DSC 423 Data Analysis and Regression

Spring 2019

Final Project

I travelled 15 hours and All I got was a Futon?

Ying Kam Chiu

Abstract

The goals of the analysis, regarding AirBnB rentals in Melbourne, Australia, was (1) to determine which independent variables have the greatest significant impact that align to the price of a nightly stay at an AirBnB listing and (2) to provide a final model that has a high degree of confidence to predict future outcomes.

The publicly available dataset was obtained from Kaggle.com. The team initialized the analysis on the dataset during the pre-processing stage by exploring which of the variables could be categorized as a dependent variable, independent variable, or a variable not to be used in the analysis. At this stage we were also able to reclassify the 250+ postal codes into 5 regions that constitute Melbourne. Subsequently we randomized and divided the 22K observations from the full data set into blocks of 2.5K observations, of which each of the 6 team members received their own unique randomized dataset. During the exploratory stage we chose to use linear regression to explain the chosen response variable Y (Price). Transforming the Y variable was necessary in order to stabilize the variance. Each team member fitted the full model to arrive at a final model that had adjusted-R2 ranging between 54% - 66%.

Qualitative variables that proved to be a constant theme amongst the team were determined to be included in a successful model were host response time, region, and room type. The number of people the property accommodates, and the rating score were the most prevalent quantitative variable that was included in all final models. Variables such as if the owner was a superhost, the bed type listed, and cancelation policy were deemed not to be as relevant and were dropped from most final models.

In order to provide a model that has a higher level of confidence to determine the price an AirBnB rental, additional variables or variables with complete information are required to be included in further improving the final models presented.

Introduction

In recent years, the sharing economy has gained popularity in many industries, such as transportation (e.g. Uber, Lyft) and accommodation (e.g. Airbnb, HomeAway). The sharing economy can be defined as "peer-to-peer-based activity of obtaining, giving and sharing the access to goods and services, coordinated through community-based online services". In the peer-to-peer (P2P) accommodation industries, Airbnb is a leader in the market, in which its business model enables hosts to offer their unoccupied properties or rooms for short-term rental. Since its establishment in 2008, Airbnb's business has grown significantly and served more than 150 million guests through over 3 million listings in more than 190 countries in less than a decade.

Melbourne ranked as the 6th on the list of top ten cities for users globally in 2016 and has been one of the top ten cities since then. Not only limited to short-term rentals, Airbnb in Melbourne has entered the market competition of long-term rentals. According to Rawnsley and Schmahmann (2018), "The median number of nights hosted per year has increased from 42 nights to 66 nights per year. Of the listings that have hosted guests, over 35 percent of listings host guests for up to 30 nights per year. Approximately 27 per cent of listings host guests for more than 180 nights per year." Even though the article stated that the impact of Airbnb on the Melbourne housing markets appears minimal, it is still interesting to understand how Airbnb's pricing has enabled itself to extend to another new market.

Listing price is often considered to be one of the critical factors that impact consumer's choices of lodging. In this paper, our objective is to assess different listing prices per night of Airbnb listings in the city of Melbourne with a number of factors (independent variables), such as number of listings that the host has, the maximum number of guests that the listing accommodates, required amount of security deposit, required amount of cleaning fee, review scores rating, the average time that the host responses to a guest's inquiry, whether the host is a super host, room types, bed types, types of cancellation policy and locations.

More importantly, we intend to go beyond the data that we already have direct access to and explore how the mentioned independent variables play a role in the pricing decision. The dataset consists of 8-year of Airbnb listing history from 2010 to 2018 in Melbourne and was consolidated in December 2018. The population of the dataset has 22,895 observations and each of us is given 2,500 observations as sample to begin our individual analysis.

We conduct the analysis through linear regression and draw conclusions on final predictive models to predict future Airbnb listing prices per night in Melbourne. Since each of us begins with different samples, we may come up with different final models. It would be interesting to identify which independent variable(s) is/are significant that will be included in most of our models and how well our models are to predict future prices with unseen dataset.

Methodology

In this paper, Melbourne, a city of Australia was chosen as the study site. The dataset of Melbourne Airbnb Open Data was obtained from website Kaggle (https://www.kaggle.com/tylerx/melbourne-airbnb-open-data#listings_dec18.csv), which provides Airbnb listing information from Airbnb.com. In accordance with the source, Melbourne was the 6th on the list of top ten cities for users globally in 2016 and has been one of the top cities for listings globally since then.

In the pre-processing stage, I identify the selected 12 predictors with a quantitative variable – listing price per night *(price)* as the dependent variable and 11 other independent variables, including quantitative: number of listings that the host has *(host_total_listings_count*), maximum number of guests that the listing accommodates *(accommodates)*, amount of security deposit *(security_deposit)*, amount of cleaning fee *(cleaning_fee)* and review scores rating *(review_scores_rating)* and qualitative variables: average time the host responses to a guest's inquiry: "N/A", "within an hour", "within a few hours", "within a day" or "within a few days' *(host_response_time),* whether the host is a super host: "t" or "f" (host_is_superhost), room type: entire apartment, private room, shared room *(room_type)*, what the type of the bed: real bed, sofa bed, futon, airbed, pull-out bed *(bed_type)*, type of cancellation policy; strict, moderate or flexible *(cancellation_policy)* and region.

Out of 22,895 observations, our team randomized the population in excel and each team member was assigned with a different sample of 2,500 observations. There are some data issues needed to be fix in excel before loading into SAS. Blanks for qualitative variables like security_deposit, cleaning_fee and review_scores_rating were replaced by zeros. For variable region, it is not an original variable from the dataset. With reference to Bob (2019), we divided Melbourne into 5 regions: Inner Melbourne (IM), Northern Suburbs Melbourne (NSM), South Eastern Melbourne (SEM), Eastern Melbourne (EM) and Western Melbourne (WM) by using given zip codes and cities. In addition, there are 9 observations removed because they have no location-identifier and therefore there is no way to verify the data. For variable cancellation_policy, there are 4 layers of strict policy, such as "strict", "strict_14_with_grace_period", "super_strict_30" and "super_strict_60". The 4 alike layers were consolidated into one layer as "strict". The blanks for host_response_time were replaced with "N/A", representing either the host did not or was never needed to respond. There is one blank for host_is_superhost, which was replaced with 'f". After all the data-preprocessing and data cleansing, there are 2491 observations left to load to SAS for further analysis.

In the data exploration stage, I took an overview into the dataset through data visualization. For univariate analysis, I built histograms and boxplots to explore quantitative variables. These figures give a general understanding about the central tendencies and the spreads of the variables. Transforming the variables will be necessary if the histogram shows a skewed distribution. With reference to Gut and Herrmann (2015), dummy variables are computed for qualitative variables: 4 dummy variables that indicate the average time the host responses to a guest's inquiry between dHRT0 and dHRT4, a dummy variable that indicates whether the host is a super host between dHIS0 and dHIS1, 2 dummy variables that indicate room type between dRT0 and dRT2, 3 dummy variables that indicate bed types between dBT0 and dBT3, 2 dummy variables that indicate cancellation policy type between dCP0 and dCP2 and 4 dummy variables that indicate regions between dRG0 and dRG4. At the same time, an interaction variable for two selected independent variables dHRT1 and dHIS was built.

For bivariate analysis, scatterplots were built for each independent variable against the dependent variable or a scatterplot matrix to observe the patterns displayed and the relationship between all independent variables within a single matrix. Boxplots are also a good visualization tool to compare attributes of the qualitative variables.

At this phase, a basic understanding of the dataset was obtained to better identify data issues. Then, I proceeded to detect if there was more missing data. If there are only a few missing values and they appeared to be random, I may proceed with the deletion of these cases. If not, I may replace the values with the median, mean or mode.

After fixing data issues, I checked for outliers through data visualization, such as scatterplots, and histograms. I kept the outlier or verify if it fit the dataset. "Verified" meant that the data does not fall too far outside our expectations and can be reasonably assumed considering this is the holiday season. Otherwise, I would report how the outlier would change the regression line or analysis result and remove it.

At the end of the exploration stage, I tried to fit a regression model. The criteria for an accurate model come from checking of assumptions and 5 indicators: (1) parameter estimates by beta weights, (2) high significance by t-test p-values (<0.05), (3) F-values and their p-values (<0.05) (4) low RMSE/ MSE and (5) a high R2 and Adj- R2 values. With reference to Christensen, we can test with the "error term, which is the residual that cannot be explained by the variables in the model". The assumptions are that "the error terms are independent from one another, identically distributed (constant variance), linear and normally distributed". If constant variance, independence and normality assumptions are violated, the regression line estimate are still unbiased but standard errors, confidence intervals and prediction intervals will be incorrect. In order to stabilize the variance, we have to transform the variables. If linearity assumption is not satisfied, a more complex regression would be used. The next step was to detect severe multicollinearity with a scatterplot matrix and a Pearson correlation matrix for each pair of the independent variables.

After fitting a full model, it entered the analysis stage to identify severe multicollinearity by computing VIF statistics. If the VIF is above 10, it indicates that the regression coefficient of the variable is poorly estimated. Then, I continue checking for outliers and influential points with the help from studentized residuals and cook's D. Observations would be flagged if they are classified as both outliers and influential points. I would remove observations one by one and then rerun regression for the remaining data to check for the next flagged observation. It would be a judgement call to either delete or keep the observation.

Following the outlier test, I checked to verify if the assumptions: constant variance, independence, linearity and normality, hold. If the residual plots of full model comparing the dependent variable against each independent variable test well, I could reasonably assume that the data observed has constant variance, independence, linearity and normality. Should they fail these tests, transformations will be invoked.

After transforming the variables, I then tested the performance of the full model by splitting the dataset into train and test sets. The train set was used to select the "best" variables to fit the final model by comparing two model selection methods, such as forward and backward methods. Checking of assumptions, the above mentioned 5 indicators of an accurate model and diagnostics for multicollinearity were performed again. Test performance was factored in to decide on the final model. The test performance trumps everything. If the test set has a lower RMSE, higher R^2 and adj-R^2 than the train set and Cross-Validated R^2 (CV-R^2) is smaller or equal to absolute 0.3, the final model is said to be a good model. Lastly, I used the validated final model to perform two predictions.

Analysis, Results and Findings

In application to my sample, I started with plotting a histogram for univariate analysis to predict Airbnb listing prices in the city of Melbourne, Australia against 11 other independent variables. Out of the 2,491 selected properties, the central tendency for predicted price was at AU$150.7, which is fairly low. Properties of an average amount AU$0 has indicated a very low in the pool, while some had much higher average price of AU$8,000. With the middle 50% of predicted price between AU$69 and AU$168 among all selected properties, the median of AU$112 is slightly closer to lower quartile. The spread of the distribution was very large with a range of AU$8,000. The mean of distribution was greater than the median, so the distribution is said to be right/positively skewed and unimodal (figure F.1). The distribution also had a peaked top with lighter tails. There is a potential outlier when predicted price was at AU$8,000, which is to the right of the graph. Skewed distribution will not generate a good predictive analysis.

