Assignment 1

MACHINE LEARNING

Fergus Walsh*

Due 5th Febuary, Epiphany Term, 2021

1 Introduction

This enquiry will analyse the `Boston` data set, which records median house values and a variety of social and environmental factors in 506 suburbs of Boston, U.S.A, first published in 1979.[1] The fourteen variables are detailed in Table 1. `medv` will be the response variable in this enquiry, with the thirteen others as predictors in the regression models.

Variable	Definition
`crim`	Crime rate per capita.
`zn`	Proportion of residential land zoned for lots over 25,000 sq. ft.
`indus`	Proportion of non-retail business acres per town.
`chas`	Charles River (`1` if town bounds river, `0` if not).
`nox`	Nitrogen oxides concentration (parts per 10 million).
`rm`	Average number of rooms per dwelling.
`age`	Proportion of owner-occupied units built before 1940.
`dis`	Weighted mean of distances to five Boston employment centres.
`rad`	Index of accessibility to radial highways.
`tax`	Full-value property tax rate per $10,000
`ptratio`	Pupil:teacher ratio by town.
`black`	Proportion of African-Americans per town. $1000 \times (Proportion - 0.63)^2$
`lstat`	Per cent lower socio-economic status.
`medv`	Median value of owner-occupied homes ($1,000s).

Table 1: Variables of `Boston` data set. 'MASS' documentation, Ripley 2020b, pp. 20–21

Four candidate models will be created using different methods, then these will then be compared using model performance statistics and by performing cross-valdiation to

*R code used in this assignment can be found at `https://github.com/FergusJPWalsh/Master-of-Data-Science/blob/main/ML%20Assignment%201%20FJPW.R`

[1] Harrison and Rubinfeld 1979.

assess their predictive performance.

2 Regression Analysis

2.1 Model 1: All Variables

The first candidate model is a least squares regression model with all thirteen predictor variables, generated using the `glm()` function in R (Model 1). Its model performance statistics are given in Table 2. It is included here as a base-line to compare the subsequent methods.

2.2 Model 2: Variable Selection with Best Subset Algorithm

The best subset algorithm is a method which compares all possible combinations of variables in the least squares regression, and selects k candidate models (where k = 1,2,3...p number of predictor variables) with the smallest residual sum of squares (R.S.S.) and highest R^2 value. Best subset selection can be prohibitively expensive computationally if p is very large, since 2^p models must be generated and compared. However, $2^{13} = 8,192$ which is a feasible number of candidate models to compare.[2] There is no reason, therefore, to use a stepwise selection algorithm instead of best subset selection.

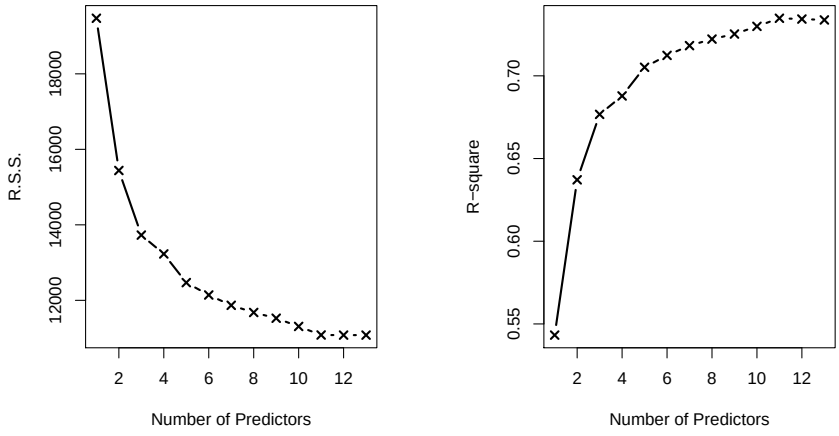

Figure 1: R.S.S. (left) and R^2 (right) values for the optimum candidate models for a given number of predictors as selected by the best subset algorithm.

