tist., 17, pp. 175–213.
aka Math. J., 3, pp. 175–213.
a Math. J., 4, pp. 41–68.
st. Assoc., 74, pp. 105–131.
by splines, unpublished manu-

0, 1] *approximation by splines*,

blem, Skand. Aktuarietids., 46,

timating the location and scale Math. and Statist., 5-11, Inst.

E, Stan Res. Rep. No. 1976-1,

elated errors, Ann. Math. *h*

ny parameters, Ann. Math.

III, Ann. Math. Statist., 41,

the Canadian Mathematical . 115–136.
ion, in Statistics and Related . Saleh, eds., North-Holland,

statistical theory, Ark. Mat., tockholm.
ple quantiles, Technometrics,

ortional stratified sampling. D. thesis, Univ. of California,

er, Ann. Math. Statist., 42,

. Statist., 2, pp. 925–934.
mal distribution from linear pp. 52–57.

SIAM REVIEW
Vol. 30, No. 3, September 1988

COMPUTER-INTENSIVE METHODS IN STATISTICAL REGRESSION*

BRADLEY EFRON†

Abstract. This is a survey of modern developments in statistical regression, written for the mathematically educated nonstatistician. It begins 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.

Key words. robust regression, bootstrap, least absolute deviations, projection pursuit, cross-validation, Mallows' C_p, Stein estimation, Poisson regression

AMS(MOS) subject classifications. 62-02, 62505

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 nonstatistician 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 1800s and was completed by Fisher in the 1920s. 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 indispensible 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 three lots of nine devices per 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 §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.

* Received by the editors May 5, 1986; accepted for publication May 26, 1987.
† Department of Statistics, Stanford University, Stanford, California 94305.

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 three lots of nine devices each (left panel). Lot C looks as if it contained considerably more of the hormone. The right panel shows the number of hours each device was worn.

Amount of hormone remaining, mgs				Hours worn			
Lot:	A	B	C	Lot:	A	B	C
Device				Device			
1.	25.8	16.3	28.8	1.	99	376	119
2.	20.5	11.6	22.0	2.	152	385	188
3.	14.3	11.8	29.7	3.	293	402	115
4.	23.2	32.5	28.9	4.	155	29	88
5.	20.6	32.0	32.8	5.	196	76	58
6.	31.1	18.0	32.5	6.	53	296	49
7.	20.9	24.1	25.4	7.	184	151	150
8.	20.9	26.5	31.7	8.	171	177	107
9.	30.4	25.8	28.5	9.	52	209	125
Mean:	23.1	22.1	28.9	Mean:	150.6	233.4	111.0
St. Error:	1.8	2.7	1.2				

The 27 data points are plotted in Fig. 1 with horizontal and vertical axes

$$(t, y) = (\text{hours worn, remaining hormone}), \tag{1.1}$$

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,

$$\mu(t) = \beta_0 + \beta_1 t. \tag{1.2}$$

This is our first example of a linear regression.

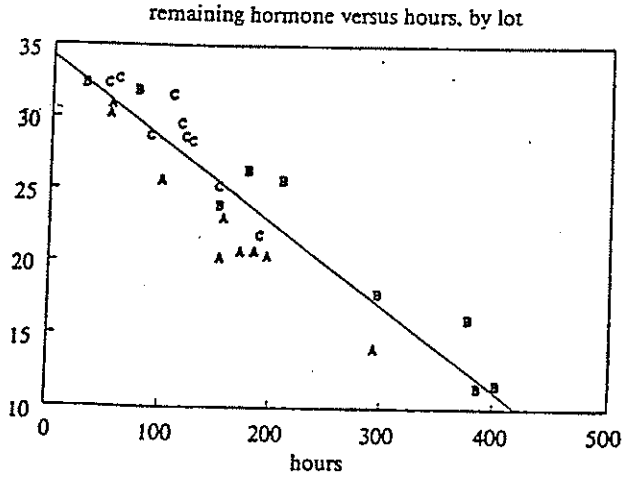

FIG. 1. *A plot of the amount of hormone remaining in each device (vertical axis) versus the number of hours the device was worn (horizontal axis). Each point is labeled with the lot of the corresponding device. The straight line is the least squares fit to the 27 data points.*

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, y) by the method of least squares: choosing $(\hat\beta_0, \hat\beta_1)$ to minimize $\sum [y - (\hat\beta_0 + \hat\beta_1 t)]^2$. This gives the least-squares line

$$\hat\mu(t) = \hat\beta_0 + \hat\beta_1 t, \qquad \begin{cases} \hat\beta_0 = 34.17, \\ \hat\beta_1 = -.0574 \end{cases}$$

shown in Fig. 1.

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

$$r_i = y_i - [\hat\beta_0 + \hat\beta_1 t_i].$$

The basic idea is simple but powerful: it is more meaningful to compare the residuals 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 Fig. 1.

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.

