6-1982

The Cost of Drilling for Oil and Gas: An Application of Constrained Robust Regression

William F. Eddy
Carnegie Mellon University, bill@stat.cmu.edu

Joseph B. Kadane
Carnegie Mellon University, kadane@stat.cmu.edu

The Cost of Drilling for Oil and Gas: An Application of Constrained Robust Regression

WILLIAM F. EDDY and JOSEPH B. KADANE*

The robust regression method of Huber (1973) is used to fit a model to the cost of drilling for petroleum. Because the model includes a categorical variable (well type), a linear constraint is imposed on the parameter estimates. Because the model was fit to the logarithm of cost and because it will be used to make repeated predictions of cost, an adjustment that approximately unbiases the predictions is imposed. The numerical values of the estimates are discussed, and a comparison is made with ordinary least squares.

KEY WORDS: Constrained regression; Cost of drilling; Iteratively reweighted least squares; Missing data; Petroleum; Robust regression.

1. INTRODUCTION

Every year the American Petroleum Institute (API) conducts a mail survey in an effort to determine the cost of drilling each oil and gas well and dry hole in the United States in the previous year. The results of the survey are published annually in aggregate form in the *Joint Association Survey of the U.S. Oil and Gas Producing Industry* (JAS). Because the survey is conducted by mail and response is strictly voluntary, a large percentage of questionnaires are not returned. However, the API also issues the *Quarterly Review of Drilling Statistics* (QRDS), which contains a summary of the detailed information (including type, depth, and location but not cost) on each well drilled in the United States. The QRDS data are gathered by responsible field officers and are widely believed to be a nearly complete list of the wells and dry holes drilled. From the data on which the QRDS is based it is possible not only to estimate the percentage of nonresponse, but also to determine specific facts about the wells with unreported costs. With this in mind the API desires each year to estimate, from the information contained in the completed questionnaires and in the QRDS, the cost of drilling wells.

It is customary in problems of this type to try, through follow-up surveys or other means, to determine if the fact of nonresponse to the original survey is related to the value of the response. For a variety of reasons, a significant one being the proprietary nature of the information, this recourse is not available.

To fulfill the purposes of the JAS it is necessary to estimate the missing costs. Historically, this has been done by one analyst, using experience and judgment to estimate the cost per foot from reported information for wells of a similar type, depth, and location. The judgment of the analyst may stem from information on geology, structure, the presence of geo-pressurized zones, and so on. Although no experiments using other analysts have been tried, obviously the exact same results would not be reproduced, which would be a significant disadvantage of this method.

The major objective of the work reported here was to provide, by a systematic method (to be used each year), point estimates of the costs of drilling typical nonreported wells. This method had to produce estimates that were not overly affected by unusually large or small reported costs, and simultaneously had to account for the variability due to the depth of the hole drilled, the geological formation where the hole was located, and the particular type of well being drilled. The method was intended only to indicate broad patterns, not to replace detailed engineering studies of particular sites.

Drilling for oil and gas is a business fraught with uncertainty. Occasionally, unpredicted, peculiar circumstances cause substantial change in the normal cost to be expected. A Bayesian analysis of the problem would require that we give a probability distribution for the cost of each well, and in fact a joint probability distribution for the costs of all the wells, since they should not necessarily be assumed to be independent. Possibly the costs of drilling wells might be assumed to be conditionally independent, given the values of certain parameters. Although desirable, such a "prevision" (in the sense of deFinetti 1974) was more detailed than good sense and the desires of our clients warranted. Consequently, we chose to make "predictions" (a single number rather than a probability distribution), and to try to predict "normal" cost. One of the fundamental difficulties of this work is that normal cost is not well defined, but clearly implies some level of insensitivity to outlying observations. Another difficulty is that although experienced statisticians examined the particular data described here, in future years the methodology will be applied directly to the data without any particular examination by statisticians or judgment by experts.

* William F. Eddy is Associate Professor of Statistics and Joseph B. Kadane is GSIA/Statistics Professor of Statistics and Social Sciences at Carnegie-Mellon University, Pittsburgh, PA 15213. This project was supported by the American Petroleum Institute.

June 1982, Volume 77, Number 378
Applications Section

2. REGIONS

The geographical regions used by the API for its publication are generally states, with the exception of Texas, which is divided into Railroad Commission Districts. Such political units are not geologically homogeneous, however. Consequently, regions were proposed that, according to geologists employed by API member companies, have reasonable geological homogeneity. A map of the regions is given in Figure 1. No wells were drilled in the blank regions, so they were unassigned.

3. THE MODEL

The cost of drilling an oil or gas well, as a function of depth, location, and type of well, has a distribution with a long upper tail. At least part of this tail could be accounted for by catastrophes, such as well-head fires, broken pipes, and so on. However, a logarithmic transformation made the distribution approximately symmetric. Henceforth, we consider only this transformed variable, the natural logarithm of cost in thousands of dollars, denoted by *log cost*. While the distribution of log cost appears more symmetric, a substantial problem with outliers remains in both tails.

The term *outlier* is used to mean an observation that, conditional on the background variables, does not *appear* to be from the same distribution as the remainder of the data. There are two general methods for dealing with outliers: rejection and accommodation. We generally prefer the second approach: modification of procedures to reduce their sensitivity to outliers. Consequently, the robust regression method of Huber (1973) was used to estimate the effects of the various variables.

The variables used as predictors were depth of the well in feet, type of well (a qualitative variable described in Table 1), and geological region (also a qualitative variable). The definitions of the various well types are complex (see API (1981)) and not important for the purposes described here, but a brief explanation follows. An exploratory well is drilled for the purpose of finding oil or gas in an area or at a depth where it is not known to be. A developmental well is drilled in a proven area to a known productive depth for the purpose of production. A multiple completion well is equipped to produce oil and/or gas separately from more than one reservoir; production is measured separately for each completion of one well.

