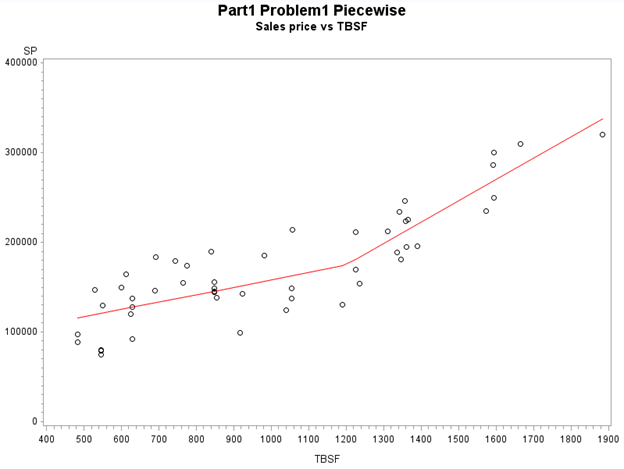
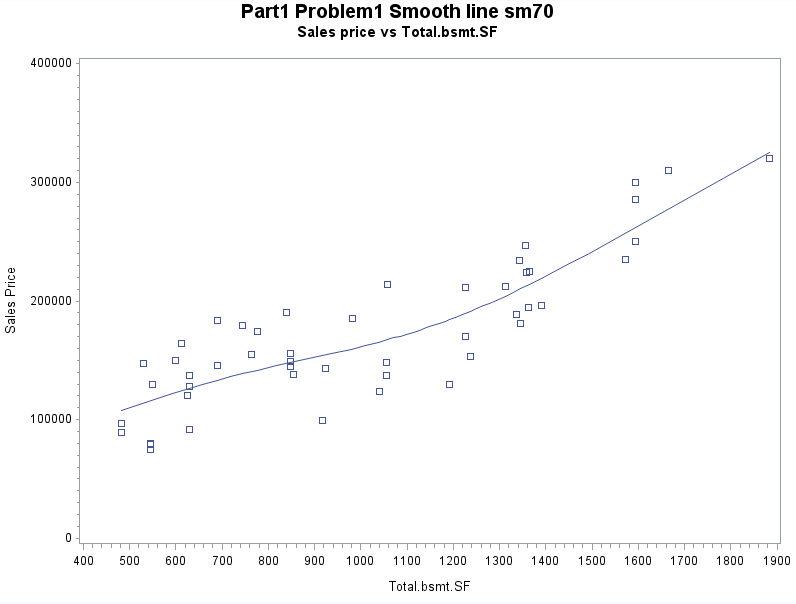
Chuyun Li; Mengzhi Zhou; Xudan Luo; Xin Li;

Data Analysis

HOUSING PRICING

**Part I**

1.

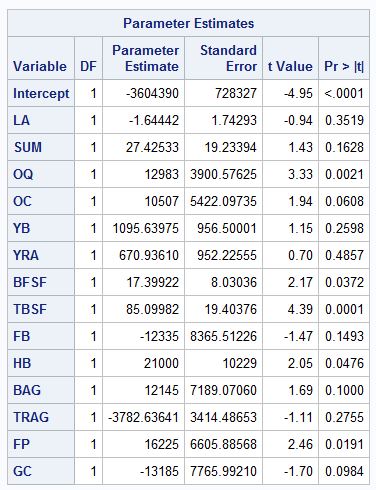
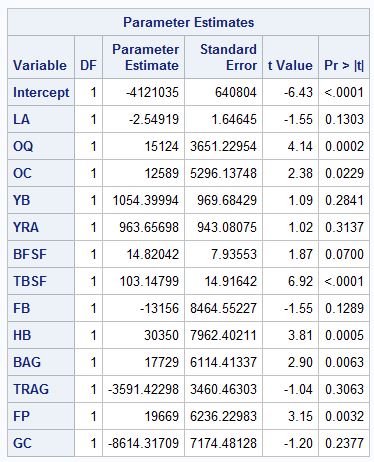


scatter plot with smoothing line piecewise SLR model

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.

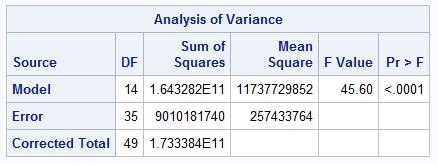
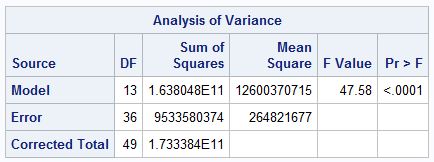


Using all the explanatory variables:

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



without SUM 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)]

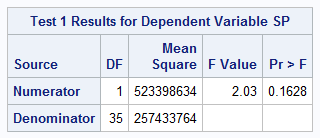
= 523398634 / (36 - 35) / (9010181740 / 35)

= 2.033139033

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

denominator = dfE for the larger model = 35

b)



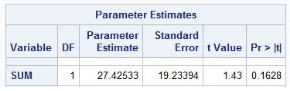
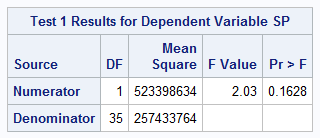
H0: the coefficient of the SUM variable is zero

Ha: 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)

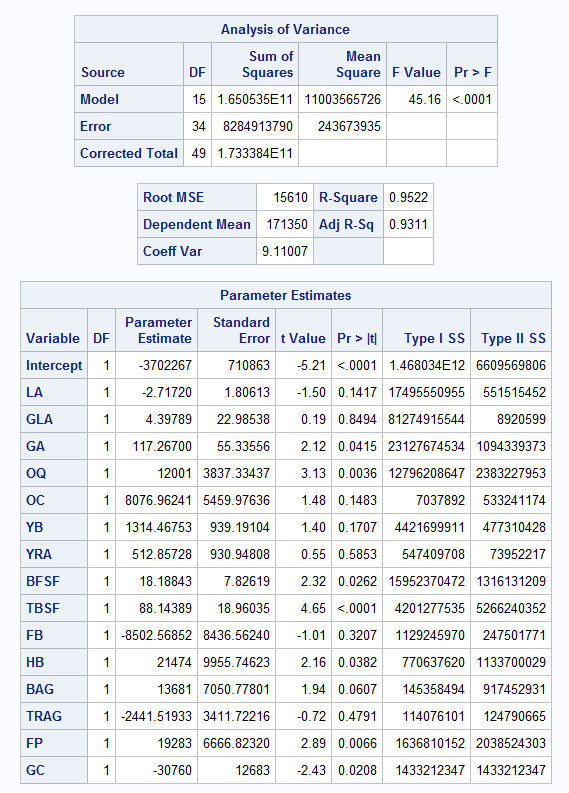


test statement: F = 2.03, p = 0.1628

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

(1.43)^2 ≈ 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.



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.

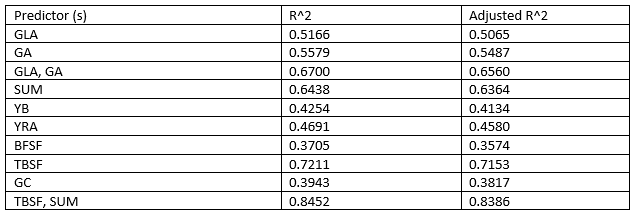
From definition, SSM(LA) + SSM(GLA|LA) + SSM(GA|LA,GLA) + SSM(OG|LA,GLA,GA) + SSM(OC|LA,GLA,GA,OQ) + SSM(YB|LA,GLA,GA,OQ,OC) + SSM(YRA|LA,GLA,GA,OQ,OC,YB) +

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).

