COMPUTER-INTENSIVE METHODS IN STATISTICAL REGRESSION

BY

BRADLEY EFRON

TECHNICAL REPORT NO. 112
APRIL 1986

PREPARED UNDER THE AUSPICES
OF
PUBLIC HEALTH SERVICE GRANT 2 RO1 GM21215-12

DIVISION OF BIOSTATISTICS
STANFORD UNIVERSITY
STANFORD, CALIFORNIA
COMPUTER-INTENSIVE METHODS IN STATISTICAL REGRESSION

BY

BRADLEY EFRON

TECHNICAL REPORT NO. 112

APRIL 1986

PREPARED UNDER THE AUSPICES

OF

PUBLIC HEALTH SERVICE GRANT 2 R01 GM21215-12

ALSO PREPARED UNDER NATIONAL SCIENCE FOUNDATION GRANT MCS80-24649 AND ISSUED AS TECHNICAL REPORT NO. 245, DEPARTMENT OF STATISTICS, STANFORD UNIVERSITY.

DIVISION OF BIOSTATISTICS

STANFORD UNIVERSITY

STANFORD, CALIFORNIA
COMPUTER-INTENSIVE METHODS IN STATISTICAL REGRESSION

Bradley Efron

Abstract

This is a survey of modern developments in statistical regression, written for the mathematically educated non-statistician. We begin with a review of the traditional theory of least-squares curve-fitting. Modern developments in regression theory have developed in response to the practical limitations of the least-squares approach. Recent progress has been made feasible by the electronic computer, which frees statisticians from the confines of mathematical tractability. Topics discussed include robust regression, bootstrap measures of variability, local smoothing and cross-validation, projection pursuit, Mallows' C_p criterion, Stein estimation, generalized regression for Poisson data, and regression methods for censored data. All of the methods are illustrated with real-life examples.

1. Introduction.

One of the oldest and most useful statistical methods, linear regression, has reemerged as a topic of intense research effort. Its renaissance reflects the impact upon statistics of modern computational equipment. Statisticians are now able to invent, investigate, and routinely use methods which require a million times the number of computations of traditional approaches, see Efron (1979). This article, which is written for the mathematically educated non-statistician with a knowledge of elementary probability theory, explores some of the new regression techniques.

The paper is based on a series of real-life examples. Our goal is to show both how and why regression theory is developing along certain new directions. The first example, which occupies the rest of this section, shows traditional least-squares regression theory in action. Sections 2 and 3 review the theory of least
squares, which began with Legendre and Gauss in the early 1800's, and was completed by Fisher in the 1920's. All of the modern developments spring from this same source, though by the end of the paper it will be clear that modern computational equipment has vastly extended the power and applicability of regression methods.

Here is a small but genuine example illustrating why regression analysis is indispensable to the intelligent interpretation of experimental data. A drug delivery device was designed to steadily release an anti-inflammatory hormone. The left side of Table 1 shows the amount of hormone remaining in 27 such devices after removal from the patient. The devices were manufactured in 3 lots, 9 devices each lot, and, since this was at an early stage of development, the question arose as to the stability of the manufacturing process. Lot C has a much higher mean than Lots A or B. The estimated standard errors of the means (how much statistical uncertainty there is in the observed average for each lot, as discussed more carefully at the end of Section 2) indicates that a difference of this magnitude cannot be attributed to random error.

A glance at the right side of Table 1 shows how flawed this conclusion is. There is a systematic difference between lots which swamps the random errors. The devices in Lot C were worn for less time than the others, and since the device releases hormone steadily during wear, it is not surprising that Lot C gave bigger numbers. Regression analysis offers a way of making this line of reasoning precise.

The 27 data points are plotted in Figure 1 with horizontal and vertical axis

\[(t,y) = \text{(hours worn, remaining hormone)}, \quad (1.1)\]

the lots being indicated by the plotting symbol. The strong dependence of \(y\) on \(t\) is quite apparent. A very simple model for this dependence assumes that \(y\) has true mean value \(\mu(t)\) depending linearly on the time of wear \(t\),

\[\mu(t) = \beta_0 + \beta_1 t . \quad (1.2)\]
Amount of Hormone Remaining, mgs

<table>
<thead>
<tr>
<th>Lot:</th>
<th>A</th>
<th>B</th>
<th>C</th>
</tr>
</thead>
<tbody>
<tr>
<td>Device</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1.</td>
<td>25.8</td>
<td>16.3</td>
<td>28.8</td>
</tr>
<tr>
<td>2.</td>
<td>20.5</td>
<td>11.6</td>
<td>22.0</td>
</tr>
<tr>
<td>3.</td>
<td>14.3</td>
<td>11.8</td>
<td>29.7</td>
</tr>
<tr>
<td>4.</td>
<td>23.2</td>
<td>32.5</td>
<td>28.9</td>
</tr>
<tr>
<td>5.</td>
<td>20.6</td>
<td>32.0</td>
<td>32.8</td>
</tr>
<tr>
<td>6.</td>
<td>31.1</td>
<td>18.0</td>
<td>32.5</td>
</tr>
<tr>
<td>7.</td>
<td>20.9</td>
<td>24.1</td>
<td>25.4</td>
</tr>
<tr>
<td>8.</td>
<td>20.9</td>
<td>26.5</td>
<td>31.7</td>
</tr>
<tr>
<td>9.</td>
<td>30.4</td>
<td>25.8</td>
<td>28.5</td>
</tr>
</tbody>
</table>

Mean: 23.1  22.1  28.9  
St. Error: 1.8  2.7  1.2

Hours Worn

<table>
<thead>
<tr>
<th>Lot:</th>
<th>A</th>
<th>B</th>
<th>C</th>
</tr>
</thead>
<tbody>
<tr>
<td>Device</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1.</td>
<td>99</td>
<td>376</td>
<td>119</td>
</tr>
<tr>
<td>2.</td>
<td>152</td>
<td>385</td>
<td>188</td>
</tr>
<tr>
<td>3.</td>
<td>293</td>
<td>402</td>
<td>115</td>
</tr>
<tr>
<td>4.</td>
<td>155</td>
<td>29</td>
<td>88</td>
</tr>
<tr>
<td>5.</td>
<td>196</td>
<td>76</td>
<td>58</td>
</tr>
<tr>
<td>6.</td>
<td>53</td>
<td>296</td>
<td>49</td>
</tr>
<tr>
<td>7.</td>
<td>184</td>
<td>151</td>
<td>150</td>
</tr>
<tr>
<td>8.</td>
<td>171</td>
<td>177</td>
<td>107</td>
</tr>
<tr>
<td>9.</td>
<td>52</td>
<td>209</td>
<td>125</td>
</tr>
</tbody>
</table>

Mean: 150.6  233.4  111.0

Table 1 - The Hormone Data. Results of an experiment concerning a device which releases anti-inflammatory hormone. The amount of hormone remaining in the device was measured for 27 devices, manufactured in 3 lots of 9 devices each (left panel). The right panel shows the number of hours each device was worn. Lot C looks like it contained considerably more of the hormone.

This is our first example of a linear regression.

The true regression line \((1.2)\) cannot be observed, at least not without assaying an infinite number of devices, but we can fit a "best" line to the 27 observed data points \((t_i,y_i)\) by the methods of least squares: choosing \((\hat{\beta}_0, \hat{\beta}_1)\) to minimize \(\sum_{i=1}^{27} [y_i - (\hat{\beta}_0 + \hat{\beta}_1 t_i)]^2\). This gives the least squares line

\[
\hat{\mu}(t) = \hat{\beta}_0 + \hat{\beta}_1 t
\]

shown in Figure 1.

Table 2 compares the residual values from the least-squares line \((1.3)\),

\[
r_i = y_i - [\hat{\beta}_0 + \hat{\beta}_1 t_i].
\]

(1.4)
remaining hormone versus hours, by lot

![Graph of remaining hormone versus hours, by lot](image)

**Figure 1.** A plot of the amount of hormone remaining in each device (vertical axis) versus the number of hours the device was worn (horizontal axis). Each point is labelled with the lot of the corresponding device. The straight line is the least squares fit to the 27 data points.

The basic idea is simple but powerful: it is more meaningful to compare the residuals $r_i$ than the raw measurements $y_i$, because we have removed from the comparison the confusing effect of "hours worn". Table 2 shows that having made this adjustment, Lot A has smaller amounts of remaining hormone than either Lots B or C. This is also evident from Figure 1.

Our regression analysis provides more information than just an improved comparison between the lots. The regression line itself is quite interesting: its slope, $\hat{\beta}_1 = -.0574$, is a good estimate of how quickly the hormone is being released from the devices, about 5.74 mg of hormone per 100 hours of wear. This example is typical in
<table>
<thead>
<tr>
<th>Lot:</th>
<th>A</th>
<th>B</th>
<th>C</th>
</tr>
</thead>
<tbody>
<tr>
<td>Device</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1.</td>
<td>-2.68</td>
<td>3.73</td>
<td>1.47</td>
</tr>
<tr>
<td>2.</td>
<td>-4.94</td>
<td>-0.45</td>
<td>-1.37</td>
</tr>
<tr>
<td>3.</td>
<td>-3.34</td>
<td>0.73</td>
<td>2.14</td>
</tr>
<tr>
<td>4.</td>
<td>-2.06</td>
<td>0.00</td>
<td>-0.21</td>
</tr>
<tr>
<td>5.</td>
<td>-2.31</td>
<td>2.20</td>
<td>1.96</td>
</tr>
<tr>
<td>6.</td>
<td>-0.02</td>
<td>0.84</td>
<td>1.15</td>
</tr>
<tr>
<td>7.</td>
<td>-2.70</td>
<td>-1.39</td>
<td>-0.15</td>
</tr>
<tr>
<td>8.</td>
<td>-3.44</td>
<td>2.50</td>
<td>3.68</td>
</tr>
<tr>
<td>9.</td>
<td>-0.78</td>
<td>3.64</td>
<td>1.51</td>
</tr>
</tbody>
</table>

Mean: -2.48 1.31 1.13
St. error: 0.48 0.60 0.50

Table 2. Residual values of the remaining hormone, after subtracting the value predicted by the least squares line. Now we see that Lot A contained less hormone than B or C.

that both the regression curve itself, in this case line (1.3), and the deviations from the regression, as in Table 2, contain important information.

Notice that the standard errors in Table 2 are considerably smaller than those in Table 1. A successful regression analysis helps explain variability. In this case we have explained a large part of the variability of the raw measurements $y_i$ in terms of an explanatory variable, "hours worn".

2. Ordinary Least Squares.

All of the new methods in regression analysis are children of the same successful parent: the traditional theory of least squares curve-fitting begun by Legendre and Gauss, and completed by Sir Ronald Fisher in the 1920's. This section and the next give a brief review of the traditional theory of regression based on the method of least squares. The presentation, which omits all proofs, assumes only elementary
probability theory and linear algebra as a background. Chapter 1 of Scheffé (1959) is an excellent reference for this material.

The data which goes into a regression analysis consists of pairs

\[(x_i, y_i) \quad i = 1, 2, \ldots, n,\]  

(2.1)

where \( n \) is the number of data points, \( n = 27 \) in the hormone example. Here \( y_i \) is the response variable and \( x_i \) is the predictor. In example (1.1), \( x_i \) was based on the single quantity "hours worn", but in general \( x_i \) can be a vector. A great advantage of regression analysis is that it easily accommodates complicated prediction models, sometimes involving \( x_i \) vectors with dozens of components. For our general discussion we will take \( x_i \) to be a \( p \)-dimensional row vector.

Fitting a line to data by least squares, as in Figure 1, is a numerical algorithm which can be carried out for any set of observed points. If we want to go deeper into the problem, to ask how accurate is the fitted line, or if least squares is in any sense a optimal procedure, we need a probabilistic model relating \( y_i \) to \( x_i \). The traditional model begins by assuming that

\[y_i = \mu_i + e_i \quad i = 1, 2, \ldots, n\]  

(2.2)

where \( \mu_i \) is the true mean (or "expectation") of \( y_i \) given the value of \( x_i \), while \( e_i \) is pure normal ("Gaussian", "bell-shaped") noise,

\[e_i \sim N(0, \sigma^2) \quad \text{independently} \quad i = 1, 2, \ldots, n.\]  

(2.3)

The notation \( N(0, \sigma^2) \) indicates that the error terms \( e_i \) are normally distributed with mean 0 and variance \( \sigma^2 \).

