UNIVERSIDAD SANTO TOMAS
PRIMER CLAUSTRO UNIVERSITARIO DE COLOMBIA

Comunicaciones en Estadística
Diciembre 2024, Vol. 17, No. 2, pp. 1–13

Case Study of Bayesian Multiple Linear Regression Using the Boston Housing Dataset

Modelado Jerárquico Bayesiano en Entornos de Regresión Lineal

Juan Sosa[a]
jcsosam@unal.edu.co

Abstract

We analyze the Boston housing dataset using multiple linear regression and ordinary least squares techniques. Various models are fitted to the data to take advantage of the predictive power of the explanatory variables, with careful evaluation of each model's assumptions and a comparative analysis among them. Relevant explanatory variables that significantly impact the response variable are identified. Additionally, cross-validation experiments are conducted on select models from the analysis. Finally, we highlight certain limitations of OLS and propose the use of ridge regression techniques as an alternative.

***Keywords*:** Bayesian statistics; Boston housing dataset; cross-validation; multiple linear regression; ordinary least squares.

Resumen

Analizamos el conjunto de datos de viviendas de Boston utilizando técnicas de regresión lineal múltiple y mínimos cuadrados ordinarios. Se ajustan varios modelos a los datos para mejorar el poder predictivo de las variables explicativas, evaluando cuidadosamente las suposiciones de cada modelo y realizando un análisis comparativo entre ellos. Identificamos las variables explicativas relevantes que tienen un impacto significativo sobre la variable de respuesta. Además, realizamos experimentos de validación cruzada en modelos seleccionados del análisis. Finalmente, destacamos ciertas limitaciones de los OLS y proponemos el uso de técnicas de regresión ridge como alternativa.

***Palabras clave*:** Estadística bayesiana; conjunto de datos de viviendas de Boston; validación cruzada; regresión lineal múltiple; mínimos cuadrados ordinarios.

[a]Departamento de Estadística, Universidad Nacional de Colombia

1. Introduction

In this study, we analyze the Boston housing dataset to investigate the factors influencing housing prices in suburban Boston. The Boston housing dataset, introduced by Harrison and Rubinfeld (1978), comprises various attributes describing housing values in the suburbs of Boston. This dataset, sourced from the StatLib library maintained at Carnegie Mellon University, is accessible online at `https://archive.ics.uci.edu/ml/datasets/Housing`. It has been extensively analyzed across diverse scenarios, and as noted by Belsley et al. (2005, p. 229), Harrison and Rubinfeld primarily aimed to study the effect of air pollution (measured by nitric oxide concentration) on the median value of owner-occupied homes (MEDV, in $1,000s). Along with MEDV, the dataset includes thirteen additional explanatory variables representing qualities influencing housing prices.

This dataset has become a foundational tool for assessing regression models and exploring relationships between socio-economic and environmental factors and property values. The original study employed hedonic pricing models to quantify the influence of air pollution, among other variables, on housing prices. Since then, the dataset has been widely utilized to evaluate statistical modeling techniques, including multiple linear regression, transformations, and diagnostic assessments, as emphasized by Belsley et al. (2005). Their work on regression diagnostics and ridge regression remains critical for addressing multicollinearity and influential observations, particularly in datasets like Boston Housing, where these challenges frequently arise.

More recently, machine learning and advanced statistical techniques have extended the analytical capabilities for this dataset. Tibshirani (1996) introduced lasso regression as a means to enforce sparsity in regression coefficients, improving interpretability and predictive performance. Zou and Hastie (2005) further developed the Elastic Net, combining the strengths of ridge and lasso regression, which is especially valuable when dealing with highly correlated predictors. James et al. (2013) showcased the utility of the dataset in demonstrating the application of linear models, tree-based methods, and ensemble techniques. Additionally, Breiman (2001) highlighted its use in comparing bagging, boosting, and random forests, illustrating the dataset's versatility in exploring advanced predictive models. These studies collectively underscore the Boston Housing dataset's continued relevance in developing and testing new methodologies, offering insights into both traditional statistical approaches and modern machine learning frameworks.

