LSTAT2120 - Linear Models

Project : Prediction of Houses Prices based on Melbourne Housing Dataset in Australia

Group members :
Abdouraman - 13291201
Janice Tedogmo - 64672000
Manuelle Ndamtang - 66732000
Email :
abdouraman.abdouraman@uclouvain.be
janice.tedogmo@student.uclouvain.be
manuelle.ndamtang@student.uclouvain.be

Professor :
Christian Hafner
Academic year :
2021-2022

Introduction

The chosen dataset is a csv file from Kaagle. It includes datapoints each representing a house with its characteristics. These houses are located in Australia and the main idea is to create a model able to predict their prices. In all, we have 13580 houses, each with 21 separate features.
Basically, our work will consist in studying the dataset and the different interactions between the variables, testing their significance of some parameters and their possibles combination that might improve models and selecting the most interesting model to predict the price of houses in Australia.

1 Presentation of the Dataset

The dataset on which we decided to carry out our work has 21 variables among which we decided to select 13 to carry out our study:

- `Bathroom`: which would represent the number of bathrooms in a house;
- `Bedroom2`: which would be the number of bedrooms in a house;
- `BuildingArea`: which would represent the area on which a house would be built;
- `Car`: which would be the number of car spaces in a garage that a house could have;
- `Landsize`: which would be the area of the exterior space of a house;
- `Distance`: which would be the distance to a nearby city center;
- `Method`: string being a qualitative variable that would represent the method of selling the house;
- `Price`: our target that gives the price of a house;
- `Propertycount`: the description of this variable remains unknown but we find it interesting to look at it;
- `Regionname`: string that would be another qualitative variable describing the region where a house would be located;
- `Rooms`: which would represent the number of rooms that a house would have;
- `Type`: which would be the type of the house;
- `YearBuilt`: the year of construction of the house.

This dataset contains non-zero values worth a total of 6479. During the whole project, the level of significance would be set at 5%.

1.1 Dataset Splitting

Before starting the in-depth study of the subject, the dataset was divided into two arbitrary parts. The first one was dedicated to testing, representing 10% of the data sizes of the whole dataset and the rest was used for training the dataset.

1.2 Analysis of the variables

For the quantitative variables, we proceeded to a general analysis through data such as the Mean, the standard deviation, the coefficient of variation, the maximum, the minimum, the skewness, the kurtosis as well as the boxplots. The conclusion is that all the selected variables do not have a symmetrical distribution. Apart from the variable Rooms and Bedroom2, which are moderately skewed, all other variables are highly skewed. Only the variable Rooms would tend towards a normal distribution. Figures 2 and 1 show the statistics and the boxplots.

The correlation matrix deduced from the quantitative variables(Figure 3) shows us possible signs of multicollinearity issue. We notice for example that the explanatory variable `Bedroom2` presents a strong correlation with the variable `Rooms`.

As for the qualitative variables, we have determined for each class belonging to it, their percentages in the dataset in order to detect possible influences that some classes would have with respect to our model. What emerges from these statistics for each of the variables is that there are no classes with more than 90% of the houses, which could cover most of the houses and would be useless in the study of the data. Figure 4 illustrates the results obtained.

Another interesting thing to study would be the variation of each variable according to each class of the qualitative variables. We notice a considerable variation of some variables according to each class. For example the houses in the Region `Nothern Victoria` have on average a bigger `LandSize` and a bigger `BuildingArea` than the other regions. And it is the same for the houses of Type `h` and sold by the method `VB` compared to the other types and methods. So the classes for each qualitative variable would have an huge influence in the study of our subject.

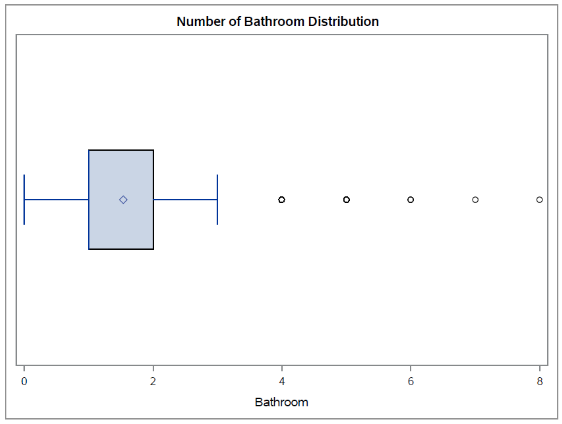
One thing to note is the average `Price` for the houses located in `Southern Metropolitan` very high (mean Price: 1377618.53) compared to other regions and especially the `Western Victoria` region (mean Price: 398663.79) which has the lowest average. And it is the same for the houses with the `VB` method having a high average price (1163393.94) compared to the one using the `SP` method (899847.00, the smallest mean). Such variations could be interpreted as possible signs of heteroskedasticty.

2 Model selection process

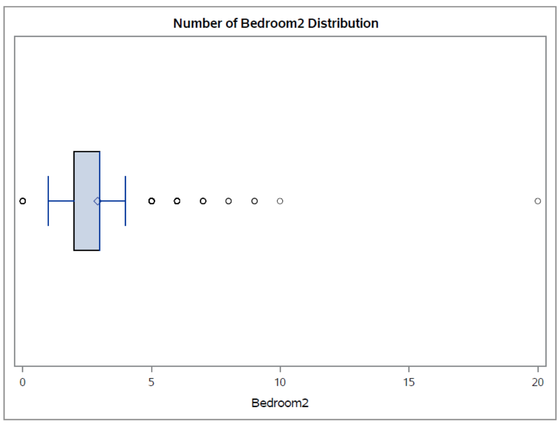
2.1 Multicollinearity handling

Before beginning with this step, we tried to set up a model only made up of the quantitative variables. For this model, we notice that the coefficients of the variables `Bedroom2` and `Propertycount` are not significant based on the p-values which are respectively 0.9132 and 0.3666.

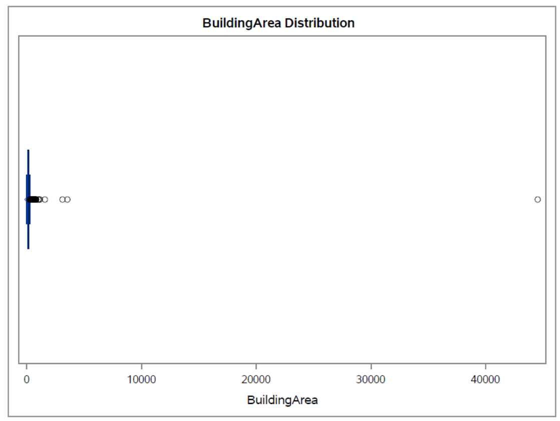
To verify the possible multicollinearity issue which could justify this, VIF was calculated for each quantitative variable. We notice that the VIF of Rooms and `Bedroom2` are indeed higher than 10, being 12.13981 for `Rooms` and 12.03043 for `Bedroom2`. This confirms the existence of a multicollinearity between these two variables. For the rest of the variables, we obtain a VIF close to 1. To counter this, two solutions are possible. The first would be to