Model (2.2) represents each \( y_i \) as signal plus noise. The heart of the regression model is an assumption that the signals \( \mu_i \) are not arbitrary, but rather that cases which are close to each other in terms of the prediction vectors \( x_i \)
are also close in terms of $\mu_i$. This relationship is expressed by the linear formula

$$\mu_i = x_i \beta, \quad i = 1, 2, \ldots, n. \quad (2.4)$$

The $1 \times p$ predictor vectors $x_i$ are known to the statistician, but the $p \times 1$ parameter vector $\beta$ is unknown, and must be estimated from the data. This is often the main point of a regression analysis.

In example (1.2), the dimension was $p = 2$ and the $i$th predictor vector was $x_i = (1, t_i)$, where $t_i$ was "hours worn" for the $i$th device. According to model (2.4) the unknown parameter vector $\beta = (\beta_0, \beta_1)'$ gives

$$\mu_i = \beta_0 + \beta_1 t_i \quad (2.5)$$

as the true mean value of remaining hormone $y_i$ for a device worn $t_i$ hours, as in (1.2). We estimated the unknown straight-line relationship (2.5) by the least-squares line (1.3), $\hat{\mu}(t) = \hat{\beta}_0 + \hat{\beta}_1 t$, shown in Figure 1.

The traditional regression model can be stated succinctly in the language of linear algebra. Let $y = (y_1, y_2, \ldots, y_n)'$ be the $n \times 1$ vector of observations, $\mu = (\mu_1, \mu_2, \ldots, \mu_n)'$ be the $n \times 1$ vector of true means, and $X$ be the $n \times p$ matrix with $i$th row $x_i$. Then (2.2)-(2.4) is equivalent to

$$y = \mu + \varepsilon \quad (2.6)$$

where

$$\mu = X\beta \quad \text{and} \quad \varepsilon \sim N(0, \sigma^2 I). \quad (2.7)$$

The notation $\varepsilon \sim N(0, \sigma^2 I)$ for an $n$-dimensional vector of pure normal noise is identical in meaning to (2.3).

The vector $y$ can take on any value in $\mathbb{R}^n$, $n$-dimensional Euclidean space. However the true mean vector $\mu = X\beta$ is constrained to lie in the column space of $X$, say
\[ \mathcal{L} \equiv \{ u = Xb, \; b \in \mathbb{R}^p \}. \quad (2.8) \]

Figure 2 illustrates the situation. To avoid algebraic difficulties we will assume that \( X \) is of full rank \( p \). Then \( \mathcal{L} \) is a \( p \)-dimensional linear subspace of \( \mathbb{R}^n \), and exactly one point \( \tilde{u} = \tilde{X}b \) in \( \mathcal{L} \) corresponds to each \( p \)-dimensional vector \( b \).

\[ \tilde{y} \]

\[ \tilde{e} \]

\[ \hat{u} = \tilde{X}\hat{\beta} \]

\[ \tilde{y} - \hat{u} \]

\[ \mu = \mu \]

\[ \mu = \mu \]

**Figure 2.** A schematic picture of the linear model. The true mean vector \( \mu \) is constrained to lie in the \( p \)-dimensional linear subspace \( \mathcal{L} \) spanned by the columns of \( X \). We observe \( y = \mu + e \), and wish to estimate \( \mu = \tilde{X} \hat{\beta} \). The least squares estimate \( \hat{\beta} = \tilde{X}^\dagger \tilde{y} \) is the point in \( \mathcal{L} \) nearest to \( y \). The residual vector \( e = y - \hat{y} \), orthogonal to \( \mathcal{L} \), is used to estimate the noise parameter \( \sigma^2 \).

The basic idea of least squares estimation is rather obvious from Figure 2: we estimate \( \mu = \tilde{X} \hat{\beta} \) by the point \( \hat{\mu} = \tilde{X} \hat{\beta} \) in \( \mathcal{L} \) nearest to \( \tilde{y} \) in Euclidean distance, that is by the minimizer of \( \| y - \tilde{X}b \|_2 \) over \( b \in \mathbb{R}^p \). The least squares solution has a neat closed-form expression, the so-called "Normal Equations"

\[ \hat{\beta} = \tilde{C}^{-1} \tilde{X}' \tilde{y} \quad (\tilde{C} = \tilde{X}' \tilde{X}) \], \quad (2.9)

going back to Legendre and Gauss in the early 1800's (who had a nasty priority fight over the invention of least squares). The solution \( \hat{\beta} = (34.17, -.0574)' \) in (1.3) was computed from (2.9); notice that in this case \( \tilde{X} \) is a \( 27 \times 2 \) matrix but \( \tilde{C} \) is only \( 2 \times 2 \) so that the numerical solution of (2.9) is easy.
Given the assumptions of the linear model (2.6), (2.7), $\hat{\beta}$ is an unbiased estimator of the vector $\beta$, by which we mean that $E(\hat{\beta}) = \beta$. Moreover it is the best unbiased estimator, in the sense that all others have greater variances for estimating every component of $\beta$.

This last point is the principal theoretical justification for the method of least squares, and is worth stating quantitatively. The covariance matrix $\Sigma$ of the estimator $\hat{\beta}$ is by definition the matrix whose ith diagonal element is the variance of the ith component of $\hat{\beta}$. (The off-diagonal elements refer to correlations between the components.) It is easy to calculate that

$$\Sigma = \sigma^2 G^{-1}.$$  \hspace{1cm} (2.10)\]

Statistical theory shows that any other unbiased estimator of $\beta$ has covariance matrix larger than (2.10), in particular having larger diagonal elements, i.e. bigger variances for the component estimates.

Formula (2.10) allows us to assess the accuracy of the least-squares estimate $\hat{\beta}_i$. The only unknown quantity in (2.10) is $\sigma^2$. This can be estimated from the residual vector

$$\tilde{r} = y - \hat{\mu}.$$  \hspace{1cm} (2.11)\]

The usual estimate is

$$\hat{\sigma}^2 = \frac{\|\tilde{x}\|^2}{n-p},$$  \hspace{1cm} (2.12)\]

which is the best unbiased estimator of $\sigma^2$ under model (2.6), (2.7). From Figure 2 we see that $\tilde{r}$ always lies in the $n-p$ dimensional space orthogonal to $L$, which accounts for the denominator of (2.12). Because $\tilde{r}$ is orthogonal to the space containing $\tilde{\mu}$, its distribution depends only on the noise $\sigma^2$ in (2.6), (2.7), and not on the signal $\mu$. In brief, the part of $y$ in $L$ estimates $\mu$ (or equivalently $\beta$), while the part orthogonal to $L$ estimates $\sigma^2$.
Most least-squares regression programs print out the estimated coefficients \( \hat{\beta}_i \) and also their estimated standard errors (square root of estimated variance) from (2.10, (2.12): \[ \hat{\sigma}^2 (G^{-1})_{ii} \]. Linear regression (1.3) gave \( \hat{\beta}_0 = 34.17 \pm .87 \) and \( \hat{\beta}_1 = -.0574 \pm .0087 \), the numbers following "\( \pm \)" being the estimated standard errors. The true value of \( \beta_j \) lies within one estimated standard error of \( \hat{\beta}_j \) with probability about 67%, and within two standard errors with probability about 95%.

The simplest of all regression problems is that of estimating a single common mean: \( y_i = \mu + e_i \) for \( i = 1, 2, \ldots, n \) in (2.2), (2.3). In this case the least squares estimate is \( \hat{\mu} = \bar{y} \), the sample average, with estimated standard error given by the famous formula \[ \left[ \sum_i (y_i - \bar{y})^2 / n * (n-1) \right]^{1/2} \]. For instance, the \( n = 9 \) numbers in the first column of Table 1 gave \( \hat{\mu} = 23.1 \pm 1.8 \).

3. Model Checking and Selection.

The simple linear regression (1.2) for the hormone data is convenient, but is it correct? The usual way to check a simple model is to challenge it with a bigger one. For instance we might replace (1.2) with the quadratic model

\[
\mu(t) = \beta_0 + \beta_1 t + \beta_2 t^2,
\]

(3.1)
estimate the parameters \((\beta_0, \beta_1, \beta_2)\) by least squares, and ask if

\[
\hat{\mu}_i = \hat{\beta}_0 + \hat{\beta}_1 t_i + \hat{\beta}_2 t_i^2 \quad i = 1, 2, \ldots, 27
\]

(3.2)
gives a significantly better fit to the data in Figure 1.

Notice that the "linear" in linear model refers to the coefficients \( \beta_j \), so that (3.1) is linear even though it involves a quadratic term in the explanatory variable "days worn". The vector \( x_i \) in (2.4) is now \((1, t_i, t_i^2)\), the matrix \( X \) is \( 27 \times 3 \), and \( \mathcal{L} \) in Figure 2 is now a 3-dimensional subspace of \( \mathbb{R}^{27} \), say

\[
\mathcal{L}(2) = \{ u: u_i = b_0 + b_1 t_i + b_2 t_i^2, \ i = 1, 2, \ldots, 27 \},
\]

(3.3)
where $b = (b_0, b_1, b_2)$ can be any point in $\mathbb{R}^3$. The name $\mathcal{L}(2)$ indicates that $\mathcal{L}$ is the space of all possible quadratic regressions in $t_i$. The space of linear regressions in $t_i$,

$$\mathcal{L}(1) = \{u: u_i = b_0 + b_1 t_i, i = 1, 2, \ldots, 27\},$$

is a 2-dimensional linear subspace contained in $\mathcal{L}(2)$.

The spaces $\mathcal{L}(1)$ and $\mathcal{L}(2)$ are indicated in Figure 3, along with the corresponding least-squares points $\hat{y}(1)$ and $\hat{y}(2)$. Because $\mathcal{L}(1) \subset \mathcal{L}(2)$ we must have

$$\|y - \hat{y}(2)\|^2 \leq \|y - \hat{y}(1)\|^2.$$  \hspace{1cm} (3.5)

In other words, increasing the explanatory space $\mathcal{L}$ decreases the residual vector $r = y - \hat{y}$. As we shall see this does not mean that big models are always better than small ones.

Figure 3 indicates the squared residual lengths $\|y - \hat{y}(1)\|^2 = 141.39$ and $\|y - \hat{y}(2)\|^2 = 129.02$ for the IUD data, from which Pythagoras' theorem gives $\|\hat{y}(2) - \hat{y}(1)\|^2 = 12.37$. Does a decrease of 12.37 in squared residual length indicate a genuine advantage to the bigger model? Here is the traditional answer to that question:

(i) $y - \hat{y}(2)$ takes its value in the 24-dimensional subspace of $\mathbb{R}^{27}$ orthogonal to the three-dimensional space $\mathcal{L}(2)$.

(ii) $\hat{y}(2) - \hat{y}(1)$ takes its value in a 1-dimensional space, the portion of $\mathcal{L}(2)$ orthogonal to $\mathcal{L}(1)$.

(iii) If there is no true quadratic effect, that is if the true mean vector $\mu$ lies in $\mathcal{L}(1)$, then we expect the ratio

$$F = \frac{\|\hat{y}(2) - \hat{y}(1)\|^2}{\|y - \hat{y}(2)\|^2 / 24}$$  \hspace{1cm} (3.6)
Figure 3. On the left is a schematic diagram comparing linear and quadratic models for the hormone data of Section 1. Because $\mathcal{X}(1) \subset \mathcal{X}(2)$, the residual vector must be larger for $\mathcal{X}(1)$. The squared lengths $\|y - \hat{\mu}(1)\|^2$ and $\|y - \hat{\mu}(2)\|^2$ are shown at right. Traditional theory indicates that there is no strong reason to believe that quadratic regression is superior to linear regression for the hormone data.

to approximately equal one (since then there is nothing special about $\hat{\mu}(2)-\hat{\mu}(1)$ compared to the other 24 components of $y - \hat{\mu}(1) = [\hat{\mu}(2) - \hat{\mu}(1)] + [y - \hat{\mu}(2)]$.)

(iv) The observed value of $F$ in this case is $12.37/(129.02/24) = 2.30$, so we have to decide whether or not $2.30$ "approximately equals one".

