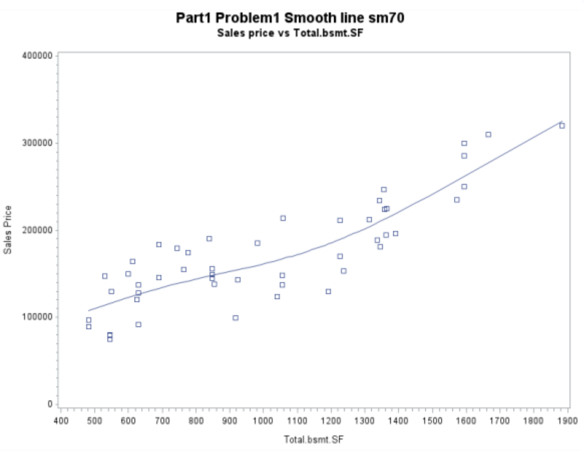
DATA ANALYSIS

HOUSING PRICING

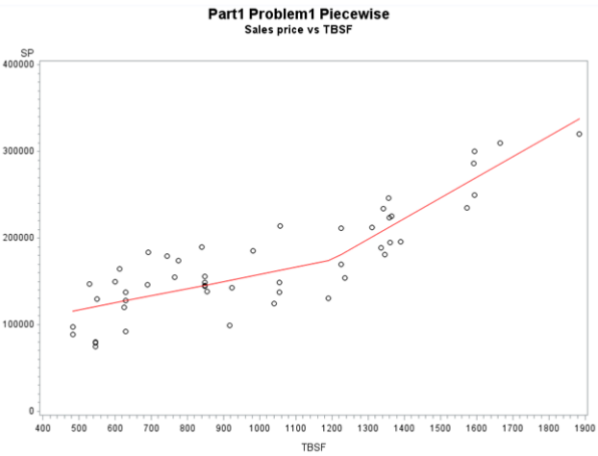
Chuyun Li; Mengzhi Zhou; Xudan Luo; Xin Li;

Part I

1.



scatter plot with smoothing line



piecewise SLR model

We choose the predictor named Total bsmt.SF and the graph of the scatter plot looks pretty similar to the piecewise SLR model. There is no extreme value or outlier in the piecewise SLR model and the points follow the line well. We add an additional explanatory variable that will add a constant to the slope whenever TBSF is greater than 1200 so the two pieces in the piecewise model are different since the slopes and intercepts change when TBSF = 1200.

2.

a) We sum up the predictors GLA and GA.

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	-4121035	640804	-6.43	<.0001
LA	1	-2.54919	1.64645	-1.55	0.1303
OQ	1	15124	3651.22954	4.14	0.0002
OC	1	12589	5296.13748	2.38	0.0229
YB	1	1054.39994	969.68429	1.09	0.2841
YRA	1	963.65698	943.08075	1.02	0.3137
BFSF	1	14.82042	7.93553	1.87	0.0700
TBSF	1	103.14799	14.91642	6.92	<.0001
FB	1	-13156	8464.55227	-1.55	0.1289
HB	1	30350	7962.40211	3.81	0.0005
BAG	1	17729	6114.41337	2.90	0.0063
TRAG	1	-3591.42298	3460.46303	-1.04	0.3063
FP	1	19669	6236.22983	3.15	0.0032
GC	1	-8614.31709	7174.48128	-1.20	0.2377

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	-3604390	728327	-4.95	<.0001
LA	1	-1.64442	1.74293	-0.94	0.3519
SUM	1	27.42533	19.23394	1.43	0.1628
OQ	1	12983	3900.57625	3.33	0.0021
OC	1	10507	5422.09735	1.94	0.0608
YB	1	1095.63975	956.50001	1.15	0.2598
YRA	1	670.93610	952.22555	0.70	0.4857
BFSF	1	17.39922	8.03036	2.17	0.0372
TBSF	1	85.09982	19.40376	4.39	0.0001
FB	1	-12335	8365.51226	-1.47	0.1493
HB	1	21000	10229	2.05	0.0476
BAG	1	12145	7189.07060	1.69	0.1000
TRAG	1	-3782.63641	3414.48653	-1.11	0.2755
FP	1	16225	6605.88568	2.46	0.0191
GC	1	-13185	7765.99210	-1.70	0.0984

Using all the explanatory variables:

SP = -4121035 - 2.54919*LA + 15124*OQ + 12589*OC + 1054.39994*YB + 963.65698*YRA + 14.82042*BFSF + 103.14799*TBSF -13156*FB + 30350*HB + 17729*BAG -3591.42298*TRAG + 19669*FP -8614.31709*GC

Using all the explanatory variables including SUM:

SP = -3604390 - 1.64442*LA + 27.42533*SUM + 12983*OQ + 10507*OC + 1095.63975*YB + 670.93610*YRA + 17.39922*BFSF + 85.09982*TBSF -12335*FB + 21000*HB + 12145*BAG - 3782.63641*TRAG + 16225*FP - 13185*GC

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	13	1.638048E11	12600370715	47.58	<.0001
Error	$\boxed{36}$	$\boxed{9533580374}$	264821677		
Corrected Total	49	1.733384E11			

without SUM

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	14	1.643282E11	11737729852	45.60	<.0001
Error	$\boxed{35}$	$\boxed{9010181740}$	257433764		
Corrected Total	49	1.733384E11			

with SUM

Extra sum of squares = SSE (R) – SSE (F) = 9533580374 - 9010181740

= 523398634

F = {[SSE (R) – SSE (F)] / [dfE (R) - dfE (F)]} / [SSE (F) / dfE (F)]

= 523398634 / (36 - 35) / (9010181740 / 35)

= 2.033139033

Degree of freedom: numerator = number of extra variables = 1

denominator = dfE for the larger model = 35

b)

Test 1 Results for Dependent Variable SP				
Source	DF	Mean Square	F Value	Pr > F
Numerator	1	523398634	2.03	0.1628
Denominator	35	257433764		

H_0: the coefficient of the SUM variable is zero

H_a: the coefficient of the SUM variable is not zero

From the SAS output, df and F-value are the same as in part (a).

Since p-value is 0.1628 which is large, we are not confident enough to reject the null hypothesis and there is evidence showing that the SUM variable has no linear relationship with SP when all other variables are included in the model.

c)

Test 1 Results for Dependent Variable SP				
Source	DF	Mean Square	F Value	Pr > F
Numerator	1	523398634	2.03	0.1628
Denominator	35	257433764		

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
SUM	1	27.42533	19.23394	1.43	0.1628

test statement: F = 2.03, p = 0.1628

individual t-test: t = 1.43, p = 0.1628

$(1.43)^2 \approx 2.04$ which is consistent with definition that $F = t^2$, and p-values for the two are the same. We get such a result because the two tests are testing the same null hypothesis that the coefficient of SUM is equal to zero when all other variables are included in the model.

3.

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	15	$\boxed{1.650535E11}$	11003565726	45.16	<.0001
Error	34	8284913790	243673935		
Corrected Total	49	1.733384E11			

Root MSE	15610	R-Square	0.9522
Dependent Mean	171350	Adj R-Sq	0.9311
Coeff Var	9.11007		

Parameter Estimates							
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Type I SS	Type II SS
Intercept	1	-3702267	710863	-5.21	<.0001	1.468034E12	6609569806
LA	1	-2.71720	1.80613	-1.50	0.1417	17495550955	551515452
GLA	1	4.39789	22.98538	0.19	0.8494	81274915544	8920599
GA	1	117.26700	55.33556	2.12	0.0415	23127674534	1094339373
OQ	1	12001	3837.33437	3.13	0.0036	12796208647	2383227953
OC	1	8076.96241	5459.97636	1.48	0.1483	7037892	533241174
YB	1	1314.46753	939.19104	1.40	0.1707	4421699911	477310428
YRA	1	512.85728	930.94808	0.55	0.5853	547409708	73952217
BFSF	1	18.18843	7.82619	2.32	0.0262	15952370472	1316131209
TBSF	1	88.14389	18.96035	4.65	<.0001	4201277535	5266240352
FB	1	-8502.56852	8436.56240	-1.01	0.3207	1129245970	247501771
HB	1	21474	9955.74623	2.16	0.0382	770637620	1133700029
BAG	1	13681	7050.77801	1.94	0.0607	145358494	917452931
TRAG	1	-2441.51933	3411.72216	-0.72	0.4791	114076101	124790665
FP	1	19283	6666.82320	2.89	0.0066	1636810152	2038524303
GC	1	-30760	12683	-2.43	0.0208	1433212347	1433212347

From the SAS output:

Type I sums of squares:

17495550955 + 81274915544 + 23127674534 + 12796208647 + 7037892 + 4421699911 + 547409708 + 15952370472 + 4201277535 + 1129245970 + 770637620 + 145358494 + 114076101 + 1636810152 + 1433212347 = 165053485882

Type II sums of squares:

551515452 + 8920599 + 1094339373 + 2383227953 + 533241174 + 477310428 + 73952217 + 1316131209 + 5266240352 + 247501771 + 1133700029 + 917452931 + 124790665 + 2038524303 + 1433212347 = 17600060803

The Type I SS is equal to themodel SS.