The characteristics potentially influencing the median value of owner-occupied homes (MEDV) were gathered at the suburb level in Boston. The explanatory variables included in the dataset are as follows:

1. **CRIM**: Per capita crime rate by town.
2. **ZN**: Proportion of residential land zoned for lots over 25,000 square feet.
3. **INDUS**: Proportion of non-retail business acres per town.

4. **CHAS**: Charles River dummy variable (1 if the tract bounds the river; 0 otherwise).
5. **NOX**: Nitric oxides concentration (parts per 10 million).
6. **RM**: Average number of rooms per dwelling.
7. **AGE**: Proportion of owner-occupied units built before 1940.
8. **DIS**: Weighted distances to five Boston employment centers.
9. **RAD**: Index of accessibility to radial highways.
10. **TAX**: Full-value property tax rate per $10,000.
11. **PTRATIO**: Pupil-teacher ratio by town.
12. **B**: $1000(B_k - 0.63)^2$, where B_k is the proportion of Black residents by town.
13. **LSTAT**: Percentage of the population with lower socioeconomic status.

Our work builds on the foundational study by Harrison and Rubinfeld (1978). We follow diagnostic and model refinement techniques outlined by Belsley et al. (2005) to identify influential data points and address multicollinearity issues. Additionally, we adopt methodological insights from Faraway (2004) regarding variable transformations and regression modeling to enhance the robustness and interpretability of our results.

That is why, here, we apply multiple linear regression and ordinary least squares (OLS) techniques to model the relationship between the median value of owner-occupied homes (MEDV) and the explanatory variables given above. Notice, however, that OLS has several limitations, which we address specifically throughout this manuscript. For starters, it is highly sensitive to outliers and multicollinearity, leading to unstable or biased coefficient estimates. OLS assumes homoscedasticity (constant variance of residuals) and normality, making it unsuitable for data with heteroscedasticity or non-normal distributions. It is also restricted to linear relationships and prone to overfitting in high-dimensional settings. Additionally, measurement errors in predictors can introduce bias, and OLS cannot handle missing data without imputation. These limitations, along with violations of assumptions or model misspecifications, reduce its reliability, necessitating the use of alternative approaches such as robust, ridge, or lasso regression in complex scenarios.

Our methodology begins with an exploratory data analysis to examine the distributions and relationships of the variables. We then fit an initial regression model and iteratively refine it by removing non-significant predictors, such as INDUS and AGE, and incorporating quadratic terms for variables like CRIM, RM, and LSTAT, which exhibit non-linear relationships. To address the issue of non-normality in residuals, we apply a log transformation to the response variable, following guidelines from Faraway (2004). Throughout the process, we use diagnostic techniques to detect influential observations and assess the robustness of the models (Belsley et al., 2005).

Our objectives are to improve the predictive quality of the models, identify the explanatory variables that significantly affect housing prices, and address the limitations of OLS by considering alternative approaches such as ridge regression (Belsley et al., 2005). Specifically, we aim to highlight the critical roles of variables like nitric oxide concentration (NOX) and the average number of rooms per dwelling (RM) in determining housing prices. To ensure the predictive power of the models, we perform a 10-fold cross-validation and compare the mean squared errors across different model configurations.

The paper is organized as follows: In **Exploratory Data Analysis**, the distributions and relationships among variables are examined to identify necessary transformations. **Initial Models and Diagnostics** outlines the baseline modeling using multiple linear regression and OLS, followed by refinement through variable selection and diagnostics. The **Transformations and Variable Selection** section addresses non-linearities and applies transformations to improve model fit. Finally, in **Discussion and Conclusions**, the results are summarized, limitations of OLS are highlighted, and ridge regression is recommended for future work.

