STAT6509 TakeHomeReport Lanqin Zhao

In the "housing price" data set, there are 522 house cases. For each case, the data contain the price, and features such as square footage of the house sqft, number of bedrooms bed, number of bathrooms bath, etc. We are interested in building a regression model to find out the relation between housing price, the responsible variable and its features, the explanatory variables, and to predict housing price with features.

1 Understand Data

First, summarized the data set. I found the minimums of the number of bed and bath are 0, which is strange. Therefore, I checked the data and found that case 108 has price=528750, sqft=2129, but bed=0 and bath=0. This must be a typo. I deleted this case from the data set since the case size is large. The data set analyzed now has 521 cases.
Second, did analysis of correlation. I found that the response variable and explanatory variables, except highway, have linear relations, so linear regression models may be appropriate. Surprisingly, price and quality are negative correlated. I think the description of quality may be wrong. I also found that there are multicollinearity in explanatory variables.
Third, drew scatter plots. I found that price and bed have curvilinear relation, and that price goes up with bath under 5, then goes down.

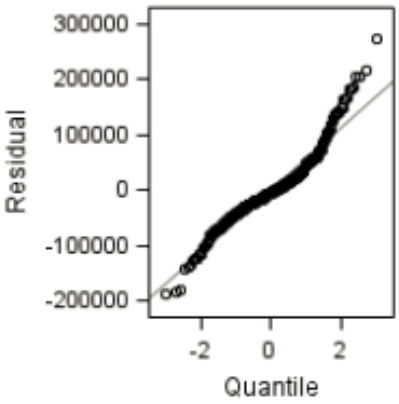
2 Prepare Data

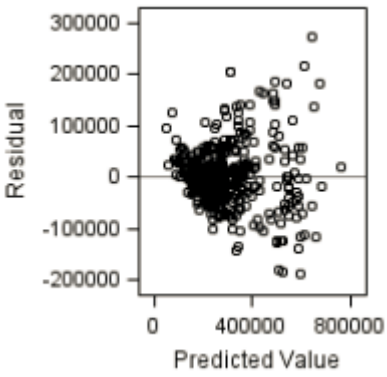
After I had a relatively clear understanding of the data, I created some variables I thought they might be related to housing price as following:
age=2016-year, has more direct meaning and is easier to compute;
quality1=1 if quality=1, otherwise quality1=0 model needed ;
quality2=2 if quality=2, otherwise quality2=0 model needed;
bed2=bed*bed the scatter plot shows price and bed have curvilinear relation
bath4=5 if bath>4, otherwise bath4=bath the scatter plot shows price goes up with bath under 5, then goes down;
bed_cars=bed/cars1 represents how many people sharing a parking spot, and I used cars=0.0001 to replace cars=0;
sqft_bed=sqft/bed represents living space per person;
bed_bath=bed/bath represents how many people sharing a bathroomt
sqft_lot=sqft/lot represents the ratio of square footage and lot size

3 Check Assumptions

Linear regression models assume that error terms are independent, normal with mean 0 and

constant variance, response variables and explanatory variables have linear relations. I build a regression model with all the above variables to check normality and constant variance assumptions.

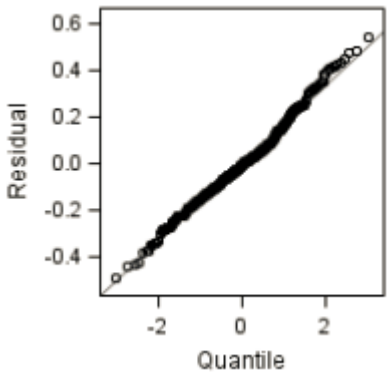


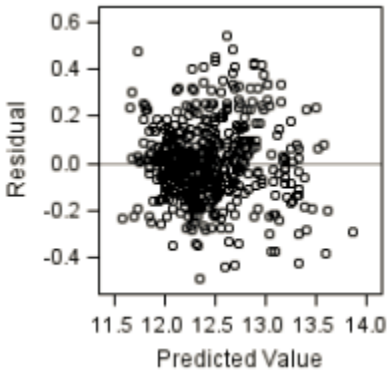


The QQ plot shows that normality is violated since residuals are heavily skewed, and Residuals against Predicted Values plot shows constant variance is also violated because variance has a megaphone shape. Then I decided to transform the response variable price to correct these. To find a appropriate transformation, I used Box-Cox Transformation and got log(price).

