```
ods html close;
ods html;
ods rtf file='c:\courses\courses\ma416\MultReg.rtf';
options nodate nonumber nocenter;
title1 'Multiple Regression of College GPA vs. Independent Variables';
data new;
    input CollegeGPA HighSchoolGPA SAT Quality;
/** Quality = Quality of letters of recommendation on a scale of 1-10 **/
/** with 10 being high quality.                                       **/
cards;
2.04  2.01  1070   5
2.56  3.40  1254   6
3.75  3.68  1466   6
1.10  1.54   706   4
3.00  3.32  1160   5
0.05  0.33   756   3
1.38  0.36  1058   2
1.50  1.97  1008   7
1.38  2.03  1104   4
4.01  2.05  1200   7
1.50  2.13   896   7
1.29  1.34   848   3
1.90  1.51   958   5
3.11  3.12  1246   6
1.92  2.14  1106   4
0.81  2.60   790   5
1.01  1.90   954   4
3.66  3.06  1500   6
2.00  1.60  1046   5
2.05  1.96  1054   4
2.60  1.96  1198   6
2.55  1.56   940   3
0.38  1.60   456   6
2.48  1.92  1150   7
2.74  3.09   636   6
1.77  0.78   744   5
1.61  2.12   644   5
0.99  1.85   842   3
1.62  1.78   852   5
2.03  1.03  1170   3
3.50  3.44  1034  10
3.18  2.42  1202   5
2.39  1.74  1018   5
1.48  1.89  1180   5
1.54  1.43   952   3
1.57  1.64  1038   4
2.46  2.69  1090   6
2.42  1.79   694   5
2.11  2.72  1096   6
2.04  2.15  1114   5
1.68  2.22  1256   6
1.64  1.55  1208   5
2.41  2.34   820   6
2.10  2.92  1222   4
1.40  2.10  1120   5
2.03  1.64   886   4
1.99  2.83  1126   7
2.24  1.76  1158   4
0.45  1.81   676   6
2.31  2.68  1214   7
2.41  2.55  1136   6
2.56  2.70  1264   6
2.50  1.66  1116   3
```

```
2.92  2.23  1292   4



2.35  2.01   604   5
2.82  1.24   854   6
1.80  1.95   814   6
1.29  1.73   778   3
1.68  1.08   800   2
3.44  3.46  1424   7
1.90  3.01   950   6
2.06  0.54  1056   3
3.30  3.20   956   8
1.80  1.50  1352   5
2.00  1.71   852   5
1.68  1.99  1168   5
1.94  2.76   970   6
0.97  1.56   776   4
1.12  1.78   854   6
1.31  1.32  1232   5
1.68  0.87  1140   6
3.09  1.75  1084   4
1.87  1.41   954   2
2.00  2.77  1000   4
2.39  1.78  1084   4
1.50  1.34  1058   4
1.82  1.52   816   5
1.80  2.97  1146   7
2.01  1.75  1000   6
1.88  1.64   856   4
1.64  1.80   798   4
2.42  3.37  1324   6
0.22  1.15   704   6
2.31  1.72  1222   5
0.95  2.27   948   6
1.99  2.85  1182   8
1.86  2.21  1000   6
1.79  1.94   910   6
3.02  4.25  1374   9
1.85  1.83  1014   6
1.98  2.75  1420   7
2.15  1.71   400   6
1.46  2.20   998   7
2.29  2.13   776   6
2.39  2.38  1134   7
1.80  1.64   772   4
2.64  1.87  1304   6
2.08  2.53  1212   4
0.70  1.78   818   6
0.89  1.20   864   2
run;
proc reg data=new;
      model CollegeGPA=HighSchoolGPA SAT Quality/p r ss1 ss2 clb;
      SAT: test SAT=0;
      SAT_Quality: test SAT=Quality=0;
      plot residual.*CollegeGPA='*';
run;
title1;
ods rtf close;
```

We already reviewed these two "test" statements. Recall they calculate F tests for testing if various sets of betas are equal 0. I just wanted to bring up here that these are called "Partial F tests".