[2] James et al. 2013, p. 209.

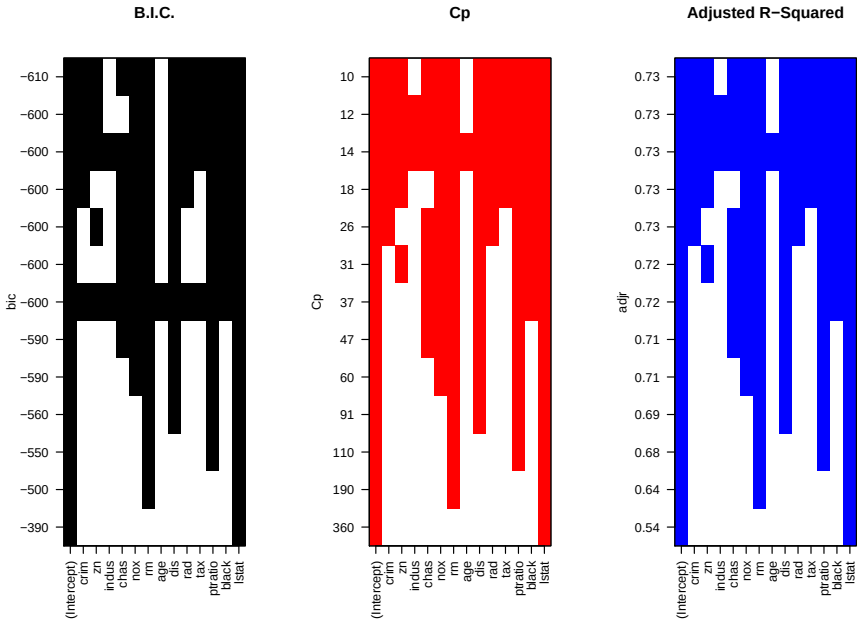

Figure 2: Variables present in each model when solving for optimum B.I.C. (left), C_p (centre) and adjusted R^2 (right) using best subset algorithm.

Statistic	Model 1	Model 2
Adjusted R^2	0.734	0.735
R.S.E.	4.745	4.736
B.I.C.	3,091.007	3,078.671

Table 2: Model Summary Statistics (3 d.p.)

Figure 1 plots the R.S.S. and R^2 values for the best candidate model for a given number of predictors. As expected, the model fit increases as more variables are added. However, to avoid over-fitting and find the correct balance between bias and variance, other model performance statistics must be considered. Figure 2 shows how the algorithm chooses amongst the variables in creating the k candidate models, using the Bayseian information criterion (B.I.C.), the C_p estimate and adjusted R^2 values. B.I.C. and C_p estimate the model's predictive performance, while R^2 measures the model fit.[3] Figure 3 shows these statistics plotted against the number of variables, and with the optimum model among the k candidates marked with a red asterix. In the case of B.I.C. and C_p the best model has the lowest value (i.e. lowest estimated error in prediction). As can be seen, in all three measures, the model with eleven predictors is the optimum

[3] James et al. 2013, pp. 211–213.

