Regression Conclusion and Synopsis

Orion Marini & Sara Chong

2023-12-14

1990 California Housing Census Data Analysis

The California Housing census data from 1990 provided by Kaggle provides a plethora of information statistics about the housing of citizens across the state. The data is divided through a manner that encapsulates overall information in each observation as one housing block. A key element of this data that is of up most importance is information about population. This information is relevant as it can be a critical influencing point for social and political movements. To illustrate some examples of it's importance, population of a block can motivate how public transportation systems are arranged to function in the optimal way, it can help businesses predict their customer base as well as their expected revenue from those customers, and population data is significantly influential in drawing political bordering in a nondiscriminatory manner. Needless to say, being able to predict and estimate population values is quite prudent. For that reason, we analyzed the census data in order to determine how other variables of housing statistics can come together and help predict population metrics.

Variables

First, it is of up most importance to look at the variables provided within the dataset. The variables in our dataset, like population, represent information of our data. In our case, the variables latitude, longitude, house median age, total rooms, total bedrooms, total households, population, median income, median house value, and ocean proximity. For the purpose of analyzing the population variable in the dataset, we will look toward the numeric variables, a type of variable that holds a number value. Therefore, we can drop the variables ocean proximity, latitude, and longitude. Now, we are left with a number of variables that may or may not have any relation to our variable of interest, population. In order to find out how these variables relate to one another, we must perform a series of tests using R.

Pair Plotting

From now on, we will be calling the variable representing our population the "response variable" because its values are what we are attempting to predict and estimate for future data in response to other variable values. Now for our first statistical analysis test, we can use R to create a type of plot called a Pairs Plot that graphs each variable's observations versus each other in order to visually look for any particular trends in the data. The Pairs Plot we are referring to can be found on page one of project step 3. We are looking within each graph to see if the plotted data points make any structure similar to a straight line. If any of these plots follow that pattern, then it is quite likely that they have a relationship that we classify as a linear relationship! This tells us that as one variable changes value, the other variable its related to changes its value at a constant rate. Looking at our data, we can see that the variables total rooms, total bedrooms, and total households all have what appear to be a distinctly positive linear relationship with population, meaning that as each one of the three variables increase, then the population variable increases as well. As for the other variables, the pairs plot does not exactly reveal how these variables are associated, so we will have to determine later whether these variables have some effect, negligible effect, or no effect at all on the

values of the population variable. Ultimately, we will be attempting to formulate an understanding of how putting these variables together will allow us to predict and estimate the value of our response variable.

What is a Regression Model

The simplest way of conveying how explanatory variables, the variables that contribute to solving the value of the response variable, is to build what's called a regression model. Regression models are essentially equations that take explanatory variables and give them coefficients that, when combined, all equate to the value of the response variable. In our case of linear regression where each variable contributes linearly to the response variable, a line of best fit called the regression line is created by the model's equation that forecasts where data points fall along given explanatory variable values. In our model, we have a coefficient for each of our variables, as well as one that determines the intercept of the regression line. Putting these together, we normally get a equation that looks something a little like the following:

$$Y = \beta_0 + \beta_1(\text{variable 1}) + ... + \beta_n(\text{variable n}) + \epsilon,\ n = 1, 2, ..., n,\ \text{for n variables}$$

As we are using the basic linear model shown above, we must understand that these variables fall under the assumptions of linearity. Each regression model has five key assumptions: Linear relationship, multivariate normality, little or no multicollinearity, no auto-correlation, and homoscedastictity. If our data follows these trends, then it is safe to follow the model built upon these assumptions. Failing or having questionable assumptions will have different effects on our model based on which assumption that fails, but if some fail, it does not immediately invalid its use, but it does impact the accuracy and precision to which it can produce values given observations.

Passing or Failing the Assumptions of Linear Models?

In project step two and three, we looked at the assumptions of linearity and how our model functioned under these assumptions. Initially, we found out in step 2 page 2 that the three explanatory variables with the greatest sign of a linear pattern in the pairs plot, total rooms, total bedrooms, and total households checked all the boxes of linearity except homoscedasticity. Homoscedasticty essentially means that the variance of residuals, the random variation of the errors within the data, is constant across all levels of the predicted values. Under normal circumstances, the residuals plotted against predicted values gives us a graph that has a random scatter but it is almost evenly spread apart from zero across the horizontal axis. In our case, we found that no matter what transformation we performed to the data, our models would produce a predicted versus residuals graph that showed a dense group towards the lower end of the predicted values, but a greate variability towards the right. This so-called "fan shape" leads us to believe our prior suspicions of heterscedasticity within the data. Fortunately, this does not completely invalidate our chances of building a linear model. We can either decide to manipulate the input data and by result have more complex prediction patterns, or choose to continue and be aware that our models accuracy will become increasingly limited as we look towards higher prediction levels. We chose to go with the ladder option, as it seemed a good portion of our data fell within what we accepted to be accurate within homoscedasticty levels, but it is pivotal to take this into consideration when later viewing our model.