It was known a priori that cost was a nonlinear function of depth. It turned out that log cost was also a nonlinear function of depth. Since the model being fit is certainly not the "true" model (it is merely being used for estimation), it seemed reasonable to account for the nonlinearity by transformation of the depth. After applying the method of Tukey (1977, Exh. 21, p. 198), we chose the natural logarithm of depth, in 10,000 feet, and named it *log depth*. We examined plots of log cost versus log depth and they appeared to be approximately linear. We concluded that log depth was the appropriate choice.

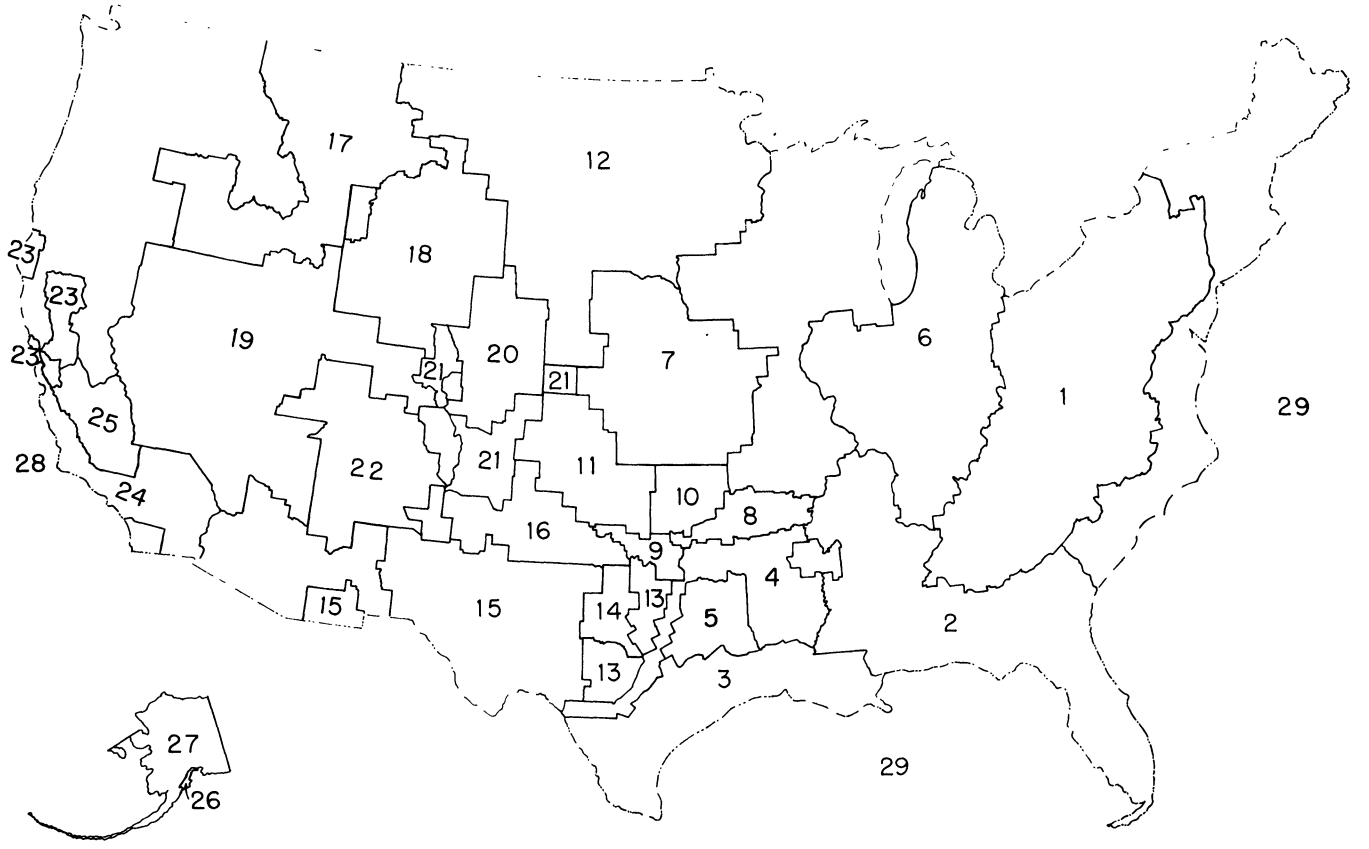

Figure 1. Geological Regions

Table 1. Type of Well

(0) Exploratory Oil (EXP OIL)
(1) Exploratory Oil, Multiple Completion (EXP OILM)
(2) Exploratory Gas (EXP GAS)
(3) Exploratory Gas, Multiple Completion (EXP GASM)
(4) Exploratory Dry (EXP DRY)
(5) Developmental Oil (DEV OIL)
(6) Developmental Oil, Multiple Completion (DEV OILM)
(7) Developmental Gas (DEV GAS)
(8) Developmental Gas, Multiple Completion (DEV GASM)
(9) Developmental Dry (DEV DRY)

Our first model fit the log-cost variable for the entire data set with three kinds of variables: log depth, type of well, and geological region (as 29 indicator variables). Subsequently, API geologists requested that we model each region separately. Thus, the basic model reported here is

$$\mathbf{Y} = X\boldsymbol{\beta} + \boldsymbol{\epsilon}, \tag{1}$$

where $\mathbf{Y} = (y_1, \ldots, y_n)^T$ is the $n \times 1$ vector of dependent variables (here, log cost) and $X = (\mathbf{X}_1, \ldots, \mathbf{X}_n)^T$ [$\mathbf{X}_i^T = (x_{i1}, \ldots, x_{ip})$, $i = 1, \ldots, n$] is the $n \times p$ matrix of explanatory variables. In the final models fitted here X has a column of ones (the constant term), a column containing log depth, and 10 indicator variables for type of well, so p is 12. In earlier attempts with these data p was as large as 42. The value of n varies across regions from a maximum of 2838 to a minimum of 23. The number of wells reported in the JAS in 1975 was 16,146. This is the total sample size summed over the 29 areas. The coefficients to be estimated are the $p \times 1$ vector $\boldsymbol{\beta}$, and the vector of errors is the $n \times 1$ vector $\boldsymbol{\epsilon}$.

