1. Some interpretation

After I import the data in SAS, I did some pre-work before I follow the instructions. Firstly, I delete the symbol "$"in front of every value of the variable "House Price". Secondly, I change the name of variables into one string for convenience as in Table 1. And I will use the changed name in the following report. Thirdly, it's easy to verify that the hp is another form of houseprice, which means hp=houseprice/1000, so it is reasonable to choose one of hp and houseprice to construct the model. Here I would like to choose House Price as the response variable. And I will delete the column of "HP in thousands" in the following report.

Original name	Changed name	Original name	Changed name
House Price	houseprice	T Bath	tbath
House Size	housesize	Age	Age
Acres	Acres	Garage	Garage
Lot Size	lotsize	Condition	Condition
Bedrooms	Bedrooms	Age Category	agecate

Table 1: The comparison name of variable

2. Summary of the response variable

(1) Summary Statistics

Moments			
N	200	Sum Weights	200
Mean	267466.265	Sum Observations	53493253
Std Deviation	115807.623	Variance	1.34114E10
Skewness	1.57054869	Kurtosis	4.48633336
Uncorrected SS	1.69765E13	Corrected SS	2.66887E12
Coeff Variation	43.2980297	Std Error Mean	8188.83555

Basic Statistical Measures			
Location		Variability	
Mean	267466.3	Std Deviation	115808
Median	242500.0	Variance	1.34114E10
Mode	300000.0	Range	874194
		Interquartile Range	100000

Figure 1: basic statistical measures

We can see from Figure 1 above, the basic statistic measurements are: Mean 267466.265, variance 1.34114E10, standard deviation 115807.623, median 242500, range 874194, interquartile range 100000, mode 300000.The mean and the median has some difference which shows that there are some extreme observations to raise the average level.

Quantiles (Definition 5)	
Level	Quantile
100% Max	887194
99%	639000
95%	494250
90%	429000
75% Q3	300000
50% Median	242500
25% Q1	200000
10%	150000
5%	136500
1%	65100
0% Min	13000

Extreme Observations			
Lowest		Highest	
Value	Obs	Value	Obs
13000	140	560000	136
60200	70	585000	184
70000	167	600000	127
100000	119	678000	110
107263	191	887194	98

Figure 2: quantiles and extreme observation

Figure 2 gives the quantiles of the houseprice and the extreme observations of the response variable.

(2) Figures of response variable

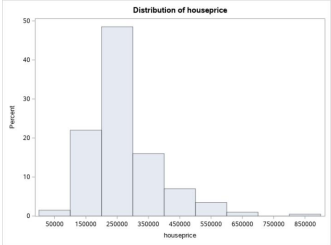
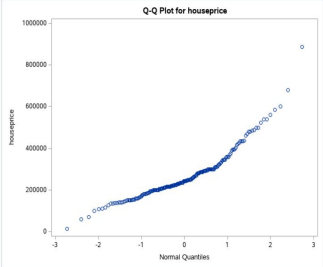


Figure 3: the hist of response variable



Tests for Normality				
Test	Statistic		p Value	
Shapiro-Wilk	W	0.894124	Pr < W	<0.0001
Kolmogorov-Smirnov	D	0.144382	Pr > D	<0.0100
Cramer-von Mises	W-Sq	0.995446	Pr > W-Sq	<0.0050
Anderson-Darling	A-Sq	5.544427	Pr > A-Sq	<0.0050

Figure 4: QQplot and the normality test

From Figure 3 we can see that the houseprice has a large range and is not distributed symmetrically. From Figure 4 we can see that QQplot is approximately like a straight line, but the Shapiro-Wilk test shows that the p-value is less than general significance level 0.05,so it seems that the data of the response can not be seen as normal distributed data.The results above show that It may be not very suitable to fit a linear regression model for this response.