From definition, SSM(LA) + SSM(GLA|LA) + SSM(GA|LA,GLA) + SSM(OG|LA,GLA,GA) + SSM(OC|LA,GLA,GA,OQ) + SSM(YB|LA,GLA,GA,OQ,OC) + SSM(YRA|LA,GLA,GA,OQ,OC,YB) + SSM(BFSF|LA,GLA,GA,OQ,OC,YB,YRA) + SSM(TBSF|LA,GLA,GA,OQ,OC,YB,YRA,BFSF) + SSM(FB|LA,GLA,GA,OQ,OC,YB,YRA,BFSF,TBSF) + SSM(HB|LA,GLA,GA,OQ,OC,YB,YRA,BFSF,TBSF,FB) + SSM(BAG|LA,GLA,GA,OQ,OC,YB,YRA,BFSF,TBSF,FB,HB) + SSM(TRAG|LA,GLA,GA,OQ,OC,YB,YRA,BFSF,TBSF,FB,HB,BAG) + SSM(GC|LA,GLA,GA,OQ,OC,YB,YRA,BFSF,TBSF,FB,HB,BAG,TRAG,FP) = SSM(LA,GLA,GA,OQ,OC,YB,YRA,BFSF,TBSF,FB,HB,BAG,TRAG,FP,GC) which is exactly the model SS.

The SS of GC in the two types are the same, since the GC is the last predictor and its SS for both types equals to SSM(GC|LA,GLA,GA,OQ,OC,YB,YRA,BFSF,TBSF,FB,HB,BAG,TRAG,FP).

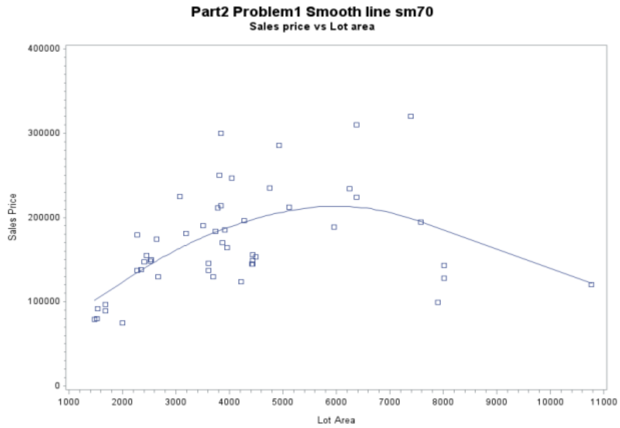
4.

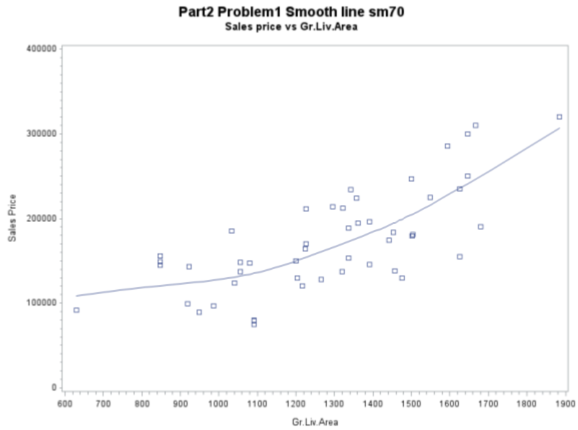
Predictor (s)	R^2	Adjusted R^2
GLA	0.5166	0.5065
GA	0.5579	0.5487
GLA, GA	0.6700	0.6560
SUM	0.6438	0.6364
YB	0.4254	0.4134
YRA	0.4691	0.4580
BFSF	0.3705	0.3574
TBSF	0.7211	0.7153
GC	0.3943	0.3817
TBSF, SUM	0.8452	0.8386

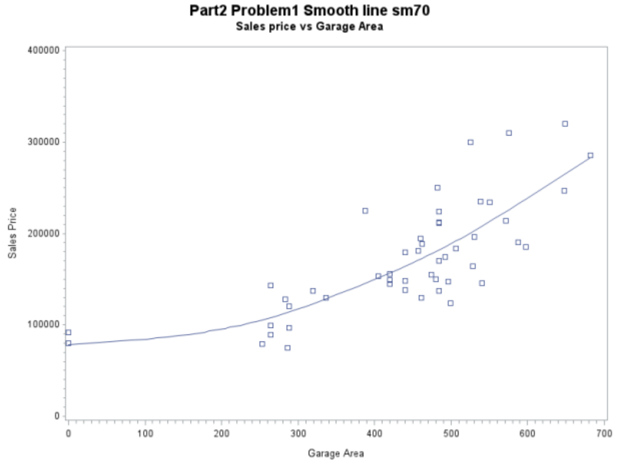
Part II

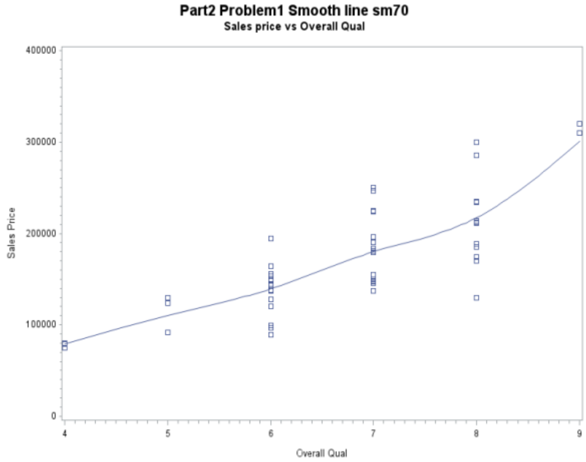
1.