According to experts at the API, only information on the depth of the well, the type of well, and the geological region just described is available to estimate the cost of drilling for purposes of the JAS, although other unavailable variables do influence cost. Consequently, we assume that conditional on these three variables, reporting of the cost is independent of the cost itself. (In fact, reporting is a function of the company drilling the well, and depends on whether it is company policy to cooperate with JAS.) Technically, we assume that the data are missing at random (Rubin 1976). An empirical test of this assumption would require data on the unreported wells, which are ipso facto unavailable. We also assume that the parameters governing cost are distinct from those governing which data are missing. Under these assumptions, Rubin shows that likelihood inference (Theorems 7.1 and 7.2) and Bayesian inference (Theorems 8.1 and 8.2) can be legitimately conducted without further explicit consideration of the missing data, as we have done. Thinking of robust estimation as a modification of the normal distribution maximum likelihood estimator justifies our reliance on the assumption of data missing at random (and distinct parameters). We discussed this assumption with API representatives and found them comfortable with it.

4. THE CONSTRAINTS

As discussed in Section 3, our original model has two sets of indicator variables, well-type and geological region. Since the sum of the indicator variables in each set is the constant column in the X matrix, the rank of the X matrix for that model is two less than the number of columns. One way of dealing with the resulting indeterminacy in the estimate of $\boldsymbol{\beta}$ is to impose two linear constraints on $\boldsymbol{\beta}$. These constraints could be imposed by eliminating one variable in each set, but this treats the categories asymmetrically. We chose to eliminate the constant term and to constrain the sum of the coefficients of the indicator variables in each set to sum to zero. Let H be a $q \times p$ matrix and constrain the vector $\boldsymbol{\beta}$ to satisfy

$$H\boldsymbol{\beta} = \mathbf{0}. \tag{2}$$

The zero on the right side is a $q \times 1$ vector of zeroes. The value of q in our final model is one. The details of implementing (2) will be given in Section 5.

5. THE ESTIMATORS

It is customary to estimate $\boldsymbol{\beta}$ by minimizing

$$(\mathbf{Y} - X\boldsymbol{\beta})^T(\mathbf{Y} - X\boldsymbol{\beta}) \tag{3}$$

subject to (2). The resulting $\boldsymbol{\beta}$'s are sensitive to unusual values of Y_i. In fact, the more unusual the value of Y_i, the greater the effect it has on the estimate of $\boldsymbol{\beta}$. Because outliers are an important characteristic of drilling costs, an alternative to least squares was critical to the success of the predictions. (We are not, of course, attempting to predict outliers or even attempting to determine which nonreported wells are outlying since there is no information in the data at hand that will help.) The choice here was Huber's (1973) method modified to allow the constraints just mentioned.

The method is to minimize

$$\sum_{i=1}^{n} \rho(y_i - \mathbf{X}_i\boldsymbol{\beta}) \tag{4}$$

subject to (2), where

$$\begin{aligned} \rho(x) &= \tfrac{1}{2}x^2, \quad |x| \le c \\ &= c|x| - \tfrac{1}{2}c^2, \quad |x| > c. \end{aligned} \tag{5}$$

This particular choice of ρ was motivated originally by Huber, but it has the additional advantages that it is fairly easy to explain to a nonstatistician and is quite fast to compute; because of the amount of data and the fact that the solution to (4) is iterative, speed can be important. The method is equivalent to minimizing

$$\sum_{i=1}^{n} \rho(y_i - \mathbf{X}_i\boldsymbol{\beta}) - \boldsymbol{\Lambda}^T H\boldsymbol{\beta}, \tag{6}$$

where $\boldsymbol{\Lambda}^T = (\lambda_1, \ldots, \lambda_q)$ is a $1 \times q$ vector of Lagrange multipliers. Differentiating with respect to $\boldsymbol{\beta}$ and $\boldsymbol{\Lambda}$ and setting the derivatives equal to zero yields

$$\sum_{i=1}^{n} x_{ij}\psi(y_i - \mathbf{X}_i\hat{\boldsymbol{\beta}}) - \mathbf{H}_j^T\hat{\boldsymbol{\Lambda}} = 0, \quad j = 1, \ldots, p \tag{7}$$

where

$$\psi(x) = \rho'(x) = \min(c, \max(-c, x)) = \text{median}(-c, x, c), \quad (8)$$

and

$$\mathbf{H}_j^T = (h_{j1}, \ldots, h_{jq}),$$
$$H\hat{\boldsymbol{\beta}} = \mathbf{0}. \quad (9)$$

Of course, these estimates are not scale invariant.

To get scale invariance in the unconstrained problem, Huber proposes to solve for $\hat{\boldsymbol{\beta}}$ and the scale estimate s in his (8.2) and (8.3). Our notation is as follows.

$$\sum_{i=1}^{n} x_{ij}\psi\left(\frac{y_i - \mathbf{X}_i\hat{\boldsymbol{\beta}}}{s}\right) = 0 \quad j = 1, \ldots, p \quad (10)$$

and

$$\frac{1}{n-p}\sum_{i=1}^{n}\psi^2\left(\frac{y_i - \mathbf{X}_i\hat{\boldsymbol{\beta}}}{s}\right) = \gamma, \quad (11)$$

where $\gamma = E\psi^2(\epsilon)$. Modifying (10) to account for the constraint, as done in equation (7), yields

$$\sum_{i=1}^{n} x_{ij}\psi\left(\frac{y_i - \mathbf{X}_i\hat{\boldsymbol{\beta}}}{s}\right) - \mathbf{H}_j^T\hat{\boldsymbol{\Lambda}} = 0 \quad j = 1, \ldots, p. \quad (12)$$

To solve (11) and (12) requires choice of the constant c in the definition of ρ and choice of the constant γ. It is clear that these choices are interrelated, but there is no guidance in the literature. Smaller values of c correspond to larger fractions of outliers. We chose $c = 1$ because this would allow as much as 30 percent of the data to be modified if the errors were normally distributed; the a priori fraction of outliers was believed to be less than 30 percent. While there might be some loss of efficiency in over-trimming (c small), we felt it was riskier to allow a smaller fraction of outliers (c large) and there were sufficient data so that efficiency was less important. Having chosen $c = 1$, we studied Huber's Table 3 to determine γ. A few trial runs with various values of γ on a sample of our data led us to choose $\gamma = .5$ so that 25 to 30 percent of the outliers would be modified in the estimation procedure.

