Colin Horgan

Ames Housing Data Analysis Project

The goal of this project is to predict the cost and age of a given house based on attributes contained within the Ames Housing Dataset, a repository of housing data from Canada.

\*the columns of independent variables shown in summary tables of multivariate regression are in the order they are listed in the exam i.e. 1 = TotalBasementSF, 2 = TotalLivingAreaSF, 3 = FullBath, 4 = HalfBath, 5 = BedroomAbvGround, 6 = TotalRoomsAbvGround, 7 = GarageCars, 8 = GarageSF, 9 = WoodDeckSF, 10 = TotalPorchSF \*

***Initial Multivariate Linear Regression Model***

The summary tables of the multivariate linear regression model are displayed in Figures 1 and 2, for SalePrice and Age respectively.

A close up of text on a white background

Description automatically generated

*Figure 1: Summary of output for multivariate regression for SalePrice*.

Table

Description automatically generated

*Figure 2: Summary of output for multivariate regression for Age*.

To make things as clear as possible, I am going to address the questions posed in The discussion of our initial model separately with respect to each dependent variable.

*Sale Price (Figure 1)*

The only two independent variables that did not significantly affect sale price in our model were TotalRoomsAbvGround and TotalPorchSF. While the intercept was significant for this model, it has a strong negative value indicating that our model will return a negative value of the home if all our independent variables are equal to zero. This does not make sense, and a better model hopefully would lack this feature. All of our independent variables positively predicted the sale price of the home (meaning as they increase, so does sale price) with the exception of BedroomsAbvGround. The coefficient of BedroomsAbvGround is negative indicating that it is inversely proportional to sale price. This is another feature of our model which does not make sense (surely bedrooms above ground are more desirable than those below). The model yields an R2 value of .7258, indicating that it explains 72.58 percent of the variance in our data. This is quite good but could surely be further improved by studying our data and transforming it where necessary.

*Age (Figure 2)*

Every independent variable was reported to be significant in our model, with respect to age. The intercept was estimated to be 76.83 meaning that if all our independent variables were zero, this is the predicted age of the house. I don’t see too much obviously wrong here as older homes tend to be smaller in square footage and the number of rooms/amenities. TotalBasementSF, FullBath, HalfBath, GarageCars, GarageSF, WoodDeckSF, all had negative coefficients in our model. As mentioned, this makes sense as older homes tend to be smaller and have less accommodations. TotalLivingAreaSF, BedroomAbvGround, TotalRoomsAbvGround, TotalPorchSF all had positive coefficients meaning as they increase, so does age. The only problem here I see is that as TotalLivingAreaSF increases, so does the age of our home. This runs counter to the common theme here that older homes are typically smaller. The R2 value is .5557, meaning we can account for 55.57% of the variance with our current model. This is certainly not very good, and so we should attempt to improve it significantly.

*Standardizing and Visualizing Data*

Examining our dependent variables SalePrice and Age reveals that both variables have some positive skew, but most importantly have drastically different variance (Figure 3). Histograms not shown to conserve space.



Figure 3: Mean and Standard deviation of SalePrice (X1) and Age (X2)

This finding necessitates a transformation, the best of which was found to be a ln transform the results of which can be found in Figure 4 (tried log10, Tukey, sqrt, and 1/x).



Figure 4: Mean and Standard deviation transformed of SalePrice (X1) and Age (X2)

Examining our dependent variables demonstrates similar positive skew compared to the independent variables, again to keep this question under 20 pages the histograms are not shown.

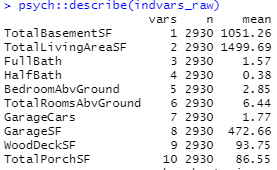


Figure 5: Mean and Standard deviation of raw independent variables

Again, we can see that our assumption of heterogeneity of variance is violated. Transforming this data using a ln-transform gives us the following distribution (Figure 5).