Therefore, predicted price must be transformed with log.

After applying log transformation, figure F.2 shows a much more symmetrical and normal distribution for lnprice (transformed predicted price). The mean of distribution of lnprice (AU$4.72) is now the same as the median. Properties with lnprice on an average amount of AU$2.64 are in the lowest, while some have a high amount of AU$8.99. With the middle 50% of lnprice between AU$4.23 and AU$5.12 among all selected properties, the median is about right in the middle of lower and upper quartiles. The spread of the distribution is much smaller than predicted price before transformation, with a range of only AU$6.35. The distribution has a flat top/ heavy tails (Kurtosis<3) with potential outliers to the right of the graph as well.

Majority of the independent variables in this dataset are qualitative and therefore dummy variables are computed to better analyze the data. At the same time, an interaction variable for two selected independent variables dHRT1 and dHIS is built. According to Shatford (2018), to qualify as a super host, one should host a minimum of 10 stays in a year, respond to guests quickly and maintain a 90% response rate or higher, have at least 80% 5-star reviews, etc. Therefore, a super host should have a joint effect with the average time the host responses to a guest’s inquiry is “within an hour” on lnprice. Consumers tend to choose Airbnb properties with high review scores and if the host is a super host. These properties are believed to be more pleasant and host is nice and reliable. Therefore, lnprice is believed to be higher. The newly created interaction term is called “HRT_superhost”.

In the interest of length limitation, I only explore the dataset for two variables dRT1 and dRG4 with boxplots. Boxplots are built to evaluate how lnprices vary by dRT1 (Private room) and dRG4 (NSM). From figure F.3, it is obvious that the middle 50% of properties of entire home/apartment (dRT1=0) and properties of private room (dRT1=1) in the sample are quite different. Both types of properties seem to have low lnprices since the boxes were closer to the lower extremes. 75% (upper quartile) of selected properties of entire home/apartment is under AU$5.5 and the remainder 25% were up to AU$8.1. For properties of private room, the box is much lower than properties of entire home/apartment, with its 75% (upper quartile) lower than the lower quartile of properties of entire home/apartment. The 75% of properties of private room is about AU$4.2 and the remainder 25% are up to AU$9. Given much longer whisker for properties of private room, it is interpreted that it varies wider in lnprice from AU$2.5 to AU$9. Properties of entire home/apartment swings less from AU$2.8 to AU$8.1. The means for both room types were very close to their medians. Mean and median for properties of entire home/apartment overlapped, which indicates the distribution of lnprice should be normal and symmetric. Relatively, the mean is higher than the median so the distribution of

Inprice for properties of private room is slightly skewed right.

From figure F.4, it is obvious that the middle 50% of properties in region IM (dRG1=0) and properties in NSM (dRG1=1) in sample are quite different. Both types of properties seemed to have low Inprices since the boxes were closer to the lower extremes. 75% (upper quartile) of selected properties in IM was under AU$5.5 and the remainder 25% are up to AU$9. For properties in NSM, the height of the box was longer than properties in IM, i.e. wider difference between its 1st and 3rd quartiles. The 75% of properties in NSM was about AU$5.5 and the remainder 25% were up to AU$7. Given much longer whisker for properties in IM, it was interpreted that it varied wider in predicted Inprice from AU$2.8 to AU$9. Properties in NSM swung less from AU$2.8 to AU$7. The means for both room types were very close to their medians. Mean and median for properties of in IM overlapped, which indicates the distribution of Inprice should be normal and symmetric. Relatively, the mean is higher than the median so the distribution of Inprice for properties in NSM is slightly skewed right.

For bivariate analysis, a scatterplot matrix was built for each variable against the dependent variable Inprice to observe the patterns displayed and the relationship between all independent variables within a single matrix. Dummy variables are excluded from the matrix for two reasons: first, there were too many dummy variables in my sample. Inclusive of all dummy variables would make all plots squeeze together within a small graph and made it difficult to observe the pattern, Secondly, dummy variables are qualitative variables and their points would just scatter along 0 or 1. Their scatterplots are not appropriate or meaningful to check for association. For interaction variable HRT_superhost in figure F.5, it is a product of two dummy variables dHRT1 and dHIS and hence it is also not meaningful to check for association as well.

With reference to figure F.5, the scatterplot showed positive linear relationship between Inprice and quantitative variables host_total_listings_count, accommodates, security_deposit, cleaning_fee and review_scores_rating. For variable *accommodates*, its scatterplot indicated a fairly strong association because most points followed a clear form. There could be outliers at the top left corner. Variable cleaning_fee has a slightly weak association because most points followed some forms, but some spread toward the center. There could be outliers at the top left corner. Scatterplots for host_total_listings_count, security_deposit and review_scores_rating indicated that they had a very weak relationship with Inprice. Most points spread out the graphs.

At the end of data exploration stage, a full model is fitted through linear regression. The full model statement is as follows:

lnprice = 4.628 – 0.0001*host_total_listings_count + 0.126*accommodates + 0.0001*security_deposit + 0.0005*cleaning_fee – 0.002*review_scores_rating – 0.038*dHRT1 + 0.04*dHRT2 – 0.013*dHRT3 + 0.153*dHRT4 + 0.011*dHIS – 0.548*dRT1 – 1.012*dRT2 – 0.162*dBT1 – 0.237*dBT3 – 0.206*dBT4 + 0.003*dCP1 + 0.001*dCP2 – 0.236*dRG1 – 0.053*dRG2 – 0.006*dRG3 – 0.268*dRG4 + 0.214*HRT_superhost + e

where dHRT1=1 when host_response_time='within an hour',
dHRT2=1 when host_response_time='within a few hours',
dHRT3=1 when host_response_time='within a day',
dHRT4=1 when host_response_time='a few days or more'
dHIS = 1 when host_is_superhost = “t”,
dRT1 = 1 when Room Type = “Private room”,
dRT2 = 1 when Room Type = “Shared room”,
dBT1=1 when bed_type = ‘Futon’,
dBT3=1 when bed_type = ‘Airbed’,
dB4=1 when bed_type = ‘Couch’
dRG1 = 1 when Region = “NSM”,
dRG2 = 1 when Region = “SEM”,
dRG4 = 1 when Region = “WM”,
HRT_superhost = dHRT1* dHIS
(dBT2 (Bed type = pull-out sofa) is set to 0 by SAS because of not enough observations)

According to the parameters estimate in figure F.6, beta weights indicates significant effect on lnprice. The higher the beta weight, the more significant effect on lnprice. However, t-test is a better methodology to measure significance of variables than the ranking of beta. Using the t-test on each model parameter, variables accommodates, security_deposit, review_scores_rating, dRT1-2, dRG1-2 and dRG4 had significant influence on lnprice (p-values for t-test are smaller than 0.05). On the other hand, the p-values for the rest of the independent variables, such as host_total_listings_count, cleaning_fee, dHRT1-4, dHIS, dBT1-4, dCP1-2, dRG3 and interaction variables were larger than 0.05. Therefore, we cannot reject the hypothesis that these variables have no effect on lnprice, which should be removed from the model.

F-Test Hypotheses:
H_o: $\beta_1 = \beta_2 = \beta_3 = \ldots = \beta_k = 0$
H_a: At least one coefficient $\beta_j \neq 0$

F=131.84 with p-value smaller than 0.05 (at alpha=0.05). The null hypothesis of no association between lnprice and other variables is rejected. We accept the alternative hypothesis that at least one coefficient of the independent variables has significant effect on lnprice. F-test gives strong support to the all variables.

The coefficients of determination of R^2 (54.06%) and adj-R^2 (53.65%) represent the amount of variation in lnprice explained by the regression model. For this analysis, about 53% of the variable in lnprice is explained by the model. To check if the model is good, we should look at adj R^2 of 53.65% because it does not increase with the addition of an independent variable that does not improve the regression model.

Next, the full model is taken to fit diagnostics to analyze if the model violates the 4 model assumptions. Referring to figure F.7 residual plots for each variable, predicted value against lnprice and figure F.8 normality graph. Residual plots for dummy variables and the interaction term are qualitative variable and their points are scattering along 0 or 1, which are not appropriate or meaningful for residual analysis and its assumptions. From figure F.7, the spread of the seven plots (host_total_listings_count, accommodates, security_deposit, cleaning_fee and review_scores_rating) are randomly scattered around the zero line, showing constant variance and independence. Even though there are some clusters in host_total_listings_count, , security_deposit and cleaning_fee towards these variable amounts are close to zero , these plots are considered to be normal because security_deposit and cleaning_fee usually will not be very expensive and within a small range and host_total_listings_count also will not be very high because the dataset only consists of 8-year data. For accommodates, it is normal as well because properties are typically small to accommodate less than 5 people. Review_scores_rating is normal because the rating is a subjective preference. The plot for predicted value is random and scattering along the zero line. The plots of host_total_listings_count, accommodates, security_deposit, cleaning_fee and review_scores_rating are linear because the pattern of the spread shows a straight line. Figure F.8 shows almost 45-degree line, which indicates the model is normally distributed and linear.

In figure F.9, severe multicollinearity will be detected if an independent variable has a correlation value more than absolute 0.9 with another for independent variable. In the model, the data does not seem to have this issue. After checking for multicollinearity, I check for outliers and influential points. In figure F.7, we could see some outliers in each independent variable residual plot, where outliers are beyond the +3/-3 bands. By referring to Studentized Residual and Cook's D (figure F.10) for lnprice, observations with arrowhead are indicated as both outliers and influential points, which are needed to removed first. After removing #413, regression is rerun again to check for the next observation to remove until there is no improvement by removing observations from the model. In total, I removed 5 observations from the model and record the changes in R^2 and adj-R^2 (figure F.11). It is realized that both R^2 (0.5569) and adj-R^2 (0.5529) are not high, meaning 55% of variables of predicted Inprice can be explained using the sample predictors. I stop removing observations because there is no further significant improvement to the model.

Following the outlier and influential point removal steps and checking for severe multicollinearity, the dataset is split into train set and test set with a ratio of 75% to 25% respectively. The train set is used to build the final model, while the test set is used to validate the model. The dependent variable is now called new_y of the train set. Figure F.12 shows the result after the split. The sample size of the train set now has 1865 observations. The train set is then taken to fit the final model by comparing two model selection methods: forward and backward methods. According to figure F.13, forward method selected 14 variables to the final model, while backward method only selected 9 variables. Looking into the summary of forward selection, the partial R^2 becomes very small after the 9th variable (dRG2) is added to the model. Comparing the selected variable from backward method and the first 9 variables from forward method, they are actually the same, i.e. accommodates, security_deposit, review_scores_rating, dHRT4, dRT1-2, dRG1-2, 4. The key to select the "best" variables/ model is to select fewer variables and a higher R^2. In this case, my final model factors in the variables that are suggested by backward selection method, except for security_deposit and review_scores_rating. I decided to exclude these two variables from my final model because the parameter estimates were so small that they are believed to have no effect on new_y. My final model is taken to regression again.

