Analysis on housing price of Melbourne

Yang Yifeng

yyangbf@ust.hk

Introduction

The housing price has always been a popular topic with the vicissitude of daily life and of great concern among the public. A variety of sophisticated methods has been employed on analyzing and forecasting housing price in different regions of the world. In this report, we utilize the traditional linear regression with appropriate modification to achieve accurate estimation of housing price in Melbourne, Australia.

The core advantage of our analysis lies in the affluence of factors in our regression model. Because of the sufficient factors in the model, we are able to achieve satisfactory R-square. In addition, by using the stepwise selection, we also succeeded in removing insignificant factors on housing price. Below is the summary of the statistics used in the analysis (see table1).

Table1. Summary of the statistics

factor name	description	factor name	description
		Bedroom2	Scraped # of Bedrooms
Suburb	Suburb	Bathroom	Number of Bathrooms
Address	Address	Car	Number of carspots
Rooms	Number of rooms	Landsize	Land Size
Type	type of the house	BuildingArea	building size
Price	price of the house	YearBuilt	Year the house was built
Method	how the property sold	CouncilArea	Governing council for the area
SellerG	real estate agent	Lattitude	lattitude of the house
Date	date sold	Longtitude	Self explanitory
Distance	Distance from CBD	Regionname	Self explanitory
Postcode	postcode of the house	Propertycount	Number of properties that exist in the suburb

Among all the variables, “suburb”,”address”,”type”,”method”,”sellerG”,”postcode”,”councilarea”,”regionname” are categorical variables. The rest, “Rooms”,”price”,”date”,”distance”,”bedroom2”,”bathroom”,”car”,”landsize”,”buildingarea”,”yearbuilt”,”latitude”,”longtitude”,”propertycount”, are continuous variable. To sum up, there are 8 categorical variables and 13 continuous variables. Notice that we modify “date” by calculating the difference between each date and the earliest date in the data set in order to transform it into continuous numerical variable. Moreover, due to the one-to-one correspondence between “postcode”, “council area” and “suburb”, we can safely delete “postcode” and “councilarea”. Due to the many-to-one relationship between “address”, “suburb” and “regionname”, we choose to delete “address” and “suburb” because the information contained in these two variables can be reflected by “regionname” and these two categorical variables have more than 100 levels.

Methodology

We first utilize the linear regression model with price as the dependent variable and all the other variables left as independent variables(including interaction terms). Please see table2 for the regression result.

