INDIAN STATISTICAL INSTITUTE

ASSIGNMENT of STATISTICS:

Multivariate Analysis

NAME: SUBHRAJYOTY ROY

ROLL: BS-1613

COURSE: B.STAT YEAR- I

SUBJECT: STATISTICAL METHODS II

Introduction

In real life problems, most of the time, the unknown dependent variable seems to be dependent on more than one or two independent variables (or predictors). Hence, to analyze such a situation and relation between those variable we need multivariate analysis.

In ISI, the student is graded in both the mid-semestral exam and the semestral exam and the composite score is taken as the weighted average of the scores in those exams. Most of the university grades their students along these rules, but have different weightages on each part of the examinations. In this assignment, I merely go in the reverse direction to predict how these weights are carefully chosen.

About the Data

The Dataset is taken from a university in USA, where all the students are taken from a chemical engineering course. I have collected it from www.openmv.net .

There are six variables, including the one dependent variable i.e. the performance in first year Final examination in college, which I want to predict about. The predictors are the grades of their assignments (Practical), Midterm examination, Problem sheet and homework, performance in Tutorial classes and their GPA awarded in high school. Clearly, this will have some effect on their performance in Final examinations.

As the grades are averaged from all the seven papers of the course and scaled down to a ratio to 100, a linear model is supposedly a good fit for the dataset.

The Univariate Analysis

The univariate analysis on the data is performed and here are the important features of the dataset, as given below:

Variable Name	Minimum	Median	Maximum	Mean	Standard deviation	Coefficient of Variation
School GPA	4	8	8	7.316	0.93698	12.7%
Homework	14.09	73.30	90.74	67.84	19.5364	28.7%
Assignments	23.45	75.08	84.02	71.62	10.4105	14.5%
Tutorial	28.41	78.03	93.82	74.72	12.7681	17.1%
Midterm	23.43	58.33	91.67	57.11	16.6929	29.2%
Final	23.38	55.32	90.74	57.12	15.7148	27.5%

From the above univariate analysis of the dataset, it seems that the coefficient of variation of Midterm and Final exams are pretty close. Also, their mean matches approximately. This makes us suspicious that midterm marks may significantly affect the final scores of the students. The relation between final scores with other variables are not eminent, but will show up as we do bivariate and multivariate analysis.

The Bivariate Analysis

The very first feature of bivariate analysis is to find how these variables depend on each other and finding out their respective pair-wise correlation coefficient. This will also help to gather more information about how much interdependence the independent variables have, and also how

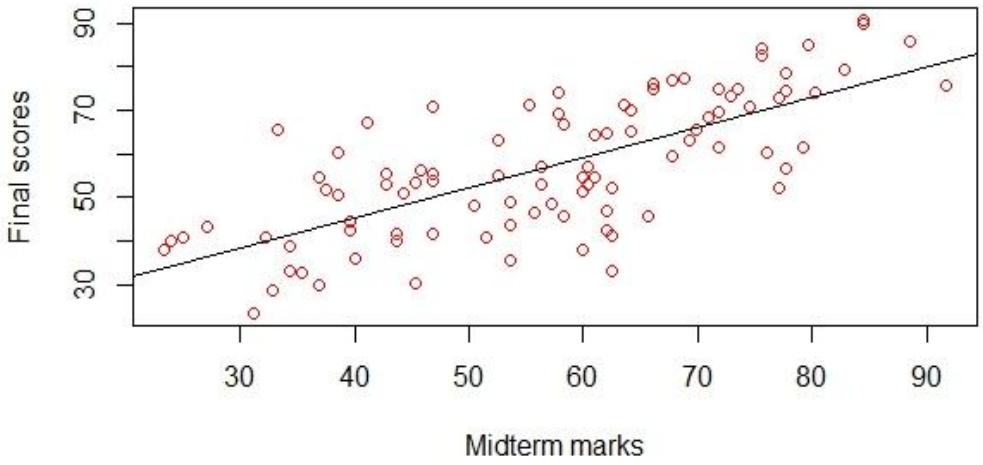
much each of the independent variable affects the dependent variable (i.e. the scores in Final examination).