According to the parameters estimate (figure F.14), 6 variables are with p-values for t-test are smaller than 0.05, which means that variables accommodates, dHRT4, dRT1, dRT2, dRG1 and dRG4 have significant influence on new_y. However, dRG2 now becomes insignificant having p-value (0.0657) of t-test larger than 0.05 and it should be removed from the model. Then regression should be rerun. In figure F.15, all 6 variables are now with p-values for t-test are smaller than 0.05. Then I can conclude my final model statement after model selection as follows:

*new_y = 4.50 + 0.14*accommodates + 0.22*dHRT4 – 0.55*dRT1 – dRT2 – 0.22*dRG1 – 0.30*dRG4 + e*

where dHRT4= 1 when host_response_time = "a few days or more",

dRT1 = 1 when Room Type = "Private room",

dRT2 = 1 when Room Type = "Shared room",

dRG1 = 1 when Region = “NSM”,
dRG4 = 1 when Region = “WM”

Since Airbnb listing prices was transformed with log in the beginning, I have to retransform independent variable new_y in order to interpret the final model statement. Variable accommodates is positively associated to new_y. Model shows that assuming all other variables constant, for any additional guests that the property accommodates, predicted price for a night increases by 15.02% computed as $100*(e^{0.14}-1) = 100*(1.1502-1) = 15.02$.

The two dummy variables for host response time show that expected price for a night varies depending on the host response time to guests’ inquiries. Thus, on average, predicted price for a property whose host responds to inquiries within a few days or more is $100*(e^{0.22}-1)$ = 24.61% higher than predicted price for a night for a property whose host does not respond.

The parameter estimates for the two dummy variables for room type show that expected price for a night varies depending on the room types. Thus, on average, predicted price for a property with a private room is $(100*(e^{-0.55}-1) = \quad -42.31\%)$ 42.31% lower than predicted price for a night for a property with the entire house/ apartment; and predicted price for a property with a shared room is $(100*(e^{-1}-1) = -63.21\%)$ 63.21% lower than for a property with the entire house/ apartment.

Similarly, for the two dummy variables for regions show that expected price for a night varies depending on the regions. Thus, on average, predicted price for a property in region NSM is $(100*(e^{-0.22}-1) = -19.75\%)$ 19.75% lower than predicted price for a night for a property in IM; and predicted price in WM is $(100*(e^{0.3}-1) = -25.92\%)$ 25.92% lower than for a property in IM.

One way to analyze the strongest or most influential variable is to refer to standardized estimate. Since the beta is normalized by standard deviation of the dependent variable, all coefficients will have the same until of measurement. Therefore, the values of the standardized coefficients can be compared. It indicates that dRT1 is the most influential variable as it has the highest absolute standardized coefficient of 0.55.

F-Test Hypotheses:
H_o: $\beta_1 = \beta_2 = \beta_3 = \ldots = \beta_k = 0$
H_a: At least one coefficient $\beta_j \neq 0$

F=350.49 with p-value less than 0.05 (at alpha=0.05). The null hypothesis of no association between new_y and other variables is rejected. We accept the alternative hypothesis that at least one coefficient of the independent variables has significant effect on predicted new_y. F-test gives strong support to 6 variables.

The coefficients of determination of R^2 (53.12%) and adj-R^2 (52.97%) represent the amount of variation in new_y explained by the regression model. For this analysis, about 53% of the variable in Inprice is explained by the model. To check if the model is good, we should look at adj-R^2 of 52.97% because it does not increase with the addition of an independent variable that does not improve the regression model. To conclude, adj-R^2 of an average rate has indicated that this is an average model.

The next step is to analyze if model built by train set has satisfied the model assumptions. Again, we can ignore scatterplots for dummy variables. The predicted value plot does not show a strong form of pattern. Like full model assumption analysis, the spread of residual plot between new_y and accommodates are randomly scattered around the zero line and shows linearity because the pattern of the spread follows a straight line. The normality graph shows almost 45-degree line, which indicates the model is normally distributed and linear. Then we should check if the train set has severe multicollinearity. In figure F.15, VIF of all variables are less than 10, which means severe multicollinearity does not seem to be a problem.

After checking for multicollinearity, I continue checking for outliers and influential points. Based on the residual plot of new_y for train set, we could see some outliers in each independent variable residual plot (figure F.16), where outliers are beyond the +3/-3 bands. By referring to Studentized Residual and Cook's D for new_y (figure F.18), observations with arrowhead are indicated as both outliers and influential points, which are needed to removed first. After removing #519, regression is rerun again to check for the next observation to remove until there is no improvement by removing observations from the model. In total, I removed 4 observations from the model and record the changes in R^2 and adj-R^2 (figure F.19). It is realized that both R^2 (0.5569) and adj-R^2 (0.5529) are not high, meaning 55% of variables of predicted new_y can be explained using the sample predictors. I stop removing observations because there is no further significant improvement to the model.

Following testing regression result of the train set, I then proceed to measure predictive performance of the final model using test set. By referring to figure F.20, it shows the model of the test set is a better case because CV-R^2 is 0.2 (less than 0.3), RMSE is smaller and both R^2 and adj-R^2 is 0.2 higher than train set. Therefore, the final model by train set is said to be validated.

After validating the final model, I run regression again to evaluate the model performance with 5 indicators, check if it satisfied model assumptions, has severe multicollinearity and more outliers or influential points needed to be removed. By referring to figure F.21, all independent variables are with p-values for t-test are smaller than 0.05. This means that variables accommodates, dHRT4, dRT1, dRT2, dRG1 and dRG4 have significant influence on new_y.

F-Test Hypotheses:
H_o: $\beta_1 = \beta_2 = \beta_3 = \ldots = \beta_k = 0$
H_a: At least one coefficient $\beta_j \neq 0$

F=371.7 with p-value less than 0.05 (at alpha=0.05). The null hypothesis of no association between new_y and other variables is rejected. We accept the alternative hypothesis that at least one coefficient of the independent variables has significant effect on predicted new_y. F-test gives strong support to 6 variables. It still indicates that dRT1 is the most influential variable as it has the highest absolute standardized coefficient of 0.55.

The coefficient of determination of R2 (54.63%) and adj-R2 (54.48%) represent the amount of variation in new_y explained by the regression model. For this analysis, about 54% of the variable in new_y is explained by the model. To check if the model is good, we should look at adj-R2 of 54.48% because it does not increase with the addition of an independent variable that does not improve the regression model. To conclude, adj-R2 of an average rate has indicated that this is an average model.

The next step is to analyze if the final model has satisfied the model assumptions. Again, we can ignore scatterplots for dummy variables. In figure F.22, the predicted value plot does not show a strong form of pattern. Similar to the full model, the spread of residual plot between new_y and accommodates are considered to be randomly scattered around the zero line and shows linearity because the pattern of the spread follows a straight line. The normality graph shows almost 45-degree line, which indicates the model is normally distributed and linear. Then we should check if the train set has severe multicollinearity. In figure F.21, VIF of all variables are less than 10, which means severe multicollinearity does not seem to be a problem.

There are outliers that are captured in the residual plot between predicted value and new_y, especially when there is a data point to the top left of the plot. I checked the source and this property offers a private room in a townhouse and the stay should be around March 2018, which is the summer in Australia. Therefore, this outlier could be explained by the holiday season. I decided not to remove this outlier since this only point cannot affect the regression line much. After removing 4 observations from train set, my final model is restated as follows:

new_y = 4.50 + 0.14*accommodates + 0.22*dHRT4 – 0.55*dRT1 – 1.11dRT2 – 0.21*dRG1 – 0.31*dRG4 + e

where dHRT4= 1 when host_response_time = “a few days or more”,
dRT1 = 1 when Room Type = “Private room”,
dRT2 = 1 when Room Type = “Shared room”,
dRG1 = 1 when Region = “NSM”,
dRG4 = 1 when Region = “WM”

Since Airbnb listing prices was transformed with log in the beginning, I have to retransform independent variable new_y in order to interpret the final model statement. Variable accommodates is positively associated to new_y. Model shows that assuming all other variables constant, for any additional guests that the property accommodates, predicted price for a night increases by 15.02% computed as 100*($e^{0.14}$-1) = 100*(1.1502-1) = 15.02.

The two dummy variables for host response time show that expected price for a night varies depending on the host response time to guests’ inquiries. Thus, on average, predicted price for a property whose host responds to inquiries within a few days or more is 100*($e^{0.22}$-1) = 24.61% higher than predicted price for a night for a property whose host does not respond.

The parameter estimates for the two dummy variables for room type show that expected price for a night varies depending on the room types. Thus, on average, predicted price for a property with a private room is *(100*(*$e^{-0.55}$*-1) = -42.31%)* 42.31% lower than predicted price for a night for a property with the entire house/ apartment; and predicted price for a property with a shared room is *(100*(*$e^{-1.11}$*-1) = -67.04%)* 67.04% lower than for a property with the entire house/ apartment.

Similarly, for the two dummy variables for regions show that expected price for a night varies depending on the regions. Thus, on average, predicted price for a property in region NSM is *(100*(*$e^{-0.21}$*-1) = -18.94%)* 18.94% lower than predicted price for a night for a property in IM; and predicted price in WM is *(100*(*$e^{-0.31}$*-1) = -26.66%)* 26.66% lower than for a property in IM.

Using the final model to predict the average Airbnb listing prices, the condition is where the property accommodates 5 guests, host responds to inquiries within a few days or more, the room type is with a private room and the property is in WM. The model predicts average price per night ($e^{4.55}$) = AU$9,363. The predicted average price is within the 95% confidence interval between AU$7,648 ($e^{4.35}$) and AU$11,458 ($e^{4.75}$). Therefore, the final model is said to be a good model.

Another condition is where the property accommodates 2 guests, host does not respond to inquiries, the room type is with entire house/ apartment and the property is in IM. The model predicts average price per night ($e^{3.45}$) = AU$3,050. The predicted average price is within the 95% confidence interval between AU$2,505 ($e^{3.26}$) and AU$3,747 ($e^{3.65}$). Therefore, the final model is said to be a good model.

Future Work

The study investigated key factors that affect Airbnb listing price and our analysis are based on 22,895 observations. From the analysis we have so far, it seems like the price of Airbnb listings have many determinants, such as different room type, bed type, location where the property is located, whether the host is a super host or not, etc., all of them are price determinants. We know that price is a vital topic and important factor when considering which listings to choose from. However, our finding suggests that none of the predictors have a high correlation with price, all predictors we examined so far have some sort of relationship with price, but evidence is still needed to determine the factors affecting the price. In the future, it is necessary to know the top factor affecting Airbnb's lodging price; something that we did not have the opportunity to explore include property amenities and reputation, which is something that should be investigated for a better understanding of the price.

The Pearson correlation value between all independent variables and price are relatively low, with the highest value of 0.65, meaning it is only moderately correlated with price. The regression model that we have been using is able to capture some critical characteristics of price. Listings that are located in Western Melbourne (WM) have the highest price, followed by Northern Suburbs Melbourne (NSM), and then, South Eastern Melbourne (SEM). Listings that are located in Western Melbourne (WM) has the lowest price, which we could consider as not having a significant effect on price. In addition, our analysis found out that majority of Airbnb listings have real bed, therefore, analyzing the price with different bed type may be unnecessary.