This statement prints a plot of the residuals (note the SAS keyword "residual." -and please note the period at the end of the word residual - this is not a mistake and must be included in this statement) on the y-axis versus the dependent variable College GPA on the x-axis. The command ='*' simply tells SAS to use an asterisk (*) as the symbol to represent each point in the plot (though for some reason, SAS used a "+" instead of an asterisk in the graph below!).

This graph is of interest because if there's truly a linear relationship between Y and the X's and if all assumptions for linear regression hold, this scatter plot of the residuals vs. College GPA will have no distinct pattern (e.g., the plot will look like a bunch of asterisks were thrown in the air and landed randomly and haphazardly on the plot). However, if there is some sort of pattern in this plot, then that means the relationship between Y and the X's may not be linear, or an assumption such as homoscedasticity or independence of observations is not met, or that there are other independent variables that should be added to the regression.

For example, the plot for these data can be found below. As you can see, as CollegeGPA increases, the residual tends to increase. This is not good! We do not want the value of the residuals to depend on the Y value. Rather we would like the mean of the residuals to be 0 at every Y value, with of course about half of the residuals above 0 and about half of the residuals below 0 at every Y value. Since this is not happening, something must not be correct in our model. See graph below for further discussion.

Multiple Regression of College GPA vs. Independent Variables

The REG Procedure
Model: MODEL1
Dependent Variable: CollegeGPA

Number of Observations Read	100
Number of Observations Used	100

Analysis of Variance					
Source	**DF**	**Sum of Squares**	**Mean Square**	**F Value**	**Pr > F**
Model	3	22.21437	7.40479	21.31	<.0001
Error	96	33.35831	0.34748		
Corrected Total	99	55.57268			

Root MSE	0.58948	**R-Square**	0.3997
Dependent Mean	1.98050	**Adj R-Sq**	0.3810
Coeff Var	29.76402		

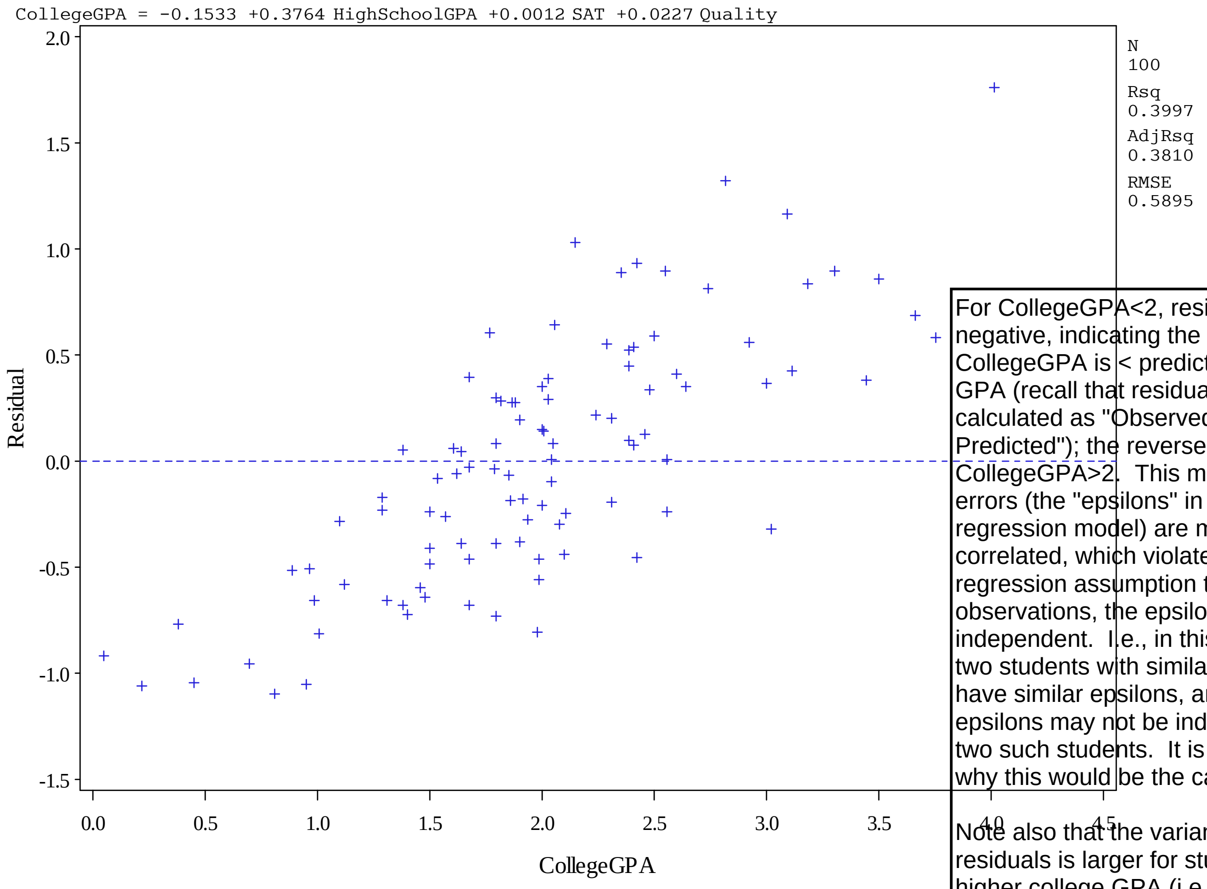
Parameter Estimates									
Variable	**DF**	**Parameter Estimate**	**Standard Error**	**t Value**	**Pr > |t|**	**Type I SS**	**Type II SS**	**95% Confidence Limits**	
Intercept	1	-0.15326	0.32294	-0.47	0.6362	392.23803	0.07827	-0.79429	0.48776
HighSchoolGPA	1	0.37635	0.11426	3.29	0.0014	16.51847	3.76981	0.14954	0.60316
SAT	1	0.00123	0.00030322	4.05	0.0001	5.62710	5.68934	0.00062505	0.00183
Quality	1	0.02268	0.05098	0.44	0.6574	0.06879	0.06879	-0.07851	0.12388