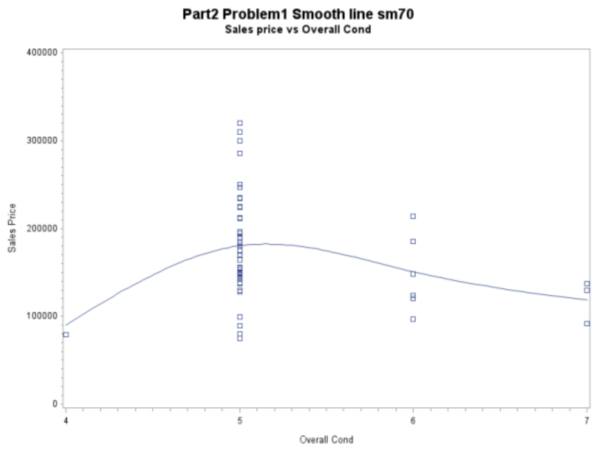
Scatter Plot

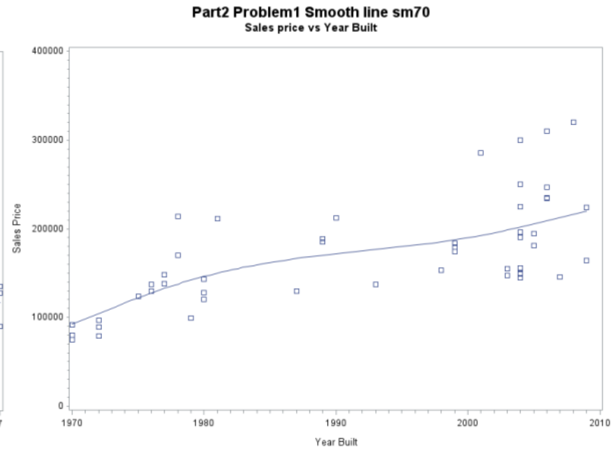


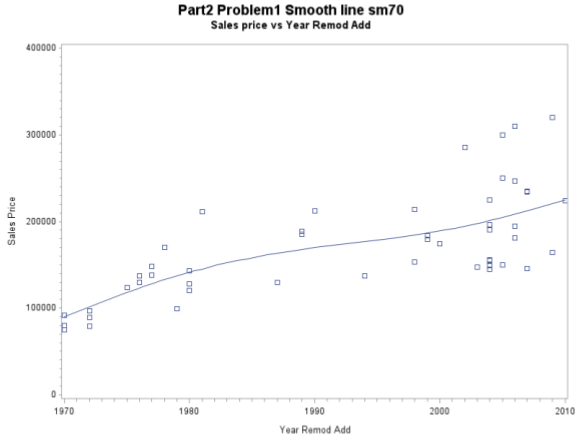


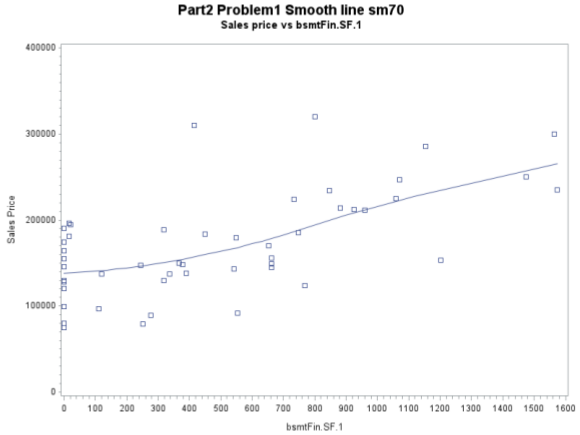


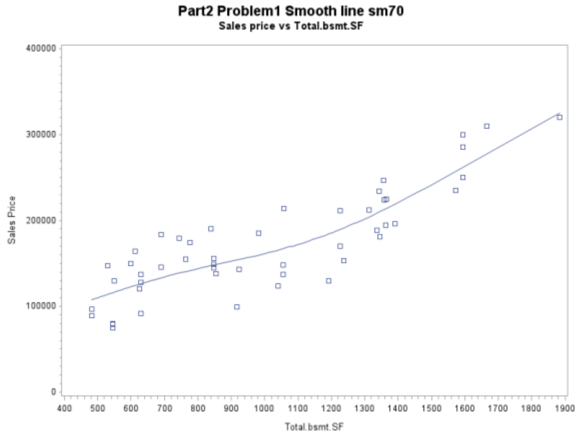


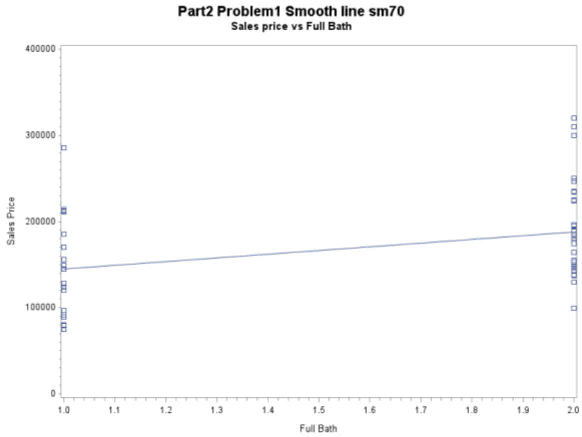


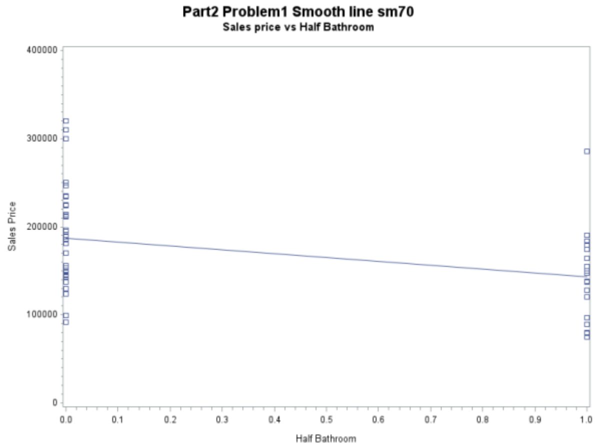


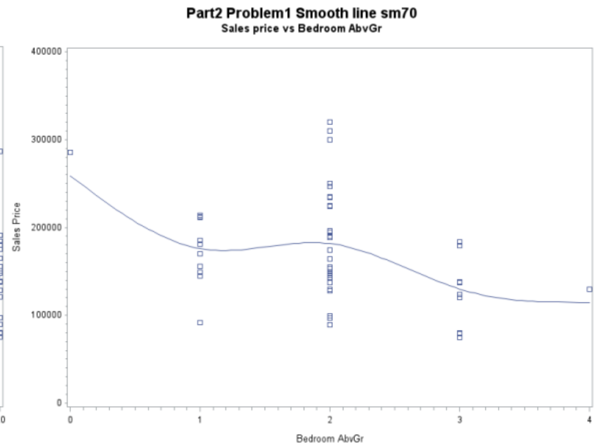

Part2 Problem1 Smooth line sm70
Sales price vs Year Remod Add
Sales Price
Year Remod Add

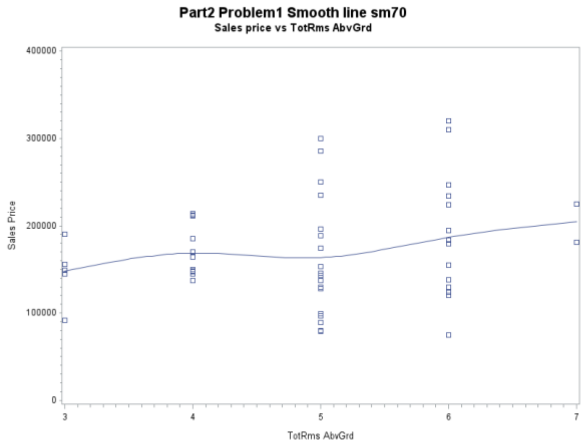

Part2 Problem1 Smooth line sm70
Sales price vs bsmtFin.SF.1
Sales Price
bsmtFin.SF.1

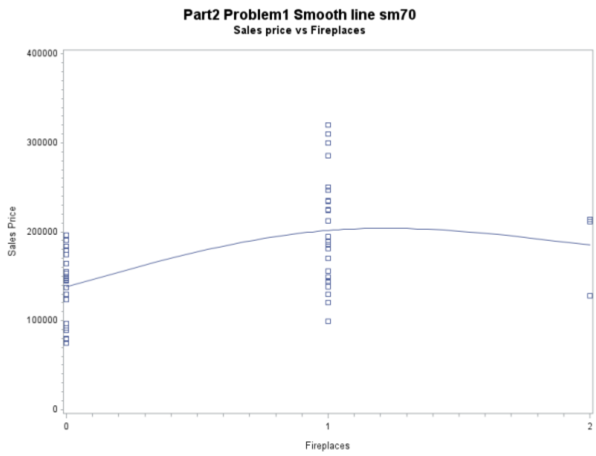

Part2 Problem1 Smooth line sm70
Sales price vs Total.bsmt.SF
Sales Price
Total.bsmt.SF

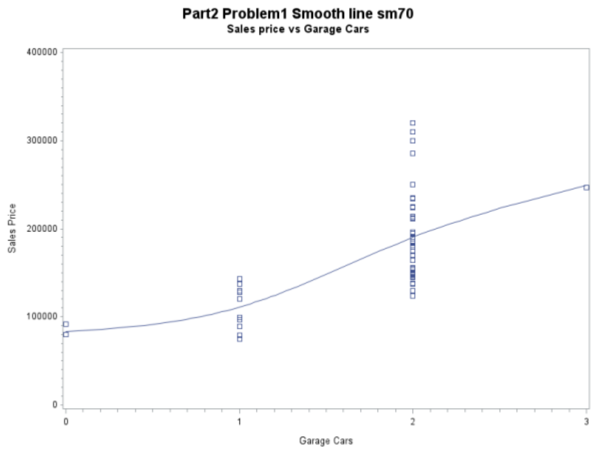

Part2 Problem1 Smooth line sm70
Sales price vs Full Bath
Sales Price
Full Bath


Part2 Problem1 Smooth line sm70
Sales price vs Half Bathroom
Sales Price
Half Bathroom


Part2 Problem1 Smooth line sm70
Sales price vs Bedroom AbvGr
Sales Price
Bedroom AbvGr







Correlation Matrix

Pearson Correlation Coefficients, N = 50

	LA	GLA	GA	OQ	OC	YB	YRA	BFSF	TBSF	FB	HB	BAG	TRAG	FP	GC	SP
LA	1.00000	0.12351	0.17550	0.24322	-0.12705	0.21886	0.21693	0.06362	0.37610	0.05842	-0.27688	-0.09719	0.17697	0.53880	0.09808	0.31770
GLA	0.12351	1.00000	0.60554	0.54325	-0.27325	0.49578	0.51169	0.24133	0.56078	0.56412	0.05700	0.19146	0.54612	0.18905	0.47699	0.71875
GA	0.17550	0.60554	1.00000	0.72980	-0.26823	0.66869	0.70478	0.35696	0.56430	0.34594	-0.17924	-0.26765	0.08967	0.21367	0.90331	0.74696
OQ	0.24322	0.54325	0.72980	1.00000	-0.11500	0.48688	0.52650	0.35696	0.67564	0.33690	-0.31534	-0.44220	0.01907	0.34913	0.61538	0.78841
OC	-0.12705	-0.27325	-0.26823	-0.11500	1.00000	-0.44437	-0.40934	-0.06436	-0.17452	-0.13023	-0.21418	0.06693	-0.14666	-0.03473	-0.38154	-0.21117
YB	0.21886	0.49578	0.66869	0.48688	-0.44437	1.00000	0.97884	0.27669	0.45500	0.49131	-0.17278	-0.27094	0.07638	0.04919	0.71339	0.65223
YRA	0.21693	0.51169	0.70478	0.52650	-0.40934	0.97884	1.00000	0.30863	0.47001	0.46290	-0.19281	-0.30999	0.05878	0.12031	0.73180	0.68488
BFSF	0.06362	0.24133	0.35696	0.35696	-0.06436	0.27669	0.30863	1.00000	0.59870	-0.04898	-0.44643	-0.31938	-0.04469	0.40751	0.38347	0.60873
TBSF	0.37610	0.56078	0.56430	0.67564	-0.17452	0.45500	0.47001	0.59870	1.00000	0.30389	-0.62387	-0.36249	0.24554	0.45301	0.50920	0.84919
FB	0.05842	0.56412	0.34594	0.33690	-0.13023	0.49131	0.46290	-0.04898	0.30389	1.00000	-0.09958	0.34408	0.39528	-0.17855	0.37302	0.34927
HB	-0.27688	0.05700	-0.17924	-0.31534	-0.21418	-0.17278	-0.19281	-0.44643	-0.62387	-0.09958	1.00000	0.37470	0.08432	-0.40278	-0.24750	-0.36238
BAG	-0.09719	0.19146	-0.26765	-0.44220	0.06693	-0.27094	-0.30999	-0.31938	-0.36249	0.34408	0.37470	1.00000	0.51768	-0.34794	-0.26433	-0.32449
TRAG	0.17697	0.54612	0.08967	0.01907	-0.14666	0.07638	0.05878	-0.04469	0.24554	0.39528	0.08432	0.51768	1.00000	0.09081	0.07729	0.20249
FP	0.53880	0.18905	0.21367	0.34913	-0.03473	0.04919	0.12031	0.40751	0.45301	-0.17855	-0.40278	-0.34794	0.09081	1.00000	0.16699	0.45711
GC	0.09808	0.47699	0.90331	0.61538	-0.38154	0.71339	0.73180	0.38347	0.50920	0.37302	-0.24750	-0.26433	0.07729	0.16699	1.00000	0.62792
SP	0.31770	0.71875	0.74696	0.78841	-0.21117	0.65223	0.68488	0.60873	0.84919	0.34927	-0.36238	-0.32449	0.20249	0.45711	0.62792	1.00000