There are certain limitations to our study. Missing data appeared repetitively which makes our analysis difficult to proceed. For example, for variable "region," we have five recorded levels: IM (Inner Melbourne), WM (Western Melbourne), EM (Eastern Melbourne), SEM (South Eastern Melbourne), and NSM (Northern Suburbs Melbourne), and our analysis have shown that, within these five levels, three of them have proved to have a significant negative impact on the response variable, "price." This shows that the place where the Airbnb located has a significant impact on price; however, there are many missing data within the variable "region," and we do not know that if our final result will be affected when all the missing data has been assigned to a specific region.

Although we already got plenty of predictors for the dependent variable, listing price; there are still ways that we can explore in our future studies for a better understanding of listing price. For future study, we would like to add in more variables that the potential customers will likely to be interested in when searching for Airbnb. For instance, proximity to public transit or pets are allowed to bring. We believe getting such kind of data could benefit us by obtaining and analyzing customer preferences and better predicting the listing price of Airbnb in a given area.

References

Terry Rawnsley and Laura Schmahmann. 2018. *What impact does Airbnb have on the Sydney and Melbourne housing markets?* https://www.sgsep.com.au/publications/what-impact-does-airbnb-have-sydney-and-melbourne-housing-markets

Dominik Gut and Philipp Herrmann. 2015. *Sharing Means Caring? Hosts' Price Reaction to Rating Visibility*

https://aisel.aisnet.org/cgi/viewcontent.cgi?referer=&httpsredir=1&article=1053&context=ecis2015_rip

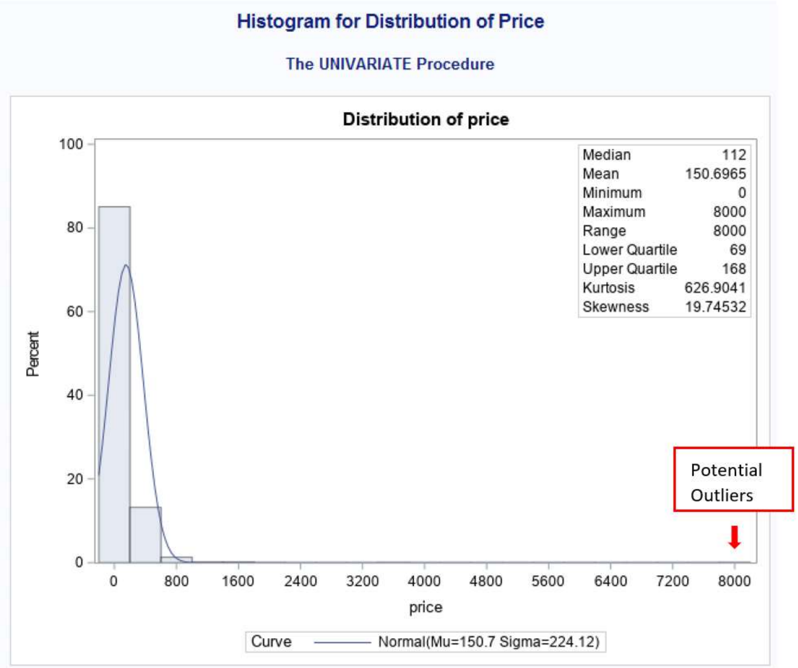
Scott Shatford. 2018. *What is Airbnb's Superhost Status Really Worth?* https://www.airdna.co/blog/airbnb_superhost_status

Leslie A. Christensen. *Introduction to Building a Linear Regression Model* The Goodyear Tire & Company, Akron Ohio

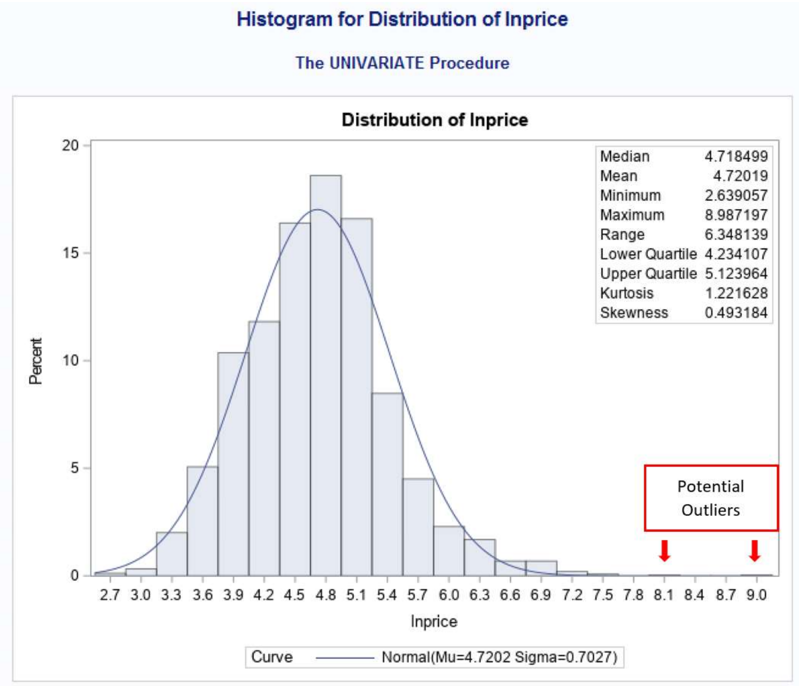
Andy Krause and Gideon Aschwanden. 2017. *A Census of Melbourne's Airbnb Market* https://cpb-ap-se2.wpmucdn.com/blogs.unimelb.edu.au/dist/b/193/files/2017/03/censusAnalysis-2hd2q7t.pdf

Bob. 2019. *Living in Melbourne* https://www.bobinoz.com/living-in-australia/melbourne

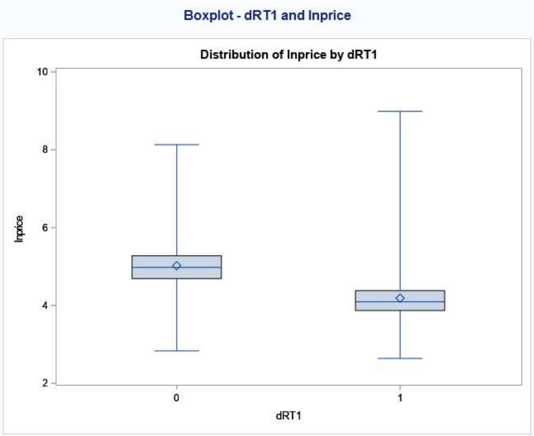
Appendix

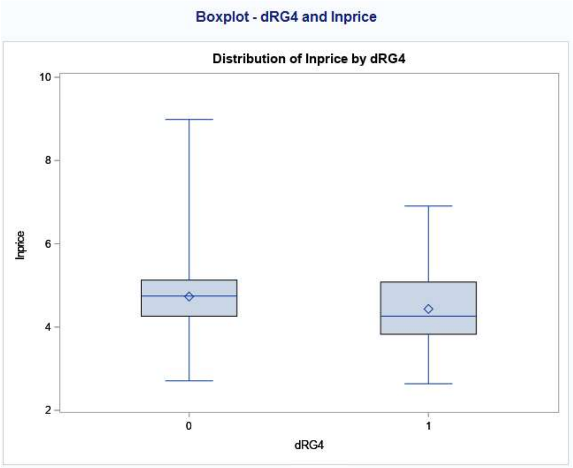


1. Histogram before log transformation



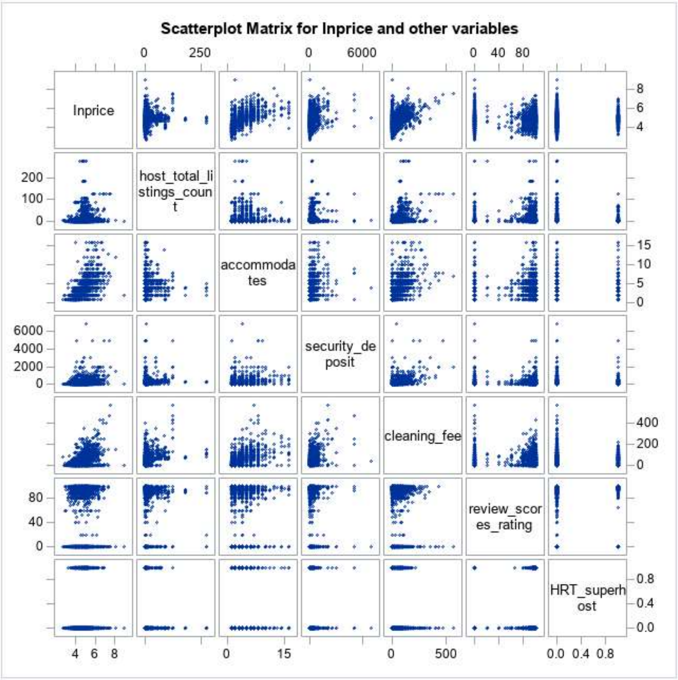
2. Histogram after log transformation



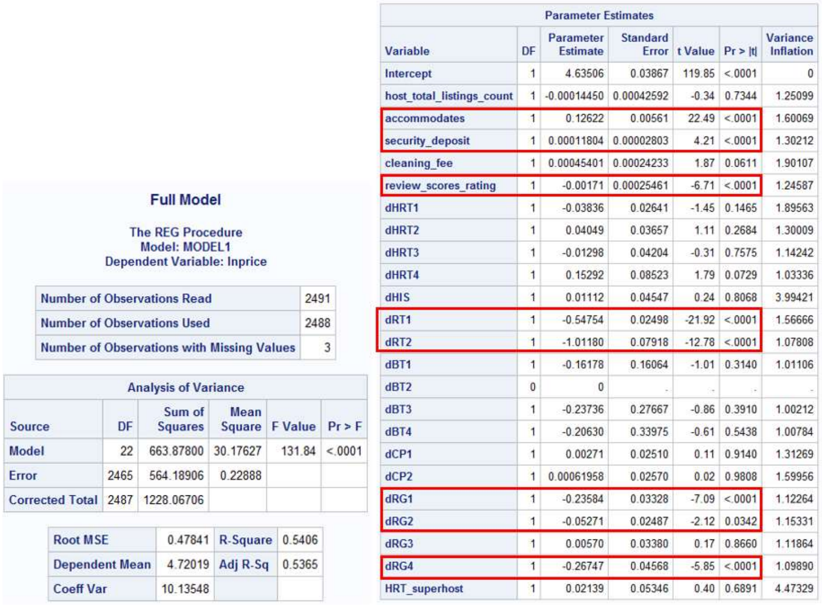


3. Boxplot – dRT1 and Inprice

4. Boxplot – dRG4 and Inprice



5. Scatterplot Mattrix for Inprice and other variables



Full Model

The REG Procedure
Model: MODEL1
Dependent Variable: lnprice

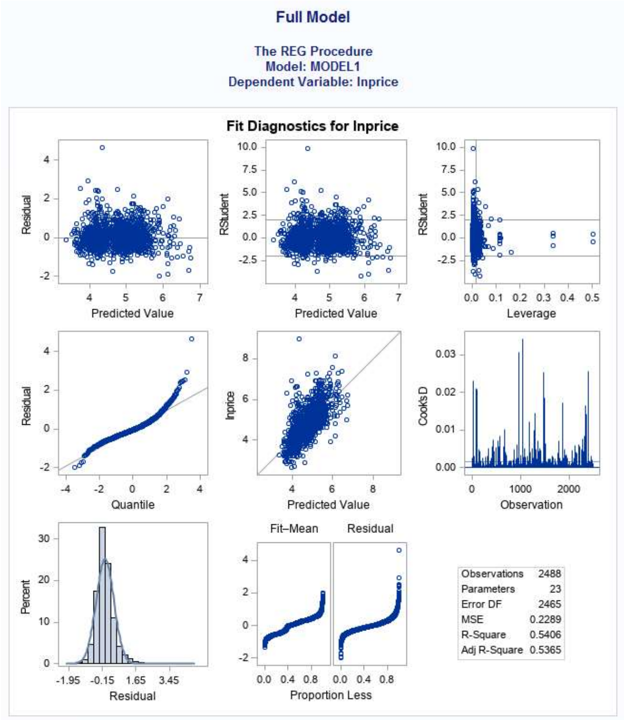
Number of Observations Read	2491
Number of Observations Used	2488
Number of Observations with Missing Values	3

