(F23) PSTAT 126: Project Summary

Anthony Cu and William Mahnke

13 December 2023

Introduction

Our project investigated the California Housing data set curated from Kaggle. The data was compiled from a 1990 California survey, providing information on a specified district in the state. The data describes homes in a district of California in 1990 in order to represent the larger population of this district during the 1990's. Some limitations of this data include the lack of data about individual houses and the individuals in each house. Each independent response corresponds to a different block within the district.

Throughout our research, we investigated the following quantitative variables: `housing_median_age` (the median age of a home within a block), `total_rooms` (the number of rooms in a home within a block), `total_bedrooms` (the number of bedrooms in a home within a block), `population` (the total number of people living within a block), `households` (the total number of groups living within a home in a block), `median_income` (the median income measured in $10,000 for households within a block), and `median_house_value` (the median house value (measured in $) for households within a block). Furthermore, we consider the qualitative variable: `ocean_proximity` (the location of the house in regards to its distance from the ocean, categorized by less than an 1 hour, inland, near bay, near ocean, or island).

Questions of Interest

Throughout our research, we used the variable of median house value as our *response*.

Some questions of interest were:

- Is there a linear relationship between any predictors and the median house value?
- What variables are significant in predicting the median house value?
- How can we use different regression methods to formulate different linear regression models to explain median house value?

Single Linear Model

We began with a Simple Linear Model to test whether median house value is linearly related with the predictor median income (measured in $10,000).

Methodology

To test the relationship between median income and median house value, we began with a partial significance test on the predictor for median income on a simple regression with only median income as a predictor. If the null hypothesis was satisfied (i.e. median income isn't a significant predictor), we would explore other potential simple regression models. If the null hypothesis was rejected, we would calculate other model statistics and confidence intervals to help explain median house value using just median income.

Procedures

After seeing that our null hypothesis was rejected, we proceeded with summarizing the essential parts of the linear model. These summary statistics for the model include the estimate coefficient for the predictor and the adjusted R squared value for the model. We proceeded by visualizing the linear relationship between median house value and median income. After visualizations of the variables, we used the model to calculate and visualize a confidence interval for the predictor's coefficient, a confidence interval for the mean median house value, and a prediction interval for a the median house value for a specific response.

Results & Analysis

Our results suggest there is a positive linear relationship between the median income and the median house value. We found that with 95% confidence, a \$10000 increase in median income will increase the average median house value between \$40390 and \$60480. This further reinforced the fact that median income is a significant predictor of median house value. Additionally, we observed that the $R^2 = 0.284$, with the adjusted $R^2 = 0.282$. This value indicates that 28.20% of variance in the median house value is explained by the model, so 71.80% of the variance in the median house value is explained by the noise.

Looking at our calculations for our confidence and prediction intervals, we came to these conclusions that with 95% confidence, the mean median house value when the median income is \$12904, is \$56062.57 and \$96204.16 and with 95% confidence, the predicted median house value when the median income is \$12904, is between -\$76892.40 and \$229159.10.

Limitations

While this evidence shows how median income is a significant predictor for median house value, the model doesn't account for other predictors of the response (by design). This lead us to explore other linear regression models that rely on more than one significant predictor.

Multiple Linear Model

Moving forward, we investigated the Multiple Linear Regression model of the response median house value with predictors: median house age, median income, population, households, number of bedrooms, number of bathrooms, and ocean proximity.

Methodology

To begin, we visualized the correlation between predictors in the data. We observed there was a high correlation between some pairs of explanatory variables, including total bedrooms and total rooms, population and total rooms, population and total bedrooms. Additionally we saw that population and households has a positive linear relationship. When finding an optimal model using the training data, we considered interaction terms for these pairs of highly correlated variables.

We split our data into two groups: training and test data. 70% of the data will be used for finding and training a model while the remaining 30% will be used to perform significance tests, analyze unusual responses, and calculating other important aspects of the model such as R^2_{adj}, R^2, and confidence and prediction intervals.

For our multiple linear model, we created two models, $\hat{Y}_1$ designed with the intention of satisfying the model assumptions and $\hat{Y}_2$ designed using backwards selection (which also satisfied the model assumptions). Using criterion based methods such as AIC, BIC, and R^2_{adj}, we determined which model was the best model based on predictive accuracy, consistency with the 'true model', and model fit.

We then analyzed the significance of each predictor in the better model and used model statistics to explain what variables were the best predictors for median house value. Further, we analyzed the unusual responses in our data and how the points with high influence affected the fit of the model. Using this same model, we then calculated a confidence interval for the mean median house price and a prediction interval for a particular neighborhood.

Procedures

For creating our first model $\hat{Y}_1$, we began by fitting a naive model, which was entering all variables of interest into our model linearly.

Looking at the diagnostic plots for our naive model, we observed that the linear assumption wasn't met for some predictors. Additionally, the residual vs fitted plot showed the constant variance assumption is also not met. We noticed that households and median income resembled a quadratic shape in its residual vs predictor plot, so we transformed our model by entering households and median income as a quadratic and a cubic function respectively.