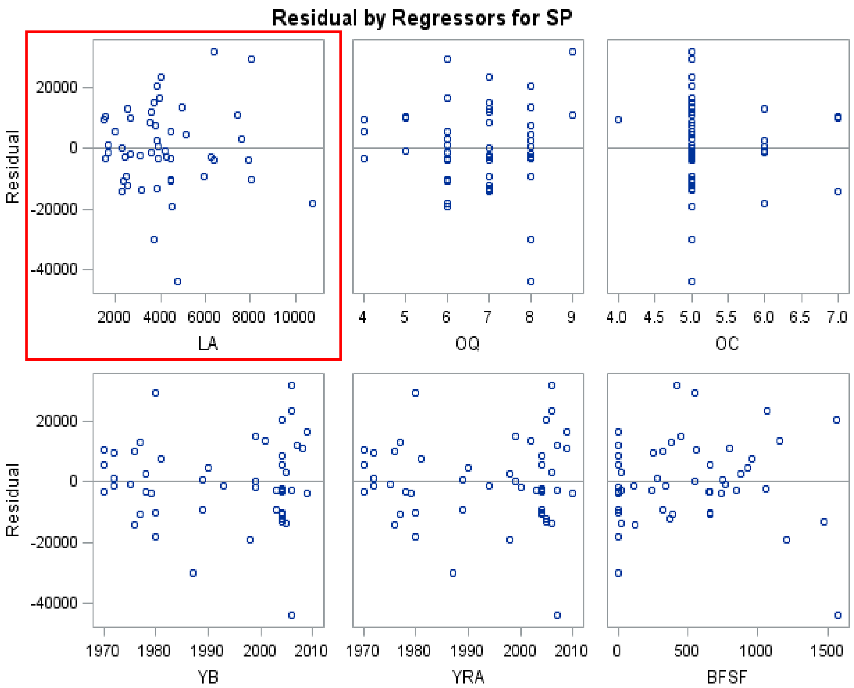
As the scatter plots shown above, most predictors have either strong or weak linear relationships with the response variable except for the full bathroom and the half bathroom. These two predictors could be categorical variables due to their behavior.

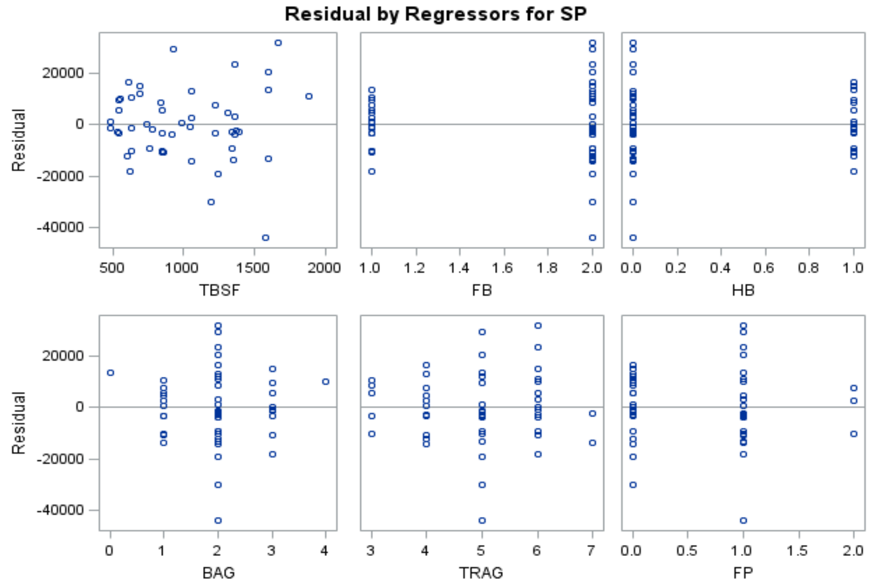
Also, according to the correlation matrix, larger absolute values show stronger linear relationships with the response variable, and some of the predictors including GLA, GA, OQ, YB, YRA, BFSF (labeled with blue), and TBSF which has the highest correlation coefficient with SP (labeled with red) each have a relative strong relationship with SP, whereas LA, OC, FB, HB, BAG, and TRAG (labeled with green) which are categorized by their low absolute correlation coefficient have weaker relationship with SP than others.

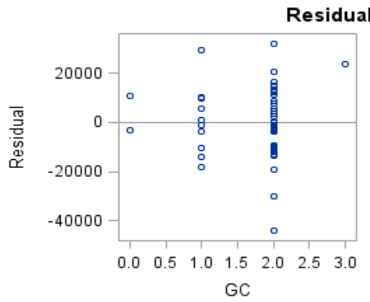
2.

According to SAS output (scatter plots and residual plots for each variable), we find that lot areas' (LA) scatterplots does not imply a linear relationship between LA itself and sale price. Also the residual plot of lot areas shows a cluster pattern which indicates that it does not follow constant variance principle, thus we need to transform this variable.

As shown in the graphs below, there is an additional variables need taking into account- BFSF. Recall its scatter plot, BFSF has a relative strong linear relationship with SP, so it is not necessary to transform BFSF in spite of its residual plot. We don't need to transform other variable since the rest shares a linear pattern with sales price and does not violate constant variance principle.







3. Original:

Prob 3 Original Cp criterion to select the best subset of variables
explanatory variables: LA GLA GA OQ OC YB YRA BFSF TBSF FB HB BAG TRAG FP GC

The REG Procedure
Model: MODEL1
Dependent Variable: SP

C(p) Selection Method

Number of Observations Read	50
Number of Observations Used	50

			Parameter Estimates															
Number in Model	C(p)	R-Square	Intercept	LA	GLA	GA	OQ	OC	YB	YRA	BFSF	TBSF	FB	HB	BAG	TRAG	FP	GC
10	9.6139	0.9471	-3145809	-2.89431	.	168.94319	11761	.	1570.22125	.	23.07147	74.21769	.	15491	9164.03323	.	19608	-42149
11	9.7155	0.9498	-3356370	-2.67932	.	144.17266	10934	7400.10783	1654.48729	.	22.06242	80.60829	.	21524	8531.55154	.	20250	-35079
8	10.0894	0.9408	-2425574	.	42.02961	153.04072	9928.56503	.	1209.30792	.	27.43719	44.90591	.	.	.	.	10430	-37025
10	10.4698	0.9459	-3104741	-2.27693	.	149.44962	11506	.	.	1548.55327	22.83539	77.99383	.	16743	9299.03484	.	16047	-38412
9	10.5910	0.9429	-2450079	.	32.61184	144.62304	9870.73801	.	1220.26047	.	27.23292	56.02616	.	9557.59604	.	.	11374	-34915
10	10.5986	0.9457	-2723985	.	29.20776	121.70125	9147.15257	7610.47916	1336.06389	.	25.81827	64.52058	.	16524	.	.	12674	-28275
10	10.7509	0.9455	-3111454	.	.	133.15778	11173	8349.53645	1524.04194	.	25.84474	77.56914	.	23333	7364.40844	.	15092	-30469
9	10.7675	0.9427	-2912109	.	.	143.79367	11838	.	.	1448.05555	26.30197	74.15514	.	17500	8393.31606	.	11772	-35738
12	10.8393	0.9510	-3672140	-2.92585	.	132.40383	12361	8057.14897	1810.42659	.	19.68734	83.72595	-7837.72301	20703	11881	.	19766	-33361
8	10.9253	0.9397	-2416327	.	42.36991	141.23278	9583.05165	.	.	1205.35490	26.44667	47.87263	.	.	.	.	8414.12781	-34953
10	10.9745	0.9452	-3113298	.	.	119.80692	10994	7084.35561	.	1528.75121	24.95478	80.93488	.	23284	7864.50602	.	12632	-28936
9	11.0214	0.9423	-2465894	.	31.63891	131.57853	9510.34527	.	.	1228.94748	26.20203	60.43458	.	10792	.	.	9473.97747	-32709
11	11.0519	0.9479	-3381710	-3.10522	.	161.34424	12953	.	1688.10818	.	21.25662	76.24279	-6223.76934	14413	11869	.	19179	-41283
11	11.1293	0.9478	-3261457	-2.05672	.	127.99699	10802	6174.36421	.	1609.16408	21.99652	83.53146	.	21857	8750.55620	.	16383	-32225
9	11.1928	0.9421	-2849005	.	.	160.38947	12137	.	1416.04323	.	27.34108	70.00562	.	16620	7978.69602	.	13889	-38108
11	11.2335	0.9477	-3188138	-2.72335	.	161.87877	11678	.	1024.54203	566.98843	22.94446	75.40942	.	15847	9299.52578	.	18502	-41248
12	11.2884	0.9504	-3392520	-2.76805	.	138.79602	10709	7680.36141	1673.62793	.	20.66507	85.72010	.	23145	10534	-2198.08635	21356	-33912
11	11.3204	0.9475	-3169094	-2.97436	.	165.27525	11601	.	1583.40153	.	21.94831	78.24200	.	16641	10839	-1816.46127	20502	-41406
12	11.4599	0.9502	-3384837	-2.54518	.	139.10892	10891	7176.95879	1203.10868	466.36506	21.98838	81.39582	.	21635	8662.07063	.	19320	-34551
9	11.4720	0.9417	-2552963	-1.21734	39.82779	159.63476	9618.12625	.	1276.95589	.	25.39446	47.78077	.	.	.	.	12738	-39088

Transformed:

Prob 3 Cp criterion to select the best subset of variables
explanatory variables: logLA GLA GA OQ OC YB YRA BFSF TBSF FB HB BAG TRAG FP GC

