California Housing Price in 1990

Phuc Lu Meghna Chandrasekar Youngju Kwon

Introduction

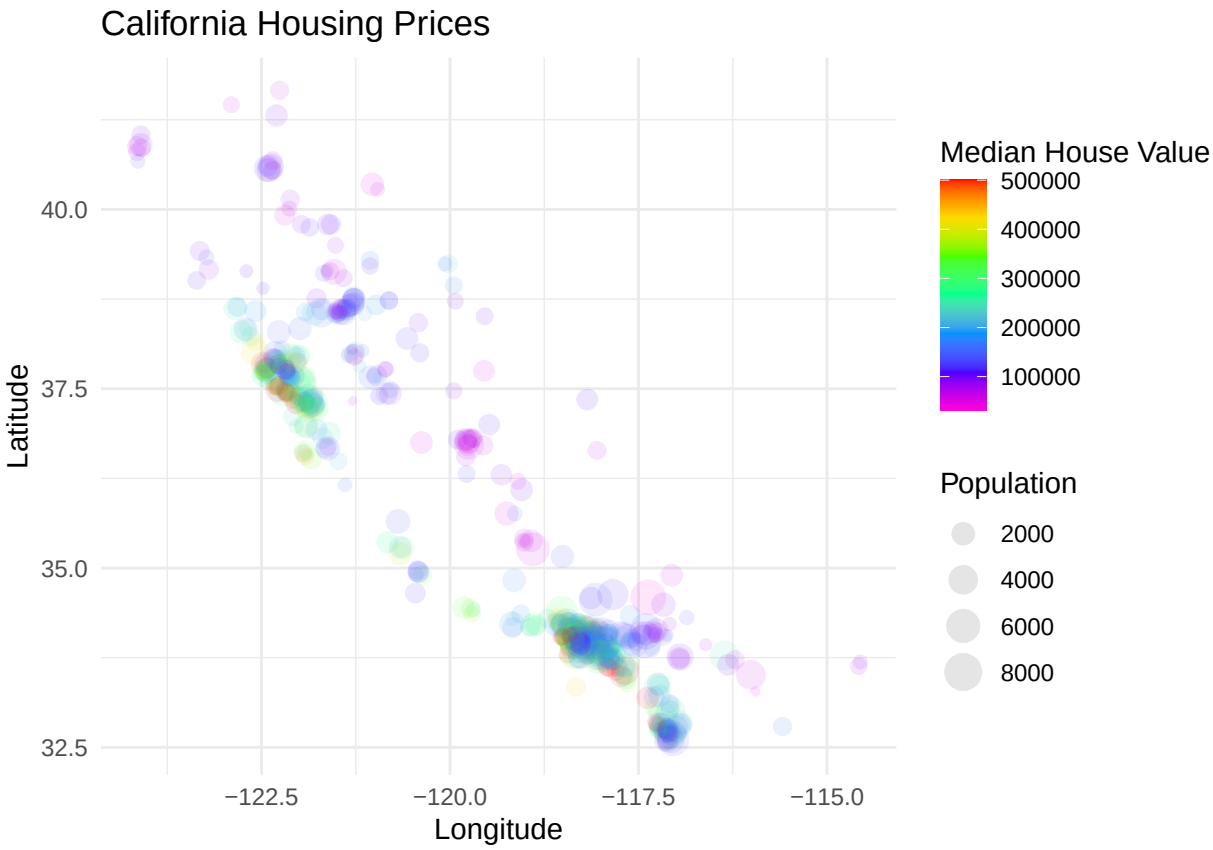
The data contains information from the 1990 California census (Source from Kaggle). This data set is a California census housing price data in the 1990s. It captures the geographical coordinates of each census tract, median house value, median house age, the proximity to the ocean, number of bedrooms, number of rooms, the population of those living in the block, the median income, and the number of households within the block.

Objectives:

- Explore the data, look for correlations and flaws.
- Test variables for significance.
- Find a good linear model with a good set of predictors.
- Develop a predictive model to predict the house values throughout California.
- Improve model via shrinkage methods.
- Incorporate one innovative techniques beyond the course to our analysis.

