JOURNAL OF ENVIRONMENTAL ECONOMICS AND MANAGEMENT **31,** 403–405 (1996)
ARTICLE NO. 0052

On the Harrison and Rubinfeld Data*

OTIS W. GILLEY[†]

Department of Economics and Finance, College of Administration and Business, Louisiana Tech University, Ruston, Louisiana 71272

AND

R. KELLEY PACE[‡]

Department of Finance, School of Management, University of Alaska, Fairbanks, Alaska 99775-6080

Received March 4, 1996

In a well-known paper, Harrison and Rubinfeld [4] investigated various methodological issues related to the use of housing data to estimate the demand for clean air. They illustrated their procedures using data from the Boston Standard Metropolitan Statistical Area with 506 observations (1 observation per census tract) on 14 nonconstant independent variables. These variables include levels of nitrogen oxides (NOX), particulate concentrations (PART), average number of rooms (RM), proportion of structures built before 1940 (AGE), black population proportion (B), lower status population proportion (LSTAT), crime rate (CRIM), proportion of area zoned with large lots (ZN), proportion of nonretail business area (INDUS), property tax rate (TAX), pupil-teacher ratio (PTRATIO), location contiguous to the Charles River (CHAS), weighted distances to the employment centers (DIS), and an index of accessibility (RAD).

Belsley *et al*. [1] used the data to examine the effects of robust estimation and published the observations in an appendix. It is also one of the few moderate sized hedonic data sets available on the Internet (via STATLIB). Many authors have used the data to illustrate various points. For example, Krasker *et al*. [5], Subramanian and Carson [8], Brieman and Friedman [3], Lange and Ryan [6], Breiman *et al*. [2], and Pace [7] have used the data to examine robust estimation, normality of residuals, and nonparametric and semiparametric estimation. Essentially, a cottage industry has sprung up around using these data to examine alternative statistical techniques.

Unfortunately, these data have some incorrectly coded observations and an unsuspected censoring problem. In the process of conducting another study, we rechecked the data against the original census data. We discovered eight miscoded dependent-variable observations, which appear in Table I. Most of the independent-variable data were not included in the 1970 Census Bureau publication, so we made no attempt to verify their accuracy. In addition to miscoded observations, we

*We thank John Boyce for his valuable comments. Both authors gratefully acknowledge the research support they have received from their respective institutions.
[†]E-mail: gilley@cab.latech.edu.
[‡]E-mail: FFRKP@aurora.alaska.edu.

0095-0696/96 $18.00

TABLE I

Miscoded Dependent Variable Observations

Observation and tract number	Median value	Corrected median value	Percentage error
8-2042	27.1	22.1	22.62%
39-2084	24.7	24.2	2.07%
119-3585	37.0	33.0	12.12%
241-3823	22.0	27.0	−18.42%
438-0905	8.7	8.2	6.1%
443-0911	18.4	14.8	24.32%
455-0923	14.9	14.4	3.47%
506-1805	11.9	19.0	−37.37%

TABLE II

Observation and Census Tract Numbers Where Censoring Occurs (Median Value $\geq$ \$50,000)

369-0107	373-0203	167-3545	226-3736
370-0108	162-3540	187-3678	258-4001
371-0201	163-3541	196-3602	268-4011
372-0202	164-3542	205-3672	284-4051

TABLE III

Estimation Results for the Harrison and Rubinfeld Data

Variable	Uncorrected OLS	Corrected OLS	TOBIT
Constant	2.84853	2.83601	1.10758
	(19.04)	(19.22)	(7.42)
CRIM	−0.01186	−0.01177	−0.01170
	(−9.53)	(−9.59)	(−9.45)
ZN	0.00008	.00009	0.00014
	(0.15)	(0.18)	(0.27)
INDUS	0.00024	0.00018	0.00101
	(0.10)	(0.08)	(0.43)
CHAS	0.09139	0.09213	0.10540
	(2.75)	(2.81)	(3.12)
NOX^2	−0.63805	−0.63724	−0.66618
	(−5.64)	(−5.71)	(−5.91)
RM^2	0.00633	0.00625	0.00666
	(4.82)	(4.83)	(5.01)
AGE	0.00009	0.00007	0.00024
	(0.17)	(0.14)	(0.45)
LDIS	−0.19125	−0.19784	−0.20454
	(−5.73)	(−6.01)	(−6.13)
LRAD	0.09571	0.08957	0.08937
	(5.00)	(4.75)	(4.69)
TAX	−0.00042	−0.00042	−0.00041
	(−3.43)	(−3.46)	(−3.38)
PTRATIO	−0.03112	−0.02960	−0.03096
	(−6.21)	(−5.99)	(−6.18)
B	0.00036	0.00036	0.00036
	(3.53)	(3.55)	(3.53)
LSTAT	−0.37116	−0.37489	−0.39122
	(−14.84)	(−15.20)	(−15.23)
σ			−0.1813
R^2	0.806	0.811	
Log-likelihood	149.955	156.979	125.532

discovered the Census Bureau censored tracts whose median value was over \$50,000. Hence, all tracts with a median value equal to or greater than \$50,000 appeared as \$50,000. Table II identifies the 16 censored observations.