Deciding on our Variables

In step 3 and 4 of our project, we dove quite deep into developing a multiple variable linear model that would be the best predictor of our data. After numerous iterations of passing in variables, we concluded that a linear regression model with all of our numeric variables excluding median_income would be the best predictor of our dataset based on the R-squared value that R was producing. R-squared is a statistical measure that represents the proportion of the variance in the response variable which is explained by the explanatory variables. Essentially, it can tell us how related our explanatory variables are to our response variable. We want this number to be as close to 1.0 as possible. When inputting these variables into R to formulate a regression model, we returned the basic output that follows outlining the R-squared value.

```
R-squared value of Regression Model:  0.8950321
```

Additionally, we can also return the coefficients of each of our inputted variables and intercept in the regression model, which we will plug in momentarily for β_i values. The coefficients, β_i, are as follows:

```
      (Intercept) xhousing_median_age        xtotal_rooms     xtotal_bedrooms
   210.6876125799        2.5329763672        0.1883613985       -1.4101888593
      xhouseholds xmedian_house_value
     3.1146630724       -0.0008948554
```

Lastly, we can also see the significance of each of our variables from the data by looking at their associated p-values. The p-values tell us if each explanatory variable does have an influential role in impacting the regression model. Like before, the p-values from our linear regression model are as follows:

```
      (Intercept) xhousing_median_age        xtotal_rooms     xtotal_bedrooms
    1.030941e-04        6.425980e-02        1.645758e-16        1.520096e-12
      xhouseholds xmedian_house_value
    2.213460e-47        2.077345e-10
```

As we can see, each of these values is significantly small, meaning that each variable plays a significant role in our model.

The Assembled Model

Putting it all together, we can finally asses our linear regression model via a simple formula. Using this formula, values of each explanatory variable can be directly plugged in and equated in order to solve what the predicted value of population is given those inputted conditions Therefore, the linear regression model along with its intercept and coefficients can be layed out as follows:

$$\begin{aligned}\text{Population } &= 210.69 + 2.53(\text{housing median age}) + 0.188(\text{total rooms}) \\ &-1.410(\text{total bedrooms}) + 3.1146(\text{households}) - 0.000895(\text{median house value})\end{aligned}$$

Further information can be gathered from this model as we can visually inspect each partition by variable and see how population increases based on a unit increase in each variable. By example, one additional household increases the population of a block by approximately 3.115 units. Similarly, the coefficients of the other variables indicate how much the population value increases per unit increase of the corresponding explanatory variable.

Further Refinement

Lastly, in project step four, we attempted to refine our model by testing out shrinkage methods to attempt to combat the existence of multicollinearity and model overfitting which drastically limit the viability of our model. We attempted two techniques, Ridge Regression and LASSO Regression, and the latter of which seemed to produce the most accurate predictions of our data in comparison to the Ridge Regression model. For this reason, we saught to replace our model with this as it provided the overall best fit and predictability of any model we developed thus far. The following code is what we inputted to produce our best fit model, followed by the output of the coefficients of the model.

```
6 x 1 sparse Matrix of class "dgCMatrix"
                                s0
(Intercept)        209.8629955603
housing_median_age   2.4957723747
total_rooms          0.1826812584
total_bedrooms      -1.3164242224
households           3.0404369559
median_house_value  -0.0008794261
```

As you can see, the coefficients in this model are quite similar to that of our previous one, but those small changes will have an even greater effect on the data as the values for the explanatory variables are increased with the dataset. Therefore, it is for those reasons that we decided to choose the LASSO Regression fitted linear model as the best model to represent the response variable for population.

What Could Be Wrong

Like stated before, the model we have formulated is not completely accurate. The failing of assumption of homoscedasticity means that as the range of values grows, the accuracy of the model decreases significantly. The graph below illustrates the fan shape of the predicted values versus the residuals of the model which leads to our failed assumption.

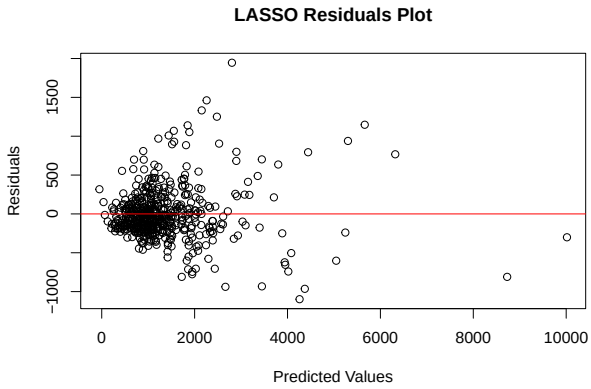

Figure 1: Predicted versus Residuals for LASSO Regression Model

These findings do not completely invalid our model, as long as the end user understands that these assumptions failed when utilizing the model itself.

Conclusion

As a result of our regression analysis, we have concluded our linear model to follow the equation of:

$$\text{Population } = 209.863 + 2.496(\text{housing median age}) + 0.183(\text{total rooms})$$

$$-1.316(\text{total bedrooms}) + 3.040(\text{households}) - 0.000880(\text{median house value})$$

This model will be sufficient for making relatively unimportant predictions on population numbers based on the values of each explanatory variable. The accuracy of the model should cover a broad range of applications despite its pitfalls , but if high precision and accuracy is need, it might be beneficial seeking an alternative model with different variables to allow for better estimates of the response variable. Because of these facts, it is of up most importance that anyone utilizing this linear regression model to understand these caveats

and base their conclusions in accordance with these notices. Nevertheless, we feel comfortable presenting this data as a relatively reliable source of prediction / estimation for the population of California Housing Blocks.