Chris Hsu
Math 242
Dr Neal
Final Project

Student Performance Data

The data I looked at was the Student Performance Data Set from the UCI website which consisted of data from two Portuguese high schools to measure their math grades. The data attributes included the student grades, demographic, social and school related features and it was collected by using school reports and questionnaires. This dataset included 33 variables with 395 observations.

	school	sex	age	address	famsize	Pstatus	Medu	Fedu	Mjob	Fjob	reason	guardian	traveltime	studytime	failures	schoolsup	famsup	paid	activities	nursery	higher	internet	romantic	famrel	freetime	goout	Dalc	Walc	health	absences	G1	G2	G3
1	GP	F	18	U	GT3	A	4	4	at_home	teacher	course	mother	2	2	0	yes	no	no	no	yes	yes	no	no	4	3	4	1	1	3	6	5	6	6
2	GP	F	17	U	GT3	T	1	1	at_home	other	course	father	1	2	0	no	yes	no	no	no	yes	yes	no	5	3	3	1	1	3	4	5	5	6
3	GP	F	15	U	LE3	T	1	1	at_home	other	other	mother	1	2	3	yes	no	yes	no	yes	yes	yes	no	4	3	2	2	3	3	10	7	8	10
4	GP	F	15	U	GT3	T	4	2	health	services	home	mother	1	3	0	no	yes	yes	yes	yes	yes	yes	yes	3	2	2	1	1	5	2	15	14	15
5	GP	F	16	U	GT3	T	3	3	other	other	home	father	1	2	0	no	yes	yes	no	yes	yes	no	no	4	3	2	1	2	5	4	6	10	10
6	GP	M	16	U	LE3	T	4	3	services	other	reputation	mother	1	2	0	no	yes	yes	yes	yes	yes	yes	no	5	4	2	1	2	5	10	15	15	15
7	GP	M	16	U	LE3	T	2	2	other	other	home	mother	1	2	0	no	no	no	no	yes	yes	yes	no	4	4	4	1	1	3	0	12	12	11
8	GP	F	17	U	GT3	A	4	4	other	teacher	home	mother	2	2	0	yes	yes	no	no	yes	yes	no	no	4	1	4	1	1	1	6	6	5	6
9	GP	M	15	U	LE3	A	3	2	services	other	home	mother	1	2	0	no	yes	yes	no	yes	yes	yes	no	4	2	2	1	1	1	0	16	18	19
10	GP	M	15	U	GT3	T	3	4	other	other	home	mother	1	2	0	no	yes	yes	yes	yes	yes	yes	no	5	5	1	1	1	5	0	14	15	15

There were many variables included in this set from which one of the portuguese schools the student went to, to their final grade. Some of the variables we looked at were Age, famrel which is quality of family relationships (numeric: from 1 - very bad to 5 - excellent), freetime - free time after school (numeric: from 1 - very low to 5 - very high), goout - going out with friends (numeric: from 1 - very low to 5 - very high), Dalc - workday alcohol consumption (numeric: from 1 - very low to 5 - very high), Walc - weekend alcohol consumption (numeric: from 1 - very low to 5 - very high), health - current health status (numeric: from 1 - very bad to 5 - very good), absences - number of school absences (numeric: from 0 to 93), G1 - first period grade (numeric: from 0 to 20), G2 - second period grade (numeric: from 0 to 20), G3 - final grade (numeric: from 0 to 20, output target). We are looking at G3 the Final Math grade for each student. Since the x variables and y variables I am looking at are continuous I am going to use a regression model to determine what variables correspond with having the best final math grade. For my hypothesis,

the null hypothesis would be that none of the variables are statistically significant, and for my null hypothesis would be at least one of the variables are statistically significant in predicting the final math grade.