The REG Procedure
Model: MODEL1
Dependent Variable: SP

C(p) Selection Method

Number of Observations Read	50
Number of Observations Used	50

Number in Model	C(p)	R-Square	Parameter Estimates															
			Intercept	logLA	GLA	GA	OQ	OC	YB	YRA	BFSF	TBSF	FB	HB	BAG	TRAG	FP	GC
10	9.5070	0.9483	-3100833	-14696	.	176.81981	11529	.	1601.25353	.	22.98070	73.66908	.	13781	9034.53405	.	20272	-41461
11	9.7471	0.9507	-3303600	-13548	.	152.45672	10759	7076.53612	1678.32116	.	22.05304	79.79332	.	19692	8431.44040	.	20793	-34725
10	10.6522	0.9467	-3061725	-11566	.	155.43129	11316	.	.	1569.10181	22.74094	77.66755	.	15423	9192.82996	.	16527	-37774
8	10.8858	0.9408	-2425574	.	42.02961	153.04072	9928.56503	.	1209.30792	.	27.43719	44.90591	.	.	.	.	10430	-37025
12	10.9370	0.9518	-3593416	-14396	.	141.55554	12105	7702.05211	1824.33235	.	19.89544	82.62106	-7424.88617	18844	11570	.	20242	-32962
11	10.9870	0.9490	-3316844	-15450	.	169.87001	12653	.	1711.84356	.	21.33168	75.48557	-5898.88315	12692	11570	.	19797	-40534
11	11.0681	0.9489	-3130336	-15346	.	172.79158	11320	.	1620.30798	.	21.56367	78.59049	.	15095	11081	-2212.25611	21450	-40569
12	11.1682	0.9515	-3346033	-14249	.	146.80182	10486	7371.88034	1703.47687	.	20.38276	85.71542	.	21452	10762	-2547.18497	22172	-33416
11	11.1941	0.9487	-3141771	-13938	.	170.06434	11466	.	1108.36908	510.73423	22.86391	74.77781	.	14188	9165.21802	.	19250	-40691
10	11.3290	0.9457	-2723985	.	29.20776	121.70125	9147.15257	7610.47916	1336.06389	.	25.81827	64.52058	.	16524	.	.	12674	-28275
9	11.3591	0.9429	-2450079	.	32.61184	144.62304	9870.73801	.	1220.26047	.	27.23292	56.02616	.	9557.59604	.	.	11374	-34915
11	11.4171	0.9484	-3214292	-10429	.	134.34235	10665	5890.43390	.	1624.29194	21.97105	82.95924	.	20437	8674.50083	.	16774	-31918
9	11.4406	0.9428	-2583251	-8571.57509	37.60580	166.31715	9378.12636	.	1324.74593	.	24.52304	49.82697	.	.	.	.	14131	-39343
10	11.4842	0.9455	-3111454	.	.	133.15778	11173	8349.53645	1524.04194	.	25.84474	77.56914	.	23333	7364.40844	.	15092	-30469
11	11.4936	0.9483	-3066333	-14235	2.52185	174.93070	11422	.	1582.18250	.	23.14688	72.38707	.	13257	8550.98862	.	19913	-41116
12	11.5363	0.9510	-3331899	-12955	.	147.54486	10727	6888.28509	1270.61124	420.35070	21.98160	80.54292	.	19870	8555.04110	.	19938	-34271
9	11.5389	0.9427	-2912109	.	.	143.79367	11838	.	.	1448.05555	26.30197	74.15514	.	17500	8393.31606	.	11772	-35738
10	11.7120	0.9452	-3113298	.	.	119.80692	10994	7084.35561	.	1528.75121	24.95478	80.93488	.	23284	7864.50602	.	12632	-28936
9	11.7196	0.9424	-3250489	-17552	.	207.49165	11927	.	1692.30112	.	21.93403	62.30572	.	.	10790	.	20483	-47803
8	11.7375	0.9397	-2416327	.	42.36991	141.23278	9583.05165	.	.	1205.35490	26.44667	47.87263	.	.	.	.	8414.12781	-34953
12	11.7470	0.9507	-3302571	-13535	0.07207	152.40791	10756	7075.03293	1677.75974	.	22.05799	79.75538	.	19676	8417.74879	.	20783	-34717

According to the tables shown above, the R Square of all subsets are close to 1.

The best subset of variables for original data is the first one shown in the first table shown above:

It has 10 predictors with C(p) = 9.6139, R-square = 0.9471

The best subset of variables for transformed data is the first one shown in the second table shown above:

It has 10 predictors with C(p) = 9.5070, R-square = 0.9483

In comparison, the best subset of variables for transformed data has a relative smaller C(p) and higher R-square. Thus, we choose the transformed one as our best model.

Linear regression model: Selling Price = -3100833 + -14696*logLA + 176.81981*GA + 11529*OQ + 1601.25353*YB + 22.98070*BFSF + 73.66908*TBSF + 13781*HB + 9034.5405*BAG + 20272*FP + -41461*GC

4.

The stepwise option to report the best subset of variables for your data:

Part 2 Problem4 Stepwise Option Original

The REG Procedure
Model: MODEL1
Dependent Variable: SP

Number of Observations Read	50
Number of Observations Used	50

Stepwise Selection: Step 8

Variable BAG Entered: R-Square = 0.9358 and C(p) = 10.0197

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	8	1.622107E11	20276335013	74.71	<.0001
Error	41	11127719568	271407794		
Corrected Total	49	1.733384E11			

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	-3115547	462052	12339843153	45.47	<.0001
OQ	12277	3120.88926	4200327499	15.48	0.0003
OC	12936	5013.83613	1806712335	6.66	0.0136
YRA	1504.72321	229.32308	11685304511	43.05	<.0001
BFSF	22.04293	6.79019	2860200423	10.54	0.0023
TBSF	91.37718	12.46231	14591532093	53.76	<.0001
HB	32586	7433.66571	5215300570	19.22	<.0001
BAG	7667.00427	3758.67680	1129285518	4.16	0.0478
FP	14262	4653.89259	2548779964	9.39	0.0038

Part 2 Problem4 Stepwise Option Transformed

The REG Procedure
Model: MODEL1
Dependent Variable: SP

Number of Observations Read	50
Number of Observations Used	50

Stepwise Selection: Step 8

Variable BAG Entered: R-Square = 0.9358 and C(p) = 9.9797

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	8	1.622107E11	20276335013	74.71	<.0001
Error	41	11127719568	271407794		
Corrected Total	49	1.733384E11			

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	-3115547	462052	12339843153	45.47	<.0001
OQ	12277	3120.88926	4200327499	15.48	0.0003
OC	12936	5013.83613	1806712335	6.66	0.0136
YRA	1504.72321	229.32308	11685304511	43.05	<.0001
BFSF	22.04293	6.79019	2860200423	10.54	0.0023
TBSF	91.37718	12.46231	14591532093	53.76	<.0001
HB	32586	7433.66571	5215300570	19.22	<.0001
BAG	7667.00427	3758.67680	1129285518	4.16	0.0478
FP	14262	4653.89259	2548779964	9.39	0.0038

According to SAS output, we can easily find that transformed data has lower Cp value rather than original data. Therefore, we will use logLA as variable in our linear regression model.

Linear regression model:

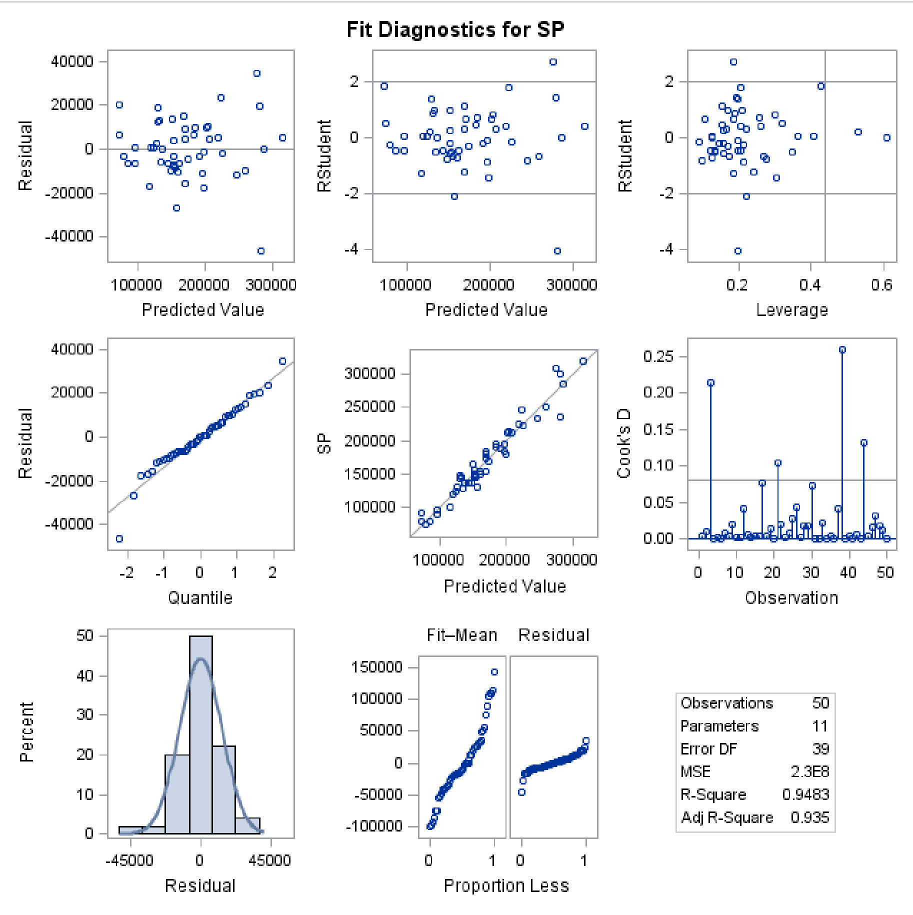
SP = -3115547 + 12277 * OQ + 12936 * OC + 1504.72321 * YRA + 22.04293 * BFSF + 91.37718 * TBSF + 32586 * HB + 7667.00427 * BAG + 14262 * FP