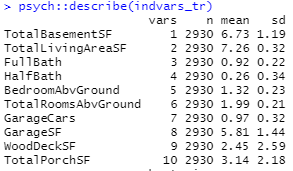
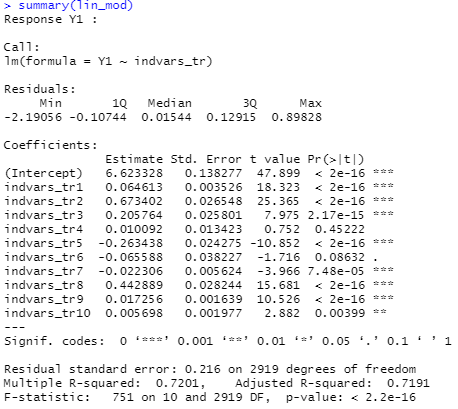
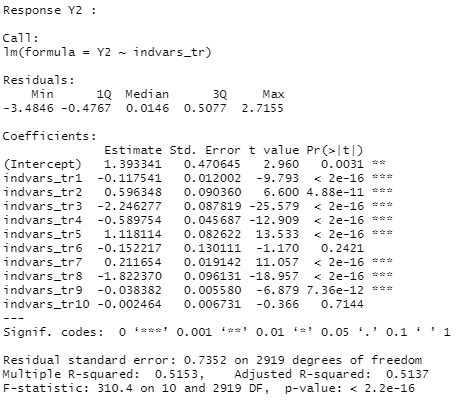


Figure 6: Mean and Standard deviation of transformed independent variables

Using these transformed variables, we can again run the multiple linear regression model to see if there are any improvements over the raw data. The results of this can be found in Figures 5 and 6. As before, I will address the results with respect to each independent variable individually.



*Figure 7: Summary of output for transformed multivariate regression for SalePrice*.



*Figure 8: Summary of output for multivariate regression for Age*.

*SalePrice (Figure 7)*

One significant improvement of this result vs. the raw data is that our intercept is much closer to 0. BedroomsAbvGround still gives a negative coefficient, which as discussed in the discussion of our initial model seems undesirable, but in this result TotalRoomsAbvGround also gives a negative coefficient. However, TotalRoomsAbvGround is not a significant predictor in this case (as in the discussion of our initial model) so this can be largely ignored. It is important to note that we observe a decrease in the R2 value - .6995, which is down from .7258 in the discussion of our initial model. I am not sure how to interpret this, but I do suspect that given more time to find a better way to transform the data I might improve this result.

*Age (Figure 8)*

The most significant change from part b in our model for Age is that TotalRoomsAbvGround and TotalPorchSF are no longer significant predictors of Age. Additionally, the coefficients for TotalLivingArea and BedroomAbvGround flipped sign from positive to negative. We see a decrease in our R2 value compared to our untransformed data. All of this is concerning given that this model is supposed to be an improvement over the raw data - but given the time I spend trying to standardize the data - I stand by the decision to keep the ln-transformed data instead of something more (like taking the inverse) or less (sqrt) extreme.

*Analysis of Residual Plots to Interpret Standardized Model.*

Scatterplots of our residuals vs. predicted values of both price and age reveal interesting findings about the fit of our model. Again, I will deal with price and age separately.

***SalePrice***

The graph of our residuals from SalePrice vs. our predicted SalePrice are shown in Figure 9. This scatterplot shows a quadratic trend around where PredictedPrice is equal to 12. This could be for a couple reasons. First, as mentioned before and as we will see in the analysis of residuals vs. predictors, while our ln-transformation worked for much of our data there are some parts of it which are still significantly skewed. This may be contributing to the quadratic trend in Figure 9. Additionally, we might be missing a feature or we could be seeing some correlation between the price of the house and the residuals.