Because (11) and (12) are nonlinear equations, an iterative method was used for their solution. Suppose $\tilde{\boldsymbol{\beta}}$ and $\tilde{s}$ are estimates of $\boldsymbol{\beta}$ and s, respectively. Let

$$r_i = (y_i - \mathbf{X}_i\tilde{\boldsymbol{\beta}})/\tilde{s} \quad (13)$$

and let

$$w_i = \psi(r_i)/r_i. \quad (14)$$

Then (12) may be written as

$$\sum_{i=1}^{n} x_{ij} w_i r_i - \mathbf{H}_j^T\hat{\boldsymbol{\Lambda}} = 0, \quad j = 1, \ldots, p. \quad (15)$$

Letting W be the diagonal matrix with w_i as the ith diagonal element, (15) becomes

$$(X^TW(\mathbf{Y} - X\tilde{\boldsymbol{\beta}})/\tilde{s}) - H^T\hat{\boldsymbol{\Lambda}} = \mathbf{0}. \quad (16)$$

Notice that (16) also results from minimizing

$$(\mathbf{Y} - X\boldsymbol{\beta})^TW(\mathbf{Y} - X\boldsymbol{\beta})$$

subject to

$$H\boldsymbol{\beta} = \mathbf{0}.$$

The solution is

$$\hat{\boldsymbol{\beta}} = [A^{-1} - A^{-1}H^T(HA^{-1}H^T)^{-1}HA^{-1}]X^TW\mathbf{Y}, \quad (17)$$

where $A = X^TWX + H^TH$.

The iterative solution proceeds as follows. Choose an initial $\tilde{\boldsymbol{\beta}}$ that satisfies the constraints and choose an initial $\tilde{s}$. Here $\tilde{\boldsymbol{\beta}} \equiv \mathbf{0}$ and $\tilde{s} = 1$ will do. Compute W according to (13) and (14). Substitute in (11) and (17) to yield new values for $\tilde{\boldsymbol{\beta}}$ and $\tilde{s}$. Repeat until convergence is obtained. We are unaware of a general proof that this method does in fact converge. However, Dempster, Laird, and Rubin (1980) have recently proven convergence when s is estimated by maximum likelihood (rather than by solving (11)) and the distribution of the errors is assumed to be a scale mixture of normal distributions. In each of the 29 situations here, satisfactory convergence was obtained in 10 to 20 iterations.

6. ADJUSTMENT

Because the fitted model will be used to make many predictions of *cost*, there is some danger in fitting log cost. Suppose for a moment that robustness is not a concern and that log cost is normally distributed with mean μ and variance σ^2. Obviously then, cost is lognormally distributed and

$$E(\text{cost}) = \exp(\mu + \tfrac{1}{2}\sigma^2) = \exp(\mu)\cdot\exp(\tfrac{1}{2}\sigma^2).$$

More generally, an unbiased estimate of log cost when exponentiated tends to underestimate cost. This is, of course, because $\exp(\cdot)$ is a convex function. Our solution to this problem was to estimate from the data an adjustment factor to correct the predictions for this underestimation. The details are given in the next section.

While the logarithm transformation of a response variable has long been used to stabilize the variance (see, e.g., Curtiss 1943), it is only recently that much attention has been paid to correcting for the resulting underestimation (see, e.g., Land 1972).

7. NUMERICAL RESULTS

This section is devoted to a description and discussion of the numerical results contained in Table 2, which were obtained by application of the robust regression method to the data collected by the API for 1975. The first column gives the number of reported wells for each of the regions. The second column gives the number of wells with reported cost data. This number is the number of observations available to estimate the parameters of the model, and the difference between the first and second columns is the number of predictions that will be made. Note that regions 26, 27, and 28 have very few wells drilled.

