On the Harrison and Rubinfeld Data

By

Otis W. Gilley

Department of Economics and Finance
College of Administration and Business
Louisiana Tech University
Ruston, Louisiana 71272
(318)-257-3468

and

R. Kelley Pace
LREC Chair of Real Estate
E.J. Ourso College of Business Administration
Louisiana State University
Baton Rouge, LA 70803
(504)-388-6238
FAX: (504)-388-6366
KelleyPace@compuserve.com

We would like to thank John Boyce for his valuable comments. Both authors gratefully acknowledge the research support they have received from their respective institutions.

This manuscript is a version of the article which appeared in:

Gilley, O.W., and R. Kelley Pace, “On the Harrison and Rubinfeld Data,” *Journal of Environmental Economics and Management,* 31 (1996), 403-405.

On the Harrison and Rubinfeld Data

I. Introduction

In a well-known paper, Harrison and Rubinfeld (1978) 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 SMSA with 506 observations (one observation per census tract) on 14 non-constant 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, Kuh, and Welch (1980) used the data to examine the effects of robust estimation and published the observations in an appendix. It also is 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, Kuh, and Welch (1983), Subramanian and Carson (1988), Brieman and Friedman (1985), Lange and Ryan (1989), Breiman *et al*. (1993), and Pace (1993) have used the data to examine robust estimation, normality of residuals, 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 1. Moreover, we 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 2 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; and (3) a TOBIT using the corrected dependent variable observations. The results of these three regressions appear in Table 3. The goodness-of-fit as measured by R^2 rises somewhat when employing the corrected observations. However, the magnitudes of the coefficients de not change much and the qualitative results from the original regression still hold.

We attempted to examine the independent variables for similar problems, but we could not replicate any of these.

References

Belsley, David. A., Edwin Kuh, and Roy. E. Welch, *Regression Diagnostics: Identifying Influential Data and Source of Collinearity,* John Wiley, New York, 1980.

Breiman, Leo, Jerome Friedman, R. Olshen, and C.J. Stone, *Classification and Regression Trees*, Chapman and Hall, New York, 1993.

Breiman, Leo, and Jerome Friedman, "Estimating Optimal Transformations for Multiple Regression and Correlation," *Journal of the American Statistical Association*, Volume 80, p. 580-619, 1985.

Harrison, David, and Daniel L. Rubinfeld, "Hedonic Housing Prices and the Demand for Clean Air," *Journal of Environmental Economics and Management*, Volume 5, p. 81-102, 1978.

Krasker, William S., Edwin Kuh, and Roy E. Welsch, "Estimation for Dirty Data and Flawed Models," *Handbook of Econometrics*, Volume 1, North-Holland, Amsterdam, p. 651-98, 1983.

Lange, Nicholas, and Louise Ryan, "Assessing Normality in Random Effects Models," *Annals of Statistics*, Volume 17, p. 624-42, 1989.

Pace, R. Kelley, "Nonparametric Methods with Application to Hedonic Models," *Journal of Real Estate Finance and Economics*, Volume 7, Number 3, November 1993, p. 185-204.

Subramanian, Shankar, and Richard T. Carson, "Robust Regression in the Presence of Heteroskedasticity," *Advances in Econometrics*, JAI Press, Volume 7, p. 85-138, 1988.

Table 1 — 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 2 — Observation and Census Tract Numbers Where Censoring Occurs (Median Value ≥ $50,000)

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

Table 3 — 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