Case study source: *Marketing Data Science* by Miller [1]

Boston Housing Study

Thomas W. Miller

Abstract
Data from the Boston Housing Study have been used to demonstrate various regression methods over the years. This brief write-up cites sources of previously published studies.

Contents

Overview

The Boston Housing Study is a market response study of sorts, with the market being 506 census tracts in the Boston metropolitan area. The objective of the study was to examine the effect of air pollution on housing prices, controlling for the effects of other explanatory variables. The response variable is the median price of homes in the census track. Table 1 shows variables included in the case. Short variable names correspond to those used in previously published studies.

The original data from the Boston Housing Study Harrison and Rubinfeld [2] were published by Belsley, Kuh, and Welsch in their book about regression diagnostics [3] . In subsequent years, versions of these data have been used by statisticians to introduce and evaluate regression methods, including classification and regression trees by Breiman et al. [4], treed regression [5], and monotone regression [6]. Miller [7] used the Boston Housing Study data to explore sample size requirements for a number of modern data-adaptive regression methods. Data provided for this case represent an updated version of the original data, following the suggested revisions of Gilley and Pace [8].

Data for the Boston Housing Study are available from the GitHub site for the *Modeling Techniques* series:

```
https://github.com/mtpa/
```

Go to the repository for the *Marketing Data Science* book and locate the file `boston.csv`:

```
mtpa
    mds
       MDS\_Appendix\_C
           MDS\_Appendix\_C\_4
                 boston.csv
```

References

[1] Thomas W. Miller. *Marketing Data Science: Modeling Techniques in Predictive Analytics with R and Python*. Pearson Education, Old Tappan, N.J., 2015. Data sets and programs available at http://www.ftpress.com/miller/ and https://github.com/mtpa/.

[2] D. Harrison and D.L. Rubinfeld. Hedonic housing prices and the demand for clean air. *Journal of Environmental Economics and Management*, 5:81–102, 1978.

[3] David A. Belsley, Edwin Kuh, and Roy E. Welsch. *Regression Diagnostics: Identifying Influential Data and Sources of Collinearity*. Wiley, New York, 1980.

[4] Leo Breiman, Jerome H. Friedman, Richard A. Olshen, and Charles J. Stone. *Classification and Regression Trees*. Chapman & Hall, New York, 1984.

[5] W.P. Alexander and S.D. Grimshaw. Treed regression. *Journal of Computational and Graphical Statistics*, 5(2):156–175, 1996.

[6] D. Dole. CoSmo: A constrained scatterplot smoother for estimating convex, monotonic transformations. *Journal of Business and Economic Statistics*, 17(4):444–455, 1999.

[7] Thomas W. Miller. The Boston splits: Sample size requirements for modern regression. *1999 Proceedings of the Statistical Computing Section of the American Statistical Association*, pages 210–215, 1999.

[8] O.W. Gilley and R.K. Pace. On the Harrison and Rubinfeld data. *Journal of Environmental Economics and Management*, 31:403–405, 1996.

Table 1. Boston Housing Study Variables

Variable Name	*Description*
neighborhood	Name of the Boston neighborhood (location of the census tract)
mv	Median value of homes in thousands of 1970 dollars
nox	Air pollution (nitrogen oxide concentration)
crim	Crime rate
zn	Percent of land zoned for lots
indus	Percent of business that is industrial or nonretail
chas	On the Charles River (1) or not (0)
rooms	Average number of rooms per home
age	Percentage of homes built before 1940
dis	Weighted distance to employment centers
rad	Accessibility to radial highways
tax	Tax rate
ptratio	Pupil/teacher ratio in public schools
lstat	Percentage of population of lower socio-economic status