Table2. Regression result of the full model

Source	DF	Type III SS	Mean Square	F Value	Pr > F
Regionname	5	260000588739	52000117748	0.62	0.6880
Rooms	1	11322263606	11322263606	0.13	0.7143
Type	2	1.4596808E12	729840391333	8.64	0.0002
Method	5	1085614145.9	217122829.17	0.00	1.0000
SellerG	56	1.0014386E13	178828321324	2.12	<.0001
datec	1	583490200.19	583490200.19	0.01	0.9338
Distance	1	130833050.75	130833050.75	0.00	0.9686
Bedroom2	1	19519524062	19519524062	0.23	0.6308
Bathroom	0	0	.	.	.
Car	1	143226562596	143226562596	1.70	0.1930
Landsize	1	57798128547	57798128547	0.68	0.4082
BuildingArea	1	180165860579	180165860579	2.13	0.1443
YearBuilt	1	947708447.75	947708447.75	0.01	0.9157
Lattitude	1	114572761.13	114572761.13	0.00	0.9706
Longtitude	1	774549309.78	774549309.78	0.01	0.9237
Propertycount	1	15052907403	15052907403	0.18	0.6730
Rooms*Regionname	0	0	.	.	.
Type*Regionname	8	3.7863965E12	473299560159	5.60	<.0001
Method*Regionname	8	578695230132	72336903767	0.86	0.5530
SellerG*Regionname	84	1.0512399E13	125147604814	1.48	0.0030
datec*Regionname	4	643389245370	160847311343	1.90	0.1069
Distance*Regionname	0	0	.	.	.
Bedroom2*Regionname	4	1.0560962E12	264024040291	3.13	0.0141
Bathroom*Regionname	0	0	.	.	.
Car*Regionname	0	0	.	.	.
Landsize*Regionname	4	523177132714	130794283179	1.55	0.1854
YearBuilt*Regionname	1	289970286748	289970286748	3.43	0.0640
Lattitude*Regionname	5	4.0408095E12	808161908858	9.57	<.0001
Longtitud*Regionname	6	5.284089E12	880681498037	10.42	<.0001
Propertyc*Regionname	6	1.6319573E12	271992887553	3.22	0.0037
Rooms*Type	2	569421868524	284710934262	3.37	0.0345
Type*SellerG	82	5.2265583E12	63738516280	0.75	0.9516
Distance*Type	2	1.242821E13	6.2141052E12	73.55	<.0001
Bathroom*Type	2	401063450052	200531725026	2.37	0.0932
Landsize*Type	2	1.1368075E13	5.6840376E12	67.28	<.0001
BuildingArea*Type	2	428412077850	214206038925	2.54	0.0793
YearBuilt*Type	2	1.3336894E12	666844679940	7.89	0.0004
Method*SellerG	150	8.4797715E12	56531810058	0.67	0.9993
Distance*Method	4	1.8023265E12	450581631007	5.33	0.0003
Bathroom*Method	4	368719139721	92179784930	1.09	0.3591
BuildingArea*Method	4	194158531442	48539632861	0.57	0.6811
YearBuilt*Method	4	2.5063518E12	626587947968	7.42	<.0001
Rooms*SellerG	47	1.7692506E13	376436288916	4.46	<.0001
datec*SellerG	52	3.2357773E12	62226486955	0.74	0.9210
Distance*SellerG	45	8.8349626E12	196332502229	2.32	<.0001
Bathroom*SellerG	55	7.183973E12	130617690129	1.55	0.0061
Car*SellerG	55	8.4775129E12	154136597476	1.82	0.0002
Landsize*SellerG	54	1.148755E13	212732407769	2.52	<.0001
BuildingArea*SellerG	55	2.1948841E13	399069836240	4.72	<.0001
YearBuilt*SellerG	54	4.2587861E12	78866410042	0.93	0.6134
Lattitude*SellerG	51	1.0167682E13	199366307659	2.36	<.0001
Longtitude*SellerG	46	7.4344803E12	161619136272	1.91	0.0002