For my first model I ran a best subset and then stepwise regression test for the variable G3, and found that there are many significant variables.

```
         schoolMS sexM age addressU famsizeLE3 PstatusT Medu Fedu Mjobhealth Mjobother Mjobservices Mjobteacher
1  ( 1 ) " "      " "  " " " "     " "        " "      " "  " "  " "       " "       " "         " "
2  ( 1 ) " "      " "  " " " "     " "        " "      " "  " "  " "       " "       " "         " "
3  ( 1 ) " "      " "  " " " "     " "        " "      " "  " "  " "       " "       " "         " "
4  ( 1 ) " "      " "  " " " "     " "        " "      " "  " "  " "       " "       " "         " "
5  ( 1 ) " "      " "  "*" " "     " "        " "      " "  " "  " "       " "       " "         " "
6  ( 1 ) " "      " "  "*" " "     " "        " "      " "  " "  " "       " "       " "         " "
7  ( 1 ) " "      " "  "*" " "     " "        " "      " "  " "  " "       " "       " "         " "
8  ( 1 ) "*"      " "  "*" " "     " "        " "      " "  " "  " "       " "       " "         " "
9  ( 1 ) "*"      " "  "*" " "     " "        " "      " "  " "  " "       " "       " "         " "
10  ( 1 ) "*"     " "  "*" " "     " "        " "      " "  " "  " "       " "       " "         " "
         Fjobhealth Fjobother Fjobservices Fjobteacher reasonhome reasonother reasonreputation guardianmother
1  ( 1 ) " "        " "       " "         " "        " "        " "         " "              " "
2  ( 1 ) " "        " "       " "         " "        " "        " "         " "              " "
3  ( 1 ) " "        " "       " "         " "        " "        " "         " "              " "
4  ( 1 ) " "        " "       " "         " "        " "        " "         " "              " "
5  ( 1 ) " "        " "       " "         " "        " "        " "         " "              " "
6  ( 1 ) " "        " "       "*"         " "        " "        " "         " "              " "
7  ( 1 ) " "        " "       "*"         " "        " "        " "         " "              " "
8  ( 1 ) " "        " "       "*"         " "        " "        " "         " "              " "
9  ( 1 ) " "        " "       "*"         " "        " "        " "         " "              " "
10  ( 1 ) " "       " "       "*"         " "        " "        " "         " "              " "
         guardianother traveltime studytime failures schoolsupyes famsupyes paidyes activitiesyes nurseryyes
1  ( 1 ) " "           " "        " "       " "      " "          " "        " "      " "           " "
2  ( 1 ) " "           " "        " "       " "      " "          " "        " "      " "           " "
3  ( 1 ) " "           " "        " "       " "      " "          " "        " "      " "           " "
4  ( 1 ) " "           " "        " "       " "      " "          " "        " "      " "           " "
5  ( 1 ) " "           " "        " "       " "      " "          " "        " "      " "           " "
6  ( 1 ) " "           " "        " "       " "      " "          " "        " "      " "           " "
7  ( 1 ) " "           " "        " "       " "      " "          " "        " "      " "           " "
8  ( 1 )
9  ( 1 ) " "           " "        " "       " "      " "          " "        " "      " "           " "
10  ( 1 ) " "          " "        " "       " "      " "          " "        " "      "*"           " "
         higheryes internetyes romanticyes famrel freetime goout Dalc Walc health absences G1  G2
1  ( 1 ) " "       " "         " "        " "    " "      " "   " "  " "  " "    " "      " " "*"
2  ( 1 ) " "       " "         " "        "*"    " "      " "   " "  " "  " "    " "      " " "*"
3  ( 1 ) " "       " "         " "        "*"    " "      " "   " "  " "  " "    "*"      " " "*"
4  ( 1 ) " "       " "         " "        "*"    " "      " "   " "  " "  " "    "*"      "*" "*"
5  ( 1 ) " "       " "         " "        "*"    " "      " "   " "  " "  " "    "*"      "*" "*"
6  ( 1 ) " "       " "         " "        "*"    " "      " "   " "  " "  " "    "*"      "*" "*"
7  ( 1 ) " "       " "         " "        "*"    " "      " "   " "  " "  "*"    "*"      "*" "*"
8  ( 1 ) " "       " "         " "        "*"    " "      " "   " "  " "  "*"    "*"      "*" "*"
9  ( 1 ) " "       " "         "*"        "*"    " "      " "   " "  " "  "*"    "*"      "*" "*"
10  ( 1 ) " "      " "         "*"        "*"    " "      " "   " "  " "  "*"    "*"      "*" "*"
```