(v) The theoretical distribution of $F$ can be computed for model (2.6), (2.7); assuming that the true quadratic effect is zero,

$$\text{Prob}\{F > 2.30\} = .14.$$  \hspace{1cm} (3.7)

(vi) An "achieving significance level" of $.14$ is not considered significant evidence for the existence of a genuine quadratic component in $y$. We expect bigger values of $F$, and so smaller achieving significance levels, if $y$ is genuinely quadratic. The conventional borderline for significance is $.05$, in this case $F \geq 4.26$. To state things in traditional language, we accept the null hypothesis
that the linear model (1.2) is correct. What we really mean is that there is no convincing evidence that a quadratic model is superior.

The letter F honors Sir Ronald Fisher, who invented this theory in the 1920's. Most of our previous results, for example (2.10), do not require \( \varepsilon \) in (2.6), (2.7) to be normally distributed, but calculations like (3.6) explicitly require normality. Fisher's theory for fitting linear models enjoys enormous and deserved popularity. It has been used literally millions of times. In experienced hands the linear model (2.6), (2.7) combined with steps (i)-(vi) help guide the scientist toward an insightful analysis of noisy data.

This doesn't mean that the recipe is perfect. Modern developments in regression theory, which are the main topic of this paper, have developed in response to certain of its deficiencies. The combination of linear models, least squares fitting, and normal error distributions leads to a mathematically elegant theory, but not necessarily a theory which is appropriate to every situation. Modern computational equipment allows us to deviate from the path of greatest elegance, and still produce numerical answers for real problems. It has also led to some interesting new theoretical results. Both theory and practice are discussed in the sections which follow.

4. Least Absolute Deviations and Robust Regression.

Our first "new development" in regression theory actually predates the method of least squares. Laplace, in the late 1700's, suggested fitting straight lines to noisy data according to the principle of Least Absolute Deviations (LAD): choosing \((\hat{\beta}_0, \hat{\beta}_1)\) to minimize \( \sum_{i=1}^{n} |y_i - (\hat{\beta}_0 + \hat{\beta}_1 t_i)| \). Applied to the data in Figure 1, LAD gives \((\hat{\beta}_0, \hat{\beta}_1) = (34.20, -0.0587)\), almost the same as the least squares solution \((34.17, -0.0574)\).

Least Absolute Deviations and Least Squares don't always agree so nicely. Table 2 shows the results of a small experiment in cell survival. A total of \( n = 14 \) plates of cells were exposed to various doses of radiation. The observed response
was the proportion of cells which survived the radiation exposure. Survival properties are not easy to measure, and the investigator expressed some doubt about the recorded response for plate 13.

<table>
<thead>
<tr>
<th>Plate #</th>
<th>Radiation Dose (Rads/100)</th>
<th>Survival Proportion</th>
<th>Log Survival Proportion</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>1.175</td>
<td>.44</td>
<td>-.821</td>
</tr>
<tr>
<td>2.</td>
<td>1.175</td>
<td>.55</td>
<td>-.598</td>
</tr>
<tr>
<td>3.</td>
<td>2.35</td>
<td>.16</td>
<td>-1.833</td>
</tr>
<tr>
<td>4.</td>
<td>2.35</td>
<td>.13</td>
<td>-2.040</td>
</tr>
<tr>
<td>5.</td>
<td>4.70</td>
<td>.0400</td>
<td>-3.219</td>
</tr>
<tr>
<td>6.</td>
<td>4.70</td>
<td>.0196</td>
<td>-3.932</td>
</tr>
<tr>
<td>7.</td>
<td>4.70</td>
<td>.0612</td>
<td>-2.794</td>
</tr>
<tr>
<td>8.</td>
<td>7.05</td>
<td>.0050</td>
<td>-5.298</td>
</tr>
<tr>
<td>9.</td>
<td>7.05</td>
<td>.0032</td>
<td>-5.745</td>
</tr>
<tr>
<td>10.</td>
<td>9.40</td>
<td>.00110</td>
<td>-6.812</td>
</tr>
<tr>
<td>11.</td>
<td>9.40</td>
<td>.00015</td>
<td>-8.805</td>
</tr>
<tr>
<td>12.</td>
<td>9.40</td>
<td>.00019</td>
<td>-8.568</td>
</tr>
</tbody>
</table>

Table 3. Cell Survival Data. Fourteen plates of cells were exposed to different levels of radiation. The observed response was the proportion of cells which survived the radiation exposure. The measurement on plate 13 was considered somewhat uncertain by the investigator. A plot of the data appears in Figure 4.

Based on theoretical models of radiation damage, the response variable \( y = \log \text{ survival proportion} \) is expected to be a low order polynomial function of the radiation dose \( t \). Letting \( \mu(t) \) equal the expected value of \( y \) at dose \( t \), a quadratic regression

\[
\mu(t) = \beta_1 t + \beta_2 t^2
\]

(4.1)
was proposed. In this case there is no constant term $\beta_0$ in the regression because we know that zero radiation gives 100% survival, so $\mu(0) = \log(1.00) = 0$.

![Figure 4](image-url)

Figure 4. The left panel plots the log survival proportion (vertical axis) versus the radiation dose in RADS/100 (horizontal axis). Least Squares regression indicated a significant quadratic component; LAD did not. The LS regression is sensitive to the questionable result for plate 13. This is illustrated in the right panel, where a gross error has been deliberately added to plate 13's response, greatly changing the LS but not the LAD regression.

It was of considerable interest to the investigator whether or not the quadratic term in (4.1) was necessary, in other words whether or not a linear model $\mu(t) = \beta_1 t$ was sufficient.

Both Least Squares and Least Absolute Deviations were used to fit model (4.1) to the log survival data. In this case the results disagree. Table 4 shows the fitted coefficients along with their estimated standard errors. For the LS fit, $\hat{\beta}_2$ is 2.4 standard errors from zero. The $F$ value is then $(2.4)^2 = 5.76$, with an achieved significance level of .03, indicating that the quadratic effect is genuine. [The $F$ statistic described in Section 3 is the square of $\hat{\beta}_2$'s distance from 0, measured in units of estimated standard error.]

The LAD estimate of $\hat{\beta}_2$ is only one standard error away from 0, giving $F = 1$, decidedly non-significant evidence for a quadratic effect. Which result
<table>
<thead>
<tr>
<th>Least Squares</th>
<th>Least Squares (without plate 13)</th>
<th>Least Absolute Deviations</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \hat{\beta}_1 ):</td>
<td>(-.970 \pm .115)</td>
<td>(-.862 \pm .094)</td>
</tr>
<tr>
<td>( \hat{\beta}_2 ):</td>
<td>(.024 \pm .010)</td>
<td>(.009 \pm .009)</td>
</tr>
</tbody>
</table>

Table 4. Least Squares and Least Absolute Deviations fitted values for \((\hat{\beta}_1, \hat{\beta}_2)\), cell survival data model (4.1), and their estimated standard errors. The quadratic coefficient is significantly non-zero in the LS fit, but not in the LAD fit. If the data from plate 13 is discarded, the quadratic coefficient for the LS fit is reduced to non-significance. The standard errors for LAD are obtained by bootstrap methods, as explained in Section 5.

should we believe? In this case there is a strong argument in favor of the LAD result: when the questionable data from plate 13 is disregarded, the least-squares estimate of \( \hat{\beta}_2 \) falls into line with the LAD estimate. (The LAD estimates are the same with or without plate 13!)

The point here is not that Least Absolute Deviations is a superior method to Least Squares. As a matter of fact if model (2.6), (2.7) is correct, then LS is both theoretically and actually the best possible fitting criterion. Our real point is that LS is more sensitive to occasional discrepant observations than is LAD. This is shown in the right panel of Figure 4. The questionable point has been moved to a grossly discrepant position. The effect on the fitted LS curve is enormous, but the LAD curve stays exactly the same.

Why not always use LAD, just to be on the safe side? Mainly because there is a considerable price to be paid in terms of the accuracy of the fitted curves if model (2.6), (2.7) is close to being correct. Typically using LAD when we should be using LS increases the variability of the fitted coefficients by about 25%.

Robust regression is a theory of curve-fitting which attempts to compromise between the perfect efficiency of Least Squares for the normal model (2.6), (2.7), and the resistance to occasional discrepant observations of LAD. The theory
begins with some function $Q[r]$ which measures the "distance" from 0 of a number $r$, for example

$$Q[r] = r^2$$  \hspace{1cm} (4.2)

or

$$Q[r] = |r|.$$  \hspace{1cm} (4.3)

We suppose we are trying to predict observations $y_i$ by fitting value $\mu_i(b) = x_i b$ as in (2.4); that $r_i(b)$ is the residual from the $i$th prediction,

$$r_i(b) = y_i - x_i b \hspace{0.5cm} i = 1,2,\ldots,n;$$  \hspace{1cm} (4.4)

and that we choose the unknown $p \times 1$ vector $b$ to minimize

$$\sum_{i=1}^{n} Q[r_i(b)] = \sum_{i=1}^{n} Q[y_i - x_i b].$$  \hspace{1cm} (4.5)

In other words that we are using "Least Q" as a fitting criterion.

Choice (4.2) gives Least Squares while (4.3) gives Least Absolute Deviations. A compromise candidate for $Q$, which makes the fitting procedure less sensitive than LS to discrepant values but more efficient than LAD for model (2.6), (2.7) is

$$Q[r] = |r|^{3/2}.$$  \hspace{1cm} (4.6)


Define $q[r] = Q'[r]$, the derivative of $Q$. At the value $\hat{b}$ of $b$ which minimizes (4.5) we must have

$$\left. \frac{3}{3b} \sum_{i=1}^{n} Q[r_i(b)] \right|_{b=\hat{b}} = 0 \hspace{0.5cm} j = 1,2,\ldots,p.$$  \hspace{1cm} (4.7)

Then (4.4) gives
\[
X' q(b) \mid_{b=\hat{b}} = \mathbb{1}, \quad q(b) \equiv (q[r_1(b)], q[r_2(b)], \ldots, q[r_n(b)])'
\]  
(4.8)

as implicit equations for solving the "Least Q" fitting problem.

Notice that for \( Q[r] = r^2 \) we have \( q[r] = 2r \), so (4.8) becomes

\[
2X' r(b) \mid_{b=\hat{b}} = 2X'(y-x_0\hat{b}) = 0
\]
(4.9)
or

\[
X'X\hat{b} = X'y \iff \hat{b} = (X'X)^{-1}X'y,
\]
(4.10)

which are just the normal equations (2.9). The computationally most difficult choice of \( Q \) is \( Q[r] = |r| \), \( q[r] = \text{sign}(r) \), for which equations (4.8) take the discontinuous form

\[
X'(\text{sign}(r_1(b)), \text{sign}(r_2(b)), \ldots, \text{sign}(r_n(b)))' \mid_{b=\hat{b}} = 0.
\]
(4.11)

Finding the LAD solution for \( \hat{b} \) is an exercise in linear programming. Laplace was ahead of his time in terms of his computational requirements!

Choice (4.6) for \( Q \) doesn't lead to a closed form solution for \( \hat{b} \), (only \( Q[r] = r^2 \) does), but the solution of (4.8) is computationally easier than for \( Q[r] = |r| \). Standard minimization routines, for example Newton-Raphson, converge quickly in most problems.

The fact is that with modern computational equipment it is practical to use almost any discrepancy function \( Q \) we want. The rapidly developing theory of robust regression is a direct response to this new-found freedom.


Traditional Least-Squares regression theory provides more than just an estimate for the unknown parameter vector \( \hat{b} \). We are also told the accuracy of \( \hat{b} \),
usually in the form of estimated standard errors for the components of \( \hat{\beta} \). The bootstrap is a computer-based method for assessing the accuracy of \( \hat{\beta} \) when the estimation procedure is more complicated than Least Squares. The standard errors for the LAD estimates in Table 4 were obtained by bootstrap methods described in this section. Efron and Tibshirani (1986) give an extensive review of the bootstrap, including examples from many different areas of application.

Figure 5 describes the bootstrap in very general terms, which include the regression situation. We have an observed data set \( \tilde{y} \) which has been generated from an unknown probability model \( P \). A statistic of interest \( \hat{\beta} \) has been calculated from \( \tilde{y} \), and we wish to assess the accuracy of \( \hat{\beta} \). "Accuracy" here refers to how little or how much \( \hat{\beta} \) would fluctuate if we could observe independently generated replicates of \( \hat{\beta} \), say \( \hat{\beta}(1), \hat{\beta}(2), \ldots \), which of course we can't do in most real situations.

