|  |
| --- |
|  |
| Evaluation of Housing Value Using Regression Analysis |
| Math 3330 Final Project |
|  |
| **Toan Cao-215230170**  **Khoa Le - 215316003**  **Alireza Shaghouzaei-215236912** |
| **11/22/2018** |

|  |
| --- |
|  |

**Introduction**

For a prospective homeowner, there are many factors that should be taken into consideration when searching for a new house. Properties ranging from the house’s appearance and structure to where it is located are all potential subjects of scrutiny in this hunt. Yet ultimately, the biggest concern for many Canadians when considering a new home is its price point. According to the UBS 2017 Global Real Estate summary, this could be felt in the housing bubble experienced by the Toronto and Vancouver regions from as far back as 2012, in which the value of the average home increased as much as 50% over five years, deterring the average house-hunter while perhaps being seen as a profitable opportunity for homeowners and realtors in these areas.

In this report we build a regression model for the expected price of a house based on its various properties. For this purpose, building class, lot area, overall quality and condition of the house, the year the house was constructed, basement and ground floor living area, and number of rooms above ground are all considered and analyzed from our data set (see references). In doing so, we hope to uncover the most important factors in predicting the value of any given house.

**Hypotheses**

From initial observations of the data set, we made the following hypotheses:

1. Newer houses will be more expensive than older houses (variable x5)
2. Houses put up for sale in good condition will be more expensive than those that are in bad condition (x3)

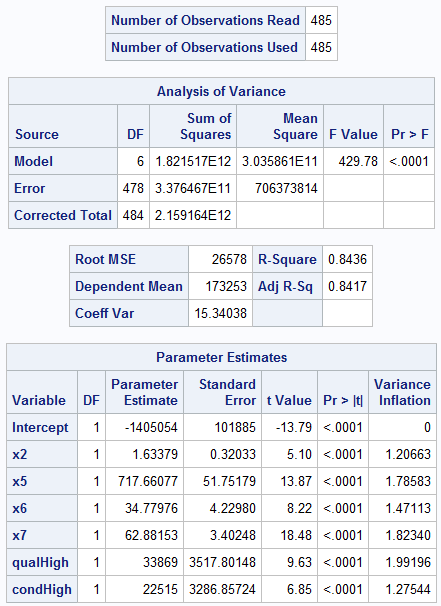
In other words, we expect that as x3 or x5 increases, so too does the dependent variable y (house price). To test these hypotheses, we use SAS to find the best fitting model and to analyze the data. It is worth noting however that, since they are not included in the data set, we will not be considering the importance of location in determining the value of a house in this model, and will be judging value simply on physical properties of the house itself.

**Methods**

To construct the initial models we used the stepwise, forward, and backward selection methods, using house price as the dependent variable. These selection methods were compared to select the potentially best fitting model. Residual analysis is conducted on these models to determine if the data contains influential outliers (removing significant outliers to achieve a higher R-value). To gauge the effect of the overall quality (x3) and the condition (x4) of the house, these variables were treated as dummy variables where a value from 0-3 was considered “low”, 4-6 was considered “moderate” and a value from 7-10 was considered “high”. We test these models for the multicollinearity problem, as well as the three assumptions of regression models: constant variance, independence, and normality.