Table 2. Numerical Results of Robust Regression

Region	# of Input Records	# of Data Points	# Huberized Residuals	Scale Est.	Mean Rho	Correction Factor	Constant	Log Depth	EXP OIL	EXP OILM	EXP GAS	EXP GASM	EXP DRY	DEV OIL	DEV OILM	DEV GAS	DEV GASM	DEV DRY
1	3854	1092	335	.3168	.07117	.98538	5.3049	1.0483	.0778	—	.2070	—	.0106	−.3228	.0259	.1343	.2803	−.4131
2	578	172	50	.5285	.20545	1.06043	6.1421	1.3623	.1092	—	.1544	.2756	−.1796	−.2432	−.0226	.1637	.2694	−.5270
3	4704	2409	748	.4898	.16503	1.02308	5.9510	1.5401	.0118	−.0920	.1647	.2171	−.3063	.1488	.1373	.0713	.0720	−.4248
4	1576	520	123	.5962	.21976	1.04968	5.6835	1.2560	.2454	—	.3190	—	−.4807	.1895	.4195	−.1390	.3658	−.9195
5	512	199	55	.3806	.09682	1.02045	5.9339	1.3678	.5053	—	.0942	—	−.4547	−.1048	.2808	.2282	—	−.5490
6	2543	712	214	.4631	.15210	1.03399	6.3122	1.6854	.1699	.0701	.1765	—	−.3906	.0781	.1098	.2413	—	−.4551
7	2338	507	155	.3404	.09158	1.03957	5.1241	1.0362	.1918	—	.1660	—	−.5321	.1998	.2820	.2300	—	−.5376
8	218	97	22	.4839	.14583	1.04892	6.1182	1.2693	−.2846	—	.0173	—	.4124	−.1406	—	.0208	.3357	−.3609
9	589	183	46	.5054	.18323	1.03743	6.1055	1.3749	−.0770	—	.4370	—	−.3455	.0774	.4939	.1553	—	−.7411
10	1530	284	76	.3500	.08192	1.02431	5.6584	1.3296	.1036	—	.5870	—	−.6169	.1436	−.0964	.2503	.2099	−.5811
11	2721	1606	492	.3280	.07912	.99911	5.9652	1.6045	−.1139	.2650	.1518	.4229	−.6270	−.0795	.1354	.0942	.2306	−.4795
12	675	300	84	.4092	.11525	1.07814	6.0354	1.5095	−.0799	−.0698	.3499	.7802	−.5045	−.2305	—	.222	—	−.4676
13	566	147	43	.3004	.06861	1.00157	5.2510	.9844	.1256	—	.2581	—	−.3778	.2031	—	.0814	.3944	−.6848
14	1898	451	144	.3210	.07425	1.00016	5.4952	1.3623	.0058	—	.2874	—	−.4408	.1766	.2850	.1751	—	−.4891
15	4934	2838	833	.3497	.08943	1.01754	5.9209	1.0988	.0374	.1662	.2912	1.1069	−.4190	−.3255	−.0186	−.2664	.0551	−.6274
16	821	247	76	.4132	.11712	1.06092	5.8042	1.4474	.0449	—	.6058	—	−.2743	−.0888	.2158	.3905	—	−.8940
17	400	140	29	.3868	.09572	.96642	4.9624	.7360	.5082	—	−.1105	−.0568	−.2367	.1924	.4307	.0957	−.2420	−.5811
18	1328	761	236	.4000	.12492	.99924	6.0400	1.1030	−.0335	—	.5795	.1049	−.5960	−.1254	.3684	.1784	.2798	−.7560
19	211	153	48	.2773	.08509	1.01628	6.3789	1.8918	.2386	.2799	.4113	—	−.6013	.0539	−.3378	.2679	—	−1.5010
20	1077	483	142	.2236	.04469	1.01364	5.3097	1.3398	.2628	.0198	.2753	—	−.9536	.3379	.6364	.2931	—	−.8321
21	137	73	20	.3907	.13987	1.00436	5.2904	.8750	.6152	—	−.0104	.4854	−.6158	.0708	−.0597	.1221	—	−.6076
22	552	321	92	.2858	.05607	1.00156	6.1890	1.5481	−.0926	—	.0382	.1458	.0756	.0087	—	−.1978	−.0289	.0511
23	187	70	14	.6196	.25742	.96517	5.4309	.8969	—	—	.3647	—	−.2574	—	—	.4356	−.0389	−.5039
24	408	305	85	.3709	.09268	1.00436	6.2341	1.0941	.1405	—	—	—	−.2408	−.0722	−.0391	—	—	.2116
25	1514	1159	357	.2915	.06173	1.01219	5.8698	.8693	−.1234	—	1.3781	—	−.7526	−.3370	.4308	−.0020	—	−.5939
26	25	23	1	.8612	.38436	1.09065	7.5822	1.2144	—	—	.3628	—	.4476	−.4243	—	—	—	−.3861
27	33	24	2	.6671	.23120	1.13031	8.2342	.3394	.4657	—	—	—	.4515	−.6286	—	—	—	−.2885
28	57	47	9	.3826	.10125	1.01069	7.3509	1.0717	.6902	—	—	—	—	−.9458	—	—	—	.2556
29	954	823	247	.5479	.20136	1.05696	6.9823	.8183	.4912	—	.0860	−.1206	.1917	−3224	.0040	.0246	.0210	−.3754

Table 3. Comparison Between Weighted and Unweighted Least Squares Regression

	Residual Mean Sq	Multiple R	PRESS	Constant	Log Depth	EXP OIL	EXP OILM	EXP GAS	EXP DRY	DEV OIL	DEV OILM	DEV GAS	DEV GASM	DEV DRY
Region 6														
Weighted	.168	.864	.170	6.32	1.69	.169	.073	.175	−.391	.076	.112	.241	—	−.455
Unweighted	.247	.823	.252	6.27	1.66	.164	.059	.202	−.392	.082	.110	.254	—	−.476
Region 13														
Weighted	.075	.857	.077	5.25	.987	.126	—	.258	−.378	.203	—	.081	.394	−.685
Unweighted	.129	.819	.139	5.29	1.034	.112	—	.248	−.353	.187	—	.099	.375	−.669

The third column gives the number of scaled residuals that are greater than $c = 1$. That is, it gives the number of observations for which w_i in (14) is less than one. The proportion of such observations fluctuates from region to region but is typically between 25 and 30 percent, conforming to the desired fraction. This suggests that the choice of $\gamma = .5$ and $c = 1$ is not unreasonable.

The column labeled "scale estimate" gives the value of s from solving (11) and (17), and the column labeled "mean rho" gives the value of (4) divided by n at the solution. In a least squares regression there would be an exact quadratic relationship between these two columns. Here the relationship is not strictly monotone and is more linear (see Figure 2). Notice that the regions with the largest residual mean rho have very few data points.

The correction factor column contains the estimate of the adjustment factor needed to correct for exponentiation. This factor is equal to the ratio of the sum of the reported costs to the sum of the estimated costs, where both sums are taken over those observations for which $w_i = 1$. The rationale is simply that for the observations that are "good," the estimates should be unbiased. If they *are* unbiased for these observations, then the predictions made for the wells with unreported costs will also be unbiased. As expected, most of the correction factors are larger than one. Because our data are not lognormally distributed, we chose not to use a method such as that of Land (1972) to make corrections. Furthermore, the poor behavior he reports for naive methods of correction makes us wary of strongly advocating our procedure without further study. A referee suggested an alternative correction factor: $\exp(\frac{1}{2}\hat{\sigma}^2)$. A comparison of these two methods of adjustment is deferred to the next section.

