Sample Work with SAS

Heidi Kang

10/05/2018

Problem 1

There are 3 types of building classes and 5 neighborhoods. The cross tabulation is done to check the mean, standard deviation, and counts of the sale prices.

```
proc tabulate data=bkhomessmall;
 class NEIGHBORHOOD BUILDING_CLASS;
 var SALE_PRICE;
 table NEIGHBORHOOD*BUILDING_CLASS,
 SALE_PRICE*(mean std n);
run;
```

In the table below, the cross tabulation shows that the counts are all the same for all of the observations at 30. All of the standard deviations are not similar which may mean that the assumption of homogeneity is violated for the general ANOVA.

		SALE_PRICE		
		Mean	Std	N
NEIGHBORHOOD	BUILDING_CLASS			
BEDFORD STUYVESANT	01 ONE FAMILY DWELLINGS	4729.88	3809.95	30
	02 TWO FAMILY DWELLINGS	4371.03	4035.54	30
	03 THREE FAMILY DWELLIN	2392.92	3502.34	30
CANARSIE	01 ONE FAMILY DWELLINGS	4396.74	1213.85	30
	02 TWO FAMILY DWELLINGS	4965.73	2095.74	30
	03 THREE FAMILY DWELLIN	5127.09	2417.06	30
CROWN HEIGHTS	01 ONE FAMILY DWELLINGS	2349.44	3365.60	30
	02 TWO FAMILY DWELLINGS	2501.15	3501.17	30
	03 THREE FAMILY DWELLIN	2956.20	3866.21	30
EAST NEW YORK	01 ONE FAMILY DWELLINGS	3445.50	1416.72	30
	02 TWO FAMILY DWELLINGS	4489.09	2664.80	30
	03 THREE FAMILY DWELLIN	4852.37	2506.07	30

		SALE_PRICE		
		Mean	Std	N
FLATBUSH-EAST	01 ONE FAMILY DWELLINGS	4860.05	2096.51	30
	02 TWO FAMILY DWELLINGS	4501.92	2734.54	30
	03 THREE FAMILY DWELLIN	5453.42	2461.53	30

Problem 2

The model shows a high amount of significance most likely because the variable neighborhood has a high amount of significance in the anova model. Building class is not significant.

```
proc anova data=bkhomessmall;
 class NEIGHBORHOOD BUILDING_CLASS;
 model SALE_PRICE=NEIGHBORHOOD BUILDING_CLASS;
run;
```

In the anova model below, the p-value is less than 0.0001 overall which means that it is significant. The p-value for neighborhood is also less than 0.0001. However, building class does not have a significant p-value on a 95% significance level. Building class has a p-value of 0.7842. The means of neighborhood and building class are very different.

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	6	326150214	54358369	6.30	<.0001
Error	443	3823753134	8631497		
Corrected Total	449	4149903348			

R-Square	Coeff Var	Root MSE	SALE_PRICE Mean
0.078592	71.78252	2937.941	4092.836

Source	DF	Anova SS	Mean Square	F Value	Pr > F
NEIGHBORHOOD	4	321950528.8	80487632.2	9.32	<.0001
BUILDING_CLASS	2	4199684.9	2099842.5	0.24	0.7842

Problem 3

Based on the anova model and the linear model, it looks like the model is significant for both overall. Both residential units and safety ranking should be kept since they are both significant.

```
proc anova data=bkhomes;
 class residentialunits safestfortotcrimeranks;
 model saleprice=residentialunits safestfortotcrimeranks;
run;
```

In the anova model, the p-value is less than 0.0001 which means that there is a very high significance. There is a high amount of variation according to this model.

