Multivariate Regression Modeling to Estimate Home Prices in Brooklyn

Abstract

Among different types of predictive modelling techniques, regression analysis is commonly employed for its ability to explore the relationship between a dependent and independent variable(s), which can then be used to estimate a value of a dependent variable given a new set of data. This paper seeks to build a predictive model that aims to predict house prices in Brooklyn using a provided dataset of 37 homes through the use of multivariate regression. We applied best subset analyses to come up with a model with explanatory variables being able to capture the most variability in the outcome variable, and thus has predictive powers in evaluating the worth of a Brooklyn house that is not on the market. We found a predictive model with the adjusted R-squared at 88%, and concluded there was a strong relationship between the predictors and the observed home prices.

Key words: multivariate regression, predictive modelling, housing prices, best subset analyses.

Introduction

Linear regression is used in predictive modelling to devise a function from a readily available dataset which, if fed a set of explanatory variables, will forecast an outcome with the best probability. This is of interest since being able to create such a tool with predictive ability aids people significantly in making more informed, accurate decisions in many disciplines, from public policy, health care, to business. In this specific case of the real estate market, such a predictive model is valuable to customers in that it may help them make more financially sound choices when it comes to making big investments in houses. Considering that Brooklyn is a big borough with a diverse offerings of real estate deals, it is important that this project aim to come

up with a multivariate that can most accurately estimate prices of houses in Brooklyn. Data was obtained from Zillow and readily available on a number of measures, including address, finished size, lot size, year built, last sold date, and last sold price. When working with these types of variables, we should decide which measures are important for us in building the right model. In addition, our goal was also for the model to be parsimonious, which means we attempted to create the best fitted model with as few independent variables as possible.

It is worth noting that the final model also suffers from issues that should be taken into consideration. These include limitations brought about by small available sample size and non-inclusive set of variables of interest, as well as the difficulty to quantify some nominal variables.

Methods

We first examined the data by looking for relationships between our dependent variable amount and the predictor variables (**Figure 1a-f**). We defined two new variables, with age as the number of years passed since the houses were built and Dayspassed as the days passed since the houses were last sold. To rule out multicollinearity, we also examined the scatterplot matrix of the predictors (**Figure 1g**) and then the VIF values for the full model that includes all predictors.

Then, we used the best subsets analysis to determine which variables we should include in our model. We used the adjusted R^2 value (R_{adj}^2), Mallow's C_p (C_p), and Bayes' information criterion (BIC) as criteria for determining the best model. The model that had a relatively low number of parameters (p), high R_{adj}^2, C_p close to p, and low BIC was chosen. Then, we examined the LINE assumption for the model: linear relationship between the dependent and independent

variables, independent observations (no auto-correlation), normality of residuals, and equal variance of residuals (homoskedasticity). We plotted our independent variable to the predictor variables (**Figure 1a-f**), the residuals against the fitted values of our model (**Figure 3a**), a normal Q-Q plot of the residuals (**Figure 3b**), histogram of the residuals (**Figure 3a**), and residuals against observation order (**Figure 3d**) to test the assumptions.

Our chosen model seemed to violate all LINE assumptions, and we therefore sought to incorporate flexible terms as a solution. We conducted two best subset analyses using quadratic terms and interaction terms, respectively. We chose one model from each best subset analysis using the same criteria as before: relatively low number of parameters (p), high R_{adj}^2, C_p close to p, and low BIC. For both models, we were faced with significantly lowering our R_{adj}^2 or incorporating higher-order terms while excluding the first-order variables within those terms (e.g. including finishedsqFt^2 but not finishedsqFt). However, the goal of the model was to predict rather than confirm a theory, and we therefore chose the latter option. We then examined the variance of the residuals for both the interaction and quadratic models. As the variance was found to be unequal in both, we *log* transformed the dependent variable and examined the variance of the residuals. However, the transformation did not significantly affect the variance, and we therefore discarded the transformed model.