Variables	School GPA	Assignment	Tutorial	Midterm	Homework
Assignment	0.0484				
Tutorial	0.4345	0.4590			
Midterm	-0.0586	0.2007	0.1486		
Homework	-0.0689	0.4832	0.2381	0.4271	
Final	0.0881	0.2863	0.2398	0.7244	0.4742

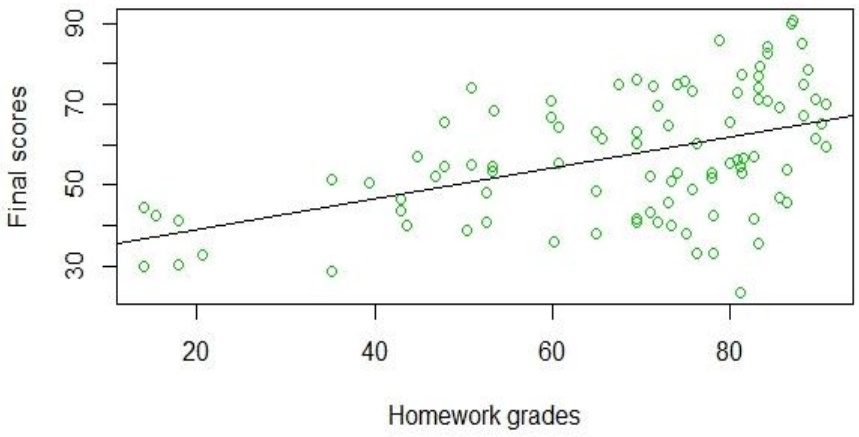
From the above correlation matrix, it can be seen that School GPA has extremely low effect on Final scores in their first year marks in college, which is quite surprising as it is always assumed that a good student would perform good in school and as well as in college. Moreover, the gap between those two standards are of merely one year.

Also, it is observed that Midterm and Homework are moderately correlated with the final scores, so it may be a good idea to take them as independent variable in the multivariate model to analyze the dependent variable. I also take Assignment and Tutorial with them to see whether the amount of explained part of final scores increases or not.

Here are a few scatter plots showing the relationship between the variables:

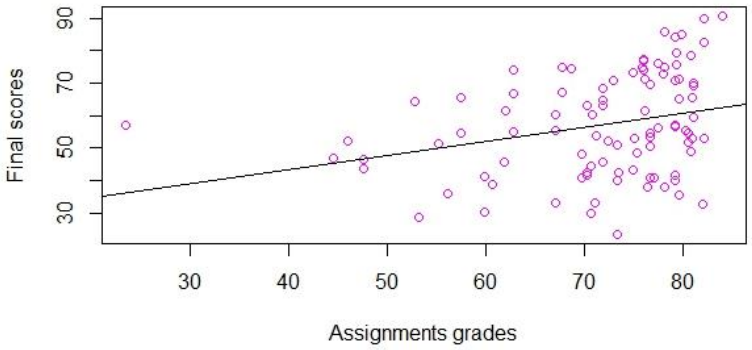


As expected, both the midterm marks and Final scores range over the same interval, following similar trend, as reflected in the scatter plots.