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	22	663.87800	30.17627	131.84	<.0001
Error	2465	564.18906	0.22888		
Corrected Total	2487	1228.06706			

Root MSE	0.47841	R-Square	0.5406
Dependent Mean	4.72019	Adj R-Sq	0.5365
Coeff Var	10.13548		

Parameter Estimates						
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	1	4.63506	0.03867	119.85	<.0001	0
host_total_listings_count	1	-0.00014450	0.00042592	-0.34	0.7344	1.25099
accommodates	1	0.12622	0.00561	22.49	<.0001	1.60069
security_deposit	1	0.00011804	0.00002803	4.21	<.0001	1.30212
cleaning_fee	1	0.00045401	0.00024233	1.87	0.0611	1.90107
review_scores_rating	1	-0.00171	0.00025461	-6.71	<.0001	1.24587
dHRT1	1	-0.03836	0.02641	-1.45	0.1465	1.89563
dHRT2	1	0.04049	0.03657	1.11	0.2684	1.30009
dHRT3	1	-0.01298	0.04204	-0.31	0.7575	1.14242
dHRT4	1	0.15292	0.08523	1.79	0.0729	1.03336
dHIS	1	0.01112	0.04547	0.24	0.8068	3.99421
dRT1	1	-0.54754	0.02498	-21.92	<.0001	1.56666
dRT2	1	-1.01180	0.07918	-12.78	<.0001	1.07808
dBT1	1	-0.16178	0.16064	-1.01	0.3140	1.01106
dBT2	0	0	.	.	.	.
dBT3	1	-0.23736	0.27667	-0.86	0.3910	1.00212
dBT4	1	-0.20630	0.33975	-0.61	0.5438	1.00784
dCP1	1	0.00271	0.02510	0.11	0.9140	1.31269
dCP2	1	0.00061958	0.02570	0.02	0.9808	1.59956
dRG1	1	-0.23584	0.03328	-7.09	<.0001	1.12264
dRG2	1	-0.05271	0.02487	-2.12	0.0342	1.15331
dRG3	1	0.00570	0.03380	0.17	0.8660	1.11864
dRG4	1	-0.26747	0.04568	-5.85	<.0001	1.09890
HRT_superhost	1	0.02139	0.05346	0.40	0.6891	4.47329

6. Full Model Regression

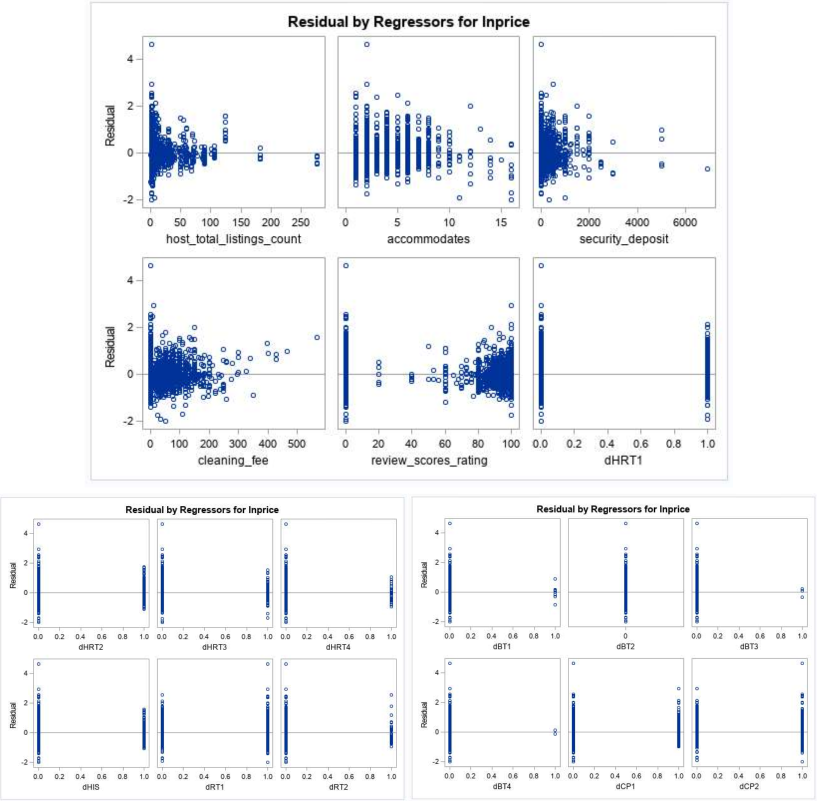


Full Model

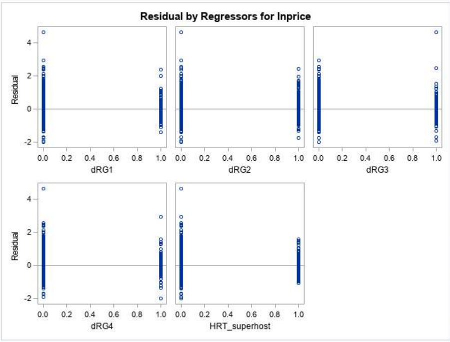
The REG Procedure
Model: MODEL1
Dependent Variable: lnprice

Fit Diagnostics for lnprice

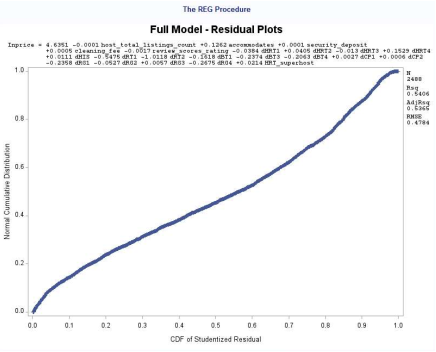
Observations	2488
Parameters	23
Error DF	2465
MSE	0.2289
R-Square	0.5406
Adj R-Square	0.5365



7. Residula plots for full model



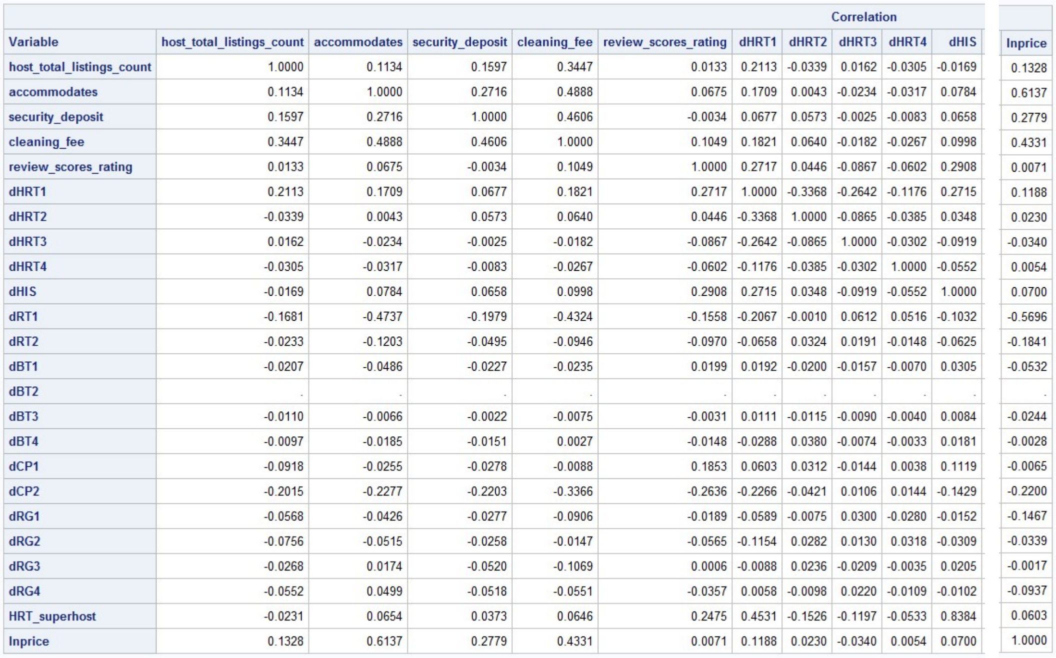
7. Residula plots for full model



```
lnprice = 4.6351 -0.0001 host_total_listings_count +0.1262 accommodates +0.0001 security_deposit
          +0.0005 cleaning_fee -0.0017 review_scores_rating -0.0384 dHRT1 +0.0405 dHRT2 -0.013 dHRT3 +0.1529 dHRT4
          +0.0111 dHIS -0.5475 dRT1 -1.0118 dRT2 -0.1618 dBT1 -0.2374 dBT3 -0.2063 dBT4 +0.0027 dCP1 +0.0006 dCP2
          -0.2358 dRG1 -0.0527 dRG2 +0.0057 dRG3 -0.2675 dRG4 +0.0214 HRT_superhost
```

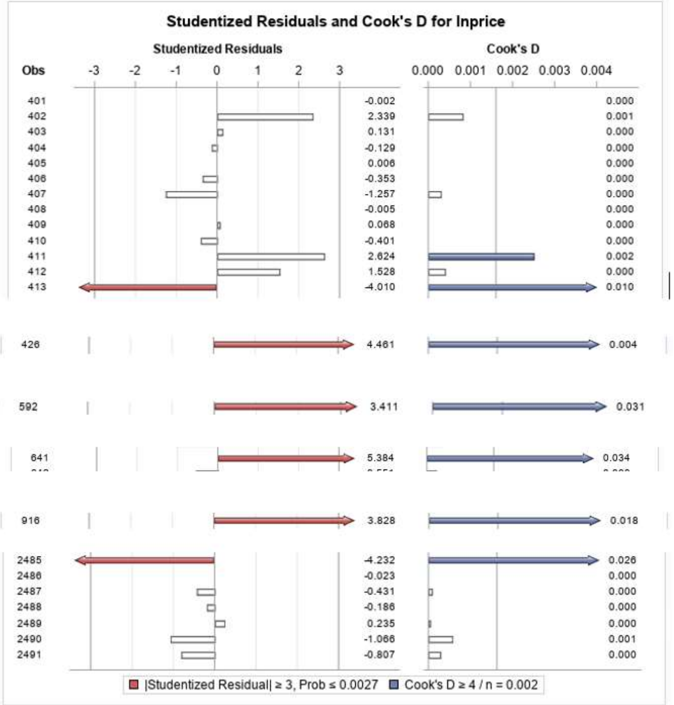
```
N
2488
Rsq
0.5406
AdjRsq
0.5365
RMSE
0.4784
```