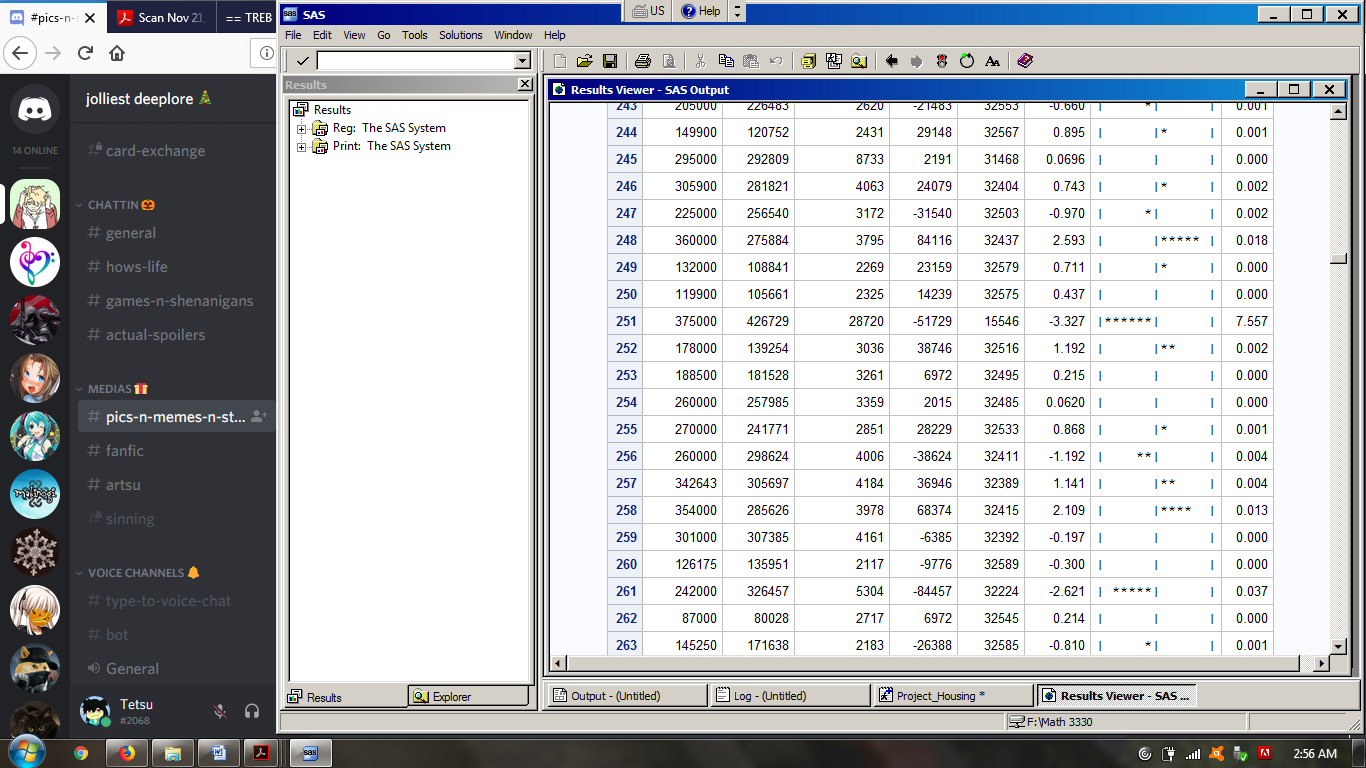
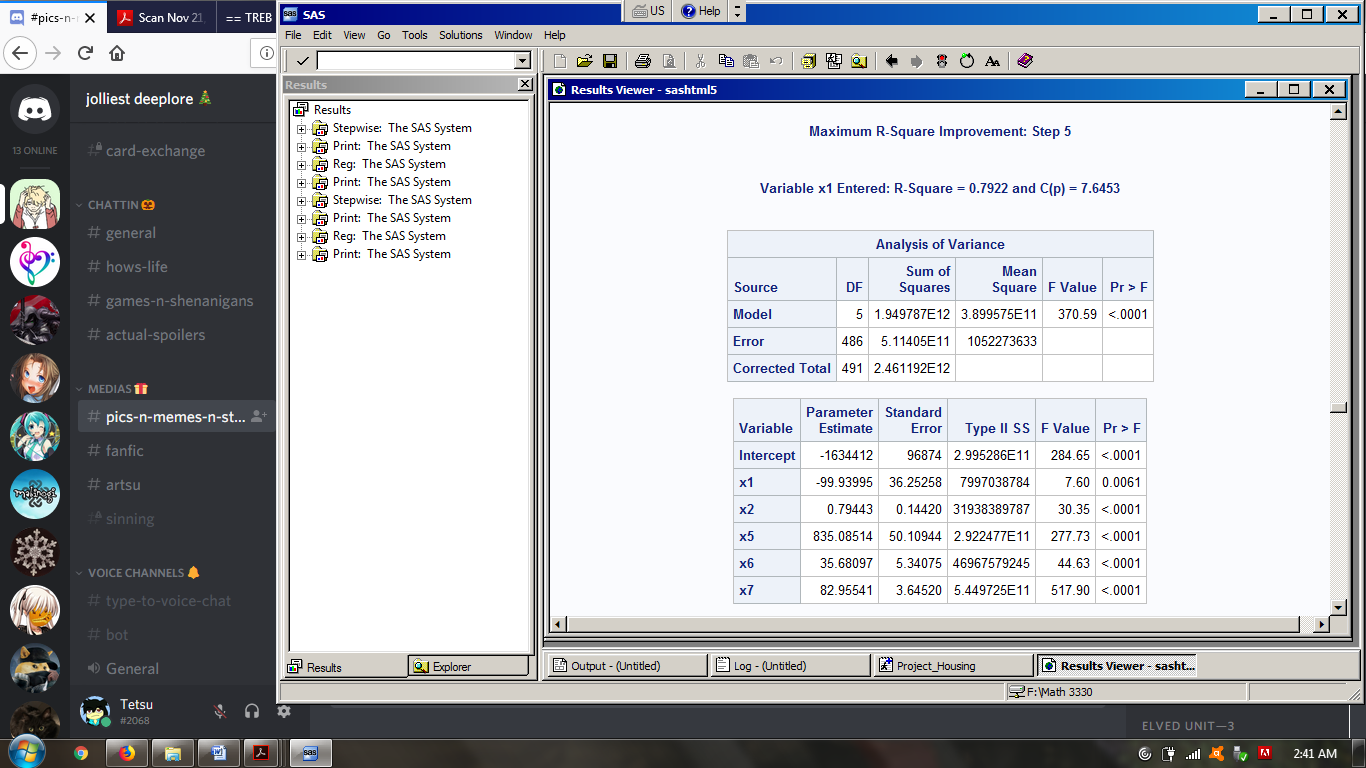
**Results**

