Predicting Property Value

Group 5

Michael Didonato,
Vysnavi Mathavaraj,
Weiyuan Fan,
Yue He

December 14, 2020

Table of Contents

1 Introduction

1.1 Problem Statement

Our group is interested in the variables that affect the property value in America when we are trying to predict the house price. So, our project problem is: What variables are good predictors for property value in America？ If we solve this problem, we could build a model, run a regression, and solve some related problems. For example: when we are trying to purchase a house with 3 bedrooms, how much we should pay? If our annual income is $70,000, how much we should prepare so that we could afford the houses price？

1.2 Data Collection

Our intended population is homeowners of the United States from 2014 to 2018. The sample size is 18046 homes, after randomized and data cleaning, we used 5277 effective data in our multiple regression model. The data set is from American Community Survey 2014-2018 ACS 5-Year PUMS Files, which is a subset of the American Community Survey (ACS) Public Use Microdata Samples (PUMS) from US Census website.

2 Variable Selection

Our data contains 237 factors (columns), but most variables are unrelated to our problems and some variables have strong correlations with other variables. We chose these variables based on trial and error using StatTools and SPSS to compare various models with different combinations. By the Forward, Stepwise, and Backward, we chose the model with the following variables (including dummy variables) that looked the strongest to us.

Dependent Variable	
VALP	Property value
Independent Variable	
BDSP	Number of bedrooms
BATH (Dummy variable)	Bathrooms 1= the house has Bathtub or Shower); 2 = the house does not have Bathtub or shower
RMSP	Number of rooms
FINCP	Family Income
NPF	Number of people in a Family
HINCP	Household Income Past 12 Months
ACR (Dummy variable)	Lot Size 1=house on less than one acre; 2=house on one to less than ten acre; 3=house on ten or more acres
ELEP	Electricity monthly cost

3 Model Analysis

3.1 Model Selection

Use the training set. Confidence interval level =95%.

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Change Statistics				
					R Square Change	F Change	df1	df2	Sig. F Change
1	.459[a]	.210	.210	168733.922	.210	983.782	1	3690	.000
2	.558[b]	.311	.311	157606.767	.101	540.426	1	3689	.000
3	.574[c]	.330	.329	155508.299	.018	101.232	1	3688	.000
4	.580[d]	.336	.336	154764.254	.007	36.546	1	3687	.000
5	.584[e]	.341	.340	154290.728	.004	23.666	1	3686	.000
6	.586[f]	.344	.343	153914.686	.003	19.033	1	3685	.000
7	.587[g]	.345	.344	153841.313	.001	4.516	1	3684	.034

a. Predictors: (Constant), RMSP
b. Predictors: (Constant), RMSP, FINCP
c. Predictors: (Constant), RMSP, FINCP, NPF
d. Predictors: (Constant), RMSP, FINCP, NPF, ACR = 2
e. Predictors: (Constant), RMSP, FINCP, NPF, ACR = 2, ACR = 3
f. Predictors: (Constant), RMSP, FINCP, NPF, ACR = 2, ACR = 3, BATH = 1
g. Predictors: (Constant), RMSP, FINCP, NPF, ACR = 2, ACR = 3, BATH = 1, BDSP

（Figure 1）

Figure 1 shows the output of forward method. With forward method, model 7 is the best model and adj, R-square is the largest. But in the process of independent variable selection, forward method can only examine whether the independent variables are statistically significant when

they enter the model and do not consider the changes in the P-value of each independent variable. Backward method can avoid this problem.