(3)1. Correlation coefficient matrix

Pearson Correlation Coefficients, N = 200 Prob > |r| under H0: Rho=0									
	houseprice	housesize	Acres	lotsize	Bedrooms	tbath	Age	Garage	Condition
houseprice	1.00000	0.71051 <.0001	0.22952 0.0011	0.22952 0.0011	0.31702 <.0001	0.57521 <.0001	-0.16976 0.0163	0.09646 0.1742	0.03407 0.6320
housesize	0.71051 <.0001	1.00000	0.36328 <.0001	0.36328 <.0001	0.25591 0.0003	0.41591 <.0001	-0.08588 0.2266	-0.07133 0.3155	-0.00502 0.9438
Acres	0.22952 0.0011	0.36328 <.0001	1.00000	1.00000 <.0001	-0.15635 0.0270	-0.04702 0.5085	0.04789 0.5007	-0.17447 0.0135	-0.03623 0.6106
lotsize	0.22952 0.0011	0.36328 <.0001	1.00000 <.0001	1.00000	-0.15635 0.0270	-0.04702 0.5085	0.04789 0.5007	-0.17447 0.0135	-0.03623 0.6106
Bedrooms	0.31702 <.0001	0.25591 0.0003	-0.15635 0.0270	-0.15635 0.0270	1.00000	0.63800 <.0001	-0.18657 0.0082	0.26504 0.0001	0.05290 0.4569
tbath	0.57521 <.0001	0.41591 <.0001	-0.04702 0.5085	-0.04702 0.5085	0.63800 <.0001	1.00000	-0.43842 <.0001	0.24689 0.0004	0.00674 0.9246
Age	-0.16976 0.0163	-0.08588 0.2266	0.04789 0.5007	0.04789 0.5007	-0.18657 0.0082	-0.43842 <.0001	1.00000	-0.31047 <.0001	0.30832 <.0001
Garage	0.09646 0.1742	-0.07133 0.3155	-0.17447 0.0135	-0.17447 0.0135	0.26504 0.0001	0.24689 0.0004	-0.31047 <.0001	1.00000	0.00386 0.9568
Condition	0.03407 0.6320	-0.00502 0.9438	-0.03623 0.6106	-0.03623 0.6106	0.05290 0.4569	0.00674 0.9246	0.30832 <.0001	0.00386 0.9568	1.00000

Figure 5: Correlation coefficient matrix

From Figure 5 we can know that: (1)Variable housesize, acres, lotsize, bedrooms, tbath, garage and condition have the positive relationship with the houseprice, while the age has the negative relationship with houseprice, which make sense in our life. (2)and we also test the coefficient for every pair of the regressor and the response, where the H0 :Two variables have no relationship. Under the significance level of 0.05, we believe that the housesize, acres,lotsize, bedrooms, tbath ,age have a relationship with houseprice. While garage and the condition have no relation under significance level of 0.05.

2. Figures

In the above coefficient matrix, we have known that most variables have relation with the response variable, here we use scatter plot to find out the relation between

the regressor and the response group by the variable agecate. For numeric variable, scatter plot enables us to see the trend. Figure 6 and 7 below only provide the scatter plot of some regressors, Fig 6 is the overview scatter for regressor, while Fig 7 is the scatter for regressor group by the variable agecate. From left to behind, the regressors are: housesize, arces,age. From Fig 6 and 7 we know that the trend is consistent with the value in the coefficient matrix.

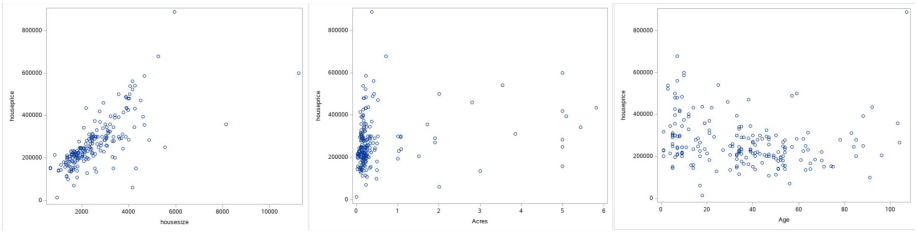


Figure 6: scatter plots of regressor with the response

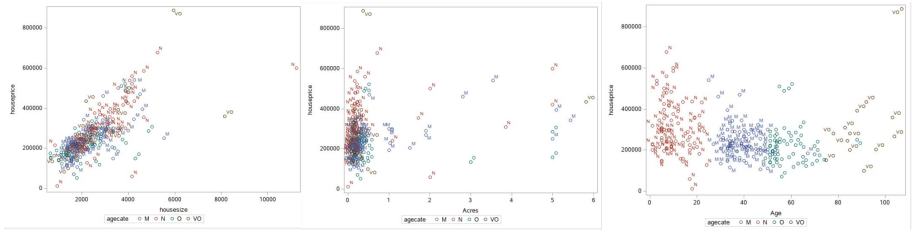


Figure 7: scatter plots of regressor with the response group by agecate

3. Regression model

First we would like to construct the Model:

$$y = b_0 + b_1 \times x_1 + b_2 \times x_2 + \ldots + b_8 \times x_8 + b_{10} \times x_{10} + b_{11} \times x_{11} + b_{12} \times x_{12}$$

As I explained in part 1, I delete the hp in thousands so there are 8 regressors left. While agecate is a categorical variable with 4 levels, so I add three variables x_{10}, x_{11}, x_{12} for this model.and you can find the meaning in the following figure 8.