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	38	1458077194	38370452	5.71	<.0001
Error	2461	16541601164	6721496		
Corrected Total	2499	17999678358			

R-Square	Coeff Var	Root MSE	saleprice Mean
0.081006	61.53942	2592.585	4212.885

Source	DF	Anova SS	Mean Square	F Value	Pr > F
residentialunits	33	389935419	11816225	1.76	0.0049
safestfortotcrimeran	5	1068141775	213628355	31.78	<.0001

```
proc glm data=bkhomes;
 class residentialunits safestfortotcrimeranks;
 model saleprice=residentialunits safestfortotcrimeranks/ ss1
```

```
 ss3;
 lsmeans residentialunits safestfortotcrimeranks/ adjust=tukey;
 ods exclude intplot;
run;
```

In the glm model, there is a p value that is less than 0.0001 which means that the model is significant at a 95% level. For Type SSI, both residential units and safety ranks have a lot of significance. However, for type SSIII, the significance for safety ranks is low since the p value is 0.0803. There is a good amount of variance in this model.

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	38	1372653442	36122459	5.35	<.0001
Error	2461	16627024916	6756207		
Corrected Total	2499	17999678358			

R-Square	Coeff Var	Root MSE	saleprice Mean
0.076260	61.69812	2599.270	4212.885

Source	DF	Type I SS	Mean Square	F Value	Pr > F
residentialunits	33	389935419.2	11816224.8	1.75	0.0053
safestfortotcrimeran	5	982718022.5	196543604.5	29.09	<.0001

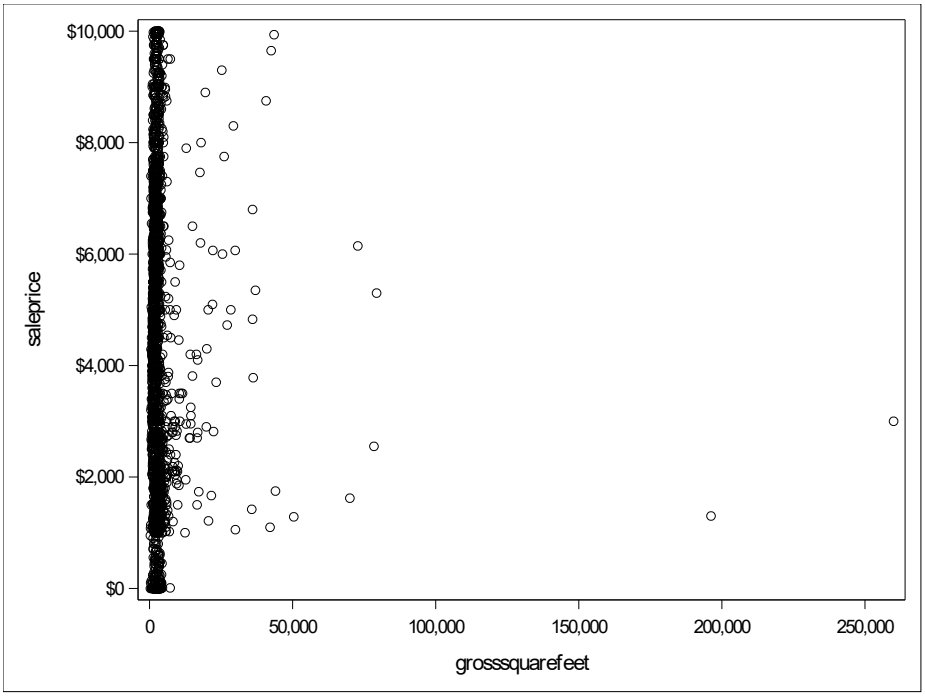
Source	DF	Type III SS	Mean Square	F Value	Pr > F
residentialunits	33	304511666.8	9227626.3	1.37	0.0803
safestfortotcrimeran	5	982718022.5	196543604.5	29.09	<.0001

Problem 4

There is not really much association between sale price and gross square footage. The sale price appears to not be affected by gross square footage much. There is not much variation in this model.

```
proc sgplot data=bkhomes;
 scatter y=saleprice x=grosssquarefeet;
run;
```

In the scatter plot below, sale price appears to have varying amounts no matter what the gross square footage is. There are more points plotted near the lower end of the amount of gross square feet.