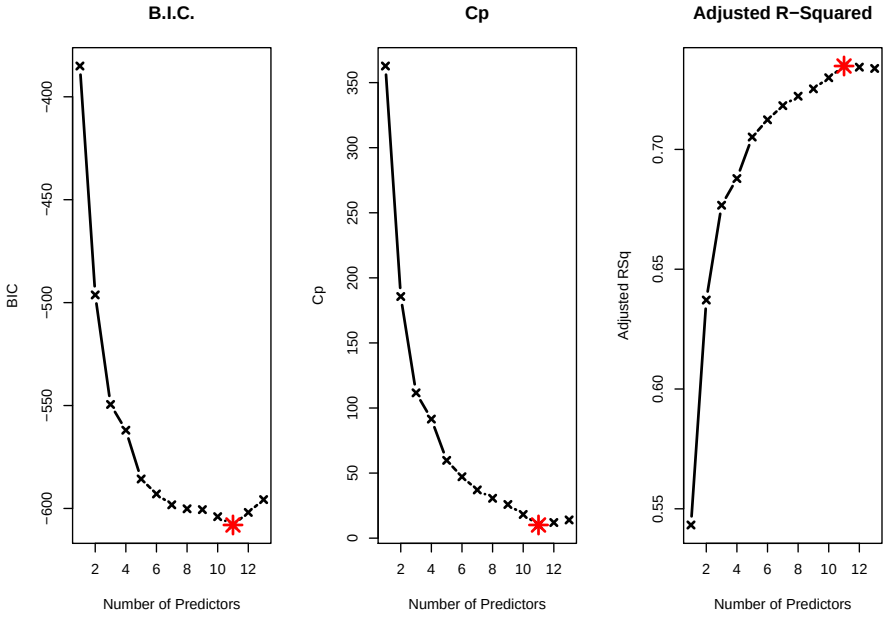

Figure 3: B.I.C. (left), C_p (centre) and adjusted R^2 (right) plotted against number of variables in k candidate models. The red asterix indicates the optimum model in each measure. In this case the model with eleven predictors is the best in each three measures.

candidate. Moreover, the predictor variables chosen in each case are the same: `indus` and `age` being left out.

If there was any doubt that the 11-variable model was the best, then cross-validation could be used to test the predictive performance of the k models. However, since all measures point to the 11-variable model, cross validation is not necessary in this instance.

As can be seen in Table 2, the regression with eleven variables as chosen by best subset selection has a slightly better fit to the data and a similarly small improvement in estimated prediction accuracy when compared to the least squares regression with all thirteen predictors.

2.3 Model 3: Ridge Regression

Rather than remove variables from the model, the ridge regression limits the value of the coefficents of each predictor through a tuning parameter λ, such that both the

residual sum of squares (R.S.S.) and the variance is minimized.[4]

$$R.S.S. + \lambda \sum_{j=1}^{p} \beta_j^2$$

The tuning parameter λ is selected by using k-fold cross-valdiation to find the value of λ where the model has the smallest mean square error (M.S.E.). Once the optimum λ value has been selected, the model is then re-fit using the whole data set to estimate the coefficents.[5]

This whole process can be performed using the `glmnet(x, y, alpha = 0)` and `cv.glmnet()` functions in R. These functions compare a range of λ values and return the minimum λ value and the minimum λ value where the cross-validation statistic is within one standard error of that of the model with the minimum λ value (1 S.E. λ). These functions were run fifty times and the mean of two λ values were calculated. These are plotted in Figure 4 and given in Table 3, along with the corresponding M.S.E. of the k-fold cross-valdiation test.