We set the significance level to be 0.01. Consequently, some insignificant interaction terms can be dropped. Then we refit the model with the remaining terms and trying to test the significance of interaction terms again. We repeat this procedure several times until no interaction term has p-value greater than 0.01. Notice that we delete "propertycount" during the iteration of the procedure because its p-value is 0.6231(greater than 0.01) after we delete its interaction terms(see table3). The final regression results are listed in table4 and table5.

Table5. Regression statistics of the final model

The SAS System

The GLM Procedure

Dependent Variable: Price Price

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1163	2.6213077E15	2.2539189E12	26.90	<.0001
Error	5666	4.7468884E14	83778474569		
Corrected Total	6829	3.0959965E15			

R-Square	Coeff Var	Root MSE	Price Mean
0.846677	26.86007	289445.1	1077604

Table3. Regression result when deleting propertycount

Rooms	1	64240706471	64240706471	0.77	0.3813
Type	2	1.9697995E12	984899737063	11.75	<.0001
Method	4	3.444112E12	861027992378	10.28	<.0001
SellerG	64	1.4177111E13	221517357723	2.64	<.0001
datec	1	1.066527E13	1.066527E13	127.29	<.0001
Distance	1	81844754.26	81844754.26	0.00	0.9751
Bedroom2	1	11821896191	11821896191	0.14	0.7072
Bathroom	1	59489723641	59489723641	0.71	0.3995
Car	1	304242395596	304242395596	3.63	0.0568
Landsize	1	94354682673	94354682673	1.13	0.2887
BuildingArea	1	2.1595748E12	2.1595748E12	25.77	<.0001
YearBuilt	1	270390942175	270390942175	3.23	0.0725
Lattitude	1	651171876891	651171876891	7.77	0.0053
Longtitude	1	720186921.42	720186921.42	0.01	0.9261
Propertycount	1	20240579374	20240579374	0.24	0.6231
Rooms*Regionname	4	1.5160199E12	379004964886	4.52	0.0012
Type*Regionname	8	8.8448032E12	1.1056004E12	13.19	<.0001
Method*Regionname	16	4.072996E12	254562251011	3.04	<.0001
SellerG*Regionname	117	1.5290401E13	130687186453	1.56	0.0001
Bedroom2*Regionname	4	1.7771178E12	444279445261	5.30	0.0003
Car*Regionname	7	1.4958479E12	213692562547	2.55	0.0128
Lattitude*Regionname	7	7.4447638E12	1.0635377E12	12.69	<.0001
Longtitud*Regionname	6	5.8488472E12	974807873417	11.63	<.0001
Distance*Type	2	1.5866212E13	7.9331058E12	94.68	<.0001
Landsize*Type	2	1.5132781E13	7.5663904E12	90.30	<.0001
YearBuilt*Type	2	1.3058281E12	652914040659	7.79	0.0004
Distance*Method	4	1.6747014E12	418675357937	5.00	0.0005
YearBuilt*Method	4	3.6942572E12	923564288823	11.02	<.0001
Rooms*SellerG	68	2.066199E13	303852799469	3.63	<.0001
Distance*SellerG	72	1.7351165E13	240988400572	2.88	<.0001
Bathroom*SellerG	68	1.3680327E13	201181275556	2.40	<.0001
Car*SellerG	71	1.1578598E13	163078850849	1.95	<.0001
Landsize*SellerG	75	2.2324616E13	297661543700	3.55	<.0001