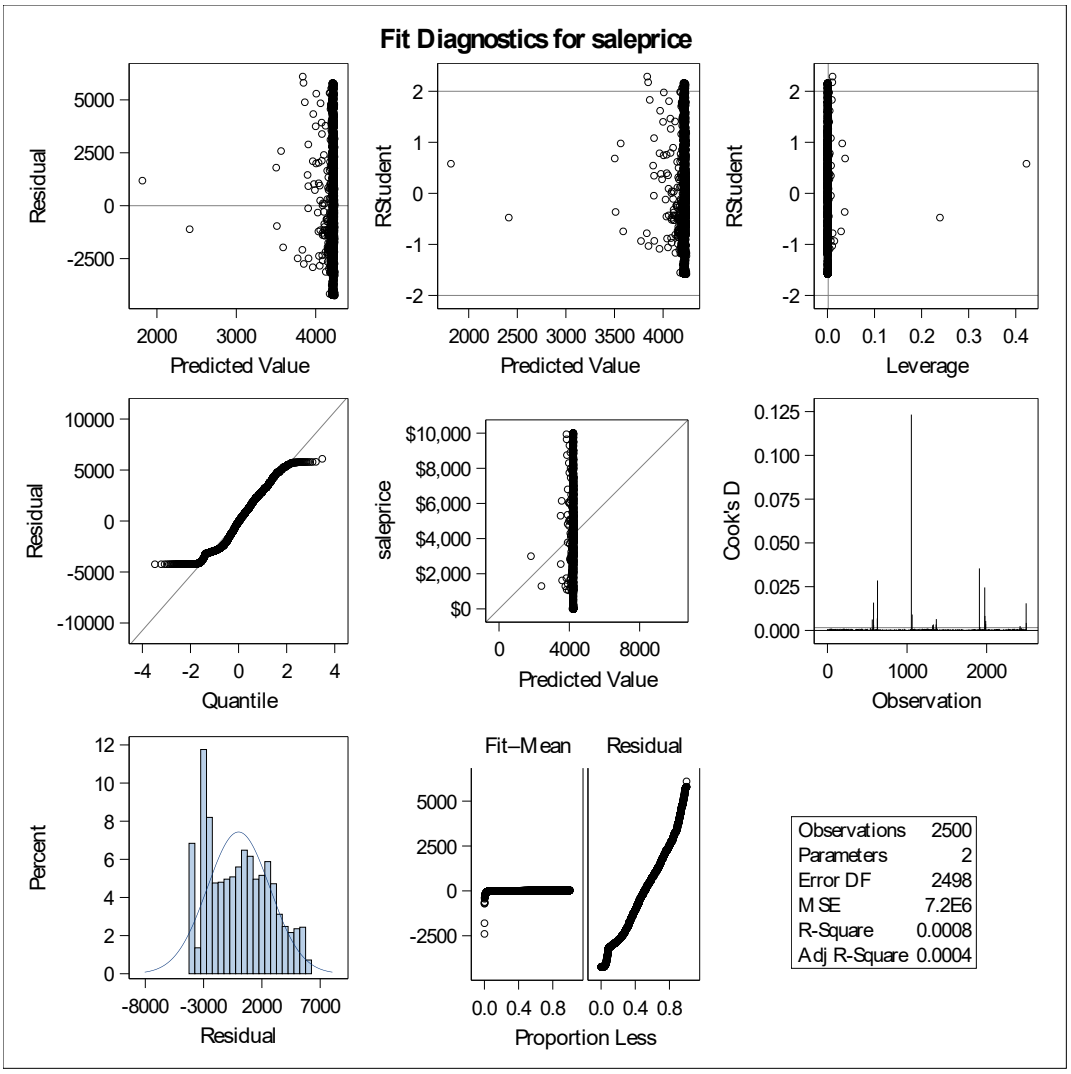

```
proc reg data=bkhomes;
 model saleprice=grosssquarefeet;
 output out=cd1 cookd=cooks1;
run;
```

There is not a lot of variation in the model. The model with gross square feet has a very low variation since the p value for the model is very high at 0.1696. The diagnostics plots also show that there is very low variation. Most of the plots show a straight up and down vertical trend. The histogram shows low variance. There were no points with a cook's distance larger than 1.