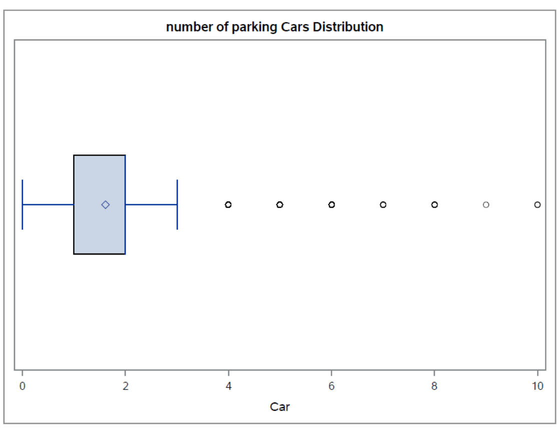
(a) Bathroom



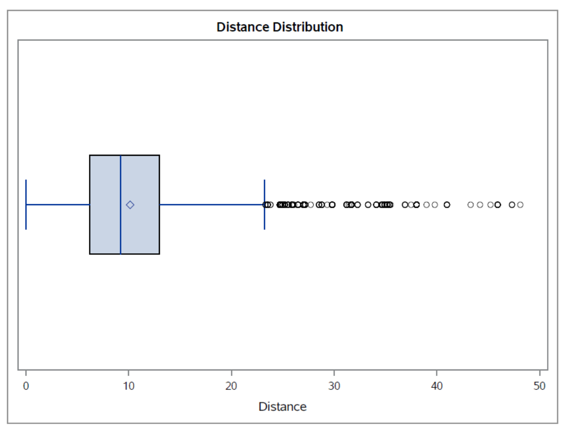
(b) Bedrooms



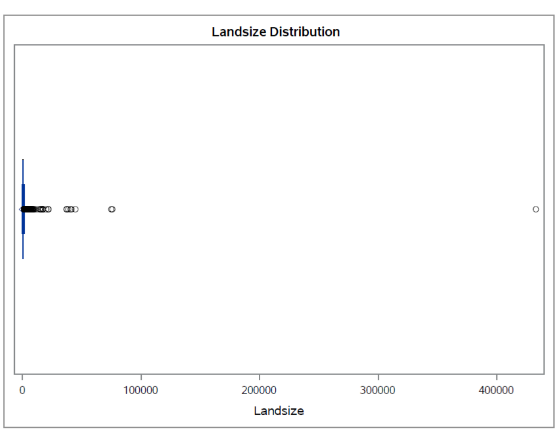
(c) BuildingArea



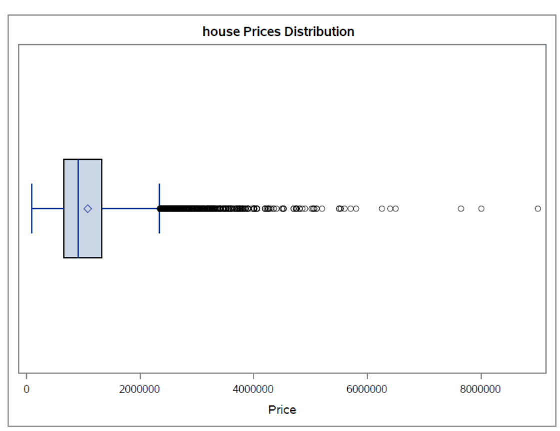
(d) Car



(e) Distance



(f) Landsize



(g) Price

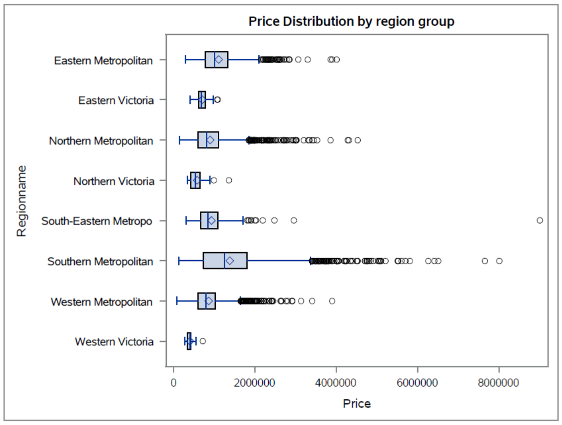
Variable	Moyenne	Ec-type	Coef. de variation	Maximum	Minimum	Skewness	Kurtosis
Price	1075054.46	642430.20	59.7579210	9000000.00	85000.00	2.2731796	10.1872083
Rooms	2.9367534	0.9590294	32.6561090	10.0000000	1.0000000	0.3808547	0.7989307
Bedroom2	2.9131893	0.9691626	33.2680954	20.0000000	0	0.8052136	8.7092476
Bathroom	1.5373098	0.6975524	45.3748778	8.0000000	0	1.4061603	3.8076091
Landsize	566.0330551	4199.06	741.8398739	433014.00	0	90.8232123	9227.30
BuildingArea	151.8377090	562.5709631	370.5080686	44515.00	0	76.2614857	6009.56
Car	1.6114215	0.9627550	59.7456969	10.0000000	0	1.3916482	5.4299604
Distance	10.1303387	5.8649018	57.8944293	48.1000000	0	1.6774828	5.2595692
YearBuilt	1964.82	37.3613682	1.9015192	2018.00	1196.00	-1.6448497	23.3902870
Propertycount	7441.86	4381.35	58.8744155	21650.00	249.0000000	1.0729600	1.2195524

Figure 2: Descriptive statistics of quantitative

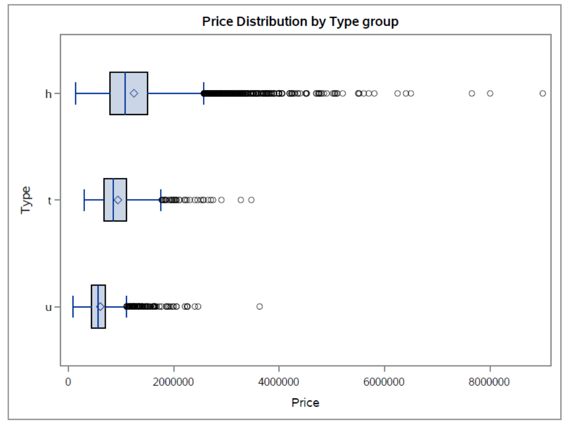
Coefficients de corrélation de Pearson
Proba > |r| sous H0: Rho=0
Nombre d'observations

	Price	Rooms	Bedroom2	Bathroom	Landsize	BuildingArea	Car	Distance	YearBuilt	Propertycount
Price	1.00000	0.49747	0.47665	0.47136	0.03773	0.08843	0.24262	-0.16086	-0.32257	-0.03982
		<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001
	12222	12222	12222	12222	12222	6439	12170	12222	7401	12222
Rooms	0.49747	1.00000	0.94278	0.59523	0.02429	0.12143	0.40842	0.29524	-0.06377	-0.08198
	<.0001		<.0001	<.0001	0.0072	<.0001	<.0001	<.0001	<.0001	<.0001
	12222	12222	12222	12222	12222	6439	12170	12222	7401	12222
Bedroom2	0.47665	0.94278	1.00000	0.58837	0.02439	0.11982	0.40536	0.29747	-0.04974	-0.08231
	<.0001	<.0001		<.0001	0.0070	<.0001	<.0001	<.0001	<.0001	<.0001
	12222	12222	12222	12222	12222	6439	12170	12222	7401	12222
Bathroom	0.47136	0.59523	0.58837	1.00000	0.03735	0.11263	0.32486	0.12701	0.15366	-0.05404
	<.0001	<.0001	<.0001		<.0001	<.0001	<.0001	<.0001	<.0001	<.0001
	12222	12222	12222	12222	12222	6439	12170	12222	7401	12222
Landsize	0.03773	0.02429	0.02439	0.03735	1.00000	0.52661	0.02535	0.02479	0.03767	-0.00761
	<.0001	0.0072	0.0070	<.0001		<.0001	0.0052	0.0061	0.0012	0.3999
	12222	12222	12222	12222	12222	6439	12170	12222	7401	12222
BuildingArea	0.08843	0.12143	0.11982	0.11263	0.52661	1.00000	0.09760	0.10321	0.01352	-0.02763
	<.0001	<.0001	<.0001	<.0001	<.0001		<.0001	<.0001	0.2876	0.0266
	6439	6439	6439	6439	6439	6439	6414	6439	6190	6439
Car	0.24262	0.40842	0.40536	0.32486	0.02535	0.09760	1.00000	0.26116	0.10393	-0.02423
	<.0001	<.0001	<.0001	<.0001	0.0052	<.0001		<.0001	<.0001	0.0075
	12170	12170	12170	12170	12170	6414	12170	12170	7375	12170
Distance	-0.16086	0.29524	0.29747	0.12701	0.02479	0.10321	0.26116	1.00000	0.24358	-0.05364
	<.0001	<.0001	<.0001	<.0001	0.0061	<.0001	<.0001		<.0001	<.0001
	12222	12222	12222	12222	12222	6439	12170	12222	7401	12222
YearBuilt	-0.32257	-0.06377	-0.04974	0.15366	0.03767	0.01352	0.10393	0.24358	1.00000	0.00631
	<.0001	<.0001	<.0001	<.0001	0.0012	0.2876	<.0001	<.0001		0.5874
	7401	7401	7401	7401	7401	6190	7375	7401	7401	7401
Propertycount	-0.03982	-0.08198	-0.08231	-0.05404	-0.00761	-0.02763	-0.02423	-0.05364	0.00631	1.00000
	<.0001	<.0001	<.0001	<.0001	0.3999	0.0266	0.0075	<.0001	0.5874	
	12222	12222	12222	12222	12222	6439	12170	12222	7401	12222

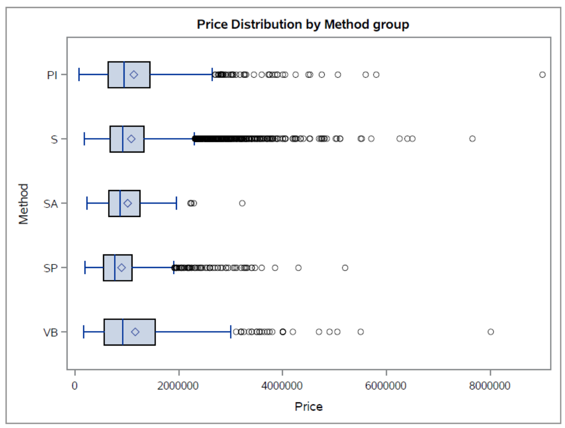
Figure 3: Correlation matrix



(a) Price per Regionname classes



(b) Price per Type classes



(c) Price per Method classes

Figure 4: Residuals for log(Price)

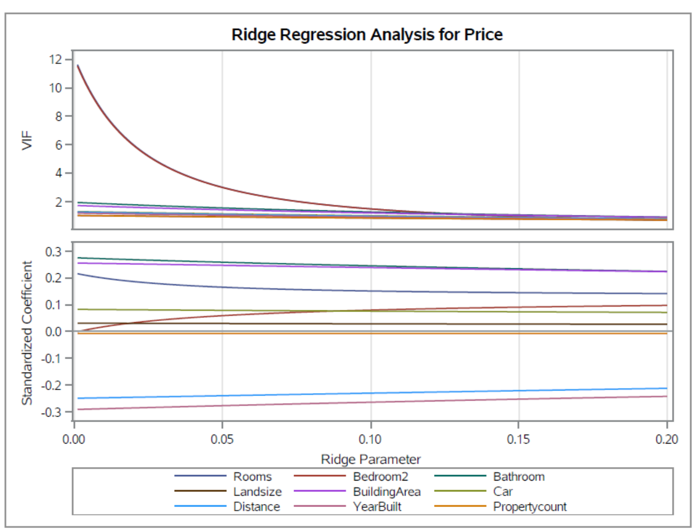


Figure 5: Ridge regression analysis for multicollinearity

use ridge regression and keep all the explanatory variables. The second would be to remove one of the variables with the highest VIF.