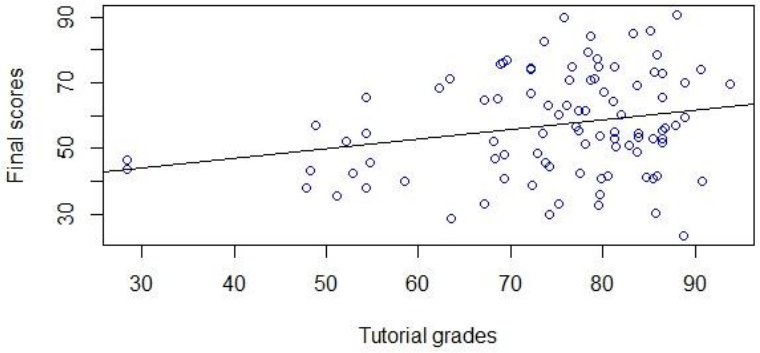


From the scatter plot between Homework grades and final scores, it is seen that most of

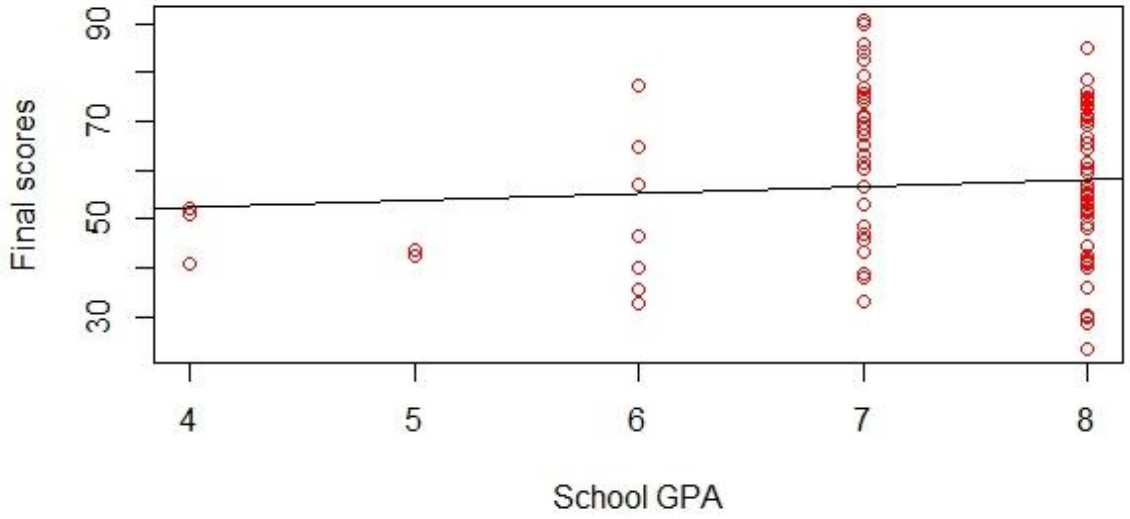
the datapoints lie in the homework grades above 70. It may be assumed that this gathering one a single side makes the correlation coefficient not to be reflected as a true measure of relationship between those two variables.



Here are the scatter plots showing the relationship between final scores in the exam with Assignment and Tutorial Grades respectively. It is seen that both of this relationship is somewhat weak.



I also tried to analyze why School GPA had so little effect on the final scores in first year exam, and the reason behind this unexpected thing is cleared by the plot below.



Clearly, School GPA's are awarded to the students in whole numbers, while after averaging out the final scores of a student over all the papers in first year, the average takes all sorts of different scores up to tenth's place after decimal. This makes the relationship between the variables to be so lower.

By the above analysis, to some extent it can be expected that taking School GPA as a linear predictor of the final scores is not a good idea. Hence, I will perform the multiple regression analysis with the exception of School GPA.

Multivariate Analysis

The Ideal Situation:

To perform multiple regression analysis, we assume a linear model satisfying the following properties:

1. $Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_p X_{pi} + \epsilon_i$, where $X_1, X_2, \ldots, X_p$'s are the predictors and ϵ_i's are the errors.
2. We assume that $\epsilon_i \sim N(0, \sigma^2)$ i.e. a normal distribution with mean 0 and variance σ^2.
3. Each of the predictors have their respective correlation close to 0, i.e. the predictors are linearly independent.

