

PROJECT REPORT

PRICES OF HOUSE RESALES IN ALBUQUERQUE NEW MEXICO



For Course:
Statistics 571 -
Statistical Methods

Instructor: Ramon V. Leon

Student's Name: Kyriakos Voutouris
ID Number: 414 95 4128
Date: 11/25/2003

PART I: **INTRODUCTION**

HISTORY OF THE DATA:

The dataset of the project is a random sample of house resales collected in Albuquerque in New Mexico. These data were maintained in the files of Albuquerque Board of Realtors and the sample was collected between the 15th of February and 30th of April 1993. The dataset contains records of the selling prices of the houses in hundreds of dollars as well as information about characteristics of the houses that were sold.

Usually this type of data is collected by multiple listing agencies in many cities and is used by realtors as an information base. Realtors can use them to find an expected selling price for a home with given characteristics such as the square feet of the house.

Some relevant information:



Albuquerque is located in central New Mexico on the upper Rio Grande, southwest of Santa Fe. It is the largest city of New Mexico with a population of about 600 thousand people.

Albuquerque Board of Realtors is a multiple listing agency that buys and sells houses. The management centers of the company are located in Albuquerque and it is considered to be one of the largest associations of realtors in the United States.

Purpose of the project:

The purpose of the project is to use the information given in the current dataset to build a multiple regression model which would be able to predict the reselling price of a house in Albuquerque based on some relevant characteristics of each house.

The complete dataset is given in the next page.

THE DATASET

PRICE	SQFT	AGE	FEATS	NE	CUST	COR	TAX
2050	2650	13	7	1	1	0	1639
2080	2600	*	4	1	1	0	1088
2150	2664	6	5	1	1	0	1193
2150	2921	3	6	1	1	0	1635
1999	2580	4	4	1	1	0	1732
1900	2580	4	4	1	0	0	1534
1800	2774	2	4	1	0	0	1765
1560	1920	1	5	1	1	0	1161
1450	2150	*	4	1	0	0	*
1449	1710	1	3	1	1	0	1010
1375	1837	4	5	1	0	0	1191
1270	1880	8	6	1	0	0	930
1250	2150	15	3	1	0	0	984
1235	1894	14	5	1	1	0	1112
1170	1928	18	8	1	1	0	600
1180	1830	*	3	1	0	0	733
1155	1767	16	4	1	0	0	794
1110	1630	15	3	1	0	1	867
1139	1680	17	4	1	0	1	750
995	1725	*	3	1	0	0	923
995	1500	15	4	1	0	0	743
975	1430	*	3	1	0	0	752
975	1360	*	4	1	0	0	696
900	1400	16	2	1	0	1	731
960	1573	17	6	1	0	0	768
860	1385	*	2	1	0	0	653
1695	2931	28	3	1	0	1	1142
1553	2200	28	4	1	0	0	1035
1250	2277	*	4	1	1	0	*
1300	2000	*	3	1	1	0	1076
1020	1478	53	3	1	0	1	626
1020	1713	30	4	1	0	1	600
922	1326	*	4	1	0	0	668
925	1050	*	2	1	0	1	553
899	1464	*	2	1	1	0	566
850	1190	41	1	1	0	0	600
876	1156	*	1	1	0	0	*
890	1746	*	2	1	0	0	591
870	1280	*	1	1	0	0	599
700	1215	*	3	1	0	0	477
720	1121	46	4	1	0	0	398
720	1050	*	1	1	0	0	*
749	1733	43	6	1	0	0	656
731	1299	*	6	1	0	0	585
725	1140	*	3	1	0	1	490
670	1181	*	4	1	0	0	440
2150	2848	4	6	1	1	0	1487
1599	2440	*	5	1	1	0	1265
1350	2253	23	4	1	1	0	939
1299	2743	25	5	1	1	1	1232
1250	2180	17	4	1	0	1	1141
1239	1706	14	4	1	0	0	810
1200	1948	*	4	1	0	0	899
1125	1710	16	4	1	1	0	800
1100	1657	*	4	1	0	0	865
1080	2200	26	4	1	0	0	1076
1050	1680	13	4	1	0	0	875
1049	1900	34	3	1	0	0	690
955	1565	*	3	1	1	0	648
934	1543	20	3	1	0	0	820

875	1173	6	4	1	0	0	456
889	1549	*	4	1	0	0	723
855	1900	*	3	1	0	0	780
835	1560	*	5	1	0	1	638
810	1365	*	2	1	0	0	673
805	1258	7	4	1	0	1	821
799	1314	*	2	1	0	0	671
750	1338	*	3	1	0	1	649
759	997	4	4	1	0	0	461
755	1275	*	5	1	0	0	*
750	1030	*	1	1	0	0	486
730	1027	*	3	1	0	0	427
729	1007	19	6	1	0	0	513
710	1083	22	4	1	0	0	504
773	1320	*	5	1	0	0	*
690	1348	15	2	1	1	0	*
670	1350	*	2	1	0	0	622
619	837	*	2	1	0	0	342
1295	3750	*	4	0	1	1	1200
975	1500	7	3	0	1	1	700
939	1428	40	2	0	0	0	701
820	1375	*	1	0	0	0	585
780	1080	*	3	0	1	0	600
770	900	*	3	0	0	0	391
700	1505	*	2	0	0	1	591
620	1480	*	4	0	0	0	*
540	1142	*	0	0	0	0	223
1070	1464	*	2	0	0	0	376
2100	2116	25	3	0	1	0	1209
725	1280	*	3	0	0	0	447
660	1159	*	0	0	0	0	225
600	1198	*	4	0	0	0	*
580	1051	15	2	0	0	0	426
1844	2250	40	6	0	1	0	915
1580	2563	*	2	0	1	0	1189
699	1400	45	1	0	1	1	481
1330	1850	5	5	0	1	1	*
1160	1720	5	4	0	0	0	867
1109	1740	4	3	0	0	0	816
1129	1700	6	4	0	0	0	725
1050	1620	6	4	0	0	0	800
1045	1630	6	4	0	0	0	750
1050	1920	8	4	0	0	0	944
1020	1606	5	4	0	0	0	811
1000	1535	7	5	0	0	1	668
1030	1540	6	2	0	0	1	826
975	1739	13	3	0	0	0	880
950	1715	*	3	0	0	0	900
940	1305	5	3	0	0	0	647
920	1415	7	4	0	0	0	866
945	1580	9	3	0	0	0	810
874	1236	3	4	0	0	0	707
872	1229	6	3	0	0	0	721
870	1273	4	4	0	0	0	638
869	1165	7	4	0	0	0	694
766	1200	7	4	0	0	1	634
739	970	4	4	0	0	1	541