The final model used contains variables x2, x5, x6, and x7 (lot area, year of construction, basement area and ground floor area, respectively), as well as dummy variables representing high quality (x3) and high condition (x4). According to the ANOVA table, we see that this model has a highly significant F statistic, explaining a large portion of the sample’s variation. We observed that the parameter estimates for all variables (minus the intercept) are positive, which suggests an increase in any variable is correlated with an increase in the dependent variable, home value. The p-values of all parameters in this model are all less than 0.0001, suggesting that the estimates for each of the parameters of the model are significant. Finally, the R-square and adjusted R-square values are 0.8436 and 0.8417, respectively, indicating the model is a good fit for the data.

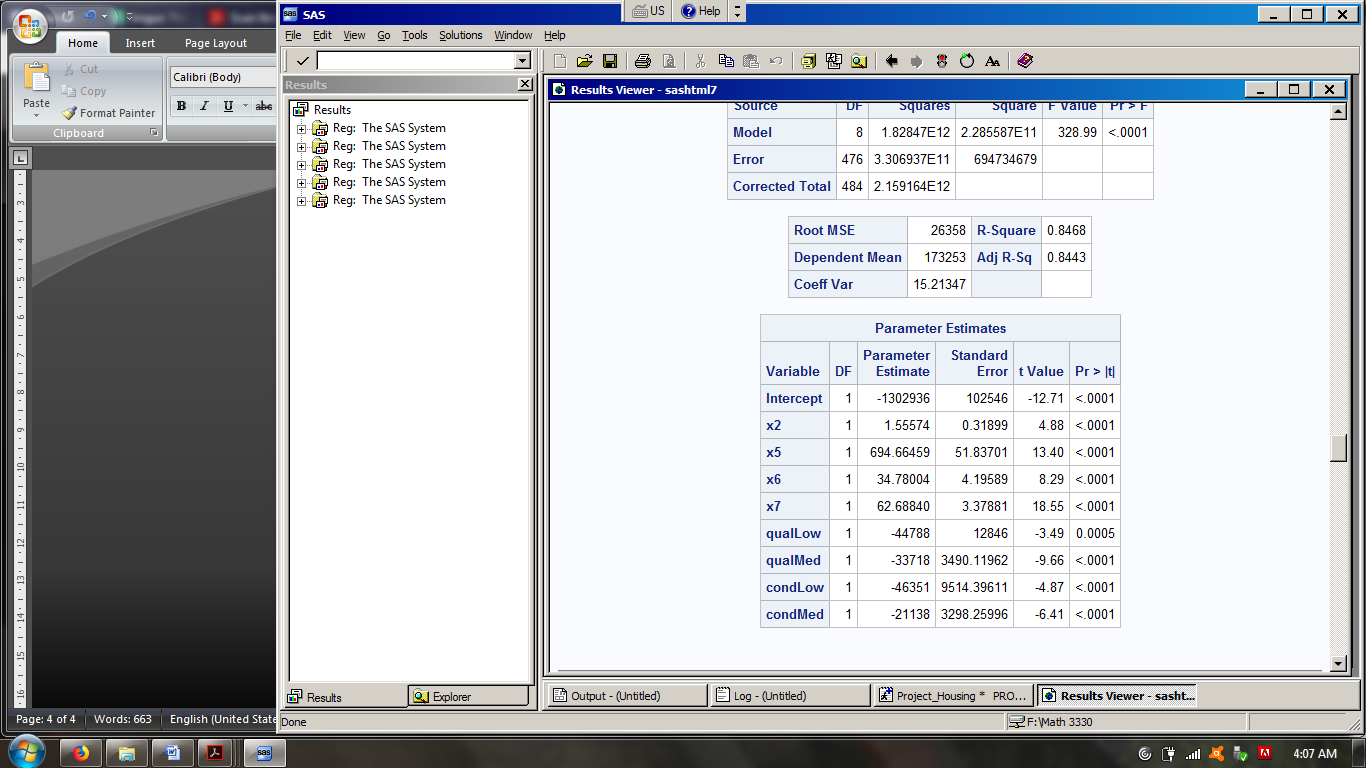


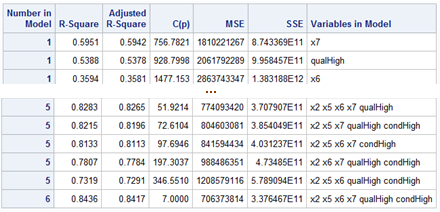
Initial model construction without inserting dummy variables revealed the most significant parameters were the year of construction (x5), ground (x7) and basement (x6) living area, and lot area (x2). Although the selection methods suggest the house type (x1) is also a significant parameter (p-value is 0.0061), we opt to exclude this from the model in an effort to create a simpler model.

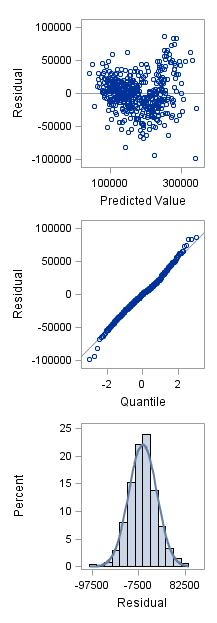
Residual analysis of the model based on x2, x5, x6, and x7 revealed one unusual entry that has a significant influence on the data set. Entry 251 has a Cook’s D of 7.557, significantly larger than the values observed in all the other points, which suggests a major influence on the data set. For this reason, the data point was expunged from the data to reduce outlier influence on the model.