Using R, I have created four predictors with 100 many data points. The predictor variables are generated independently from different distributions. Then the errors are generated from a normal distribution with mean 0 and variance 25. Y is computed according to the formula given above and then on the whole set of data points, multiple regression analysis is performed to see how things behaves in case of the ideal situation.

The correlation matrix of the variables is given below:

Variables	Y	X1	X2	X3	X4
X1	0.5817				
X2	0.3316	0.0790			
X3	0.8568	0.0236	0.00075		
X4	0.0589	0.0538	0.0387	0.0201	
e	0.2876	0.0359	0.1028	0.1446	0.0677

Performing the multiple regression analysis on y with respect to x1, x2, x3 and x4, we get the following linear combination as the regression subspace;

Y = -1.3572 + 0.9971 x1 +1.0481 x2 + 1.0264 x3 + 5.9080 x4

After performing this regression, the following features of the data is found to come out:

1. Multiple R squared: 0.9821
2. Adjusted R squared: 0.9814
3. F statistic: 1304 on 4 (due to regression) and 95 (due to residuals) degrees of freedom.
4. Residual Standard Error: 5.048 on 95 degrees of freedom.
5. The following table is generated to provide the summary for the coefficients of the linear model;

Coefficients:

	Estimate	Std. Error	t value	Probability (>|t|)
(Intercept)	-1.35722	1.14574	-1.185	0.239
x1	0.99708	0.03681	27.084	$<2e^{-16}$
x2	1.04807	0.04815	21.767	$<2e^{-16}$
x3	1.02636	0.01879	54.628	$<2e^{-16}$
x4	5.90802	8.08299	0.731	0.467

6. For the residuals, we have the following summary;

```
Residuals:
     Min       1Q   Median       3Q      Max
-11.6311  -3.5176  -0.2039   2.9058  12.5863
```

With the standard deviation of the residuals being **4.94**.

7. The ANOVA table for this case becomes the following;

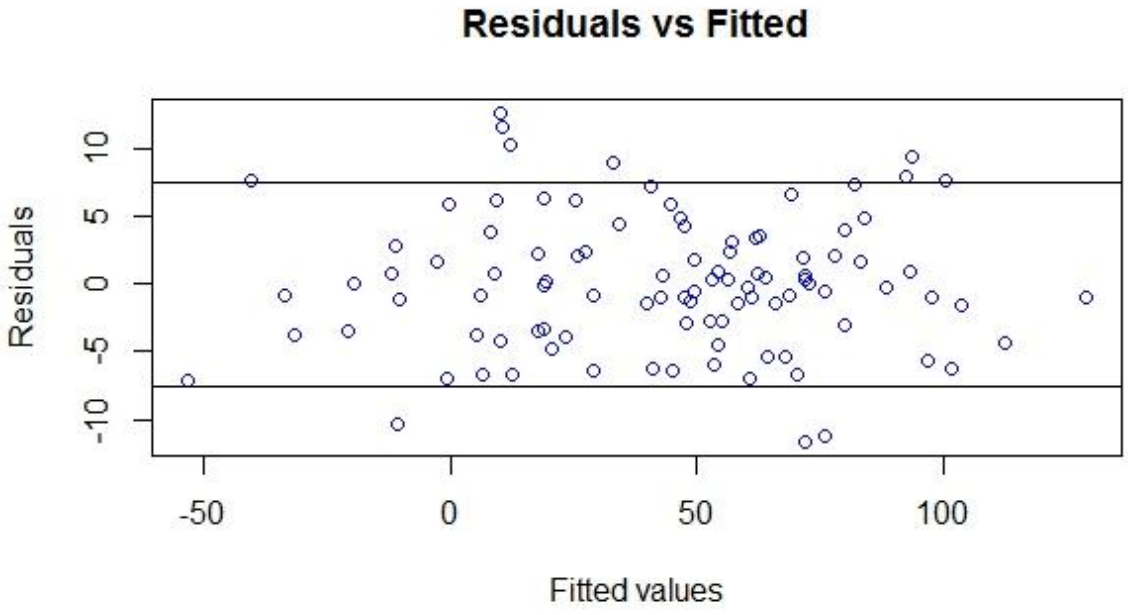
```
Analysis of Variance Table:

Response: y
          Df  Sum Sq  Mean Sq  F value  Pr(>F)
x1         1   45790   45790  1796.9150 <2e-16
x2         1   11073   11073   434.5119 <2e-16
x3         1   76069   76069  2985.1329 <2e-16
x4         1      14      14     0.5342 0.4666
Residuals 95    2421      25
```