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Change Statistics				
					R Square Change	F Change	df1	df2	Sig. F Change
1	.587[a]	.345	.343	153873.829	.345	215.344	9	3682	.000
2	.587[b]	.345	.343	153856.538	.000	.172	1	3682	.678
3	.587[c]	.345	.344	153841.313	.000	.271	1	3683	.603

a. Predictors: (Constant), ELEP, FINCP, ACR = 3, ACR = 2, BATH = 1, NPF, BDSP, RMSP, HINCP

b. Predictors: (Constant), ELEP, FINCP, ACR = 3, ACR = 2, BATH = 1, NPF, BDSP, RMSP

c. Predictors: (Constant), FINCP, ACR = 3, ACR = 2, BATH = 1, NPF, BDSP, RMSP

(Figure 2)

Figure 2 shows the output of backward method. The optimal prediction model obtained by the backward method is the same as that from the forward method. But in the backward method, if the collinearity between the independent variables is significant, the conclusion may be incorrect. Based on our analysis (mmulticollinearity part), there are some collinear relationships between our variables, but they are not enough to be considered. Just in case, we still use the stepwise method for verification.

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Change Statistics				
					R Square Change	F Change	df1	df2	Sig. F Change
1	.459[a]	.210	.210	168733.922	.210	983.782	1	3690	.000
2	.558[b]	.311	.311	157606.767	.101	540.426	1	3689	.000
3	.574[c]	.330	.329	155508.299	.018	101.232	1	3688	.000
4	.580[d]	.336	.336	154764.254	.007	36.546	1	3687	.000
5	.584[e]	.341	.340	154290.728	.004	23.666	1	3686	.000
6	.586[f]	.344	.343	153914.686	.003	19.033	1	3685	.000
7	.587[g]	.345	.344	153841.313	.001	4.516	1	3684	.034

a. Predictors: (Constant), RMSP

b. Predictors: (Constant), RMSP, FINCP

c. Predictors: (Constant), RMSP, FINCP, NPF

d. Predictors: (Constant), RMSP, FINCP, NPF, ACR = 2

e. Predictors: (Constant), RMSP, FINCP, NPF, ACR = 2, ACR = 3

f. Predictors: (Constant), RMSP, FINCP, NPF, ACR = 2, ACR = 3, BATH = 1

g. Predictors: (Constant), RMSP, FINCP, NPF, ACR = 2, ACR = 3, BATH = 1, BDSP

(Figure 3)

Figure 3 shows the output of stepwise method. The optimal model obtained by the stepwise method is the same as that obtained by the previous two methods. Therefore, our prediction model is the model indicated in Figure 4. Finally, there are 7 independent variables in our predictive model.

Coefficients[a]

Model		Unstandardized Coefficients B	Unstandardized Coefficients Std. Error	Standardized Coefficients Beta	t	Sig.	95.0% Confidence Interval for B Lower Bound	95.0% Confidence Interval for B Upper Bound
7	(Constant)	-9632.985	14293.995		-.674	.500	-37657.907	18391.937
	RMSP	24467.135	1515.157	.299	16.148	.000	21496.506	27437.764
	FINCP	.751	.033	.326	22.934	.000	.687	.815
	NPF	-12887.431	1489.888	-.121	-8.650	.000	-15808.518	-9966.343
	ACR = 2	38058.460	5765.392	.090	6.601	.000	26754.785	49362.135
	ACR = 3	80860.430	15906.249	.068	5.084	.000	49674.510	112046.351
	BATH = 1	51297.679	12156.561	.059	4.220	.000	27463.427	75131.931
	BDSP	7542.916	3549.496	.039	2.125	.034	583.746	14502.087

a. Dependent Variable: VALP

(Figure 4)

VALP = - 9632.985 + 38058.460 (ACR = 2) + 80860.430 (ACR = 3) + 51297.679 (BATH = 1) + 7542.916 BDSP + 24467.135 RMSP + 0.751 FINCP - 12887.431 NPF

3.2 Model Summary

Multiple Regression for VALP Summary	Multiple R	R-Square	Adjusted R-square	Std. Err. of Estimate
	0.587	0.345	0.344	153841.313

(Figure 5)

Multiple coefficient of determination (R-Square = 0.345) implies that 34.5% of the variation in the dependent variable, VALP is explained by the independent variables. It also indicates that almost 65.5% of variations in the dependent variable can be attributed to other possible independent variables which are not included in the regression specification.