For this best subset test I got that G2, famrel, absences, G1, age, Fjob (services), Walc, School (MS), romantic(yes), and activities were the top ten significant variables in this dataset. Then I ran a stepwise regression test.

```
Step:  AIC=501.92
G3 ~ G2 + famrel + absences + G1 + age + activities + Walc +
    romantic + school

            Df Sum of Sq    RSS    AIC
<none>                  1338.0 501.92
+ failures   1    5.5442 1332.5 502.28
+ schoolsup  1    5.5306 1332.5 502.28
+ Dalc       1    2.3951 1335.6 503.21
+ famsup     1    2.3777 1335.6 503.21
+ health     1    2.2750 1335.7 503.24
+ internet   1    2.1553 1335.8 503.28
+ studytime  1    1.3703 1336.6 503.51
+ freetime   1    1.2613 1336.7 503.54
+ traveltime 1    1.2457 1336.8 503.55
+ Medu       1    1.1679 1336.8 503.57
+ nursery    1    1.1527 1336.8 503.58
+ Pstatus    1    1.1134 1336.9 503.59
+ sex        1    0.9004 1337.1 503.65
+ higher     1    0.7921 1337.2 503.68
+ Fedu       1    0.7102 1337.3 503.71
+ paid       1    0.6183 1337.4 503.73
+ famsize    1    0.2163 1337.8 503.85
+ goout      1    0.1263 1337.9 503.88
+ address    1    0.0228 1338.0 503.91
+ guardian   2    5.6634 1332.3 504.24
+ Fjob       4   16.5231 1321.5 505.01
+ reason     3    9.3473 1328.7 505.15
+ Mjob       4    0.5407 1337.5 509.76

Call:
lm(formula = G3 ~ G2 + famrel + absences + G1 + age + activities +
    Walc + romantic + school, data = students)

Coefficients:
  (Intercept)            G2         famrel       absences             G1            age  activitiesyes
      0.65012        0.96153        0.38361        0.04714        0.18058       -0.26017       -0.32049
         Walc  romanticyes       schoolMS
      0.11979       -0.32668        0.45495
```

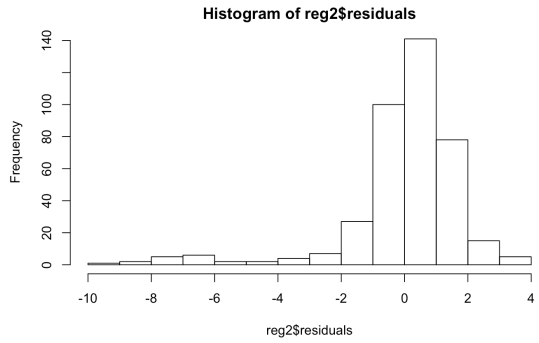
For the stepwise regression test it gave me a model with G2, famrel, absences, G1, age, activities, Walc, romantic, and school as variables for my model. So I ran a linear regression model with these variables.

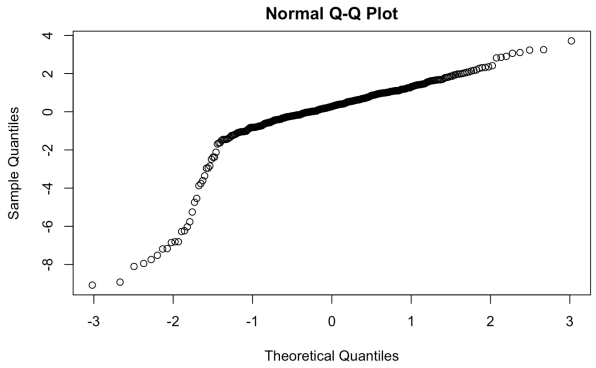
```
lm(formula = G3 ~ G2 + famrel + absences + G1 + age + activities +
    Walc + romantic + school, data = students)

Residuals:
    Min      1Q  Median      3Q     Max
-8.8416 -0.4534  0.2645  1.0247  4.0315

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   0.65012    1.45806   0.446 0.655933
G2            0.96153    0.04904  19.607  < 2e-16 ***
famrel        0.38361    0.10668   3.596 0.000365 ***
absences      0.04714    0.01232   3.826 0.000152 ***
G1            0.18058    0.05517   3.273 0.001159 **
age          -0.26017    0.08405  -3.095 0.002109 **
activitiesyes -0.32049    0.18989  -1.688 0.092278 .
Walc          0.11979    0.07521   1.593 0.112050
romanticyes  -0.32668    0.20622  -1.584 0.113989
schoolMS      0.45495    0.32293   1.409 0.159693
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.864 on 385 degrees of freedom
Multiple R-squared:  0.8382,    Adjusted R-squared:  0.8344
F-statistic: 221.6 on 9 and 385 DF,  p-value: < 2.2e-16
```