Note: * is used for missing values

VARIABLES INCLUDED IN THE DATASET:

Explanatory Variable:

PRICE – Selling price of the house in hundreds of dollars.

Predictor Variables:

1. SQFT – Square feet of living space
2. AGE – Age of the house in years
3. FEATS - Number out of 11 possible features (Dishwasher, refrigerator, microwave, disposer, washer, intercom, skylight(s), compactor, dryer, handicap fit, cable TV access)
4. NE - House located in the North East sector of the city (1), or not (0).
5. CUST – Custom built of the house, yes (1) or no (0).
6. COR – Corner location of the house, yes (1) or no (0).
7. TAX - Annual property taxes in dollars

COMMENTS ON THE DATASET:

Variables PRICE, SQFT, AGE, FEATS and TAX are considered to be continuous while variables NE, CUST and COR are categorical with values 0 and 1. NE, CUST and COR are dummy variables.

As stated before, the response variable is the selling price of the houses and the remaining variables are the predictors. We are going to analyze the data and try to find the best fitting a multiple regression model for this data. Our effort is try to construct a multiple regression linear model that would be able to predict the selling price of a house in Albuquerque, New Mexico. It is important to say that because of the fact that the data were collected in Albuquerque, we can not make general inferences about the reselling prices of houses in other cities.

All of the above predictor variables have a reasoning to be associated with the response variable which is the selling price of the house. The **amount of living space** of a house is an important factor that possibly affects the selling price of a house. The **age of the house** could also be a factor associated with the selling price as well as **the number of some luxurious features** that the house contains.

It is believed that whether or not a house is **custom built** affects the selling price of the house. The uniqueness of the house and probably the special features that each owner has added to his house probably could increase the price that he is willing to sell it. The location of the house is believed that could be a possible reason that could affect the selling price. It is accepted that houses **built in corner locations** of the streets are preferred and are generally more expensive. For some reasons also it is suspected that whether or not a house is **built on the North East sector of the city** could be associated with the reselling price of a house.

Finally, the **annual amount of property taxes** that each owner pays should be strongly correlated with the selling price of the house.

POSSIBLE PROBLEMS ASSOCIATED WITH THE DATA

One problem about this dataset is that there are missing values mostly concerning the age of the houses that have been resold. This fact should be considered in the further analysis of the data.

Another possible problem that could confound our conclusions is the possibility that the predictor variables in the dataset are correlated. This could cause multicollinearity problems. The following analysis should check if multicollinearity is present and if so, find ways that multicollinearity problems are solved.

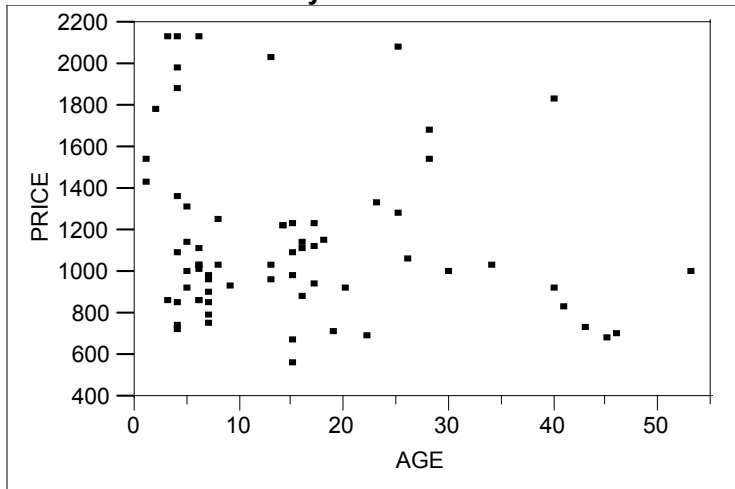
PART II
ANALYSIS OF THE DATA:

MISSING VALUES:

As a first step in this analysis we will consider the problem of the missing values in the dataset. The variable AGE has a large number of missing variables and this creates problems in our data as JMP ignores all the observations that have missing values in the analysis. This results in losing 41 observations because of the AGE variable and consequently losing important information which could help in the selection of the best model. In order to examine the nature of the relationship between the age and the price of the house we can construct a scatter plot of PRICE against AGE.

Scatter plot of PRICE VS AGE:

Bivariate Fit of PRICE By AGE



The scatter plot does not support a strong linear relationship between PRICE and AGE. This is further supported by the calculated Pearson's Correlation Coefficient:

Correlations

	PRICE	AGE
PRICE	1.0000	-0.1687
AGE	-0.1687	1.0000

49 rows not used due to missing values.

