Tan-Hsiu Lan
NEAS Regression Analysis
Spring 2012

Home Sale Prices

Introduction

This project aims at investigating some variables which might influence home prices and finding out the most suitable "Regression Analysis" in order to apply the models in predicting average home prices.

Data

Source： http://lib.stat.cmu.edu/DASL/Datafiles/homedat.html

The data are a random sample of records of resale of homes from Feb 15 to Apr 30, 1993 from the files maintained by the Albuquerque Board of Realtors. This type of data is collected by multiple listing agencies in many cities and is used by realtors as an information base.

There are about 117 raw data. However, some data (AGE and TAX) are incomplete and omitted. Therefore, the 66 complete data to the use of "Regression Analysis ".

Variable names

Dependent variable: (Y)

PRICE = Home Selling price ($hundreds)

Explanatory Variables: (X_i)

1. SQFT = Square feet of living space
2. AGE = Age of home (years)
3. FEATS = Number out of 11 features (dishwasher, refrigerator, microwave, disposer, washer, intercom, skylight(s), compactor, dryer, handicap fit, cable TV access
4. NE = Located in northeast sector of city (1) or not (0)
5. CUST= Custom built(1) or not (0)
6. COR = Corner location (1) or not (0)
7. TAX = Annual taxes ($)

Initial Regression Equation

Hypothesis

The null hypothesis which is on trial by the researcher shows that all the regression coefficents (the β_i) are zero： $\beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = \beta_6 = \beta_7 = 0$

Based on the data, I defined a regression equation of home sales price (Y) by the follow:

$$Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \beta_6 X_6 + \beta_7 X_7 + \varepsilon$$

Where Y is the home sales price, X_1 , X_2 , X_3 , X_4 , X_5 , X_6 , X_7 are Square feet of living space(SQFT), Age of home (AGE), Features (FEATS), Located in northeast sector of city (NE), Custom built (CUST) ,Corner location (COR), Annual taxes (TAX) respectively.

Using the 7 Variables to proceed regression analysis, the result is as follows：

Table 1: 7 Variables Regression Model

Regression Statistics	
Multiple R	0.9286
R Square	0.8623
Adjusted R Square	0.8456
Standard Error	158.8811
Observations	66

ANOVA

	df	SS	MS	F	Significance F
Regression	7	9164659	1309237	51.8650	0.0000
Residual	58	1464105	25243		
Total	65	10628764			

	Coefficients	Standard Error	t Stat	P-value
Intercept	92.7448	101.6070	0.9128	0.3651
SQFT	0.3522	0.0957	3.6786	0.0005
AGE	-0.5651	2.0025	-0.2822	0.7788
FEATS	4.3896	18.5550	0.2366	0.8138
NE	-17.3853	47.2746	-0.3678	0.7144
CUST	174.9411	53.7237	3.2563	0.0019
COR	-73.5823	49.1301	-1.4977	0.1396
TAX	0.4989	0.1585	3.1477	0.0026

This model can be represented by the following equation:

$Y = 92.7448 + 0.3522X_1 - 0.5651X_2 + 4.3896X_3 - 17.3853X_4 + 174.9411X_5 - 73.5823X_6 + 0.4989X_7$

The R^2 and adjusted R^2 of the full model are 86.23% and 84.56%, indicating that the model totally significant. But almost each explanatory variable is not significant (AGE, FEATS, NE, COR). Therefore, the model must be adjusted.

Explanatory Variables Screen

1. Correlation

Table 2 : Correlation Matrix

	PRICE	SQFT	AGE	FEATS	NE	CUST	COR	TAX
PRICE	1.0000							
SQFT	0.8839	1.0000						
AGE	-0.1667	-0.0377	1.0000					
FEATS	0.3663	0.3574	-0.1835	1.0000				
NE	0.2892	0.3625	0.2164	0.3096	1.0000			
CUST	0.5821	0.4919	0.0085	0.3122	0.1502	1.0000		
COR	-0.1876	-0.0785	0.1627	-0.2491	-0.0237	-0.0537	1.0000	
TAX	0.8775	0.8752	-0.2918	0.3040	0.3024	0.4370	-0.1532	1.0000

As seen in the correlation matrix above, the variable SQFT and TAX has high correlation of 0.8752.IT indicates that the two explanatory variables has Multi co-linearity. Further analyzed as follows：