Analysis of Variance					
Source	**DF**	**Sum of Squares**	**Mean Square**	**F Value**	**Pr > F**
Model	1	13592531	13592531	1.89	0.1696
Error	2498	1798608582	7200194		
Corrected Total	2499	1799967835			

Root MSE	2683.31781	**R-Square**	0.0008
Dependent Mean	4212.88468	**Adj R-Sq**	0.0004
Coeff Var	63.69312		

Parameter Estimates					
Variable	**DF**	**Parameter Estimate**	**Standard Error**	**t Value**	**Pr > |t|**
Intercept	1	4242.82678	57.92223	73.25	<.0001
grosssquarefeet	1	-0.00933	0.00679	-1.37	0.1696




```
proc print data=cd1;
 where cooks1 > 1;
run;
```

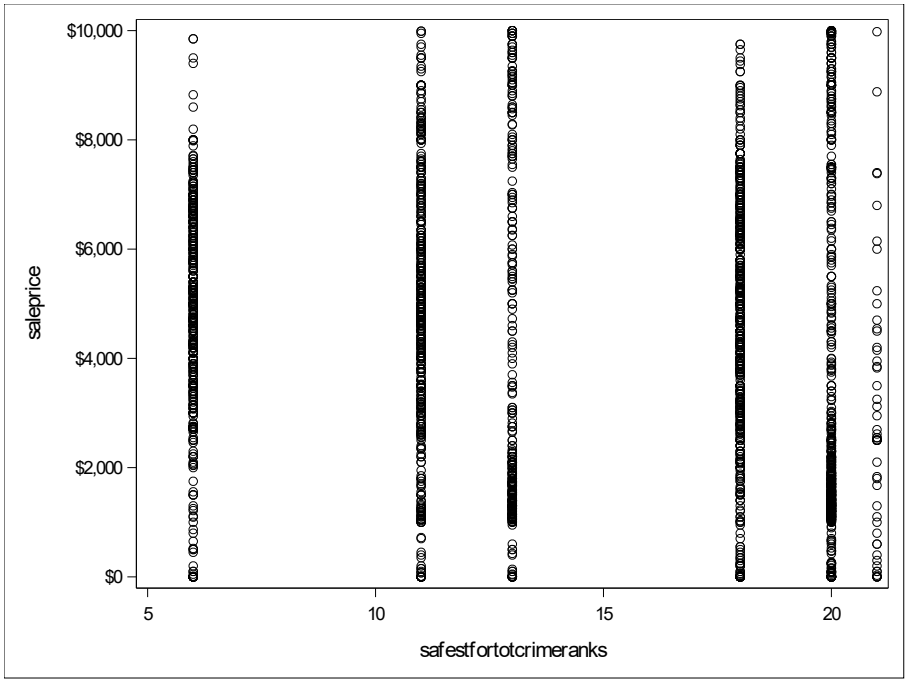
```
NOTE: No observations were selected from data set WORK.CD1.
 NOTE: There were 0 observations read from the data set WORK.CD1.
       WHERE cooks1>1;
```

Problem 5

There is no association between sale price and the safety rank. There is no linear trend. However, there appears to be some variation.

```
proc sgplot data=bkhomes;
 scatter y=saleprice x=safestfortotcrimeranks;
run;
```

The scatter plot below shows that there may be some variation between sale price and safety rank but no linear trend that shows any association. The sale prices appear to have around the similar frequency for many different safety ranks. Since it does not appear to really have a pattern, it is very unlikely that there is association.



