Project Step 4: Summary

Keon Dibley & Lucy Zhao

2023-12-15

Introduction to Dataset: Dataset: California Housing Prices Data Source: Kaggle

For our project, we decided to examine a California Housing Data Set. We wanted to look at different factors, like Distance to Coast, Median Income and Proximity to Big Cities and see how these served as predictors of Median House Value.

One observational unit in our data set is a block of California Houses, and the following are a few of the interesting variables from our data set: - Median House Value: Median house value for households within a block (measured in US Dollars) [$]

- Median Income: Median income for households within a block of houses (measured in tens of thousands of US Dollars) [10k$]
- Distance to coast: Distance to the nearest coast point [m]
- Distance to San Jose: Distance to the centre of San Jose [m]
- Distance to San Francisco: Distance to the centre of San Francisco [m]

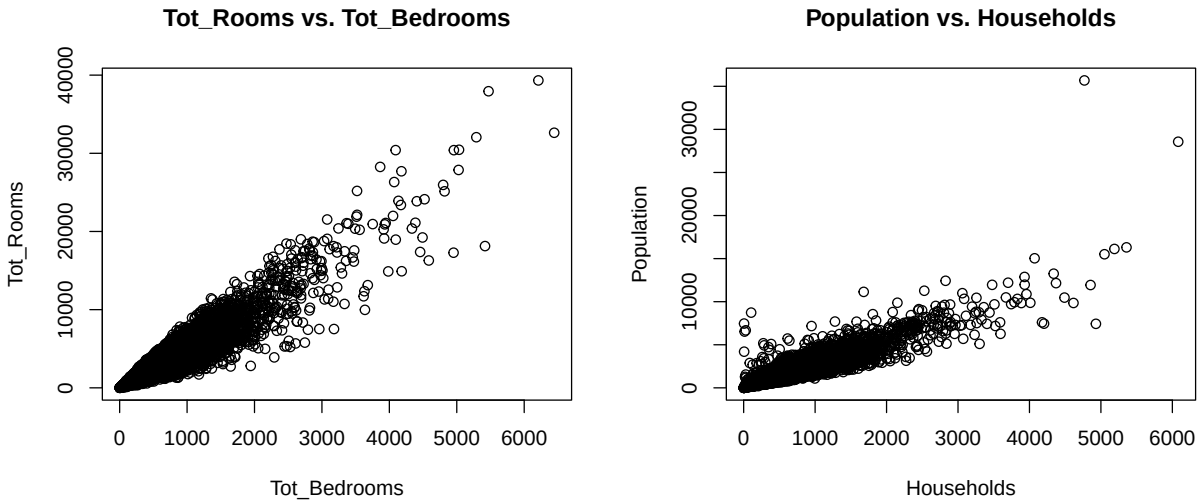
Initial Data Description and Investigation

To start our investigation, we looked at the distributions of some of the variables, and came to the conclusion that we could create some new variables that would provide an interesting perspective on our data.

We created 3 new variables, north_south_ca (to see the difference between houses in NorCal and SoCal), near_top4_big_city (a categorical variable with 4 levels to indicate how close a block of houses is to a big city), and min_distance (to return a value for the distance between a block of houses and its closest big city). We felt that these variables, especially min_distance, would be helpful in the four variables that returned the distances to the top four biggest cities in California.

Next, we looked at how near_top4_big_city impacted Median_House_Value, Median_Income and Median_Age, and found that houses that are very close to city centers tend to be newer (younger in age).

In this initial examination, we found a couple more interesting relationships. First, a roughly linear relationship between Median_Income and Median_House_Value. Also, we found similar patterns in Total Rooms, Total Bedrooms, Population, and Households, which made us think they should be considered together. This is shown below:



Simple Linear Regression

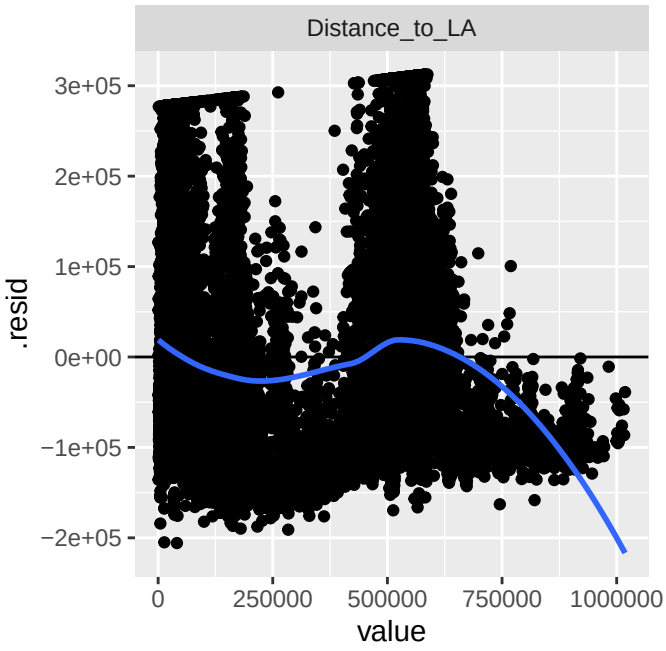
After this initial investigation, we wanted to make simple linear models for each of our predictors and our response, and examine these models.

We examined each linear model at a significance value of alpha = 0.05, and found that Median Income, Distance to Coast, Median Age, and Longitude were significant predictors of Median House Value. Additionally, we noticed that Longitude and Median Age had low R^2 values, which indicates that they are less significant than Median Income and Distance to Coast.