8. Here is the corrected ANOVA Table reflecting the regression sum of squares (SSR) and residual sum of squares (RSS) contributing the total corrected sum of squares.

Variables	Degrees of Freedom	Sum squared	Mean squared
Regression	4	132946	33236.5
Residuals	95	2421	25
Total	99	135367	

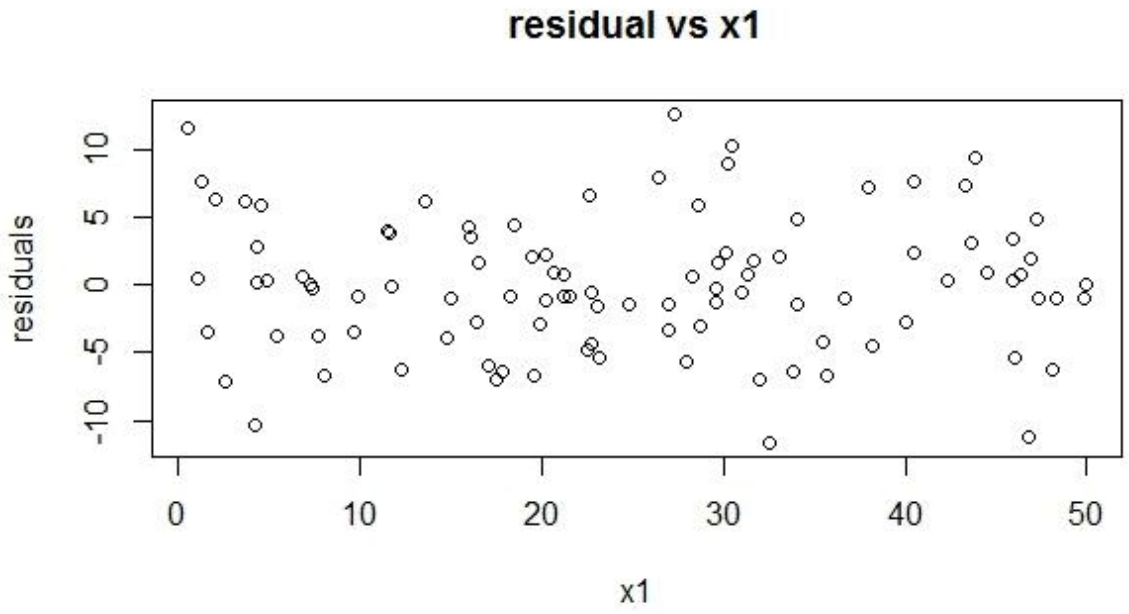
9. The residual plot between the residual and the fitted values is given below:

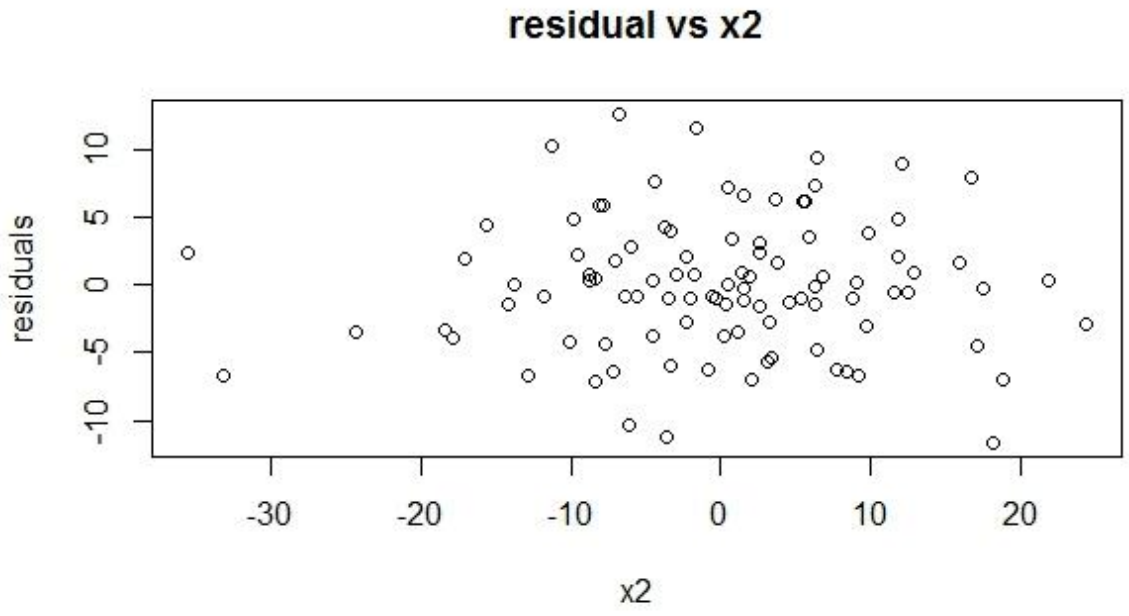


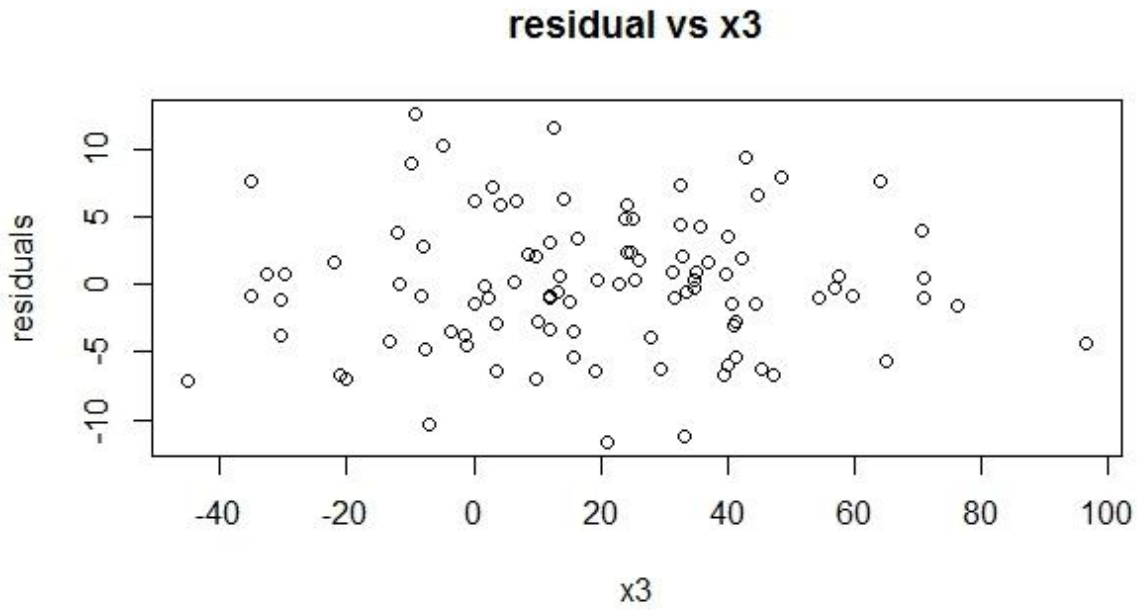
From the above plot, one can see that about 90% of the data points lie in the interval $(-1.5 * \sigma, 1.5 * \sigma)$. About 95% of the data points lie in the interval $(-2 * \sigma, 2 * \sigma)$, where σ is the standard deviation of the errors assumed in the linear model. Also, in the ideal case, as the theory