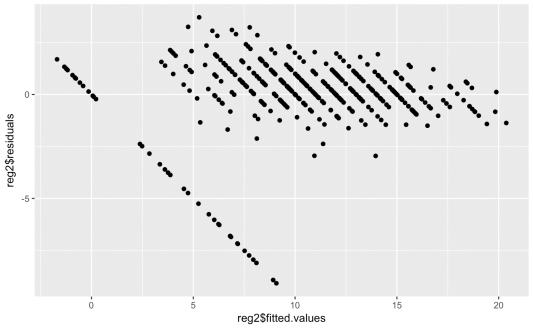
With this model above I found out that G2, famrel, absences, G1, and Age are significant which would lead me to create my first model with these significant variables. I ran a linear regression model with these significant variables.

```
Call:
lm(formula = G3 ~ G2 + famrel + absences + G1 + age, data = students)

Residuals:
    Min      1Q  Median      3Q     Max
-9.0811 -0.4081  0.2733  0.9927  3.7111

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.07765    1.37626  -0.056 0.955033
G2           0.97804    0.04895  19.981  < 2e-16 ***
famrel       0.35725    0.10622   3.363 0.000847 ***
absences     0.04365    0.01205   3.623 0.000329 ***
G1           0.15794    0.05503   2.870 0.004329 **
age         -0.20167    0.07679  -2.626 0.008978 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.881 on 389 degrees of freedom
Multiple R-squared:  0.8336,    Adjusted R-squared:  0.8315
F-statistic: 389.8 on 5 and 389 DF,  p-value: < 2.2e-16
```

With this model we see that it is pretty good with all the variables significant and a p-value less than 2.2e-16 which meets our 0.05 cutoff. We also see our R squared value at .83 or 83% which show us that 83% of our data is represented with this model which is really good. We also see a very high F-statistic of 389.8 which is really good. Next we run some tests to test for normality, homoscedasticity, and autocorrelation.







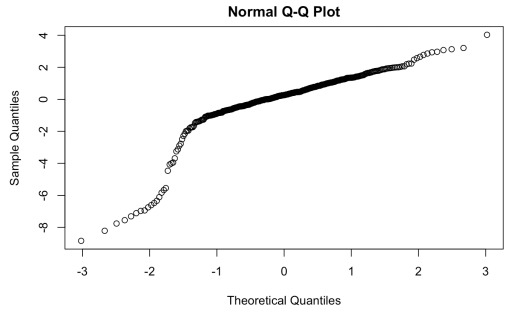

```
        Shapiro-Wilk normality test

data:  reg2$residuals
W = 0.77152, p-value < 2.2e-16

Non-constant Variance Score Test
Variance formula: ~ fitted.values
Chisquare = 87.37829    Df = 1     p = 8.962888e-21
```

What we see from these tests are not good. First with the histogram of the residuals to test for normality, we do not really see a nice bell shaped curve, and instead we see some over dispersion with some residuals being as low as -10 which exceeds our -2 to 2 goal. Also for normality, we ran a shapiro test and we wanted to p value to be as high as possible, but instead we get a very low p value of less than 2.2e-16, but that could also be because we have way more than 100 data points. For our QQnorm plot we do not really see a straight line which further strengthens our case against normality. Then we ran a NCV test which would test for homoscedasticity which we want, but when we ran the NCV test we got a very low p value rejecting homoscedasticity. Then we ran a residual vs fitted plot to test for homoscedasticity and autocorrelation, and since the data points were not complete random and scattered we have autocorrelation and we have heteroscedastic which is not good. In order to combat these problems I added more variables to try and prevent the problems.

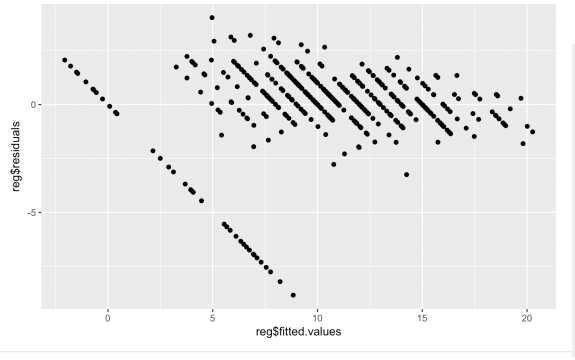
For our second model I would add back the variables from the original stepwise model and run a linear regression test.