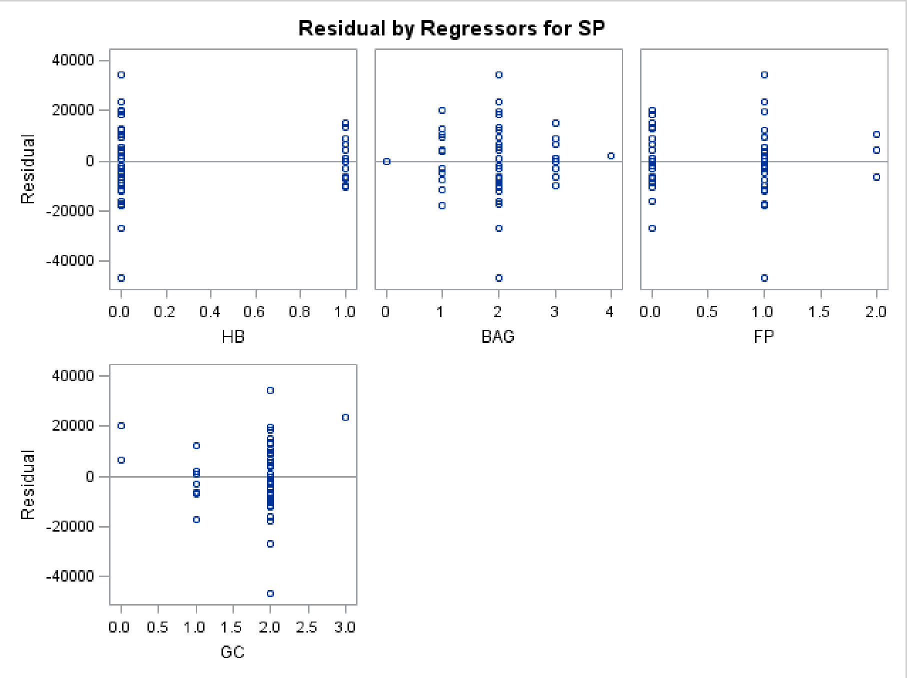
5.

Model: Selling Price = -3100833 + -14696*logLA + 176.81981*GA + 11529*OQ + 1601.25353*YB + 22.98070*BFSF + 73.66908*TBSF + 13781*HB + 9034.5405*BAG + 20272*FP + -41461*GC

Prob 5 Check assumption of this best model using all the usual plots
Transformed data

The REG Procedure
Model: MODEL1
Dependent Variable: SP





a. Linear relationship

From problem 1, we could know SP has a linearly relationship with most of the variables individually except FB & HB. Therefore, the linear relationship is somewhat satisfied.

b. Constant Variance

According to the residuals plots, there is no obvious pattern in each plots and each has a constant variance. Therefore, there is no violation of the constant variance assumption.

c. Residual Normally Distributed

As shown in the histogram, the residual looks approximately normally distributed. Our qq plot also verifies the assumption.

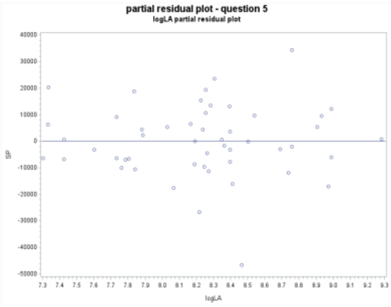
d. Independence

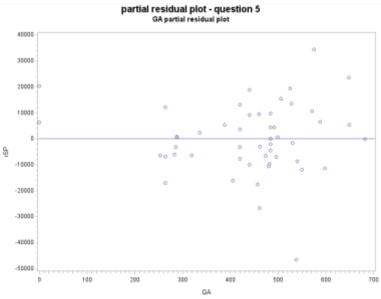
According to the SAS output, the predictors are independent. Therefore, the independence assumption is satisfied.

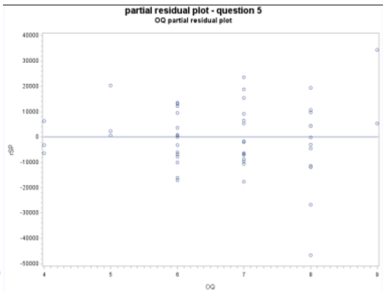
6.Examining other diagnostics of the same “best” model:

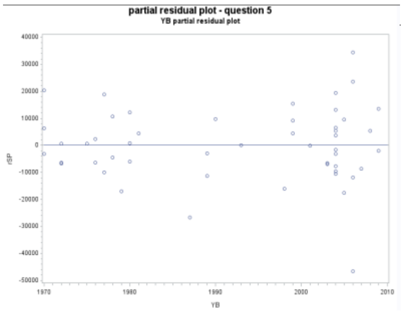
Selling Price = -3100833 + -14696*logLA + 176.81981*GA + 11529*OQ + 1601.25353*YB + 22.98070*BFSF + 73.66908*TBSF + 13781*HB + 9034.5405*BAG + 20272*FP + -41461*GC

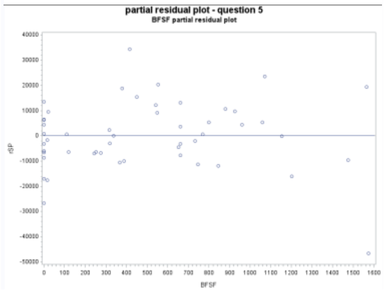
The partial residual plots for the predictors (logLA, GA, OQ, YB, BFSF, TBSF, HB, BAG, FP, and GC) are shown in the figures below. There are no obvious outlier, but one point exist as a potential outlier, so we further check with Cook’s D method.

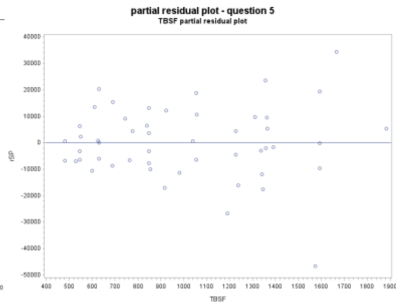


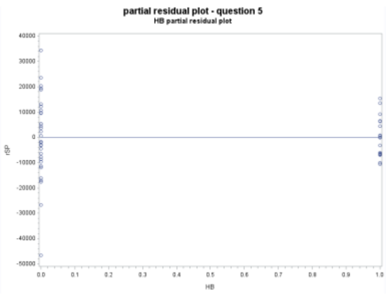


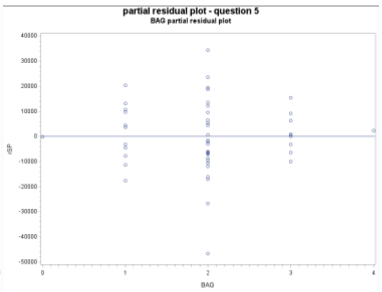


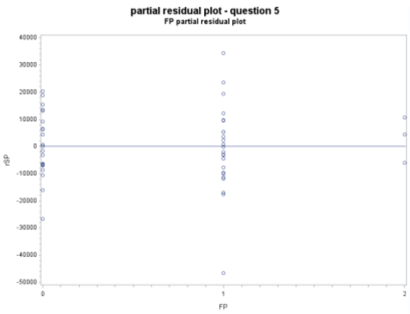


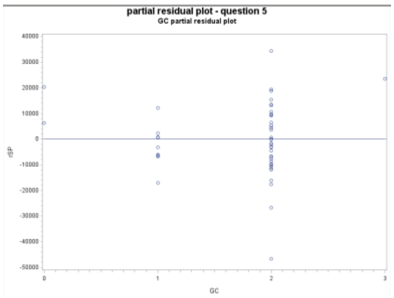












Cook'sD:

Simple Linear Regression with Diagnostic Plots

The REG Procedure
Model: MODEL1
Dependent Variable: SP

Output Statistics

Obs	Dependent Variable	Predicted Value	Std Error Mean Predict	Residual	Std Error Residual	Student Residual	-2-1 0 1 2	Cook's D	RStudent	Hat Diag H	Cov Ratio	DFFITS
1	234250	246064	4753	-11814	14402	-0.820	| *| |	0.007	-0.8168	0.0982	1.2185	-0.2695
2	246578	222951	6896	23627	13508	1.749	| |*** |	0.072	1.7985	0.2067	0.6829	0.9180
3	147110	154062	5573	-6952	14105	-0.493	| | |	0.003	-0.4881	0.1360	1.4363	-0.1928
4	225000	219691	6837	5309	13637	0.392	| | |	0.004	0.3879	0.2032	1.5991	0.1959
5	212000	202280	4977	9720	14326	0.678	| |* |	0.005	0.6737	0.1077	1.3087	0.2340
6	179400	170336	6518	9064	13693	0.662	| |* |	0.009	0.6571	0.1847	1.4415	0.3128
7	120000	119125	9136	875.0150	12106	0.072	| | |	0.000	0.0714	0.3629	2.0857	0.0539
8	224000	225945	4522	-1945	14476	-0.134	| | |	0.000	-0.1327	0.0889	1.4532	-0.0414
9	196000	197594	6707	-1594	13602	-0.117	| | |	0.000	-0.1157	0.1956	1.6479	-0.0570
10	320000	314686	7727	5314	13049	0.407	| | |	0.005	0.4028	0.2596	1.7150	0.2385
11	185000	196380	7001	-11380	13453	-0.846	| *| |	0.018	-0.8428	0.2131	1.3795	-0.4386
12	164500	150925	6230	13575	13827	0.982	| |* |	0.018	0.9813	0.1688	1.2157	0.4422
13	183900	168573	5979	15327	13937	1.100	| |** |	0.020	1.1027	0.1554	1.1142	0.4730
14	137900	147914	7979	-10014	12897	-0.776	| *| |	0.021	-0.7724	0.2768	1.5602	-0.4778
15	235000	281676	6701	-46676	13606	-3.431	|******| |	0.260	-4.0630	0.1963	0.0318	-1.9964
16	310090	275667	6557	34423	13675	2.517	| |***** |	0.132	2.7151	0.1869	0.2328	1.3018
17	130000	156742	7154	-26742	13372	-2.000	| ***| |	0.104	-2.0837	0.2225	0.5206	-1.1147