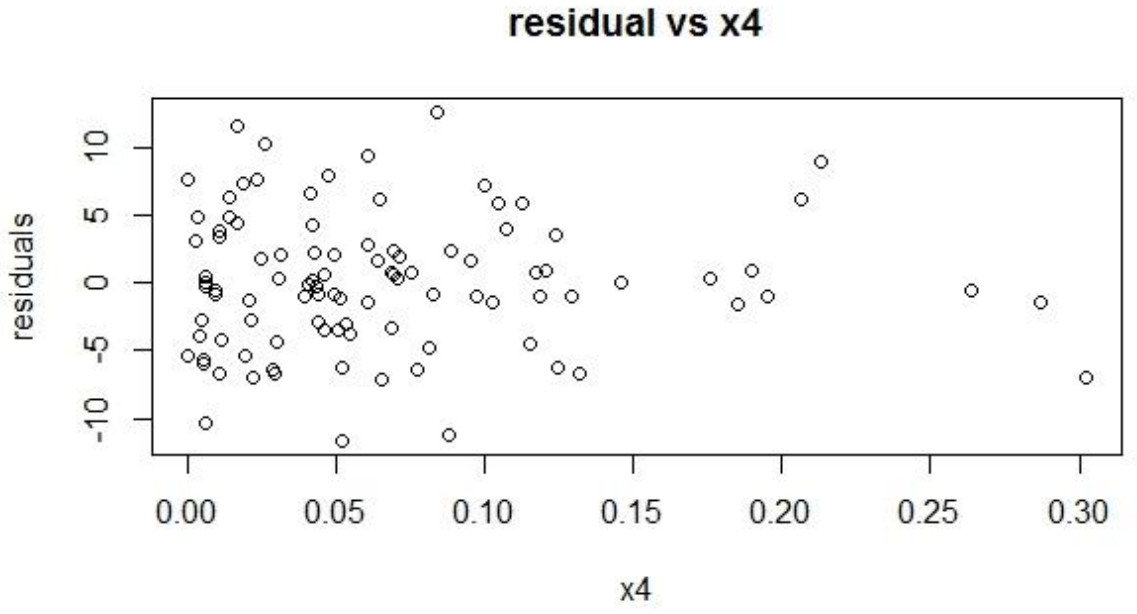
suggests there is not certain pattern in the residual plot, as if the points are distributed randomly over the whole range.

Here is given the individual residuals plots for each predictors;









Most of the residual plots are almost similar to that of 'Fitted' vs 'Residuals' plot, except the variable x4, as it came from an exponential distribution.

Here are the partial correlation found between the dependent variable and the independent ones, removing the effect of the rest.

Partial Correlation	X1	X2	X3	X4
Y	0.949249	0.912678	0.984453	0.074781

By the above partial correlation, it seems apparent that these factors indeed controls (to some extent) the value of the dependent variable. Also, the 4th variable is generated from an exponential distribution that shows a really low level relationship between normal and exponential distribution.

Now, after performing the multiple regression on the actual dataset, we can compare the result with the result of the ideal situation and based on that one can analyze the multivariate data.