Standard Least Squares Regression:

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob> t	VIF
Intercept	92.744798	101.607	0.91	0.3651	.
SQFT	0.3522184	0.095748	3.68	0.0005	6.1974316
AGE	-0.565082	2.002529	-0.28	0.7788	1.6923362
FEATS	4.3896066	18.55499	0.24	0.8138	1.4591479
NE	-17.38534	47.27462	-0.37	0.7144	1.3749655
CUST	174.94108	53.72371	3.26	0.0019	1.3859059
COR	-73.58234	49.13007	-1.50	0.1396	1.1083272
TAX	0.4988701	0.158485	3.15	0.0026	6.4765975

Type III Sum of Squares Tests:

Effect Tests

Source	Nparm	DF	Sum of Squares	F Ratio	Prob > F
SQFT	1	1	1600.8139	15.0040	0.0003
FEATS	1	1	658.0312	6.1675	0.0159
NE	1	1	836.6556	7.8417	0.0069
CUST	1	1	91.0944	0.8538	0.3592
COR	1	1	0.1512	0.0014	0.9701
TAX	1	1	2904.0601	27.2189	<.0001

By running the standard least squares regression and by using Type III tests we can see that the AGE variable does not contribute significantly in reducing Square Error as the relevant test shows that AGE is not significant indicated by the large p-value.

There is a possibility that the fact that we have quite many missing values for AGE create bias to the above results about the significance of AGE but since these are the only data available, we do not have another option than excluding AGE of the model. In this way, we will also benefit for having more observations available for our analysis.

A method that could be use to solve the problem of the missing values could be to predict those values of AGE by using multiple regression techniques with AGE as response variable and the predictors to be the rest of the previous model predictors. This would not be a good solution though because of the multicollinearity problems that would be caused, as AGE would be strongly correlated with the rest of the predictors.

Therefore we decide to exclude AGE out of the model not only because it has many missing variables, but because as seen in the above, it does not contribute significantly in reducing the sum of squares error.

VARIABLE SELECTION:

I. Stepwise Regression:

As a next step, we will see what the stepwise regression suggests about which predictors are significant in the model:

Stepwise Fit

Response:
PRICE

Stepwise Regression Control

Prob to Enter 0.250
Prob to Leave 0.100

Direction:

10 rows not used due to missing values.

Current Estimates

		SSE	DFE	MSE	RSquare	RSquare Adj	Cp	AIC
		2687729	102	26350.284	0.8280	0.8213	3.7680763	1094.057
Lock	Entered	Parameter	Estimate	nDF	SS	"F Ratio"	"Prob>F"	
	X	Intercept	175.166031	1	0	0.000	1.0000	
	X	SQFT	0.20760441	1	304735.9	11.565	0.0010	
		FEATS	0	1	20228.72	0.766	0.3836	
		NE	0	1	1730.411	0.065	0.7992	
	X	CUST	156.814814	1	327300.1	12.421	0.0006	
	X	COR	-83.401261	1	114214.9	4.334	0.0399	
	X	TAX	0.67706796	1	1183801	44.926	0.0000	

Step History

Step	Parameter	Action	"Sig Prob"	Seq SS	RSquare	Cp	p
1	TAX	Entered	0.0000	11984585	0.7668	33.658	2
2	CUST	Entered	0.0000	586774.1	0.8043	13.658	3
3	SQFT	Entered	0.0027	256275.7	0.8207	6.0502	4
4	COR	Entered	0.0399	114214.9	0.8280	3.7681	5

As seen in the above output, the stepwise regression drops variables FEATS and NE out of the model.

II Standard Least Squares Regression:

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob> t	VIF
Intercept	152.11631	62.85299	2.42	0.0173	.
SQFT	0.2065093	0.061436	3.36	0.0011	4.2347481
FEATS	10.839246	12.92587	0.84	0.4037	1.2943801
NE	3.4134871	34.72503	0.10	0.9219	1.0799406
CUST	154.78142	44.91536	3.45	0.0008	1.4080978
COR	-81.48848	40.37073	-2.02	0.0462	1.0313466
TAX	0.657407	0.104234	6.31	<.0001	4.10081

Effect Tests

Source	Nparm	DF	Sum of Squares	F Ratio	Prob > F
SQFT	1	1	301369.1	11.2989	0.0011
FEATS	1	1	18756.0	0.7032	0.4037
NE	1	1	257.7	0.0097	0.9219
CUST	1	1	316745.4	11.8754	0.0008
COR	1	1	108673.0	4.0744	0.0462
TAX	1	1	1060997.8	39.7788	<.0001

The above output supports the results of the stepwise regression as the partial F-Tests or equivalently the t-tests show that the coefficients of NE and FEATS variables are not significant. The p-values of the tests are very large causing fail of rejection of the null hypothesis that the coefficients are equal to zero. Therefore the number of features of the selling houses as well as the whether or not the selling house is located on the North East sector of the city will be excluded from further consideration.