Of the two models--one that included interaction terms and another that included quadratic terms--the former one had a higher R_{adj}^2 and was therefore chosen as our model. Since heteroskedaticity was observed in our chosen model, we conducted a bootstrap analysis on the model to determine coefficient values and the 95% confidence interval for each coefficient.

Results

No significant correlation was found between predictors based on the scatterplot matrix of the variables (**Figure 1g**) and the VIF values (all VIF < 5). From our initial best subsets analysis, we decided to choose the following model, as it had a high R_{adj}^2 of 0.77, a low C_p of 2.5, and a low BIC of -41 (**Figure 2**):

$$\text{E[House price]} = \beta_0 + \beta_1 \times \text{age} + \beta_2 \times \text{lotSizeSqFt} + \beta_3 \times \text{finishedSqFt} + \beta_4 \times \text{lastSoldPrice} \quad \ldots (1)$$

However, upon examination of residuals, we found that the model violated the LINE assumptions (**Figure 3**). We therefore performed two additional best subset analyses incorporating flexible terms: one with interaction terms, and another with quadratic terms. We used the same criteria to choose the model with interaction terms (**Figure 4;** Equation 2; R_{adj}^2=0.88, C_p=3.8, BIC=-63) and the one with quadratic terms (**Figure 5;** Equation 3; R_{adj}^2=0.84, C_p=4.8, BIC=-51). Upon examination of the LINE assumptions for both models, we found that both had unequal variance of residuals (**Figure 6**). We therefore log-transformed the models, only to find that the residuals' variance did not change much from the original models (**Figure 7**). Instead, we decided to choose the model with interaction terms (Equation 2)

$$\begin{aligned}\text{E[House price]} = {} & \beta_0 + \beta_1 \times \text{age} + \beta_2 \times \text{lotSizeSqFt} \times \text{finishedSqFt} + \beta_3 \times \text{lotSizeSqFt} \times \\ & \text{lastSoldPrice} + \beta_4 \times \text{finishedSqFt} \times \text{lastSoldPrice} + \\ & \beta_5 \times \text{lotSizeSqFt} \times \text{age} \end{aligned} \quad \ldots (2)$$

$$\begin{aligned}\text{E[House price]} = {} & \beta_0 + \beta_1 \times \text{age} + \beta_2 \times \text{lotSizeSqFt} + \beta_3 \times \text{age}^2 + \beta_4 \times \text{lotsizeSqFt}^2 + \\ & \beta_5 \times \text{finishedSqFt}^2 + \beta_6 \times \text{lastSoldPrice}^2 \end{aligned} \quad \ldots (3)$$

For model evaluation, we used bootstrap to find the 95% Confidence Interval and coefficient of the bootstrap distribution, which provides an approximation of those of the sampling distribution (Kuiper and Sklar, 2013). Without having to worry about violating the LINE assumptions--in our case, unequal variance of residuals--the coefficients for both the bootstrap and normality-based are similar. The coefficients for both the bootstrap sampling distribution (which attempts to asymptotically “mimic" the distribution of the population) and our model are fairly similar. In other words, the coefficients are not too off from what is the model's true coefficients, which validates the use of our current model for predicting house prices

Discussion

As seen from the regression model, all the variables in the final model are significant and contributive to the predictive model. As a result of the best subset analysis, we excluded variables “days passed”, “bathrooms", “bedrooms", which means when the house was purchased, and number of bedrooms and bathrooms might also affect the price, but to a very limited degree. Although inspection of the residuals versus fitted values still lacks consistency, the bootstrap method can be used to calculate the same parameters as well as the confidence intervals by constructing the distribution that resembles the sampling distribution, which we think, in this case, shows that we have calculated the parameter estimates well enough.

Despite the fact that it is generally agreed that a regression should not include higher-order terms without the lower-order terms that are part of the higher-order terms (Aiken & West, 1991), the goal at hand is to find the model with the highest predictive ability.