Obs	DFBETAS Intercept	logLA	GA	OQ	YB	BFSF	TBSF	HB	BAG	FP	GC
1	0.1220	-0.0404	-0.0609	-0.0898	-0.1105	-0.0751	0.0748	0.0809	-0.1099	-0.0122	0.1158
2	0.1339	-0.1715	-0.1133	-0.3090	-0.1136	0.0287	0.1261	-0.0062	0.1202	0.2088	0.4989
3	0.0675	0.0334	-0.0046	-0.0642	-0.0679	-0.0194	0.1019	-0.0124	-0.0020	0.0028	0.0046
4	-0.0716	-0.0746	-0.1207	0.0314	0.0752	0.0187	0.0463	0.0140	0.0569	0.0678	0.0797
5	0.1067	0.0913	-0.0778	0.0854	-0.1142	0.0718	-0.0051	0.0321	-0.0923	-0.0670	0.0778
6	-0.0629	-0.0379	-0.1413	0.1590	0.0578	0.0886	-0.0569	0.0843	0.1609	-0.0111	0.1093
7	0.0047	0.0406	-0.0109	0.0109	-0.0091	0.0066	-0.0081	0.0181	0.0098	-0.0057	0.0024
8	0.0242	-0.0061	0.0055	0.0055	-0.0228	0.0004	-0.0070	0.0031	-0.0106	-0.0077	0.0018
9	0.0049	0.0052	-0.0138	0.0181	-0.0057	0.0349	-0.0289	0.0121	0.0032	0.0149	0.0089
10	-0.0476	0.0064	0.0956	0.0194	0.0427	-0.0163	0.0839	-0.0001	0.0652	-0.0139	-0.1299
11	-0.0849	0.0303	-0.2774	-0.0474	0.0761	-0.0789	0.2432	0.2002	0.0990	-0.0118	0.1971
12	-0.1653	0.0632	0.1627	-0.1651	0.1608	-0.0730	-0.0758	0.0583	-0.0999	-0.0427	-0.1011
13	-0.0505	0.1553	-0.0321	0.1963	0.0241	0.1933	-0.1981	0.0892	0.2194	-0.1416	0.0288
14	-0.1873	0.1469	0.1096	0.0746	0.1703	0.0919	-0.1479	-0.2017	-0.1275	-0.2213	-0.2449
15	0.6131	-0.0066	-0.1272	-0.4290	-0.5702	-1.3632	0.0715	0.0016	-0.6646	0.0909	0.5881
16	-0.3429	-0.0943	0.0515	0.4896	0.3165	-0.4772	0.3994	-0.0816	0.4228	0.1294	-0.3217
17	-0.4511	-0.1510	0.3942	-0.4907	0.4675	0.4452	-0.1850	0.1793	-0.0502	0.4715	-0.4582

……

There are 50 observations in total, and according to the fomulation, 2*p/n=11/50=0.22

Only 15th observation has cook's D larger than 0.2.

Observation #15 (0.26) seems to have a lot of influence.

Summary of Tests

Diagnostic Test	Significant Values	Conclusion
Partial Residual Plots	No outlier	All predictors are of value
Cook's D	Observation 15 is an outlier with a lot of influence	Need to take further investigate for that outlier

7.

I. Equation of the regressional model

Selling Price = -3100833 + -14696*logLA + 176.81981*GA + 11529*OQ + 1601.25353*YB + 22.98070*BFSF + 73.66908*TBSF + 13781*HB + 9034.5405*BAG + 20272*FP + -41461*GC

II. 90% confidence interval for the mean of the response variable

The 1st fifteen observations are shown below:

Part 2 Problem7 Confidence interval and Prediction interval
alpha=0.1

The REG Procedure
Model: MODEL1
Dependent Variable: SP

Output Statistics

Obs	Dependent Variable	Predicted Value	Std Error Mean Predict	90% CL Mean		90% CL Predict		Residual
1	92000	82204	11445	62882	101527	48605	115803	9796
2	79500	82208	10193	64999	99418	49778	114638	-2708
3	137500	151435	9561	135294	167577	119559	183312	-13935
4	97000	98646	6972	86875	110418	68744	128548	-1646
5	89000	88503	7399	76011	100995	58310	118696	496.9408
6	79400	71053	10117	53973	88133	38691	103414	8347
7	75000	68522	7637	55629	81415	38161	98883	6478
8	99500	102883	9444	86939	118827	71106	134660	-3383
9	143000	113400	10486	95695	131104	80704	146095	29600
10	120000	140162	11268	121138	159186	106733	173591	-20162
11	129500	117890	11757	98040	137740	83985	151795	11610
12	128000	137354	9798	120812	153895	105273	169435	-9354
13	145000	139401	8006	125885	152917	108770	170032	5599
14	190000	180721	9041	165458	195985	149280	212162	9279
15	148500	136825	8319	122780	150871	105957	167693	11675

III. 90% prediction interval for individual observations

The 1st fifteen observations are shown below:

Part 2 Problem7 Confidence interval and Prediction interval
alpha=0.1

The REG Procedure
Model: MODEL1
Dependent Variable: SP

Output Statistics								
Obs	Dependent Variable	Predicted Value	Std Error Mean Predict	90% CL Mean		90% CL Predict		Residual
1	92000	82204	11445	62882	101527	48605	115803	9796
2	79500	82208	10193	64999	99418	49778	114638	-2708
3	137500	151435	9561	135294	167577	119559	183312	-13935
4	97000	98646	6972	86875	110418	68744	128548	-1646
5	89000	88503	7399	76011	100995	58310	118696	496.9408
6	79400	71053	10117	53973	88133	38691	103414	8347
7	75000	68522	7637	55629	81415	38161	98883	6478
8	99500	102883	9444	86939	118827	71106	134660	-3383
9	143000	113400	10486	95695	131104	80704	146095	29600
10	120000	140162	11268	121138	159186	106733	173591	-20162
11	129500	117890	11757	98040	137740	83985	151795	11610
12	128000	137354	9798	120812	153895	105273	169435	-9354
13	145000	139401	8006	125885	152917	108770	170032	5599
14	190000	180721	9041	165458	195985	149280	212162	9279
15	148500	136825	8319	122780	150871	105957	167693	11675

IV. 90% confidence intervals for the regression coefficients

Parameter Estimates							
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	90% Confidence Limits	
Intercept	1	-4054961	628719	-6.45	<.0001	-5116427	-2993496
logLA	1	-11542	7512.32386	-1.54	0.1332	-24225	1140.62706
OQ	1	15042	3650.16613	4.12	0.0002	8879.72159	21205
OC	1	12634	5298.02237	2.38	0.0225	3689.60867	21579
YB	1	1057.42935	971.07221	1.09	0.2834	-582.02964	2696.88834
YRA	1	968.83899	943.19376	1.03	0.3112	-623.55287	2561.23086
BFSF	1	15.00658	7.90286	1.90	0.0656	1.66420	28.34896
TBSF	1	103.15117	14.92946	6.91	<.0001	77.94580	128.35655
FB	1	-12911	8441.78136	-1.53	0.1349	-27163	1341.38350
HB	1	29590	8023.39947	3.69	0.0007	16044	43136
BAG	1	17698	6115.12803	2.89	0.0064	7374.27282	28023
TRAG	1	-3880.83061	3487.56083	-1.11	0.2732	-9768.87159	2007.21038
FP	1	19825	6308.23122	3.14	0.0033	9174.88468	30475
GC	1	-6677.37014	7041.62286	-0.95	0.3493	-18566	5210.98563

APPENDIX

PART I

```
data HPD;
infile 'W:\STAT512\data\HPD.DAT';
input LA GLA GA OQ OC YB YRA BFSF TBSF FB HB BAG TRAG FP GC SP;
```

1.

```
proc sort data=HPD;
        by TBSF;
title1'Part1 Problem1 Piecewise';
title2'Sales price vs TBSF';
axis1 label=('Gr.Liv.Area');
axis2 label=(angle=90 'Sales Price');
symbol1 v=square i=sm70;
proc gplot data=HPD;
        plot SP*TBSF/haxis=axis1 vaxis=axis2;
run;
data pie;
        set HPD;
        if TBSF le 1200
        then cslope=0;
        if TBSF gt 1200
        then cslope = TBSF-1200;
proc print data=pie;
run;
proc reg data=pie;
        model SP=TBSF cslope;
        output out=pieceout p=costhat;
symbol1 v=circle i=none c=black;
symbol2 v=none i=join c=red;
proc sort data=pieceout;
        by TBSF;
proc gplot data=pieceout;
        plot (sp costhat)*TBSF/overlay;
run;
```

2.

a)

```
data HPD;
        set HPD;
        SUM = GLA + GA;
proc reg data = HPD;
        model SP = LA OQ OC YB YRA BFSF TBSF FB HB BAG TRAG FP GC;
proc reg data = HPD;
```

```
        model SP = LA SUM OQ OC YB YRA BFSF TBSF FB HB BAG TRAG FP GC;
run;
b)
data HPD;
        set HPD;
        SUM = GLA + GA;
proc reg data = HPD;
        model SP = LA SUM OQ OC YB YRA BFSF TBSF FB HB BAG TRAG FP GC;
        test SUM;
run;
```