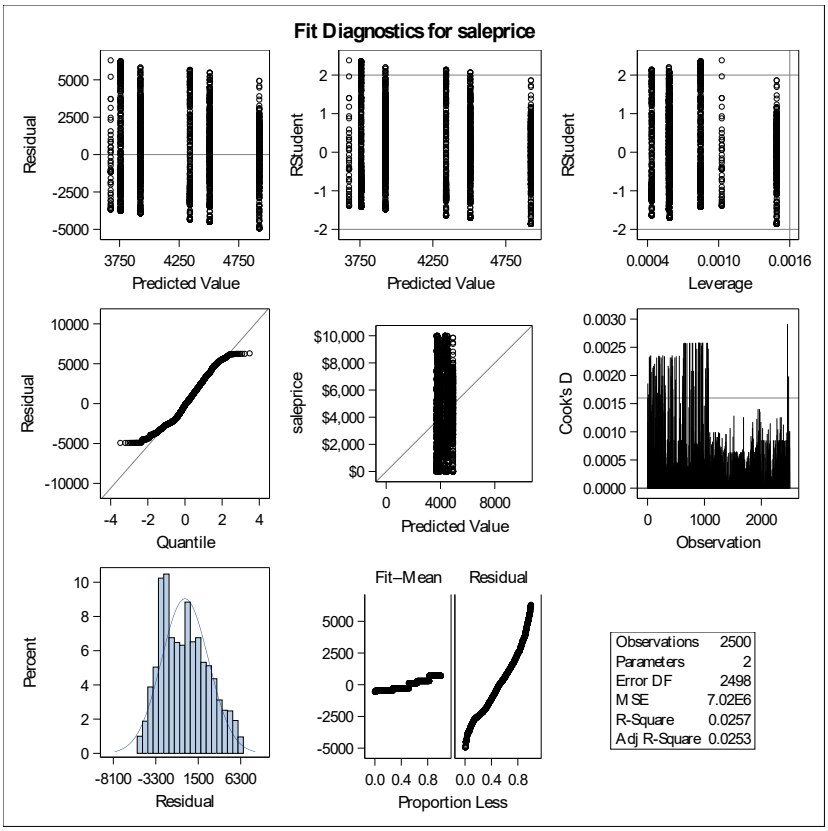
```
proc reg data=bkhomes;
 model saleprice=safestfortotcrimeranks;
 output out=cd2 cookd=cooks2;
run;
```

The model has some variation and shows that there may be association since the p value is less than 0.0001. The diagnostics plots show that there is a lot of variation and the histogram is approximately normal. There are no points in the model that have a cook's distance greater than 1.

Analysis of Variance					
Source	**DF**	**Sum of Squares**	**Mean Square**	**F Value**	**Pr > F**
Model	1	461764393	461764393	65.77	<.0001
Error	2498	1753791396	7020782		
Corrected Total	2499	1799967835			

Root MSE	2649.67587	**R-Square**	0.0257
Dependent Mean	4212.88468	**Adj R-Sq**	0.0253
Coeff Var	62.89457		

Parameter Estimates					
Variable	**DF**	**Parameter Estimate**	**Standard Error**	**t Value**	**Pr > |t|**
Intercept	1	5421.06830	158.12045	34.28	<.0001
safestfortotcrimeranks	1	-83.13953	10.25156	-8.11	<.0001




```
proc print data=cd2;
 where cooks2 > 1;
run;
```

```
NOTE: No observations were selected from data set WORK.CD2.

 NOTE: There were 0 observations read from the data set WORK.CD2.

       WHERE cooks2>1;
```

Problem 6

The model for 5 seems much better since there was much more variation and the model had a higher significance level. There was also much better diagnostics plots so sale price and safety ranking makes a better model.

Problem 7

The model appears to be at a high significance level. There is a high amount of variation but the model could probably be without total units, land square feet, or gross square feet.