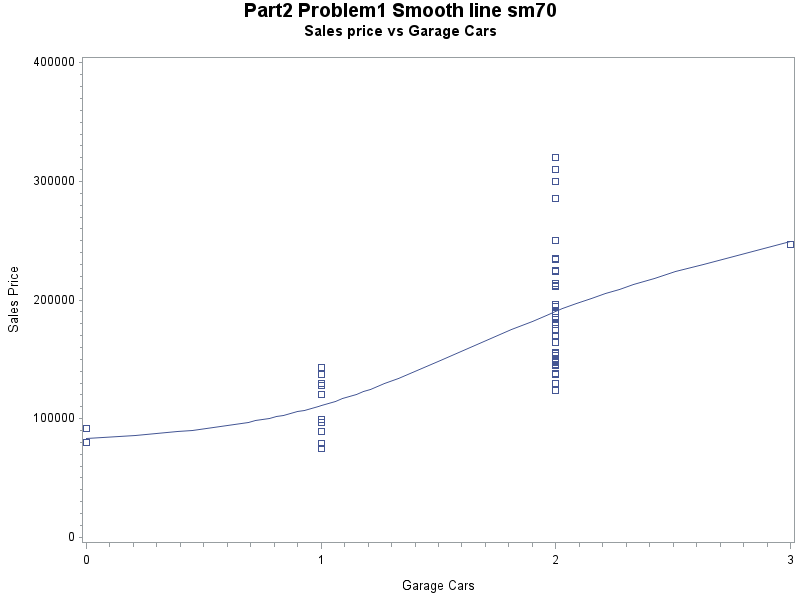
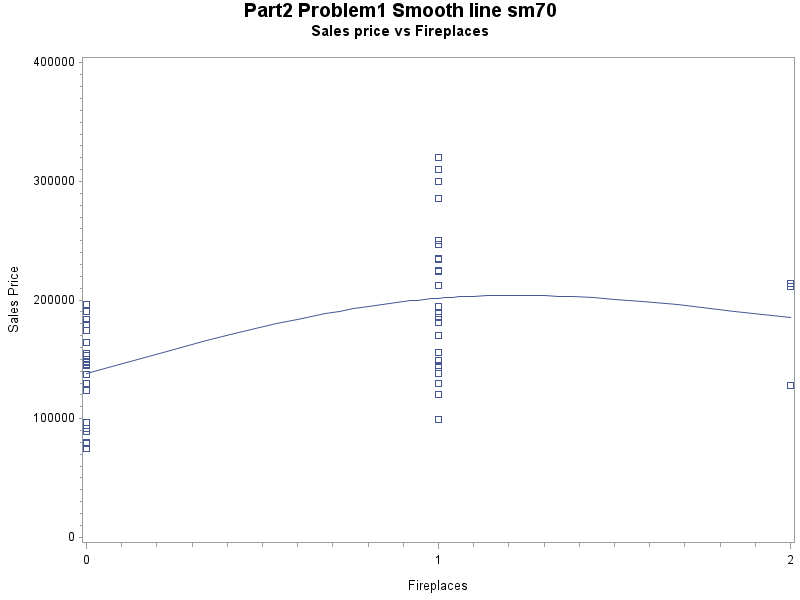
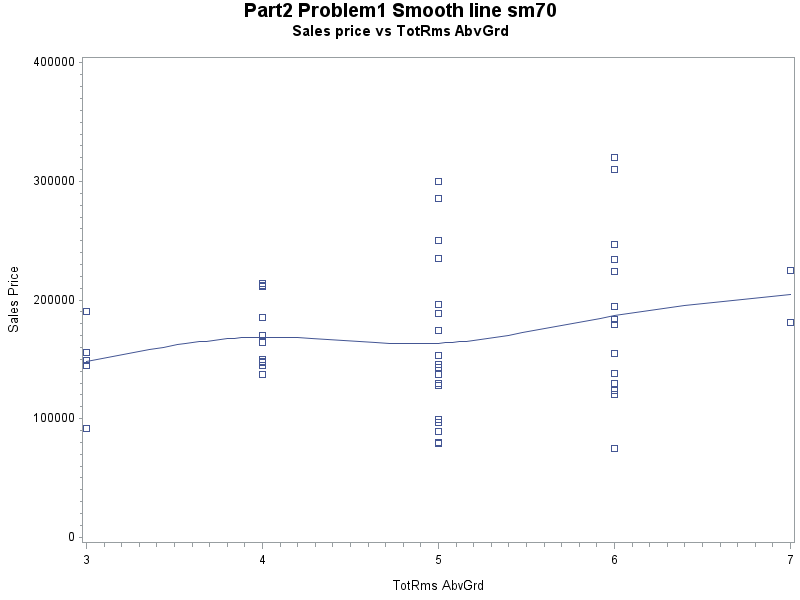
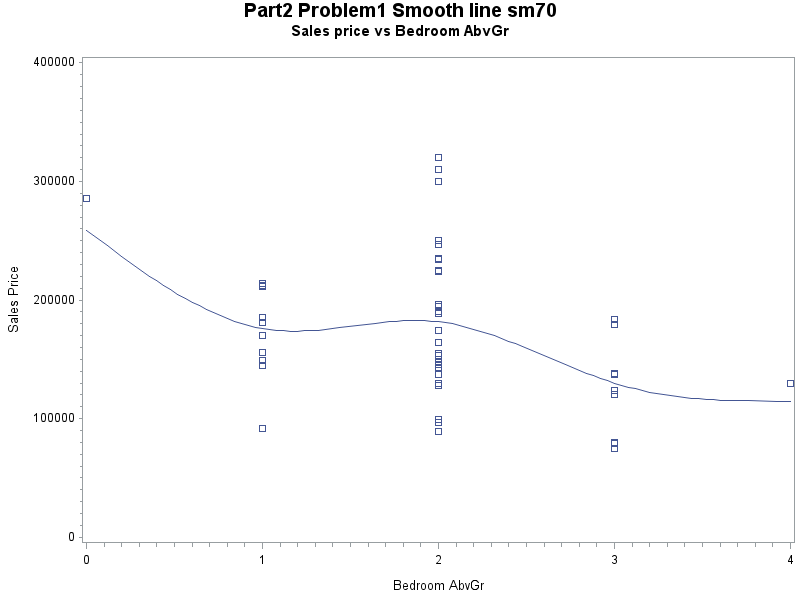
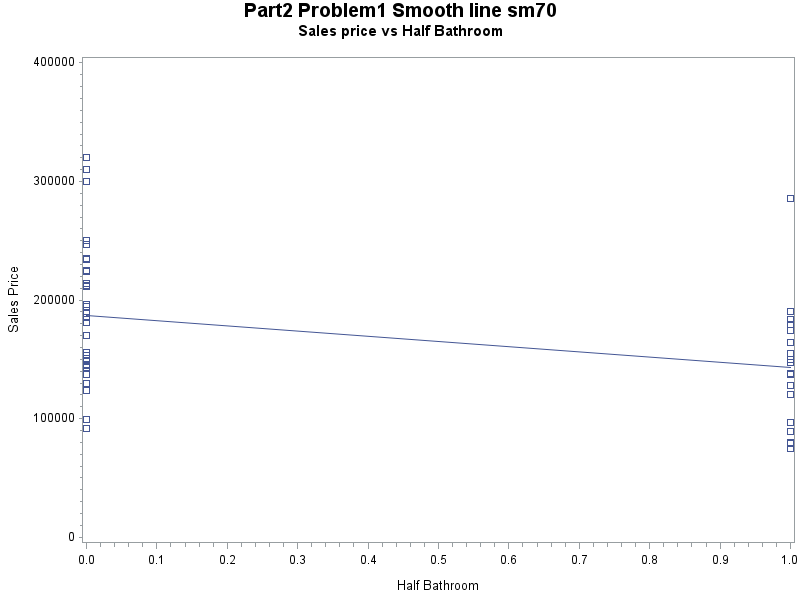
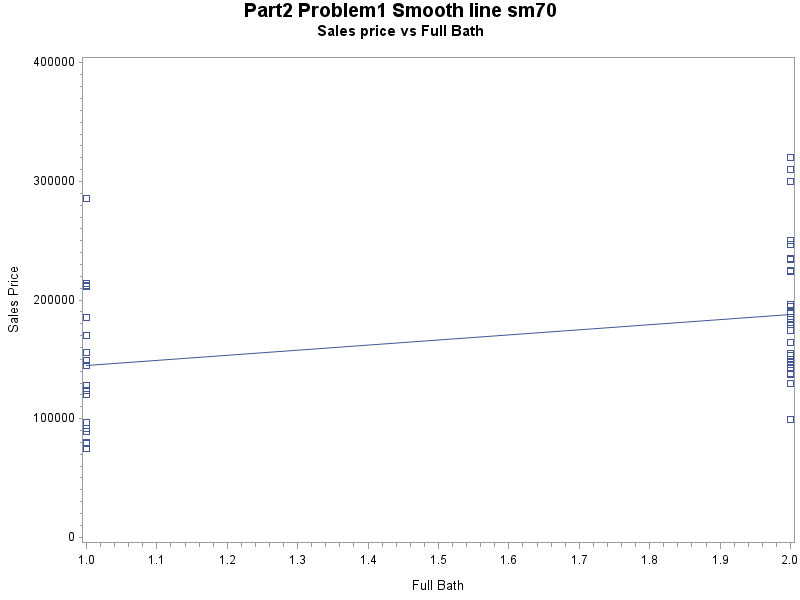
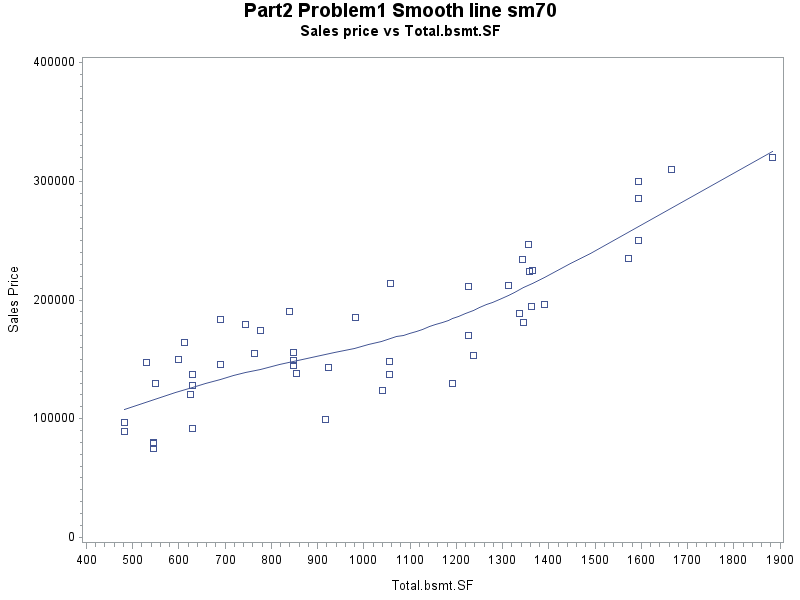
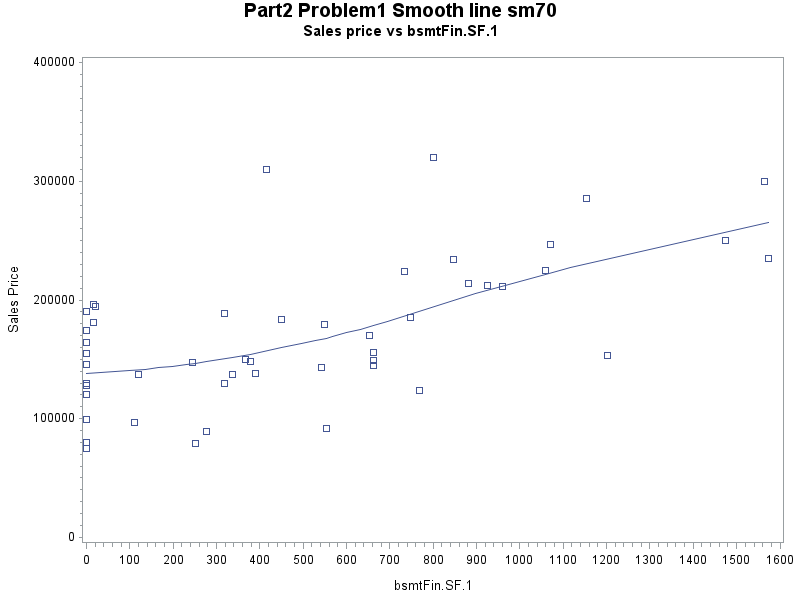
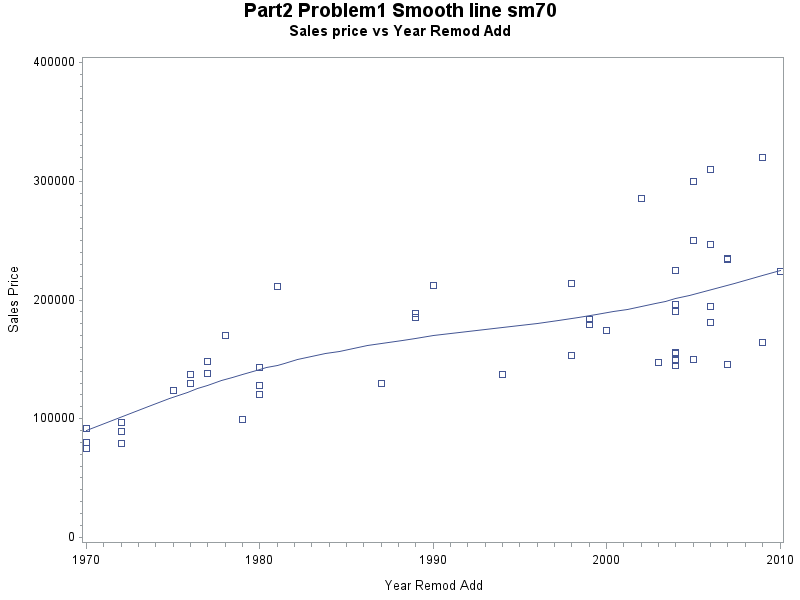
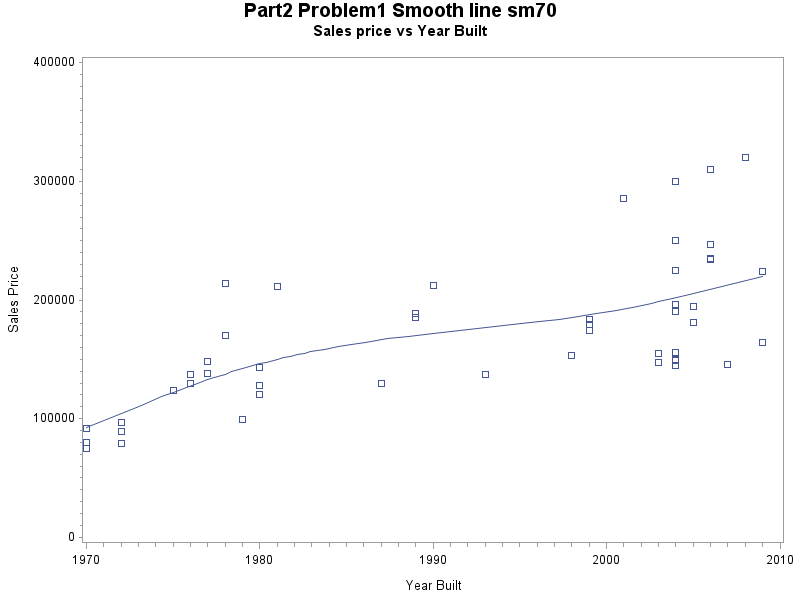
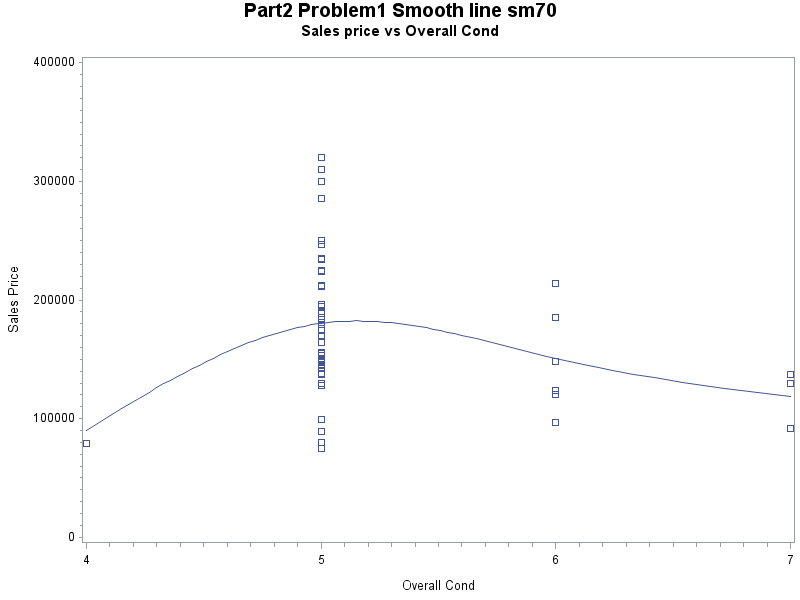
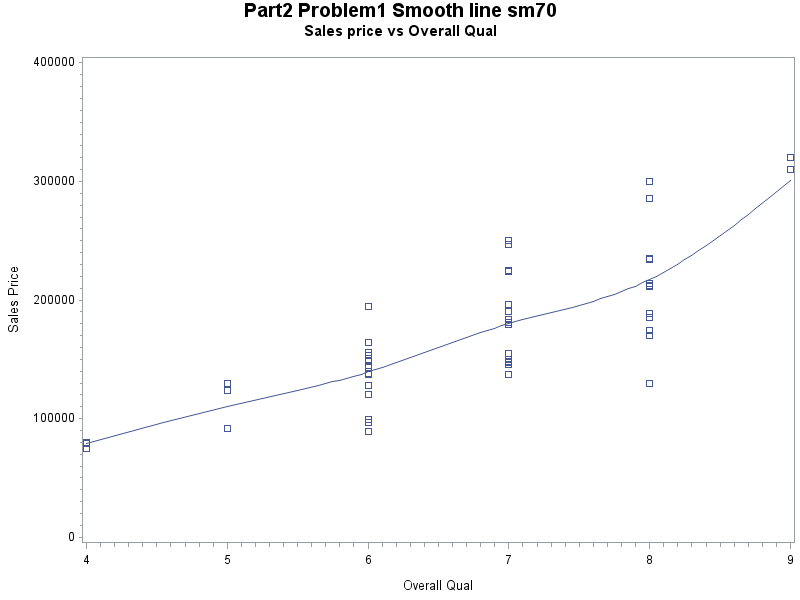
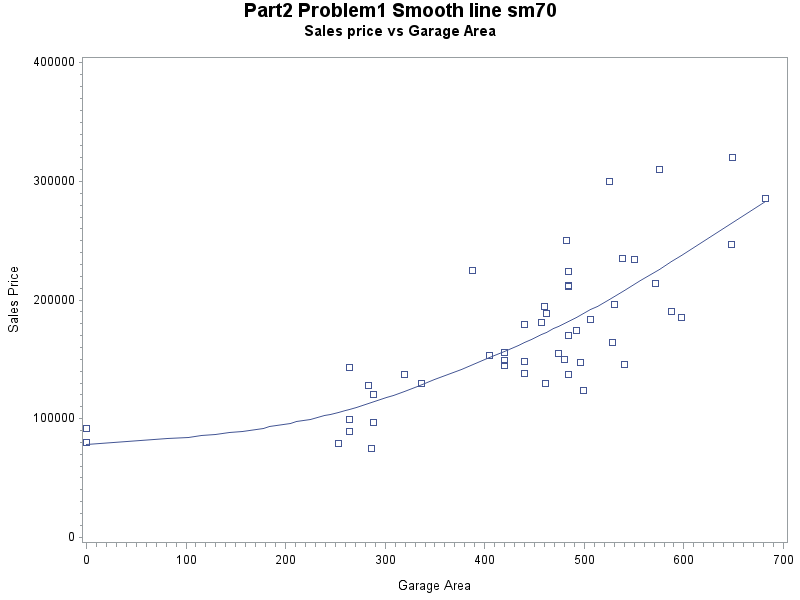
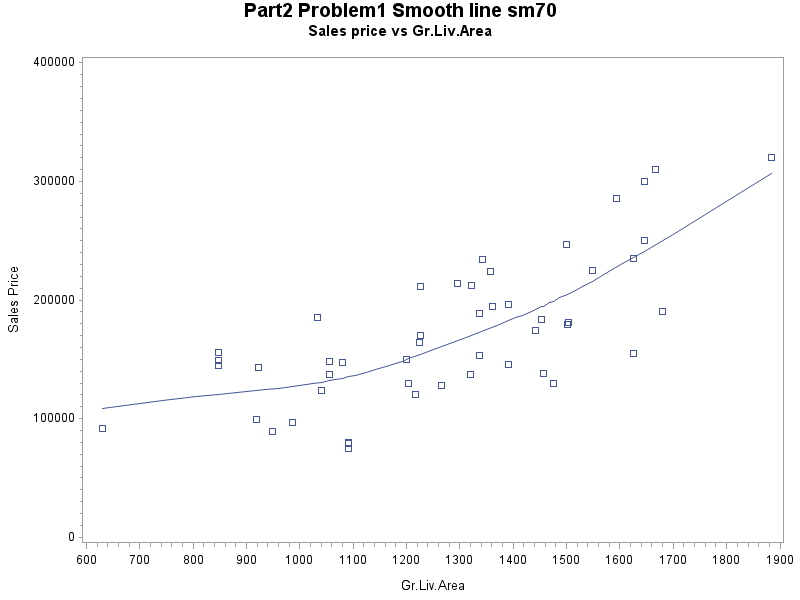
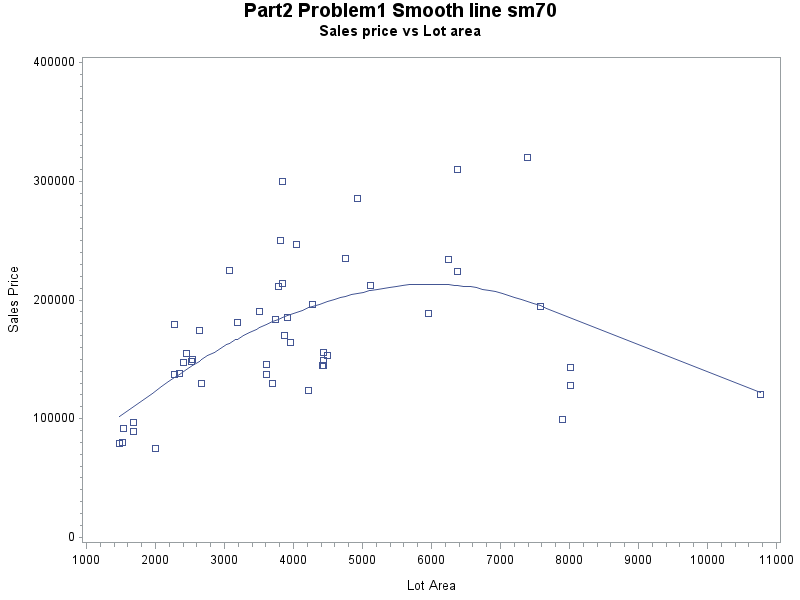
4.