![Diagram](image)

**Figure 5.** A general diagram of the bootstrap method for assessing statistical accuracy. An unknown probability model \( P \) has given us some observed data \( \tilde{y} \), from which we calculated a statistic of interest \( \hat{\beta} \). We wish to know the accuracy of \( \hat{\beta} \). The bootstrap method assumes that \( \tilde{y} \) allows us to estimate the entire model \( P \), say by \( \hat{P} \). "Bootstrap" data sets \( \tilde{y}^* \) are repeatedly generated from \( \hat{P} \) by Monte Carlo sampling, and for each \( \tilde{y}^* \) the statistic of interest \( \hat{\beta}^* \) is recalculated. The accuracy of the original estimate \( \hat{\beta} \) is indicated by the variability of the \( \hat{\beta}^* \)'s.
What we can do, in most situations, is to use \( \hat{y} \) to estimate the entire probability model \( P \), say by \( \hat{P} \). For example in the traditional regression model (2.6), (2.7) the only unknowns are \( (\beta, \sigma^2) \), which can be estimated by \( (\hat{\beta}, \hat{\sigma}^2) \) as in (2.9), (2.12); in this case \( P = (\beta, \sigma^2) \) and \( \hat{P} = (\hat{\beta}, \hat{\sigma}^2) \).

The basic idea of the bootstrap is simple: use Monte Carlo methods to repeatedly simulate "bootstrap" data sets \( \hat{y}^* \) from \( \hat{P} \); recalculate \( \hat{\beta}^* \) for each \( \hat{y}^* \); and observe how little or how much variability there is in the sequence \( \hat{\beta}^*(1), \hat{\beta}^*(2), \ldots \).

In the traditional regression model (2.6), (2.7), the steps \( \hat{P} \rightarrow \hat{y}^* \rightarrow \hat{\beta}^* \) in Figure 5 are carried out as follows:

\[
y^* = X\hat{\beta} + \hat{\epsilon}^* \quad \text{where} \quad \hat{\epsilon}^* \sim N(0, \hat{\sigma}^2 I), \quad \text{and} \quad \hat{\beta}^* = G^{-1}X'y^*. \tag{5.1}
\]

Notice that \( \hat{\beta} \) and \( \hat{\sigma} \) are fixed quantities in (5.1). The Monte Carlo part of (5.1) consists of drawing the \( n \) independent errors \( \hat{\epsilon}_i^* \sim N(0, \hat{\sigma}^2) \). We independently repeat (5.1) some large number "B" times, calculate \( \hat{\beta}^*(1), \hat{\beta}^*(2), \ldots, \hat{\beta}^*(B) \), and compute the covariance matrix of the \( \hat{\beta}^* \)'s. An easy theoretical calculation shows that as \( B \rightarrow \infty \), this bootstrap covariance matrix always converges to \( \hat{\sigma}^2 \hat{G}^{-1} \), as in (2.10), (2.13). In other words, the bootstrap method shown in Figure 5 gives the usual assessment of accuracy for \( \hat{\beta} \) in situation (2.6), (2.7).

This is almost the only case the which the bootstrap results can be obtained theoretically. However the bootstrap calculations indicated in Figure 5 can be carried out numerically, by pure computational force, for complicated situations where theoretical analysis is hopeless. Here is how the standard errors for the Least Absolute Deviations estimates \( (\hat{\beta}_1, \hat{\beta}_2) \) in Table 4 were obtained:

(i) The unknown probability model \( P \) was taken to be

\[
y_i = \mu_i + \epsilon_i, \quad \mu_i = \beta_1 t_i + \beta_2 \epsilon_i^2 \quad (i=1,2,\ldots,14) \tag{5.2}
\]
as in (4.1), with \(e_1, e_2, \ldots, e_{14}\) independent errors drawn from some unknown probability distribution \(F\).

(ii) The LAD coefficients \((\hat{\beta}_1, \hat{\beta}_2) = (-.813, .009)\) gave estimated mean values \(\hat{\mu}_i = \hat{\beta}_1 t_{1i} + \hat{\beta}_2 t_{2i}^2\) and estimated errors (residuals)

\[
\tilde{e}_i = y_i - \hat{\mu}_i \quad (i=1,2,\ldots,14),
\]

from which was obtained an estimated version of \(F\),

\[
\hat{F} \text{: probability } \frac{1}{14} \text{ on } \tilde{e}_i, \quad i = 1,2,\ldots,14.
\]

(iii) Bootstrap data vectors \(y^*\) were generated according to

\[
y_i^* = \hat{\mu}_i + e_i^* \quad (i=1,2,\ldots,14),
\]

with \(e_1^*, e_2^*, \ldots, e_{14}^*\) independent observations drawn from \(\hat{F}\) by Monte Carlo.

(iv) Having generated \(y^*\), the bootstrap LAD estimates \((\hat{\beta}_1^*, \hat{\beta}_2^*)\) were calculated exactly as in Section 4: as the minimizers of \(\sum_{i=1}^{14} |y_i^* - (\hat{\beta}_1^* t_{1i} + \hat{\beta}_2^* t_{2i}^2)|\).

(v) This whole process was repeated \(B = 40\) times. The observed variability of the vectors \(\hat{\beta}^*(1), \hat{\beta}^*(2), \ldots, \hat{\beta}^*(40)\) gave the standard errors reported in Table 4. The first 20 of the 40 bootstrap replicates \(\hat{\beta}^*\) are shown in Table 5.

Stopping the bootstrap process at \(B = 40\) replications is an arbitrary choice, and may seem premature to the reader. In fact \(B = 40\) is almost as good as \(B = \infty\) for estimating the standard errors of \(\hat{\beta}_1\) and \(\hat{\beta}_2\), according to Section 9 of Efron and Tibshirani (1986).

The bootstrap is not the only technique for estimating standard errors in complicated situations, see Efron (1982). However all of the techniques ultimately rely on the simple idea illustrated in Figure 5, and none of them would be feasible without the massive power of modern computational equipment.
<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\hat{\beta}_1^*$</td>
<td>-.796</td>
<td>-.814</td>
<td>-.838</td>
<td>-.809</td>
<td>-.795</td>
<td>-.915</td>
<td>-.807</td>
<td>-.889</td>
<td>-.807</td>
<td>-.804</td>
</tr>
<tr>
<td>$\hat{\beta}_2^*$</td>
<td>.0077</td>
<td>.0088</td>
<td>.0126</td>
<td>.0070</td>
<td>.0075</td>
<td>.0095</td>
<td>.0075</td>
<td>.0263</td>
<td>.0075</td>
<td>.0081</td>
</tr>
<tr>
<td>11</td>
<td></td>
<td>12</td>
<td></td>
<td>13</td>
<td></td>
<td>14</td>
<td></td>
<td>15</td>
<td></td>
<td>16</td>
</tr>
<tr>
<td></td>
<td></td>
<td>17</td>
<td></td>
<td>18</td>
<td></td>
<td>19</td>
<td></td>
<td>20</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\hat{\beta}_1^*$</td>
<td>-.814</td>
<td>-.807</td>
<td>-.814</td>
<td>-.827</td>
<td>-.771</td>
<td>-.706</td>
<td>-.914</td>
<td>-.901</td>
<td>-.949</td>
<td>-.703</td>
</tr>
<tr>
<td>$\hat{\beta}_2^*$</td>
<td>.0088</td>
<td>.0061</td>
<td>.0088</td>
<td>.0116</td>
<td>.0042</td>
<td>-.0086</td>
<td>.0201</td>
<td>.0273</td>
<td>.0238</td>
<td>.0010</td>
</tr>
</tbody>
</table>

Table 5. The first 20 bootstrap replications of the LAD estimates ($\hat{\beta}_1^*$, $\hat{\beta}_2^*$) for the cell survival data, Table 4.


Our examples so far all make an important assumption: that the regression function $\mu(t)$ can be expressed by a simple formula, for example $\mu(t) = \beta_0 + \beta_1 t + \beta_2 t^2$, which holds true across the entire observed range of the explanatory variable $t$. **Local smoothing** avoids this assumption. This section and the next discuss the main ideas involved.

Figure 6 displays some microbiological data due to Dr. T. A. Roberts. [The same figure appears in an excellent review article of local smoothing based on spline functions, Silverman (1985).] There are $n = 33$ points $(t_i, y_i)$, equally spaced at half-hour intervals along the $t$ axis. A smooth curve has been fit to the data using "supersmoother", a local smoothing algorithm developed by J. Friedman and W. Stuetzle. The details of supersmoother are discussed in Friedman (1984).

The main ideas involved are simple to describe:

(i) An integer $L < n/2$ which controls the amount of local smoothing is selected; $L$ can be chosen directly by the user or, as was actually done in Figure 6, automatically by the algorithm using a clever method called **cross-validation**, described later in this section.
Figure 6. Microbiological data, collected by Dr. T. Roberts. Vertical axis is the log (base 10) of population count per millilitre of the organism Staphylococcus aureus; horizontal axis is the incubation time of the broth containing the organism. A smooth curve has been fit to the data by the "supersmooth" method due to J. Friedman and W. Stuetzle. This is a local smoothing technique which does not involve global formulas for the regression function $\mu(t)$.

(ii) For convenient notation suppose that the values on the horizontal axis are listed in increasing order, $t_1 \leq t_2 \leq t_3, \ldots, \leq t_n$. The fitting procedure at each $(t_i, y_i)$ considers only a neighborhood $N(i,L)$ consisting of $2L+1$ points,

$$N(i,L) \equiv \{(t_j, y_j); i-L \leq j \leq i+L\}.$$  \hfill (6.1)

[For $i \leq L$ or $i > n-L$, definition (6.1) is modified in an obvious way so that $N(i,L)$ still contains $2L+1$ points.]
(iii) A simple linear regression \( \hat{\mu}^{(1)}(t) = \hat{\beta}_0^{(1)} + \hat{\beta}_1^{(1)} t \) is fit to the data in \( N(i,L) \) by the method of Least Squares.

(iv) The fitted value \( \hat{\mu}_i \) is set equal to \( \hat{\mu}^{(1)}(t_i) \).

(v) This process is repeated for each \( i = 1, 2, \ldots, n \). The smooth "curve" \( (\hat{\mu}_1, \hat{\mu}_2, \ldots, \hat{\mu}_n) \) is filled in by linear interpolation, as in Figure 6.

If \( L = (n-1)/2 \) then \( N(i,L) \) equals all of the data, and the smooth curve \( \{\hat{\mu}_1, \hat{\mu}_2, \ldots, \hat{\mu}_n\} \) is nothing more than a simple linear regression fit to all \( n \) points, as in (1.3). This is very smooth indeed, but may not fit the data, as in the case of Figure 6. At the opposite extreme the choice \( L = 0 \) results in \( \hat{\mu}_i = y_i \). This fits the data perfectly, but at the expense of being unacceptably unsmooth. Intermediate values of \( L \) are needed if we want sensibly smooth curves which fit the data well, as in Figure 6.

How should we choose \( L \)? For any given \( L \), let \( \hat{\mu}_i(L) \) be the prediction for the \( i \)th case, and let \( r_i(L) \) equal the residual \( y_i - \hat{\mu}_i(L) \). An overall measure of "prediction error" is

\[
D(L) = \sum_{i=1}^{n} r_i(L)^2 ,
\]

which we could minimize over the choice of \( L \). Unfortunately this always leads to the choice \( L = 0 \), since \( \hat{\mu}_i(0) = y_i \) and \( r_i(0) = 0 \). The trouble here is that each point \( (t_i, y_i) \) is being used twice: first as part of data leading to the prediction \( \hat{\mu}_i(L) \), and second to evaluate its own prediction error, in (6.2).

What we would really like to know is the value of \( L \) which minimizes the prediction error for some new data, not involved in the construction of \( \hat{\mu}_i(L) \). Cross-validation is a way of estimating such an \( L \) when no new data is available.