$$x_{10} = \begin{cases} 1 & \text{if the agecate is M} \\ 0 & \text{if otherwise} \end{cases} \quad x_{11} = \begin{cases} 1 & \text{if the agecate is O} \\ 0 & \text{if otherwise} \end{cases} \quad x_{12} = \begin{cases} 1 & \text{if the agecate is N} \\ 0 & \text{if otherwise} \end{cases}$$

Figure 8: the meaning of added variables

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	118995	72909	1.63	0.1043
housesize	1	50.15435	5.32396	9.42	<.0001
Acres	B	6024.49150	5456.40937	1.10	0.2709
lotsize	0	0	.	.	.
Bedrooms	1	-7146.45943	6475.90192	-1.10	0.2712
tbath	1	57901	10401	5.57	<.0001
Age	1	-444.79485	751.97030	-0.59	0.5549
Garage	1	26627	13410	1.99	0.0485
Condition	1	2075.07512	16530	0.13	0.9002
x10	1	-91667	43881	-2.09	0.0380
x11	1	-73278	32997	-2.22	0.0276
x12	1	-93656	63228	-1.48	0.1402

Figure 9： results of the regression

The fitted model is:

$$y = 118995 + 50.15 \times x_1 + 6024.49 \times x_2 - 7146.46 \times x_4 + 57901 \times x_5 - 444.79 \times x_6 + 26627 \times x_7 + 2075.08 \times x_8 - 91667 \times x_{10} - 73278 \times x_{11} - 93656 \times x_{12}$$

Fig9 gives t-test for each estimated parameter, from the last column of Fig9 we can see that only the parameters for housesize and tbath are significant under

significance level 0.05.

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	10	1.675167E12	1.675167E11	31.86	<.0001
Error	189	9.937024E11	5257684666		
Corrected Total	199	2.66887E12			

Root MSE	72510	R-Square	0.6277
Dependent Mean	267466	Adj R-Sq	0.6080
Coeff Var	27.10992		

Figure 10: ANAOV table

1. The sum of regression square is 1.675167E12 with freedom 10, the sum of residual square is 9.937024E11 with freedom 189, the total sum square is 2.66887E12 with freedom 199.

2. The Estimator of σ^2 is 5257684666.

3. R-square is 0.6277, R- square adjusted is 0.608, which do not show the appropriateness of the model.

4. Figure 10 also gives the F test results, where

$$H_0 : b_1 = b_2 = \cdots = b_8 = b_{10} = b_{11} = b_{12} = 0 \quad \text{vs} \quad H_1 : b_j \neq 0, \quad \text{for at least one } j$$

The F-statistic is 31.86,p-value is <0.0001,so we reject the H_0,that's to say there is a linear relationship between the response and any of the regressor variables.

4. adequacy of the model

We can see from Fig11 shapior-wilk test that the Standardized Residuals is not normal distribution but t-test shows that the Mu for SR is 0. From the hist and qqplot for SR and SR vs Fitted value , we can easily find that there are some outliers and the SR&Fitted is a satisfactory pattern. So I would like to delete the outliers to regress again. Figure 12 gives a further proof that the model is not bad. But I still want to make some progress in the model.

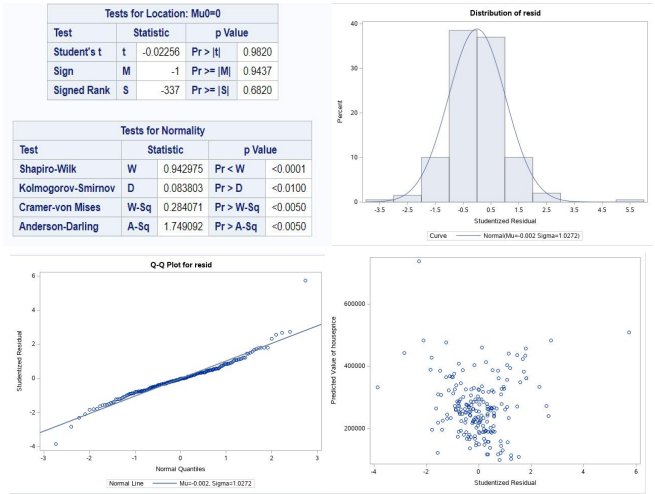


Figure11: test for SR , hist and qqplot for SR, SR vs Fitted value

5. Improvement

1. I delete the obvious outlier observations whose absolute value of SR is greater than 4.5 ,cook's D is greater than 0.5.The outlier is the observation 98.

2. Figure 9 shows the parameter of lotsize is 0,that's because the lotsize is a linear

function of acres: lotsize =43560 * acres, so I would like to delete the variable lotsize.