Data Exploration

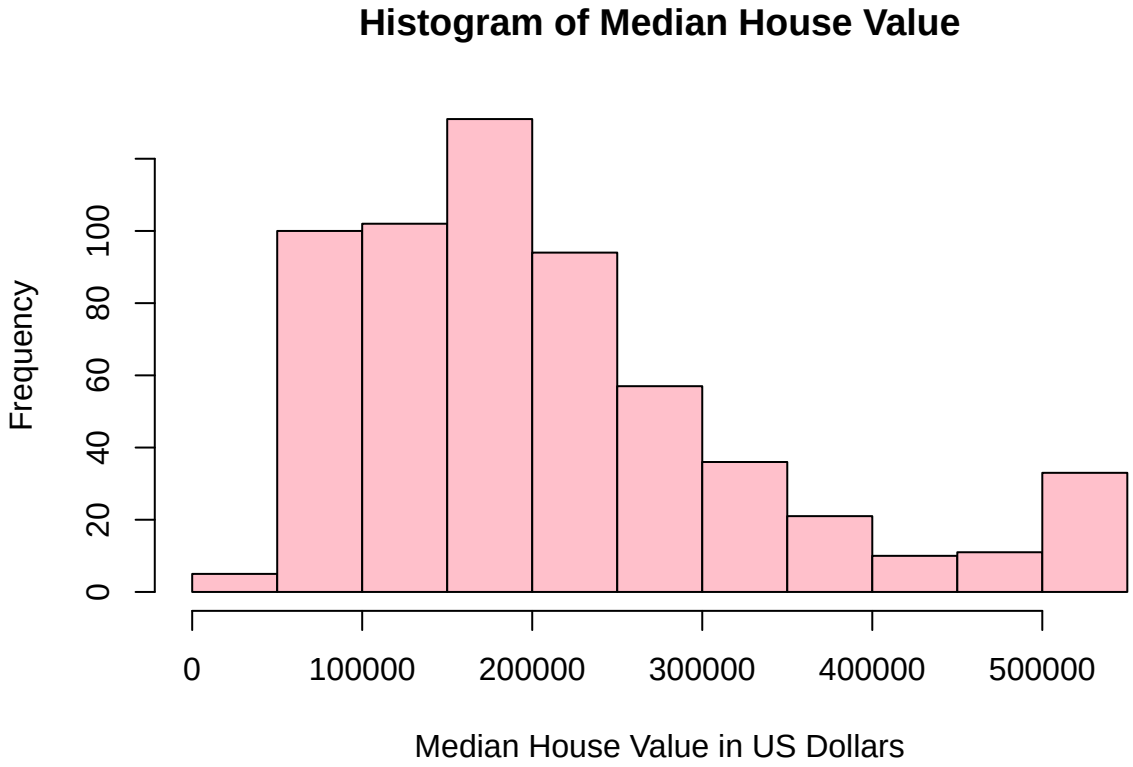
Distribution of Data Points Throughout California?



With this graph, we can see that most of the blocks are clustered around Bay Area and LA. In addition, houses that are inland are more cheap than houses along the ocean. We notice that there aren't that many data points in the middle of California (the valley region, defined by two large mountain ranges on each side). Nowadays, there are houses in this area but perhaps back in the 1990s, this area was area is mostly for farming and agriculture.

We use this map to helps us visualize in real life where our data is the most densily clustered.That is around 122 longitude Bay Area and 118 longitude LA area.

How are the Median House Values Distributed?

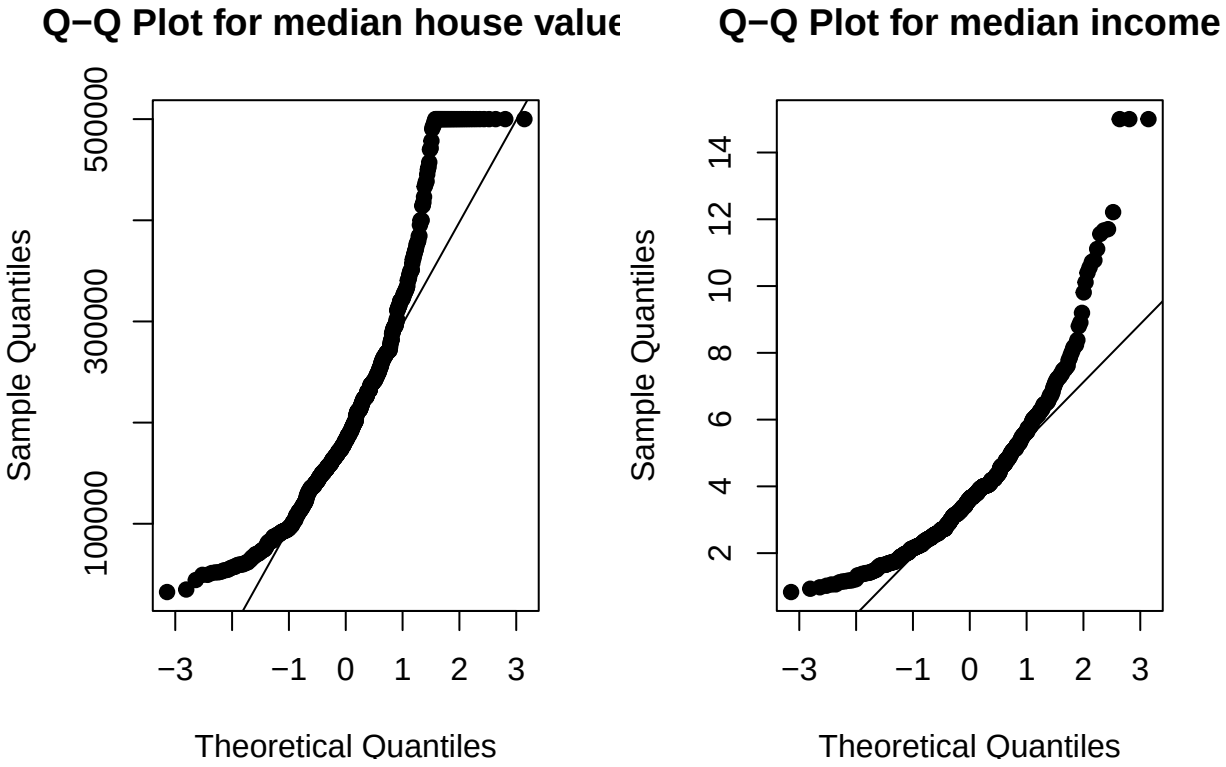


Center: Not centered at zero, but is around 0.5 standard deviation below that. Approx $150,118.9.