8. Normality graph for full model



							Correlation				
Variable	host_total_listings_count	accommodates	security_deposit	cleaning_fee	review_scores_rating	dHRT1	dHRT2	dHRT3	dHRT4	dHIS	Inprice
host_total_listings_count	1.0000	0.1134	0.1597	0.3447	0.0133	0.2113	-0.0339	0.0162	-0.0305	-0.0169	0.1328
accommodates	0.1134	1.0000	0.2716	0.4888	0.0675	0.1709	0.0043	-0.0234	-0.0317	0.0784	0.6137
security_deposit	0.1597	0.2716	1.0000	0.4606	-0.0034	0.0677	0.0573	-0.0025	-0.0083	0.0658	0.2779
cleaning_fee	0.3447	0.4888	0.4606	1.0000	0.1049	0.1821	0.0640	-0.0182	-0.0267	0.0998	0.4331
review_scores_rating	0.0133	0.0675	-0.0034	0.1049	1.0000	0.2717	0.0446	-0.0867	-0.0602	0.2908	0.0071
dHRT1	0.2113	0.1709	0.0677	0.1821	0.2717	1.0000	-0.3368	-0.2642	-0.1176	0.2715	0.1188
dHRT2	-0.0339	0.0043	0.0573	0.0640	0.0446	-0.3368	1.0000	-0.0865	-0.0385	0.0348	0.0230
dHRT3	0.0162	-0.0234	-0.0025	-0.0182	-0.0867	-0.2642	-0.0865	1.0000	-0.0302	-0.0919	-0.0340
dHRT4	-0.0305	-0.0317	-0.0083	-0.0267	-0.0602	-0.1176	-0.0385	-0.0302	1.0000	-0.0552	0.0054
dHIS	-0.0169	0.0784	0.0658	0.0998	0.2908	0.2715	0.0348	-0.0919	-0.0552	1.0000	0.0700
dRT1	-0.1681	-0.4737	-0.1979	-0.4324	-0.1558	-0.2067	-0.0010	0.0612	0.0516	-0.1032	-0.5696
dRT2	-0.0233	-0.1203	-0.0495	-0.0946	-0.0970	-0.0658	0.0324	0.0191	-0.0148	-0.0625	-0.1841
dBT1	-0.0207	-0.0486	-0.0227	-0.0235	0.0199	0.0192	-0.0200	-0.0157	-0.0070	0.0305	-0.0532
dBT2	.	.	.	.	.	.	.	.	.	.	.
dBT3	-0.0110	-0.0066	-0.0022	-0.0075	-0.0031	0.0111	-0.0115	-0.0090	-0.0040	0.0084	-0.0244
dBT4	-0.0097	-0.0185	-0.0151	0.0027	-0.0148	-0.0288	0.0380	-0.0074	-0.0033	0.0181	-0.0028
dCP1	-0.0918	-0.0255	-0.0278	-0.0088	0.1853	0.0603	0.0312	-0.0144	0.0038	0.1119	-0.0065
dCP2	-0.2015	-0.2277	-0.2203	-0.3366	-0.2636	-0.2266	-0.0421	0.0106	0.0144	-0.1429	-0.2200
dRG1	-0.0568	-0.0426	-0.0277	-0.0906	-0.0189	-0.0589	-0.0075	0.0300	-0.0280	-0.0152	-0.1467
dRG2	-0.0756	-0.0515	-0.0258	-0.0147	-0.0565	-0.1154	0.0282	0.0130	0.0318	-0.0309	-0.0339
dRG3	-0.0268	0.0174	-0.0520	-0.1069	0.0006	-0.0088	0.0236	-0.0209	-0.0035	0.0205	-0.0017
dRG4	-0.0552	0.0499	-0.0518	-0.0551	-0.0357	0.0058	-0.0098	0.0220	-0.0109	-0.0102	-0.0937
HRT_superhost	-0.0231	0.0654	0.0373	0.0646	0.2475	0.4531	-0.1526	-0.1197	-0.0533	0.8384	0.0603
Inprice	0.1328	0.6137	0.2779	0.4331	0.0071	0.1188	0.0230	-0.0340	0.0054	0.0700	1.0000

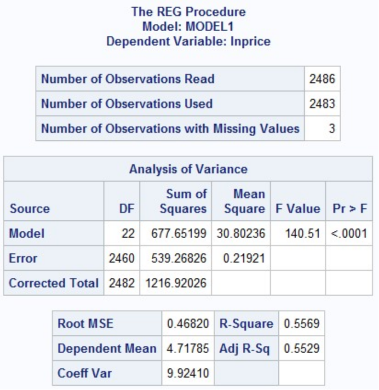
9. Correlation for Inprice and other variables



10. Full Model Extract of Studentized Residual and Cook’s D

Full Model

	n	Type	Current#	R Sq		adj R Sq	
	2491			0.5406		0.5365	
1st	2490	O & IF	413	0.5435	↑	0.5394	↑
2nd	2489	O & IF	2419	0.5489	↑	0.5448	↑
3rd	2488	O & IF	2009	0.5509	↑	0.5468	↑
4th	2487	O & IF	1586	0.5545	↑	0.5505	↑
5th	2486	O & IF	915	0.5569	↑	0.5529	↑



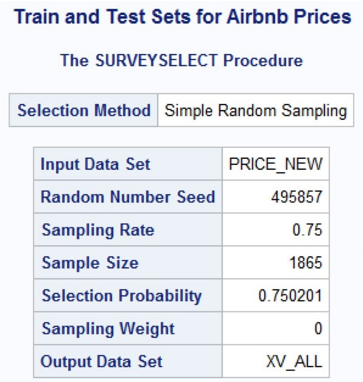
The REG Procedure
Model: MODEL1
Dependent Variable: lnprice

Number of Observations Read	2486
Number of Observations Used	2483
Number of Observations with Missing Values	3

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	22	677.65199	30.80236	140.51	<.0001
Error	2460	539.26826	0.21921		
Corrected Total	2482	1216.92026			

Root MSE	0.46820	R-Square	0.5569
Dependent Mean	4.71785	Adj R-Sq	0.5529
Coeff Var	9.92410		

11. Outliers and Influential Points removal record and final regression



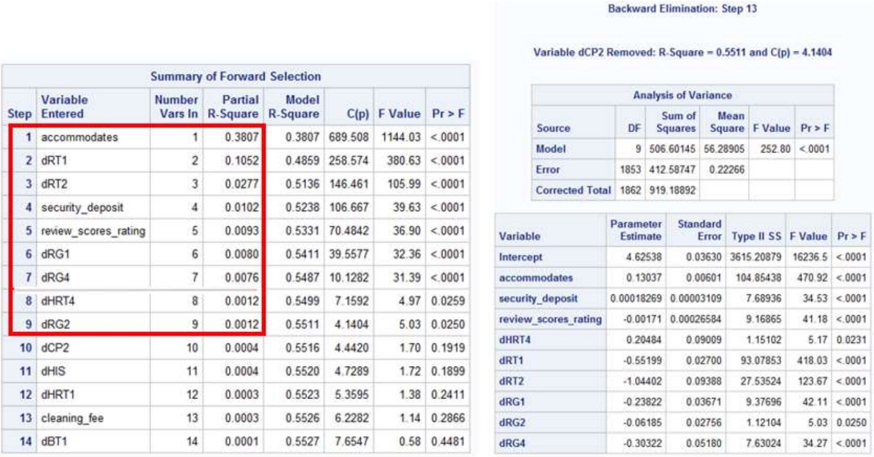
Train and Test Sets for Airbnb Prices

The SURVEYSELECT Procedure

Selection Method	Simple Random Sampling

Input Data Set	PRICE_NEW
Random Number Seed	495857
Sampling Rate	0.75
Sample Size	1865
Selection Probability	0.750201
Sampling Weight	0
Output Data Set	XV_ALL

12. Splitting of Train set and Test set



Summary of Forward Selection							
Step	Variable Entered	Number Vars In	Partial R-Square	Model R-Square	C(p)	F Value	Pr > F
1	accommodates	1	0.3807	0.3807	689.508	1144.03	<.0001
2	dRT1	2	0.1052	0.4859	258.574	380.63	<.0001
3	dRT2	3	0.0277	0.5136	146.461	105.99	<.0001
4	security_deposit	4	0.0102	0.5238	106.667	39.63	<.0001
5	review_scores_rating	5	0.0093	0.5331	70.4842	36.90	<.0001
6	dRG1	6	0.0080	0.5411	39.5577	32.36	<.0001
7	dRG4	7	0.0076	0.5487	10.1282	31.39	<.0001
8	dHRT4	8	0.0012	0.5499	7.1592	4.97	0.0259
9	dRG2	9	0.0012	0.5511	4.1404	5.03	0.0250
10	dCP2	10	0.0004	0.5516	4.4420	1.70	0.1919
11	dHIS	11	0.0004	0.5520	4.7289	1.72	0.1899
12	dHRT1	12	0.0003	0.5523	5.3595	1.38	0.2411
13	cleaning_fee	13	0.0003	0.5526	6.2282	1.14	0.2866
14	dBT1	14	0.0001	0.5527	7.6547	0.58	0.4481

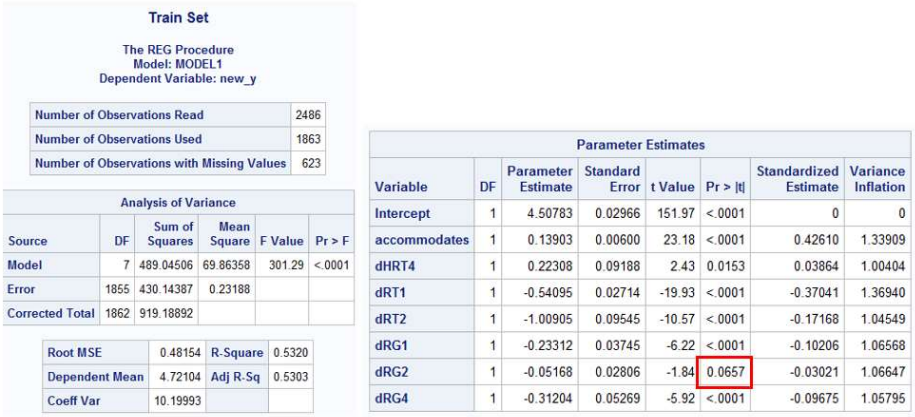
Backward Elimination: Step 13

Variable dCP2 Removed: R-Square = 0.5511 and C(p) = 4.1404

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	9	506.60145	56.28905	252.80	<.0001
Error	1853	412.58747	0.22266		
Corrected Total	1862	919.18892			