2. Exploratory Data Analysis

The dataset comprises $N = 506$ instances and 13 predictors, capturing various attributes relevant to housing values. While most predictors are continuous, CHAS is a binary variable indicating proximity to the Charles River, included in the model as a dummy variable with two levels. Importantly, the dataset is complete, with no missing values, ensuring reliable statistical analysis.

Preliminary numerical and graphical summaries, including quantiles, boxplots, and histograms, highlight significant differences in the range and variability of the explanatory variables. For example, NOX values range from 0.385 to 0.871, while B spans from 0.32 to 396.9, reflecting the diverse scales of these attributes, such as concentrations, indexes, proportions, and physical measures. Overall, the 13 explanatory variables exhibit varied distributional shapes, from clear symmetry, as seen with AGE, to pronounced skewness, observed in variables like CRIM, NOX, B, and LSTAT.

The response variable exhibits a long right tail, a characteristic commonly observed in economic variables such as values and incomes, which naturally display a heavy decay toward higher values. This marginal behavior suggests that a transformation may be necessary to satisfy certain model assumptions, enhancing the precision of inferences and predictions. Additionally, scatterplots of MEDV against individual predictors reveal varied relationships: some regressors, like LSTAT, show a clear non-linear impact, while others, such as AGE, present a less defined dependency.

3. Initial Models and Diagnostics

We begin by fitting a multiple linear regression model that includes all 13 predictors in a linear form, as follows:

$$\text{MEDV}_i = \beta_0 + \beta_1 \text{CRIM}_i + \beta_2 \text{ZN}_i + \ldots + \beta_{13} \text{LSTAT}_i + \epsilon_i \,, \tag{1}$$

where $\epsilon_i \overset{\text{iid}}{\sim} \mathsf{N}(0, \sigma^2)$, for $i = 1, \ldots, 506$. This initial model is statistically significant ($p \approx 0.00$) and achieves an adjusted R^2 value of 0.734. However, two predictors—INDUS ($p = 0.738$) and AGE ($p = 0.958$)—are not statistically significant. Conversely, key variables central to the study, such as NOX ($p \approx 0.00$) and RM ($p \approx 0.00$), show high significance, emphasizing their relevance in explaining the response variable.

Before checking assumptions, we remove the predictors INDUS and AGE from the model since they turned out to be nonsignificant (this is confirmed by a stepwise procedure), and then we readjust model (1) without considering these two variables. By doing so, this model with 11 regressors is still significant ($p \approx 0.00$) with an $R^2_{\text{adj}} = 0.734$. We note that now each independent variable is significant, and the variability explained by the model has not changed in comparison with model (1). We also perform an F-test to evaluate the assessment of this reduced model in comparison to the saturated model, as in (Faraway, 2004, pp. 26–31), and we fail to reject the null hypothesis; therefore, the reduced model is preferred ($p = 0.944$).

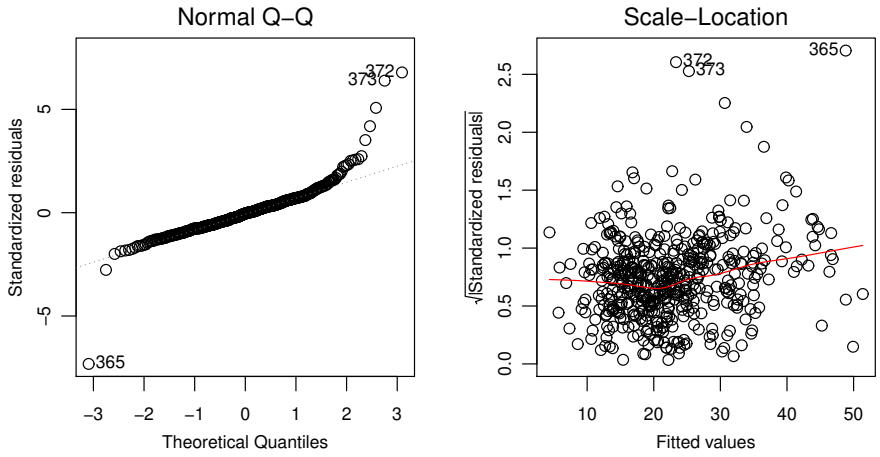
Now, we check the assumptions in this (untransformed) model with 11 explanatory variables (all but INDUS and AGE). The nonlinear pattern in the plot of standardized residuals versus fitted values in the first panel of Figure 1 strongly suggests that some predictors should enter the model in a quadratic or cubic fashion, for example. Furthermore, the normal Q-Q plot of the residuals in the second panel of Figure 1 shows clear deviations from the normality assumption: the distribution of the residuals is clearly skewed with a heavy right tail. Both figures reveal potential outliers and/or influential observations.