Table 3: Use explanatory variable TAX regress on dependent variable SQFT

Regression Statistics	
Multiple R	0.8752
R Square	0.7661
Adjusted R Square	0.7624
Standard Error	249.7531
Observations	66

ANOVA

	df	SS	MS	F	Significance F
Regression	1	13072670	13072670	209.5765	0.0000
Residual	64	3992102	62377		

Total	65	17064772

	Coefficients	Standard Error	t Stat	P-value
Intercept	514.0332	90.7944	5.6615	0.0000
TAX	1.4172	0.0979	14.4768	0.0000

According to the table displayed above, it states that the SQFT and Tax of the explanatory variables are apparently related to each other (P-value=0), and consider that the SQFT has the higher correlation coefficient with the dependent variables, TAX will be removed from the model and then the implementation will be relied upon the 6 remaining variables to proceed the regression.

Adjusted Regression Equation

6 Variables Model

Table 4: 6 Variables Regression Model

Regression Statistics	
Multiple R	0.9158
R Square	0.8387
Adjusted R Square	0.8223
Standard Error	170.4541
Observations	66

ANOVA

	df	SS	MS	F	Significance F
Regression	6	8914543	1485757	51.1367	0.0000
Residual	59	1714221	29055		
Total	65	10628764			

	Coefficients	Standard Error	t Stat	P-value
Intercept	161.0689	106.4918	1.5125	0.1357
SQFT	0.6134	0.0512	11.9700	0.0000
AGE	-4.1068	1.7772	-2.3108	0.0244
FEATS	-8.5281	19.4136	-0.4393	0.6621
NE	9.1935	49.9026	0.1842	0.8545
CUST	189.4036	57.4258	3.2982	0.0017

COR	-96.7459	52.1141	-1.8564	0.0684

This model can be represented by the following equation:

$$Y = 161.0689 + 0.6134X_1 - 4.1068X_2 - 8.5281X_3 + 9.1935X_4 + 189.4036X_5 - 96.7459X_6$$

The t- statistic associated with NE (0.1842) is the lowest (in absolute value terms) of the six explanatory variables. Since a higher t-statistic indicates a better estimate of the true coefficient, the variable associated with the lowest t-statistic would be the least likely to be a good estimator. In addition, the p-value associated with NE (0.8545) is the highest of the p-values of the six explanatory variables. In general, the lower the p-value, the more likely it is that the result is significant. Thus, both of these statistics indicate that NE is the least significant variable of those shown above. Thus, NE will be removed from the model and then the implementation will be relied upon the 5 remaining variables to proceed the regression testing.

5 Variables Model

Table 5: 5 Variables Regression Model

Regression Statistics	
Multiple R	0.9158
R Square	0.8386
Adjusted R Square	0.8252
Standard Error	169.0763
Observations	66

ANOVA

	df	SS	MS	F	Significance F
Regression	5	8913557	1782711	62.3614	0.0000
Residual	60	1715207	28587		
Total	65	10628764			

	Coefficients	Standard Error	t Stat	P-value
Intercept	156.6329	102.8953	1.5223	0.1332
SQFT	0.6162	0.0485	12.7049	0.0000
AGE	-4.0094	1.6830	-2.3823	0.0204
FEATS	-7.5435	18.5126	-0.4075	0.6851
CUST	188.3923	56.7008	3.3226	0.0015
COR	-96.5201	51.6786	-1.8677	0.0667

This model can be represented by the following equation:

$Y = 156.6329 + 0.6162X_1 - 4.0094X_2 - 7.5435X_3 + 188.3923X_4 - 96.5201X_5$

The p-value associated with FEATS (0.6851) is the highest of the p-values of the five explanatory variables. Thus, FEATS will be removed from the model and then the implementation will be relied upon the 5 remaining variables to proceed the regression testing.