To examine the sensitivity of the Harrison and Rubinfeld results to these changed data, we ran (1) the original uncorrected OLS regression, (2) the OLS regression on the corrected dependent-variable observations in the presence of censoring, and (3) a TOBIT to correct for censoring using the corrected dependent-variable observations. The results of these three regressions appear in Table III. The figures in parentheses below each estimated coefficient are t-statistics. The goodness-of-fit as measured by R^2 rises somewhat when employing the corrected observations. However, the magnitudes of the coefficients do not change much, and the qualitative results from the original regression still hold.

REFERENCES

1. David A. Belsley, Edwin Kuh, and Roy E. Welch, ''Regression Diagnostics: Identifying Influential Data and Source of Collinearity,'' Wiley, New York (1980).
2. Leo Breiman, Jerome Friedman, R. Olshen, and C. J. Stone, ''Classification and Regression Trees,'' Chapman and Hall, New York (1993).
3. Leo Breiman, and Jerome Friedman, Estimating optimal transformations for multiple regression and correlation, *J. Amer. Statist. Assoc.*, **80**, 580–619 (1985).
4. David Harrison, and Daniel L. Rubinfeld, Hedonic housing prices and the demand for clear air, *J. Environ. Econom. Management* **5**, 81–102 (1978).
5. William S. Krasker, Edwin Kuh, and Roy E. Welsch, Estimation for dirty data and flawed models, *in* ''Handbook of Econometrics,'' (Z. Griliches and M. D. Intriligator, Eds.), Vol. 1, pp. 651–698, North-Holland, Amsterdam (1983).
6. Nicholas Lange, and Louise Ryan, Assessing normality in random effects models, *Ann. Statist.* **17**, 624–42 (1989).
7. R. Kelley Pace, Nonparametric methods with application to hedonic models, *J. Real Estate Finance Econom.* **7**(3), 185–204 (1993).
8. Shankar Subramanian, and Richard T. Carson, Robust regression in the presence of heteroskedasticity, *Adv. Economet.* **7**, 85–138 (1988).

Statement of ownership, management, and circulation required by the Act of October 23, 1962, Section 4369, Title 39, United States Code: of

JOURNAL OF ENVIRONMENTAL ECONOMICS AND MANAGEMENT

Published bimonthly by Academic Press, Inc., 6277 Sea Harbor Drive, Orlando, FL 32887-4900. Number of issues published annually: 6. Editor: Dr. Ronald G. Cummings, Policy Research Center, College of Business Administration, Georgia State University, University Plaza, Atlanta, GA 30303-3083.
Owned by Academic Press, Inc., 525 B Street, Suite 1900, San Diego, CA 92101-4495. Known bondholders, mortgagees, and other security holders owning or holding 1 percent or more of total amount of bonds, mortgages, and other securities: None.
Paragraphs 2 and 3 include, in cases where the stockholder or security holder appears upon the books of the company as trustee or in any other fiduciary relation, the name of the person or corporation for whom such trustee is acting, also the statements in the two paragraphs show the affiant's full knowledge and belief as to the circumstances and conditions under which stockholders and security holders who do not appear upon the books of the company as trustees, hold stock and securities in a capacity other than that of a bona fide owner. Names and addresses of individuals who are stockholders of a corporation which itself is a stockholder or holder of bonds, mortgages, or other securities of the publishing corporation have been included in paragraphs 2 and 3 when the interests of such individuals are equivalent to 1 percent or more of the total amount of the stock or securities of the publishing corporation.
Total no. copies printed: average no. copies each issue during preceding 12 months: 1933; single issue nearest to filing date: 2100. Paid circulation (a) to term subscribers by mail, carrier delivery, or by other means: average no. copies each issue during preceding 12 months: 886; single issue nearest to filing date: 918. (b) Sales through agents, news dealers, or otherwise: average no. copies each issue during preceding 12 months: 582; single issue nearest to filing date: 672. Free distribution (a) by mail: average no. copies each issue during preceding 12 months: 85; single issue nearest to filing date: 85. (b) Outside the mail: average no. copies each issue during preceding 12 months: 6; single issue nearest to filing date: 6. Total no. of copies distributed: average no. copies each issue during preceding 12 months: 1559; single issue nearest to filing date: 1681. Percent paid and/or requested circulation: average percent each issue during preceding 12 months: 94%; single issue nearest to filing date: 95%.

(Signed) Janice M. Peterson, Director, Fulfillment and Special Projects