SCATTER PLOTS AND CORRELATION COEFFICIENTS OF PREDICTORS

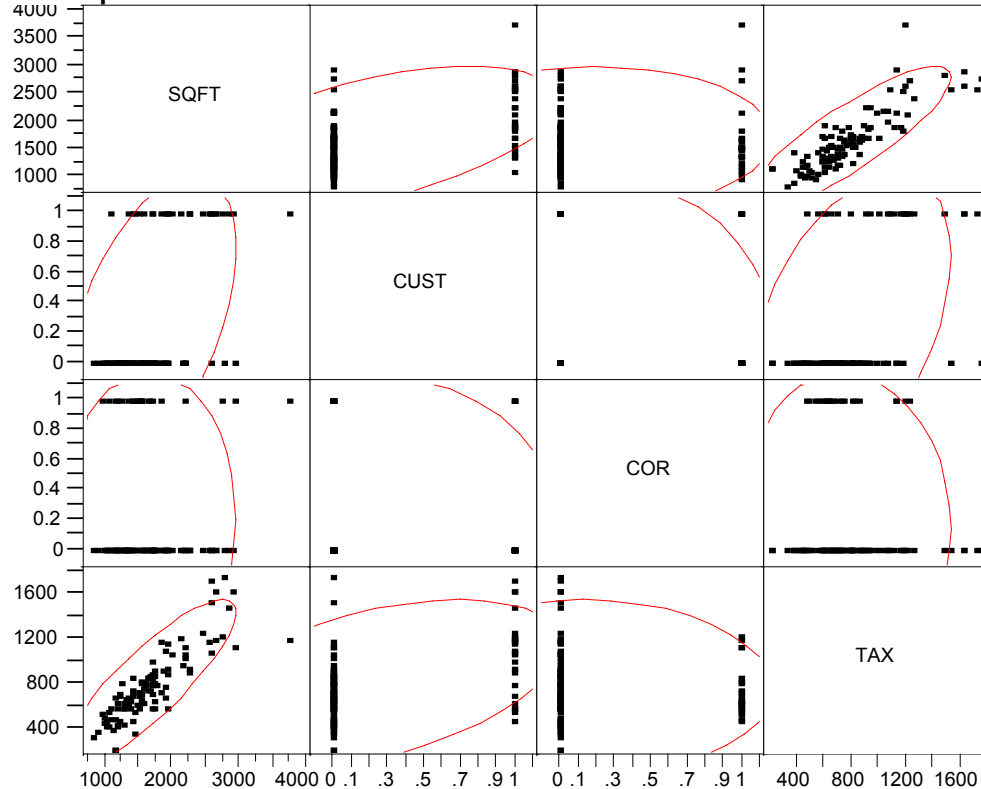
As we mentioned before a possible problem that could occur in the data analysis is multicollinearity. Multicollinearity is present in multiple regression when the predictor variables are linearly related and causes severe problems in regression analysis. High correlation coefficients between predictor variables could indicate if there is strong linear relationship between the predictors.

Correlation matrix between predictor as well as scatter plots follow:

Correlations

	SQFT	CUST	COR	TAX
SQFT	1.0000	0.5311	0.0217	0.8586
CUST	0.5311	1.0000	-0.0401	0.4699
COR	0.0217	-0.0401	1.0000	-0.0600
TAX	0.8586	0.4699	-0.0600	1.0000

Scatterplot Matrix



As seen in the above outputs there is a high correlation between TAX and SQFT. This is logical as we expect that the amount of annual taxes that a house owner pays is dependent on the amount of living space of the house.

Because of this linear relationship of the two variables, in the next steps we are going to examine if multicollinearity problem exists and if so, find ways to solve it.

Let's consider the model

MODEL I:

$$\hat{PRICE} = \beta_0 + \beta_1 SQFT + \beta_2 CUST + \beta_3 COR + \beta_4 TAX$$

Response PRICE - Summary of Fit

RSquare	0.828036
RSquare Adj	0.821292
Root Mean Square Error	162.3277
Mean of Response	1077.346
Observations (or Sum Wgts)	107

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	4	12941849	3235462	122.7866
Error	102	2687729	26350	Prob > F
C. Total	106	15629578		<.0001

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob> t	VIF
Intercept	175.16603	56.31157	3.11	0.0024	.
SQFT	0.2076044	0.061047	3.40	0.0010	4.2324957
CUST	156.81481	44.49454	3.52	0.0006	1.3987293
COR	-83.40126	40.05934	-2.08	0.0399	1.027913
TAX	0.677068	0.101015	6.70	<.0001	3.8985328

Effect Tests

Source	Nparm	DF	Sum of Squares	F Ratio	Prob > F
SQFT	1	1	304735.9	11.5648	0.0010
CUST	1	1	327300.1	12.4211	0.0006
COR	1	1	114214.9	4.3345	0.0399
TAX	1	1	1183801.5	44.9256	<.0001

Press
3339156.4306

The F-test of the hypothesis $H_0 : \beta_1 = \beta_2 = \beta_3 = \beta_4 = 0$ indicates rejection of H_0 . Also each of the coefficients in the model is statistically significant as the p-values of the tests $H_0 : \beta_j = 0$ for $j=1,2,3,4$ are small. The R-Square value is 0.828 and therefore, 82.8% of the variability in the selling price of the houses is explained by regression on the predictors. The PRESS statistic which is a measure of the predictive ability of the model is calculated to be 3,339,156.

As we can also see in the above problem it does not seem that we have a serious multicollinearity problem as all the VIF's are less than 10. Yet, we can see that the VIF of SQFT and TAX is kind of high but quite lower than 10.

Overall, we can say that this model seems quite reasonable as a large proportion in the variability in the selling prices of the houses is accounted for by regression on the predictor variables. Another advantage of this model is the fact that is quite simple. Also, we do not seem to have any serious multicollinearity problem in this model.

As a next step we are going to consider the model with all the possible interactions and quadratic terms and see if we have an improvement in our results

MODEL II:

Let's consider the model,

$$\hat{PRICE} = \beta_0 + \beta_1 SQFT + \beta_2 CUST + \beta_3 COR + \beta_4 TAX + \beta_{11} SQFT^2 + \beta_{22} TAX^2 + \beta_{12} SQFT * CUST + \beta_{13} SQFT * COR + \beta_{14} SQFT * TAX + \beta_{23} CUST * COR + \beta_{24} CUST * TAX + \beta_{34} COR * TAX$$

By running Least Square Regression in JMP we get the following results:

Response PRICE - Summary of Fit

RSquare	0.877467
RSquare Adj	0.861824
Root Mean Square Error	142.7372
Mean of Response	1077.346
Observations (or Sum Wgts)	107