Considering “there is nothing wrong with including interaction effects by themselves” (Cleves et al., 2008), we allowed for exclusion of some main effect terms while keeping the flexible terms. It was expected that the final model (with introduced flexible terms) would provide a good explanation of the observed home prices, which was confirmed by high R-squared value of approximately 0.88

We also need to take into consideration the fact that this model is based off a relatively small dataset with a limited number of measures (most are qualities and conditions of the homes themselves). We expect a model with better predictive power to include information about external conditions (environment quality, neighborhood crime rate, school API, et.c) on top of variables denoting innate properties of homes.

Model's limitations and issues that need further consideration

We suppose location of the house could be a good predictor in the model. However, the nominal variable “address” is hard to quantify directly. We have considered using zip code or neighborhood (“Greenpoint", Williamsburg", etc.), but given the dataset provided, some zip code or neighborhood would only have one datapoint, which can be problematic since we cannot tell whether or not such datapoint is representative of its group. Another reason why we think in this specific case, zipcode as a predictor would not work that well is that Brooklyn has more zip codes than provided in the dataset. Thus, any zip code other than those five in the dataset fed to the model would make this categorical variable rather useless.

We converted what otherwise could have been categorical variables “date last sold” and “year built" into continuous variables “days passed since last sold" and “home age". “Date last

sold" is another example of difficulty in potential interpretation of categorical variable, whereas the conversion of "year built" into "age" was merely attempting to make the predictive model more "functional" and not prone to failure when fed new information that is not part of the set of categories - "year built"'s - the model used to explain house price.

The model with log transformation of the dependent variable "amount" we obtained in an effort to fix heteroskedasticity had a lowered R-squared. And while this transformation made the data points in the residuals versus fitted values plot neater, it did not very effectively fix unequal variance problem. Therefore, we decided to go with the original model with interaction terms but without log transformation.

In the end, even though we attempted to find the best predictive model given the available set of explanatory variables, we believe that there are other external factors that, if accounted for, might have allowed for a better multivariate regression fit and thus stronger predictive power.

Conclusion

The strength of popularity of multivariable regression in making models with predictive powers is undeniable. In this paper, we attempted to build such a predictive model by fitting a multivariable regression on a set of observed available data points. We used adjusted R-squared, Mallow's Cp, as well as Bayesian Information Criterion (BIC) as the metrics to select the best subset of predictors. We also attempted to correct violations of linear regression assumptions by including flexible terms and transforming variables. In the end, we obtained statistical significance of coefficients and high adjusted R-squared value, which indicated high strength

between the estimated outcome variable and the predictors in our final best fit regression model. The process of building such regression models may be applied to many fields where the ability to forecast outcomes from a given set of predictors is desired.

References

Aiken, L.S., West, S.G. and Reno, R.R., 1991. *Multiple regression: Testing and interpreting interactions.* Sage.

Kuiper, S. and Sklar, J., 2012. *Practicing statistics: Guided investigations for the second course.* Pearson Higher Ed.

Mario, C., Gould, W. W., Gutierrez, R.G. and Marchenko, Y., 2008. *An introduction to survival analysis using Stata.* StataCorp LP.

Figure 1. Testing for linearity between independent/dependent variables and multicollinearity.