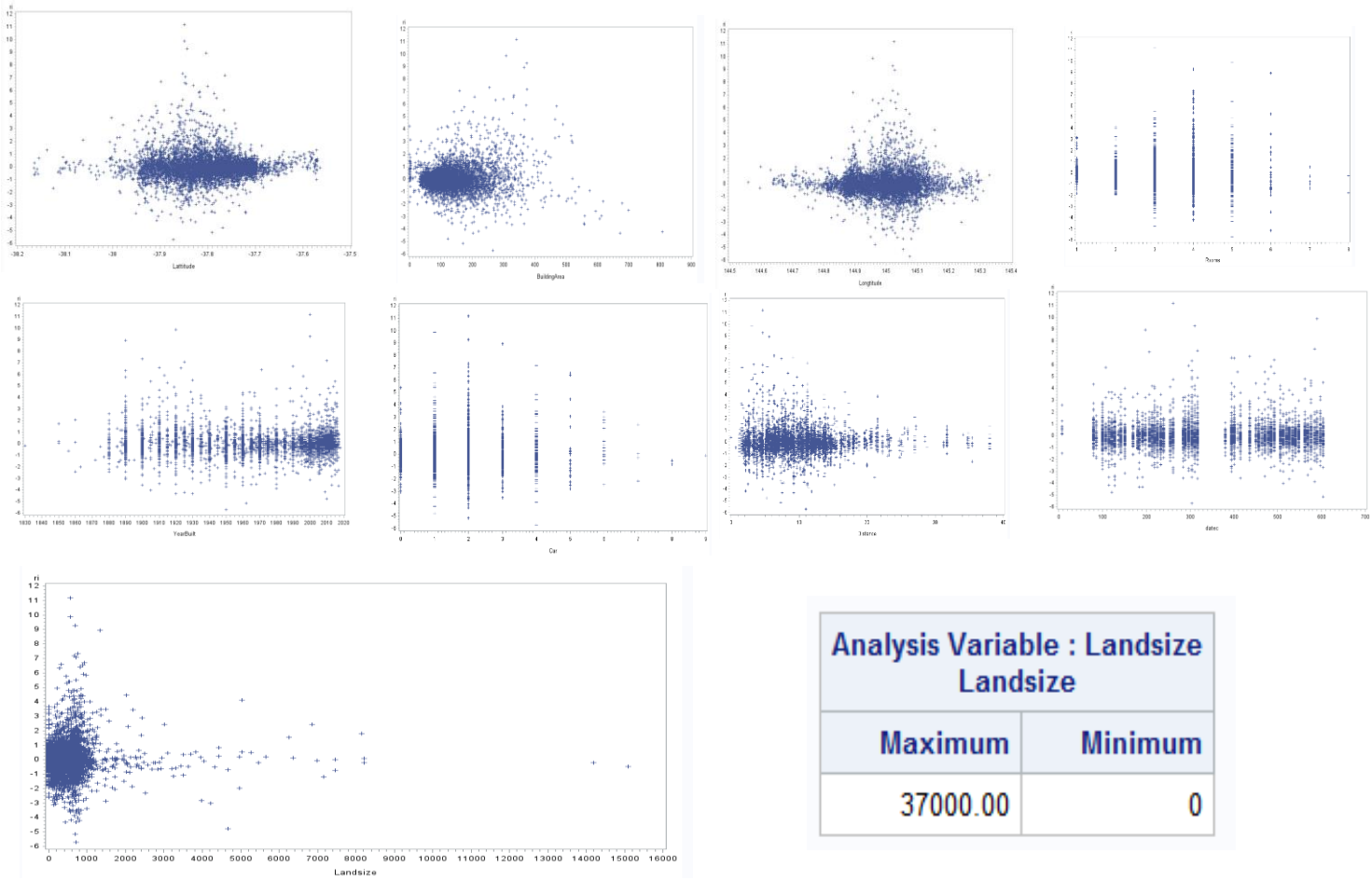
Table4. Final regression result

Rooms	1	8.6239185E14	8.6239185E14	10293.7	<.0001
Type	2	1.5467476E14	7.7337381E13	923.12	<.0001
Method	4	2.7466416E12	686660398749	8.20	<.0001
SellerG	213	3.8582657E14	1.8113924E12	21.62	<.0001
datec	1	232118278646	232118278646	2.77	0.0961
Distance	1	1.1827171E14	1.1827171E14	1411.72	<.0001
Bedroom2	1	1.3953061E12	1.3953061E12	16.65	<.0001
Bathroom	1	7.2953029E13	7.2953029E13	870.78	<.0001
Car	1	1.2171885E13	1.2171885E13	145.29	<.0001
Landsize	1	1.7948492E12	1.7948492E12	21.42	<.0001
BuildingArea	1	5.8683697E13	5.8683697E13	700.46	<.0001
YearBuilt	1	4.4620281E13	4.4620281E13	532.60	<.0001
Lattitude	1	1.0670194E13	1.0670194E13	127.36	<.0001
Longtitude	1	3.6773721E12	3.6773721E12	43.89	<.0001
Rooms*Regionname	7	7.5255322E13	1.075076E13	128.32	<.0001
Type*Regionname	9	1.1067575E13	1.2297305E12	14.68	<.0001
Method*Regionname	24	4.4289869E12	184541122448	2.20	0.0006
SellerG*Regionname	157	6.974777E13	444253309301	5.30	<.0001
Bedroom2*Regionname	4	2.9590121E12	739753026589	8.83	<.0001
Car*Regionname	7	8.2444396E12	1.1777771E12	14.06	<.0001
Lattitude*Regionname	7	7.7189359E12	1.1027051E12	13.16	<.0001
Longtitud*Regionname	6	1.4531852E13	2.4219753E12	28.91	<.0001
Distance*Type	2	2.8246079E13	1.4123039E13	168.58	<.0001
Landsize*Type	2	1.6769728E13	8.3848642E12	100.08	<.0001
YearBuilt*Type	2	1.2443609E12	622180429407	7.43	0.0006
Distance*Method	4	1.494909E12	373727253787	4.46	0.0013
YearBuilt*Method	4	3.4625193E12	865629836779	10.33	<.0001
Rooms*SellerG	113	5.9094303E13	522958432069	6.24	<.0001
Distance*SellerG	112	2.3904088E13	213429353987	2.55	<.0001
Bathroom*SellerG	94	2.5820531E13	274686499232	3.28	<.0001
Car*SellerG	84	2.3415348E13	278754138060	3.33	<.0001
Landsize*SellerG	87	2.932643E13	337085397829	4.02	<.0001
BuildingArea*SellerG	77	3.7844134E13	491482263884	5.87	<.0001
Lattitude*SellerG	68	1.0636502E13	156419152426	1.87	<.0001

After we delete all the insignificant interaction terms and all the insignificant main factors(only one in fact). There are still many interaction terms in the model. We observe that the residual seems not to be normally distributed(see graph1). To further test the normality assumption, we calculate externally studentized residual and internally studentized residual of the model from last step. After that, normality test is performed on externally studentized residual(see table6).