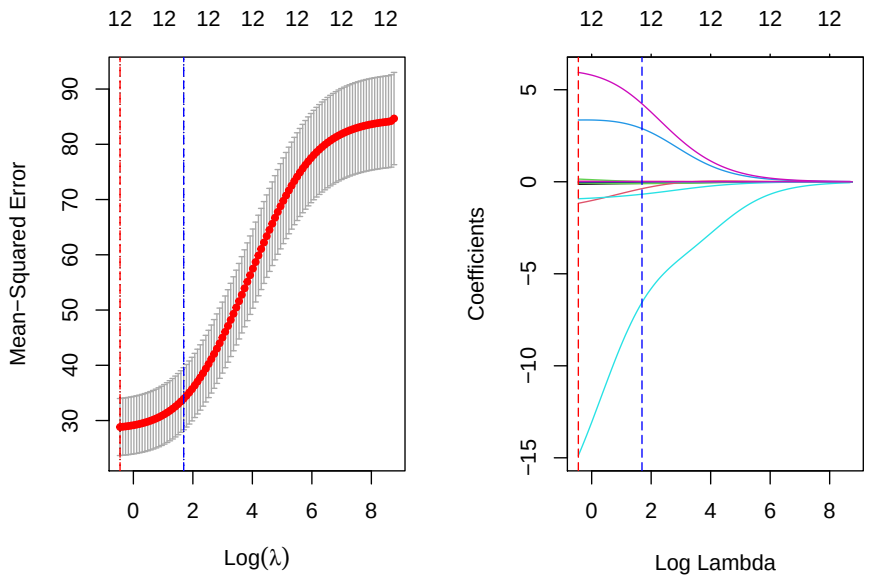

Figure 4: M.S.E. of Model 3 as λ increases (left). Values of Model 3 coefficents as λ increases (right). Minimum λ (red) and 1 S.E. λ (blue).

The 1 S.E. λ value gives Model 3 markedly different coefficents and a worse prediction performance. It may be asked why the 1 S.E. value should be considered at all,

[4] James et al. 2013, p. 215; Friedman, Hastie, and Tibshirani 2009, pp. 61–62.
[5] James et al. 2013, p. 227.

	Min λ	1 S.E. λ
λ Value	0.678	4.041
M.S.E.	24.089	27.548

Table 3: Optimum λ Values and Model 3 M.S.E. (3 d.p.).

considering a smaller λ value gives better predictive performance in this case, as can be seen in Figure 4. However, the purpose of the ridge regression is to reduce the variability of the model, and the more conservative approach of using the 1 S.E. λ further reduces the variability of the coefficents, though the bias of the model suffers a slight increase as a result.[6]

2.4 Model 4: Lasso Regression

The lasso regression functions much like the ridge regression, except that the ℓ_1 norm of the coefficents β_j are used:

$$R.S.S. + \lambda \sum_{j=1}^{p} |\beta_j|$$

This leads to sparse models, where certain coefficents equal zero.[7] As with the ridge regression, the R function `glmnet(x, y, alpha = 1)` and `cv.glmnet` were used. In this case however, the argument `alpha = 1` sets the α value equal to 1, which gives the constraint region a straight edge and hence can result in a β value being equal to zero.[8]

Figure 5 is similar to Figure 4 above, except that here the scale is log λ. It is notable that the reduction in the value of the coefficents in this plot is less extreme than in the ridge regression, and that one (`rm`) actually peaks at the 1 S.E. value before decreasing again. Again, the values for the minimum and 1 S.E. λ are displayed in Table 4, along with their corresponding M.S.E. values from the cross-validation selection process.

	Min λ	1 S.E. λ
λ Value	0.026	0.353
M.S.E.	24.287	26.712

Table 4: Optimum λ Values and Model 4 M.S.E. (3 d.p.).

3 Predictive Performance

In this section, the predictive performance of the three models will be compared. All three models were trained on the whole `Boston` data set and fit that data well (recall the

[6] Friedman, Hastie, and Tibshirani 2010, p. 18; Friedman, Hastie, and Tibshirani 2009, p. 244.
[7] James et al. 2013, p. 219.
[8] Friedman, Hastie, and Tibshirani 2009, p. 91; James et al. 2013, p. 222.