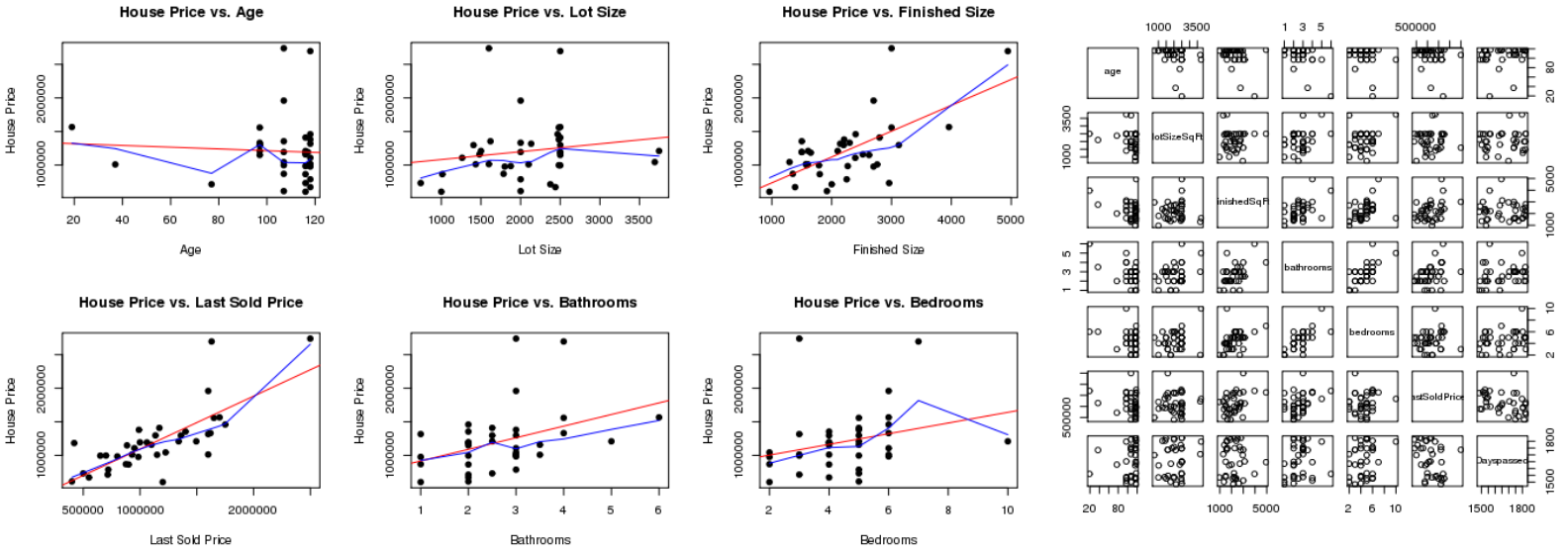
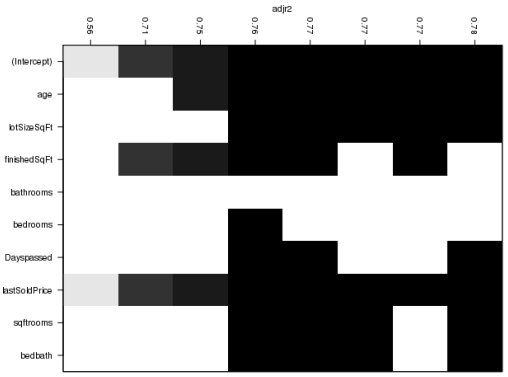
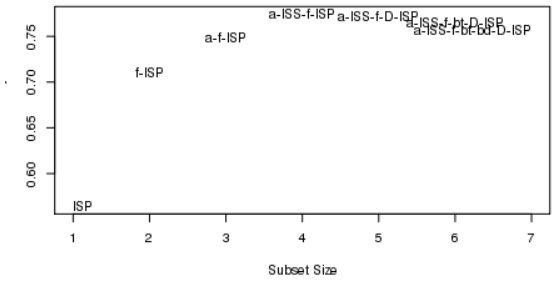


Figure 2. Initial best subsets analysis results. Black indicates terms that are included.

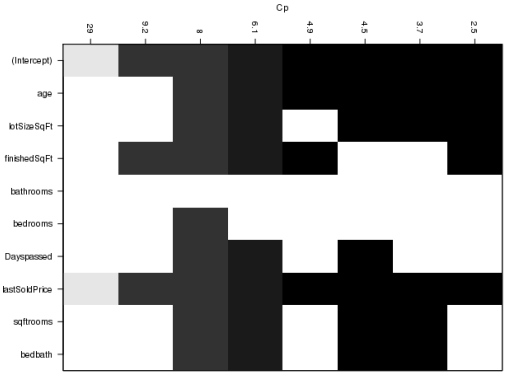
a.