Spread: This data is right skewed when more observations in the right tail.

We expect a long trailing tail that drag on far right. The bump at the end is due to the data cap at $500,001.

Normality



With this Q-Q plot, we can see that the data for median house value is distributed normally in the middle but falls off towards the end. The line of points at the top is caused by the data cap that we mentioned earlier. As for the median income, the data is normal in the middle, but falls off towards the end.

We are aware that failure in this assumption should be addressed for our purpose of predicting. However, none of our transformations made sense, so we proceed without transformation.

Issue:

Since our data is capped at $500,001, the actual housing values beyond this limit are unclear. In addition, although the median is resistant to outliers, capping the observations like this does suppress the true median housing values by reporting it to be lower than it should be. These are the red flags, which could be signs of data censoring.

We believe that during the data collection process, the survey as used to collect housing prices above $500,001 was stated as "above $500,001".

Although there is a flaw in our data, we will not remove data because doing so that would considered data tampering, so we'll proceed with the analysis. As a result, we **caution** the reader when using our inferential and prediction results.

Finding Our Model

Goal: Predictive model to predict California housing prices in the 1990s.

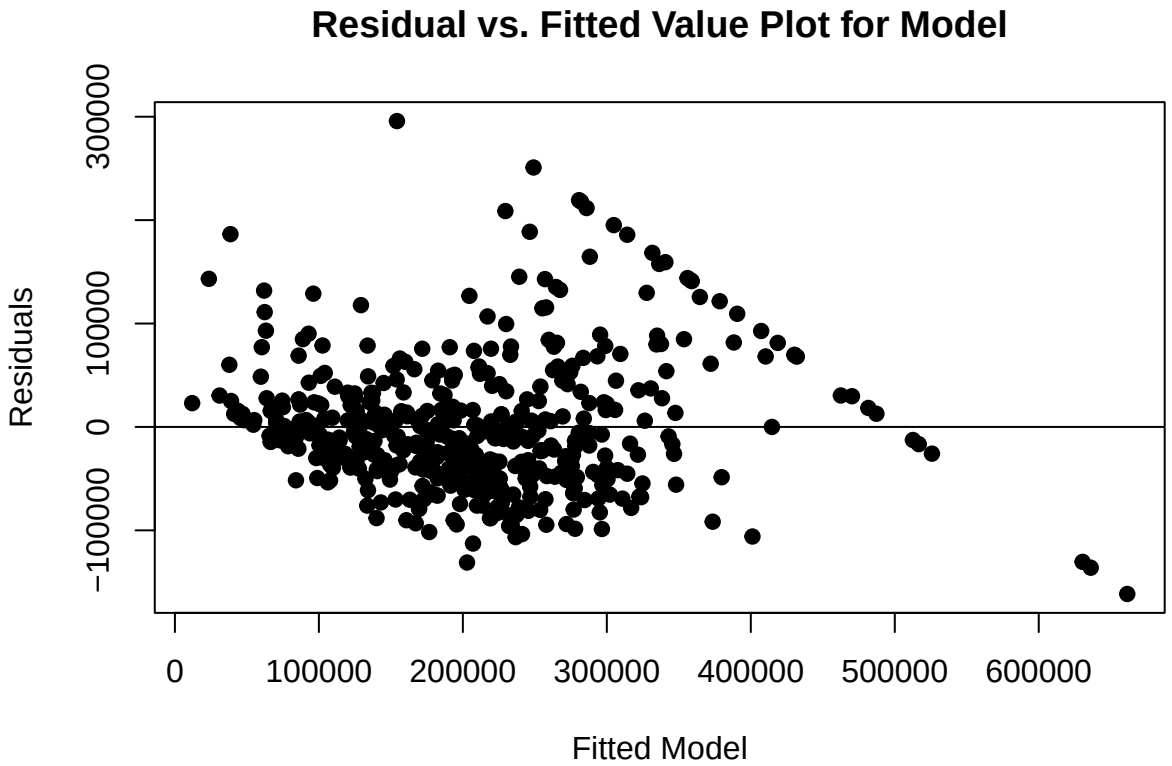
We'll be using **stepwise variable selection** algorithm to select the predictors for our model and **prioritize predictive accuracy** by minimizing the AIC (Akaike Information Criterion) at at 12008.5057.