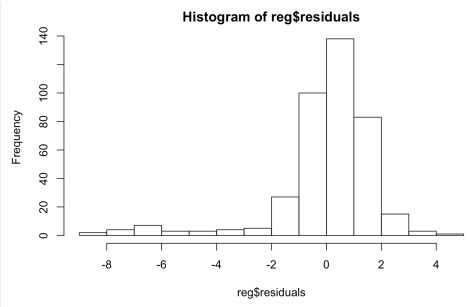
```
lm(formula = G3 ~ G2 + famrel + absences + G1 + age + activities +
    Walc + romantic + school, data = students)

Residuals:
    Min      1Q  Median      3Q     Max
-8.8416 -0.4534  0.2645  1.0247  4.0315

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   0.65012    1.45806   0.446 0.655933
G2            0.96153    0.04904  19.607  < 2e-16 ***
famrel        0.38361    0.10668   3.596 0.000365 ***
absences      0.04714    0.01232   3.826 0.000152 ***
G1            0.18058    0.05517   3.273 0.001159 **
age          -0.26017    0.08405  -3.095 0.002109 **
activitiesyes -0.32049    0.18989  -1.688 0.092278 .
Walc          0.11979    0.07521   1.593 0.112050
romanticyes  -0.32668    0.20622  -1.584 0.113989
schoolMS      0.45495    0.32293   1.409 0.159693
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.864 on 385 degrees of freedom
Multiple R-squared:  0.8382,    Adjusted R-squared:  0.8344
F-statistic: 221.6 on 9 and 385 DF,  p-value: < 2.2e-16
```

In this model we see that G2, farmel, absences,G1, and age are the significant variables because they have a p-value below .05. We then look at our R squared which is 83% again, so 83% of our data points are accounted for in this model. Again we get a p value of 2.2e-16 which is extremely good and get a high F-statistic of 221.6 which is really good. So now we run all the tests again to test for normality, homoscedasticity, and autocorrelation.




```
        Shapiro-Wilk normality test

data:  reg$residuals
W = 0.79675, p-value < 2.2e-16

Non-constant Variance Score Test
Variance formula: ~ fitted.values
Chisquare = 91.92734    Df = 1     p = 8.991798e-22
```



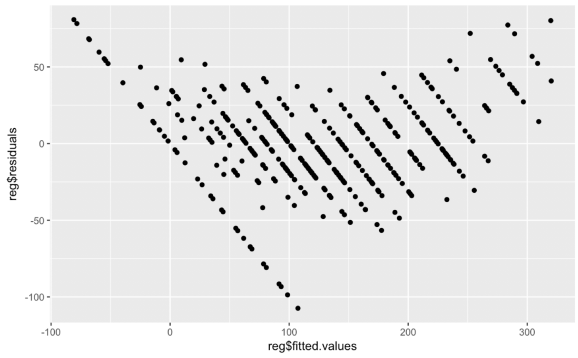


Again we see almost the same results with this model having autocorrelation, heteroscedasticity, and not being normal. We see that the QQnorm plot is not linear hurting normality. Our Shapiro test is below .05 rejecting normality. The NCV test being below .05 rejecting homoscedasticity, the fitted vs residual plot showing us that we have autocorrelation and not being homoscedastic. Then the histogram of the residuals not showing us a normal bell shaped curve showing us it does not have normality. So we move on to our third model to use transformations to see if we can correct this problem.

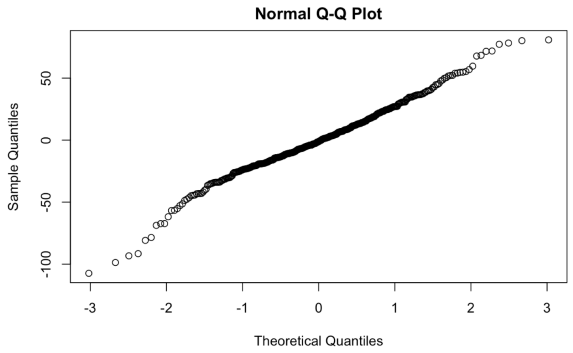
For our third model we decide to make a transformation and mutate G3 by cubing G3 of it to see if it changes anything. So we run another linear regression model with the same variables as before.