Let \( \hat{\mu}^{(1)}(L) \) be the prediction for the \( i \)th case based on all the data except \( (t_i, y_i) \). In the case of supersmoother this means deleting \( (t_i, y_i) \) from \( N(i,L) \)

24
in (6.1); fitting a simple linear regression to the remaining 2L points; and reading off the height of this regression at $t = t_i$. The cross-validated residual is defined to be $r_{(i)}(L) \equiv y_i - \hat{\mu}_{(i)}(L)$, giving the cross-validated prediction error

$$D^\dagger(L) \equiv \sum_{i=1}^{n} r_{(i)}(L)^2.$$  
(6.3)

Because $(t_i, y_i)$ is not used in the construction of $\hat{\mu}_{(i)}(L)$, $D^\dagger(L)$ is a nearly unbiased estimate of how well $(\hat{\mu}_1(L), \hat{\mu}_2(L), \ldots, \hat{\mu}_n(L))$ would predict a new set of data. The cross-validation criterion chooses $L$ to minimize $D^\dagger(L)$.

For the data in Figure 6, cross-validation led to the choice $L = 3$, $2L+1 = 7$, resulting in the smooth fit shown. [Actually supersmoother incorporates some additional smoothing techniques, which were unimportant for the not very noisy data of Figure 6.]

Like the bootstrap (to which it is closely related, see Efron (1982)), cross-validation can be carried out by sheer computing power for any collection of possible fitting rules. It happens to be particularly easy to implement cross-validation for rules which are linear in the data, say

$$\hat{\mu}(L) \sim M(L)y$$  
(6.4)

where $M(L)$ is an $n \times n$ matrix which may depend on $L$ and the predictors $\tilde{t}$, but not on the data $\tilde{y}$. An interesting argument which appears in Craven and Wahba (1979) shows that in this case

$$r_{(i)}(L) = r_i(L)/[1-M_{ii}(L)]$$  
(6.5)

so that the cross-validated residuals are obtained directly from the ordinary residuals $r_i(L)$, and no data deletion is necessary.
Smoothing by local linear regressions, as in steps (i)-(v) above, is of form (6.4). In this case (6.5) gives

$$D^+(L) = \frac{n}{i=1} \left\{ \frac{[y_i - \hat{\mu}_i(L)]}{[1 - \frac{1}{2L+1} - \frac{(t_i - \bar{t}_i)^2}{V_i}]} \right\}^2$$

(6.6)

where

$$\bar{t}_i = \sum_{j \in N(i, L)} t_j / (2L + 1), \quad V_i = \sum_{j \in N(i, L)} (t_j - \bar{t}_i)^2.$$  

(6.7)

The choice $L = 3$ minimized (6.6) for the microbiological data.

How accurate is the smooth curve in Figure 6? The bootstrap was used to obtain an answer. The probability model $P$ in Figure 5 was taken to be

$$y_i = \mu_i + \epsilon_i, \quad \epsilon_i \sim N(0, \sigma_i^2) \text{ independently for } i = 1, 2, \ldots, n.$$  

(6.8)

Both $\mu_i$ and $\sigma_i^2$ are unknown in (6.8), but assumed to vary smoothly as a function of time $t_i$. The step $y^* \sim \hat{P}$ in Figure 5 is, in this case, the calculation of $(\hat{\mu}_1, \hat{\mu}_2, \ldots, \hat{\mu}_n)$ from the data using supersmoother (including the cross-validated selection of $L$).

The bootstrap probability model $\hat{P}$ was taken to be

$$y_i^* = \hat{\mu}_i + \epsilon_i^*, \quad \epsilon_i^* \sim N(0, \hat{\sigma}_i^2) \text{ independently for } i = 1, 2, \ldots, n.$$  

(6.9)

The estimates $\hat{\sigma}_i^2$ were obtained by applying supersmoother to the squared residuals $[y_i - \hat{\mu}_i]^2$. Here are some of the values of $\hat{\sigma}_i$:

<table>
<thead>
<tr>
<th>time $t_i$:</th>
<th>0</th>
<th>4</th>
<th>8</th>
<th>12</th>
<th>16</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\hat{\sigma}_i$:</td>
<td>.054</td>
<td>.091</td>
<td>.239</td>
<td>.227</td>
<td>.058</td>
</tr>
</tbody>
</table>

(6.10)

Applying supersmoother (including cross-validation) to $y^*$ yields a bootstrap smooth curve $(\hat{\mu}_1^*, \hat{\mu}_2^*, \ldots, \hat{\mu}_n^*)$. This entire process was repeated $B = 100$ times, giving a good idea of both the variability and bias of the smoothing process. For
example at \( t = 4 \) hours, corresponding to index \( i = 9 \), the 100 values of \( \hat{\mu}_9^* \) had variability (square root of variance) equal \( .031 \). The mean value of the \( \hat{\mu}_9^* \)’s was \( 3.955 \), compared to \( \hat{\mu}_9 = 3.849 \). This may seem reasonably close, but in fact it indicates a substantial upward bias in \( \hat{\mu}_9 \),

\[
3.955 - 3.849 = 0.106 \quad (6.11)
\]
nearly three times the magnitude of the estimated standard error \( .031 \). The bias is quite evident in Figure 6, upward near \( t = 4 \) and downward near \( t = 10 \).

The center curve in Figure 7 was obtained by subtracting the estimated bias from each corresponding value \( \hat{\mu}(t) \) in Figure 6. Upper and lower bounds

\[
\text{center curve} \pm 1.65 \cdot \text{estimated standard error} \quad (6.12)
\]
are also indicated. Statistical theory indicates that the true value \( \mu(t) \) of the regression will lie inside this range about 90% of the time. We now have a quite reasonable estimate of the smooth curve, and also the accuracy of our estimate. The amount of computation going into Figure 7 is truly prodigious from Fisher’s 1920 point of view, but takes less than a minute on a modern minicomputer.

![Figure 7](image_url)

**Figure 7.** The smooth curve in Figure 6 has been corrected by subtracting the bootstrap estimate of bias from each original value \( \hat{\sigma}_1^* \) [the center curve]. The upper and lower curves are the corrected value plus or minus 1.65 times the bootstrap estimate of standard error. The true regression \( \mu(t) \) lies in the band with probability about \( .90 \), for any particular value of \( t \).

Regression analysis is at its most useful when there are many predictor variables at work, rather than just one as in our previous examples. In the whitecap data set, Table 6, there are three predictors: \( x(1), x(2), x(3) \) windspeed, air temperature, and the air-water temperature difference. The response variable \( y \) is minus the logarithm of the percentage whitecap coverage at a particular oceanic site. Data for \( n = 30 \) sites is given. Each column has been centered at zero by subtracting its mean for the 30 cases. [Table 6 is part of a larger data set analyzed by O'Muircheartaigh and Gaver (1985).]

<table>
<thead>
<tr>
<th></th>
<th>( x(1) )</th>
<th>( x(2) )</th>
<th>( x(3) )</th>
<th>( y )</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>-0.6</td>
<td>-2.62</td>
<td>-4.97</td>
<td>0.07</td>
</tr>
<tr>
<td>2</td>
<td>0.0</td>
<td>-3.42</td>
<td>-4.77</td>
<td>-1.05</td>
</tr>
<tr>
<td>3</td>
<td>1.4</td>
<td>-2.72</td>
<td>-3.77</td>
<td>-0.36</td>
</tr>
<tr>
<td>4</td>
<td>2.8</td>
<td>-2.92</td>
<td>-3.77</td>
<td>-2.27</td>
</tr>
<tr>
<td>5</td>
<td>6.0</td>
<td>-3.22</td>
<td>-2.87</td>
<td>-1.69</td>
</tr>
<tr>
<td>6</td>
<td>3.3</td>
<td>-3.42</td>
<td>-3.27</td>
<td>-2.17</td>
</tr>
<tr>
<td>7</td>
<td>5.1</td>
<td>-3.62</td>
<td>-2.57</td>
<td>-2.33</td>
</tr>
<tr>
<td>8</td>
<td>6.5</td>
<td>-3.72</td>
<td>-3.07</td>
<td>-3.49</td>
</tr>
<tr>
<td>9</td>
<td>8.7</td>
<td>-3.82</td>
<td>-1.07</td>
<td>-3.10</td>
</tr>
<tr>
<td>10</td>
<td>-4.1</td>
<td>-1.42</td>
<td>0.53</td>
<td>0.71</td>
</tr>
<tr>
<td>11</td>
<td>-2.8</td>
<td>-2.02</td>
<td>3.03</td>
<td>0.63</td>
</tr>
<tr>
<td>12</td>
<td>-3.5</td>
<td>-1.72</td>
<td>2.13</td>
<td>3.11</td>
</tr>
<tr>
<td>13</td>
<td>-4.4</td>
<td>-1.22</td>
<td>1.43</td>
<td>3.11</td>
</tr>
<tr>
<td>14</td>
<td>-1.4</td>
<td>-0.92</td>
<td>1.83</td>
<td>3.11</td>
</tr>
<tr>
<td>15</td>
<td>-1.5</td>
<td>-1.12</td>
<td>2.03</td>
<td>1.16</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>( x(1) )</th>
<th>( x(2) )</th>
<th>( x(3) )</th>
<th>( y )</th>
</tr>
</thead>
<tbody>
<tr>
<td>16</td>
<td>0.0</td>
<td>-1.12</td>
<td>2.93</td>
<td>0.55</td>
</tr>
<tr>
<td>17</td>
<td>-1.8</td>
<td>-1.12</td>
<td>1.83</td>
<td>0.91</td>
</tr>
<tr>
<td>18</td>
<td>0.1</td>
<td>-0.82</td>
<td>0.73</td>
<td>1.50</td>
</tr>
<tr>
<td>19</td>
<td>-3.2</td>
<td>3.58</td>
<td>2.23</td>
<td>1.50</td>
</tr>
<tr>
<td>20</td>
<td>-2.4</td>
<td>3.68</td>
<td>1.43</td>
<td>1.72</td>
</tr>
<tr>
<td>21</td>
<td>-1.7</td>
<td>3.88</td>
<td>0.63</td>
<td>-0.32</td>
</tr>
<tr>
<td>22</td>
<td>-1.8</td>
<td>3.68</td>
<td>0.43</td>
<td>0.71</td>
</tr>
<tr>
<td>23</td>
<td>-3.9</td>
<td>4.08</td>
<td>1.43</td>
<td>1.16</td>
</tr>
<tr>
<td>24</td>
<td>-1.7</td>
<td>4.18</td>
<td>1.93</td>
<td>0.22</td>
</tr>
<tr>
<td>25</td>
<td>-1.2</td>
<td>4.28</td>
<td>0.83</td>
<td>-0.82</td>
</tr>
<tr>
<td>26</td>
<td>-1.0</td>
<td>3.88</td>
<td>0.23</td>
<td>-1.26</td>
</tr>
<tr>
<td>27</td>
<td>1.8</td>
<td>2.98</td>
<td>1.73</td>
<td>-1.06</td>
</tr>
<tr>
<td>28</td>
<td>1.9</td>
<td>1.98</td>
<td>0.73</td>
<td>-0.98</td>
</tr>
<tr>
<td>29</td>
<td>0.9</td>
<td>2.18</td>
<td>0.93</td>
<td>-1.41</td>
</tr>
<tr>
<td>30</td>
<td>-1.4</td>
<td>2.48</td>
<td>1.23</td>
<td>2.01</td>
</tr>
</tbody>
</table>

Table 6. Whitecap data. Three meteorological variates \( x(1), x(2), x(3) \) [windspeed, air temperature, and air-water temp. difference] are available for predicting \( y \) [\(-\log \) whitecap coverage.] All four variates have been centered at zero by subtracting their respective means.

The linear regression model

\[
y_i = \mu_i + e_i \\
\mu_i = \beta_1 x_{i(1)} + \beta_2 x_{i(2)} + \beta_3 x_{i(3)}
\]  \hspace{1cm} (7.1)

gave least squares solution
\( (\hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3) = (-.428, -.182, .251) \), \( (7.2) \)

with estimated standard errors .065, .071, .094 respectively. No intercept term \( \beta_0 \) is needed in (7.1) because the data is centered at zero. Centering does not change the fitted values \( (\hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3) \), but makes \( \hat{\beta}_0 = 0 \), so it doesn't matter whether or not \( \beta_0 \) is included in the regression. Omitting \( \beta_0 \) simplifies the description of projection pursuit.