Analysis of Variance

Source	Source	DF	Sum of Squares	Mean Square	F Ratio
Model	Model	12	13714431	1142869	56.0948
Error	Error	94	1915147	20374	Prob > F
C. Total	C. Total	106	15629578		<.0001

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob> t	VIF
Intercept	122.22379	58.90648	2.07	0.0407	.
SQFT	0.3849865	0.069086	5.57	<.0001	7.0106501
CUST	111.75955	43.72844	2.56	0.0122	1.7472673
COR	-93.68516	39.14121	-2.39	0.0187	1.269195
TAX	0.3712074	0.113609	3.27	0.0015	6.3777086
(SQFT-1667.26)*(SQFT-1667.26)	-0.00014	0.000135	-1.03	0.3043	26.276868
(SQFT-1667.26)*(CUST-0.2243)	-0.029	0.163355	-0.18	0.8595	11.026588
(SQFT-1667.26)*(COR-0.19626)	-0.174582	0.171265	-1.02	0.3106	9.7689291
(SQFT-1667.26)*(TAX-793.486)	0.0006735	0.000445	1.52	0.1331	53.419549
(CUST-0.2243)*(COR-0.19626)	-317.1657	110.1259	-2.88	0.0049	1.6036481
(CUST-0.2243)*(TAX-793.486)	0.167528	0.241084	0.69	0.4888	7.5670212
(COR-0.19626)*(TAX-793.486)	0.0117742	0.328989	0.04	0.9715	5.6620771
(TAX-793.486)*(TAX-793.486)	-0.000537	0.000392	-1.37	0.1741	23.232734

Effect Tests

Source	Nparm	DF	Sum of Squares	F Ratio	Prob > F
SQFT	1	1	632673.44	31.0531	<.0001
CUST	1	1	133080.83	6.5319	0.0122
COR	1	1	116720.52	5.7289	0.0187
TAX	1	1	217512.39	10.6760	0.0015
SQFT*SQFT	1	1	21735.35	1.0668	0.3043
SQFT*CUST	1	1	642.10	0.0315	0.8595
SQFT*COR	1	1	21170.79	1.0391	0.3106
SQFT*TAX	1	1	46763.40	2.2953	0.1331
CUST*COR	1	1	168992.53	8.2946	0.0049
CUST*TAX	1	1	9838.13	0.4829	0.4888
COR*TAX	1	1	26.10	0.0013	0.9715
TAX*TAX	1	1	38217.80	1.8758	0.1741

Press
4722219.2569

We can see that now the R-Square value has increased but we have other problems like multicollinearity problems. The VIF values mostly of the quadratic terms SQFT and TAX and the interaction between those two exceed 10 by far. This was kind of expected as multicollinearity increases when quadratic terms and interactions are inserted in the model as this creates high correlations between the predictors.

We can also see that the predictive ability of the model is reduced now as the PRESS statistic has a larger value than the one before.

The Type III tests of the above model suggest that the only polynomial term that is statistically significant is the interaction between CUST and COR. Thus, we are going to consider the model with the four predictors and the interaction between CUST and COR and see the results.

MODEL III

$$PRICE = \beta_0 + \beta_1 SQFT + \beta_2 CUST + \beta_3 COR + \beta_4 TAX + \beta_{23} CUST * COR$$

Response PRICE - Summary of Fit

RSquare	0.860959
RSquare Adj	0.854076
Root Mean Square Error	146.6845
Mean of Response	1077.346
Observations (or Sum Wgts)	107

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio	Prob > F
Model	5	13456427	2691285	125.0810	
Error	101	2173151	21516		
C. Total	106	15629578			<.0001

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob> t	VIF
Intercept	159.61305	50.98423	3.13	0.0023	.
SQFT	0.2826403	0.057259	4.94	<.0001	4.559937
CUST	141.64845	40.32612	3.51	0.0007	1.4070511
COR	-104.0697	36.44479	-2.86	0.0052	1.0419249
TAX	0.5445691	0.095216	5.72	<.0001	4.2420028
(CUST-0.2243)*(COR-0.19626)	-458.3832	93.7319	-4.89	<.0001	1.1000451

Effect Tests

Source	Nparm	DF	Sum of Squares	F Ratio	Prob > F
SQFT	1	1	524271.43	24.3662	<.0001
CUST	1	1	265472.38	12.3382	0.0007
COR	1	1	175446.96	8.1541	0.0052
TAX	1	1	703802.11	32.7101	<.0001
CUST*COR	1	1	514577.83	23.9157	<.0001

Press
2723427.243

This model has a quite high R-Square value (0.861) which is almost as high as the model with all the possible quadratic terms and interactions. We know that every time we add predictors in a model the R-Square value increases but this does not mean that the model with the highest R-Square value is the best as the simplicity of a model is essential.

All the VIF values in this model are lower than 10 and so, no multicollinearity problems are present in this model. Also, the PRESS statistic value is lower than the one of both the previous models. This means that this model has a better predictive ability than both the previous models.

Use of Stepwise Regression in MODEL II

As a next step we are going to use Stepwise Regression on the model with all the quadratic terms and interactions to see which model Stepwise Regression suggests.