Because the normality test fails at significance level 0.01, we seek to perform transformation. As a result of the presence of interaction terms, it is impossible to use Box-Cox, Box-Tidewell or Spline method by usual method. Instead, we plot residual against each continuous variable to see whether some non-linearlity can be spotted (see graph2). Note that all the residuals are randomly distributed except the residual of landsize. Also note that the maximum of landsize is also 10 times greater than its minimum value. Therefore, we take logarithm of landsize to shrinkage its scale. To avoid the case that landsize is 0, we add one to landsize before taking logarithm. The normality test result after taking logarithm of landsize is given in table7. We see the F-value improves slightly.

Graph2. Residual plot and max&min of landsize



Analysis Variable : Landsize Landsize	
Maximum	Minimum
37000.00	0

Graph1. Diagnosis plot

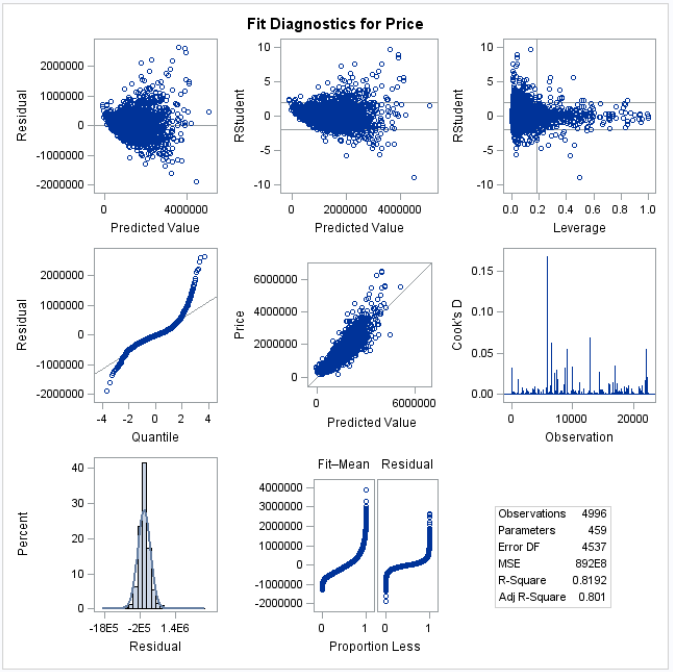


Table6. Normality test on external residual

Tests for Normality				
Test	Statistic		p Value	
Kolmogorov-Smirnov	D	0.110642	Pr > D	<0.0100
Cramer-von Mises	W-Sq	24.47321	Pr > W-Sq	<0.0050
Anderson-Darling	A-Sq	142.3783	Pr > A-Sq	<0.0050

Table7. Normality test after taking logarithm on landsize

Tests for Normality				
Test	Statistic		p Value	
Kolmogorov-Smirnov	D	0.114131	Pr > D	<0.0100
Cramer-von Mises	W-Sq	27.69261	Pr > W-Sq	<0.0050
Anderson-Darling	A-Sq	159.2387	Pr > A-Sq	<0.0050

Again, we need to refit the model after the modification on landsize. Interaction terms are checked again. During the iteration, “bedroom2” and “bathroom” are deleted after their interaction terms being deleted completely.

The same analysis of normality and residual is performed again. However, the residual still does not follow normal distribution. See graph3. The residual has dramatic deviance from normal distribution at two ends.

We then look for outliers and influential points. Externally studentized residuals, leverage, cookd are calculated based on the new model. Outliers are dropped when externally studentized residuals are greater than t(alpha/n) , where alpha is chosen to be 0.01 and n is the sample size. Influential points are deleted when cookd>4/n or the data points lie in the top right corner of the Rstudent-leverage plot.

We find that the normality assumption still does not hold even after deletion of outliers and influential points. We then try to perform cubic spline transformation only on continuous variable terms. After we get the coefficients of different continuous variables, we refit the model with transformed continuous variables, categorical variables and interaction terms(either between two categorical variables or between one categorical variable and one continuous variable without transformation). The coefficients are given in table8. The reason that we do not include categorical variables in the model is for the reason of simplicity. “SellerG” has more than 100 levels. “Regionname” has 8 levels. “Method” has 11 levels. “Type” has 6 levels. Performing cubic spline transformation on each combination of the categorical variables is thus cumbersome and less efficient. Therefore, we try to test whether we can achieve better results by only using transformation on continuous variables(excluding continuous variables in interaction terms).