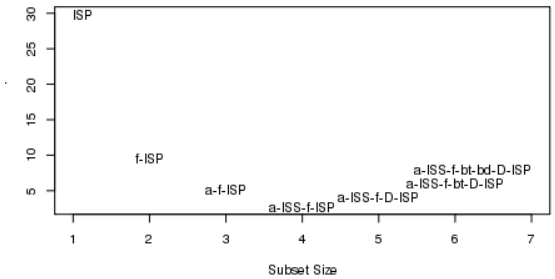
d.



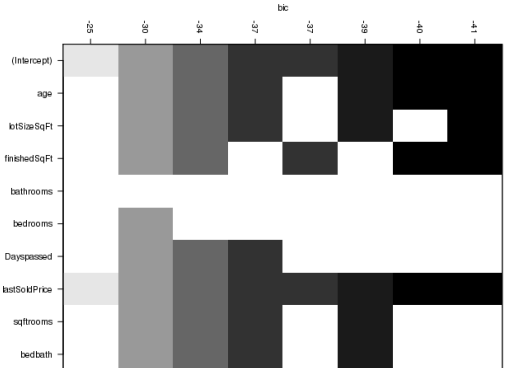
b.



e.



c.



f.

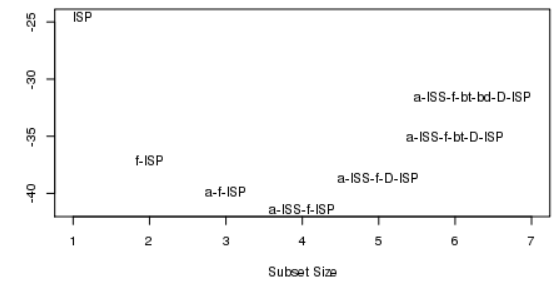


Figure 3. Testing for LINE assumptions for initial model.

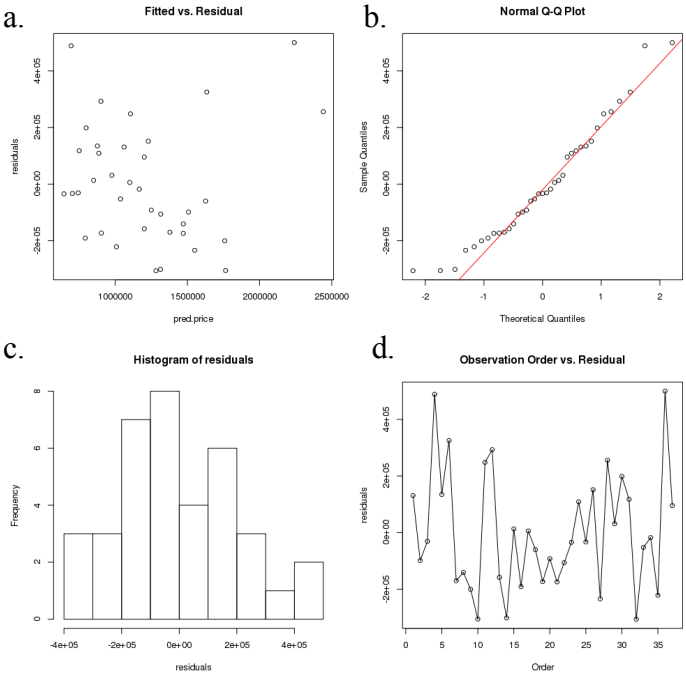


Figure 4. Best subsets analysis including interaction terms.