Lot:	A	B	C
Device			
1.	−2.68	3.73	1.47
2.	−4.94	−0.45	−1.37
3.	−3.34	0.73	2.14
4.	−2.06	0.00	−0.21
5.	−2.31	2.20	1.96
6.	−0.02	0.84	1.15
7.	−2.70	−1.39	−0.15
8.	−3.44	2.50	3.68
9.	−0.78	3.64	1.51
Mean:	−2.48	1.31	1.13
St. error:	0.48	0.60	0.50

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 that both the regression curve itself, in this case line (1.3), and the deviations from the regression, shown 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 1920s. 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

(2.1) $$(x_i, y_i), \qquad i = 1, 2, \cdots, n$$

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 (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 Fig. 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 an optimal procedure, we need a probabilistic model relating y_i to x_i. The traditional model begins by assuming that

(2.2) $$y_i = \mu_i + e_i, \qquad i = 1, 2, \cdots, n$$

where μ_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,

(2.3) $$e_i \sim N(0, \sigma^2) \quad \text{independently } i = 1, 2, \cdots, n.$$

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

Model (2.2) represents each y_i as signal plus noise. The heart of the regression model is an assumption that the signals μ_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 μ_i. This relationship is expressed by the linear formula

(2.4) $$\mu_i = x_i\beta, \qquad i = 1, 2, \cdots, n.$$

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

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

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

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 Fig. 1.

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

(2.6) $$\mathbf{y} = \boldsymbol{\mu} + \mathbf{e}$$

where

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

The notation $e \sim N(0, \sigma^2 I)$ 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

$$\mathscr{L} \equiv \{u = Xb, b \in \mathbb{R}^p\}. \tag{2.8}$$

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

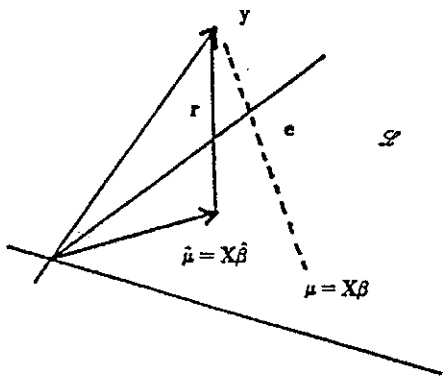

FIG. 2. *A schematic picture of the linear model. The true mean vector μ is constrained to lie in the p-dimensional linear subspace $\mathscr{L}$ spanned by the columns of X. We observe $y = \mu + e$, and wish to estimate $\mu = X\beta$. The least squares estimate $\hat{\mu} = X\hat{\beta}$ is the point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

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

$$\hat{\beta} = G^{-1}X'y \qquad (G \equiv X'X), \tag{2.9}$$

going back to Legendre and Gauss (who had a nasty priority fight over the invention of least squares) in the early 1800s. The solution $\hat{\beta} = (34.17, -.0574)'$ in (1.3) was computed from (2.9); note that in this case X is a 27×2 matrix, but G is only 2×2, 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 β, 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 β.

This last point is the principal theoretical justification for the method of least squares, and is worth stating quantitatively. The covariance matrix Σ 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

$$\Sigma = \sigma^2 G^{-1}. \tag{2.10}$$

Statistical theory shows that any other unbiased estimator of β has covariance matrix larger than (2.10), in particular, 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 $\hat{\beta}$. The only unknown quantity in (2.10) is σ^2. This can be estimated from the residual vector

$$\mathbf{r} \equiv \mathbf{y} - \hat{\mu}. \tag{2.11}$$

The usual estimate is

$$\hat{\sigma}^2 = \frac{\|\mathbf{r}\|^2}{n-p}, \tag{2.12}$$

which is the best unbiased estimator of σ^2 under model (2.6), (2.7). From Fig. 2 we see that $\mathbf{r}$ always lies in the $(n-p)$-dimensional space orthogonal to $\mathscr{L}$, which accounts for the denominator of (2.12). Because $\mathbf{r}$ is orthogonal to the space containing μ, its distribution depends only on the noise σ^2 in (2.6), (2.7), and not on the signal μ. In brief, the part of $\mathbf{y}$ in $\mathscr{L}$ estimates μ (or equivalently β), while the part orthogonal to $\mathscr{L}$ estimates σ^2.

Most least-squares regression programs print out the estimated coefficients $\hat{\beta}_j$ and also their estimated *standard errors* (square root of estimated variance) from (2.10), (2.12): $[\hat{\sigma}^2(G^{-1})_{jj}]^{1/2}$. Linear regression (1.3) gives $\hat{\beta}_0 = 34.17 \pm .87$ and $\hat{\beta}_1 = -.0574 \pm .0087$, the numbers following "$\pm$" being the estimated standard errors. The true value of β_j lies within one estimated standard error of $\hat{\beta}_j$ with probability about 67 percent, and within two standard errors with probability about 95 percent.

The simplest of all regression problems is that of estimating a single common mean: $y_i = \mu + e_i$ for $i = 1, 2, \cdots, n$ 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 $[\sum_1^n (y_i - \bar{y})^2/n \cdot (n-1)]^{1/2}$. For instance, the $n = 9$ numbers in the first column of Table 1 give $\hat{\mu} = 23.1 \pm 1.8$.