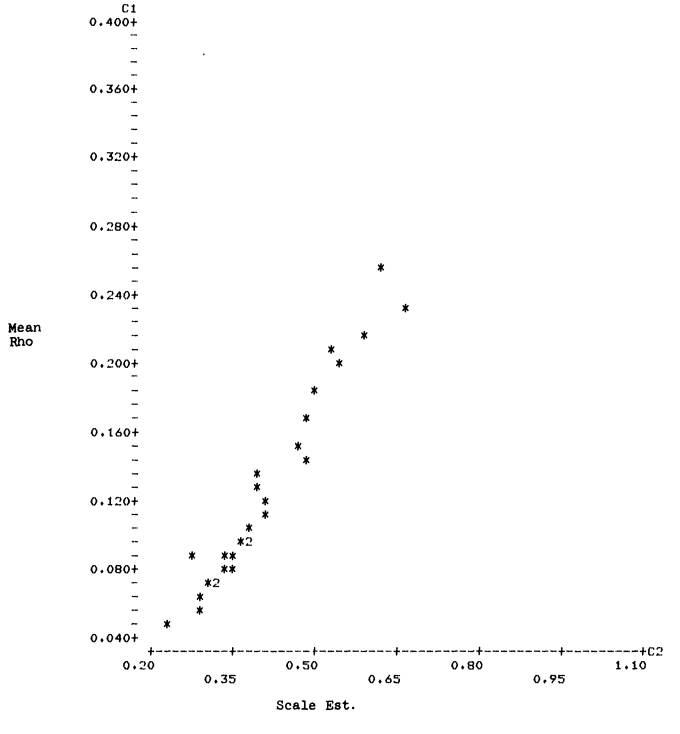

Figure 2. Plot of Mean Rho Against Scale Estimate

The constant, in the next column, is the estimated log cost of a 10,000 foot well, neglecting the effect of well type and the correction factor. The next column gives the coefficient of "log depth." It is generally the case that cost per foot should be an increasing function of depth. If so, then all these coefficients should be larger than one. For the few regions with coefficients smaller than one, it appears either (a) that cost per foot is a decreasing function of depth, (b) that there are so few data that the estimates are unreliable, or (c) that an important variable is missing from the model.

The last 10 columns give the differential effects of well type. There are several comforting features of these columns. Dry wells are generally cheaper to drill than successful wells despite the fact that they are generally drilled deeper. This is no doubt because the definition of cost includes items not installed on dry holes, such as the "Christmas tree." Also, exploratory wells are often more expensive than development wells in the same region, again for fairly obvious reasons. It is difficult to make more explicit comparisons of well-type effects across regions since the constraint differs from region to region; there are different well types, and numbers of each type, in the various regions.

As a hypothetical example of the use of Table 2, suppose we wish to estimate the cost of a 5,000 foot development oil well (single completion) in region 15. From Table 2, we have

$$\underset{\text{(in thousands of dollars)}}{\log \text{cost}} = \underset{\text{(constant)}}{5.9209} + \underset{\text{(log depth in 10,000 feet)}}{\log(.5)} \times 1.0988 - \underset{\text{(development oil)}}{.3255}.$$

$$\text{Unadjusted log cost} = 4.8338.$$

$$\text{Unadjusted cost} = \exp(4.8338) = \$125{,}684.$$

$$\text{Adjusted cost} = \$125{,}684 \times 1.06092 = \$133{,}341.$$

8. COMPARISON WITH LEAST SQUARES

To assess the value of the robust regression method used here, ordinary least squares regressions were also calculated. Rather than present all the results, two regions (13 and 6) were chosen as representative. To make the comparisons as direct as possible we used the weights from the robust regression (14) and also computed a weighted least squares regression using the same program used to compute the ordinary least squares output. This provided not only a check on the computations, but also parallel output for the least squares regression and robust regression. A summary of the numerical information is given in Table 3.

The residual mean squares for the weighted regression are larger than the mean rho values reported in Table 2 because the mean rho values have been divided by $n - p$ rather than by the sum of the weights. The column

labeled PRESS is the predicted residual mean square obtained as the (weighted) average squared difference of the observation and the prediction that would have been obtained if the observation were not included in the regression. This statistic was suggested by Allen (1971) and considered by Geisser and Eddy (1979, Sec. 3.2) for selecting predictor variables in multiple regression. The larger the ratio of PRESS to the residual mean square, the less adequate the fit. The column labeled "Multiple R" is the ordinary (unadjusted) correlation between observed and predicted values of log cost. For both regions the least squares regression appears to fit well, but the robust regression fits slightly better. In the final paragraph of this article, we give our view of the usefulness of robust regression for this problem.

Plots of the standardized residuals versus the predicted values for each of the four regressions are given as Figures 3 through 6. The standardization divides each residual by the square root of the corresponding diagonal element in

$$W^{-1} - X[A^{-1} - A^{-1}H^T(HA^{-1}A^T)^{-1}HA^{-1}]X^T, \quad (18)$$

which is the covariance matrix of the residuals. This is a widely recommended procedure; see, for example, Prescott (1975) and the references therein.

Generally, the larger absolute residuals from a robust regression are larger in magnitude than their counterparts from a least squares regression. Here it is easy to see that the major impact of the robust regression is to reduce the larger absolute *standardized* residuals. This is not surprising. The diagonal element in (18) is dominated by W^{-1}. Thus, the standardized residual is approximately

$$w_i^{1/2}\tilde{s}r_i = \operatorname{sgn}(r_i)\,\tilde{s}\,[r_i\,\psi(r_i)]^{1/2}.$$

When $|r_i| \le c$, (we have chosen $c = 1$) it is just $\tilde{s}r_i$; when $|r_i| \ge c$, it equals $\operatorname{sgn}(r_i)\tilde{s}\,|r_i|^{1/2}$, using $c = 1$. Therefore, we are not surprised that the larger absolute standardized residuals are smaller in magnitude than their counterparts from a least squares regression.

Plots of the standardized residuals versus log depth were also examined for the four regressions. Because no structure was apparent, we concluded that the relationship between log cost and log depth was approximately linear. This confirmed our choice of transformations.