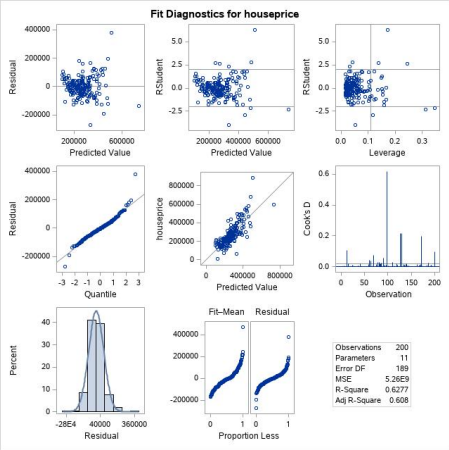


Figure 12: Fit Diagnostics for response

3. it's not difficult to find that the variable agecate is just derived from the variable age, and we spare a lot of efforts to design new variables for the qualitative variable agecate, which I think is redundant, so in the new model I would like to use the variable age while deleting the variable x_{10}, x_{11}, x_{12}.

4. So the formula of new model is :

$y = b_0 + b_1 \times x_1 + b_2 \times x_2 + \ldots + b_7 \times x_7$

The fitted model is: $y = 45204 + 45.88 \times x_1 + 6969.69 \times x_2 - 12204 \times x_3 + 62575 \times x_4 - 40.56 \times x_5 + 15455 \times x_6 - 2446.9 \times x_7$

5. The sum of regression square is 1.433104E12 with freedom 7, the sum of residual square is 8.497732E11 with freedom 191,the total sum square is 2.282877E12 with freedom 198.

6. The Estimator of σ^2 is 4449074413.

7. R-square is 0.6278,R- square adjusted is 0.6141,which dose not show the appropriateness of the model.

8. Figure 13 also gives the F test results, where

$$H_0 : b_1 = b_2 = \cdots = b_7 = 0 \quad \text{vs} \quad H_1 : b_j \neq 0, \quad \text{for at least one } j$$

The F-statistic is 46.02,p-value is <0.0001,so we reject the H_0,that's to say there is a linear relationship between the response and any of the regressor variables.

Analysis of Variance

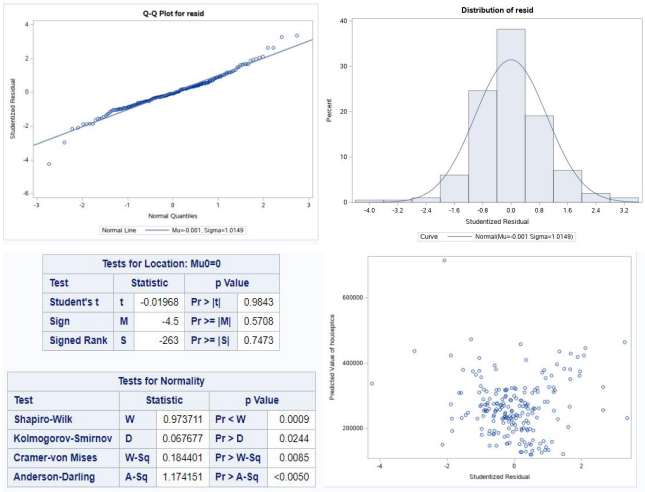
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	7	1.433104E12	2.047292E11	46.02	<.0001
Error	191	8.497732E11	4449074413		
Corrected Total	198	2.282877E12			

Root MSE	66701	R-Square	0.6278
Dependent Mean	264352	Adj R-Sq	0.6141
Coeff Var	25.23203		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	45204	23006	1.96	0.0509
housesize	1	45.87783	4.81332	9.53	<.0001
Acres	1	6969.68693	4875.59630	1.43	0.1545
Bedrooms	1	-12204	5792.54087	-2.11	0.0364
tbath	1	62575	9472.75896	6.61	<.0001
Age	1	-40.56464	245.71648	-0.17	0.8690
Garage	1	15455	12320	1.25	0.2112
Condition	1	-2446.89631	15110	-0.16	0.8715

Figure 13: ANOVA and parameter estimate for new model



Tests for Location: Mu0=0				
Test	Statistic		p Value	
Student's t	t	-0.01968	Pr > |t|	0.9843
Sign	M	-4.5	Pr >= |M|	0.5708
Signed Rank	S	-263	Pr >= |S|	0.7473

Tests for Normality				
Test	Statistic		p Value	
Shapiro-Wilk	W	0.973711	Pr < W	0.0009
Kolmogorov-Smirnov	D	0.067677	Pr > D	0.0244
Cramer-von Mises	W-Sq	0.184401	Pr > W-Sq	0.0085
Anderson-Darling	A-Sq	1.174151	Pr > A-Sq	<0.0050

Figure 14: test for new SR , hist and qqplot for new SR, new SR vs new Fitted value

9. Fig13 gives t-test for each estimated parameter, from the last column of Fig13 we can see that only the parameters for housesize and tbath are significant under significance level 0.05.