3.3 Hypothesis Test

ANOVA Table	Degrees of Freedom	Sum of Squares	Mean of Squares	F	p-Value
Explained	7	4.59E+13	6.55E+12	2.77E+02	< 0.0001
Unexplained	3684	8.72E+13	2.37E+10		

Regression Table	Coefficient	Standard Error	t-Value	p-Value	Confidence Interval 95% Lower	Confidence Interval 95% Upper	Multicollinearity Checking VIF	Multicollinearity Checking R-Square
Constant	-9632.98	14293.99	-0.67	0.500	-37657.91	18391.94		
ACR = 2	38058.46	5765.39	6.60	0.000	26754.78	49362.13	1.053	0.050
ACR = 3	80860.43	15906.25	5.08	0.000	49674.51	112046.35	1.020	0.019
BATH = 1	51297.68	12156.56	4.22	0.000	27463.43	75131.93	1.092	0.084
BDSP	7542.92	3549.50	2.13	0.034	583.75	14502.09	1.853	0.460
RMSP	24467.13	1515.16	16.15	0.000	21496.51	27437.76	1.928	0.481
FINCP	0.75	0.03	22.93	0.000	0.69	0.82	1.137	0.121
NPF	-12887.43	1489.89	-8.65	0.000	-15808.52	-9966.34	1.101	0.092

(Figure 6)

H_0: $\beta_j = 0$ (i.e.) The slope coefficients of all independent variables are equal to zero.

H_a: $\beta_j \neq 0$ (i.e.) The slope coefficient of at least one independent variable is not equal to zero.

Rejection rule: If significance F (p-value) < Alpha, Reject H_0.

- Joint Test - From ANOVA table (Figure 6), p-value is 0.000 which is less than alpha (0.05), we reject H_0. This implies that the regression model is significant.
- Individual Test – From the regression table (Figure 6), p-value of all independent variables are less than alpha (0.05), we reject H_0. This implies that all the independent variables are significantly related to the dependent variable, VALP.

4 Assumptions

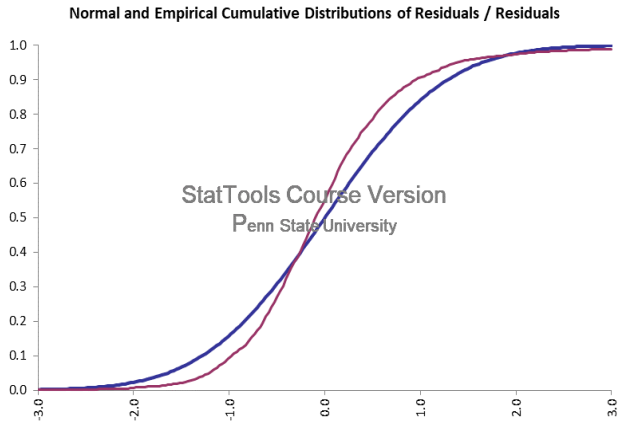
- Normality:

H_0: The data are normally distributed; H_a: The data are NOT normally distributed

Rejection rule: Test statistic (t_{stat}) > CVal, Reject H_0.

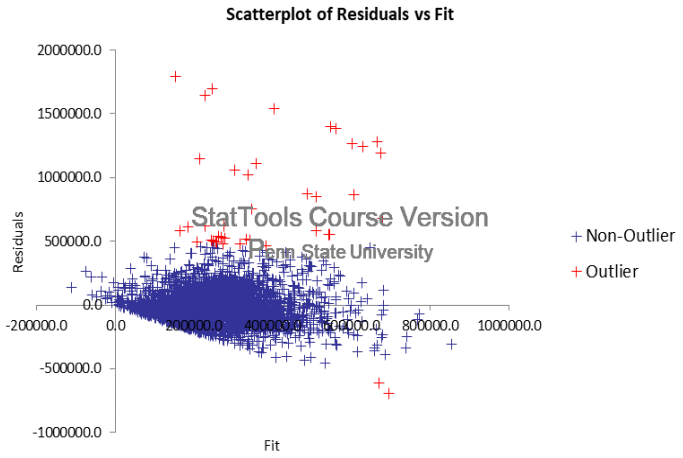
In numerical output (Figure 7), the test statistic (t_{stat}) is greater than the CVal values at 5%, 10%, 15%, 2.5% and 1% significance levels. Therefore, we reject the null hypothesis of normality. The graphical output (Figure 7) shows that there is a gap between the expected curve and the observed curve, the probability of the residual distribution is not normal.