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	4.62538	0.03630	3615.20879	16236.5	<.0001
accommodates	0.13037	0.00601	104.85438	470.92	<.0001
security_deposit	0.00018269	0.00003109	7.68936	34.53	<.0001
review_scores_rating	-0.00171	0.00026584	9.16865	41.18	<.0001
dHRT4	0.20484	0.09009	1.15102	5.17	0.0231
dRT1	-0.55199	0.02700	93.07853	418.03	<.0001
dRT2	-1.04402	0.09388	27.53524	123.67	<.0001
dRG1	-0.23822	0.03671	9.37696	42.11	<.0001
dRG2	-0.06185	0.02756	1.12104	5.03	0.0250
dRG4	-0.30322	0.05180	7.63024	34.27	<.0001

13. Model Selections – Forward and Backward



Train Set

The REG Procedure
Model: MODEL1
Dependent Variable: new_y

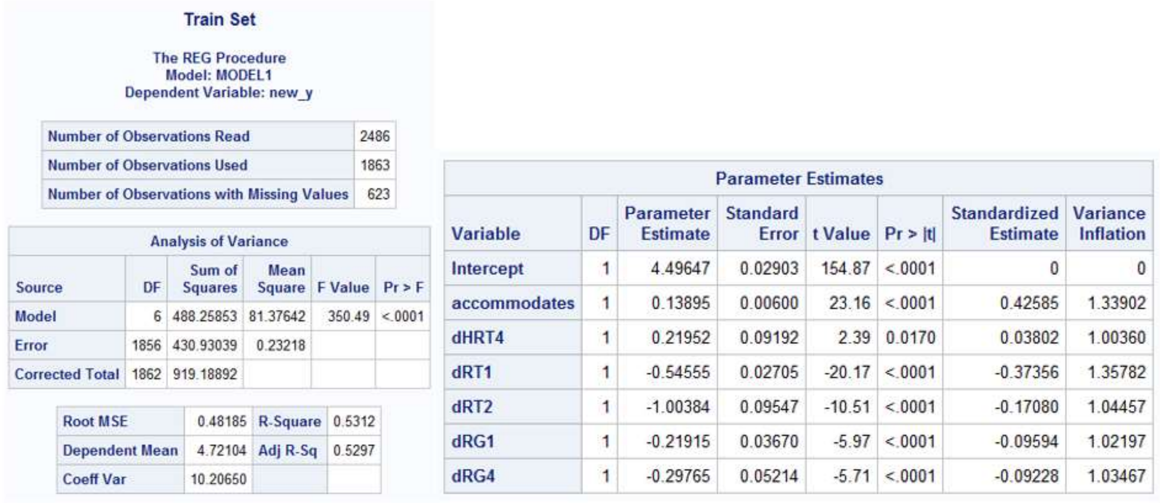
Number of Observations Read	2486
Number of Observations Used	1863
Number of Observations with Missing Values	623

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	7	489.04506	69.86358	301.29	<.0001
Error	1855	430.14387	0.23188		
Corrected Total	1862	919.18892			

Root MSE	0.48154	R-Square	0.5320
Dependent Mean	4.72104	Adj R-Sq	0.5303
Coeff Var	10.19993		

Parameter Estimates							
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Standardized Estimate	Variance Inflation
Intercept	1	4.50783	0.02966	151.97	<.0001	0	0
accommodates	1	0.13903	0.00600	23.18	<.0001	0.42610	1.33909
dHRT4	1	0.22308	0.09188	2.43	0.0153	0.03864	1.00404
dRT1	1	-0.54095	0.02714	-19.93	<.0001	-0.37041	1.36940
dRT2	1	-1.00905	0.09545	-10.57	<.0001	-0.17168	1.04549
dRG1	1	-0.23312	0.03745	-6.22	<.0001	-0.10206	1.06568
dRG2	1	-0.05168	0.02806	-1.84	0.0657	-0.03021	1.06647
dRG4	1	-0.31204	0.05269	-5.92	<.0001	-0.09675	1.05795

14. Regression for Train set



Train Set

The REG Procedure
Model: MODEL1
Dependent Variable: new_y

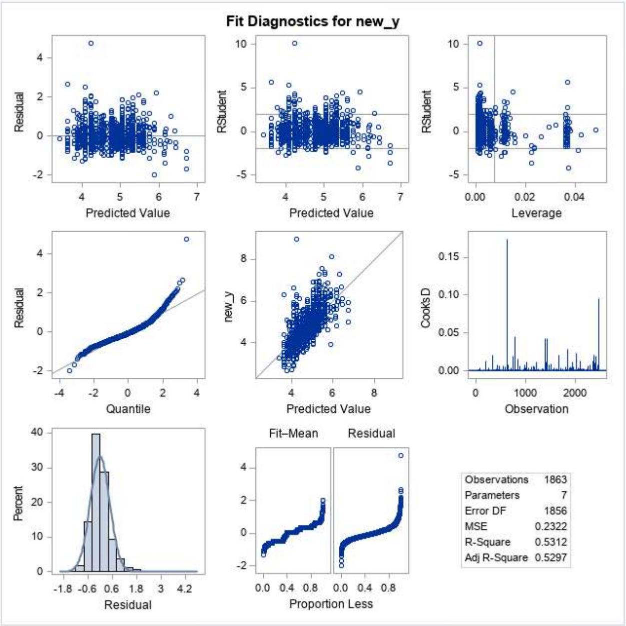
Number of Observations Read	2486
Number of Observations Used	1863
Number of Observations with Missing Values	623

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	6	488.25853	81.37642	350.49	<.0001
Error	1856	430.93039	0.23218		
Corrected Total	1862	919.18892			

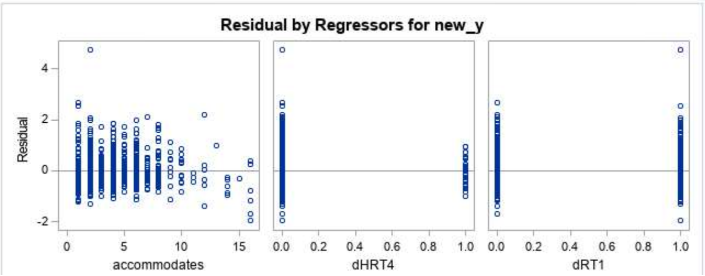
Root MSE	0.48185	R-Square	0.5312
Dependent Mean	4.72104	Adj R-Sq	0.5297
Coeff Var	10.20650		

Parameter Estimates							
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Standardized Estimate	Variance Inflation
Intercept	1	4.49647	0.02903	154.87	<.0001	0	0
accommodates	1	0.13895	0.00600	23.16	<.0001	0.42585	1.33902
dHRT4	1	0.21952	0.09192	2.39	0.0170	0.03802	1.00360
dRT1	1	-0.54555	0.02705	-20.17	<.0001	-0.37356	1.35782
dRT2	1	-1.00384	0.09547	-10.51	<.0001	-0.17080	1.04457
dRG1	1	-0.21915	0.03670	-5.97	<.0001	-0.09594	1.02197
dRG4	1	-0.29765	0.05214	-5.71	<.0001	-0.09228	1.03467

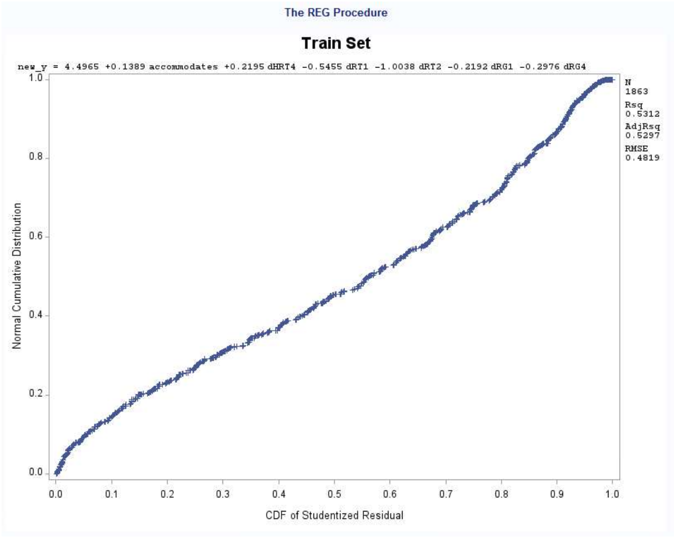
15. Regression for Train set (removed dRG2)



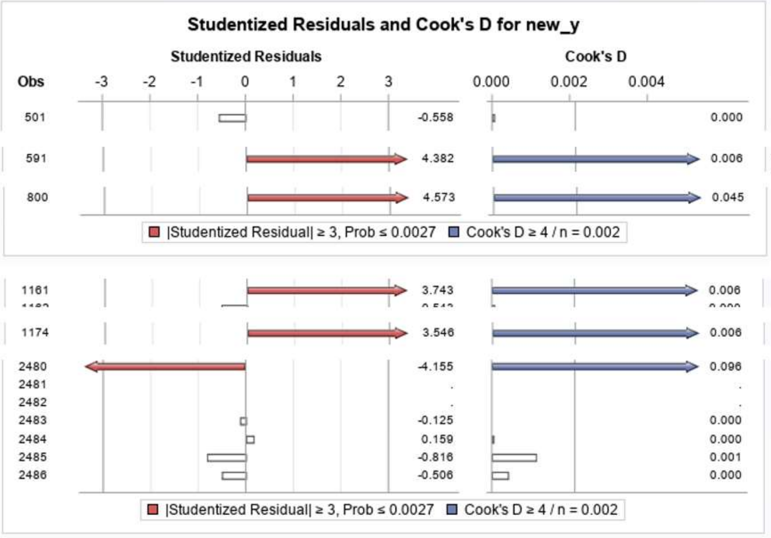
16. Residual Plots for Train set (removed dRG2)



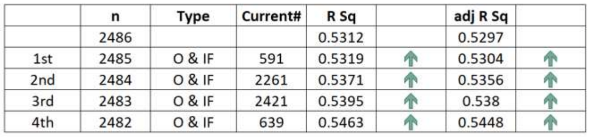
17. Residual Plots for Train set (removed dRG2)



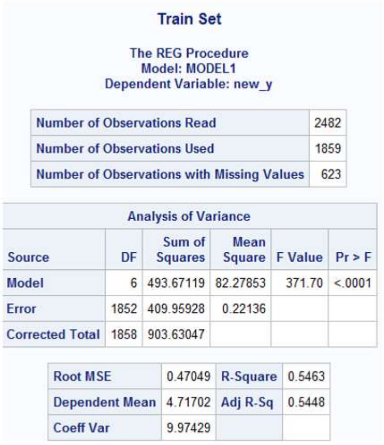
17. Final Model - Normality graph for Train set (removed dRG2)