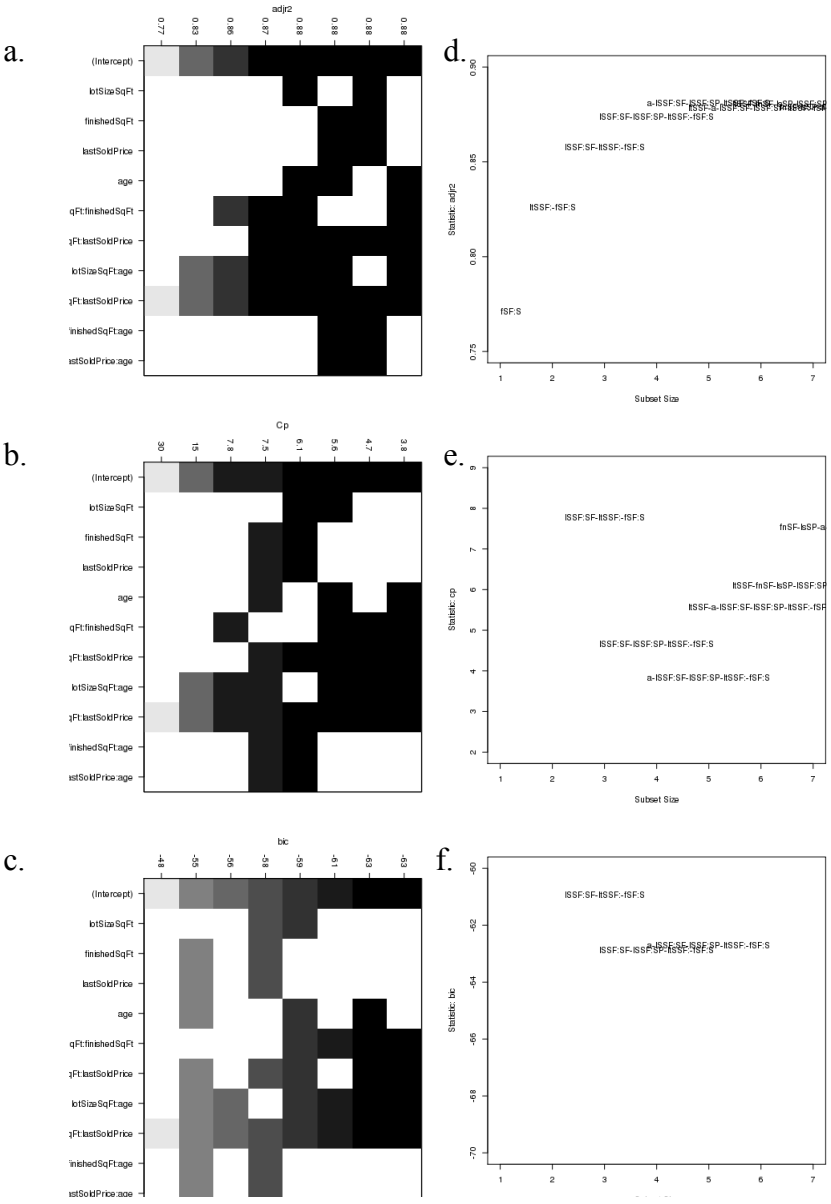


Figure 5. Best subsets analysis including quadratic terms.