We have explored these two methods. After running a Ridge regression for the model, plots the VIF against the Rigde parameter and also obtained the Ridge trace, we notice that Standardized coefficients and VIFs stabilize around a ridge parameter worth 0.10. We can deduce that 0.10 would be a reasonable choice(Figure 5).

For the second method, we preferred to delete the `Bedroom2` variable, because it is already explained by the `Rooms` variable. To solve this multicollinearity issue, we finally opted for the second solution.

2.2 Variables selection process

The qualitative variables were transformed into dummy variables. And after performing a regression taking into account the different classes corresponding to these variables, it appears that some coefficients of the different classes would be non-significant. This is the case for the `Nothern Victoria` region (with a p-value of 0.1771), the `Eastern Metropolitan` region (p-value of 0.8194), the `SA` method (p-value of 0.1485), the `SP` method (p-value of 0.1766) and the `PI` method (p-value of 0.3433). The quantitative variable `Propertycount` is also non-significant with a p-value of 0.8749.

For the selection of the model, we proceeded by three methods: a type 1 variable selection using the Mallows criterion and the same type using the adjusted R-square as criterion, a type 2 variable selection and a type 3 variable selection with Lasso. Starting from the selection of type 1 variables with the Mallows criterion, we selected the 10 best models ordered by importance. The best model includes the 17 explanatory variables such as: `Rooms`, `Bathroom`, `Landsize`, `BuildingArea`, `Car`, `Distance`, `YearBuilt`, `Typeh`, `Typet`, `MethodS`, `MethodSP`, `RegionEV`, `RegionNM`, `RegionNV`, `RegionSEM`, `RegionSM`, `RegionWM`. The same process, this time using the R-square as criterion, still gives us the same variables but with the class corresponding to the selected `SA` method (represented by variable `methodSA`). An interesting thing we could notice is that the variable `Propertycount` is not among them.

To confirm this choice of variables, the selection method of type 2 has also been realized. In addition to the variables mentioned before, we notice that by performing the forward selection, we obtain the same variables as those mentioned for the type 1 variables section using the Mallows criterion. We also obtain the same result by performing the backward selection. Using Lasso for the selection of the variables, we have instead 20 explanatory variables selected(figure 6).

Following the results obtained through the different methods, our choice of explanatory variables is the one obtained using the Mallows criterion, given the fact that most of the methods used before present this result.

2.3 Heteroskedasticty and normality checking

We analyzed the residuals to detect whether there was homoskedasticity and normality in the data. When checking the normality of the error terms using the residuals histogram and the QQ-Plot, we notice that the distribution is not normal. There is a strong asymmetry of the

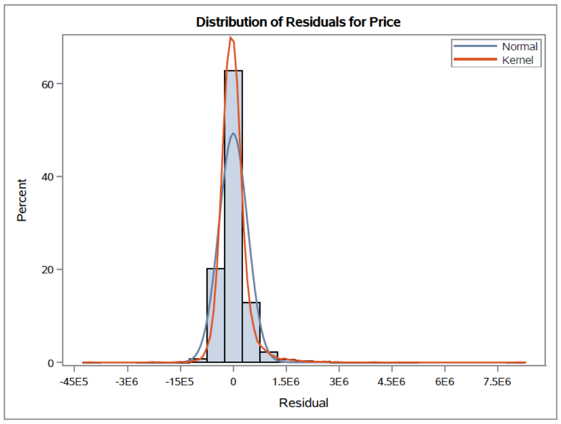
LASSO Selection Summary				
Step	Effect Entered	Effect Removed	Number Effects In	CVEX PRESS
0	Intercept		1	4.56131E11
1	BuildingArea		2	4.49922E11
2	Rooms		3	4.12924E11
3	Bathroom		4	3.67593E11
4	Regionname_SM		5	3.48206E11
5	YearBuilt		6	3.09696E11
6	Type_h		7	2.64091E11
7	Distance		8	2.4772E11
8	Type_u		9	2.00919E11
9	Regionname_WM		10	1.93517E11
10	Car		11	1.84153E11
11	Regionname_NM		12	1.81463E11
12	Regionname_SEM		13	1.77524E11
13	Landsize		14	1.77133E11
14	Method_S		15	1.73785E11
15	Method_PI		16	1.73084E11
16	Regionname_EV		17	1.72493E11
17	Method_SA		18	1.71634E11
18	Regionname_NV		19	1.71106E11
19	Method_VB		20	1.70649E11
20	Regionname_WV		21	1.70646E11*
21	Propertycount		22	1.70647E11
* Optimal Value of Criterion				

Figure 6: Variables selection result using Lasso

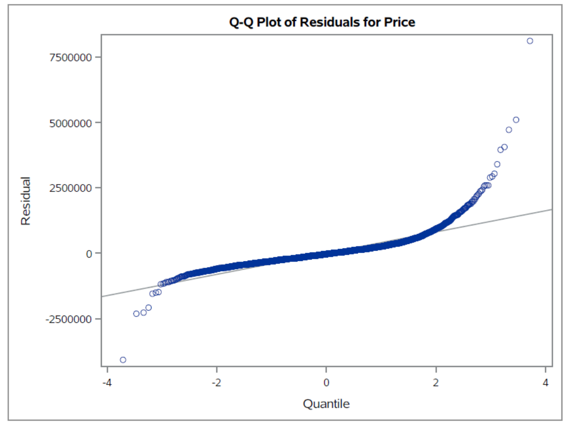
error distribution.(Figure 7) To confirm this, we performed the Jarque-Bera test. Indeed, the p-value is less than 5%, so the null hypothesis given that errors are normally distributed is rejected.(Figure8)

A plot of the residuals against the predicted prices shows an exponential increase in the variance of the error terms(Figure 9). From this, we could conclude that there is indeed heteroskedasticty in the error terms. After running the white test under the null hypothesis that the residuals are homoskedastic, we found that the p-value is less than 5%, so we reject the null hypothesis that errors are homoskedastic(Figure 10).

Assuming that there is heteroskedasticty in the error terms, an interesting solution would be to transform our target value into a logarithm. Instead of predicting house prices, our model should predict the logarithm of house prices instead. One could justify this choice because of the observed exponential increase in residuals as a function of the predicted values. Once this modification is made, we notice a clear improvement. The residuals seem to have constant variance over the predicted values showing the homoskedasticity (Figure 11). To



(a) Distribution of Residuals for Price



(b) Q-Q Plot of Residuals for Price

Figure 7: Residuals for Price

The AUTOREG Procedure

Ordinary Least Squares Estimates			
SSE	1.01103E15	**DFE**	6148
MSE	1.64448E11	**Root MSE**	405522
SBC	176879.654	**AIC**	176758.571
MAE	259252.031	**AICC**	176758.682
MAPE	28.000951	**HQC**	176800.562
Durbin-Watson	1.7081	**Total R-Square**	0.6411

Miscellaneous Statistics			
Statistic	**Value**	**Prob**	**Label**
Normal Test	501639.847	<.0001	Pr > ChiSq

Figure 8: Jarque-Bera's test result

confirm this, we ran the white test under the null hypothesis that there is homoskedasticity in the errors terms. But for some reasons, the hypothesis is rejected(Figure 12). To fix this persistent heteroskedasticty issues, two solutions can be used: either we can proceed to weighted least squares by estimating the variance from the auxilary white model either continue to use the OLS method by way of robust inference statistics. This will reduce the underestimation of the variance generated by the OLS method. In the following we will adopt the second solution, which consists of using OLS with robust inference. However, Region `Nothern Victoria` is no longer indivudually significant in this transformed model.

We also notice a considerable improvement in the distribution of the error terms, which now appears symmetrical and normally distributed. However, To confirm this we run the Jarque-Bera test, but for some reason, this hypothesis is rejected (Figure 14).