$$
\begin{aligned}
\text{medianHouseValue}_i = \beta_0 &+ \beta_1 I\{\text{oceanProximity} = \text{<1 HourOcean}\}_i \\
&+ \beta_2 I\{\text{oceanProximity} = \text{nearOcean}\}_i \\
&+ \beta_3 I\{\text{oceanProximity} = \text{island}\}_i \\
&+ \beta_4 I\{\text{oceanProximity} = \text{nearBay}\}_i \\
&+ \beta_5 \text{medianIncome}_i \\
&+ \beta_6 I\text{longitude}_i \\
&+ \beta_7 I\text{latitude}_i \\
&+ \beta_8 I\{\text{housingMedianAge} = \text{NEW}\}_i \\
&+ \beta_9 I\{\text{housingMedianAge} = \text{MODERATE}\}_i \\
&+ \beta_{10} I\text{population}_i \\
&+ \beta_{11} I\text{households}_i
\end{aligned}
$$

With our model, we believe that it can predict the median house value of a particular block in California by using the following predictors:

- How far the block is from the ocean.
- The median income of the people living in the block.
- The geographical coordinates (longitude and latitude).
- The age of the house.
- The number of households in the block.

Residual vs. Fitted Value Plot for Model



Generally seems like the data points are spread with no discernible pattern. However, the data points towards the right form a diagonal line across the graph. This caused by the data cap in median house value data at $500,001.

The Goodness of Fit for this Model:

Multiple R-squared: $R^2 = 0.6965$

Adjusted R-squared: (adjusted)$R^2 = 0.6893$

This is also roughly the same fit as our full model, so algorithm just removed some insignificant the predictors.