I drew scatter plots again to see the relations between the new response variable and explanatory variables, and found the relations almost stay same.

Finally, I used transformed response variable log(price) and all original and created explanatory variables to model to check the assumptions. The following QQ plot shows that normality is nearly met, and sample size 521 is large, based on Central Limit Theorem, so normality is not an issue here. And Residuals against Predicted Values plot shows variance is much closer to be constant. I also did lack of fit test to test whether the linear regression model is a good fit, and the p-value= 0.5488 approved the assumption is met. Now, I'm ready to build linear regression models.





4 Build Models

I used stepwise, forward, and adjusted R^2 methods to select variables. Stepwise and forward

methods gave the same results. The 9 explanatory variables: sqft, bath, ac, cars, lot, quality1, quality2, age, bath4 were selected. This result is good. Grouping variables quality1 and quality2 both were selected. Adjusted method selected 11 variables: sqft, bed, bath, ac, cars, lot, quality1, age, bath4, sqft_bed, and bed2. Since grouping variables quality1 and quality2 both should be included, I added quality2. I built 2 models with the selected 2 subset of explanatory variables to validate the selected regression models by comparing PRESS with SSE, CP with the number of explanatory variables for each model. Both models were valid and performed comparably in the validation study. Based on the principle of parsimony, the model with 9 variables was ultimately chosen as the final model.

5 Explain the Model

The fitted regression function is:
log(price).hat=11.5274+ 0.00025864sqft- 0.18324bath+ 0.055902 ac + 0 .033071cars+ 0.000005173lot+ 0.35473quality1+ 0.072719 quality2- 0.003683415 age+ 0.24994 bath4
This function indicates:
Holding all other factors constant, if we increase the square footage of the house by 1, we estimate that the average log(price) increases by 0.00025864. Holding all other factors constant, if we increase the number of bathrooms in the house with 4 or fewer bathrooms by 1, we estimate that the average log(price) increases by 0.24994-.0.18324=0.0667; if we increase the number of bathrooms in the house with 5 or more bathrooms by 1, we estimate that the average log(price) decreases by 0.18324. Holding all other factors constant, if we add air conditioning to the house, we estimate that the average log(price) increases by 0.055902. Holding all other factors constant, if we increase the number of cars the garage can hold by 1, we estimate that the average log(price) increases by 0 .033071. Holding all other factors constant, if we increase the lot size by 1, we estimate that the average log(price) increases by 0.000005173. Holding all other factors constant, we estimate that the average log(price) increases by 0.35473 for low quality houses than for high quality houses. Holding all other factors constant, we estimate that the average log(price) increases by 0.072719 for medium quality houses than for high quality houses. Holding all other factors constant, if we increase the age the house by 1, we estimate that the average log(price) decreases by 0.003683415.
This model is almost like what we expected. The interesting thing is about house quality. Again, I think this is caused by the quality description, which is wrong. Furthermore, Housing price is not associated with the number of bedroom, whether having a pool, whether the house is next to a highway. This is a little surprising. .

6 Apply the Model

I used the model to estimate sale prices of the two houses. The results are:
I'm 95% confident that the estimated average log(price) for house 1 with 1500 square footage is between 12.3921 and 12.4547, and that for house 2 with 3500 square footage is between

12.6415 and 12.9126. Transformed log(price) and got that the estimated average sale price for house 1 is between $240891 and $256452.8 and for house 2 is between $309124.7 and $405388. There's about a 99% chance that the estimated log(price) for house 1 is between 12.0310 and 12.8158, and that for house 2 is between 12.3631 and 13.1910. Transformed log(price) and got that the estimated sale price for house 1 is between $167879.2 and $367985.9 and for house 2 is between $234005.5 and $535523.5.