Since Region `Nothern Victoria` is no longer individually significant in this transformed model, we have proceed to variables selection from the new model with forward selection.

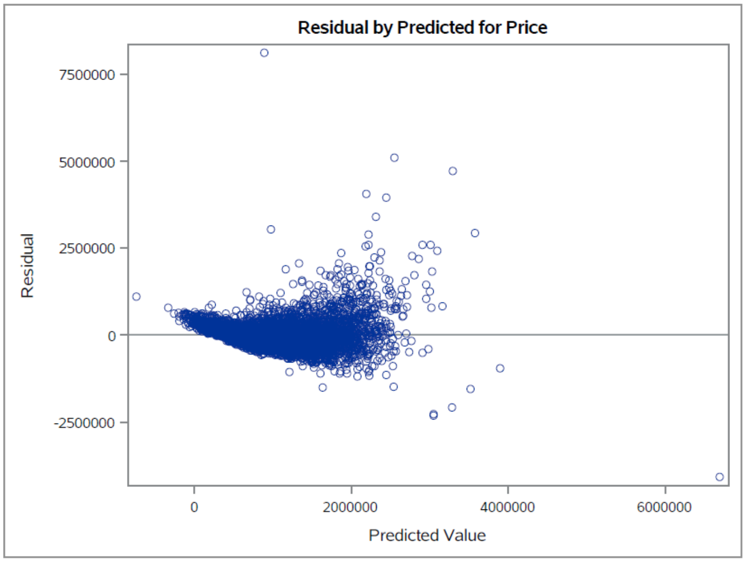


Figure 9: Residual by Predicted for Price

Number of Observations		Statistics for System	
Used	6166	Objective	1.6397E11
Missing	6056	Objective*N	1.011E15

Heteroscedasticity Test					
Equation	Test	Statistic	DF	Pr > ChiSq	Variables
Price	White's Test	1124	139	<.0001	Cross of all vars

Figure 10: White test on untransformed dataset

The result is that the variable RegionNV is indeed no longer in our model. Our final model with robust inference, is the one shown on figure 15. It is worth to precise that the estimated parameters don't change compared to the model without robust inference but the standard error increases as can be shown on figure 15. That give as consequence the p-value increases and the t-value decreases compared to the non-robust model.

2.4 Autocorrelation

Since we are not dealing with time series, there is not autocorrelation issue to be tested here.

2.5 Outliers and influential observations management

For the detection of outliers, we proceeded to two methods: outliers observation in the explanatory variables by calculating the leverage and at 5% level of significance using the Studentized residuals for outliers of the Price our dependent variable.

For outliers observation with respect to independant explanatory variables, the calculation of the leverage of each datapoint was performed. Then we considered as outliers, the houses

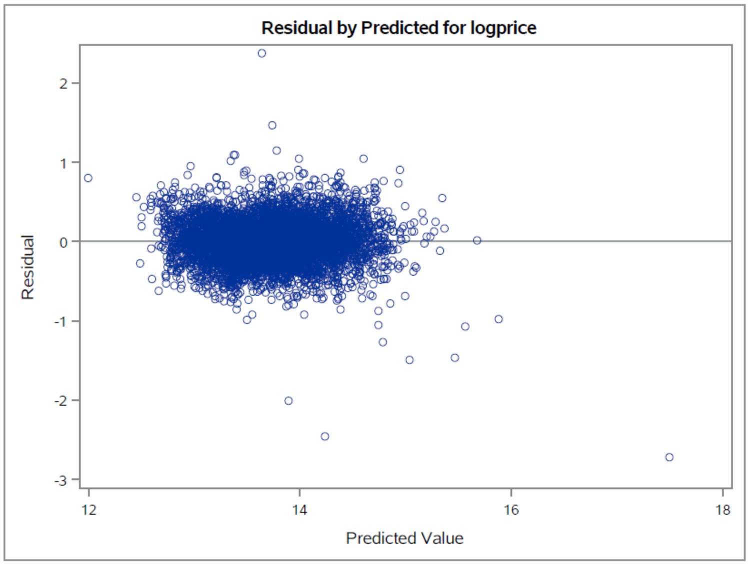


Figure 11: Residual by Predicted for log(Price)

Number of Observations		Statistics for System	
Used	6166	**Objective**	0.0740
Missing	6056	**Objective*N**	456.4338

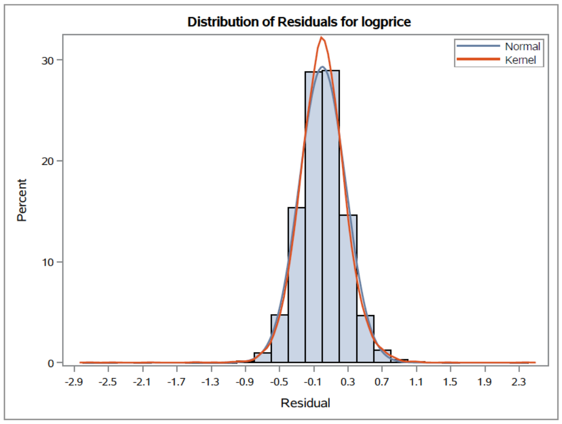
Heteroscedasticity Test					
Equation	**Test**	**Statistic**	**DF**	**Pr > ChiSq**	**Variables**
logprice	White's Test	2166	139	<.0001	Cross of all vars

Figure 12: White test on transformed dataset

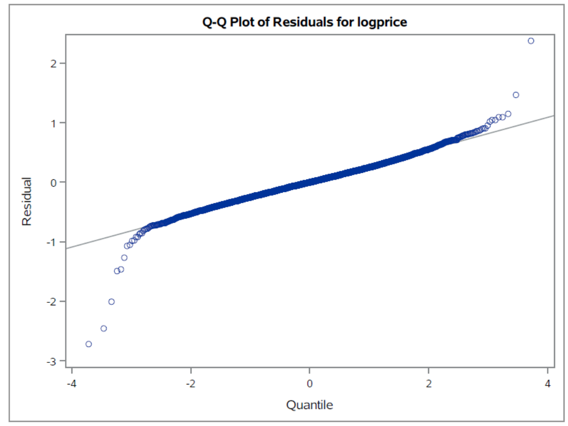
whose leverage would be higher than $2 * \frac{17}{6166}$ with 17 the number of coefficients and 6166 the size of the dataset. For outliers observation with respect to dependant target value Price, we determined instead the Studentized residuals for each datapoint at 5% significance level. What were considered outliers were datapoints having Studentized residuals averaging greater than the 0.975-th quantile from he Student's t distribution with degrees of freedom $61669 - 17 - 1$. We decided not to remove the outliers because this would also remove important information for our model.

The detection of influential observations was done in order to find explanatory variables as well as the dependent variable that could be outliers for a particular datapoint. We calculated the DFFITS for each dwelling and determined which ones were greater than $2 * \sqrt{\frac{17}{6166}}$ or less than $-2 * \sqrt{\frac{17}{6166}}$.

We then determined the influential observations using the Cook's distance(Figure 16). These observations are not only influential on a particular predicted value and a specific coefficient belonging to a variable, but on all predictions and all coefficients. This shows that observation 1410 is one of them.We decide to keep these influential observations.



(a) Distribution of Residuals for log(Price)



(b) Q-Q Plot of Residuals for log(Price)

Figure 13: Residuals for log(Price)

The AUTOREG Procedure

Ordinary Least Squares Estimates			
SSE	456.433825	**DFE**	6148
MSE	0.07424	**Root MSE**	0.27247
SBC	1603.10314	**AIC**	1482.02063
MAE	0.20675674	**AICC**	1482.13191
MAPE	1.50719469	**HQC**	1524.01101
Durbin-Watson	1.5975	**Total R-Square**	0.7495

Miscellaneous Statistics			
Statistic	**Value**	**Prob**	**Label**
Normal Test	5773.5741	<.0001	Pr > ChiSq

Figure 14: Jarque-Bera's test result on transformed data

Let's nevertheless mention that to handle with influential observations, one solution might be either add other explanatory variables, or add quadratic terms or add interaction terms. Another radical one could be to use a robust estimation method such as minimising the absolute residuals rather than the square residuals. To make things easier, we won't neither use the first option (because this require to find out other explanatory variables that we do not have) nor the second (we won't try to add all quadratic terms to see which one is significant) nor the radical solution (because it's very complex to minimise absolute residuals since there is no close form solution). But rather we will add some interaction terms in the following session (as this is the purpose of this question) without unfortunately checking further again if those influential observations are still influential.