Signicance of Each Predictor

```
Coefficients:
                    Estimate    Std. Error t value            Pr(>|t|)
β̂_0 (Intercept)   -1853065.5185  509561.2606  -3.637            0.000307 ***
β̂_1 lessHour         35758.1710   11356.6029   3.149            0.001745 **
β̂_2 nearOcean        35357.7477   14299.2159   2.473            0.013764 *
β̂_3 island          223066.6114   66474.9845   3.356            0.000856 ***
β̂_4 nearBay          25248.2196   14904.0379   1.694            0.090921 .
β̂_5 median_income        3.7815       0.1517  24.922 < 0.0000000000000002 ***
β̂_6 longitude      -21717.5323    5942.4925  -3.655            0.000287 ***
β̂_7 latitude       -19444.5835    5805.0155  -3.350            0.000875 ***
β̂_8 new_house      -49445.8192    9536.2489  -5.185   0.0000003220097517 ***
β̂_9 moderate_house -29781.3877    8079.2219  -3.686            0.000254 ***
β̂_10 population        -47.1829       6.9407  -6.798   0.0000000000324261 ***
β̂_11 households        157.5535      19.7807   7.965   0.0000000000000127 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The coefficients are estimates of what the slopes of our predictors should be. For one additional unit of change in the predictors, the median house value also changes by the estimate values.

From this output, we see that the median income of a particular block is a **very significant predictor** of the median housing price of a given block. Meanwhile, the P-value for houses with proximity near the bay is not a very significant predictor of housing price. However, this is only when you observe at the bay specifically. If you're curious about proximity near the ocean, then it is a significant predictor for housing values.

We believe that this model is slightly overfit, but it is the best that we can do.

Which Predictors Did The Model Dropped?

Predictor:	Estimate	Standard Error	t-value	p-value
total_rooms	0.3586	4.8920	0.073	0.941596
total_bedrooms	13.5249	38.1067	0.355	0.722810

Since the variable selection algorithm works with comparing p-values, it decided to drop total rooms and total bedrooms.

These two predictors are actually not significant when it comes to the median house value. One might assume that there is a positive correlation between house price and number of rooms or bedrooms in the house. The assumption is that houses with more rooms tend to me bigger and thus more expensive.

But our variable selection algorithm tells us that these two predictors are not significant and the data shows that there almost no correlation.