Lilliefors Test Results	Residuals
Sample Size	3692
Sample Mean	0.00
Sample Std Dev	153695.36
Test Statistic	0.101
CVal (15% Sig. Level)	**0.013**
CVal (10% Sig. Level)	**0.013**
CVal (5% Sig. Level)	**0.015**
CVal (2.5% Sig. Level)	**0.016**
CVal (1% Sig. Level)	**0.022**



(Figure 7)

- Linearity: According to the residual plot (Figure 8), there is a fan-shaped pattern indicating the residuals are not randomly dispersed, linearity (mean of zero) assumption is violated.
- Homoscedasticity: From the scatterplot of residuals vs fit (Figure 8), it is evident that residuals appear to fan out as $\hat{y}$ (predicted y) increases. It has an increasing error variance. Therefore, this model violates homoscedasticity assumption.
- Multicollinearity: From the Regression table (Figure 8), the VIF for all the independent variables are between 1.00 and 2.00, which indicates that there is no significant multicollinearity.



(Figure 8)

5 Residual Analysis

	R-Square	Adj. R-Square	Std. Error	SSE	RMSE
Validation Set	0.2965	0.2934	172700.36	4.70347E+13	172263.98
Training Set	0.3448	0.3435	153841.31	3.73231E+13	153452.58

(Figure 9)

Run the prediction model in the validation set. By calculating residuals, residuals square, and others, we get R Square, Adj R Square, standard errors, SSE and RMSE (Root-mean-square error). These implies that in validation set, 29.65% of the variation in the dependent variable VALP is explained by the independent variables which are ACR=2, ACR=3, BATH=1, BDSP, RMSP, FINCP and NPF.

Comparing the validation set with the training set, it is found that the R-square and Adj R-square have become smaller, indicating that the independent variable's interpretation of the dependent variable has become smaller. But this change is only 4.83% and 5.01%. The standard errors and SSE become bigger. This shows that the residual of the validation set is greater than that of the training set. It shows that the model fitting is not very good, and the data prediction is not very successful. RMSE is still large, indicating that the predictive results are unstable.

6 Conclusion

To predict property value in America, we use 6 explanatory variables: number of bedrooms, number of rooms, family income, number of people in a family, types of bathrooms, and types lot size. Upon inspection, number of bedrooms and number of total rooms had minor collinearity according to the correlation matrix but far below VIF threshold. Number of people in a family decreased the overall property value (NPF coefficient = -$12,887 per person). Besides, the amount of land had the largest contributing factor to Property Value (ACR = 3 More than 10 acres of land) coefficient of $80,860

7 Future Directions

- Data collection of alternate variables that subject matter experts believe to have effect on property value. This will increase our R2 from 35%. Such variables could be crime rate in the area, credit score of homeowners, age of the home, rural, suburban, urban categorical variables.
- Using prior years PUMS property value data (2014, 2015, 2016, 2017 and 2018) we could set-up a forecasted analysis to predict average future property value using Holt's trend methodology
- Set-up a logistical regression to guess if a homeowner has a mortgage or owns the property outright. Based on low interest rates currently, the firm can target mortgage holders to try and get them to refinance.
    a) Make property value a dependent variable
    b) Add independent variable TEN for analysis

```
TEN         Character   1
      Tenure

            1 .Owned with mortgage or loan (include home equity loans)
            2 .Owned free and clear
```

Works Cited

"American Community Survey 2014-2018 ACS 5-Year PUMS Files." 2020, www.census.gov/programs-surveys/acs/technical-documentation/pums/documentation.html.