This output comes out because of the "p" and "r" options in our model statement above. "p" gives us the "Predicted Value" below, and the "r" option gives us the "Residual". You can ignore the other columns here.

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**
1	2.0400	2.0294	0.0620	0.0106	0.586	0.0180	| | |	0.000
2	2.5600	2.8010	0.1353	-0.2410	0.574	-0.420	| | |	0.002
...								
98	2.0800	2.3767	0.1212	-0.2967	0.577	-0.514	| *| |	0.003
99	0.7000	1.6564	0.0970	-0.9564	0.581	-1.645	| ***| |	0.019
100	0.8900	1.4038	0.1415	-0.5138	0.572	-0.898	| *| |	0.012

Sum of Residuals	0
Sum of Squared Residuals	33.35831
Predicted Residual SS (PRESS)	36.75093

Multiple Regression of College GPA vs. Independent Variables



For CollegeGPA<2, residuals are negative, indicating the observed CollegeGPA is < predicted College GPA (recall that residuals are calculated as "Observed minus Predicted"); the reverse is true for CollegeGPA>2. This means that the errors (the "epsilons" in the population regression model) are most likely correlated, which violates the regression assumption that for any two observations, the epsilons are independent. I.e., in this analysis, any two students with similar CollegeGPA have similar epsilons, and hence the epsilons may not be independent for two such students. It is unknown as to why this would be the case.

Note also that the variance of the residuals is larger for students with higher college GPA (i.e., as CollegeGPA increases, the spread of the residuals increases). This can be shown to indicate lack of homoscedasticity.

PLEASE REVIEW SECTION 3.3 IN THE BOOK FOR FURTHER DISCUSSION OF RESIDUALS (this section is for simple linear regression but applies for multiple linear regression).

Test SAT Results for Dependent Variable CollegeGPA

Source	DF	Mean Square	F Value	Pr > F
Numerator	1	5.68934	16.37	0.0001
Denominator	96	0.34748		

Test SAT_Quality Results for Dependent Variable CollegeGPA

Source	DF	Mean Square	F Value	Pr > F
Numerator	2	2.84795	8.20	0.0005
Denominator	96	0.34748		