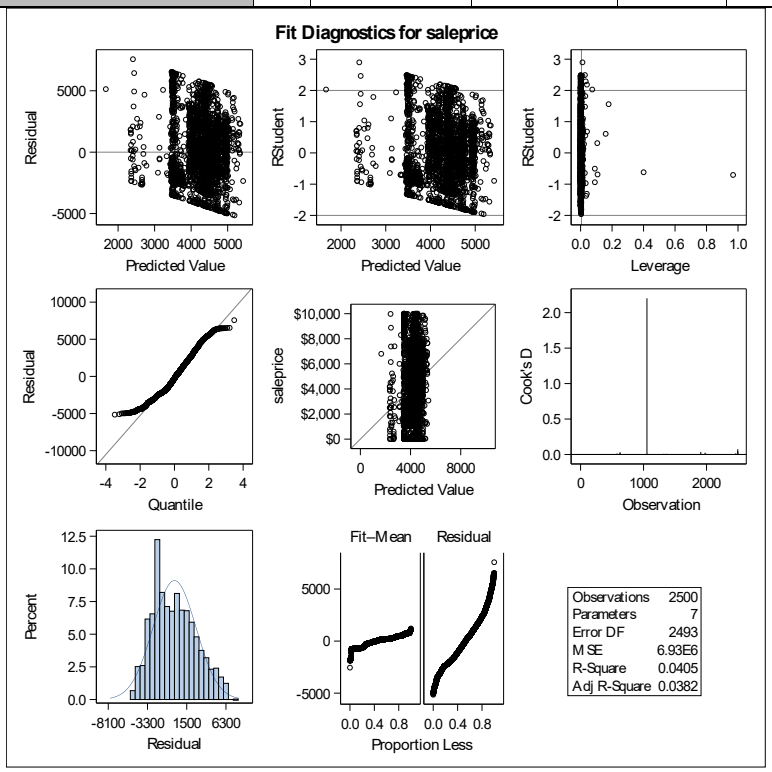
```
proc reg data=bkhomes;
 model saleprice=safestfortotcrimeranks yearbuilt
 medianhouseincome totalunits landsquarefeet
 grosssquarefeet;
run;
```

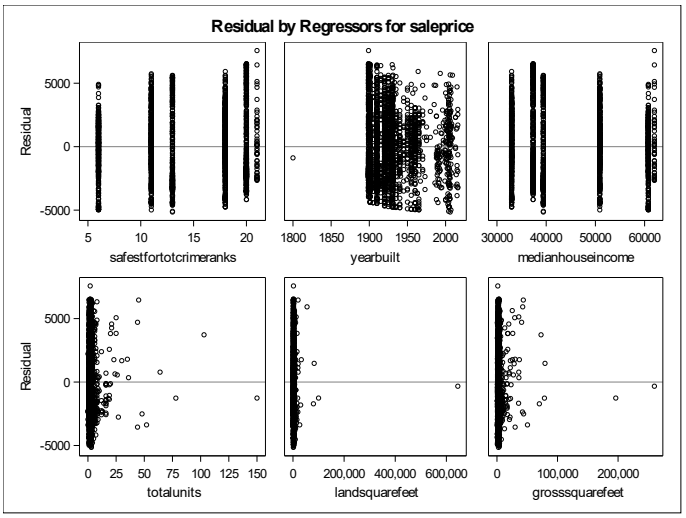
The p value is low since it is less than 0.0001 which means the significance is high. The diagnostics plots show a lot of variation. The residuals appear to show concentrated frequency of points in the center.

Analysis of Variance					
Source	**DF**	**Sum of Squares**	**Mean Square**	**F Value**	**Pr > F**
Model	6	728395797	121399300	17.52	<.0001
Error	2493	1727128256 0	6927911		
Corrected Total	2499	1799967835 8			

Root MSE	2632.09255	**R-Square**	0.0405
Dependent Mean	4212.88468	**Adj R-Sq**	0.0382
Coeff Var	62.47720		

Parameter Estimates					
Variable	**DF**	**Parameter Estimate**	**Standard Error**	**t Value**	**Pr > |t|**
Intercept	1	-6611.65020	3665.81416	-1.80	0.0714
safestfortotcrimeranks	1	-137.13583	19.13714	-7.17	<.0001
yearbuilt	1	7.55399	1.83201	4.12	<.0001
medianhouseincome	1	-0.03969	0.00932	-4.26	<.0001
totalunits	1	18.31450	21.68550	0.84	0.3984
landsquarefeet	1	0.00919	0.01076	0.85	0.3932
grosssquarefeet	1	-0.02934	0.02318	-1.27	0.2057





Problem 8

The variation is very high in problem 7. The relationship between sale price and the safety ranks, years built, and the median household incomes are all significant at a 95% level.