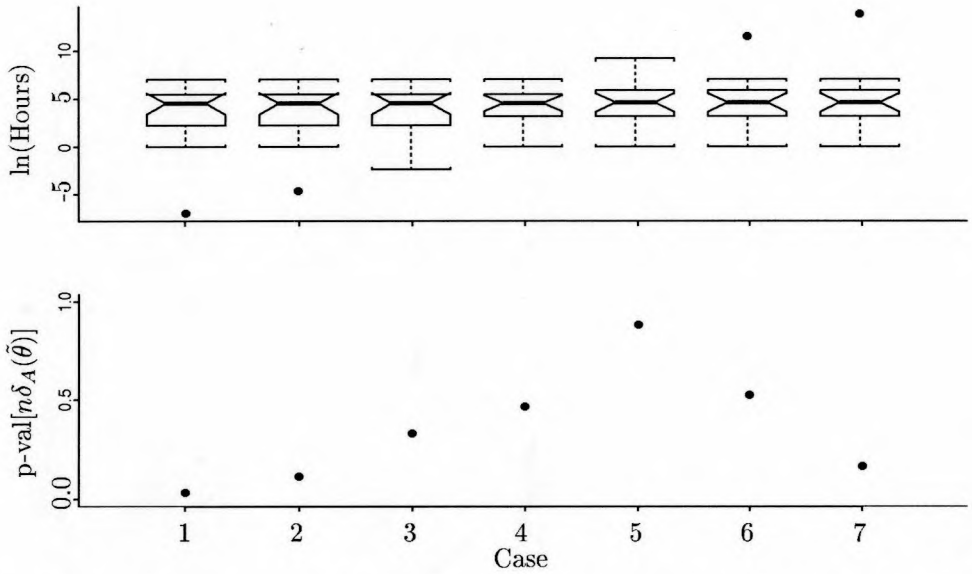


Figure 6: Boxplots of the Perturbed Time-to-Failure Datasets and Anderson-Darling p-values for the cases listed in Table 4.

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: $Y_i = m(\mathbf{X}_i) \circ \epsilon_i$.