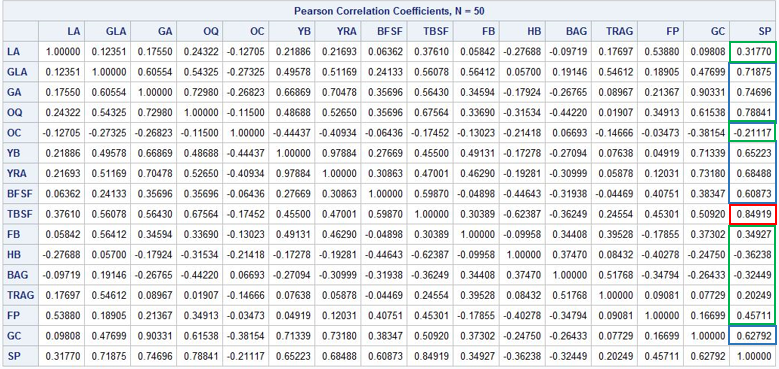
**Part II**

1.

Scatter Plot



Correlation Matrix



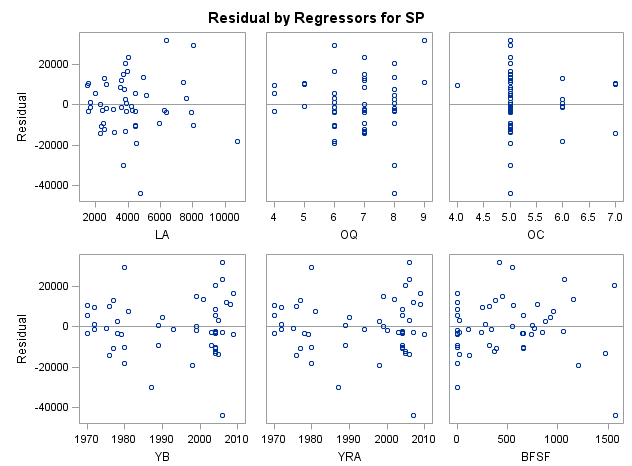
As the scatter plots shown above, most predictors have either strong or weak linear relationships with the response variable except for the full bathroom and the half bathroom. These two predictors could be categorical variables due to their behavior.

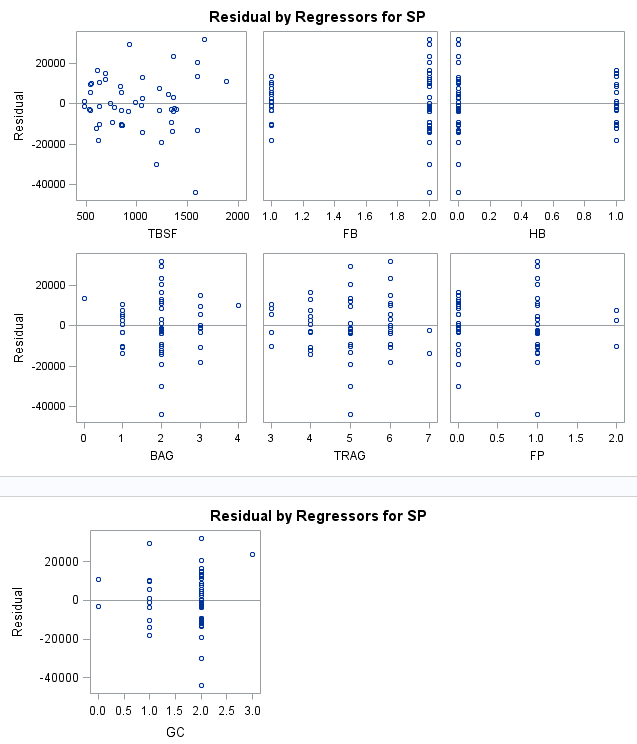
Also, according to the correlation matrix, larger absolute values show stronger linear relationships with the response variable, and some of the predictors including GLA, GA, OQ, YB, YRA, BFSF (labeled with blue), and TBSF which has the highest correlation coefficient with SP (labeled with red) each 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.