4. Transformations and Variable Selection

Before checking more assumptions in detail, we recognize the need to modify the model to improve its fit and satisfy key assumptions. This involves, first, incorporating quadratic or cubic terms for predictors that exhibit non-linear relationships, and second, applying an appropriate transformation to the response variable to address skewness and improve normality.

To improve the model, we first examine the plots of residuals versus each regressor to identify terms that should be included in a higher-order form. Based on this analysis, we add quadratic terms for CRIM, RM, and LSTAT to the model (with

Figure 1: Diagnostic plots (untransformed response variable).



11 predictors), as these variables exhibit relationships beyond linearity with the residuals. Cubic terms are not included because they turn out to be nonsignificant (as justified in Faraway (2004, p. 123)). After incorporating these quadratic terms, the residuals versus fitted values plot indicates that no further higher-order terms are necessary. The resulting model, which now includes 14 predictors, is significant ($p \approx 0.00$) with an $R^2_{\text{adj}} = 0.814$, reflecting a substantial increase in the explained variability. However, the normal Q-Q plot of residuals continues to display serious deviations from normality, including skewness and a heavy right tail, with residuals ranging from -7.318 to 6.788.

Before going any further, we proceed to transform the response variable to correct the skewness observed in the residuals (while acknowledging that the long tail behavior may persist, as transformations primarily aim to symmetrize distributions). We choose to work with MEDV in the log scale, a popular choice for positive data, partly for interpretability. MEDV is better analyzed on a multiplicative rather than additive scale since it measures value. For instance, \$1,000 holds far greater relative importance to a poor person than to a millionaire, as \$1,000 represents a much larger fraction of the poor person's wealth (Faraway, 2004, p. 165). Additionally, we apply the Box-Cox transformation to the response in the model with 14 regressors, yielding $\hat{\lambda}_{\text{MLE}} = 0.298$ and $\text{CI}_{95\,\%}(\lambda) = (0.423, 1.181)$, which confirms that the log transformation is a suitable choice.

Thus, we adjust a model using the log scale for MEDV, which remains significant ($p \approx 0.00$) with an $R^2_{\text{adj}} = 0.819$, nearly the same as in the previous case. However, ZN now turns out to be nonsignificant ($p = 0.215$), allowing us to remove it from the model. Notably, the Box-Cox transformation with $\lambda = 0.298$ yields the same result. Consequently, we arrive at a transformed model with the response variable in the log scale and 13 significant predictors, three of which are included

in quadratic form. The final model is expressed as:

$$\begin{aligned}\log(\text{MEDV}_i) = \beta_0 &+ \beta_1 \text{CRIM}_i + \beta_2 \text{CHAS}_i + \beta_3 \text{NOX}_i \\ &+ \beta_4 \text{RM}_i + \beta_5 \text{DIS}_i + \beta_6 \text{RAD}_i + \beta_7 \text{TAX}_i \\ &+ \beta_8 \text{PTRATIO}_i + \beta_9 \text{B}_i + \beta_{10} \text{LSTAT}_i \\ &+ \beta_{11} \text{CRIM}_i^2 + \beta_{12} \text{RM}_i^2 + \beta_{13} \text{LSTAT}_i^2 + \epsilon_i \,, \end{aligned} \tag{2}$$