18. Final Model Extract of Studentized Residual and Cook's D



	n	Type	Current#	R Sq		adj R Sq	
	2486			0.5312		0.5297	
1st	2485	O & IF	591	0.5319	↑	0.5304	↑
2nd	2484	O & IF	2261	0.5371	↑	0.5356	↑
3rd	2483	O & IF	2421	0.5395	↑	0.538	↑
4th	2482	O & IF	639	0.5463	↑	0.5448	↑



Train Set

The REG Procedure
Model: MODEL1
Dependent Variable: new_y

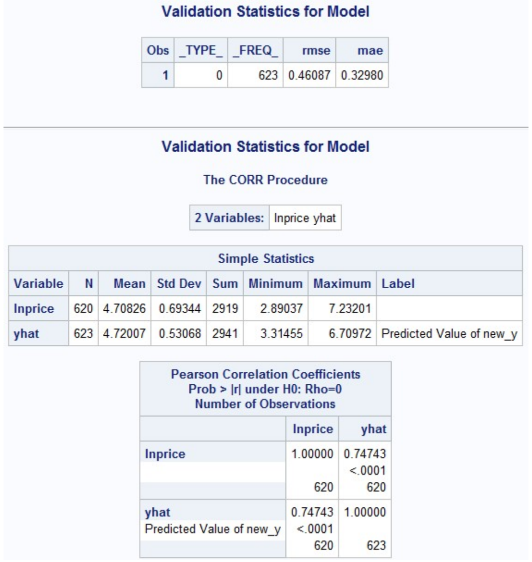
Number of Observations Read	2482
Number of Observations Used	1859
Number of Observations with Missing Values	623

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	6	493.67119	82.27853	371.70	<.0001
Error	1852	409.95928	0.22136		
Corrected Total	1858	903.63047			

Root MSE	0.47049	R-Square	0.5463
Dependent Mean	4.71702	Adj R-Sq	0.5448
Coeff Var	9.97429		

19. Outliers and Influential Points removal record and final regression



Validation Statistics for Model

Obs	_TYPE_	_FREQ_	rmse	mae
1	0	623	0.46087	0.32980

Validation Statistics for Model

The CORR Procedure

2 Variables:	lnprice yhat

Simple Statistics

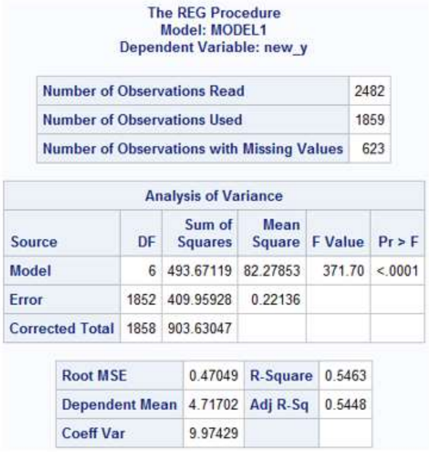
Variable	N	Mean	Std Dev	Sum	Minimum	Maximum	Label
lnprice	620	4.70826	0.69344	2919	2.89037	7.23201	
yhat	623	4.72007	0.53068	2941	3.31455	6.70972	Predicted Value of new_y

Pearson Correlation Coefficients
Prob > |r| under H0: Rho=0
Number of Observations

	lnprice	yhat
lnprice	1.00000 620	0.74743 <.0001 620
yhat Predicted Value of new_y	0.74743 <.0001 620	1.00000 623

Train	RMSE	0.47049
	R-square	0.5463
	Adj-R sq	0.5448
	GOF	OK
	Residuals	OK
Test	RMSE	0.46087
	MAE	0.3298
	R-square	0.74743
	Adj-R sq	0.7449699
	CV-R sq	0.20113

20. Test set vs Train set



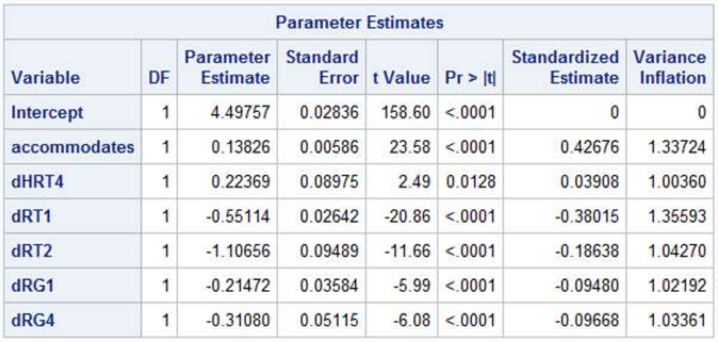
The REG Procedure
Model: MODEL1
Dependent Variable: new_y

Number of Observations Read	2482
Number of Observations Used	1859
Number of Observations with Missing Values	623

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	6	493.67119	82.27853	371.70	<.0001
Error	1852	409.95928	0.22136		
Corrected Total	1858	903.63047			

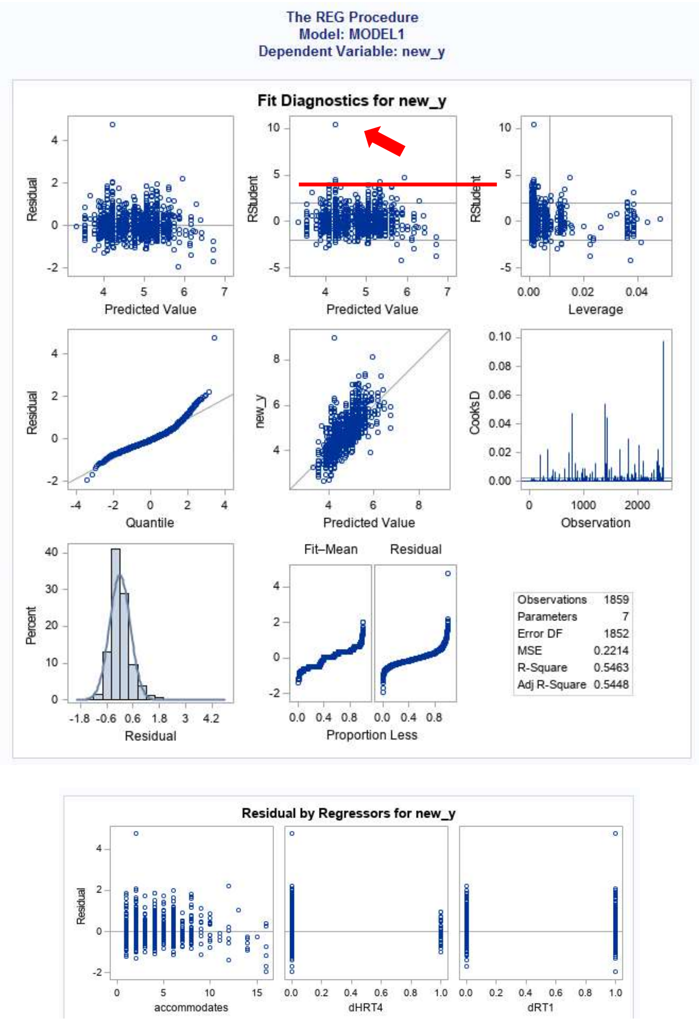
Root MSE	0.47049	R-Square	0.5463
Dependent Mean	4.71702	Adj R-Sq	0.5448
Coeff Var	9.97429		



Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Standardized Estimate	Variance Inflation
Intercept	1	4.49757	0.02836	158.60	<.0001	0	0
accommodates	1	0.13826	0.00586	23.58	<.0001	0.42676	1.33724
dHRT4	1	0.22369	0.08975	2.49	0.0128	0.03908	1.00360
dRT1	1	-0.55114	0.02642	-20.86	<.0001	-0.38015	1.35593
dRT2	1	-1.10656	0.09489	-11.66	<.0001	-0.18638	1.04270
dRG1	1	-0.21472	0.03584	-5.99	<.0001	-0.09480	1.02192
dRG4	1	-0.31080	0.05115	-6.08	<.0001	-0.09668	1.03361

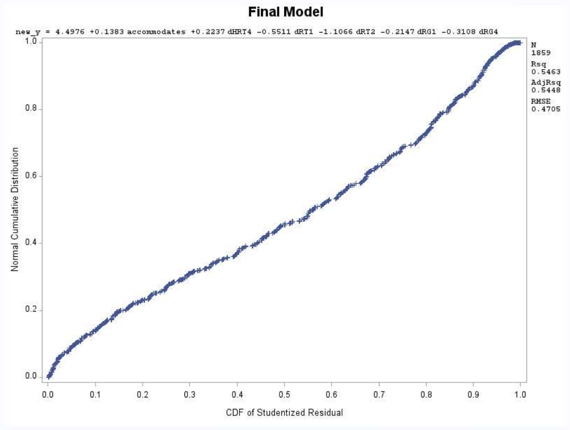
21. Final Model Regression



The REG Procedure
Model: MODEL1
Dependent Variable: new_y

Observations	1859
Parameters	7
Error DF	1852
MSE	0.2214
R-Square	0.5463
Adj R-Square	0.5448

22. Final Model Residual Plots

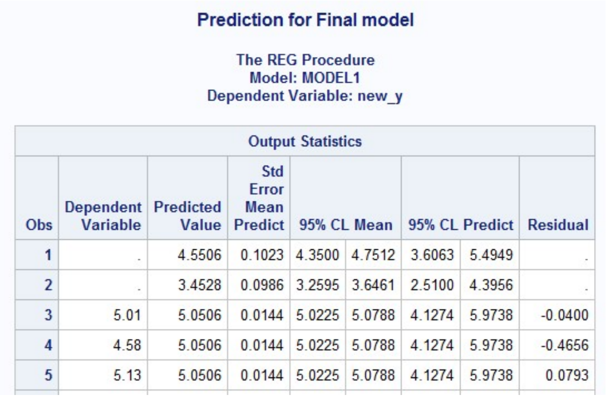


Final Model

new_y = 4.4976 +0.1383 accommodates +0.2237 dHRT4 -0.5511 dRT1 -1.1066 dRT2 -0.2147 dRG1 -0.3108 dRG4

N 1859
Rsq 0.5463
AdjRsq 0.5448
RMSE 0.4705

23. Final Model Normality Graph



Prediction for Final model

The REG Procedure
Model: MODEL1
Dependent Variable: new_y

Output Statistics

Obs	Dependent Variable	Predicted Value	Std Error Mean Predict	95% CL Mean		95% CL Predict		Residual
1	.	4.5506	0.1023	4.3500	4.7512	3.6063	5.4949	.
2	.	3.4528	0.0986	3.2595	3.6461	2.5100	4.3956	.
3	5.01	5.0506	0.0144	5.0225	5.0788	4.1274	5.9738	-0.0400
4	4.58	5.0506	0.0144	5.0225	5.0788	4.1274	5.9738	-0.4656
5	5.13	5.0506	0.0144	5.0225	5.0788	4.1274	5.9738	0.0793

24. Final Model Predictions