2.6 Interaction

Regarding possible interactions of categorical variables, we tried to see the averages of the variables by class. A class that had a higher average of a variable than the other classes gave

Parameter Estimates									
							Heteroscedasticity Consistent		
Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	17.37080	0.22750	76.35	<.0001	0.48985	35.46	<.0001
Rooms		1	0.13774	0.00584	23.58	<.0001	0.01180	11.67	<.0001
Bathroom		1	0.12403	0.00673	18.43	<.0001	0.01304	9.51	<.0001
Landsize		1	0.00001378	0.00000384	3.59	0.0003	0.00000496	2.78	0.0055
BuildingArea		1	0.00198	0.00012323	16.06	<.0001	0.00087102	2.27	0.0231
Car		1	0.04243	0.00419	10.13	<.0001	0.00513	8.27	<.0001
Distance		1	-0.03888	0.00082214	-47.29	<.0001	0.00107	-36.49	<.0001
YearBuilt		1	-0.00227	0.00011594	-19.55	<.0001	0.00024487	-9.26	<.0001
Typeh	Type h	1	0.56583	0.01679	33.70	<.0001	0.07808	7.25	<.0001
Typet	Type t	1	-2.16332	1.11722	-1.94	0.0529	1.22622	-1.76	0.0777
MethodS	Method S	1	0.09752	0.00877	11.12	<.0001	0.00997	9.78	<.0001
MethodSP	Method SP	1	0.06256	0.01194	5.24	<.0001	0.01243	5.03	<.0001
RegionEV	Regionname EV	1	0.30584	0.05828	5.25	<.0001	0.05329	5.74	<.0001
RegionNM	Regionname NM	1	-0.26245	0.01354	-19.38	<.0001	0.01461	-17.97	<.0001
RegionSEM	Regionname SEM	1	0.20303	0.02311	8.79	<.0001	0.02832	7.17	<.0001
RegionSM	Regionname SM	1	0.11280	0.01308	8.62	<.0001	0.01439	7.84	<.0001
RegionWM	Regionname WM	1	-0.31257	0.01350	-23.15	<.0001	0.01440	-21.70	<.0001

Figure 15: Model estimation with robust inference

an indication of a possible interaction. So we think that the variable `BuildingArea` has more effect on the price for the houses of type `h`. In the same way, the variable `YearBuilt` has more effet on the houses of type `t`. We will test the following interactions: `typeh*BuildingArea` and `typet*YearBuilt` in order to confirm our assumptions. Indeed, from result given by figure 17, the estimated associated to `tt_year`(`typet*YearBuilt`) is significant, although the coefficient associated to `th_Build` (`typeh*BuildingArea`) is not. But this latest interaction variable has been kept by the variable selection procedure.

3 Test on the obtained model

Let's recall that all the hypothesis test to be correct have been done using the robust inference.

3.1 Test for significance of the estimated coefficients

We tested the null hypothesis that all coefficients are zero. With a P-value lower than 5% and the Fisher value (1042.82) higher than the corresponding critical value $f_{0.025,17,6147}$, we

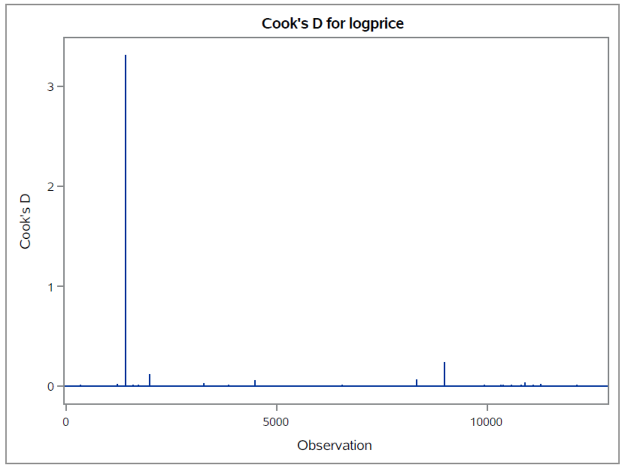


Figure 16: Cook's D for log(Price)

reject this hypothesis(Figure 18).

The numeric variables `Distance` and `YearBuilt` have negative coefficients. So increasing one of that variable will tend to decrease the price. Whereas numeric variables `Rooms`, `Bathroom`, `Landsize`, `BuildingArea` and `Car` have positive coefficients. So increasing one of that variables, will tend to increase the `Price`.

Concerning the qualitative variables, those who have the negative coefficients as Type `t` and Regionname `Northern Metropolitan`, Regionname `Western Metropolitan` have lower price than the level of reference. But the qualitative variables such as Method `S` and `SP`, Type `h`, Regionname `Eastern Victoria`, `South-Eastern Metropolitan` and `Southern Metropolitan` have positive estimated coefficients. So the house having these features will have higher price compared to the reference price for this model.

Since we deal with log-level model, where the logarithm of price (`P`) is explained by a certain number of explanatory variable `X`, the marginal effect with respect to `X` is given by: $\beta * E(P|X)$ where β is the corresponding estimated coefficient for a given `X`. So if the number of rooms increases by one unit(everything else remained constant), the price will increase of $0.13774 * P$. The same explanation holds for all the other numeric variables with positive coefficients.

However the increasing of the variable distance by one unit(everything else remained constant) decrease the price of the house of $-0.03888 * P$. The same explanation holds for all the other numeric variables with negative coefficients.

A house located in Regionname `Western Metropolitan` will have $-0.31527 * P$ lower than a house located in reference region(everything else remained constant). While a house of `h` type will have $0.56583 * P$ higher than a house with the reference type (everything else remained constant).

Given that the variable `YearBuilt` contains an interaction term with type `t` variable(significant coefficient), the increasing of the variable `YearBuilt` by one unit(everything else remained constant) decreases the price of the house not by $-0.00227 * P$ but by $(-0.00227 + 0.00122) * P$

where 0.00122 is the estimated parameter of the interaction variable between `YearBuilt` and type `t`.

3.2 Test of linear combination of at least two coefficients

We have the null hypothesis that the increase in rooms and the increase in bathrooms would have the same influence on the price of housing. Interestingly, this hypothesis is not rejected because the p-value is higher than our significance level. So we can conclude that the increase of a room or a bathroom would lead to the same price increase(Figure 19).

3.3 Test of subset of coefficients of qualitative variables

We also tested the null hypothesis that all the coefficients of the categorical variables in our model would be zero. We notice a p-value lower than our significance level. So we can reject the null hypothesis, all categorical variables are jointly significant(Figure 20).

Parameter Estimates									
							Heteroscedasticity Consistent		
Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Standard Error	t Value	Pr > |t|
th_Build		1	-0.00119	0.00012789	-9.29	<.0001	0.00084469	-1.41	0.1596
tt_year		1	0.00122	0.00055909	2.18	0.0297	0.00062285	1.95	0.0509

Figure 17: Parameters estimated of the interaction variables

Test 1 Results for Dependent Variable logprice				
Source	DF	Mean Square	F Value	Pr > F
Numerator	18	76.25641	1042.82	<.0001
Denominator	6147	0.07312		

(a) Results for Dependent Variable logprice

Test 1 Results using Heteroscedasticity Consistent Covariance Estimates		
DF	Chi-Square	Pr > ChiSq
18	15818.2	<.0001

(b) Heteroscedasticity Consistent Covariance

Figure 18: Residuals for log(Price)

4 Predictions for the observations

The figure 21 illustrated the 20 first results obtained after applying the model on the testset. The column `logprice_Obs` presents the logarithm of the observation price. The column `predicted` give the predicted value. The `lower_value` and `upper_value` give the the lower and the upperbound of the prediction interval for each observation.
We observe that all the prediction observations of the transformed variable `logprice_Obs` are indeed in the confidence interval as illustrated by these 20 observations.

Test 1 Results for Dependent Variable logprice				
Source	DF	Mean Square	F Value	Pr > F
Numerator	1	0.12200	1.67	0.1965
Denominator	6147	0.07312		

(a) Results for Dependent Variable logprice

Test 1 Results using Heteroscedasticity Consistent Covariance Estimates		
DF	Chi-Square	Pr > ChiSq
1	0.80	0.3701

(b) Heteroscedasticity Consistent Covariance

Figure 19: Residuals for log(Price)

Test 1 Results for Dependent Variable logprice				
Source	DF	Mean Square	F Value	Pr > F
Numerator	1	18.59173	254.25	<.0001
Denominator	6147	0.07312		

(a) Results for Dependent Variable logprice

Test 1 Results using Heteroscedasticity Consistent Covariance Estimates		
DF	Chi-Square	Pr > ChiSq
1	5.09	0.0241

(b) Heteroscedasticity Consistent Covariance

Figure 20: Residuals for log(Price)

Conclusion

At the end of our analysis, we can say that we have achieved our goal to create a model to predict prices for houses in Melbourne, Australia. Overall, our model consisting of predicting the logarithm of house prices as a function of some explanatory variables is a linear regression, from which we have verified some classical assumptions such as: independence, normality and homoscedasticity of the error terms.
With this model, we show that it is possible to well predict the logarithm of price of a new house according to its characteristics (number of rooms, bathrooms, size of the land, year of construction, build area, region's name, building area, number of place in parking, method of buying, distance with the town, the type of the habitation) located in Melbourne, with fairly precise intervals and an interesting average coefficient of determination .