We also used the regressions for Regions 6 and 13 to compare our adjustment for the convexity of exp(·) with

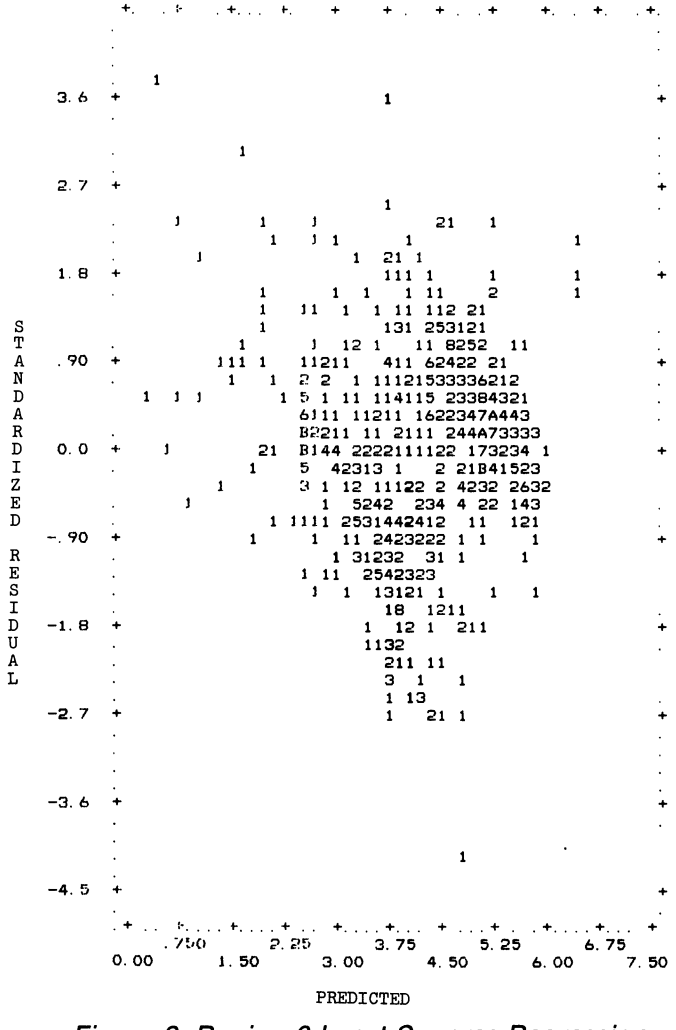

Figure 3. Region 6 Least Squares Regression

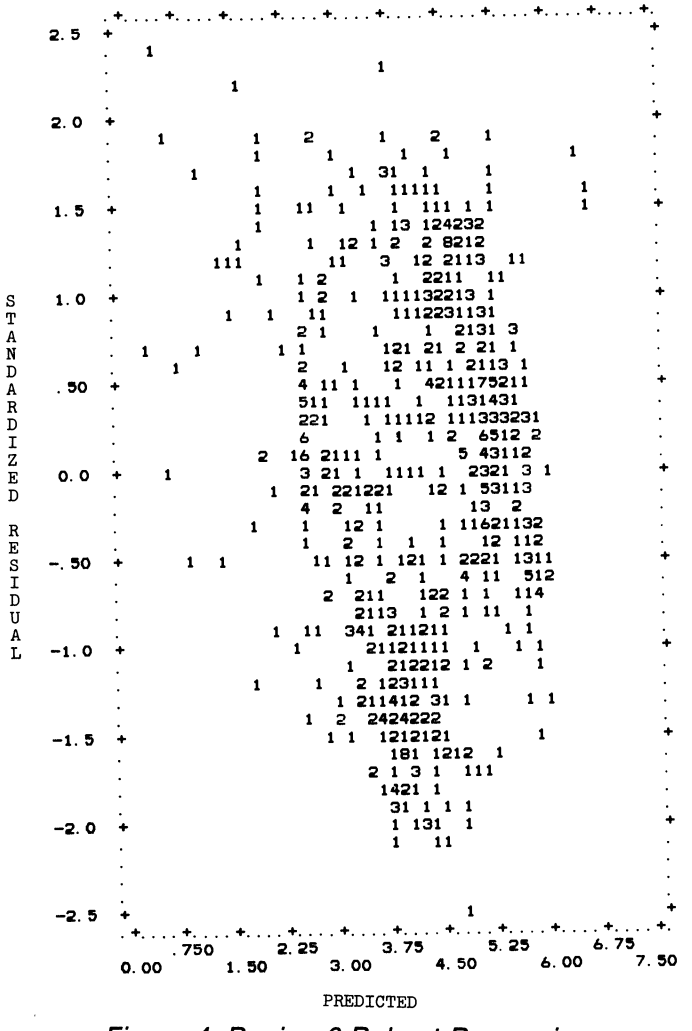

Figure 4. Region 6 Robust Regression

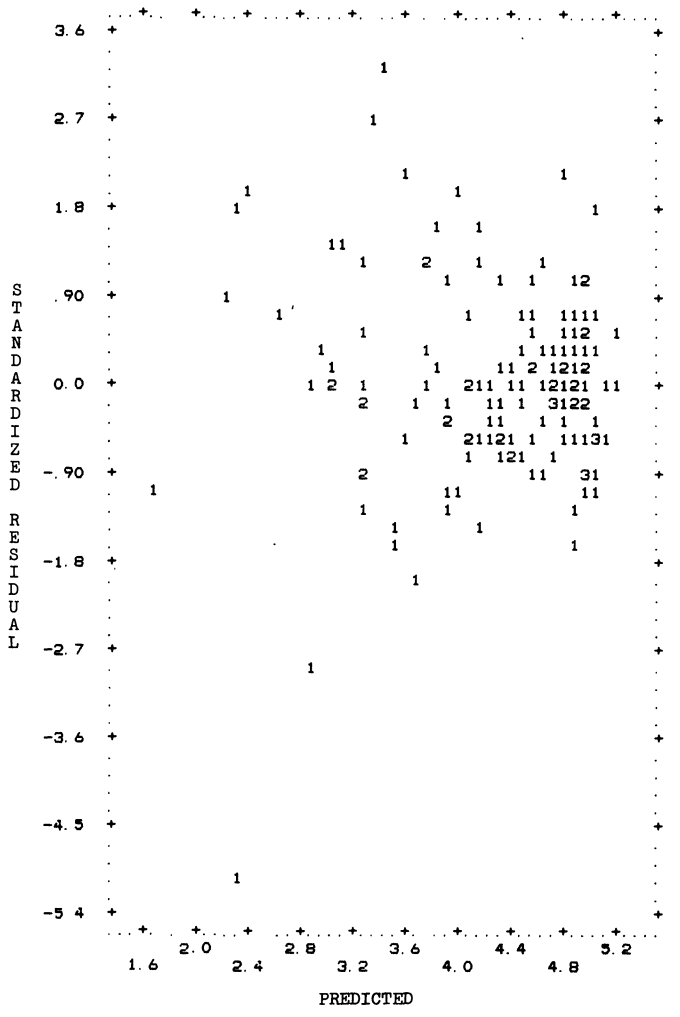

Figure 5. Region 13 Least Squares Regression

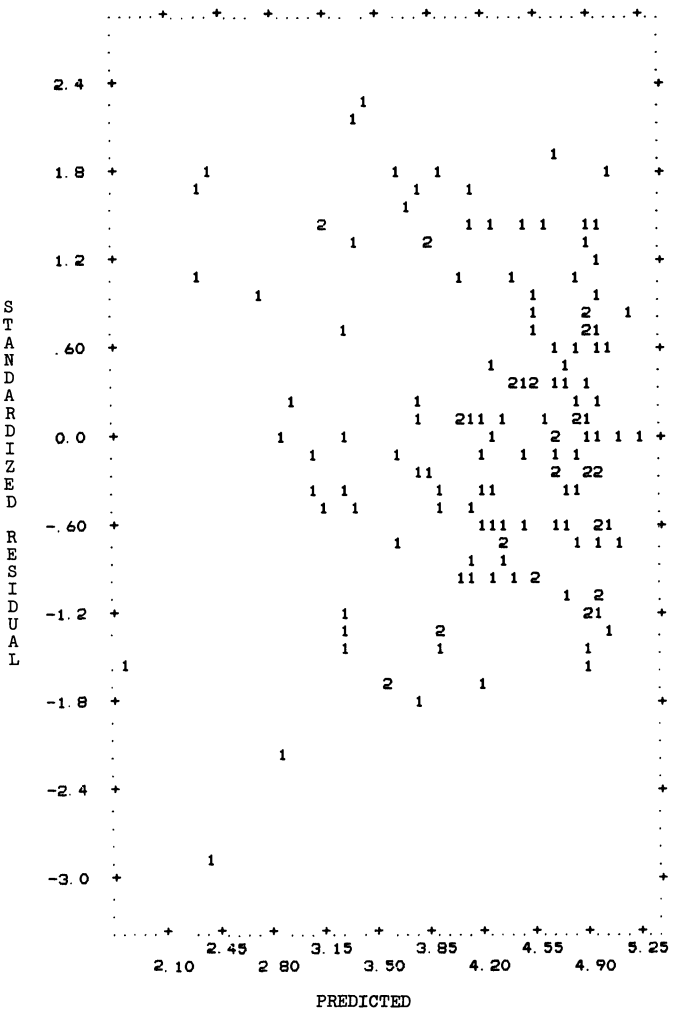

Figure 6. Region 13 Robust Regression

that suggested by the referee. In region 6, exp $\frac{1}{2}$ (.168) = 1.088 would be the correction factor instead of 1.034; in region 13, the correction factor would be exp $\frac{1}{2}$ (.075) = 1.038 instead of 1.002. These numbers suggest that both correction ideas are reasonable and that neither appears to make a substantial difference.

It appears that between 5 and 10 percent of these data are outliers. While this is much smaller than the 25 to 30 percent that led to our original choices of $c = 1$ and $\gamma = .5$, it must be emphasized that the method will subsequently be applied to data that we have not examined. Because efficiency does not seem to be as important an issue, we felt the best choice was to protect ourselves against much larger fractions of outliers and we thus retained our original choices for c and γ.

For this data set, the use of robust methods led to only a slight change in the estimates and predictions. Nonetheless, given that our task was to propose a method to be used on data as yet unseen, we feel that the protection provided by robust methods is an important advantage and worth the premium paid in loss of efficiency.