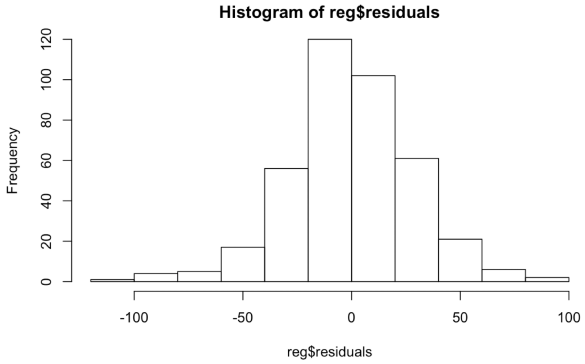
```
Call:
lm(formula = G3cube ~ G2 + famrel + absences + G1 + age, data = students4)

Residuals:
     Min       1Q   Median       3Q      Max
-107.397  -17.689   -1.042   17.460   80.844

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -75.2160    21.2868  -3.533 0.000459 ***
G2           14.0884     0.7571  18.609  < 2e-16 ***
famrel        5.5300     1.6429   3.366 0.000839 ***
absences     -0.0974     0.1863  -0.523 0.601492
G1            8.7203     0.8512  10.245  < 2e-16 ***
age          -3.7552     1.1878  -3.162 0.001692 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 29.09 on 389 degrees of freedom
Multiple R-squared:  0.8845,    Adjusted R-squared:  0.883
F-statistic: 595.8 on 5 and 389 DF,  p-value: < 2.2e-16
```

We now see some changes with all the variables being statistically significant except for absences. We see that that R squared has increased to .88 or 88%, so now 88% of our data point are represented by this model. For our F statistic we had a large increase up to 595.8 which very good and a p value below .05 which means our model is statistically significant. Now we will test for normality, homoscedasticity, and autocorrelation.




```
        Shapiro-Wilk normality test

data:  reg$residuals
W = 0.9854, p-value = 0.0005179

Non-constant Variance Score Test
Variance formula: ~ fitted.values
Chisquare = 2.637871    Df = 1     p = 0.1043433
```





In these tests we see a little improvement, but not good enough. With the shapiro test we see that our p-value increased but it is not above .05, so we still have to reject normality. With our histogram of residuals we see a better bell shaped curve, but we see residuals from 100 to -100 which is way out of our 2 to -2 range so we have overdispersion still. With our residual vs fitted plot we see that the data points are more spread out but still can see autocorrelation and can tell it is not homoscedastic. With the NCV test we can actually see that it is above 0.05 so we cannot reject normality, but it is still not as high as we would like.

In conclusion we definitely had flaws in our model. Our models did not have homoscedasticity, normality, and it had autocorrelation. To try and combat these flaws I tried adding variables and logarithmic and quadratic transformations, but the models still had heteroscedasticity, autocorrelation, and was not normal. So these were the main problems with our models. We found that G2(second period grade), famrel (family relationship), number of absences, G1(first period grade), and Age are significant in predicting the final math grade in these two portuguese schools. With our first model we found that the mean grade was 10.41519

on the scale out of 20, and when we used our first model formula **(-.07765+(.97804*G2)+(.35725+Famrel)+(.04365*Absences)+(.15794*G1)+(-.20167*Age)**) to predict the average grade we get 13.30756 which is not too bad or too far off. We can tell by the formula as your second period grade increases and family relationship gets better/increases, this will most positively directly correlated with better grades, but as you get older in high school it seems that your grade would get worse. These variables in a way make sense because if you are doing well in your first and second period it would mean you are doing well in the math class. If you have a good family relationship, you do better in school makes sense because these family probably make sure these kids do well in school and if the student have a bad relationship they probably rebel at home and do not do their homework or pay attention in class. Age could have a negative correlation because they have people who have failed grades and are just way older than other kids and just do not care about school. What just confuses me a little is how absences have a positive correlation because I would think that the more absences you have the more likely you are to have a worse grade that a person with zero absences. Some future questions could be could we come up with some other variables to include in this data set like something to do with social life outside of class. Another question is could this model fit with students math grades in the United States and not just these two school in Portugal. If we have a dataset for the United States would we have a technology variable, like how much time spent playing video games or on their phone because technology is getting to be a huge distraction for students today in the United States. Overall our model was not the best but was not bad, it just could not pass normality, homoscedasticity, and autocorrelation.