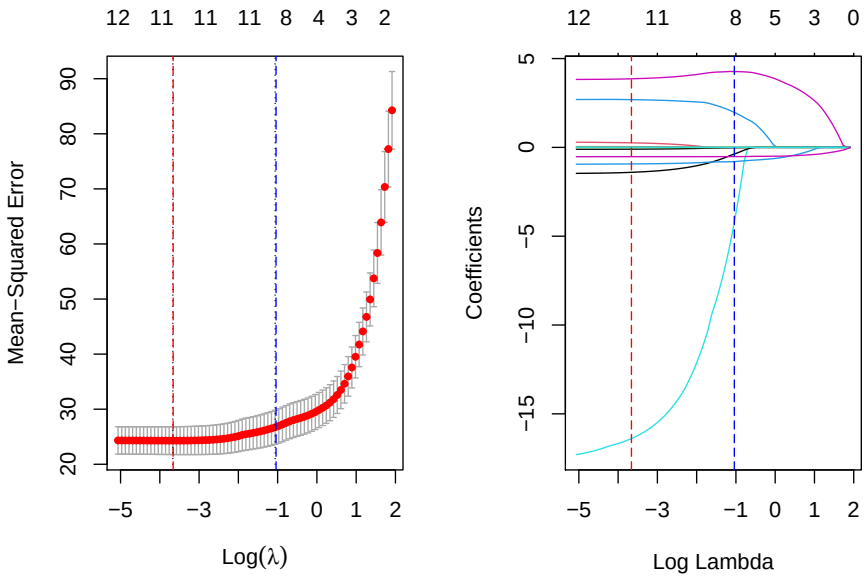

Figure 5: M.S.E. of Model 4 as log λ increases (left). Values of Model 4 coefficents and log λ (right). Minimum λ (red) and 1 S.E. λ (blue).

adjusted R^2 for Model 2 was 0.735). However, by testing the predictive performance, it can be ascertained which model would perform best using different data for the Boston area, or indeed similar data for other cities.

3.1 All Predictor Variables

The predictive performance of Models 1 and 2 can be easily computed using the `cv.glm` function in R, which automatically divides the data into k folds, refits the model on each fold and validates it against the remaining folds.[9] The average M.S.E. of each k validations is then calculated, and the whole process repeated fifty times to account for the variation in splitting the data. The mean cross-validation statistic is 24.093 (3 d.p.).

3.2 Best Subset Selection

As explored in section 2.2, the best subset algorithm chose a model with eleven predictor variables. A similar process to the one used above can be used to assess the predictive performance of this candidate model, except now using a `glm()` object with

[9] Ripley 2020a, cv.glm{boot} in R Documentation.

only the eleven predictor variables chosen. Likewise k-fold cross-validation was repeated fifty times. The mean cross-validation statistic is 23.698 (3 d.p.).

3.3 Ridge Regression

To test the predictive performance of Model 3, a validation set approach was used. The data was sampled, and half assigned to the training set, with the other half kept back for the test set. The model was then fitted on the training set and tested against the test set, using both the minimum λ and 1 S.E. λ value. This process was repeated fifty times. The M.S.E. values were 22.904 and 22.991 respectively (3 d.p.).

Model 1		24.093
Model 2		23.698
Model 3	(min. λ)	22.904
Model 3	(1 S.E. λ)	22.991
Model 4	(min. λ)	22.990
Model 4	(1 S.E. λ)	22.859

Table 5: Valdiation Statistics (3 d.p.).

As can be seen from Table 5, the ridge regression outperforms the two least squares regression models. Notably also, the minimum λ value results in a slightly lower prediction accuracy in comparison to the 1 S.E. λ value. By using the 1 S.E. λ value a lower variance would also be expected. Figure 6 plots the correlation of the response estimated using Model 3 with the the test set for the fifty valdiation tests. The minimum λ value results in correlations that are outliers, despite the mean being centred higher than that of the 1 S.E. λ. Indeed, the range of the correlations for the minimum λ model range between 0.781 – 0.886 (0.105 difference), whereas those for the 1 S.E. λ model range 0.769 – 0.872 (0.103 difference). In this particular case, despite having a smaller range in the fifty correlations, the use of the 1 S.E. λ value does not offer any significant improvements in prediction variability, and so the minimum λ value can be used with confidence.

3.4 Lasso Regression

As with the ridge regression, a validation set approach was used and repeated fifty times. The cross-validation statistics for both λ values are 22.990 (min.) and 22.859 (1 S.E.), and also given in Table 6. Again the mean prediction correlation for the minimum λ value is higher than that of the 1 S.E. λ value, but the difference between the two extremes is likewise negligible (0.077 and 0.094 respectively). As Table 5 and 6 show, even though the ridge regression outperforms the lasso in terms of M.S.E., it is not by much; while the two models have virtually identical correlation statistics.