We can think of \( \hat{\beta} = (\hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3) \) as comprised of a unit direction vector \( \hat{\alpha} \equiv \hat{\beta}/\|\hat{\beta}\| \), and a slope \( \hat{\phi} \equiv \|\hat{\alpha}\| \), so that

\[
\hat{\beta} = \hat{\phi} \hat{\alpha} \quad \text{where} \quad \|\hat{\alpha}\| = 1
\] \( (7.3) \)

The least squares estimate \( \hat{\mu}_1 \) equals \( x_i \hat{\beta} = \hat{\phi} x_i \hat{\alpha} \), or

\[
\hat{\mu}_1 = \hat{\phi} t_i \quad \text{where} \quad t_i \equiv x_i \hat{\alpha}.
\] \( (7.4) \)

For the whitecap data writing (7.2) in the form \( \hat{\beta} = \hat{\phi} \hat{\alpha} \) results in

\[
\hat{\alpha} = (-.810, -.345, .474), \quad \hat{\phi} = .528.
\] \( (7.5) \)

This is illustrated in the left panel of Figure 8: the horizontal axis is \( t_i = x_i \hat{\alpha} \); the straight line shows the LS values \( \hat{\mu}_1 = \hat{\phi} t_i \); and the 30 data points \( (t_i, y_i) \) are indicated by '+'s. The advantage of formulation (7.4) is that we can draw simple scatterplots of the data \( (t_i, y_i) \). This is possible because we have reduced the 3-dimensional predictor \( x_i \) down to one dimension, \( t_i = x_i \hat{\alpha} \).

The left side of Figure 8 suggests an obvious next step: why not fit the data points \( (t_i, y_i) \) with something more ambitious than a straight line? Projection pursuit regression uses cross-validated local smoothing, as described in Section 6.

Here is an outline of the projection pursuit algorithm, taken from the original article by Friedman and Stuetzle (1981): (i) choose a direction \( \alpha \); (ii) reduce
Figure 8. The horizontal axis of the left panel is $t_i = x_i \hat{\alpha}$, the projection of $x_i$ along the direction $\hat{\alpha} = \beta / ||\beta||$ of the LS vector $\beta$. The straight line indicates the LS values $\hat{\mu}_i = ||\beta|| t_i$. The data $(t_i, y_i)$ is indicated by '+'s. In the right panel the horizontal axis is $x_i \tilde{\alpha}$, where $\tilde{\alpha}$ is the optimal direction selected by the projection pursuit algorithm. The projection pursuit regression function is indicated by the bumpy line.

the data to $\{(t_i(\alpha), y_i) : i = 1, \ldots, n\}$, where $t_i(\alpha) = x_i \alpha$; (iii) fit a smooth curve $\hat{f}_\alpha(t)$ to this data by cross-validated smoothing; (iv) compute the residual prediction error $D(\alpha) = \Sigma[y_i - \hat{f}_\alpha(t_i)]^2$; (v) repeat steps (i)-(iv) for different choices of the direction vector $\alpha$, using standard methods of optimization to find the best direction $\tilde{\alpha}$, the direction with the minimum value $D(\tilde{\alpha})$.

The right side of Figure 8 shows the application of projection pursuit to the whitecap data. The optimum direction was

$$\tilde{\alpha} = (-.856, -.373, .358), \quad (7.6)$$

not much different from (7.5). However the fitted curve $\hat{f}_\alpha(t)$ has an interesting-looking bump, which might be important in the physical interpretation of the results.
A small bootstrap analysis, \( B = 10 \) bootstrap replications, checked the accuracy of the projection pursuit fit. [The simplest possible method of bootstrapping a regression, by resampling entire cases \((x_i, y_i)\) rather than residuals, was used; see the discussion after (5.11) in Efron and Tibshirani (1986).] This gave approximate standard errors \((.12, .10, .14)\) for the components of \( \tilde{\alpha} \), so that the observed difference between \( \tilde{\alpha} \) and \( \hat{\alpha} \) could easily be due to statistical noise.

Figure 9 shows the 10 bootstrap replications of the projection pursuit regression curve. The bump seen in the right panel of Figure 8 appears in 9 of the 10 bootstraps. This is only marginally convincing evidence that the bump is genuine, rather than an artifact of the fitting process. We shouldn't be too surprised if a much larger data set had no bump at all.

![Figure 9. Ten bootstrap replications of the projection pursuit curve shown in the right panel of Figure 8 (indicated by asterisks here). The curves have been offset for clarity. The bump in Figure 8 appears in 9 of the 10 bootstrap curves.](image-url)

Traditional regression theory concentrates on unbiased estimators \( \hat{\mu}_1 \) for the true means \( \mu_1 : E(\hat{\mu}_1) = \mu_1 \). Unbiasedness, or at least approximate unbiasedness, is often desirable in scientific applications since it enforces the objectivity of the estimation process. In estimating the beneficial effect \( \mu \) of a new drug, an estimator \( \hat{\mu} \) which exceeded \( \mu \) 75% of the time would make doctors nervous, no matter how good its other statistical properties might be. On the other hand statisticians have begun to realize that deliberately induced biases can be a powerful ally in estimation problems involving many parameters. This section and the next give a brief look at biased estimation. See Efron (1975), (1982a) for more details.

We continue to consider the traditional regression model (2.6), (2.7): \( y = \mu + \varepsilon \) where \( \mu \in \mathcal{L} \), a known p-dimensional subspace of \( \mathbb{R}^p \), and \( \varepsilon \sim N(0, \sigma^2 I) \). [The normality of \( \varepsilon \) is not needed for the results quoted here, but will be required again in Section 9.] Suppose that contained in \( \mathcal{L} \) is a smaller linear subspace of interest \( \mathcal{L}^0 \),

\[
\mathcal{L}^0 \subset \mathcal{L},
\]

of dimension \( p^0 < p \). We can consider using \( \hat{\mu}^0 \), the point in \( \mathcal{L}^0 \) nearest to \( y \), as an estimate of the true mean vector \( \mu \), even if we don't necessarily believe that \( \mu \in \mathcal{L} \). In Figure 3 for example, we might consider using \( \hat{\mu}(1) \) instead of \( \hat{\mu}(2) \) to estimate \( \mu \), even though we think there is a good choice that the true regression is quadratic.

How good is the possibly biased estimator \( \hat{\mu}^0 \)? A natural measure of goodness is the Mean Squared Error of \( \hat{\mu}^0 \),

\[
\text{MSE} = E\{|\hat{\mu}^0 - \mu|^2\}.
\]

An easy calculation shows that

\[
\text{MSE} = |\mu - \mu^0|^2 + p^0 \sigma^2,
\]

(8.3)
where \( \hat{\mu}^0 \) is the point in \( \mathcal{X}^0 \) nearest \( \hat{\mu} \). Notice that (8.3) has a bias term

\[
\| \hat{\mu} - \hat{\mu}^0 \|^2,
\]

which is likely to get bigger as \( \mathcal{X}^0 \) gets smaller, and a variance term \( \tilde{\sigma}^2 \), which decreases as \( \mathcal{X}^0 \) gets smaller. Hopefully we can trade off bias for variance in order to minimize mean squared error.

Mallows' \( C_p \) formula (1974),

\[
C_p \equiv \| \hat{y} - \hat{\mu}^0 \|^2 + (2\tilde{\sigma}^2 - n)\tilde{\sigma}^2,
\]

is an unbiased estimator of the MSE for \( \hat{\mu}^0 \),

\[
E(C_p) = \text{MSE}.
\]

Here \( \tilde{\sigma}^2 = \| \hat{y} - \hat{\mu} \|^2 / (n-p) \) as in (2.12). The basic idea is to compare different choices of \( \mathcal{X}^0 \) in terms of their \( C_p \) statistics, with small values of \( C_p \) presumably indicating small MSE.

Formula (8.4), which is not difficult to derive, has an interesting feature: the term relating to variance, \( (2\tilde{\sigma}^2 - n)\tilde{\sigma}^2 \), punishes a model \( \mathcal{X}^0 \) in proportion to twice its dimension \( p^0 \). Readers familiar with Akaike's Information Criterion, see Stone (1977), will recognize this result. In fact the AIC reduces to Mallows' \( C_p \) in model (2.6), (2.7).

Table 7 shows \( C_p \) applied to the IUD data of Table 1. Polynomial models

\[
\hat{\mu}_i = \beta_0 + \beta_1 t_i + \beta_2 t_i^2 + \ldots + \beta_m t_i^m \quad (i = 1, 2, \ldots, 27)
\]

were compared for five choices of \( m, m = 0, 1, 2, 3, 4 \). The space \( \mathcal{X} \) (which enters the \( C_p \) formula only in the determination of \( \tilde{\sigma}^2 \)) was taken to be the biggest of these, \( m = 4 \). Differences between the three lots were ignored in this analysis.

The choice \( m = 2 \), the quadratic model, minimizes \( C_p \). The linear model, \( m = 1 \), does almost as well, but higher-order polynomials are noticeably worse. We can quite comfortably settle on the quadratic model in this case.
\[ m: \quad 0 \quad 1 \quad 2 \quad 3 \quad 4 \]
\[ C_p^m: \quad 939.1 \quad 13.6 \quad 12.4 \quad 22.6 \quad 27.8 \]
\[ \| y - \hat{\mu}(m) \|^2: \quad 1077.9 \quad 141.4 \quad 129.0 \quad 128.1 \quad 122.2 \quad (\hat{\sigma}^2 = 5.56) \]

Table 7. \( C_p^m \) analysis of the IUD data of Table 1. Polynomial regressions of degree \( m, m = 0, 1, 2, 3, 4 \), were compared. The quadratic model, \( m = 2 \), gives the smallest value of \( C_p^m \).

Suppose \( L \) is a regression model with \( p = 15 \) predictors. There are \( 2^{15} \) possible submodels, including or not including each predictor in the regression equation. A modern computer can find the minimum \( C_p^m \) among all \( 2^{15} \) models. Many regression packages include a single \( C_p^m \) command to do this calculation. As we discuss in the next section, this is not necessarily a good way to estimate the true regression function.


Table 8 shows the batting averages of \( n = 18 \) major league players after their first 45 at bats in the 1970 season. For example player 1 batted successfully in 18 of his first 45 tries, so \( y_1 = 18/45 = .400 \). The observed average \( y_1 \) for player 1 estimates his true average \( \mu_i \), how successful he would be in a very large number of tries. To a good approximation

\[ y_i = \mu_i + e_i \quad e_i \sim N(0, \sigma^2) \quad \text{independently} \quad i = 1, 2, \ldots, n, \quad (9.1) \]

where \( \sigma = [.265 \cdot .735/45]^{\frac{1}{2}} = .0658 \). See Efron (75) or Efron and Morris (75a).