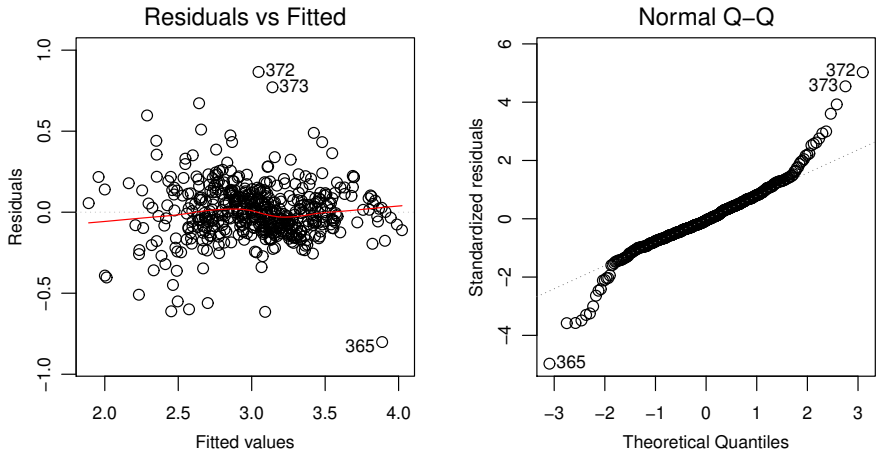
where $\epsilon_i \overset{\text{iid}}{\sim} \mathsf{N}(0, \sigma^2)$, for $i = 1, \ldots, 506$.

5. Final Model and Diagnostics

Working with model (2), the plot of residuals versus fitted values in the first panel of Figure 2 is well behaved, indicating no issues with nonlinearity. However, the normal Q-Q plot of the residuals in the second panel of Figure 2 reveals substantial departures from normality due to the heavy tails in the distribution. The standardized residuals now range from -4.943 to 5.014, reflecting the correction of skewness in the residuals. Additionally, a simple linear regression of the fitted values versus the residuals (as described in Faraway (2004, p. 59)) does not suggest any symptoms of nonconstant variance in the residuals. From these diagnostic plots, we identify observations 365, 372, and 373 as hightailed, marking them as influential and/or outliers.

The Bonferroni corrected test (Monahan, 2011, p. 143), based on externally studentized (or crossvalidated) residuals, indicates that observations 365, 372, 373,

Figure 2: Diagnostic plots (transformed response variable).



and 410 could be considered outliers (see Table 1) and warrant further examination. Observations 365, 372, and 373 had already been identified previously, and their residuals and p-values confirm them as critical cases. Observation 410, however, appears to be an outlier primarily due to the heavy tails in the residual distribution.

Additionally, following the rule of thumb (Faraway, 2004, p. 69), which states that leverages greater than $2p/N = 0.0553$ should be examined more closely, we identify 41 observations exceeding this threshold (including observation 365). Are all of these values unusual? A half-normal plot of these leverage values (Faraway, 2004, p. 70) suggests that observations 381 and 366 stand out significantly, indicating that their leverage is unusually high compared to the others.

Table 1: Bonferroni corrected test based on externally studentized residuals.

Obs.	r_i	p
372	5.16	0.000
365	-5.10	0.000
373	4.64	0.002
410	3.98	0.039

An influential point may or may not be an outlier and may or may not have large leverage, but it will tend to have at least one of these two properties (Faraway, 2004, p. 75). We identify 37 observations with D-values (Cook's distance values) greater than $4/(N-p) = 0.008$. Once again, which of these values are truly unusually high? Influential plots, such as the one shown in Figure 3, indicate that special care should be taken with observations 365, 372, 373, 406, 410, and 413. Therefore, considering the heavy tail distribution of the residuals, and based on their residuals and D-values, we conclude that observations 365, 373, and 372 are indeed influential.