Correlation of house value & total rooms : 0.1291

Correlation of house value & total bedrooms: 0.0219

Thus, we conclude that wealthier people tend to own more expensive houses, but the price of the house is no associated with and how many rooms or bedrooms the house has.

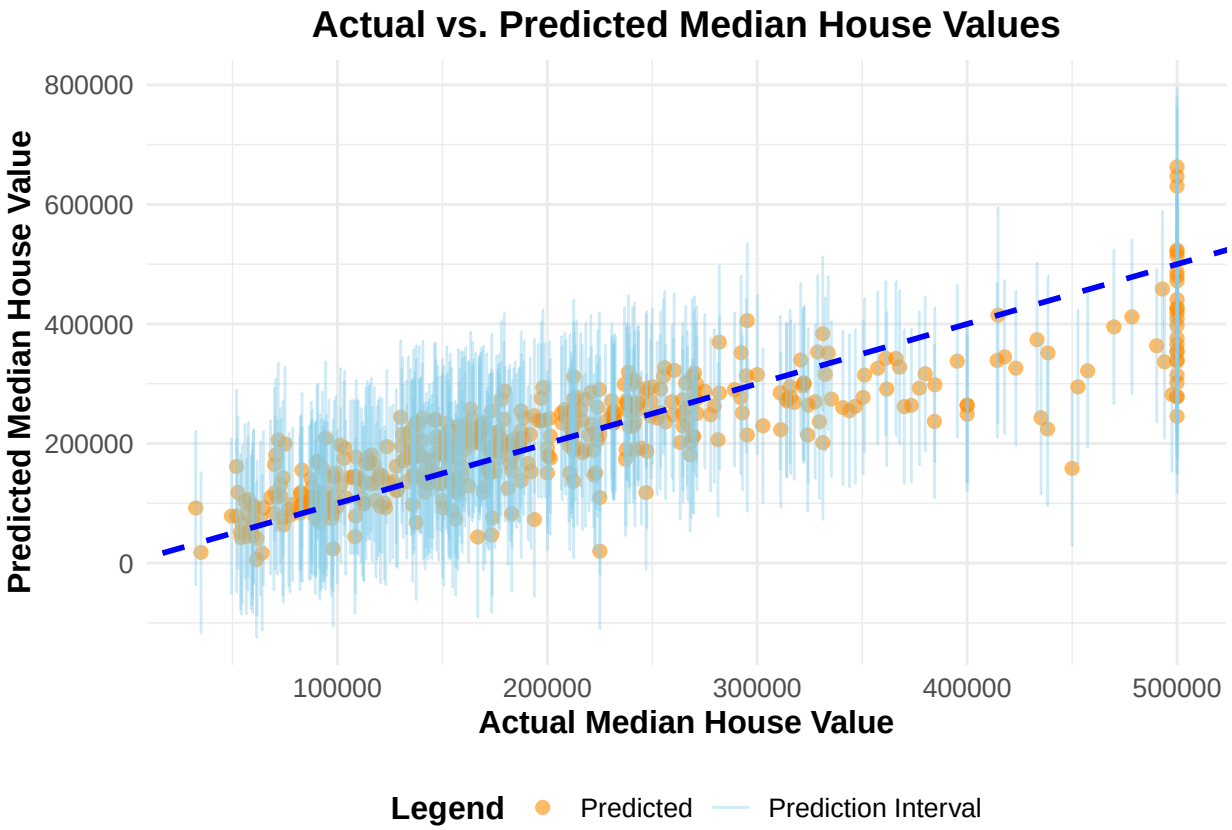
Mean House Price and Price Predicion

Confidence for Mean Value

With 95% confidence, the mean median house values of blocks throughout California in 1990s with attributes similar to the average block value is estimated to be between $118,694.0 and $185,402.7.

So if you were to buy a house in California in the 1990s, this interval captures the true average median values of houses 95% of the time.

Prediction Interval



We see that the model is pretty good at prediction values up until $300,000 and then the quality falls off after that.

We test predictive power of our model, we sampled 2 random blocks from test data.

First block's attributes:

- 37.85 latitude, -122.27 longitude, near Bay.
- The median age of houses in this block are considered old.
- This block contains 843 people and 319 households.
- The median income in this block is $13,785.

With 95% confidence, the median house value a block with these attributes is estimated to be between $27,008.13 and $28,4750.9.

Our second block has these attributes:

- 38.55 latitude, -121.47 longitude, Inland
- The median age of houses in this block are considered moderately new.
- Block population: 291 people with 546 households
- Block's median income is $11,860 a year,

With 95% confidence, the median house value in a block with these attributes is estimated to be between $-57,362.67 and $198,418.4. The predicted value is $70527.89 and actual value is $67000.

This means that if you seeking in buying a California home within this particular block in the 1990s, this interval predicts the block's median house value 95% of the time.

Important Caution: The range of the prediction interval stretched from negative to positive value. This means that any value between the interval bounds are equally as likely, so there is a possibility of getting a free house. Or that block owe houses.

As we know, this scenario is almost impossible in real life, so our prediction interval is as good as not having one at all.

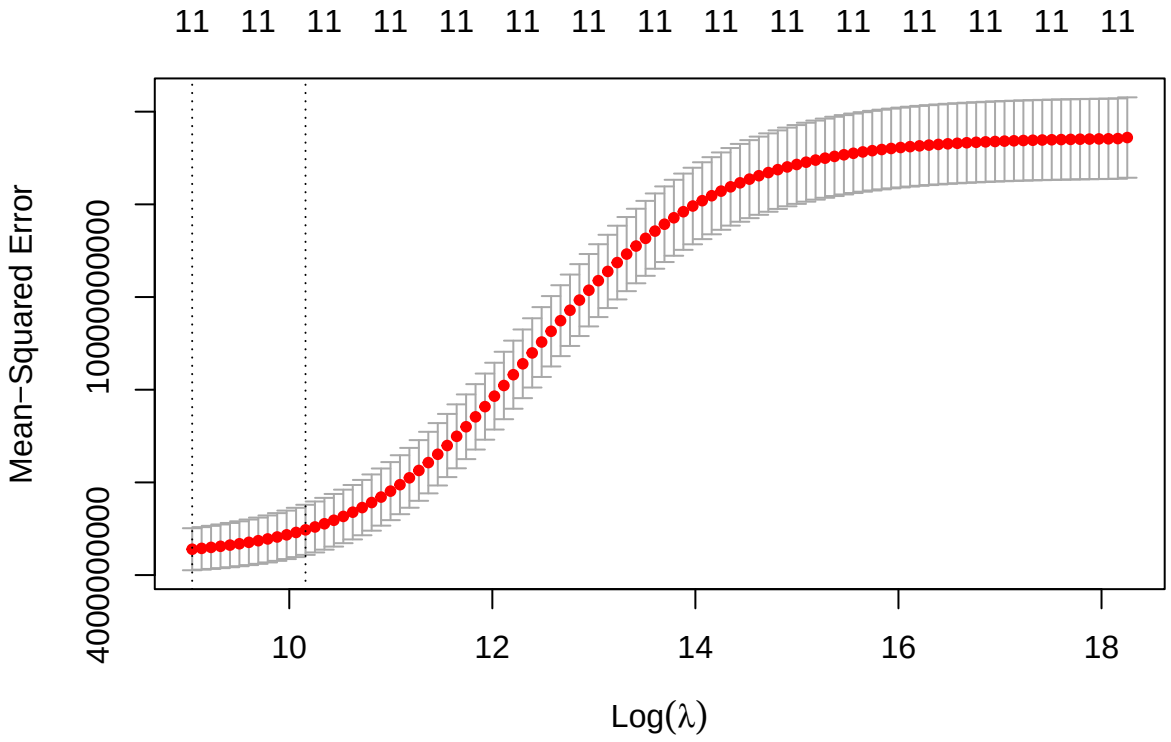
This is a case where our model would fail to give an accurate prediction for the median price of houses in a particular block.