Chart, scatter chart

Description automatically generated

*Figure 9: Residuals vs. predicted values for SalePrice*

***Age***

Our scatterplot of the residuals vs. predicted values of age harkens back to what I said about the intercept in part b. Our residuals look as though they are normally distributed about the mean, however there is a clear linear trend with positive slope as our predicted values for age increase. As I said in part b, the intercept (~6 years) for age seems very low, and if it took on a higher value we would see exactly what we want in our scatterplot i.e. no trend in the residuals.

Chart, scatter chart

Description automatically generated

*Figure 10: Residuals and predicted values for Age*

Now we move on to a discussion about the plots of residuals vs. our independent variables. There are 10 predictor variables in this model, and to keep this problem from being 25 pages long (and from having to say the same things over and over again) I will stick to describing the figures that I find most demonstrative of the relationship between the residuals and independent variables. You can find all the plots in the folder labled “residuals vs. independent variables”. Also, since these scatterplots contain histograms of the independent variables, I am going to count them as histograms of our transformed data as well. I hope that is ok.

***Residuals vs. Independent Variables***

The big takeaway from plotting our residuals for SalePrice and Age vs. our independent variables is that our data transformation did a really great job of standardizing data, unless the value of the data could be 0. TotalBasementSF, GarageCars, WoodPorchSF, and TotalPorchSF all suffered from being well standardized except at zero. For example, we can see that our data for TotalPorchSF is nicely distributed except for the buildup of values at zero (Figure 11). Despite our troubles with normalization, our residuals look evenly distributed around a mean of zero. The plots of residuals vs. independent variables that were standardized well look very good. These variables include FullBath, HalfBath, BedroomAboveGround, TotalRoomsAboveGround, GarageSF. It seems with predictors that can only take discrete values our residuals are well distributed about a mean of zero, of which BedroomAboveGround is a good example (Figure 12). One independent variable which does not fit any of the trends discussed so far is TotalLivingSF. For this variable, the residuals are centered around a mean of zero but are concentrated quite densely near it (Figure 13). It is not clear to me whether or not this is a problem – or why we observe this – but it seemed necessary to point it out.

Chart

Description automatically generated

*Figure 11: Residuals of SalePrice vs. TotalPorchSF*

*Chart, histogram

Description automatically generated*

*Figure 12: Residuals of Age vs. BedroomAbvGround*

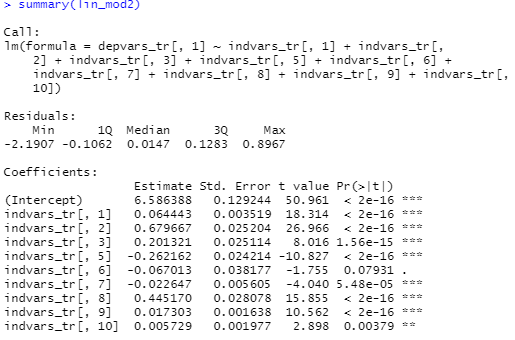
*Chart, histogram

Description automatically generated*

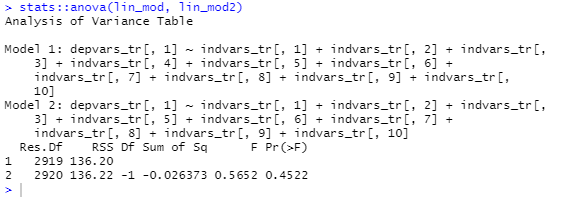
*Figure 13: Residuals of SalePrice vs. TotalLivingArea*

***Stepwise Multivariate Linear Regression***

Using stepwise regression for SalePrice, we first see that the only variable which lies above our inclusion threshold is HalfBath. After excluding this variable in our next model, we observe that no more variables lie outside our inclusion threshold so we stop here (not sure if that is right, but its what I got) (Figure 14). ANOVA of both our full and reduced model reveals no significant difference (p>.05) meaning the exclusion of the HalfBath variable did not affect the overall performance of the model (Figure 15).



*Figure 14: Coefficients of reduced linear model. Note that all of our p < .15*



*Figure 15: ANOVA of full vs. reduced linear model for SalePrice*