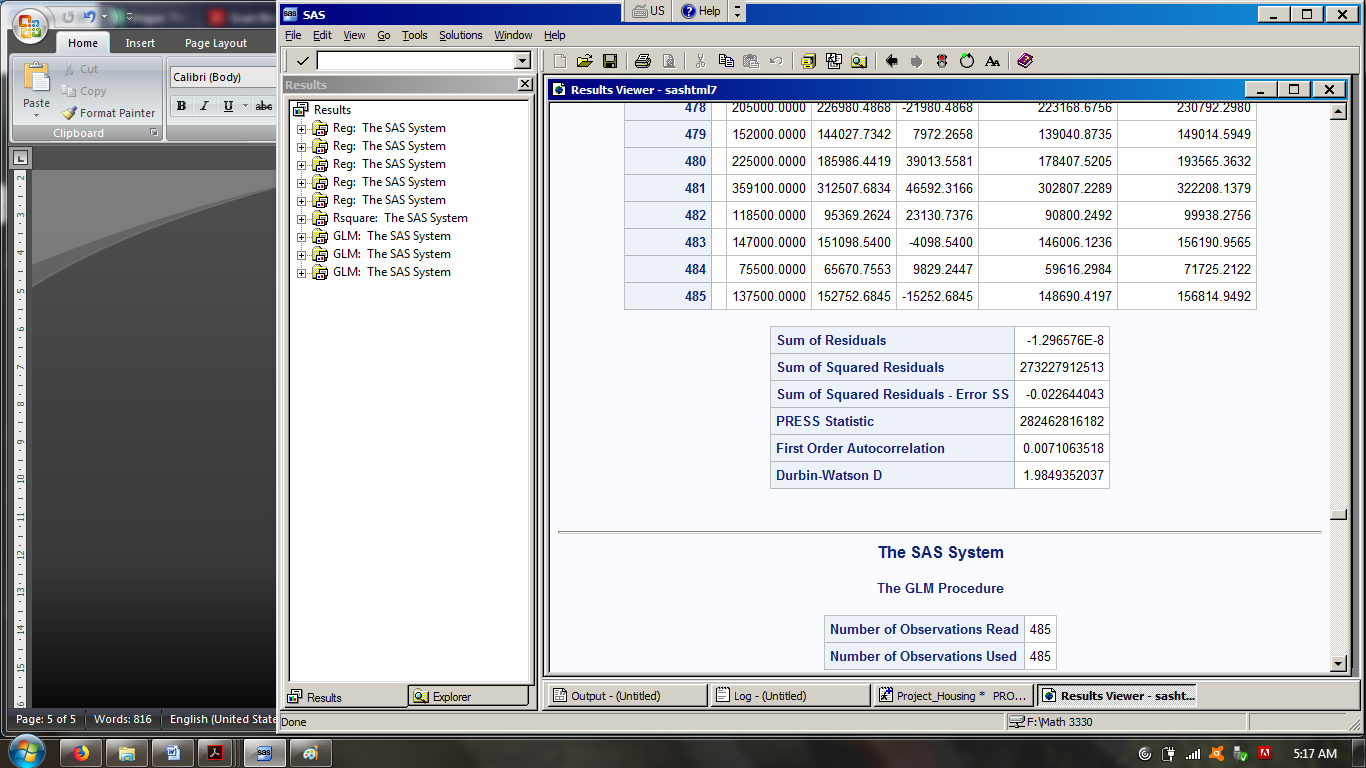


On including the dummy variables to the model, we see a similar increase in the R2 value between adding the medium and low quality and condition dummy variables and adding the high quality and condition dummy variables. In the interest of developing a simpler model we favour the use of the High Quality and High Condition dummy variables. Although removing the x2 parameter from the model would result in only a slight reduction in the amount of variance explained by the model, we see that the C statistic is not greater than the number of parameters for the most complex model (indicating unbiasness), while also boasting the lowest SSE of the possible models. Thus, the model containing parameters x2, x5, x6, x7, and the dummy variables High Quality and High Condition is the best choice.