Obs	Selected	Price	logprice	logprice_Obs	predicted	lower_pred	upper_pred
7	1	1876000	.	14.4447	14.1640	13.6300	14.6981
17	1	1200000	.	13.9978	14.1947	13.6606	14.7288
18	1	1176500	.	13.9781	14.0430	13.5090	14.5769
20	1	890000	.	13.6990	13.6637	13.1296	14.1977
36	1	1195000	.	13.9937	13.9052	13.3709	14.4395
39	1	840000	.	13.6412	13.3792	12.8432	13.9151
89	1	2120000	.	14.5669	14.2304	13.6964	14.7644
125	1	2840000	.	14.8593	14.6831	14.1481	15.2181
135	1	390000	.	12.8739	12.9742	12.4401	13.5083
145	1	1120000	.	13.9288	13.5417	13.0078	14.0756
167	1	447000	.	13.0103	12.7149	12.1809	13.2490
180	1	857000	.	13.6612	13.3965	12.8621	13.9309
187	1	1085000	.	13.8971	14.3553	13.8204	14.8902
190	1	421000	.	12.9504	13.0447	12.5099	13.5795
195	1	588000	.	13.2845	13.5032	12.9693	14.0371
208	1	620000	.	13.3375	13.5310	12.9968	14.0652
218	1	1200000	.	13.9978	13.7910	13.2565	14.3254
225	1	1205000	.	14.0020	14.0230	13.4887	14.5574
267	1	710000	.	13.4730	13.4979	12.9639	14.0320
273	1	3625000	.	15.1034	14.5540	14.0199	15.0881