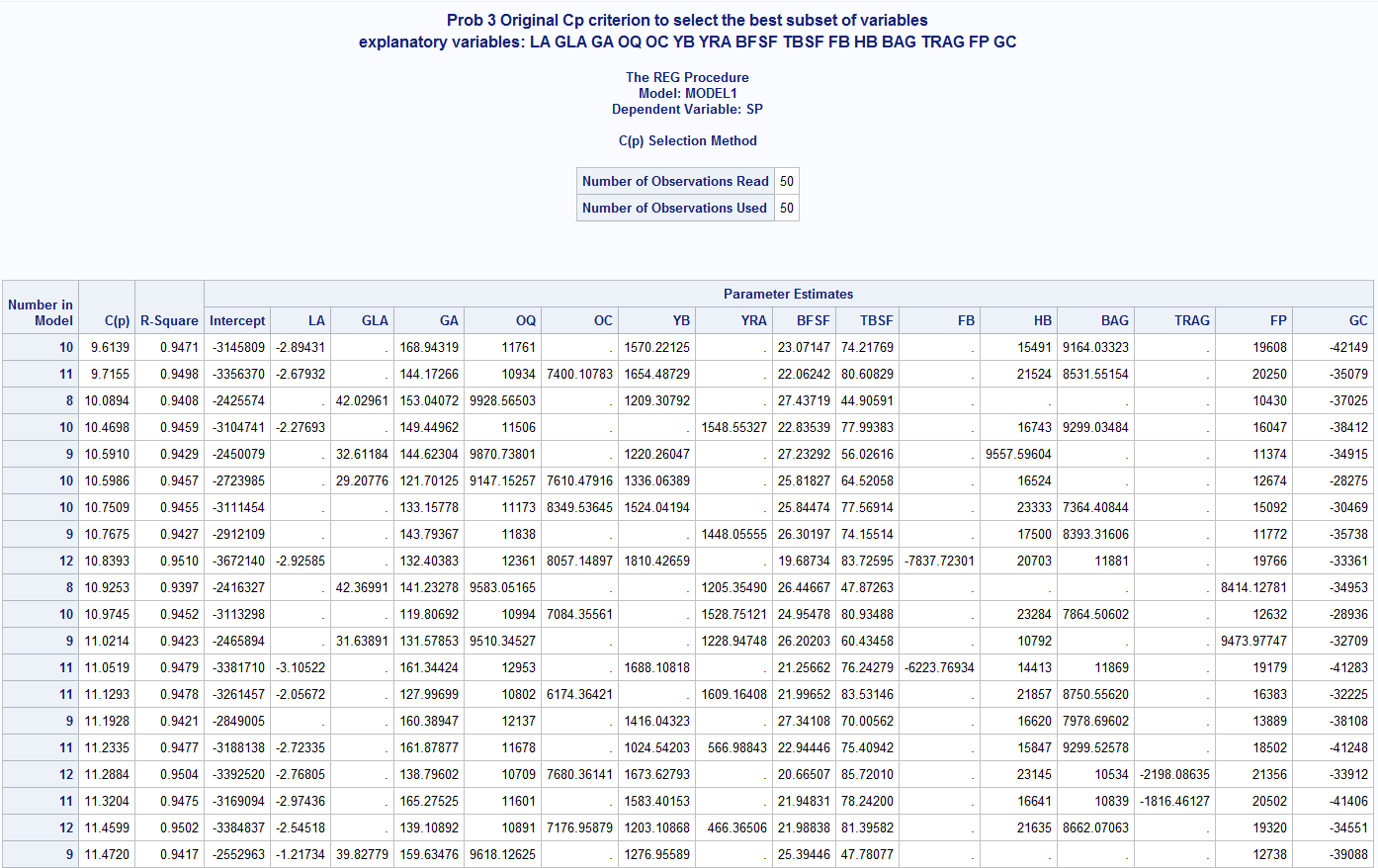
2.

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.

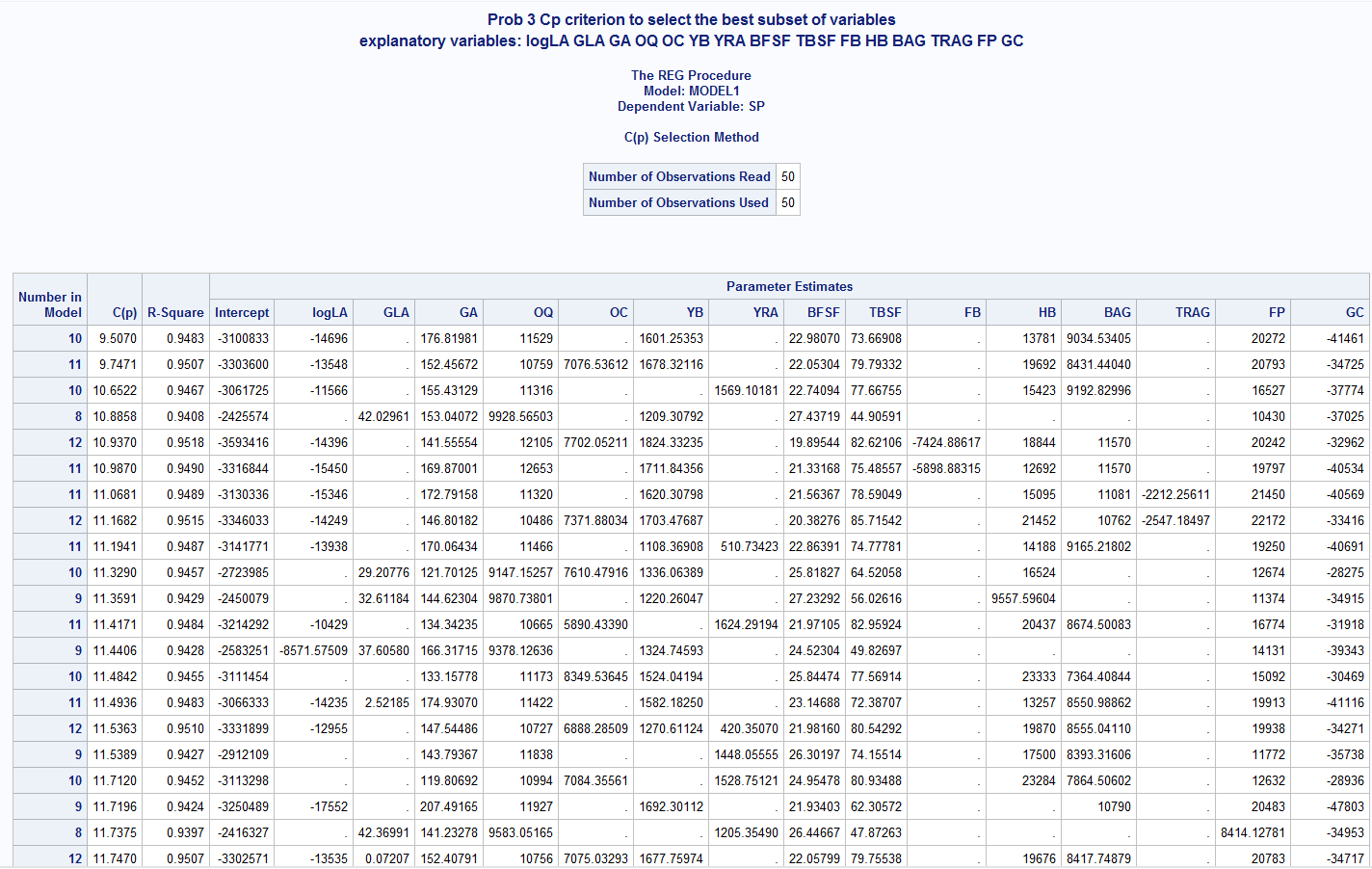
As shown in the graphs below, there is an additional variables need taking into account- BFSF. Recall its scatter plot, BFSF has a relative strong linear relationship with SP, so it is not necessary to transform BFSF in spite of its residual plot. We don’t need to transform other variable since the rest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





3. Original: 

Transformed:



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:

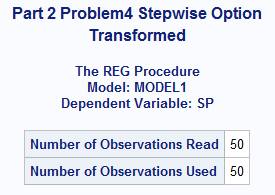
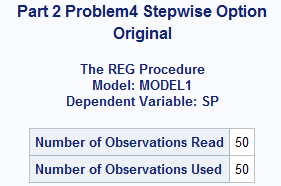
It has 10 predictors with C(p) = 9.5070, R-square = 0.9483

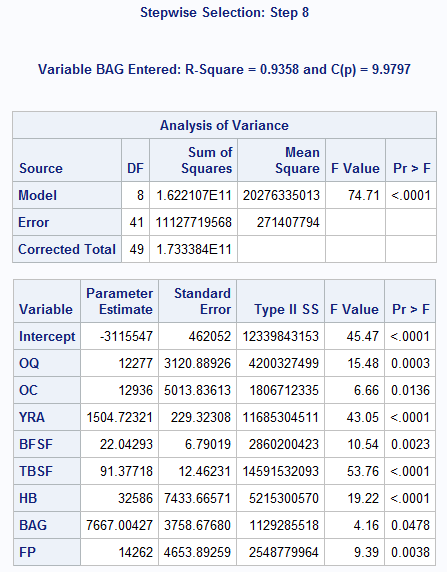
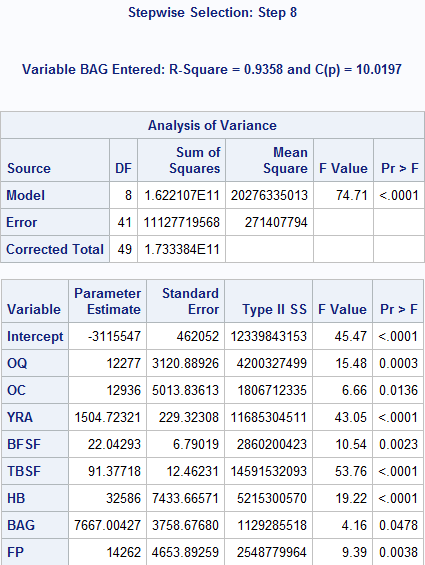
In comparison, the best subset of variables for transformed data has a relative smaller C(p) and higher R-square. Thus, we choose the transformed one as our best model.