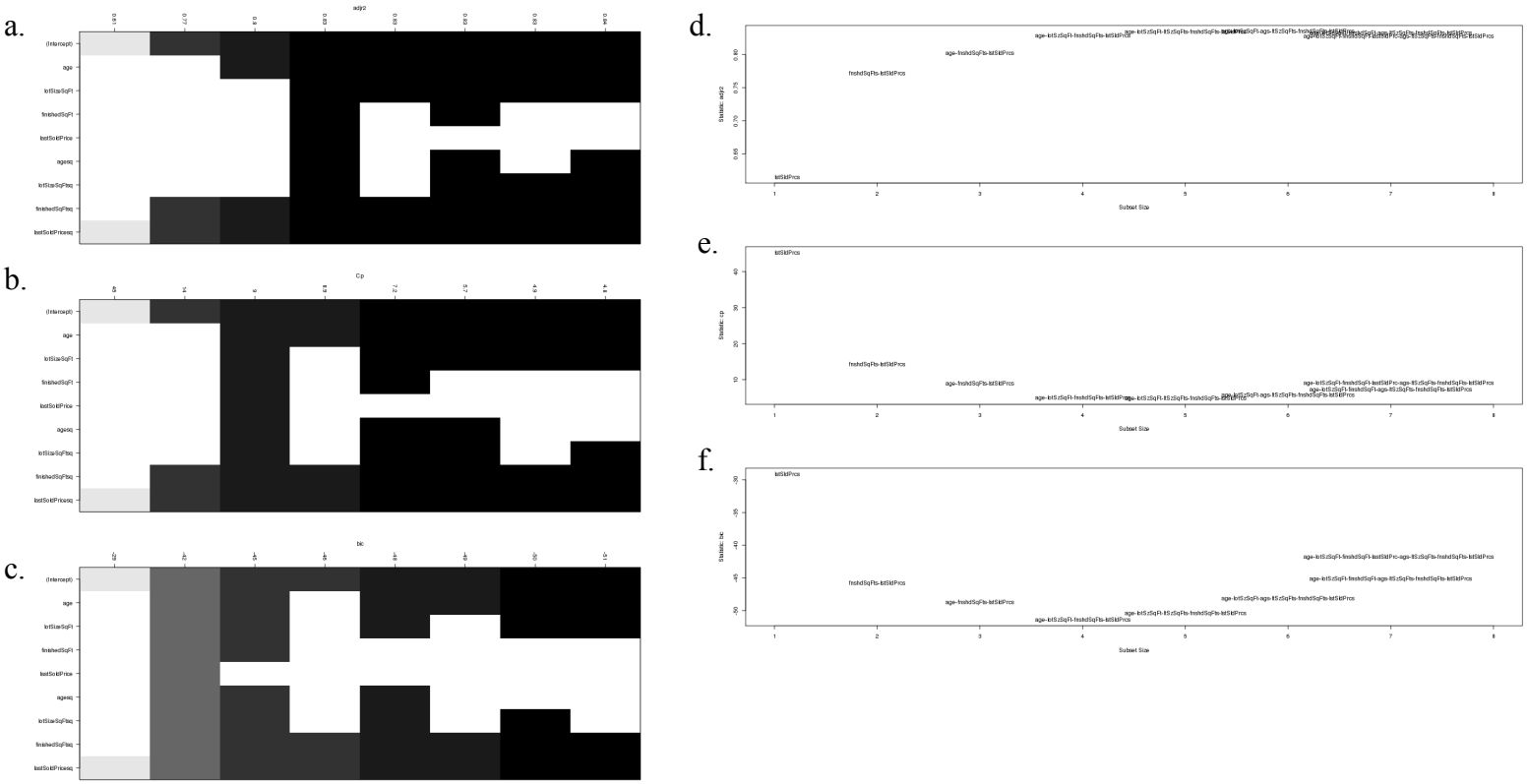


Figure 6. Testing LINE assumptions for the two chosen models with quadratic (a-d) and interaction (e-h) terms