Even though this is not a regression situation, since the \( \mu_i \) are not related as in (2.4), it will help us make an important and surprising point about the biased estimation of a regression. [This section can be skipped by readers anxious to get back to topics relating directly to regression methodology.]
The obvious estimate of \( \mu_i \) is \( \hat{\mu}_i = y_i \), the observed average, which is also the best unbiased estimate in situation (9.1). In other words, we estimate each player's true average by his observed proportion of successes during the period of observation. There doesn't seem to be any room for improvement here since, unlike the regression situation, there is no model linking the \( \mu_i \) to each other, and the \( y_i \) are statistically independent. The obvious is wrong in this case! Charles Stein showed that there can be a considerable amount of information about player \( i \) in the observed values of \( y_j, j \neq i \), and that this information can be extracted

<table>
<thead>
<tr>
<th>( i )</th>
<th>( y_i - \hat{\mu}_i )</th>
<th>( \hat{\mu}_i )</th>
<th>( \mu_i )</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.400</td>
<td>0.293</td>
<td>0.346</td>
</tr>
<tr>
<td>2</td>
<td>0.378</td>
<td>0.289</td>
<td>0.298</td>
</tr>
<tr>
<td>3</td>
<td>0.356</td>
<td>0.284</td>
<td>0.276</td>
</tr>
<tr>
<td>4</td>
<td>0.333</td>
<td>0.279</td>
<td>0.221</td>
</tr>
<tr>
<td>5</td>
<td>0.311</td>
<td>0.275</td>
<td>0.273</td>
</tr>
<tr>
<td>6</td>
<td>0.311</td>
<td>0.275</td>
<td>0.270</td>
</tr>
<tr>
<td>7</td>
<td>0.289</td>
<td>0.270</td>
<td>0.263</td>
</tr>
<tr>
<td>8</td>
<td>0.267</td>
<td>0.265</td>
<td>0.210</td>
</tr>
<tr>
<td>9</td>
<td>0.244</td>
<td>0.261</td>
<td>0.269</td>
</tr>
<tr>
<td>10</td>
<td>0.244</td>
<td>0.261</td>
<td>0.230</td>
</tr>
<tr>
<td>11</td>
<td>0.222</td>
<td>0.256</td>
<td>0.264</td>
</tr>
<tr>
<td>12</td>
<td>0.222</td>
<td>0.256</td>
<td>0.256</td>
</tr>
<tr>
<td>13</td>
<td>0.222</td>
<td>0.256</td>
<td>0.304</td>
</tr>
<tr>
<td>14</td>
<td>0.222</td>
<td>0.256</td>
<td>0.264</td>
</tr>
<tr>
<td>15</td>
<td>0.222</td>
<td>0.256</td>
<td>0.226</td>
</tr>
<tr>
<td>16</td>
<td>0.200</td>
<td>0.251</td>
<td>0.285</td>
</tr>
<tr>
<td>17</td>
<td>0.178</td>
<td>0.247</td>
<td>0.319</td>
</tr>
<tr>
<td>18</td>
<td>0.156</td>
<td>0.242</td>
<td>0.200</td>
</tr>
</tbody>
</table>

Table 8. Batting averages for 18 major league players after their first 45 at bats in 1970. The true value \( \mu_i \) is taken to be the batting average for the remainder of the season. The estimates (9.6), which are strongly biased toward the grand average \( \bar{y} = .265 \), have much smaller total squared error than the obvious estimates \( \hat{\mu}_i = y_i : \Sigma(\bar{\mu}_i - \mu_i)^2 / \Sigma(\hat{\mu}_i - \mu_i)^2 = .29. \)
by the use of biased estimators. Here we will be following Stein's 1981 paper, though his initial results were published in 1961.

Suppose that we estimate each true mean \( \mu_i \) in (9.1) by a possibly biased estimator

\[
\hat{\mu}_i = y_i - g_i(y), \tag{9.2}
\]

where \( g_i \) is allowed to be a function of the entire data vector \( y = (y_1, y_2, \ldots, y_n)' \).

We define the mean squared error of \( \hat{\mu} = (\hat{\mu}_1, \hat{\mu}_2, \ldots, \hat{\mu}_n)' \) to be

\[
\text{MSE}_{\hat{\mu}} = E\{\parallel \hat{\mu} - \mu \parallel^2\} = E \frac{n}{\sum_{i=1}^{n} (\hat{\mu}_i - \mu_i)^2}, \tag{9.3}
\]

as in (8.2). The variance \( \sigma^2 \) in (9.1) will be considered known in what follows, for instance \( \sigma^2 = .0658^2 \) in the baseball example, but this is just for convenient discussion, and is not essential to Stein's theory.

**Theorem (Stein, 1981).** Let \( S_{\hat{\mu}}(y) \) be the statistic

\[
S_{\hat{\mu}}(y) = n \sigma^2 + \sum_{i=1}^{n} g_i(y_i)^2 - 2\sigma^2 \sum_{i=1}^{n} \frac{\partial g_i(y_i)}{\partial y_i}. \tag{9.4}
\]

Then

\[
\text{MSE}_{\hat{\mu}} = E\{S_{\hat{\mu}}(y)\}. \tag{9.5}
\]

In other words, \( S_{\hat{\mu}}(y) \) is an unbiased estimate of the mean squared error for the estimation rule \( \hat{\mu} \). This generalizes Mallows' \( C_p \) result, (8.5). [Stein's theorem requires that the partial derivatives \( \frac{\partial g_i(y_i)}{\partial y_i} \) exist and have finite absolute expectations.]

The obvious estimation rule \( \hat{\mu}_1 = y_1 \) amounts to taking \( g_1(y) = 0 \) in (9.2). Then \( \hat{\mu} = (\hat{\mu}_1, \hat{\mu}_2, \ldots, \hat{\mu}_n)' \) has mean squared error \( \text{MSE}_{\hat{\mu}} = n\sigma^2 \) according to (9.4), (9.5). Rules like those discussed in Section 8, based on minimizing the \( C_p \) statistic, can have MSE either greater or less than \( n\sigma^2 \), depending on the actual
value of $\bar{\mu}$. Much to everyone's surprise, Stein was able to produce biased estimators having MSE always less than $n\sigma^2$:

**Corollary.** For $n \geq 4$, the estimator

$$
\tilde{\mu}_i = \bar{y} + \left[ 1 - \frac{(n-3)\sigma^2}{n \sum_{j=1}^{n} (y_j - \bar{y})^2} \right] (y_i - \bar{y}) \quad i = 1, 2, \ldots, n, \quad (9.6)
$$

where $\bar{y}$ is the grand average $\sum_{i=1}^{n} y_i / n$, has $\text{MSE}_{\tilde{\mu}_i} < n\sigma^2$ for every value of $\mu$.

[Proof: $S_{\tilde{\mu}}(y)$ in (9.4), (9.5) equals $n\sigma^2 \left( 1 - (n-3)^2 \sigma^2 / (n \sum_{j=1}^{n} (y_j - \bar{y})^2) \right) < n\sigma^2$.]

The reason for considering the baseball players is that we eventually learned the true values $\mu_1$, which allowed us to directly compare the estimators $\tilde{\mu}_i$ and $\hat{\mu}_i$. The third column of Table 8 shows each player's average during the remainder of the 1970 season. This is based on a much larger number of at bats, 369.3 on average, and is a reasonable approximation to the true value $\mu_i$. We see that the actual ratio of squared errors is

$$
\frac{\sum_{i=1}^{n} (\tilde{\mu}_i - \mu_i)^2}{\sum_{i=1}^{n} (\hat{\mu}_i - \mu_i)^2} = .29 \quad , \quad (9.7)
$$

strongly favoring the biased estimator (9.6).

Stein's results have profound implications for future developments in regression theory. They imply that in situation (2.6), (2.7), with $p \geq 3$, there are biased estimates of $\beta$ which are uniformly superior to the least squares estimate $\hat{\beta}$ in terms of MSE-type criterion. These biased estimators are smooth functions of the data vector $y$, unlike rules based on $C_p$ which discontinuously set some coordinates of the estimated regression vector equal to zero, and others equal to their least-squares values $\hat{\beta}_i$.

So far Stein's theory has had a relatively minor effect on the practice of statistical regression. The difficulties are interpretive ones; using (9.6) can
seem very much like combining apples, oranges, and watermelons if the \( \mu_i \) refer to completely unrelated quantities. Some of these difficulties are discussed in the last section of Efron (1982a).

10. Regression with Poisson Data.

Very often regression analysis must be applied to situations where the response variable is counted rather than measured. The normal response model (2.2), (2.3), which presumes a continuous response, must be wrong in such situations, sometimes dangerously so. We will discuss one such example here.

The upper panel of Table 9 shows the data, which consists of counts \( y_i \), \( i = 0, 1, 2, \ldots, 99 \). The counts refer to four poems by Christopher Marlowe, of total length 495 words. For reasons which will become apparent, each of the 495 words was ranked according to its rarity of appearance in Shakespeare's known works. The count \( y_i \) was the number of distinct words occurring at the \( i \)-th level of rarity. For example there were 10 distinct words among the 495 that appeared zero times in all of the 884,647 known words attributed to Shakespeare, so \( y_0 = 10 \); 8 distinct words among the 495 that appeared exactly once in the 884,647, so \( y_1 = 8 \); 8 which appeared twice, 9 which appeared three times, etc.

The lower panel of Table 9 gives expected values \( \nu_i \) for the \( y_i \), \( i = 0, 1, \ldots, 99 \), which would apply if the four poems were actually representative of Shakespeare's vocabulary. The details of this analysis (how can there be a prediction for the zero case?!) appear in Thisted and Efron (1986). Here we will just take the \( \nu_i \) as given, and ask whether or not they could possibly be the expected values of the \( y_i \) shown above them. In other words, did Shakespeare write Marlowe? The methods of regression analysis, modified to handle Poisson count data, will give us a nice answer.
A reasonable probability model for the counts $y_i$ is that they are independent Poisson observations with different true means $\mu_i$ (expected values),

$$y_i \sim \text{Poisson}(\mu_i) \quad \text{independently} \quad i = 0, 1, 2, \ldots, 99,$$  

(10.1)

so that $\text{Prob} \{y_i = y\} = e^{-\mu_i} \mu_i^y / y!$ for $y = 0, 1, 2, \ldots$. Model (10.1) replaces the normal model (2.2), (2.3) in the analysis which follows. But how can we replace the linear regression equation (2.4) in a reasonable way (for example to exclude the prediction of negative $\mu_i$, which is impossible in the Poisson situation)?

A nice theory, discussed and extended extensively in McCullagh and Nelder (1983), says that the correct procedure for Poisson regression is to linearly model $\log(\mu_i)$, rather than $\mu_i$ itself as in (2.4). The actual model used to test the data in Table 9 was

$$\log(\mu_i) = \log(v_i) + \beta_0 + \beta_1 \log(i+1) \quad i = 1, 2, \ldots, 99,$$  

(10.2)
where the $\nu_i$ are the constants appearing in the lower panel of the Table. Notice that $(\beta_0, \beta_1) = (0,0)$ corresponds to $\mu_i = \nu_i$, the case where the predictions based on Shakespearian usage exactly fit Marlowe's usage.

Model (10.2) was fit to the observed counts $y_i$ in Table 9 by the method of maximum likelihood. As will be discussed, this is the natural analogue of least squares for the Poisson situation, and indeed for all generalized forms of data-fitting. The method provides estimates of the parameters and also estimated standard errors. These came out to be

$$\hat{\beta}_0 = 1.14 \pm 0.22 \quad \text{and} \quad \hat{\beta}_1 = -0.32 \pm 0.08 \quad (10.3)$$

when model (10.2) was applied to the Marlowe data.

We see right off that neither $\beta_0$ nor $\beta_1$ equal 0, so we don't have grounds for worrying that Shakespeare wrote Marlowe's poems, or by implication vice-versa. Exponentiating (10.2) and plugging in the maximum likelihood estimates $\hat{\beta}_0, \hat{\beta}_1$ gives

$$\mu_i = \nu_i \cdot 3.13(i+1)^{-0.32} \quad (10.4)$$

Marlowe's usage declines at about the cube root of the rarity index $(i+1)$, as we move in the direction of larger $i$, that is toward less-rare Shakespearian usage. In other words, Marlowe's usage increases toward the rare end of Shakespeare's vocabulary. This doesn't mean that Marlowe was writing from a bigger vocabulary, just one selected with different frequencies. We might see the same result if a Shakesperian poem was analyzed with respect to its Marlowe rarity index.

What about analyzing Shakespeare with respect to Shakespeare? Table 10 shows the rarity counts and expected values for Sonnets 12-15. These are of total length 448 words, so the expected values are a little smaller than for the 495 Marlowe words. Model (10.2) was applied to this situation, giving maximum likelihood estimates
Table 10. Observed and expected word counts for Shakespeare's sonnets 12-15. In this case the expected counts are based on the rest of Shakespeare's works.

\[ \hat{\beta}_0 = 0.12 \pm 0.29 \quad \text{and} \quad \hat{\beta}_1 = -0.03 \pm 0.09. \] (10.4)

In this case we can easily accept the hypothesis \( (\beta_0, \beta_1) = (0,0) \), i.e. \( \mu_i = v_i \).

Thisted and Efron (1986) give several more examples with generally the same results: Shakespeare fits Shakespeare, but other writers don't.

Here are a few comments about model (10.2):

- It is a close analogue of the simple linear regression (2.5), with \( t_i \) equal to \( \log(i+1) \).

- The zero category, \( i = 0 \), was not included in the analysis on purely pragmatic grounds: in several of the cases considered in Thisted and Efron (1986), \( y_0 \) was grossly discrepant from the model (10.2) fit to \( y_i \) for \( i = 1, 2, \ldots, 99 \).