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:





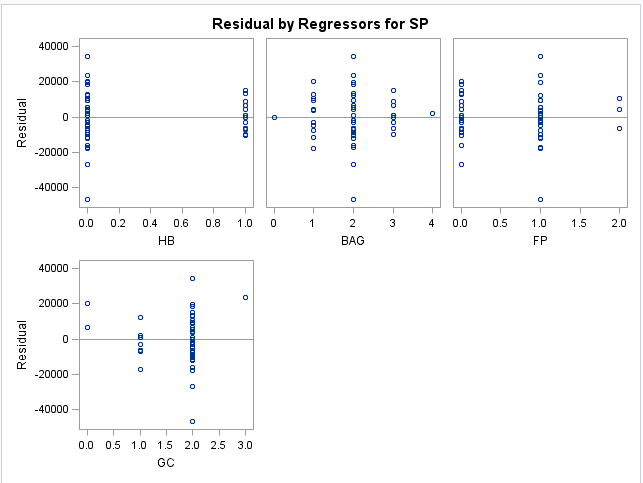
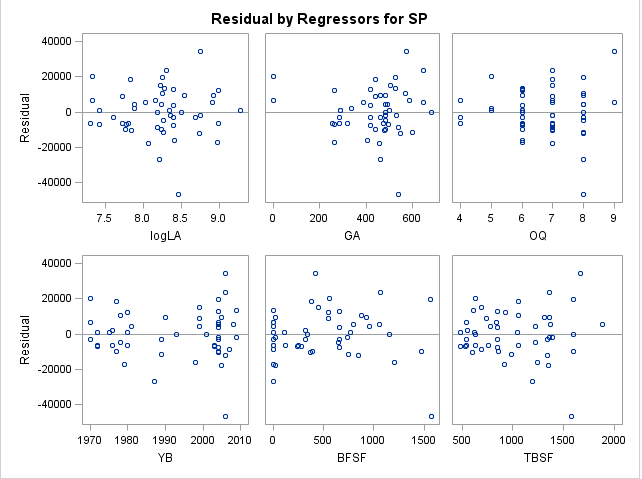
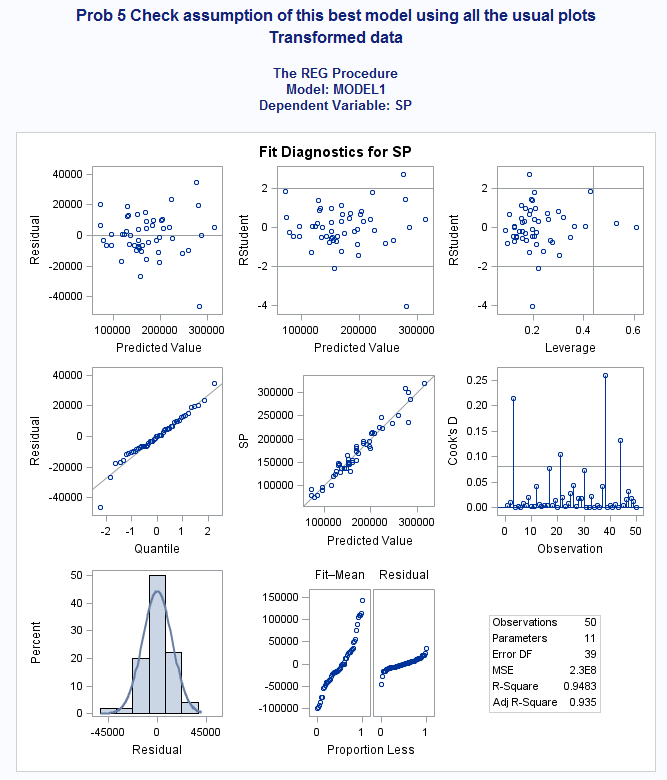
According to SAS output, we can easily find that transformed data has lower Cp value rather than original data. Therefore, we will use logLA as variable in our linear regression model.

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



## 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.

1. 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.