Figure 21: The 20 first predictions on the observations

Annexes

Links

Dataset: `https://www.kaggle.com/peterkmutua/housing-dataset`

Code

```
ods graphics on;
ods pdf file="/home/u59968750/Project/sorties_projetLSTAT2120.pdf";
libname modlin "/home/u59968750/Project/";
/*creating data set in worklibrary from the dataset of the libname*/
data melbourne_housing;
        set modlin.melbourne_housing;
run;
/*visualizing dataset*/
proc contents data=melbourne_housing; run;
/*Question 1*/
```

```
/*splitting dataset randomly in 2 groups*/
/*predicting dataset 10% and estimating dataset 90%*/
proc surveyselect data=melbourne_housing samprate=0.10 seed=2021 out=Sample
        outall method=srs noprint;
run;
/*  predicting dataset*/
data melbourne_housing_pred (drop=Selected);
        set Sample;
        where Selected=1;
run;
/*estimating dataset*/
data melbourne_housing_est (drop=Selected);
        set Sample;
        where Selected=0;
run;
/*Question2*/
/*Question3*/
/**Quantitatives variables*/
/***Statistics*/
proc means data=melbourne_housing_est mean std CV max min skew kurt;
        var Price Rooms Bedroom2 Bathroom Landsize BuildingArea Car Distance
        YearBuilt Propertycount;
run;
/***boxplot*/
proc sgplot data=melbourne_housing_est;
        title "house Prices Distribution"; hbox Price;
run;
proc sgplot data=melbourne_housing_est;
        title "Number of Rooms Distribution"; hbox Rooms;
run;
proc sgplot data=melbourne_housing_est;
        title "Number of Bedroom2 Distribution"; hbox Bedroom2;
run;
proc sgplot data=melbourne_housing_est;
        title "Number of Bathroom Distribution"; hbox Bathroom;
run;
proc sgplot data=melbourne_housing_est;
        title "Landsize Distribution"; hbox Landsize;
run;
proc sgplot data=melbourne_housing_est;
        title "BuildingArea Distribution"; hbox BuildingArea;
run;
proc sgplot data=melbourne_housing_est;
        title "number of parking Cars Distribution"; hbox Car;
run;
proc sgplot data=melbourne_housing_est;
        title "Distance Distribution"; hbox Distance;
run;
proc sgplot data=melbourne_housing_est;
```

```
        title "Year of Built Distribution"; hbox YearBuilt;
run;
proc sgplot data=melbourne_housing_est;
        title "Propertycount Distribution"; hbox Propertycount;
run;
/***Correlation matrix*/
proc corr data=melbourne_housing_est;
        var Price Rooms Bedroom2 Bathroom Landsize BuildingArea Car Distance
        YearBuilt Propertycount;
run;
*conclusion: room is highly correlated with bedrooms(we suspect multicollinearity;
/**qualitatives variables*/
proc freq data=melbourne_housing_est;
        tables Type Method Regionname;
run;
/**Quantitatives*qualitatives variables*/
/**Statistics*/
proc sort data=melbourne_housing_est out=data_by_Type;
        by Type;
run;
proc means data=data_by_Type mean std CV max min skew kurt;
        var Price Rooms Bedroom2 Bathroom Landsize BuildingArea Car Distance
        YearBuilt Propertycount;
        by Type;
run;
/*higher mean in type=h cathegories for Landsize and BuildingArea
can be a good candidat for interaction term: typeh*Landsize and typeh*BuildingArea*/
/*type t seems to be specialist of selling recent houses and t for older houses
interaction term typet*year*/
proc sort data=melbourne_housing_est out=data_by_Method;
        by Method;
run;
proc means data=data_by_Method mean std CV max min skew kurt;
        var Price Rooms Bedroom2 Bathroom Landsize BuildingArea Car Distance
        YearBuilt Propertycount;
        by Method;
run;
proc sort data=melbourne_housing_est out=data_by_Regionname;
        by Regionname;
run;
proc means data=data_by_Regionname mean std CV max min skew kurt;
        var Price Rooms Bedroom2 Bathroom Landsize BuildingArea Car Distance
        YearBuilt Propertycount;
        by Regionname;
run;
/*Region Nothern victoria has high mean for landsize and building area
interaction term RegionNV*landsize et RegionNV*buildingarea
Overall the house prices seems to have different variance and mean by group of
qualitatives variables namely for variable region. this can be a sign of
```

```
heteroscedasticity*/
/***Boxplot*/
proc sgplot data=melbourne_housing_est;
        title "Price Distribution by Type group"; hbox Price / category=Type;
run;
proc sgplot data=melbourne_housing_est;
        title "Price Distribution by Method group"; hbox Price / category=Method;
run;
proc sgplot data=melbourne_housing_est;
        title "Price Distribution by region group"; hbox Price / category=Regionname;
run;
/*Question4*/
/*model with only quantitatives variables*/
proc reg data=melbourne_housing_est plots=none;
        model Price=Rooms Bedroom2 Bathroom Landsize BuildingArea Car Distance
                YearBuilt propertycount;
run;
/*conclusion: Bedroom2 and propertycount are individually not significant*/
/*may be Bedroom is not significant because of multicollinearity stated above*/
/*let's calculate VIF to confirm multicollinearity*/
proc reg data=melbourne_housing_est plots=none;
        model Price=Rooms Bedroom2 Bathroom Landsize BuildingArea Car Distance
                YearBuilt propertycount/vif;
run;
/*Indeed Rooms and bedrooms have VIF larger than 10 */
/*we can conclude that there is a multicollinearity issue*/
/*We can either use ridge regression or (keeping all the involved variable)*/
/*or simply  remove one of the variables way to deal with multicollinearity.*/
/*with large VIF.*/
/*just for illustration we run ridgeregression to see how it works*/
proc reg data=melbourne_housing_est outest=data_ridge plots(only)=ridge;
        model Price=Rooms Bedroom2 Bathroom Landsize BuildingArea Car Distance
                YearBuilt propertycount/ridge=(0.001 to 0.2 by 0.001);
run;
/*Both VIF and standard coefficient stabilised aroung ridge_parameter =0.15*/
/*but In this project we adopt the second strategy and remove Bedrooms
from the model for what follows*/
/*model with all quantitatives (except Bedroom2) and qualitative variables*/
/**lets first renomme modalities for Regioname to shorten them*/
data melbourne_housing_est2;
        set melbourne_housing_est;
        if Regionname="Eastern Metropolitan" then Regionname="EM";
        if Regionname="Eastern Victoria" then Regionname="EV";
        if Regionname="Northern Metropolitan" then Regionname="NM";
        if Regionname="Northern Victoria" then Regionname="NV";
        if Regionname="South-Eastern Metropolitan" then Regionname="SEM";
        if Regionname="Southern Metropolitan" then Regionname="SM";
        if Regionname="Western Metropolitan" then Regionname="WM";
        if Regionname="Western Victoria" then Regionname="WV";
```

```
run;
proc transreg data=melbourne_housing_est2 plots=none;
        model identity(Price)=identity(Rooms Bathroom Landsize BuildingArea Car
                Distance YearBuilt Propertycount) class(Type Method Regionname)/ss2;
        output out=data_transformed;
run;
/*conclusion: propertycount and some dumy variable are not significant*/
/*variable selection type 1 using cp as criteria*/
proc reg data=data_transformed plots=none;
        model Price=Rooms Bathroom Landsize BuildingArea Car Distance YearBuilt
        propertycount Typeh Typet MethodPI MethodS MethodSA MethodSP RegionEM
        RegionEV RegionNM RegionNV RegionSEM RegionSM RegionWM/selection=cp best=10;
run;
/*variable selection type 1 using adjRsq as criteria*/
proc reg data=data_transformed plots=none;
        model Price=Rooms Bathroom Landsize BuildingArea Car Distance YearBuilt
        propertycount Typeh Typet MethodPI MethodS MethodSA MethodSP RegionEM
        RegionEV RegionNM RegionNV RegionSEM RegionSM RegionWM/selection=adjrsq
        best=10;
run;
/*en plus des variables choisis précedement methodSA est aussi choisie*/
/*variable selection type2 stepwise*/
proc reg data=data_transformed outest=selected_modelforward tableout plots=none;
        model Price=Rooms Bathroom Landsize BuildingArea Car Distance YearBuilt
        propertycount Typeh Typet MethodPI MethodS MethodSA MethodSP RegionEM
        RegionEV RegionNM RegionNV RegionSEM RegionSM RegionWM/noprint
        selection=stepwise;
run;
proc print data=selected_modelforward; run;
/*variable selection type2 stepwise*/
proc reg data=data_transformed outest=selected_modelbackward tableout plots=none;
        model Price=Rooms Bathroom Landsize BuildingArea Car Distance YearBuilt
        propertycount Typeh Typet MethodPI MethodS MethodSA MethodSP RegionEM
        RegionEV RegionNM RegionNV RegionSEM RegionSM RegionWM/noprint
        selection=backward;
run;
proc print data=selected_modelbackward; run;
/*Type2 meme variable choisis que type1 avec cp comme critère*/
/*variable selection type3 Lasso*/
proc glmselect data=melbourne_housing_est2 plots(stepaxis=normb)=all seed=123;
        class Type Method Regionname;
        model Price=Rooms Bathroom Landsize BuildingArea Car Distance YearBuilt
        propertycount Type Method Regionname/ selection=lasso(stop=none choose=cvex)
        cvmethod=random(5);
run;
/*pour la suite nous decidons de garder le model resultant de la majorité de*/
/*methode de selection ici celui avec 17 variables*/
/*analyse de residus afin de detecter s'il ya une structure dans les données:
        homoscedasticity et normalité*/
```

```
/*normalité des termes d'erreur*/
proc reg data=data_transformed plots(MAXPOINTS=7000 only)=(qq residualhistogram);
        model Price=Rooms Bathroom Landsize BuildingArea Car Distance YearBuilt Typeh
        Typet MethodS MethodSP RegionEV RegionNM RegionNV RegionSEM RegionSM RegionWM;
run;
/*The histogramm and QQ-Plot are not good*/
/*We could conclude that the distribution of the error is not Normal */
/*test de Jarque - Bera ou test for normality*/
proc autoreg data=data_transformed plots=none;
        model Price=Rooms Bathroom Landsize BuildingArea Car Distance YearBuilt
        Typeh Typet MethodS MethodSP RegionEV RegionNM RegionNV RegionSEM
        RegionSM RegionWM/normal;
run;
/*homoscedasticité des termes d'erreur*/
proc reg data=data_transformed plots(MAXPOINTS=7000 only)=(residuals(unpack)
                RESIDUALBYPREDICTED);
        model Price=Rooms Bathroom Landsize BuildingArea Car Distance YearBuilt Typeh
        Typet MethodS MethodSP RegionEV RegionNM RegionNV RegionSEM RegionSM RegionWM;
run;
/*le plot de residus en fonction des y predits montrent que l'accroissement
de la variance des termes d'erreur est exponentiel en Y:  Les termes
d'erreur sont heteroscedastique */
/*we do a white test to confirm homoscedasticity*/
proc model data=data_transformed;
        parms b0 b1 b2 b3 b4 b5 b6 b7 b8 b9 b10 b11 b12 b13 b14 b15 b16 b17;
        Price=b0 + b1*Rooms + b2*Bathroom + b3*Landsize + b4*BuildingArea + b5*Car
        + b6*Distance + b7*YearBuilt + b8*Typeh + b9*Typet + b10*MethodS
        + b11*MethodSP + b12*RegionEV + b13*RegionNM + b14*RegionNV
        + b15*RegionSEM+ b16*RegionSM + b17*RegionWM;
        fit Price/white;
run;
/*we expect the p-value to be higher than 5%, such a way H0:
errors are homoscedastic will be rejected*/
/*remedial action*/
/*etant donné que les residus augmentent exponentiellement en fonction des valeurs
ajustées, afin de rendre cette variance constante nous faisons la transformation
suivante y'=logy*/
data data_transformed2;
        set data_transformed;
        logprice=log(price);
run;
/*visualisation à nouveau l'homoscedasticité des residus*/
proc reg data=data_transformed2 plots(MAXPOINTS=7000 only)=(residuals(unpack)
                RESIDUALBYPREDICTED);
        model logprice=Rooms Bathroom Landsize BuildingArea Car Distance YearBuilt
                Typeh Typet MethodS MethodSP RegionEV RegionNM RegionNV RegionSEM