Adjusting the model with these new quadratic terms, we saw that the linear, constant variance, and normality assumptions were met. Additionally, by the nature of the data, we concluded that the independence assumption was met. So our model $\hat{Y}_1$ satisfied all assumptions.

Using our correlation plot, we thought it was important to include interaction terms when creating $\hat{Y}_2$. So, we used backward selection to fit our second linear model that would take into account for certain interaction terms between predictors. Some of the interaction terms are reflected in the variables with high correlation from the pairs plot that we had constructed, as mentioned in our *Methodology* section.

We defined our full model to consider all of the predictors of interest as well as all of the interactions between them. Applying backwards selection left us with only the significant predictors and interaction terms. Checking the model assumptions for this new model, we saw that linearity, constant variance, and normality were satisfied. Independence followed too since the nature of the data didn't changed, so $\hat{Y}_2$ satisfied all assumptions.

We compared AIC, BIC, and R^2_{adj} of $\hat{Y}_1$ and $\hat{Y}_2$ to select a single 'best' model to use on the testing data we set aside earlier. In all three comparisons, we deduced that $\hat{Y}_2$ minimizes AIC and BIC as well as maximizes R^2_{adj}, leading us to conclude that it is the best model. Furthermore, this model would reflect the high correlation we saw between some of the explanatory variables in our pairs plot.

Results

Looking at the summary of our model using the testing data and assuming a p-value of 0.1, the statistically significant predictors in our model were median income, households, the indicator variable when ocean proximity is near the ocean, total bedrooms interacting with households, total bedrooms interacting with median income, population interacting with median_income, median income when the ocean proximity is inland, and median income when the ocean proximity is near ocean. The significance of certain predictors over others indicated that median income, households, and ocean proximity are the most important factors when estimating the median house price of the sample.

On the test data, the $R^2 = 0.852$ and the $R^2_{adj} = 0.807$. This means about 80% of the variance was explained by our model. The high R^2 is not a guarantee that the model will accurately describe the population because the calculated R^2 value only accounts for values within the sample and doesn't include the entire population.

Our model showed that fixing all other variables, a \$10000 increase in median income increases the average median house value by about \$52000. Looking at the interaction terms involving median income and ocean proximity, we saw that the association between mean median house value and median income decreases by about \$7118 per \$10000 of median income when the neighborhood is considered inland. Additionally, the

association between mean median house value and median income increases by about $47670 per $10000 of median income when the neighborhood is considered near the ocean.

Analyzing the residuals, leverage values, and cook's distances for all of the testing data, we noticed there was one row with a significantly higher cook's distance than other points (more than the other two points that unusual cook's distances) When visualizing the model fit with and without the influential point, we observed that the fit was pulled a little towards the influential point. However, we agreed that the fit didn't change a considerable amount, so we continued with our confidence and prediction interval calculations including the row in our data.

Using our model, we calculated that with 95% confidence, the mean median house value for a block with measurements equal to the mean of average in the data is estimated to be between $184570.20 and $219761.10. Additionally, with 95% confidence, the median house value for the particular sampled block is estimated to be between $73722.94 and $313399.90 (the particular block with the following measurements: `housing_median_age` = 44, `total_rooms` = 2526, `total_bedrooms` = 579, `population` = 1423, `households` = 573, `median_income` = 2.5363, and `ocean_proximity` of <1H OCEAN).

While the results of our backwards selection method were strong, we wanted to test how other regression techniques would affect our model, and compare with the model we created using backwards selection.

Shrinkage Methods & Random Forest

Finally, we applied shrinkage methods to find the best regression models for our data. Additionally, we compared the predicted responses of shrinkage methods to those from a random forest.

Methodology

Before applying regression techniques, we looked at the correlation between variables again to determine if there was collinearity in our data. Once determining the presence of collinearity in our data, we performed Lasso and Ridge regression to create two different models. Additionally, we also implemented random forests and compared the predicted response values to that of the other three models.

Results

We observed that there was collinearity between the variables, which indicates that Ridge Regression would be an effective method to create a model. Since the number of predictors is significantly less than the number of observations, we predicted that Lasso regression will not be as effective as Ridge Regression.

Applying cross validation to our model, we got the best lambda value to apply to our ridge regression was about 110.185. Using this optimal shrinkage factor, we got coefficients for our Ridge regression model. Applying cross validation again for our lasso regression, the best lambda value was about 0.001, which we used for our lasso regression model. Superimposing our results from backwards selection, ridge regression, and lasso regression onto a response versus predicted response graph, we observed that the model from backwards selection and ridge regression were far better than the model from lasso regression. Additionally, after applying a random forest to the data set, we noticed that random forests were just as effective at predicting response values as our backwards selection and ridge regression models.

Limitations

The accuracy in predictions from the lasso regression model were never going to do well because of the intended use for lasso regression, while ridge regression was a perfect application to the model given the qualities of the random variable.