1. Residual Normally Distributed

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

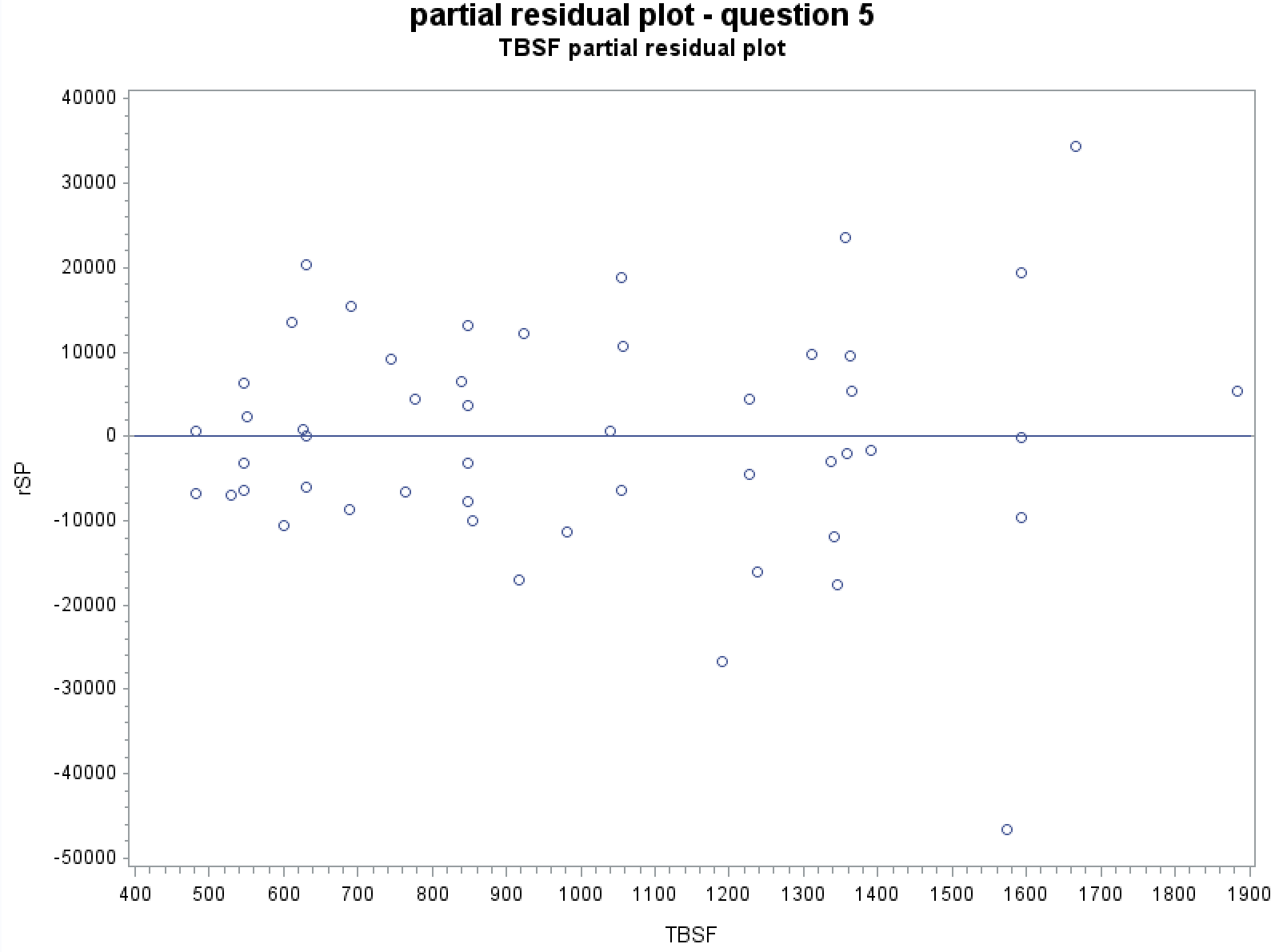
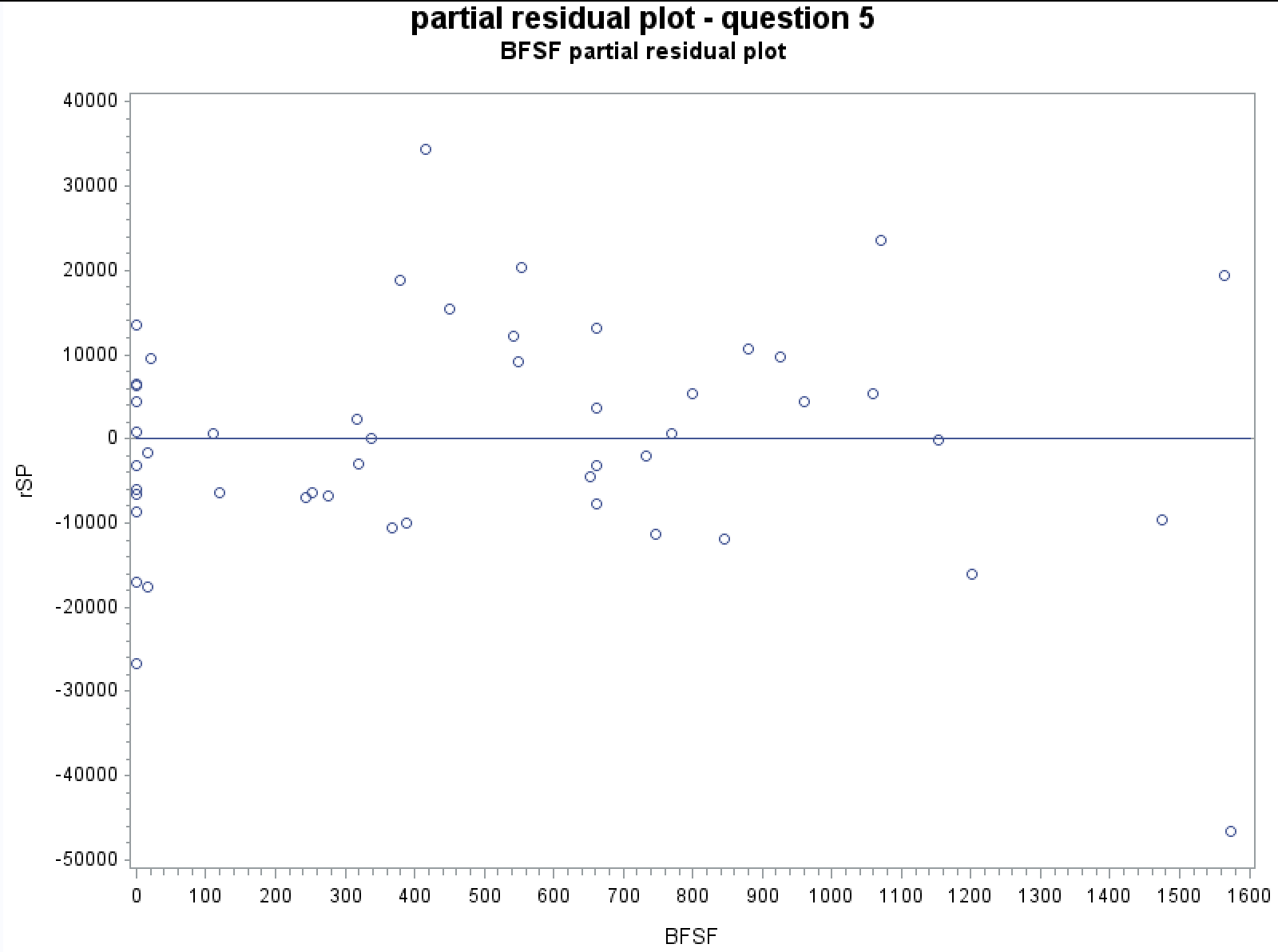
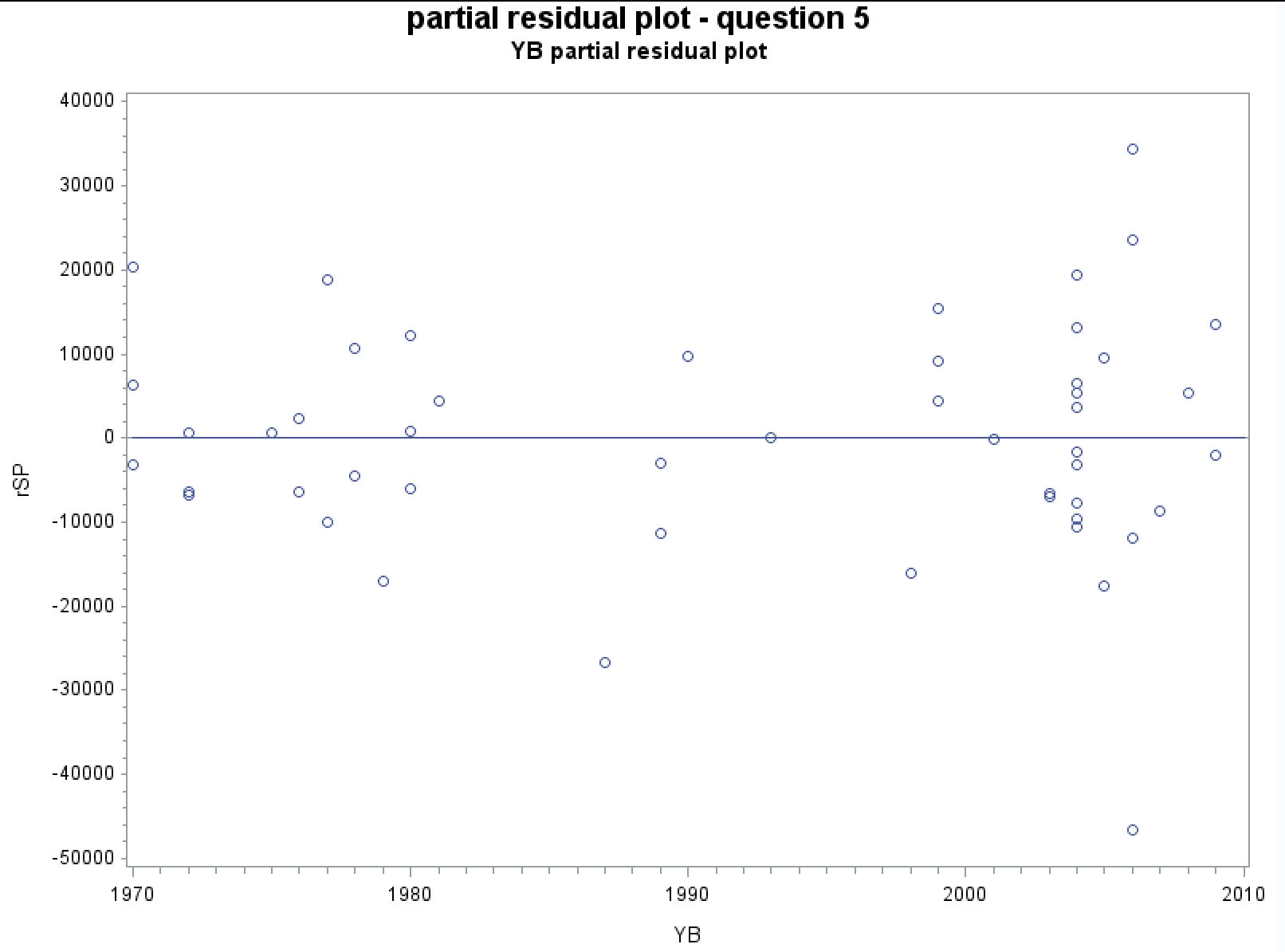
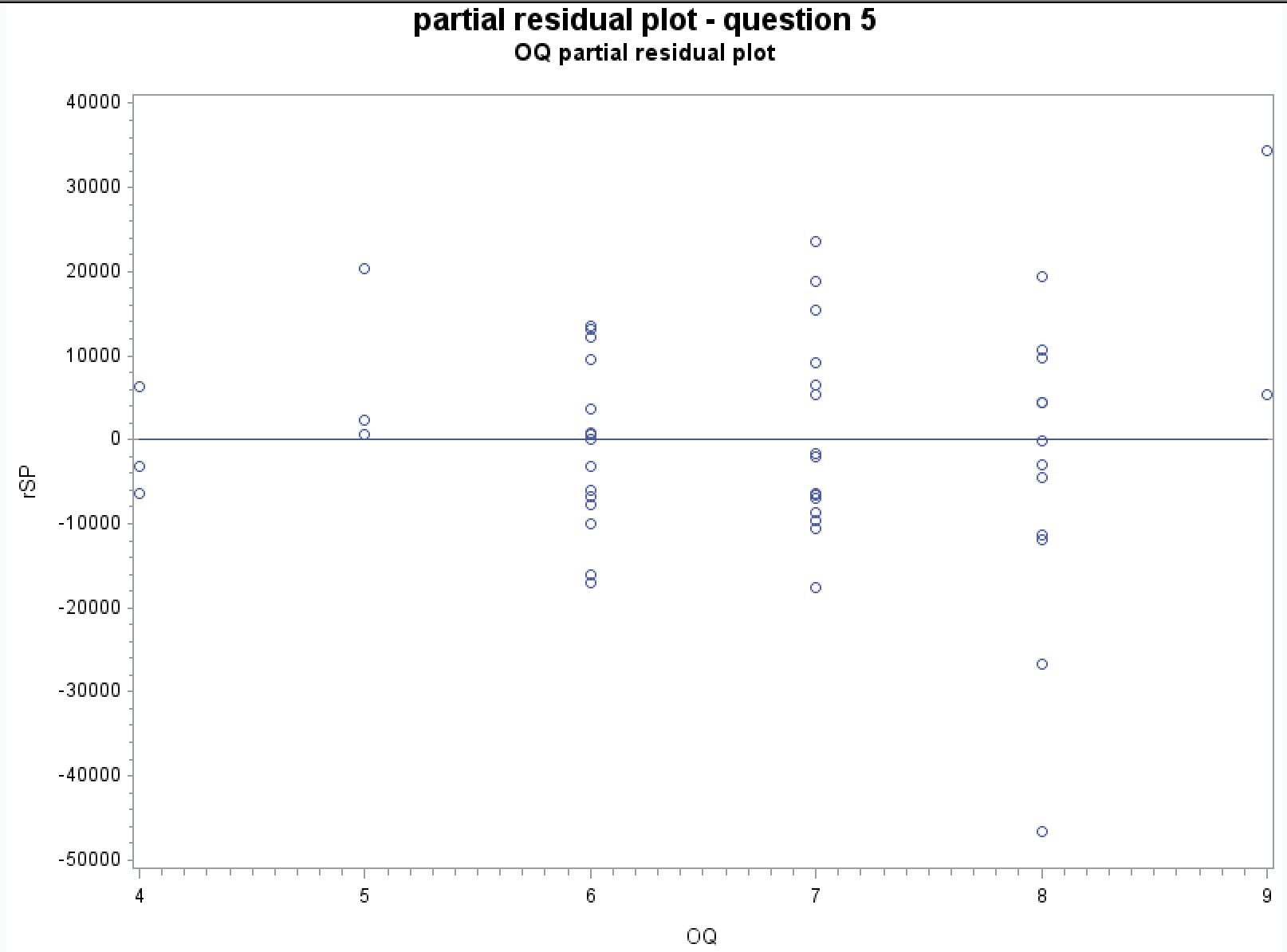
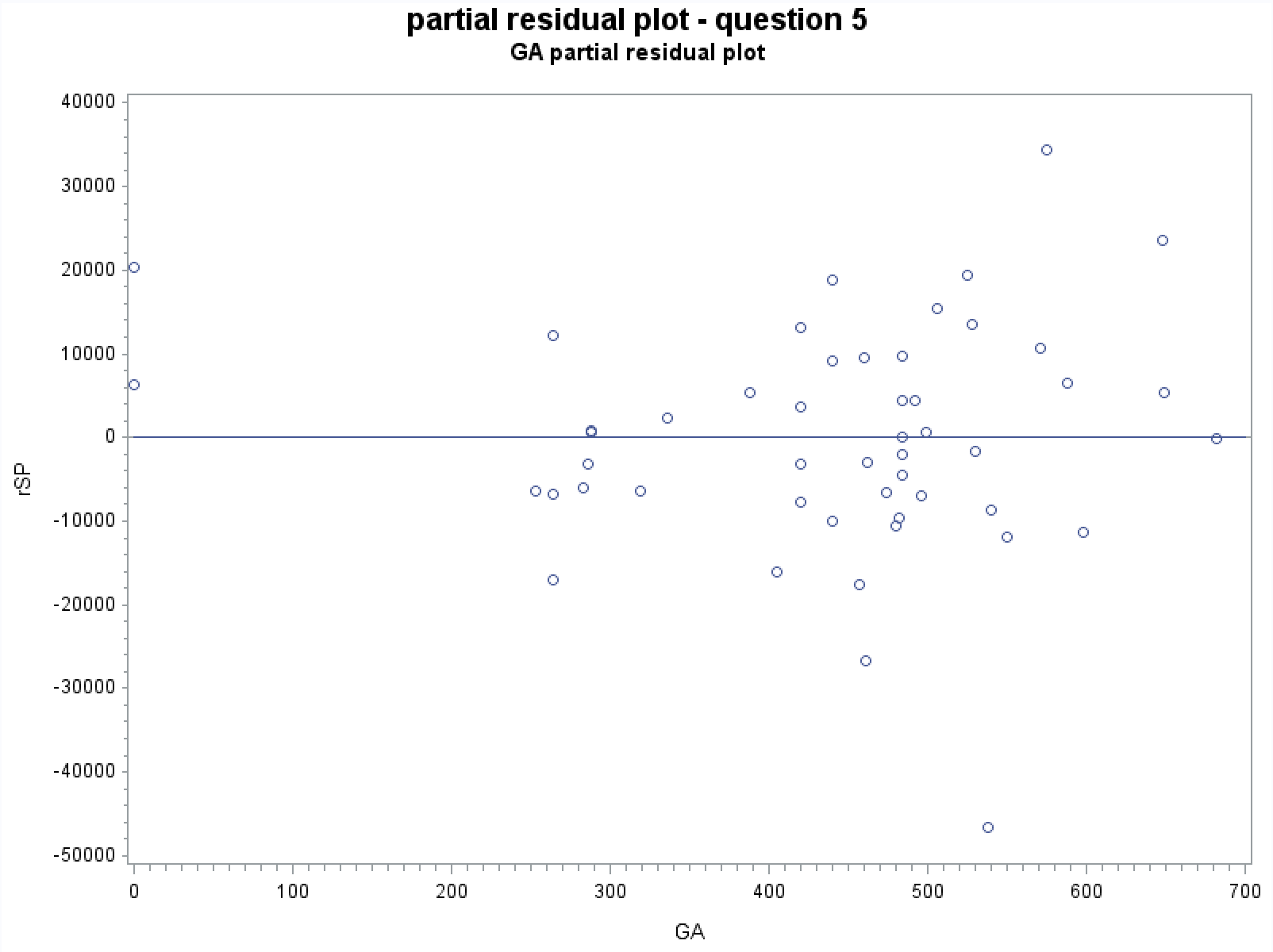
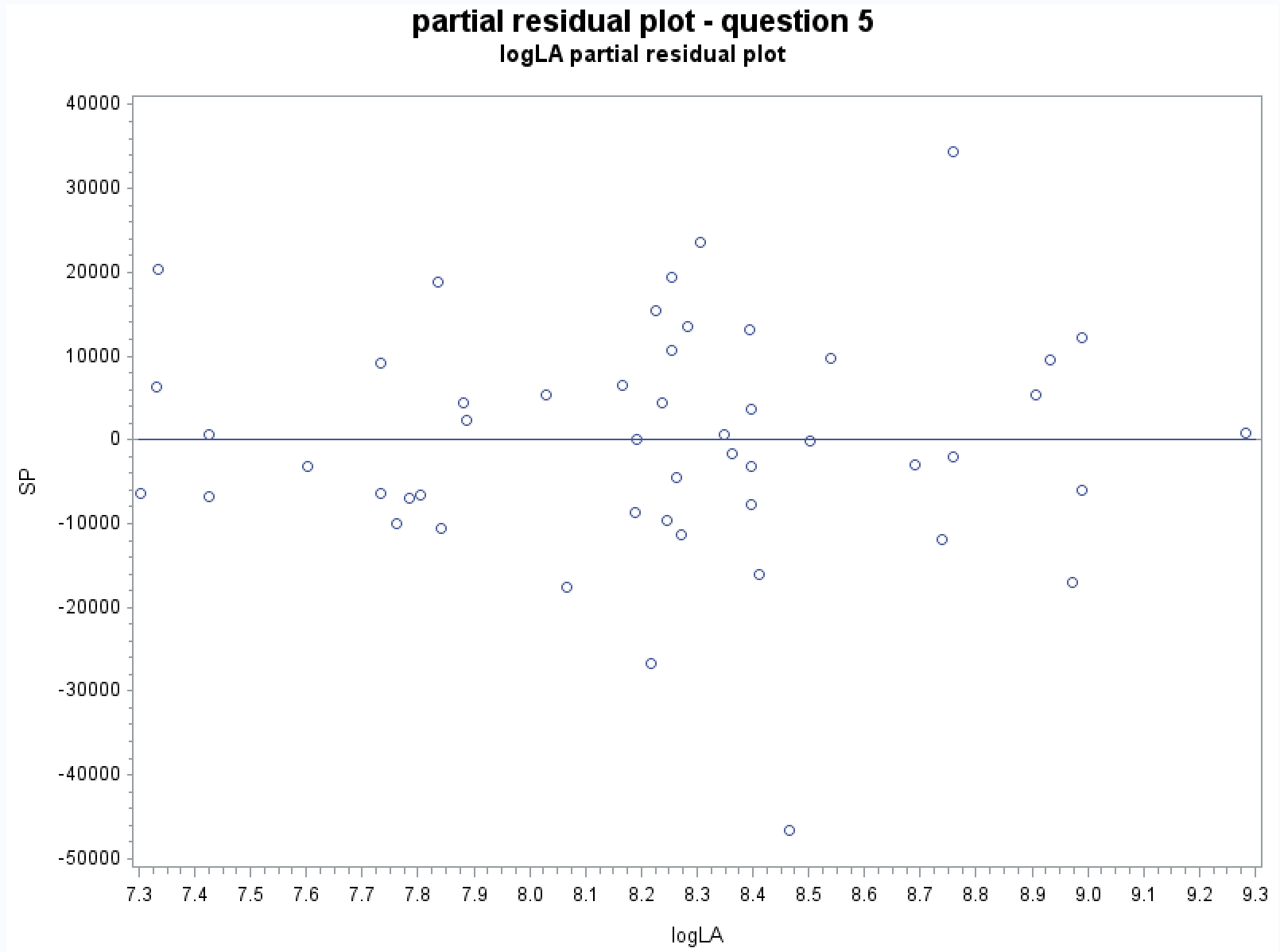
1. Independence