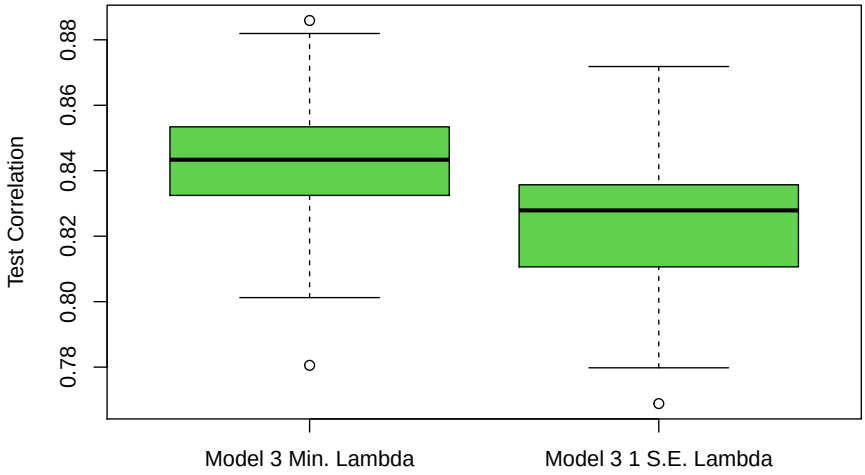

Figure 6: Box Plot of Correlation of Model 3 Predictions with Test Data.

Model 3	(min. λ)	0.842
Model 3	(1 S.E. λ)	0.825
Model 4	(min. λ)	0.845
Model 4	(1 S.E. λ)	0.823

Table 6: Predictive Performance as Correlation (3 d.p.).

4 Conclusion

In the previous sections, the performances of the four candidate models were analysed, and the models that used shrinkage methods were found to be better than the least squares regressions, with the ridge regression with a minimum λ value slightly outperforming the other permutations of the ridge and lasso regressions. Here, the coefficents of the models will briefly be examined and in turn what this modelling process reveals about the relationship between house prices and environmental factors in Boston and the surrounding area.

Table 7 details the coefficents of the four candidate models. One notes how the greatest reduction in the value of the coefficents caused by the tuning parameter in the shrinkage models is for the most influential predictors, such as `nox` and `dis`. The values of these predictors also differs markedly between the two λ values, again pointing to the aim of the shrinkage method to reduce model variability. The variables `indus` and `age` were excluded from Model 2, and it is unsuprising to see them reduced to

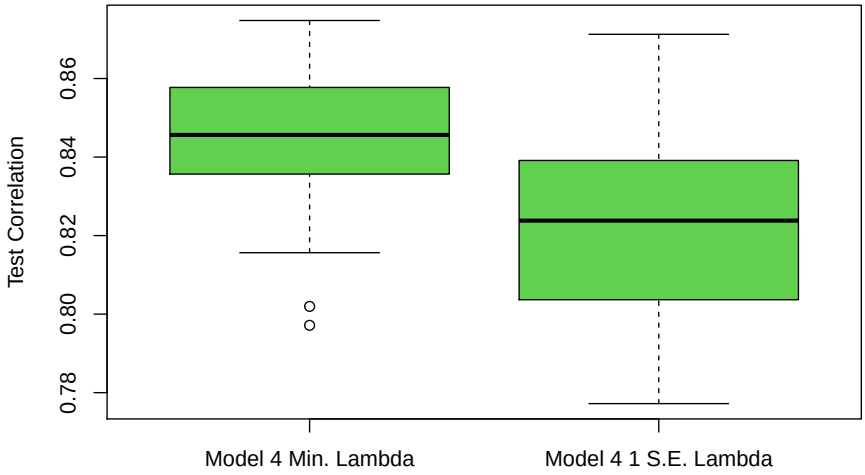

Figure 7: Box Plot of Correlation of Model 3 Predictions with Test Data.

either negliable or null values in Models 3 and 4. It is notable too that `zn`, `rad` and `dis`, like `indus`, have either very small or null (Model 4 with 1 S.E. λ) coefficents. From these variables, it might be inferred that the nature of the built environment does not particularly effect house prices. `rad` and `dis` may also point to the nature of transportation and working patterns in the Boston area, if access to highways and distance from employment centres does not exert a significant influence on house prices.