Stepwise Fit

Response: PRICE

Stepwise Regression Control

Prob to Enter 0.250
Prob to Leave 0.100

Current Estimates

Lock	Entered	Parameter	Estimate	nDF	SS	"F Ratio"	"Prob>F"
		Intercept	128.272044	1	0	0.000	1.0000
X	X	SQFT	0.36276417	2	677349.5	16.983	0.0000
	X	CUST	102.938259	3	522809.4	8.739	0.0000
	X	COR	-99.929163	3	720933.6	12.051	0.0000
	X	TAX	0.41420097	2	453061	11.360	0.0000
		(SQFT-1667.26)*(SQFT-1667.26)	0	1	1094.992	0.054	0.8161
		(SQFT-1667.26)*(CUST-0.2243)	0	1	2121.833	0.105	0.7461
	X	(SQFT-1667.26)*(COR-0.19626)	-0.1991435	1	158547.5	7.951	0.0058
		(SQFT-1667.26)*(TAX-793.486)	0	1	11765.5	0.588	0.4452
	X	(CUST-0.2243)*(COR-0.19626)	-315.56644	1	193688	9.713	0.0024
	X	(CUST-0.2243)*(TAX-793.486)	0.18172256	1	57837.76	2.900	0.0917
		(COR-0.19626)*(TAX-793.486)	0	1	2350.408	0.117	0.7332
		(TAX-793.486)*(TAX-793.486)	0	1	4198.745	0.209	0.6487

Step History

Step	Parameter	Action	"Sig Prob"	Seq SS	RSquare	Cp	p
1	(SQFT-1667.26)*(COR-0.19626)	Entered	0.0000	12716015	0.8136	44.005	4
2	(CUST-0.2243)*(TAX-793.486)	Entered	0.0000	745638.9	0.8613	13.407	7
3	(CUST-0.2243)*(COR-0.19626)	Entered	0.0024	193688	0.8737	5.9002	8

The results of stepwise regression are different from the results that we get if we use the Type III tests. Stepwise Regression suggests that we should keep the interactions between SQFT and COR, CUST and COR and between CUST and TAX in the model.

We are also going to consider both the model that Stepwise Regression suggests and compare the results among the other candidate models.

MODEL IV

$$\hat{PRICE} = \beta_0 + \beta_1 SQFT + \beta_2 CUST + \beta_3 COR + \beta_4 TAX + \beta_{13} SQFT * COR + \beta_{23} CUST * COR + \beta_{24} CUST * TAX$$

Response PRICE - Summary of Fit

RSquare	0.873686
RSquare Adj	0.864755
Root Mean Square Error	141.2154
Mean of Response	1077.346
Observations (or Sum Wgts)	107

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	7	13655342	1950763	97.8229
Error	99	1974236	19942	Prob > F
C. Total	106	15629578		<.0001

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob> t	VIF
Intercept	128.27204	52.00222	2.47	0.0154	.
SQFT	0.3627642	0.062259	5.83	<.0001	5.8168835
CUST	102.93826	42.01761	2.45	0.0160	1.648179
COR	-99.92916	35.23521	-2.84	0.0055	1.0508092
TAX	0.414201	0.100757	4.11	<.0001	5.1251249
(SQFT-1667.26)*(COR-0.19626)	-0.199144	0.070627	-2.82	0.0058	1.6972974
(CUST-0.2243)*(COR-0.19626)	-315.5664	101.2561	-3.12	0.0024	1.3851064
(CUST-0.2243)*(TAX-793.486)	0.1817226	0.106705	1.70	0.0917	1.5144988

Effect Tests

Source	Nparm	DF	Sum of Squares	F Ratio	Prob > F
SQFT	1	1	677026.09	33.9501	<.0001
CUST	1	1	119689.16	6.0019	0.0160
COR	1	1	160396.40	8.0432	0.0055
TAX	1	1	337002.65	16.8993	<.0001
SQFT*COR	1	1	158547.50	7.9505	0.0058
CUST*COR	1	1	193687.98	9.7127	0.0024
CUST*TAX	1	1	57837.76	2.9003	0.0917

Press
2617793.0813

The above model has an R-Square value of 0.874 and an Adjusted R-Square value of 0.865. The adjusted R-Square value of this model is even larger than the model with all the possible quadratic terms and interactions. Also, all the VIF values are smaller than ten and so there is no multicollinearity present.

As we can see the p-values of the tests $H_0 : \beta_j = 0$ for $j=1,2,3,4$ as well as for the tests $H_0 : \beta_{ij} = 0$ for $(i, j) = (1, 3), (2,3)$ and $(2,4)$ are very small and therefore we conclude that all the coefficients in the model are significant.

The PRESS statistic value of this model is calculated to be 2,617,793 which is the smallest among all the other candidate models. So, we can say that this model has the best predictive ability among all of the candidate model that we examined.

Leaving either TAX or SQFT out of the model:

One might think that it would be a good solution to leave either TAX or SQFT out of the model after the output in Page 10 that indicated multicollinearity problems. The fact is that after excluding the quadratic terms of TAX and SQFT as well as the interaction between them out of the model which are not significant, we do not have any multicollinearity problems. This is indicated by the fact that the VIF values are less than ten. Also we can see that if we make the model without TAX or SQFT we have an important decrease in the R-Square values and an increase in the PRESS statistic value.

JMP OUTPUT:

Model without TAX:

Summary of Fit

RSquare	0.830383
RSquare Adj	0.81839
Root Mean Square Error	163.6404
Mean of Response	1077.346
Observations (or Sum Wgts)	107

Press
4808233.1124

Model without SQFT:

Summary of Fit

RSquare	0.823099
RSquare Adj	0.811739
Root Mean Square Error	165.0681
Mean of Response	1062.735
Observations (or Sum Wgts)	117

Press
3249051.658

By leaving one of TAX or SQFT out of the model, the variation of PRICE that is accounted for by regression on the predictors is decreasing as is the predictive ability of the model. Therefore it would not be a good idea if we exclude either one of the two variables.

CONCLUSION

According to the analysis based above the best model is model IV. This model has the largest adjusted R-Square value ($r_{adj,p}^2$) and almost as large R-Square value as the model with all the possible interactions and quadratic terms. A large amount of variation of the price of the reselling houses is accounted for by regression on the predictors that were used in the model (87.4%).

Also, model IV has the lower PRESS statistic value among all the possible multiple regression models. This means that this model has the best predictive ability among all the other possible models.