Shrinkage

Ridge Regression (RR) and LASSO We will perform Ridge Regression and LASSO using cross-validation to find the optimal lambda (regularization parameter). We will compare these models to a previously developed Multiple Linear Regression (MLR) model.The results of this analysis are inferred to the housing market in California, representing various geographical and socio-economic conditions within the state.

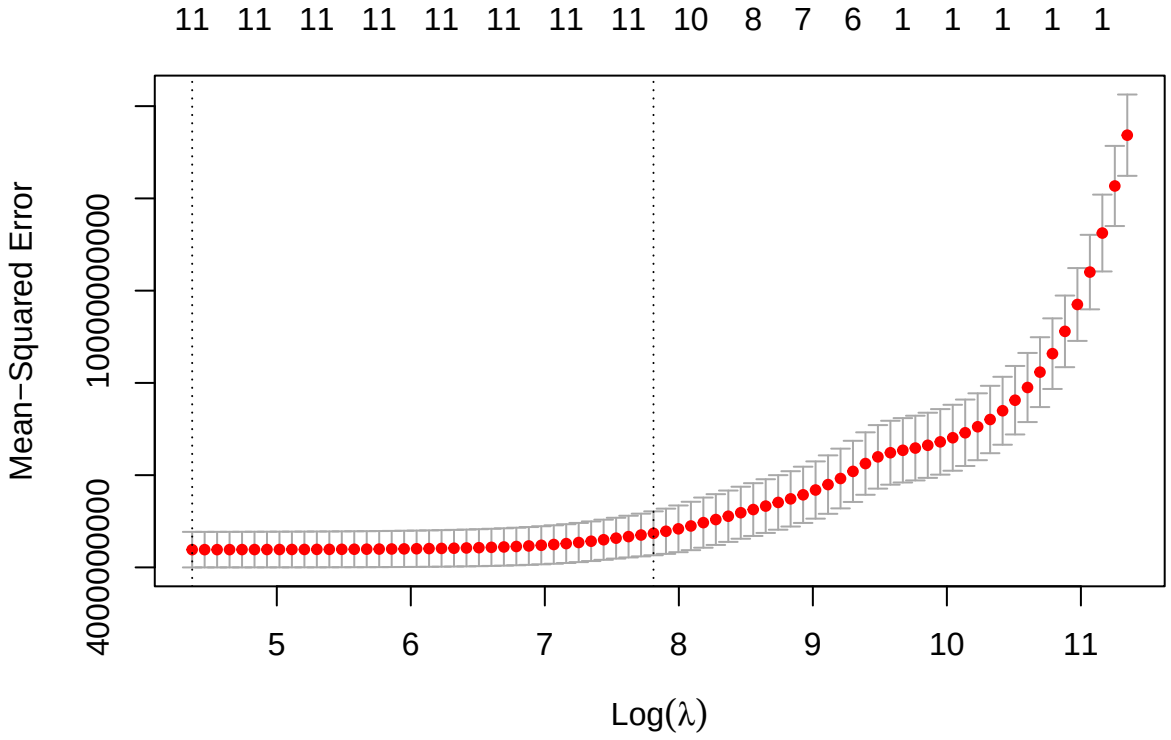
Ridge Regression

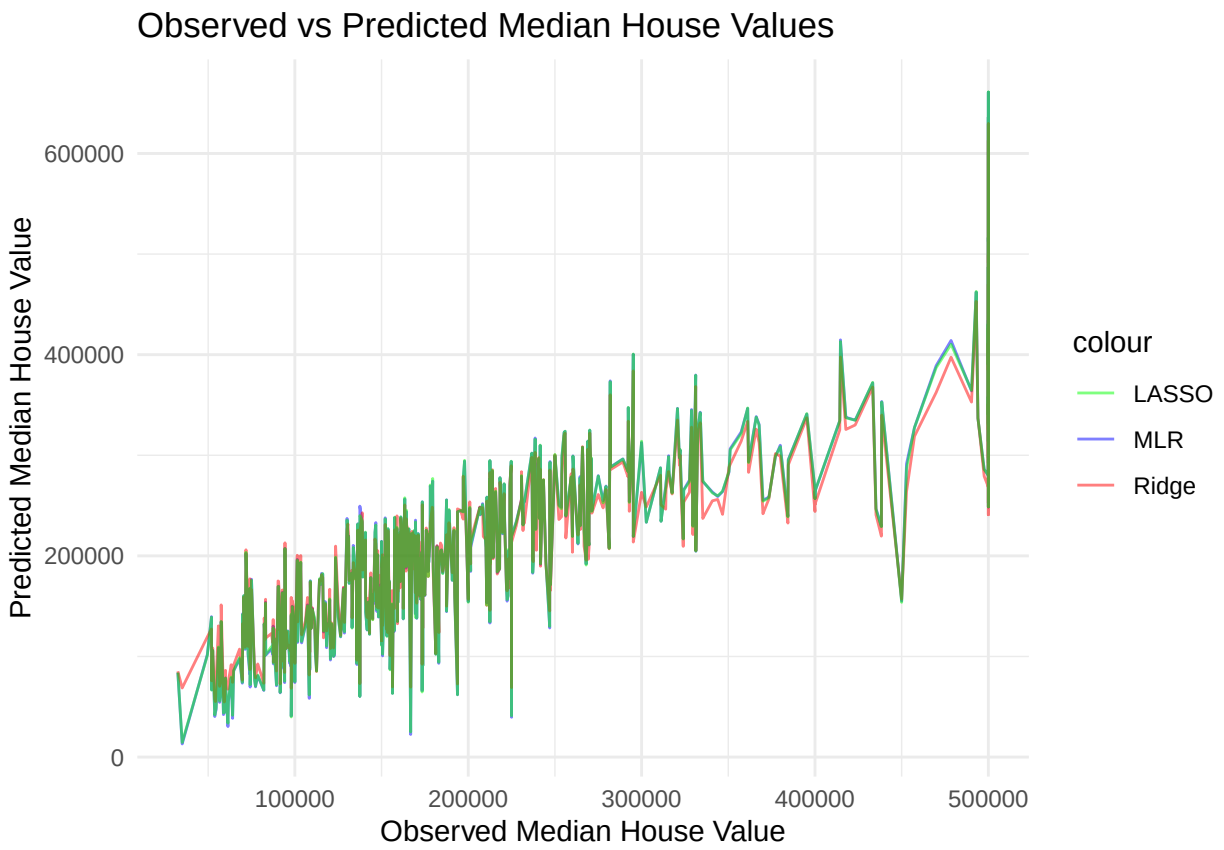
Ridge regression adds a penalty equal to the sum of the squared coefficients to the loss function. This helps in dealing with multicollinearity (when predictor variables are highly correlated) by shrinking the coefficients of correlated predictors towards zero but not exactly zero, thereby stabilizing the estimates.



Lasso Regression

LASSO (Least Absolute Shrinkage and Selection Operator) Regression adds a penalty similar to Ridge Regression and equal to the sum of the absolute values of the coefficients to the loss function. This can drive some coefficients exactly to zero, effectively performing variable selection. This is particularly useful when dealing with high-dimensional data where many predictors may be irrelevant.





Through analysis, it seems that the LASSO test yielded a greater variance than that of the Ridge Test. The coefficients from Ridge and LASSO regression are likely to differ due to the regularization applied. Although Ridge tends to shrink coefficients towards zero but not exactly zero, LASSO can shrink some coefficients to exactly zero, effectively performing variable selection.

Predictive Performance: By examining the plot, we can visually assess which model provides the best predictions. The model whose points lie closest to the 45-degree line (where observed equals predicted) performs best. In this examination, it seemed the values produced by the LASSO Test produced the most values.

Conclusion

Overall, both tests seemed to perform more efficiently than that of the MLR because of the standardization. A surprising thing is noticing how every predictor seemed to indicate a near negative association when it comes to estimating the median house value, which was usually believed to be the opposite. Any further queries that could result from performing seemed to be common.

Innovation