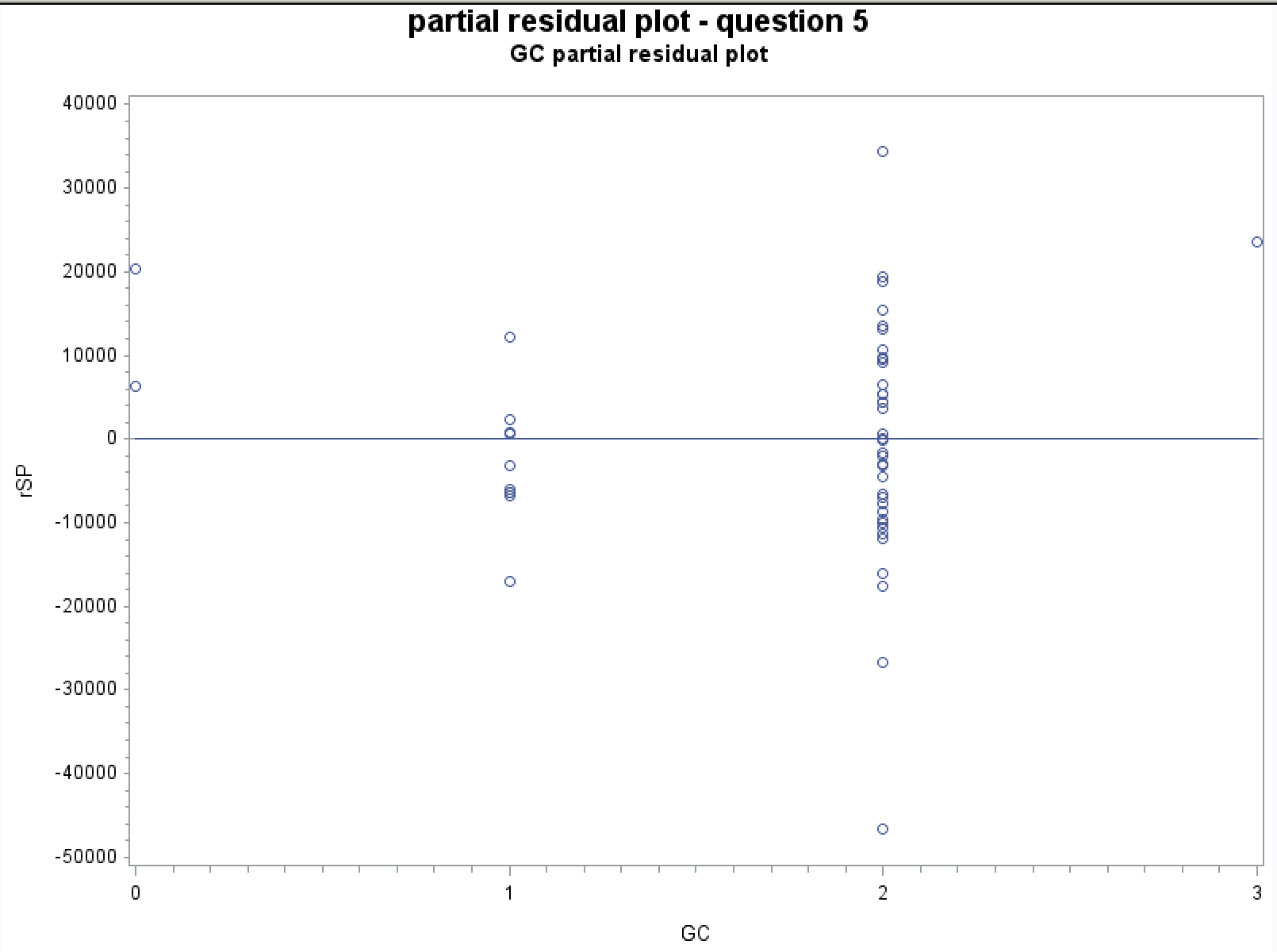
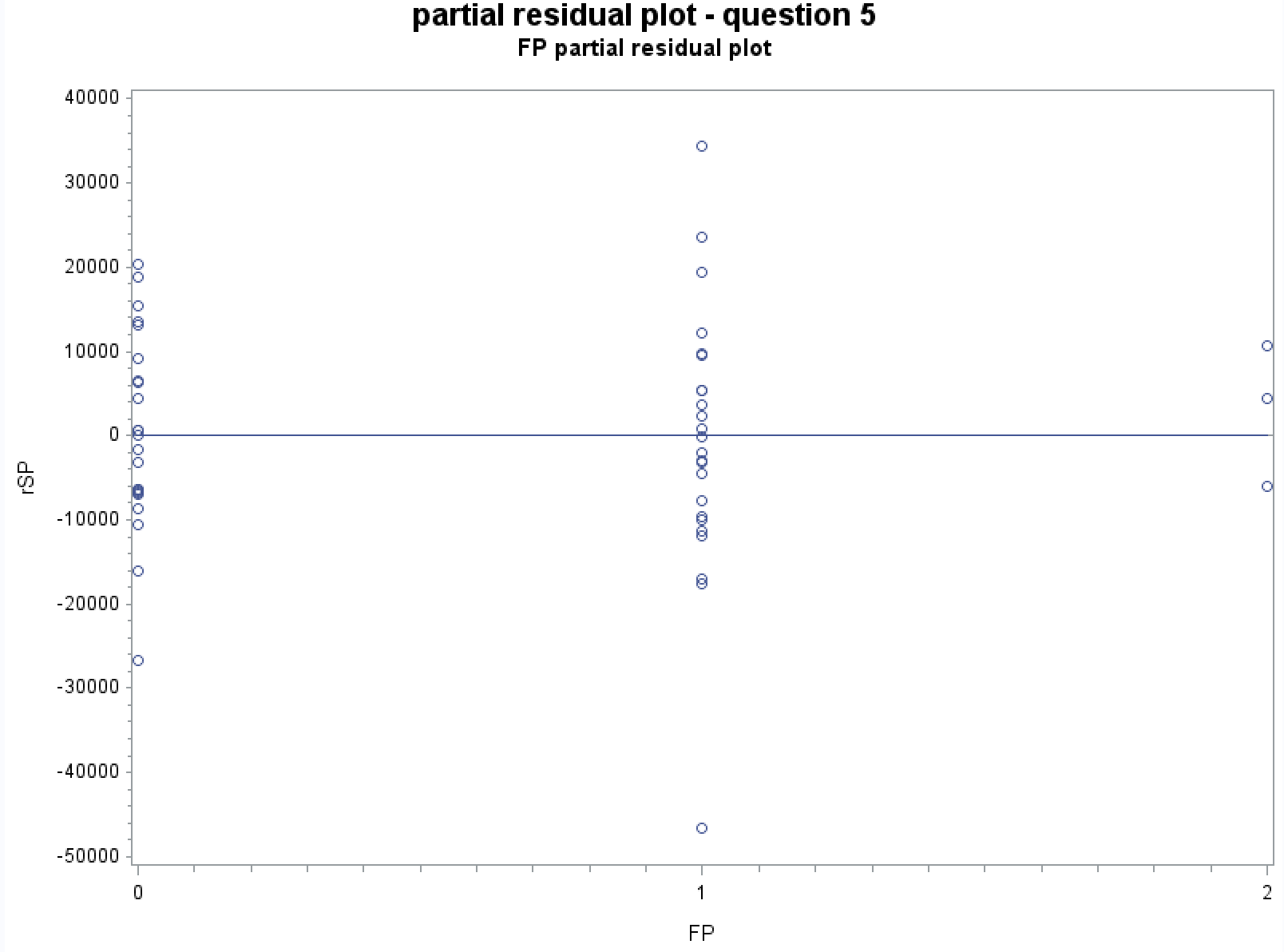
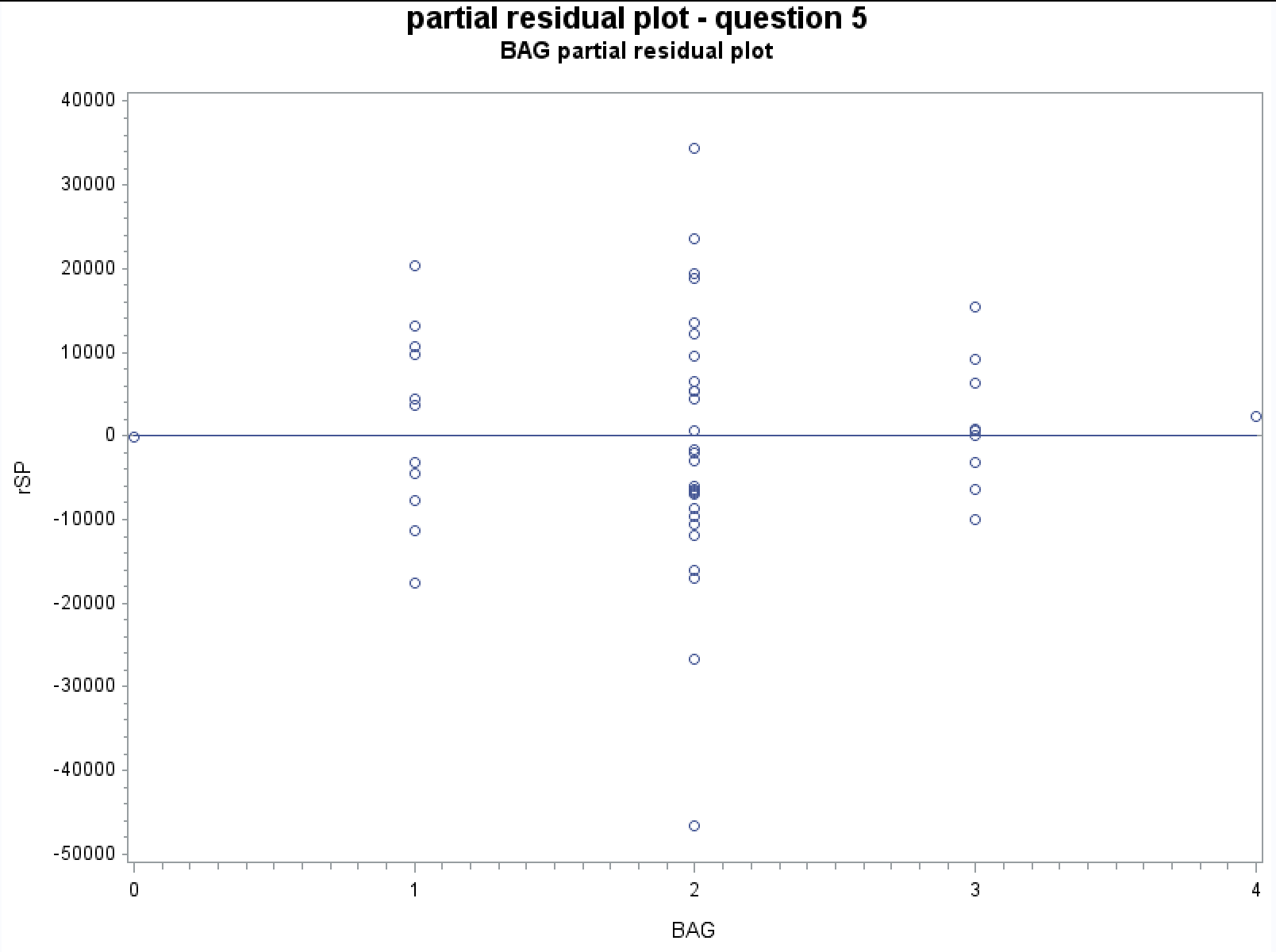
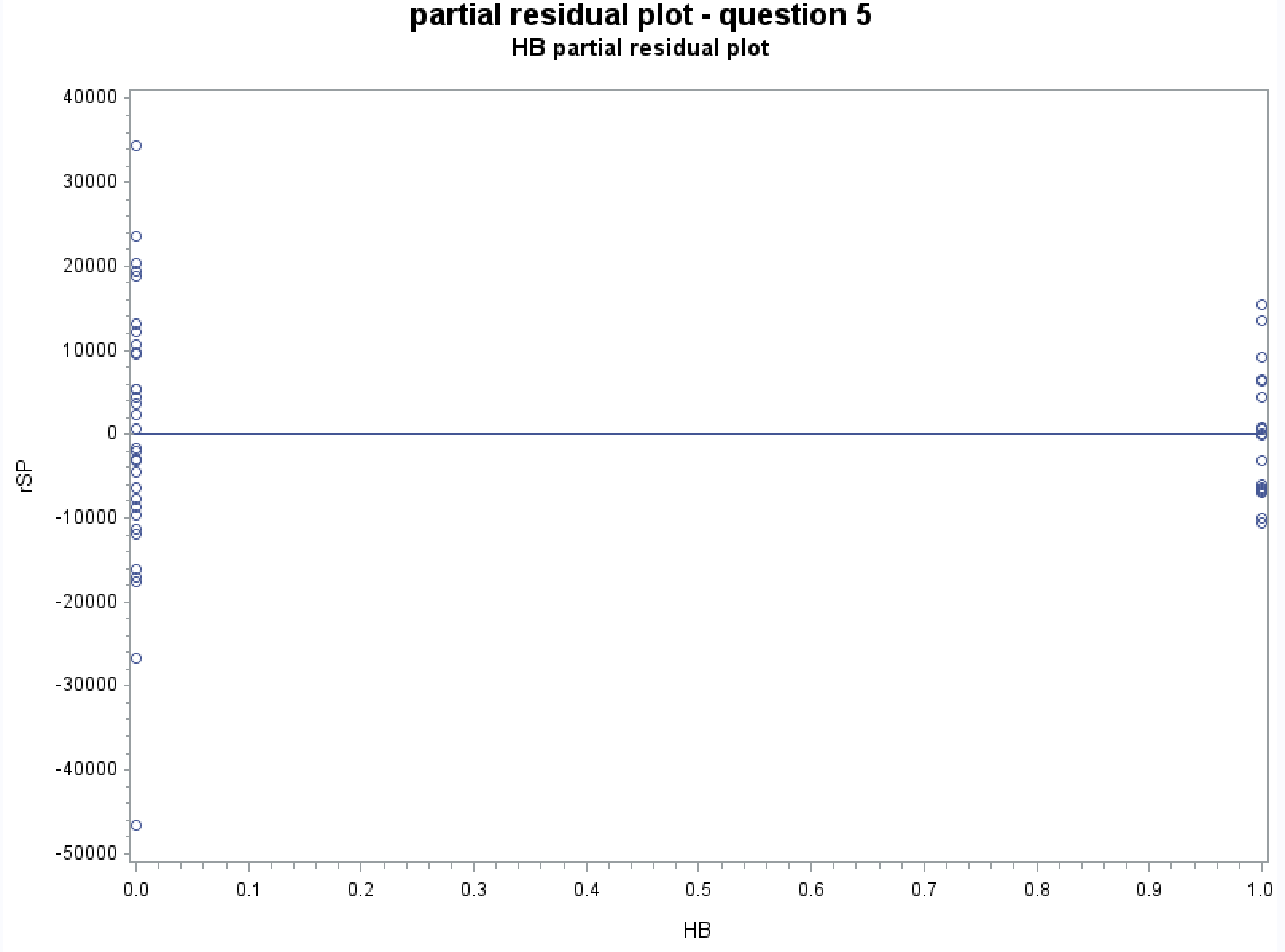
According to the SAS output, the predictors are independent. Therefore, the independence assumption is satisfied.