By running the Type III tests, we get that all the terms in the model are statistically significant in a 0.1 level as they significantly decrease the Sum of Squares error (p 13).

Therefore we conclude that the best multiple regression model for the data of the reselling houses in Albuquerque is MODEL IV:

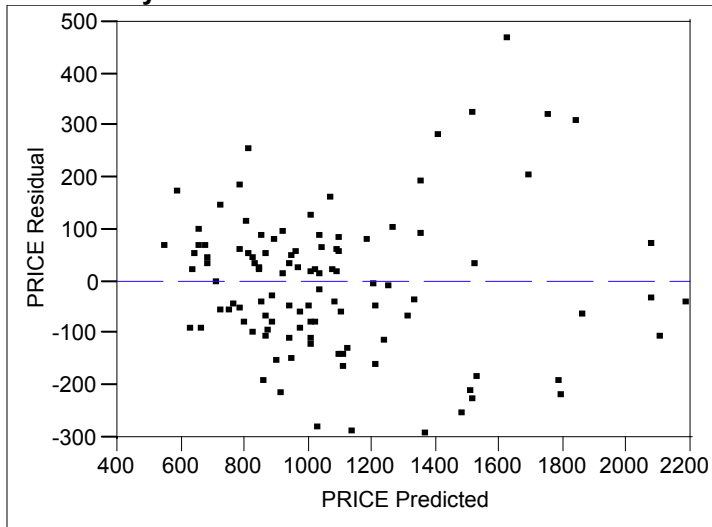
$$\hat{PRICE} = \beta_0 + \beta_1 SQFT + \beta_2 CUST + \beta_3 COR + \beta_4 TAX + \beta_{13} SQFT * COR + \beta_{23} CUST * COR + \beta_{24} CUST * TAX$$

In the following section we are going to use Residual Analysis to examine if the model assumptions are not violated in this specific multiple regression model.

ANALYSIS OF THE RESIDUALS (For the selected Model):

Residual VS Predicted Plot:

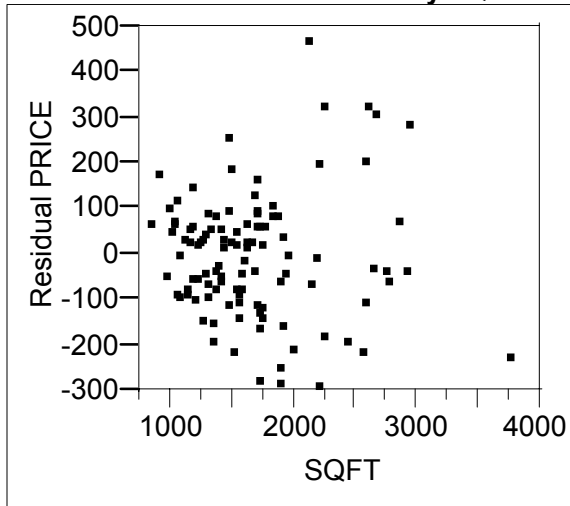
Residual by Predicted Plot



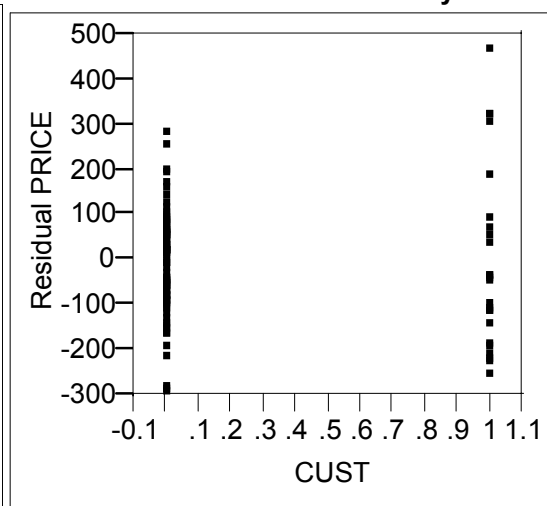
As we can see in the above plot the residuals are randomly scattered around zero. Because of this, the assumption of **constant variance** is not violated.

Plots of the Residuals versus the Predictors:

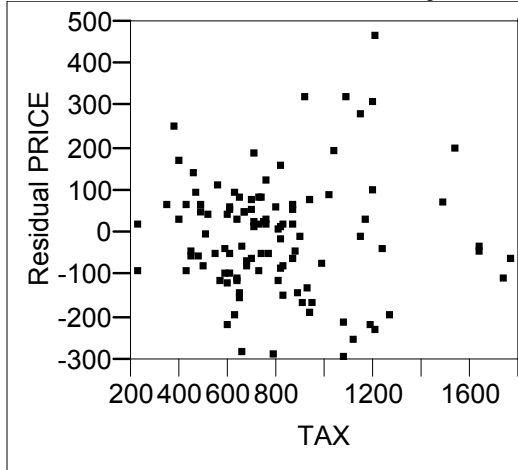
Bivariate Fit of Residual PRICE By SQFT



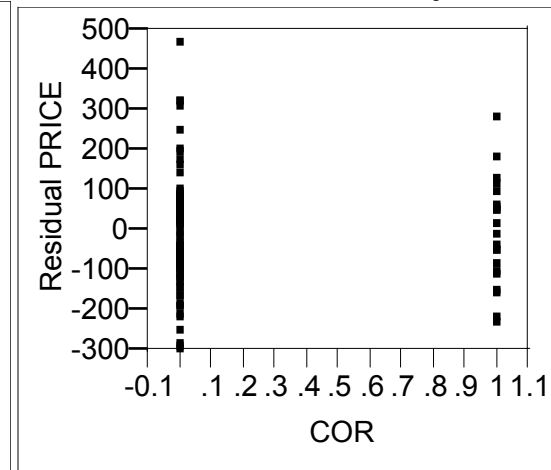
Bivariate Fit of Residual PRICE By CUST



Bivariate Fit of Residual PRICE By TAX



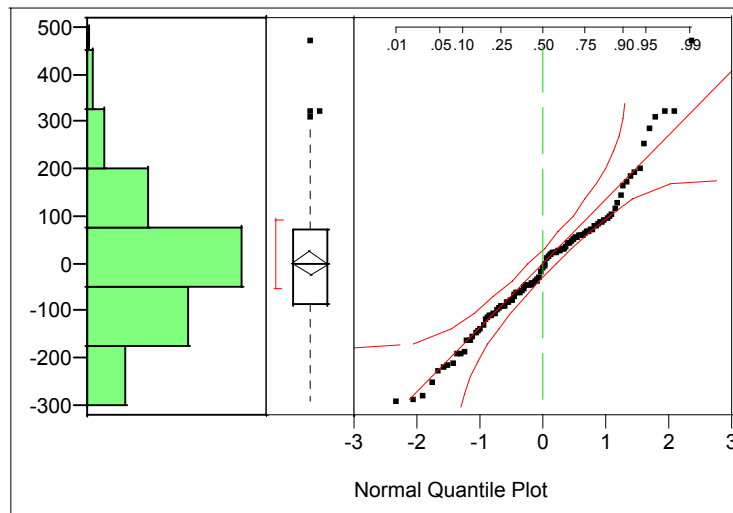
Bivariate Fit of Residual PRICE By COR



The above plots of the residuals against the predictors are randomly scattered around zero. Also, there is no systematic pattern in the above plots. Therefore, the **assumption of linearity** is not violated.

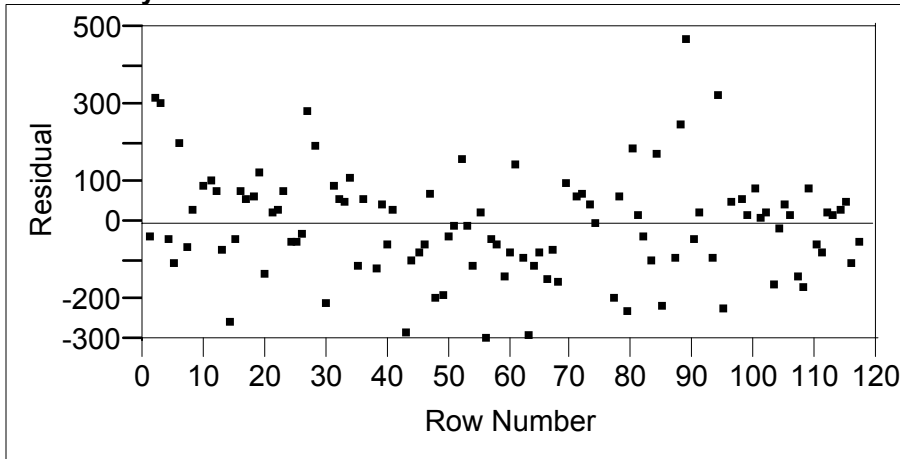
Histogram and Normal Plot of the Residuals:

Residual PRICE



The normal plot of the residuals is very close to a straight line and therefore the **assumption of normality** seems quite reasonable. This is also supported by the shape of the histogram as well as the outlier box plot.

Run Chart of the Residuals:
Residual by Row Plot



There are no signs of correlation introduced by time order as seen in the above run chart of the residuals and therefore the **assumption of independence** is not violated.

Durbin – Watson Test:

In order to further check the assumption of independence of the residuals we can use the Durbin – Watson test.

Durbin-Watson

Durbin-Watson	Number of Obs.	AutoCorrelation
1.7421603	107	0.0511

The autocorrelation statistic is calculated to be 0.051 which indicates that we do not have a problem of autocorrelation and therefore we have further evidence that the **assumption of independence** seems quite reasonable.

CONCLUSION

All the assumptions of linearity, normality, constant variance and independence seem quite reasonable based on the above residual analysis. Therefore, we conclude that none of the underlying assumptions of the model is being violated.

PART III

SUMMARY – FINAL CONCLUSIONS:

The best model for the data of house resales in Albuquerque, according to the above analysis, is the following:

$$PRICE = \beta_0 + \beta_1 SQFT + \beta_2 CUST + \beta_3 COR + \beta_4 TAX + \beta_{13} SQFT * COR + \beta_{23} CUST * COR + \beta_{24} CUST * TAX$$

where PRICE is the price of the reselling houses, SQFT are the square feet of living space in the house, CUST is whether a house is custom built or not, COR is whether or not the house is built on a corner location and TAX is the annual amount of taxes that the owner of the house pays.

This kind of model can be used by realtors to predict the selling price of a house in Albuquerque by using some basic information that they have about the house like the footage of the house and the amount of taxes that the owner pays. It can also be used to help individual residents of Albuquerque who are interested in buying or selling a house to gain insight about the levels that the price of a reselling house could vary.

It should be noted again that this model is useful only for predicting the reselling price of a house in Albuquerque in New Mexico. The price that homes are resold in other cities in the United States can not be predicted by this model as many other possible reasons could affect this price. It is also a fact that house prices are relatively cheap in some places whereas they are quite expensive in others. Therefore, if someone uses this model which is based on the data collected in Albuquerque, to predict the selling price of a house in some other place, it would be completely wrong.

Another fact is that this data that the analysis was based on were collected in 1993. Many possible factors could affect the selling price of a house during a ten year period. The advance in economy and the increase of the average salary could be two of them. So, questions arise about the actual predictive ability of the model. Lastly, we can say that this model would be reliable to predict the reselling price of houses in Albuquerque for a relatively short time period after the data were collected.