Finally, we decide to adjust a model excluding observations 365, 373, and 372. This model (see Table 2) remains significant ($p \approx 0.00$) with an $R^2_{\text{adj}} = 0.8423$, indicating an improvement compared to model 2. Diagnostic plots exhibit similar behavior as before, but this time the standardized residuals range from -3.802 to 4.245, a smaller range than in model 2. Examining other diagnostic plots, we still detect candidates for outliers and/or influential observations, but these cases are attributed to the long-tail distribution of the residuals rather than other issues. Notably, one of the key reasons for excluding these three observations is the significant changes detected in the coefficients, as illustrated in Table 3. For example, we observe changes of 10 % and 20 % for NOX and RM, respectively, which are variables of primary interest in this study.

Figure 3: Plot of influential observations.

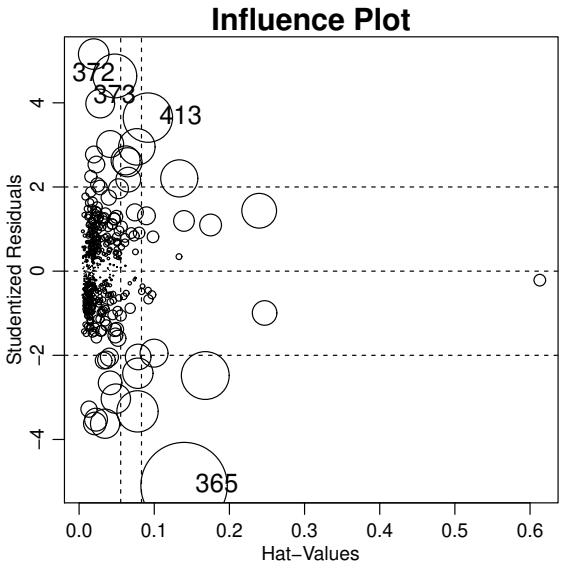


Table 2: Summary of the final model.

Variable	Estimate	Std. Error	t value	p
(Intercept)	7.1007	0.3815	18.61	0.0000
CRIM	-0.0273	0.0034	-8.10	0.0000
CRIM^2	0.0002	0.0000	5.01	0.0000
CHAS1	0.0910	0.0299	3.05	0.0025
NOX	-0.6203	0.1212	-5.12	0.0000
RM	-0.9109	0.1164	-7.83	0.0000
RM^2	0.0787	0.0091	8.63	0.0000
DIS	-0.0359	0.0056	-6.41	0.0000
RAD	0.0156	0.0024	6.50	0.0000
TAX	-0.0004	0.0001	-3.94	0.0001
PTRATIO	-0.0303	0.0043	-7.08	0.0000
B	0.0003	0.0001	2.89	0.0040
LSTAT	-0.0431	0.0049	-8.87	0.0000
LSTAT^2	0.0005	0.0001	3.66	0.0003

6. Discussion and Concussions

In this section, we summarize all our findings and discuss the results of the analysis. First, we observe that the distribution of the residuals in all adjusted models exhibits heavier tails than the normal distribution. Consequently, it becomes natural to consider estimators that are more robust than OLS for handling departures from normality in the error structure, as suggested by Belsley et al. (2005, pp. 229–

Table 3: Percentual change in the coefficients due to deletion of influential observations. A negative sign indicates a decrease.

$CRIM^2$	NOX	RM	RM^2
9.54	-10.17	20.34	20.41
DIS	PTRATIO	LSTAT	$LSTAT^2$
-14.84	-8.63	-9.46	-15.50

244). Furthermore, we note that the large proportion of observations flagged by the diagnostics as requiring further attention is attributable not only to a high number of leverage points but also to the non-Gaussian nature of the error distribution.