3.

```
proc reg data = HPD;
        model SP = LA GLA GA OQ OC YB YRA BFSF TBSF FB HB BAG TRAG FP GC
/ SS1 SS2;
run;
```

4.

```
data HPD;
        set HPD;
        SUM = GLA + GA;
proc reg data = HPD;
        model SP = GLA / ADJRSQ;
        model SP = GA / ADJRSQ;
        model SP = GLA GA / ADJRSQ;
        model SP = SUM / ADJRSQ;
        model SP = YB / ADJRSQ;
        model SP = YRA / ADJRSQ;
        model SP = BFSF / ADJRSQ;
        model SP = TBSF / ADJRSQ;
        model SP = GC / ADJRSQ;
        model SP = TBSF SUM / ADJRSQ;
run;
```

PARTII

1.

```
data HPD;
infile 'W:\STAT512\data\HPD.DAT';
input LA GLA GA OQ OC YB YRA BFSF TBSF FB HB BAG TRAG FP GC SP;

/*LA*/
proc sort data=HPD;
by LA;
title1'Part2 Problem1 Smooth line sm70';
```

```
title2'Sales price vs Lot area';
axis1 label=('Lot Area');
axis2 label=(angle=90 'Sales Price');
symbol1 v=square i=sm70;
proc gplot data=HPD;
plot SP*LA/haxis=axis1 vaxis=axis2;
run;
proc reg data=HPD;
/*GlA*/
proc sort data=HPD;
by GlA;
title1'Part2 Problem1 Smooth line sm70';
title2'Sales price vs Gr.Liv.Area';
axis1 label=('Gr.Liv.Area');
axis2 label=(angle=90 'Sales Price');
symbol1 v=square i=sm70;
proc gplot data=HPD;
plot SP*GlA/haxis=axis1 vaxis=axis2;
run;
```

Same pattern for the other variables.

```
/*Correlation Matrix*/
proc corr data = HPD noprob;
        var LA GLA GA OQ OC YB YRA BFSF TBSF FB HB BAG TRAG FP GC SP;
run;
data trans; set HPD;
logLA = log(LA);
rsLA = LA**(-0.5);
proc print data = trans;
run;
```

2.

```
title1'Part 2 Problem2 Transformed Variable';
title2'Transformed Variable: LA';
title1'Log Transformation';
proc reg data = trans;
model logLA = SP;
output out = logtrans r = logresid;
symbol1 i=rl;
proc gplot data = logtrans;
plot logLA * SP;
```

3.

```
title1'Prob 3 Original Cp criterion to select the best subset of variables';
```

```
title2'explanatory variables: LA GLA GA OQ OC YB YRA BFSF TBSF FB HB BAG
TRAG FP GC';
proc reg data=HPD;
model SP = LA GLA GA OQ OC YB YRA BFSF TBSF FB HB BAG TRAG FP GC/
selection = cp b;
run;
title1'Prob 3 transformed Cp criterion to select the best subset of variables';
title2'explanatory variables: logLA GLA GA OQ OC YB YRA BFSF TBSF FB HB BAG
TRAG FP GC';
proc reg data=trans;
model SP = logLA GLA GA OQ OC YB YRA BFSF TBSF FB HB BAG TRAG FP GC/
selection = cp b;
run;
```

4.

```
title1'Part 2 Problem4 Stepwise Option';
title2'Original';
proc reg data = HPD;
model SP = LA OQ OC YB YRA BFSF TBSF FB HB BAG TRAG FP GC/selection =
stepwies;
run;
title1'Part 2 Problem4 Stepwise Option';
title2'Transformed';
proc reg data = trans;
model SP = logLA OQ OC YB YRA BFSF TBSF FB HB BAG TRAG FP GC/selection =
stepwies;
run;
```

5.

```
title1'Prob 5 Check assumption of this best model using all the usual plots';
title2'Transformed data';
proc reg data=trans;
model SP = logLA GLA GA OQ OC YB YRA BFSF TBSF FB HB BAG TRAG FP GC/
selection = cp b;
run;
```

6.

```
title1 'partial residual plot - question 5';
title2  'logLA partial residual plot';
proc reg data=trans;
model SP = logLA GA OQ YB BFSF TBSF HB BAG FP GC;
output out = plLA r=residSP;
symbol1 v=circle i=rl;
axis1 label=('logLA');
```

```
axis2 label=(angle=90 'rSP');
proc gplot data=plLA;
plot residSP*logLA/ haxis = axis1 vaxis=axis2 vref=0;
Run;
title1 'partial residual plot - question 5';
title2  'GA partial residual plot';
proc reg data=trans;
model SP = logLA GA OQ YB BFSF TBSF HB BAG FP GC;
output out = pGA r=residSP;
symbol1 v=circle i=rl;
axis1 label=('GA');
axis2 label=(angle=90 'rSP');
proc gplot data=pGA;
plot residSP*GA/ haxis = axis1 vaxis=axis2 vref=0;
Run;
title1 'partial residual plot - question 5';
title2  'OQ partial residual plot';
proc reg data=trans;
model SP = logLA GA OQ YB BFSF TBSF HB BAG FP GC;
output out = pOQ r=residSP;
symbol1 v=circle i=rl;
axis1 label=('OQ');
axis2 label=(angle=90 'rSP');
proc gplot data=pOQ;
plot residSP*OQ/ haxis = axis1 vaxis=axis2 vref=0;
Run;
title1 'partial residual plot - question 5';
title2  'YB partial residual plot';
proc reg data=trans;
model SP = logLA GA OQ YB BFSF TBSF HB BAG FP GC;
output out = pYB r=residSP;
symbol1 v=circle i=rl;
axis1 label=('YB');
axis2 label=(angle=90 'rSP');
proc gplot data=pGA;
plot residSP*YB/ haxis = axis1 vaxis=axis2 vref=0;
Run;
title1 'partial residual plot - question 5';
title2  'BFSF partial residual plot';
proc reg data=trans;
model SP = logLA GA OQ YB BFSF TBSF HB BAG FP GC;
output out = pBFSF r=residSP;
symbol1 v=circle i=rl;
axis1 label=('BFSF');
```

```
axis2 label=(angle=90 'rSP');
proc gplot data=pBFSF;
plot residSP*BFSF/ haxis = axis1 vaxis=axis2 vref=0;
Run;
title1 'partial residual plot - question 5';
title2  'TBSF partial residual plot';
proc reg data=trans;
model SP = logLA GA OQ YB BFSF TBSF HB BAG FP GC;
output out = pTBSF r=residSP;
symbol1 v=circle i=rl;
axis1 label=('TBSF');
axis2 label=(angle=90 'rSP');
proc gplot data=pTBSF;
plot residSP*TBSF/ haxis = axis1 vaxis=axis2 vref=0;
Run;
title1 'partial residual plot - question 5';
title2  'HB partial residual plot';
proc reg data=trans;
model SP = logLA GA OQ YB BFSF TBSF HB BAG FP GC;
output out = pHB r=residSP;
symbol1 v=circle i=rl;
axis1 label=('HB');
axis2 label=(angle=90 'rSP');
proc gplot data=pHB;
plot residSP*HB/ haxis = axis1 vaxis=axis2 vref=0;
Run;
title1 'partial residual plot - question 5';
title2  'BAG partial residual plot';
proc reg data=trans;
model SP = logLA GA OQ YB BFSF TBSF HB BAG FP GC;
output out = pBAG r=residSP;
symbol1 v=circle i=rl;
axis1 label=('BAG');
axis2 label=(angle=90 'rSP');
proc gplot data=pBAG;
plot residSP*BAG/ haxis = axis1 vaxis=axis2 vref=0;
Run;
title1 'partial residual plot - question 5';
title2  'FP partial residual plot';
proc reg data=trans;
model SP = logLA GA OQ YB BFSF TBSF HB BAG FP GC;
output out = pFP r=residSP;
symbol1 v=circle i=rl;
axis1 label=('FP');
```

```
axis2 label=(angle=90 'rSP');
proc gplot data=pHB;
plot residSP*FP/ haxis = axis1 vaxis=axis2 vref=0;
Run;
title1 'partial residual plot - question 5';
title2  'GC partial residual plot';
proc reg data=trans;
model SP = logLA GA OQ YB BFSF TBSF HB BAG FP GC;
output out = pGC r=residSP;
symbol1 v=circle i=rl;
axis1 label=('GC');
axis2 label=(angle=90 'rSP');
proc gplot data=pGC;
plot residSP*GC/ haxis = axis1 vaxis=axis2 vref=0;
Run;
proc reg data=trans;
model SP = LA GA OQ YB  BFSF TBSF HB BAG FP GC/r influence;
run;
proc reg data=trans;
model SP = logLA GA OQ YB BFSF TBSF HB BAG FP GC/r influence;
run;
```

7.

```
title1'Part 2 Problem7 Confidence interval and Prediction interval';
title2'alpha=0.1';
proc reg data = trans alpha = 0.1;
model SP = logLA OQ OC YB YRA BFSF TBSF FB HB BAG TRAG FP GC/clb cli clm;
run;
```

Notation:
Abbreviation of variables
Lot.Area= LA
Gr.Liv.Area=GLA
Garage.Area=GA
Overall.Qual=OQ
Overall.Cond=OC
Year.Built=YB
Year.Remod.Add=YRA
BsmtFin.SF.1=BFSF
Total.Bsmt.SF=TBSF
Full.Bath= FB
Half.Bath=HB
Bedroom.AbvGr= BAG
TotRms.AbvGrd=TRAG

Fireplaces=FP
Garage.Cars=GC
SalePrice=SP