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

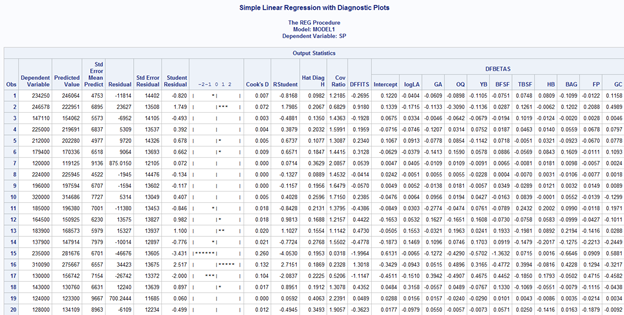
Selling Price = -3100833 + -14696\*logLA + 176.81981\*GA + 11529\*OQ + 1601.25353\*YB + 22.98070\*BFSF + 73.66908\*TBSF + 13781\*HB + 9034.5405\*BAG + 20272\*FP + -41461\*GC

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:



…...

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 |

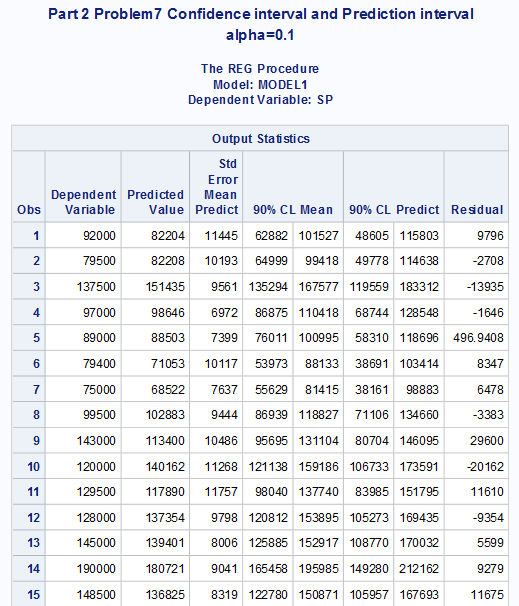
7.

1. 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

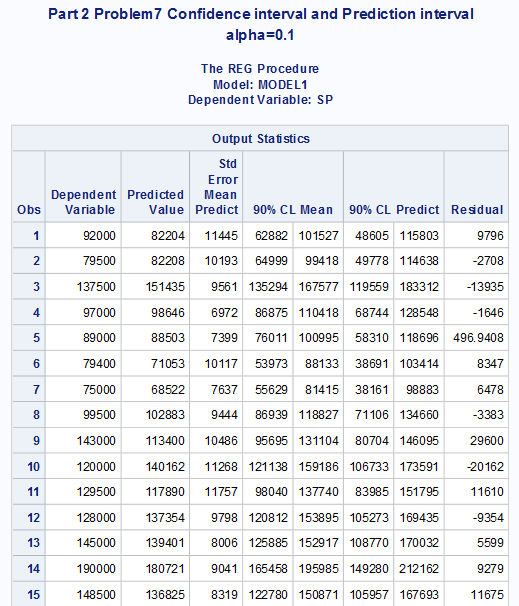
1. 90% confidence interval for the mean of the response variable

The 1st fifteen observations are shown below:

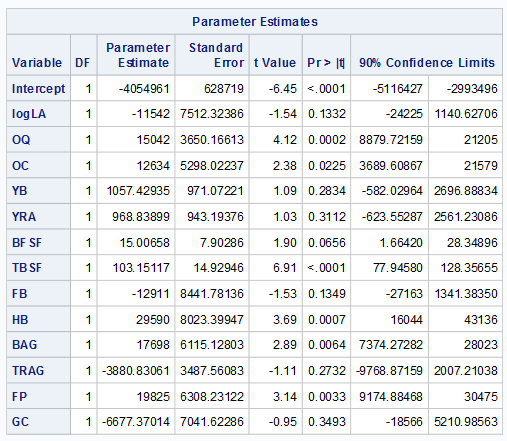


1. 90% prediction interval for individual observations

The 1st fifteen observations are shown below:



1. 90% confidence intervals for the regression coefficients



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