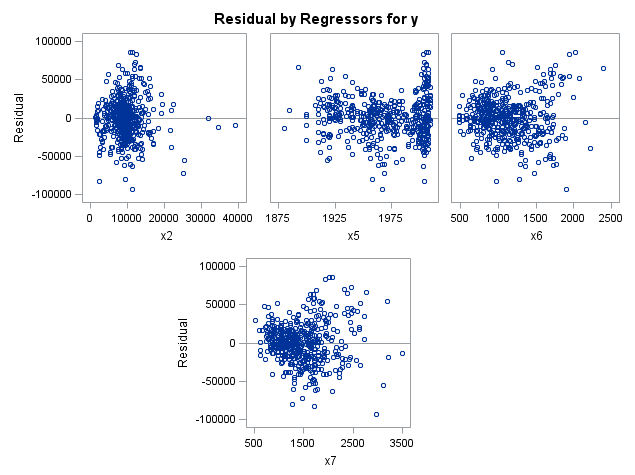


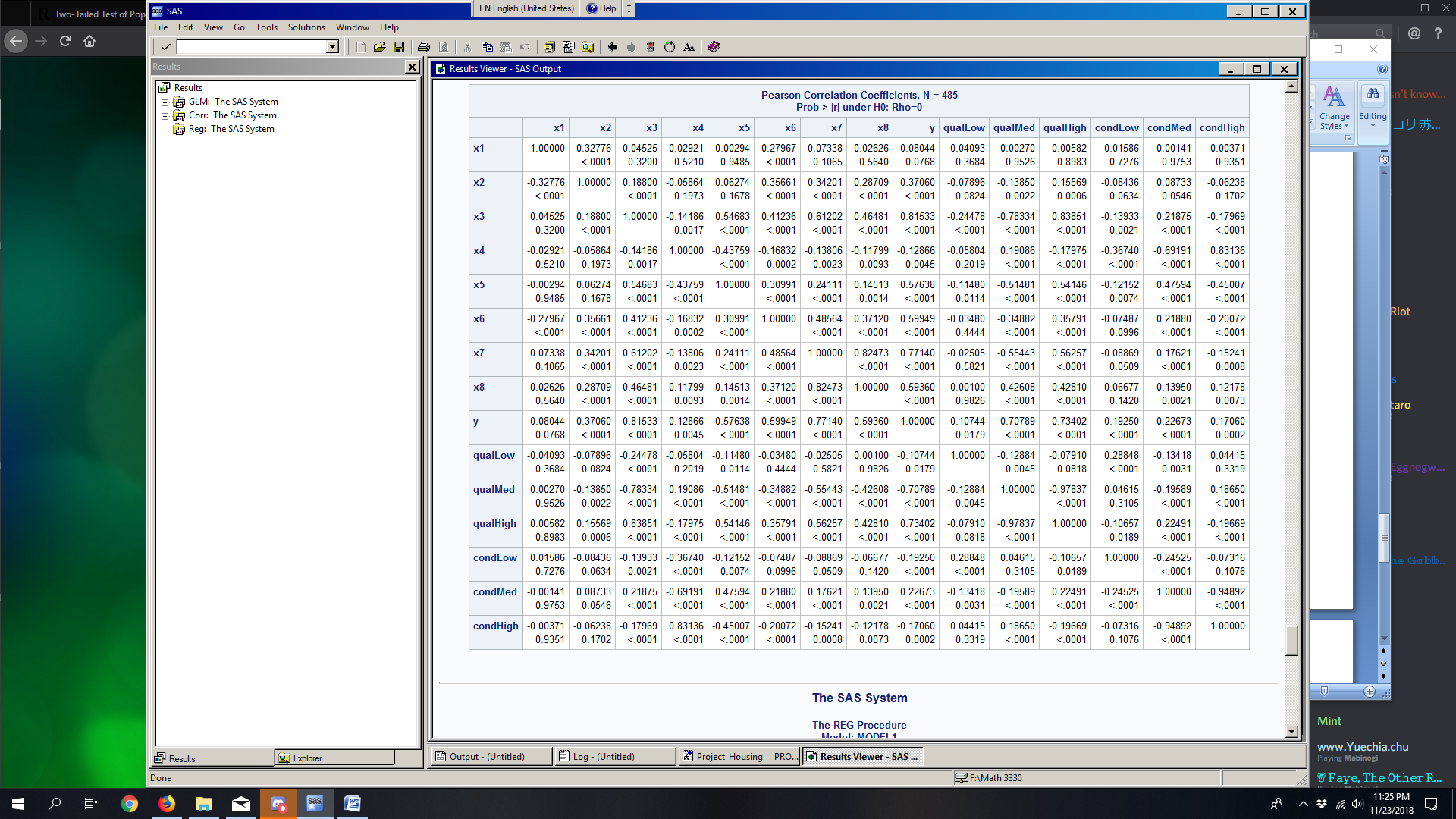


In addition, we find that this model is able to satisfy all three inference assumptions, as indicated by the fit diagnostic graphs for the dependent variable. The graph of residual against predicted value does not appear to form a fanning or funneling pattern, suggesting constant variance. The residual-quartile plot, together with the residual histogram, suggests a near normal distribution in the data, which strongly suggests normality in the population. Conducting the Durbin-Watson test we see the D value is significantly large, which suggests there is no autocorrelation for the data set and thereby satisfying the independence assumption.



Residual plots for the parameters used in the model (barring the dummy variables) shows a random distribution of points in x2, x5, x6, and x7 around 0. This suggests the model fits very well with the data. Although there are a number of points that lie beyond the majority in each plot, we believe that they are outliers that do not represent a significant proportion of the sample.



On constructing the correlation matrix for the housing value data set (pictured below), we discover that there are no significant correlation coefficients (ie. greater than 0.9) between the parameters used in our model. This indicates that there is a lack of evidence of multicollinearity between x2, x5, x6, and x7, further supporting the importance of these parameters in the model. Above the diagonal there are some significantly large correlation coefficients, but these values can be ignored as they are associated with variables that have been removed from the final model. Additionally, the variance inflation factors found in the final model’s parameter estimate table support the findings of the coefficient table, as none of the VIF values are greater than 10.

**Conclusion**

From the final model, it is safe to assume that there is strong evidence supporting our hypotheses, as an increase in x5, x3, or x4 is correlated with an increase in housing value. In other words, newer homes, homes considered high quality, or houses that have been well maintained are generally valued higher than houses that are none of the three. Of course, while the model we have arrived at supports this claim there are many more factors to consider in determining the value of a house. Variables that go beyond the physical characteristics of the house and its surroundings, such as location (country, city, district, etc.), mean area income, housing demand and market trends, access to city services or commodities, and many others are likely to also influence the value a house may have. In the absence of this information there is considerable room for improvement in the model, though with the data provided we believe the model we have arrived at is the most conclusive.

**Sources:**

https://www.ubs.com/global/en/asset-management/insights/asset-class-research/real-assets/2017/2017-3-global-real-estate-summary.html