Graph3. Diagnosis plot

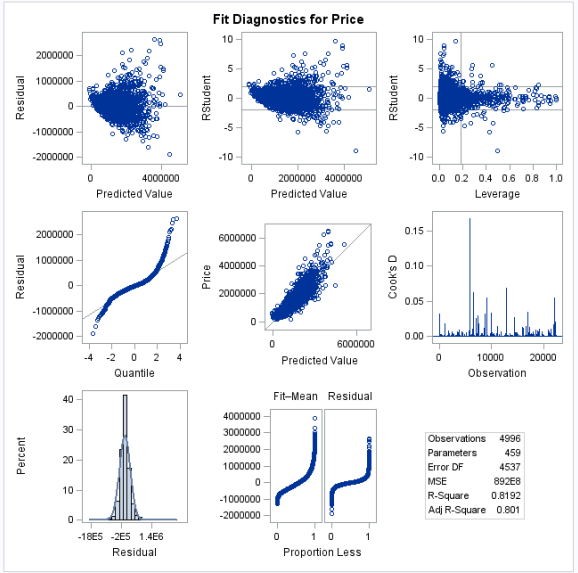


Table8. Coefficients of cubic spline

Univariate Regression Table Based on the Usual Degrees of Freedom									
Variable	DF	Coefficient	95% Confidence Limits		Type II Sum of Squares	Mean Square	F Value	Pr > F	Label
Intercept	1	-281712497	-294363830	-269061164	2.33E14	2.33E14	1905.45	<.0001	Intercept
Spline(Rooms)	3	138090	.	.	4.14E13	1.38E13	112.64	<.0001	Rooms
Spline(datec)	3	237	.	.	8.68E12	2.89E12	23.65	<.0001	datec
Spline(Distance)	3	-22574	.	.	6.67E13	2.22E13	181.79	<.0001	Distance
Spline(Car)	3	54875	.	.	1.13E13	3.78E12	30.88	<.0001	Car
Spline(loglandsize)	3	40121	.	.	2.56E13	8.53E12	69.66	<.0001	loglandsize
Spline(BuildingArea)	3	3680	.	.	2.05E14	6.82E13	557.13	<.0001	BuildingArea
Spline(YearBuilt)	3	-4038	.	.	9.13E13	3.04E13	248.58	<.0001	YearBuilt
Spline(Lattitude)	3	-2045175	.	.	8.19E13	2.73E13	223.12	<.0001	Lattitude
Spline(Longtitude)	3	1464263	.	.	9.33E13	3.11E13	254.20	<.0001	Longtitude

Table9. Regression on partial transformed variables

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	808	2.1898516E15	2.7102123E12	42.72	<.0001
Error	5663	3.5928187E14	63443735418		
Corrected Total	6471	2.5491334E15			

R-Square	Coeff Var	Root MSE	Price Mean
0.859057	23.72978	251880.4	1061453

Source	DF	Type III SS	Mean Square	F Value	Pr > F
TRooms	1	885779835073	885779835073	13.96	0.0002
Tdatec	1	9.0268279E12	9.0268279E12	142.28	<.0001
TDistance	1	1.0226934E12	1.0226934E12	16.12	<.0001
TCar	1	30492328099	30492328099	0.48	0.4882
Tloglandsize	1	4.6502464E12	4.6502464E12	73.30	<.0001
TBuildingArea	1	3.2267425E12	3.2267425E12	50.86	<.0001
TYearBuilt	1	1.819966E13	1.819966E13	286.86	<.0001
TLattitude	1	774539880506	774539880506	12.21	0.0005
TLongtitude	1	2.2412463E12	2.2412463E12	35.33	<.0001
Rooms*Regionname	7	7.2355721E12	1.0336532E12	16.29	<.0001
Regionname*Type	8	7.8179294E12	977241180537	15.40	<.0001
Regionname*Method	20	1.0905057E13	545252873597	8.59	<.0001
Regionname*SellerG	147	1.4760217E13	100409639768	1.58	<.0001
Lattitude*Regionname	7	4.1848772E12	597839604123	9.42	<.0001
Longtitud*Regionname	6	3.3224479E12	553741323417	8.73	<.0001
Distance*Type	2	1.0208061E13	5.1040303E12	80.45	<.0001
loglandsize*Type	3	7.3957252E12	2.4652417E12	38.86	<.0001
Rooms*SellerG	62	8.8725747E12	143106043902	2.26	<.0001
Distance*SellerG	60	8.0174426E12	133624043639	2.11	<.0001
Car*SellerG	62	9.6180669E12	155130110710	2.45	<.0001
BuildingArea*SellerG	67	1.2492055E13	186448585268	2.94	<.0001
Lattitude*SellerG	61	1.1240159E13	184264909477	2.90	<.0001
Longtitude*SellerG	52	6.1846562E12	118935697009	1.87	0.0002

After refitting the model on partial-transformed variables(see table9), we delete the outliers and influential points using the same method mentioned previously.