[*Received October 1978. Revised July 1981.*]

REFERENCES

ALLEN, DAVID M. (1971), "The Prediction Sum of Squares as a Criterion for Selecting Prediction Variables," Technical Report No. 23, University of Kentucky, Dept. of Statistics.

API (1981), "Standard Definitions for Petroleum Statistics," Technical Report No. 1, Third Edition, Washington, D.C.

API (1977), "1975 Joint Association Survey of the U.S. Oil and Gas Producing Industry," Washington, D.C.

CURTISS, J.H. (1943), "On Transformations Used in the Analysis of Variance," *Annals of Mathematical Statistics*, 14, 107–122.

DEFINETTI, B. (1974), *Theory of Probability, Vol. 1*, New York: John Wiley.

DEMPSTER, A.P., LAIRD, N.M., and RUBIN, D.B. (1980), "Iteratively Reweighted Least Squares for Linear Regression When Errors are Normal/Independent Distributed," in *Multivariate Analysis V*, ed. P.R. Krishnaiah, Amsterdam: North-Holland, 35–57.

GEISSER, SEYMOUR, and EDDY, WILLIAM F. (1979), "A Predictive Approach to Model Selection," *Journal of the American Statistical Association*, 74, 153–160.

HUBER, P.J. (1973), "Robust Regression: Asymptotics, Conjectures, and Monte Carlo," *Annals of Statistics*, 1, 799–821.

LAND, CHARLES E. (1972), "An Evaluation of Approximate Confidence Interval Estimation Methods for Lognormal Means," *Technometrics*, 14, 145–158.

PRESCOTT, P. (1975), "An Approximate Test for Outliers in Linear Models," *Technometrics*, 17, 129–132.

RUBIN, D.B. (1976), "Inference and Missing Data," *Biometrika*, 63, 581–592 (with discussion).

TUKEY, J.W. (1977), *Exploratory Data Analysis*, Reading, Mass.: Addison-Wesley.