We consider six models since the analysis itself naturally leads us to consider several alternatives:

1. Saturated model including all baseline predictors and no transformations of any kind, see equation (1).
2. Model as in 1, but removing two predictors, namely, INDUS and AGE.
3. Model as in 2, adding quadratic predictors, namely, $CRIM^2$, RM^2, and $LSTAT^2$.
4. Model as in 3, adjusting the response variable to the log scale.
5. Model as in 4, removing one predictor, ZN (see equation (2)).
6. Model as in 5, removing influential observations.

Table 4 summarizes models 1 to 6 in detail. We observe that the strategy of removing nonsignificant variables and transforming the response variable works effectively, as it consistently increases R^2_{adj} while reducing both $\hat{\sigma}$ and the information criteria (AIC and BIC). We include an additional division in the table to explicitly highlight the change in the response variable's scale. Furthermore, we note that model 3 represents the best option for interpretability without using the log scale. On the other hand, model 6 emerges as the most suitable for predictive purposes without resorting to ridge regression, as suggested by Belsley et al. (2005).

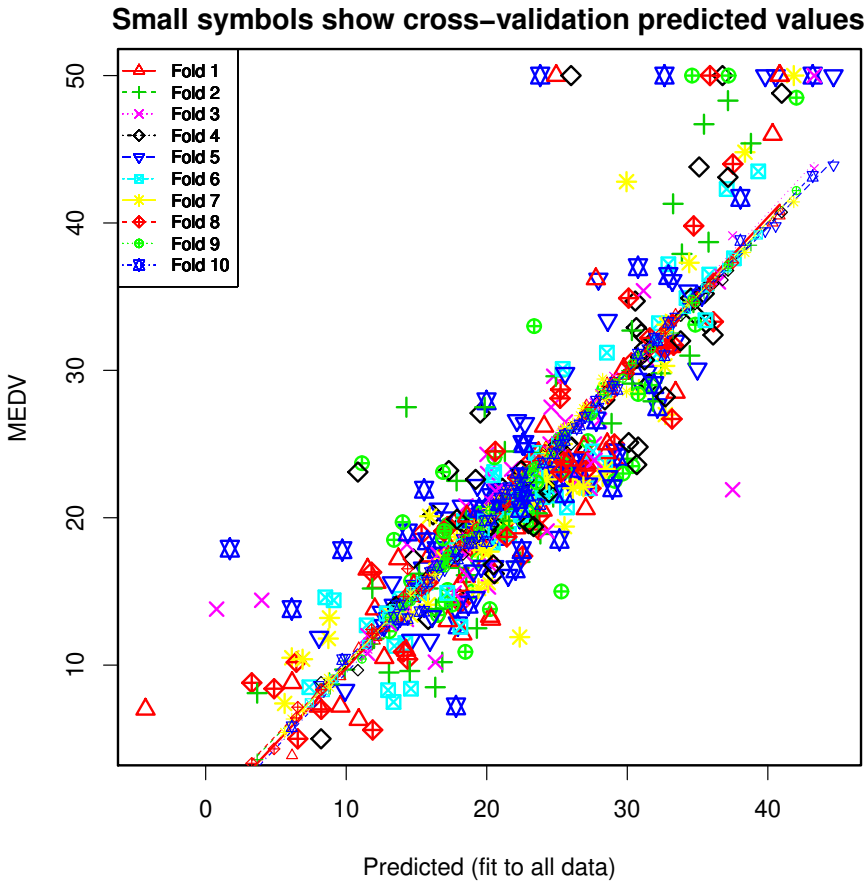
In relation to the predictors included in the model, we observe that NOX (the most important regressor according to the research) is significant in all the models considered and has a negative effect on MEDV. This result is expected, as higher levels of NOX are likely to decrease the value of owner-occupied homes. The NOX coefficient appears to be well-determined based on row-deletion diagnostics, indicating that its magnitude is not strongly influenced by data perturbations. Furthermore, since $N \gg p$, we find that both inferences and predictions are robust to departures from normality. Lastly, we note that the proportion of owner-occupied units

Table 4: Models' information. M := Model. Tr.? := Is the model using MEDV in log scale? #N := Number of nonsignificant predictors in the model.