Notice that we achieve better R-square(0.861386) after using cubic-spline transformation on continuous variables, compared to R^square(0.8467) in the last model. Now, our model can be written as:

Y= intercept+ cubic spline continuous variable + categorical variable + categorical*categorical + continuous*categorical

Finally, we use stepwise selection to reduce the number of variables in the model by using significance level as the entry criteria and using BIC as the staying criteria. In order to reduce as many variables as possible, we set the entry alpha as 0.05 and staying alpha as 0.01. The results are given in the table10. Note that some levels of “sellerG” and “regionname” are deleted. As a consequence, our model are simplified to some extent.

Table10. Final selected model

The GLMSELECT Procedure

Stepwise Selection Summary

Step	Effect Entered	Effect Removed	Number Effects In	Number Parms In	SBC	F Value	Pr > F
0	Intercept		1	1	131313.057	0.00	1.0000
1	BuildingArea*SellerG		2	61	127588.974	110.50	<.0001
2	Regionname*Type		3	76	126413.727	97.90	<.0001
3	Distance*Type		4	79	125670.649	272.83	<.0001
4	loglandsize*Type		5	82	125250.935	152.74	<.0001
5	Lattitude*Regionname		6	87	124986.202	62.29	<.0001
6	Rooms*Regionname		7	92	124804.985	44.92	<.0001
7	YearBuilt*Type		8	95	124646.800	61.18	<.0001
8	Longtitud*Regionname		9	101	124549.101	24.65	<.0001
9	datec		10	102	124395.011	161.88	<.0001
10	Method		11	106	124329.244*	24.66	<.0001

* Optimal Value of Criterion

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value
Model	105	1.521375E15	1.448929E13	181.41
Error	4815	3.845839E14	79872052137	
Corrected Total	4920	1.905959E15		

Root MSE	282616
Dependent Mean	1157879
R-Square	0.7982
Adj R-Sq	0.7938
AIC	128563
AICC	128568
SBC	124329

Class Level Information

Class	Levels	Values
Regionname	6	Eastern Metropolitan Northern Metropolitan Northern Victoria South-Eastern Metropolitan Southern Metropolitan Western Metropolitan
Type	3	h t u
Method	5	PI S SA SP VB
SellerG	60	Alexkarbon Barlow Barry Bells Biggin Brad Buckingham Burnham Buxton C21 Castran Cayzer Chisholm Collins Considine Darren Douglas Edward Eview Fletchers Frank Gary Greg HAR Harcourts Harrington Hodges Jas Jellis Kay LITTLE Love Marshall McDonald McGrath ...

Conclusion

From table10, for the main factors, we can see that the most significant variables in determining price are “date” and “method” since they have the optimal value of criterion in stepwise selection. Moreover, for the interaction terms, we see that “buildingarea*sellerg” and “regionname*type” are also very significant, which is consistent with our common sense. The area of the building, together with the real estate agent, determines the price. In different regions, the price of different types of housing is also significantly different.

To further find the type and region of the highest housing price. We check the regression coefficient at different levels of combination. We see that “Southern Metropolitan u” and “ Southern Metropolitan t “ have the largest t-value of 11.53, indicating that regression coefficients are the most significant. That is to say, the houses of type u and of type t in region Southern Metropolitan are the most expensive in Melbourne.

Besides, we also report that “BuildingArea*SellerG Barry” has t-value of 19.85, indicating that among all the houses sold by Barry, the price increases significantly when building area increases.

We also see that the t-value for “date” is 12.97, indicating a dramatic positive relation between date and price. In other words, the price keeps increasing during the years tested.

Therefore, we can conclude that

1. The price increases significantly during the years tested.
2. The method the house is sold influences housing price significantly
3. In different regions, different types of houses are sold for significantly different price
4. For different sellers, the prices increases significantly with building area