                RegionSM RegionWM;
run;
/*les erreurs semblent homoscedastique*/
```

```
/*confirmons cela par le test de white de white */
proc model data=data_transformed2;
        parms b0 b1 b2 b3 b4 b5 b6 b7 b8 b9 b10 b11 b12 b13 b14 b15 b16 b17;
        logprice=b0 + b1*Rooms + b2*Bathroom + b3*Landsize + b4*BuildingArea
      + b5*Car + b6*Distance + b7*YearBuilt + b8*Typeh + b9*Typet + b10*MethodS
      + b11*MethodSP + b12*RegionEV + b13*RegionNM + b14*RegionNV + b15*RegionSEM
      + b16*RegionSM + b17*RegionWM;
        fit logprice/white;
run;
* bien que les erreurs semblent homoscédastique sur les plots des résidus,
le test de white rejette néanmoins l'homoscedasticité(avec une p-valeur < 0.0001);
* dans le modele final, la résolution que nous adoptons est de continuer
d'utiliser la methodes des moindres carrées ordinaires, mais nous utiliserons
l'inférence robuste afin de pouvoir tenir compte de l'hétéroscedasticité
ceci aura pour conséquence une augmentation des erreurs standards et donc,
des tests statistiques et des p-valeurs, mais les valeurs des coefficients
estimés resteront les mêmes;
/*visualisation à nouveau la normalité des residus*/
proc reg data=data_transformed2 plots(MAXPOINTS=7000 only)=(qq residualhistogram);
        model logprice=Rooms Bathroom Landsize BuildingArea Car Distance YearBuilt
        Typeh Typet MethodS MethodSP RegionEV RegionNM RegionNV RegionSEM
        RegionSM RegionWM;
run;
/*les erreurs semblent être distribuées normalement.
Confirmons cela par le test de JB */
proc autoreg data=data_transformed2 plots=none;
        model logprice=Rooms Bathroom Landsize BuildingArea Car Distance YearBuilt
        Typeh Typet MethodS MethodSP RegionEV RegionNM RegionNV RegionSEM
        RegionSM RegionWM/normal;
run;
/*we expect this time the p-value to be lower than 5%, such a way H0:
errors are normally distributed is not rejected*/
/*Etant donné que Region NV n'est plus indivuduellement significatif
dans ce model transformé, refaisons une selection de variables du
nouveau model avec la selection forward*/
proc reg data=data_transformed2 outest=selected_modelforward2 tableout plots=none;
        model logprice=Rooms Bathroom Landsize BuildingArea Car Distance YearBuilt
        Typeh Typet MethodS MethodSP RegionEV RegionNM RegionNV RegionSEM RegionSM
        RegionWM/noprint selection=stepwise;
run;
proc print data=selected_modelforward2; run;
/*on voit effectivement que cette variable Region NV est rejeté */
/*donc the final model is:*/
proc reg data=data_transformed2 plots=None;
        model logprice=Rooms Bathroom Landsize BuildingArea Car Distance YearBuilt
        Typeh Typet MethodS MethodSP RegionEV RegionNM RegionSEM RegionSM RegionWM;
run;
/*outliers observation with respect to independant variables X*/
/*consideroutliers if leverahge hii>threshol=2*p/N =2*16/6166 */
```

```
proc reg data=data_transformed2;
        model logprice=Rooms Bathroom Landsize BuildingArea Car Distance YearBuilt
        Typeh Typet MethodS MethodSP RegionEV RegionNM RegionSEM
        RegionSM RegionWM/noprint;
        plot h.*obs.;
        output out=outliers_X h=lev;
run;
proc print data=outliers_X;
        var lev;
        where lev > 2*17/6166;
run;
/*outliers observation with respect to dependant variable Y*/
/*using the Studentized residuals at 5% level*/
proc reg data=data_transformed2;
        model logprice=Rooms Bathroom Landsize BuildingArea Car Distance YearBuilt Typeh
        Typet MethodS MethodSP RegionEV RegionNM RegionSEM RegionSM RegionWM/noprint;
        plot rstudent.*obs.;
        output out=outliers_Y rstudent=stud;
run;
proc print data=outliers_Y;
        var stud;
        where abs(stud) > tinv(0.975, 61669-17-1);
run;
/*Influentiel observation (because Xi is outlier or Yi is outliers or both of them*/
/*using norm(DFFITS)>2*sqrt(p/n)=2*sqrt(17/6166)*/
proc reg data=data_transformed2 plots(MAXPOINTS=7000 only)=dffits;
        model logprice=Rooms Bathroom Landsize BuildingArea Car Distance YearBuilt
        Typeh Typet MethodS MethodSP RegionEV RegionNM RegionSEM RegionSM RegionWM;
        output out=influentials dffits=df;
run;
proc print data=influentials;
        var df;
        where (df > 2*sqrt(17/6166) or df < -2*sqrt(17/6166)) and df ne .;
run;
/*influential observations using the Cook's distance
influence not only prediction i or beta k but all the prediction
and all the betas*/
proc reg data=data_transformed2 plots(MAXPOINTS=7000 only)=COOKSD;
        model logprice=Rooms Bathroom Landsize BuildingArea Car Distance YearBuilt
        Typeh Typet MethodS MethodSP RegionEV RegionNM RegionSEM RegionSM RegionWM;
        output out=influentialscook cookd=cd;
run;
proc print data=influentialscook;
        var cd;
        where cd > finv(0.95, 17, 6166-17);
run;
/*une seule observation numero=1410*/
/*Interaction*/
data data_interaction2;
```

```
        set data_transformed2;
        th_Build=typeh*BuildingArea;
        tt_year=typet*YearBuilt;
run;
proc reg data=data_interaction2 plots=None;
        model logprice=Rooms Bathroom Landsize BuildingArea Car Distance YearBuilt
        Typeh Typet MethodS MethodSP RegionEV RegionNM RegionSEM RegionSM RegionWM
        th_Build tt_year/white;
run;
proc reg data=data_interaction2 outest=selected_modelforward3 tableout plots=none;
        model logprice=Rooms Bathroom Landsize BuildingArea Car Distance YearBuilt
        Typeh Typet MethodS MethodSP RegionEV RegionNM RegionSEM RegionSM RegionWM
        th_Build tt_year/white noprint selection=stepwise;
run;
proc print data=selected_modelforward3;
run;
* Bien que la variable ait une p-valeur légèrement supérieur à 5%, cette variable
n'est pas rejetée par la méthode de sélection des variables;
*comme dit plus haut, dans ce modele final, nous utiliserons l'inférence robuste
afin de remedier au problème persistant d'hétéroscedasticité en incluant l'option
"white" dans le modèle Sas;
*Inférence robuste;
proc reg data=data_interaction2 plots=None;
        model logprice=Rooms Bathroom Landsize BuildingArea Car Distance YearBuilt
        Typeh Typet MethodS MethodSP RegionEV RegionNM RegionSEM RegionSM RegionWM
        th_Build tt_year/white;
run;
* on retrouve les mêmes coefficients estimés mais avec des ecart-type différents
 (écart-types augmentent), ce qui modifie vraisemblablement les statistiques
 des tests et des p-valeurs. Par exemple on voit que certaines variables
 deviennent non significatives(le "th_build" avec un p-valeur de 15%,
 alors que sans inférence robuste, on avait une p-valeur <0.0001 );
/*Question5*/
/**Test de significativité joint*/
proc reg data=data_interaction2 plots=none;
        model logprice=Rooms Bathroom Landsize BuildingArea Car Distance YearBuilt
        Typeh Typet MethodS MethodSP RegionEV RegionNM RegionSEM RegionSM RegionWM
        th_Build tt_year /white noprint;
        test Rooms=0, Bathroom=0, Landsize=0, BuildingArea=0, Car=0, Distance=0,
        YearBuilt=0, Typeh=0, Typet=0, MethodS=0, MethodSP=0, RegionEV=0, RegionNM=0,
        RegionSEM=0, RegionSM=0, RegionWM=0, th_Build=0, tt_year=0;
run;
/*Label Price-valeur associer à Landsize F-statistique est inferieur à 5%, on rejette
l'hypothese nulle H0: tous les coefficient sont nuls ie dire aucune variable n'est
significative ou bien la prediction se reduit à une simple moyenne des prix*/
/*interpretation des signes/interpretations des coefficients*/

/*Question6*/
/*extra room or extra bathroom has the same effect of the price*/
```

```
proc reg data=data_interaction2 plots=none;
        model logprice=Rooms Bathroom Landsize BuildingArea Car Distance YearBuilt
        Typeh Typet MethodS MethodSP RegionEV RegionNM RegionSEM RegionSM RegionWM
        th_Build tt_year /white noprint;
        test Rooms-Bathroom=0;
run;
/*on ne reject pas H0, ie à dire l'idée à laquelle l'augmentation d'une
pièce ou d'une toilette conduirait à la meme augmentation du prix*/
/*extra m2 on land or extra m2 on building area has the same effect of the price*/
proc reg data=data_interaction2 plots=none;
        model logprice=Rooms Bathroom Landsize BuildingArea Car Distance YearBuilt
        Typeh Typet MethodS MethodSP RegionEV RegionNM RegionSEM RegionSM RegionWM
        th_Build tt_year /white noprint;
        test Landsize-BuildingArea=0;
run;
/*on reject H0, ie à dire l'idée à laquelle l'augmentation d'un m2 de terrain
ou d'un m2 de la surface contruite conduirait à la meme augmentation du prix*/
/*Question7*/
/*test subset of coefficient for exemple the qualitative variables are
jointly non significant*/
proc reg data=data_interaction2 plots=none;
        model logprice=Rooms Bathroom Landsize BuildingArea Car Distance YearBuilt
        Typeh Typet MethodS MethodSP RegionEV RegionNM RegionSEM RegionSM RegionWM
        th_Build tt_year /white noprint;
        test Typeh=0, Typet=0, MethodS=0, MethodSP=0, RegionEV=0, RegionNM=0,
                RegionSEM=0, RegionSM=0, RegionWM=0;
run;
/*H0: tous les variables qualitatives sont jointivement non significative, est rejeté*/
/*question8* and *Question 9*/
/*overall our model which consists of predicting logprice of house with respect
to some explanatory variables is adequate for linear regression, we've been
verified classical hypothesis: independance,normality and homoscedasticity
of error terms. With this model we do show that one can predict a price of
a new house dependant on its characteristics (number of rooms, bathrooms,
landsize etc...)*/
/*prediction*/
data predict;
        set sample;
        if Regionname="Eastern Metropolitan" then Regionname="EM";
        if Regionname="Eastern Victoria" then Regionname="EV";
        if Regionname="Northern Metropolitan" then Regionname="NM";
        if Regionname="Northern Victoria" then Regionname="NV";
        if Regionname="South-Eastern Metropo" then Regionname="SEM";
        if Regionname="Southern Metropolitan" then Regionname="SM";
        if Regionname="Western Metropolitan" then Regionname="WM";
        if Regionname="Western Victoria" then Regionname="WV";
run;
data predict;
        set predict;
```

```
        if Type="h" then typeh=1; else typeh=0;
        if Type="t" then typet=1; else typet=0;
        if Method="S" then MethodS=1; else MethodS=0;
        if Method="SP" then MethodSP=1; else MethodSP=0;
        if Regionname="EV" then RegionEV=1; else RegionEV=0;
         if Regionname="NM" then RegionNM=1; else RegionNM=0;
        if Regionname="SEM" then RegionSEM=1; else RegionSEM=0;
        if Regionname="SM" then RegionSM=1; else RegionSM=0;
        if Regionname="WM" then RegionWM=1; else RegionWM=0;
        th_Build=typeh*BuildingArea;
        tt_year=typet*BuildingArea;
run;
data predict;
        set predict;
        logprice=log(Price);
        if selected=1 then do;
                logprice_Obs=logprice;
                logprice=.;
        end;
run;
proc print data=predict;
run;
proc reg data=predict plots=none;
        model logprice=Rooms Bathroom Landsize BuildingArea Car Distance YearBuilt
        Typeh Typet MethodS MethodSP RegionEV RegionNM RegionSEM RegionSM RegionWM
        th_Build tt_year/white;
        output out=my_pred p=predicted ucl=upper_pred lcl=lower_pred;
run;
quit;
data toprint(keep=selected Price logprice logprice_Obs predicted lower_pred upper_pred);
        set my_pred;
run;
proc print data=toprint (obs=20);
        where selected=1 and lower_pred ne . ;
run;
quit;
* on observe que toutes les valeurs observées de la variable transformée "logprice";
*sont l'intervalle de confiance comme illustré dans les 20 premières observations
ci dessus;
ods pdf close;
ods _all_ close;
```