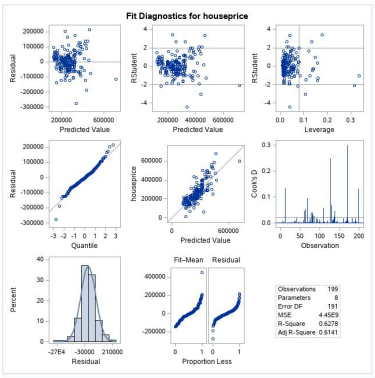


Figure15 :Fit Diagnostics for response of new model

10. Compare the Fig 11 and Fig14, we can find the new SR are more closely to a standard normal distribution because the qqplot is more like a straight line.The new-SR&new-fitted value is more concentrate. The p-value of shapiro-wilk test has a larger value than the old model. And the Cook's D in Fig15 are all less than 0.3. The total sum square is less than the old model. We have a simpler model than before. That's all the reason why the new model is better than the old one.

APPENDIX

```
/* Generated Code (IMPORT) */
/* Source File: house_selling_prices_OR.csv */
/* Source Path: /home/u59857001 */
/* Code generated on: 11/8/21, 8:19 PM */
%web_drop_table(WORK.house_data);
FILENAME REFFILE '/home/u59857001/house_selling_prices_OR.csv';

PROC IMPORT DATAFILE=REFFILE
```

```
      DBMS=CSV
      OUT=WORK.house_data;
      GETNAMES=YES;
RUN;

PROC CONTENTS DATA=WORK.house_data; RUN;

%web_open_table(WORK.house_data);

/*change the name of variable*/
data house_data;
set house_data(rename=('House Price'n=houseprice 'HP in thousands'n=hp 'House
Size'n=housesize    'Lot Size'n=lotsize 'T Bath'n=tbath 'Age Category'n=agecate));
run;
porc corr data=house_data nosimple;
title"the corrlation matrix for the response variable";
var houseprice lotsize condition;
run;

proc univariate data=house_data;
var houseprice;
run;
proc univariate data=house_data;
  histogram houseprice;
run;

proc univariate data=house_data;
var houseprice;
qqplot;

proc univariate normal;
var houseprice;
run;

proc corr data=house_data nosimple;
title "Correlation coefficient matrix";
var houseprice housesize acres lotsize bedrooms tbath age garage condition ;
run;

/* Scatter plot by agecate */
proc sgscatter    data = house_data;
     plot houseprice * age;
     / datalabel = agecate group = agecate;
run;
```

```
/*give variable agecate an indicator (dummy) variable*/
data house_data;
set house_data;
if agecate = "M" then x10= 1;
else x10= 0;
run;
data house_data;
set house_data;
if agecate = "O" then x11= 1;
else x11= 0;
run;
data house_data;
set house_data;
if agecate = "N" then x12= 1;
else x12= 0;
run;

proc reg data=house_data;
   model houseprice = housesize acres lotsize bedrooms tbath age garage condition
x10 x11 x12;
   output out=analysis P =yhat R =residual STUDENT = resid cookd= cooks H =
leverage;
/*a dataset named "analysis" is stored with all information about data and the
above variables;*/
/*P = fitted, R = raw residuals, student is standardized residuals ;*/
run;
quit;

proc univariate data=analysis normal ;
var resid;
histogram resid /normal;
qqplot /normal (mu=est sigma=est) ;
run;

proc sgscatter data=analysis;
plot yhat*resid;
run;

/*delete the outlier */
data   work.house_data1;
set   house_data;
if houseprice   =   887194     then   delete;
run;
```

```
proc print data=work.house_data1;
run;

proc reg data=house_data1;
   model houseprice = housesize acres  bedrooms tbath age garage condition ;
   output out=analysis1 P =yhat R =residual STUDENT = resid cookd= cooks H = leverage;
/*a dataset named "analysis" is stored with all information about data and the above variables;*/
/*P = fitted, R = raw residuals, student is standardized residuals ;*/
run;
quit;

proc univariate data=analysis1 normal ;
var resid;
histogram resid /normal;
qqplot /normal (mu=est sigma=est) ;
run;
proc sgscatter data=analysis1;
plot yhat*resid;
run;
```