4 Variables Model

Table 6: 4 Variables Regression Model

Regression Statistics	
Multiple R	0.9155
R Square	0.8382
Adjusted R Square	0.8276
Standard Error	167.9165
Observations	66

ANOVA

	df	SS	MS	F	Significance F
Regression	4	8908810	2227202	78.9901	0.0000
Residual	61	1719954	28196		
Total	65	10628764			

	Coefficients	Standard Error	t Stat	P-value
Intercept	133.0733	84.5312	1.5742	0.1206
SQFT	0.6116	0.0468	13.0661	0.0000
AGE	-3.9007	1.6504	-2.3636	0.0213
CUST	184.3325	55.4357	3.3252	0.0015
COR	-92.0147	50.1357	-1.8353	0.0713

This model can be represented by the following equation:

$Y = 133.0733 + 0.6116X_1 - 3.9007X_2 + 184.3325X_3 - 92.0147X_4$

The p-value associated with COR (0.0713) is the highest of the p-values of the four explanatory variables. Thus, COR will be removed from the model and then the implementation will be relied upon the 5 remaining variables to proceed the regression testing.

3 Variables Model

Table 7: 3 Variables Regression Model

Regression Statistics	
Multiple R	0.9106
R Square	0.8292
Adjusted R Square	0.8210
Standard Error	171.0937
Observations	66

ANOVA

	df	SS	MS	F	Significance F
Regression	3	8813835	2937945	100.3635	0.0000
Residual	62	1814928	29273		
Total	65	10628764			

	Coefficients	Standard Error	t Stat	P-value
Intercept	111.0316	85.2569	1.3023	0.1976
SQFT	0.6161	0.0476	12.9367	0.0000
AGE	-4.3885	1.6596	-2.6442	0.0104
CUST	186.6379	56.4701	3.3051	0.0016

This model can be represented by the following equation:

$Y = 111.0316 + 0.6161X_1 -4.3885X_2 + 186.6379X_3$

The P-value of AGE is 0.0104. In 95% confident level is significant, but in 99% confident level is not significant. Thus, AGE will be removed from the model and then the implementation will be relied upon the 5 remaining variables to proceed the regression testing.

2 Variables Model

Table 8: 2 Variables Regression Model

Regression Statistics	
Multiple R	0.9000
R Square	0.8100
Adjusted R Square	0.8040

Standard Error	179.0451
Observations	66

ANOVA

	df	SS	MS	F	Significance F
Regression	2	8609163	4304582	134.2784	0.0000
Residual	63	2019600	32057		
Total	65	10628764			

	Coefficients	Standard Error	t Stat	P-value
Intercept	35.1922	84.0191	0.4189	0.6767
SQFT	0.6222	0.0498	12.4982	0.0000
CUST	181.9942	59.0659	3.0812	0.0031

This model can be represented by the following equation:

$Y= 35.1922 +0.6222X_1 + 181.9942X_2$

The R^2 (81.00%) and adjusted R^2 (80.40%) in this model are lower than those in the rest of the models examined above.

Conclusions

Raw Data can't be used directly. We should consider the multi co-linearity and adjusted the regression model. After, the null hypothesis which is on trial by the researcher shows that all the regression coefficents (the β_i) are zero.

To sum up, the number of not significant explanatory variables have to be 0 in order to have the best model outcomes. Meanwhile, the F Statistics, R Square, and Adjusted R Square must be as higher as it could be, while the Standard Error needs to be as lower as possible.

The table below assembles the outcomes of all the models. By comparison, the best regression model is found out. Because the Number of not significant explanatory variables must be 0, it shows that only model 2 and 3 match to the result. Also, in model 2 and 3, the statistics of the R Square, Adjusted R Square and Standard Error point out that the 3 explanatory variables model is better. However, the 2

explanatory variables model of F Statistic is also acceptable, but the 3 explanatory variables model is apparently outstanding.
Due to what has been mentioned above, it evidences that the three explanatory variables (SQFT, AGE and CUST) will considered to be the best choice.

Table 9: Summarizes the results of the regression analysis performed

Number of explanatory variables in models	F Stat	R Square	Adjusted R Square	Standard Error	Number of not significant explanatory variables (95% confident level)
7	51.86	86.23%	84.56%	159	4
6	51.14	83.87%	82.23%	170	3
5	62.36	83.86%	82.52%	169	2
4	78.99	83.82%	82.76%	168	1
3	**100.36**	**82.92%**	**82.10%**	**171**	**0**
2	134.28	81.00%	80.40%	179	0

According to the statement mentioned above, it can be noted that the regression model, which is composed of the following explanatory variables, Square feet of living space, Age of home, and Custom built (SQFT, AGE and CUST), likely being reasonable illustrate the level of house prices. The regression model is as below.

$Y = 111.0316 + 0.6161X_1 - 4.3885X_2 + 186.6379X_3$