The variables `age`, `rm` and `tax` are the only measures of houses themselves, as opposed to the environment. The insignificance of `age` can be taken to imply that properties do not depreciate in value with age, while `tax` likewise indicates that the property tax has very little correlation with property value (perhaps a fact to be brought to lawmakers' attention!). Instead, the number of rooms per dwelling, again unsuprisingly, seems to be a significant predictor of house price. Although lower than `nox`, say, the coefficent remains remarkably constant across all models, and actually increases in value in Model 4 between the minimum λ and 1 S.E. λ values (cf. Figure 5). While the built environment does not seem to influence house prices, the proximity to the Charles River does: the variable `chas` has almost as strong an influence as the number of rooms in the property, and likewise remians relatively constant across all models.

Finally, there are variables which record social and economic factors, such as `crim`, `ptratio`, `black` and `lstat`. Although `lstat` has a small but significant influence, it is notable that `crim` and `black` do not, which raises further research questions into urban decline and "white flight" that plagued many American cities in the late 20$^{\text{th}}$ century.

Predictor	Model 1	Model 2	Model 3		Model 4	
			min. λ	1 S.E. λ	min. λ	1 S.E. λ
Intercept	36.460	36.341	28.001	20.709	34.595	18.567
`crim`	– 0.108	– 0.108	– 0.088	– 0.067	– 0.099	– 0.024
`zn`	0.046	0.046	0.033	0.020	0.042	
`indus`	0.021		– 0.038	– 0.069		
`chas`	2.687	2.719	2.900	2.729	2.688	1.994
`nox`	– 17.770	– 17.376	– 11.913	– 5.174	– 16.401	– 4.486
`rm`	3.810	3.802	4.011	3.552	3.861	4.272
`age`	0.001		– 0.004	– 0.008		
`dis`	-1.476	– 1.493	– 1.119	– 0.486	– 1.405	– 0.390
`rad`	0.306	0.300	0.154	0.028	0.257	
`tax`	– 0.012	– 0.012	– 0.006	– 0.003	– 0.010	
`ptratio`	– 0.953	– 0.946	– 0.855	– 0.657	– 0.931	– 0.802
`black`	0.009	0.009	0.009	0.007	0.009	0.007
`lstat`	– 0.525	– 0.522	– 0.472	– 0.339	– 0.523	– 0.519

Table 7: Coefficents of Models (3 d.p.)

Yet the data recorded by Harrison and Rubinfeld here does not seem to suggest, at first glance, that this was particularly pronounced in Boston in the late 1970's.

References

Friedman, K., Hastie, T., and Tibshirani, R. (2009). *The Elements of Statistical Learning*. 2nd ed. Springer Texts in Statistics. New York: Springer.

— (2010). "Regularization Paths for Generalized Linear Models via Coordinate Descent". In: *Journal of Statistical Software* 33.1, pp. 1–22. URL: 10.18637/jss.v033.i01.

Harrison, D. and Rubinfeld, D. (1979). "Hedonic Prices and the Demand for Clean Air". In: *Journal of Environmental Economics and Management* 5, pp. 81–102.

James, G., Witten, D., Hastie, T., and Tibshirani, R. (2013). *An Introduction to Statistical Learning: with Applications in R*. 8th ed. Springer Texts in Statistics. New York: Springer.

Ripley, B. (2020a). *Package 'boot'*. Version 1.3–25. URL: CRAN.R-project.org/packages/boot/index.html.

— (2020b). *Package 'MASS'*. Version 7.3–53. URL: CRAN.R-project.org/web/packages/MASS.