M	N	p	Tr.?	R^2_{adj}	#N	$\hat{\sigma}$	AIC	BIC	NOX
1	506	14	No	0.73	2	4.75	3027.6	3091.0	-17.8
2	506	12	No	0.73	0	4.74	3023.7	3078.7	-17.4
3	506	15	No	0.81	0	3.97	2847.1	2914.7	-14.6
4	506	15	Yes	0.82	1	0.17	-318.7	-251.1	-0.67
5	506	14	Yes	0.82	0	0.17	-319.1	-255.7	-0.68
6	503	14	Yes	0.84	0	0.16	-392.0	-328.7	-0.62

built prior to 1940 and the proportion of non-retail business acres per town do not significantly impact the value of owner-occupied homes. Additionally, we have serious doubts regarding the relevance of the proportion of residential land zoned

Figure 4: 10-fold cross-validation using model 1.



for lots over 25,000 sq.ft. in predicting MEDV.

All models report three strong influential observations, namely, observations 365, 372, and 373. Referring to the source of these observations (Belsley et al., 2005, p. 230), we find that observation 365 originates from Back Bay, while observations 372 and 373 are from Beacon Hill. Examining all candidates identified as influential and/or outliers in model 5, we detect 37 such observations, which are clustered within specific towns. This clustering supports our suspicion that a geographical factor influencing housing values has not been accounted for in the model. As noted by Belsley et al. (2005, p. 243), ridge regression highlights that influential data points often concentrate heavily within a few neighborhoods, suggesting that the housing-price equation may not be as well specified as it could be. Lastly, none of the models presented issues with homoscedasticity.

Finally, the `DAAG` package in `R`, which provides a function called `CVlm`, allows us to perform K-fold cross-validation. This method randomly removes K-folds for the testing set and fits the model using the remaining (training set) data. At the bottom of the output, the cross-validation residual sums of squares (overall MS) is reported. The ten-fold cross-validation shown in Figure 4 corresponds to model 6, for which the overall MSE is 0.0282. According to this plot, these experiments yield consistent results, and the low value of the reported MSE confirms the high prediction quality of this model. Additionally, we perform ten-fold cross-validation on models 1 and 3 (both with the same scale for the response variable) and obtain MSE values of 24 and 17.5, respectively. This indicates that, as expected, removing nonsignificant predictors and incorporating nonlinear terms significantly improves the model's predictive capacity.

Statements and Declarations

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Recibido: Julio 3 de 2024
Aceptado: Febrero 10 de 2025

Referencias

D. Belsley, E. Kuh, and R. Welsch. *Regression Diagnostics: Identifying Influential Data and Sources of Collinearity*. Wiley Series in Probability and Statistics. Wiley, 2005. ISBN 9780471725145. URL `http://books.google.com/books?id=GECBEUJVNe0C`.

L. Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001.

J. Faraway. *Linear Models with R*. Chapman & Hall/CRC Texts in Statistical Science. Taylor & Francis, 2004. ISBN 9780203507278. URL `http://books.google.com/books?id=fvenzpofkagC`.

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

G. James, D. Witten, T. Hastie, and R. Tibshirani. *An Introduction to Statistical Learning with Applications in R*. Springer, 2013.

J. Monahan. *A Primer on Linear Models*. Chapman & Hall/CRC Texts in Statistical Science. Taylor & Francis, 2011. ISBN 9781420062045. URL `http://books.google.com/books?id=ysdjfK6nrXkC`.

R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 58(1):267–288, 1996.

H. Zou and T. Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67 (2):301–320, 2005.