When we use our best model with 11 predictors, we see heteroscedasticity in the residual distribution. This can lead to several issues in a linear model, including: Inefficient OLS estimators that do not have the smallest possible variance, incorrect standard errors of the estimated coefficients, and invalid inferences using test statistics.

Weighted Least Squares is appropriate to use when homoscedasticity in OLS linear regression is violated.

The Breusch-Pagan Test for heteroscedacisity tests for it statistically. The p-value for the Breusch-Pagan test for heteroskedacity is very low, at 0.00002107. We can reject the null hypothesis that the variance of the residuals is constant.

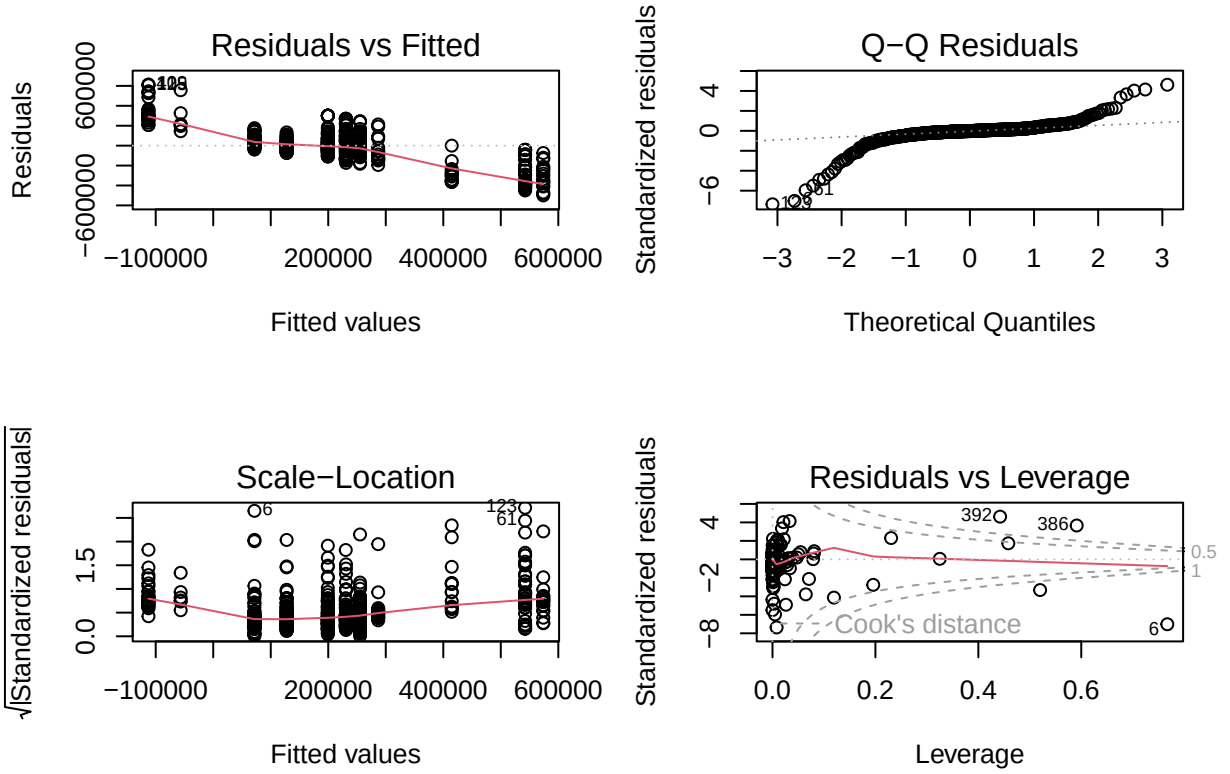
In weighted least squares, the weight is defined by the reciprocal of each variance, $\hat{\sigma_i}^2$

$$w_i = \frac{1}{\hat{\sigma_i}^2}$$

An observation with a large error variance will have a smaller weight, and vice versa.

When we perform weighted least squares by weighted the residuals, the R^2 increases from 0.696451 to 0.9726604, indicating that our model's explanatory power of the variance improved.

However, the AIC (Aikake Information criterion) in the weighted least squared model by residuals is 13584.62 vs 12008.51 in the OLS model, and the BIC (Bayesian Information Criterion) is 13613.84 in the WLS versus 12062.76 in the OLS model, were lower values indicate a better model.



The residuals vs fitted plot in the WLS model shows that the data is fitted into clusters. A binned plot groups the data into bins and plots the average residuals within each bin. This may show a more representative view of the nature of the residuals variance.

In the weighted least squared model by residuals, all of the binned residuals fall outside the 95% CI bounds, indicating that the residuals are not evenly distributed.

Overall, the cross validation failed in WLS method improving our linear model, and further methods should be implemented to fix the heteroskedasticity in variances.

Final Note

Overall, this data set is indeed terrible for predictive purposes because it fails a bunch of model assumptions and we made some poor choices such as not transforming some of the predictors. Though, we believe that our analysis could be slightly better if we use the full data set instead of using a subset of 600 observations. This is all of our first ever data analysis project, so we wanted to keep everything as it is and be fine with making mistakes. There are a lot of aspects that we would've done better on. However, with the time restraint, this is the final product of our hard work throughout the quarter.