We knew that although the other predictors weren't significant on their own, we could make transformations so that they explained our response variable better.

First, we looked at the residuals vs predictors plots, and saw that we could use a third degree polynomial to explain Distance to LA. However, we found that as we increased the degree of our polynomial, R^2 also increased, and we were able to find that the best representation of this variable was through a ninth degree polynomial.

When examining this graph of residuals vs Distance to LA, this makes sense because the very high peaks in this graph suggest a high degree polynomial being a good fit to explain it. This plot is shown below:



Also, we examined confidence and prediction intervals for some of our predictors, and found that high Median Income and low Distance to Coast were both associated with high Median House Values.

Multiple Linear Regression

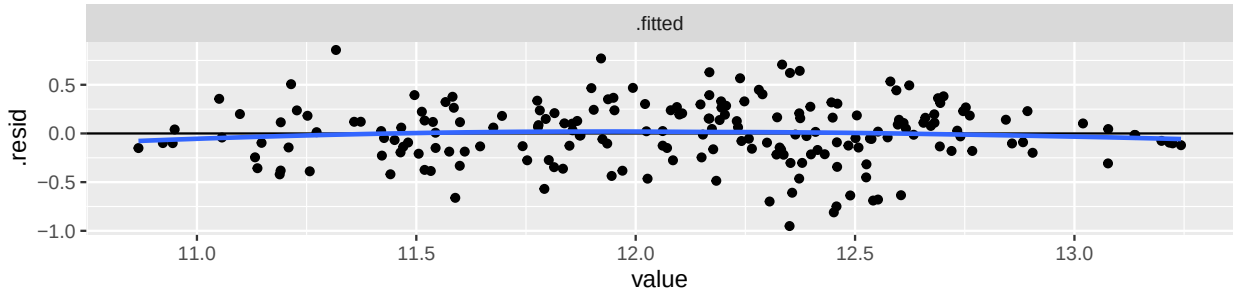
Next, we wanted to find one model that included multiple predictors and could best explain Median House Value.

After examining a pairs plot, and looking at the significance of our simple linear regression models, we decided to fit two different models, fit 1 and fit 2. After examining the R^2 and σ values, we decided to that fit 2 better explained the data.

fit2 :$\ln(\hat{\text{MedianHouseValue}}) = -14.894 + 4.663 * \text{MedianIncome} - 1.491 * \text{MedianIncome}^2 - 0.125 * \ln(\text{MinDistance}) - 0.162 * \ln(\text{Distancetocoast}) + 0.004 * \text{MedianAge}$

This model includes Median Income, Min Distance, Distance to Coast and Median Age as predictors of Median House Value. We decided to take the log of Min Distance, Distance to Coast and Median House Value because these observed values are often very large, which justifies a log.

This model has a high adjusted R^2 value of 0.748 with only five predictors, which indicates its appropriateness. However, we wanted to perform diagnostic checks to further confirm our model's accuracy. The fitted vs residual plot for our model is shown below.



This residual vs fitted value graph has points that are scattered randomly around the residual = 0 line, so all these estimates appear to be unbiased.

Also, when examining the residual vs predictor graphs, we found overall constant variance. These two findings indicate that our model was fitted well.

To conclude this step in our investigation, we found our final model to be:

fit2: $\ln(\text{Median}\hat{\text{H}}\text{ouseValue}) = -14.894 + 4.663 * \text{MedianIncome} - 1.491 * \text{MedianIncome}^2 - 0.125 * \ln(\text{MinDistance}) - 0.162 * \ln(\text{Distancetocoast}) + 0.004 * \text{MedianAge}$
with **all coefficients significant**
numeber of predictors: 5
Adjusted R-squared: 0.748
$\hat{\sigma}$: 0.312
Global F test p-value smaller : <2e-16.

Beyond the Linear Model: Shrinkage and Principal Component Regression

After finding our final MLR model, fit2, we wanted to try other methods to see if we could find a better model for Median House Value.

Through our investigation of Lasso Regression, we found that it didn't help us create an accurate model. We think that it is because of the high lambda value that we found through cross validation, which caused our model to sacrifice a lot of bias for a smaller variance.

When we tried Ridge Regression, we hoped that it would shrink our coefficients and decrease the variance of our model, but cross validation actually found that our best lambda value was actually 0, meaning that Ridge Regression actually returned the OLS model. Nevertheless, this was still a good model, but not quite as accurate to the observed values as our MLR model, fit2.

Next, we wanted to innovate further using Principal Component Regression to shrink our variable size, and this proved to be more successful than Ridge and Lasso Regression. This method helped us point out

colinearities in our predictors, and combine our predictors so that we could have fewer variables.

In the end, we found a model with 6 predictors and a similar adjusted R^2 value to fit2. However, this model is more difficult to interpret than fit2, so it is our opinion that fit2 is the best model for explaining Median House Value in this data set.

Inferences/Final Thoughts

Overall, we found many interesting associations and ways to explain California House Values, some more surprising than others.

It was less surprising to us that Median Income and min distance had large impacts on Median House Value, because this felt rather intuitive. However, we didn't expect that Median Age and Distance to Coast would be significant variables, so this was interesting for us to discover. Completing this analysis was very interesting to me, as a California resident, because I have never thought so deeply about housing trends before, especially in relation to distance to coast.

Although not all of the methods we tried proved fruitful in generating the best model, they all helped us get a greater understanding of the data set and allowed us to move forward with greater insight in future project steps.

Thanks for a great quarter! :)