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

$$\mu(t) = \beta_0 + \beta_1 t + \beta_2 t^2, \tag{3.1}$$

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

$$\hat{\mu}_i = \hat{\beta}_0 + \hat{\beta}_1 t_i + \hat{\beta}_2 t_i^2, \qquad i = 1, 2, \cdots, 27 \tag{3.2}$$

gives a significantly better fit to the data in Fig. 1.

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

$$\mathscr{L}(2) = \{\mathbf{u}: u_i = b_0 + b_1 t_i + b_2 t_i^2, i = 1, 2, \cdots, 27\} \tag{3.3}$$

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

$$\mathscr{L}(1) = \{\mathbf{u}: u_i = b_0 + b_1 t_i, i = 1, 2, \cdots, 27\}, \tag{3.4}$$

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

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

$$\|y - \hat{\mu}(2)\|^2 \leqq \|y - \hat{\mu}(1)\|^2. \tag{3.5}$$

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

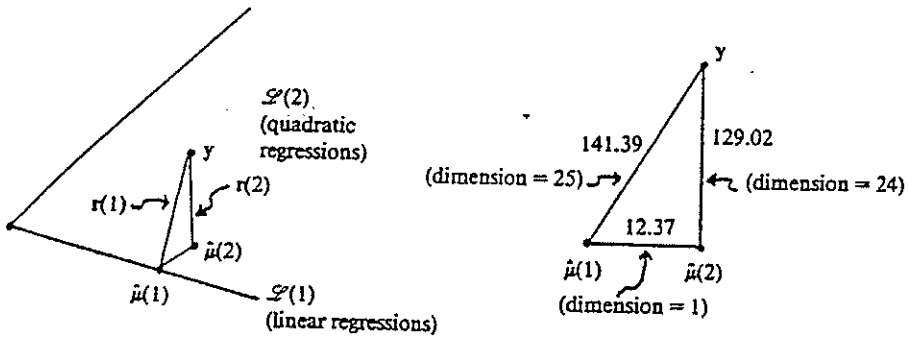

FIG. 3. *On the left is a schematic diagram comparing linear and quadratic models for the hormone data of §1. Because $\mathscr{L}(1) \subset \mathscr{L}(2)$, the residual vector must be larger for $\mathscr{L}(1)$. The squared lengths $\|y - \hat{\mu}(1)\|^2$ and $\|y - \hat{\mu}(2)\|^2$ are shown at right. Traditional theory indicates that there is no strong reason to believe that quadratic regression is superior to linear regression for the hormone data.*

Figure 3 indicates the squared residual lengths $\|y - \hat{\mu}(1)\|^2 = 141.39$ and $\|y - \hat{\mu}(2)\|^2 = 129.02$ for the hormone data, from which Pythagorus' theorem gives $\|\hat{\mu}(2) - \hat{\mu}(1)\|^2 = 12.37$. 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) $y - \hat{\mu}(2)$ takes its value in the 24-dimensional subspace of $\mathbb{R}^{27}$ orthogonal to the three-dimensional space $\mathscr{L}(2)$.

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

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

$$F = \frac{\|\hat{\mu}(2) - \hat{\mu}(1)\|^2}{\|y - \hat{\mu}(2)\|^2/24} = \frac{(SSE_R - SSE_C)/(dfE_R - dfE_C)}{SSE_C / dfE_C} \tag{3.6}$$

to approximately equal 1 (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 1."

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

$$\text{Prob}\{F > 2.30\} = .14. \tag{3.7}$$

(vi) An "achieved significance level" of .14 is not considered significant evidence for the existence of a genuine quadratic component in μ. We expect bigger values of F, and smaller achieved significance levels, if μ is genuinely quadratic. The conventional borderline for significance is .05, in this case $F \geqq 4.26$. 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 introduced this theory in the 1920s. Most of our previous results, for example (2.10), do not require e in (2.6), (2.7) to be normally distributed, but calculations like (3.7) 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), helps to guide the scientist toward an insightful analysis of noisy data.

This does not 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 one 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, along with other writers in the late 1700s, 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 Fig. 1, LAD gives $(\hat{\beta}_0, \hat{\beta}_1) = (34.20 - .0587)$, almost the same as the the least-squares solution $(34.17, -.0574)$. Stigler (1986) provides a fascinating history of regression theory.

Least absolute deviations and least squares do not always agree so nicely. Table 3 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

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 Fig. 4.

Plate number	Radiation dose (Rads/100)	Survival proportion	Log survival proportion
1.	1.175	.44	− .821
2.	1.175	.55	− .598
3.	2.35	.16	−1.833
4.	2.35	.13	−2.040
5.	4.70	.0400	−3.219
6.	4.70	.0196	−3.932
7.	4.70	.0612	−2.794
8.	7.05	.0050	−5.298
9.	7.05	.0032	−5.745
10.	9.40	.00110	−6.812
11.	9.40	.00015	−8.805
12.	9.40	.00019	−8.568
13.	14.10	.00070	−7.264
14.	14.10	.00006	−9.721