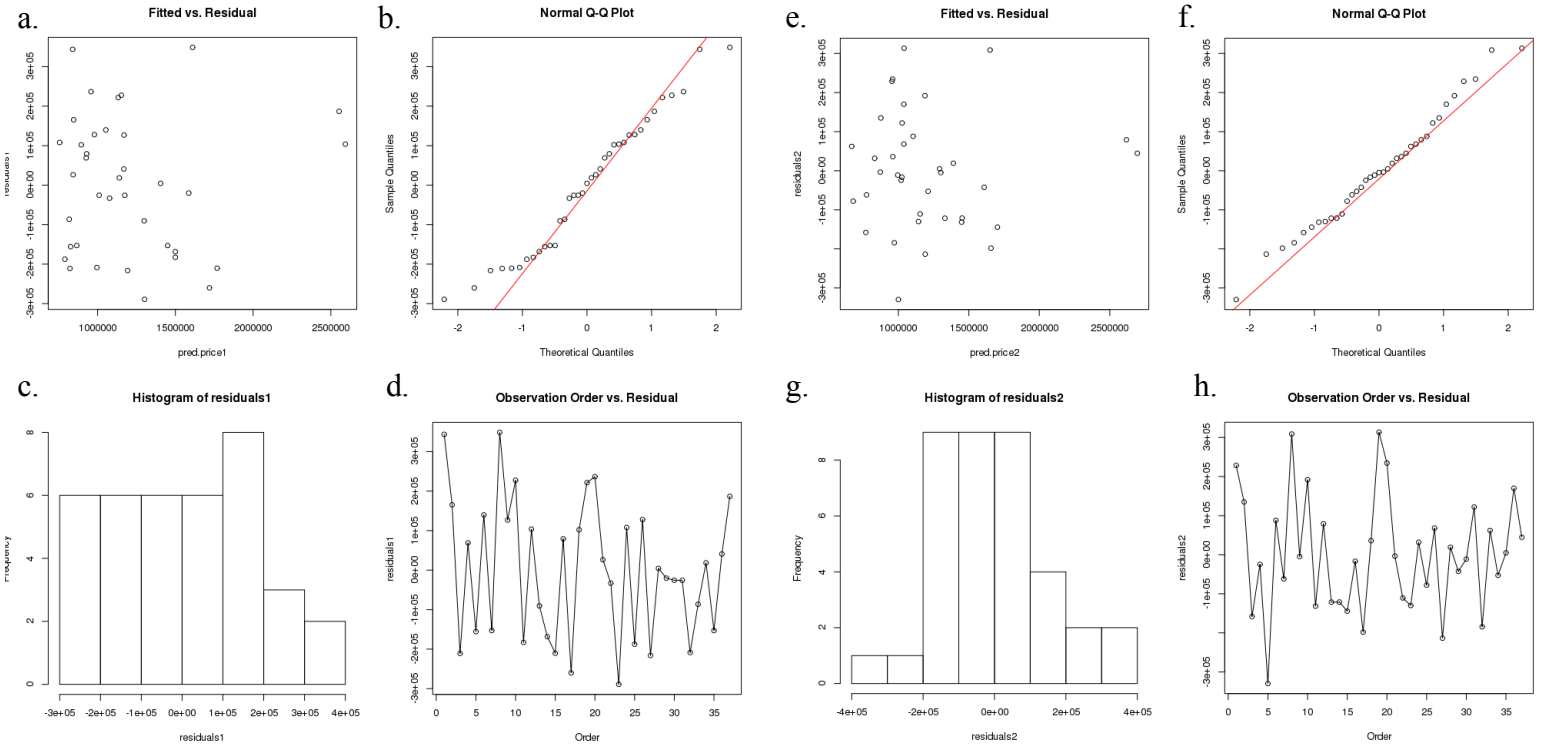
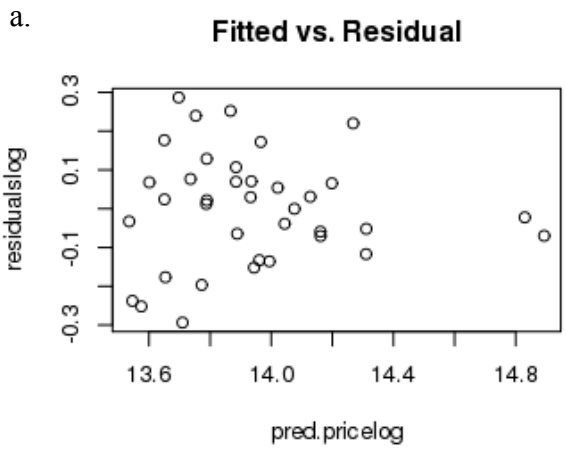


Figure 7. Examining residuals for the log-transformed models of the quadratic (a) and interaction (b) models.

a.



b.