- Otherwise model (10.2) fits the data quite well. For instance, adding a quadratic term \( \beta_2 [\log(i+1)]^2 \) did not significantly improve the fit.
No single Poisson count, for instance \( y_1 = 8 \) compared to \( n_1 = 4.86 \), contains very much information. A regression model like (10.2) pools the information from a large number of cases, here \( i = 1, 2, \ldots, 99 \), in order to achieve a sensitive analysis. Such a model does not have to be perfectly accurate in order to be useful for this purpose.

Here is a brief description of maximum likelihood estimation, see Efron (1982d) or Nelder and McCullagh (1983). The basic idea is simply to choose as estimates those values of the unknown parameters which maximize the probability density of the observed data. In situation (10.1), (10.2) this means choosing \((\beta_0, \beta_1)\) to maximize \( \prod_{i=1}^{99} e^{-\mu_i} \frac{\mu_i^{y_i}}{y_i!} \), where \( \mu_i = n_i e^{\beta_0(i+1)} \), and the \( y_i \) are fixed at their observed values. For the traditional linear model (2.6), (2.7), the maximum likelihood estimate of \( \beta \) equals \( \hat{\beta} = (X^T X)^{-1} X^T y \), the least squares solution (2.9). In many ways maximum likelihood is the natural analogue of least squares for non-normal data.

The Kullback-Leibler distance between two Poisson distributions (10.1), with true means \( \mu_1 \) and \( \mu_2 \) respectively, is

\[
I(\mu_1, \mu_2) = E \left\{ \log \frac{\text{Prob}_{\mu_1}(y)}{\text{Prob}_{\mu_2}(y)} \right\} = \sum_{y=0}^{\infty} \frac{e^{-\mu_1} \mu_1^y}{y!} \left[ y \log\left( \frac{\mu_1}{\mu_2} \right) - (\mu_1 - \mu_2) \right]
\]

\[
= \mu_1 \log\left( \frac{\mu_1}{\mu_2} \right) - (\mu_1 - \mu_2).
\]

It turns out that the maximum likelihood estimate \( \hat{\beta} \) is the minimizer of

\[
\sum_{i} I(y_i, \mu_i(\hat{\beta})) ,
\]

the total Kullback-Leibler distance from the observed points \( y_i \) to the fitted values \( \mu_i(\hat{\beta}) = n_i e^{\hat{\beta}_0(i+1)} \) in our case. Maximum likelihood is equivalent to Least Kullback-Leibler distance estimation.

For normal observations (2.2), (2.3), the Kullback-Leibler distance

\[
E \left\{ \log \frac{\text{Prob}_{\mu_1}(y)}{\text{Prob}_{\mu_2}(y)} \right\} = \frac{(\mu_1 - \mu_2)^2}{2\sigma^2},
\]

so that least squares, least
Kullback-Leibler distance, and maximum likelihood are equivalent fitting methods. Any error distribution generates its own Kullback-Leibler distance function, and hence its own analogy of least squares fitting. The binomial case, where maximum likelihood fitting is known as logistic regression, is particularly useful in practice, see Cox (1970). Only the normal case has a simple closed-form solution for $\hat{\theta}$. For other error distributions the maximum likelihood estimate must be found by an iterative search algorithm. McCullagh and Nelder (1983) give a nice description of this theory and its generalizations. They also provide a sophisticated computer package called GLIM which automates the fitting of generalized regressions.

11. Regression With Censored Data.

Table 11 shows a small portion of the Stanford heart transplant data, reported in full in Miller and Halpern (1982). Full data was available on $n = 157$ patients. The response variable $y_i$ is the survival length of patient $i$ following the transplant. The covariate $t_i$ is the patient's age at the time of the operation. The doctors wished to know the effect of age on survival length.

<table>
<thead>
<tr>
<th>Patient #</th>
<th>Age (years)</th>
<th>Survival Length (days)</th>
</tr>
</thead>
<tbody>
<tr>
<td>101</td>
<td>46</td>
<td>1393+</td>
</tr>
<tr>
<td>102</td>
<td>38</td>
<td>1202</td>
</tr>
<tr>
<td>103</td>
<td>41</td>
<td>1378+</td>
</tr>
<tr>
<td>104</td>
<td>41</td>
<td>1373+</td>
</tr>
<tr>
<td>105</td>
<td>31</td>
<td>274</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>106</td>
<td>33</td>
<td>31</td>
</tr>
<tr>
<td>107</td>
<td>50</td>
<td>1341+</td>
</tr>
<tr>
<td>108</td>
<td>19</td>
<td>42</td>
</tr>
<tr>
<td>109</td>
<td>45</td>
<td>381</td>
</tr>
<tr>
<td>110</td>
<td>52</td>
<td>1264+</td>
</tr>
</tbody>
</table>

Table 11. Ten patients, Stanford heart transplant data. The full data set of $n = 157$ patients, appears in Miller and Halpern (1982). The "+" after a survival length indicates that the patient was still alive at the end of the observation period. This is an example of censored data. The doctors wished to know the effect of the patient's age on his length of survival.
Censoring of the response variable keeps this from being a standard exercise in regression analysis. The data was tabulated in February 1980, at which time 40% of the patients were still alive. Patient #101 for instance had survived for 1393 days by February 1980, and was still doing fine. His data is reported as \( y_{101} = 1393^+ \), indicating that we only know that his true survival length exceeded 1393 days. Data censoring of this sort is common in medical experimentation.

It is obvious that it won't do to simply run a regression of \( y_i \) as some function of \( t_i \), ignoring the censoring. This can badly bias the results. Nor can we discard all of the censored cases, which would lead to even worse biases. D. R. Cox (1972) introduced a clever form of regression analysis which takes into account all of the data, both censored and uncensored, in an unbiased way. Cox's method is particularly computer-intensive.

The analysis begins with predictor-response pairs \( (x_i, y_i), i = 1, 2, \ldots, n \), exactly as in (2.1). The predictor vectors \( x_i \) are \( p \)-dimensional; for the heart transplant data we are going to take

\[
x_i = (1, t_i, t_i^2),
\]

(11.1)

a quadratic regression on age \( t_i \), so \( p = 3 \).

In an ordinary regression analysis the predictor vector determines the mean of the response variable \( y_i \), as in (2.4). In Cox's analysis the predictor determines the survival curve

\[
S_i(y) \equiv \text{Prob}\{y_i > y\}
\]

(11.2)

of \( y_i \). [As an example of a survival curve, if \( y_i \) has density function \( (1/\mu)e^{-y/\mu} \) for \( y > 0 \), then \( S_i(y) = e^{-y/\mu} \).] Let \( S_0(y) \) be the survival curve of a patient with covariate vector \( x = 0 \). Then Cox's regression model is

44
\[ \log \{ S_i(y) \} = e^{x_i \beta} \cdot \log \{ S_0(y) \}. \] (11.3)

Here \( \beta \) is an unknown vector of regression parameters that we wish to estimate from the data. A little bit of numerical experimentation shows that if \( x_i \beta \) is positive then \( S_i(y) < S_0(y) \) for all \( y \); i.e. the survival tends to be shorter than in the "par" case \( S_0 \). Conversely, negative \( x_i \beta \) leads to longer survivals.

Model (11.3) looks strange but it turns out to be just what is needed to handle censored data. Here is the key idea. Suppose that \( \mathcal{R} \) is a set of integers, and to each \( j \in \mathcal{R} \) there corresponds an \( (x_j, y_j) \) pair, the \( y_j \) being independent random quantities satisfying (11.3). Suppose also that we know that all of the \( y_j \)'s for \( j \in \mathcal{R} \) survived longer than a certain fixed length \( y \), but that exactly one \( y_j \) died in the infinitesimal interval \( (y, y+dy) \). Then the probability that the death happened to correspond to the pair with predictor \( x_i \) is

\[ e^{x_i \beta} / \sum_{j \in \mathcal{R}} e^{x_j \beta}. \] (11.4)

Proving (11.4) is a nice exercise in elementary probability theory.

Cox's analysis proceeds as follows:

(i) List the response values \( \{ y_{i_1}, y_{i_2}, y_{i_3}, \ldots, \} \) corresponding to uncensored observations. For instance Table 11 shows that the list of uncensored observations for the heart transplant data includes \{1202, 274, 31, 42, 381\}.

(ii) Corresponding to each value \( y_{i_k} \) compute the risk set \( \mathcal{R}_k \), this being the set of subjects in the study who might possibly have been observed to die on day \( y_{i_k} \). For patient 102, \( y_{i_k} = 1202 \), the risk set \( \mathcal{R}_k \) includes patient 101, since that patient was still under observation on day 1202; it does not include patient 105 who left the study (due to death) on day 274.

(iii) Estimate the regression vector \( \beta \) in (11.3) by that value \( b = \hat{\beta} \) which maximizes the product, over all the uncensored observations,
\[ \prod_{k} \sum_{j \in \mathcal{R}_k} e^{x_{jk}b} \]  \hspace{1cm} (11.5)

The rationale behind this procedure is simple: according to (11.4), the product (11.5) is the probability, if \( \beta = b \), of seeing the sequence of deaths actually observed; we select the estimate \( \hat{\beta} \) by the method of maximum likelihood. Notice that

- Censoring does not affect the computation, except that censored observations disappear from all risk sets after their censoring time. The method can be shown to give efficient, unbiased estimates of \( \beta \) no matter what the censoring pattern may be, see Chapter 5 of Miller (1981).

- The "par" survival curve, \( S_0(t) \) never enters the computation of \( \hat{\beta} \). In this sense model (11.3) is partially nonparametric.

The left panel of Figure 10 shows the result of fitting a quadratic regression in age, (11.1), to the heart transplant data by Cox's method. The dashed line shows the fitted quadratic curve \( \hat{\beta}_0 + \hat{\beta}_1 t + \hat{\beta}_2 t^2 \). It declines until age 35 and then increases rapidly, indicating much poorer survival for the older patients.

The physicians were interested in the decline of the fitted curve from ages 15 to 35. Taken literally, the result indicates that a 15 year old has a worse heart transplant prognosis than a 35 year old. Of course this result could easily be an artifact due to using a quadratic model.

Hastie and Tibshirani (1986) checked this by fitting model (11.3) with the quadratic function \( x_1 = \beta_0 + \beta_1 t + \beta_2 t^2 \) replaced by a smooth function of \( t \). Essentially the local smoothing works as described in Section 6, though the implementation is more complicated for Cox's model. The solid line in the left panel of Figure 10 shows the fitted smooth curve. Now the minimum risk occurs at age 45. The decline from age 15 to 45 is much less pronounced.
Figure 10. Stanford Heart Transplant Data. The left panel, dashed line, shows the quadratic regression function $\hat{\beta}_0 + \hat{\beta}_1 t + \hat{\beta}_2 t^2$ obtained by Cox's method. The solid line indicates a Cox regression based on local smoothing, rather than the quadratic model. The right panel displays 20 bootstrap replications of the locally smoothed Cox regression. The rise in risk for patients older than 45 is genuine, but the observed decreases in risk up to age 45 is probably an artifact.

The right panel of Figure 10 displays 20 bootstrap replications of the locally smoothed Cox regression. A substantial proportion of the bootstrap curves are flat, rather than decreasing, up to age 45. It is clear that the rise in risk after age 45 is genuine, but we cannot put much confidence in the observed decline up to age 45. This example is discussed in Section 3 of Efron and Tibshirani (1986).

A staggering amount of computation went into Figure 10. The Cox quadratic regression takes perhaps 100 times as long as the corresponding least squares fit. Local smoothing and cross-validation adds a factor of 20. The bootstrap analysis adds another factor of 20. All together we have done at least 40,000 times more computation than in the classical analysis.

Our payoff is a more believable result, based on a theory that makes minimal assumptions about the regression model. The power of modern equipment is so great that the entire analysis took only a few minutes on a minicomputer.
All of the theory behind Figure 10 was invented since 1972. The pace of new statistical theory is accelerating, as statisticians learn how to take better advantage of the electronic computer. Ordinary least squares regression, as perfected by Fisher in the 1920's, exhausted the capabilities of the mechanical calculator. We are nowhere near the limits of the electronic computer.

References


Efron, B. (1982). The jackknife, the bootstrap, and other resampling plans. SIAM Monograph #38.


Efron, B. and Thisted, R. (1986). Did Shakespeare write a new poem?