Analysis of the Dataset:

The similar features that are found after regressing Final scores with all the independent regressors like School GPA, Assignments and Midterm marks etc. are given below:

1. The regression equation is as follows;

   Final = -11.36494 + 2.10736* School GPA + 0.08453* Assignment + 0.02727* Tutorial + 0.61906* Midterm + 0.14182* Homework

2. Multiple R squared: 0.5811
3. Adjusted R squared: 0.5576

4. F -statistic: 24.7 on 5 (due to regression) and 89 (due to residuals) degrees of freedom.

5. Residual Standard Error: 10.45 on 89 degrees of freedom.

6. The summary of the linear model is given by the table as follows:

   Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-11.36494	11.42154	-0.995	0.3224
School GPA	2.10736	1.31517	1.602	0.1126
Assignment	0.08453	0.13033	0.649	0.5183
Tutorial	0.02727	0.10807	0.252	0.8014
Midterm	0.61906	0.07348	8.425	$5.8e^{-13}$
Homework	0.14182	0.06860	2.067	0.0416

7. For the residuals after performing the multiple linear regression, we have the following summary;

Min.	1st Qu.	Median	3rd Qu.	Max.
-27.5100	-8.2970	0.8203	6.7510	26.490

   with a standard deviation of **10.17039**.

8. The ANOVA table for this regression analysis is given below;

   Analysis of Variance Table:

   Response: Final Scores

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
School GPA	1	180.5	180.5	1.6521	0.20201
Assignment	1	1850.8	1850.8	16.9415	$8.595e^{-5}$
Tutorial	1	226.7	226.7	2.0752	0.15322
Midterm	1	10765.7	10765.7	98.5434	$4.533e^{-16}$
Homework	1	466.9	466.9	4.2739	0.04161
Residuals	89	9723.1	109.2		

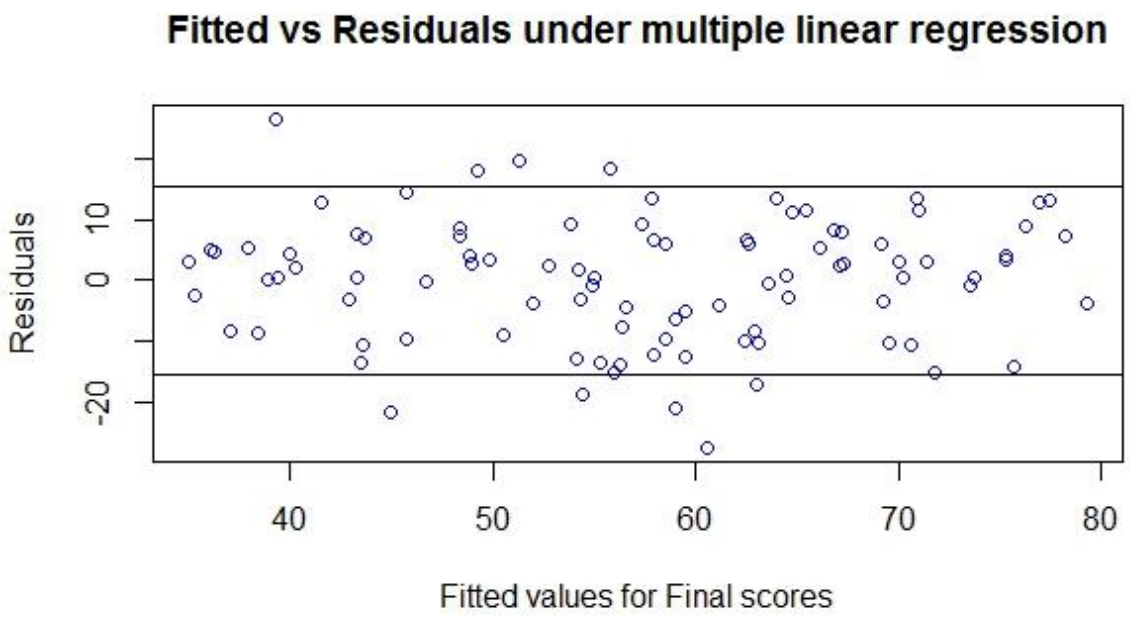
9. The partial correlations between the final scores and the independent factors are given below in a table:

Partial Correlation	School GPA	Assignments	Tutorial	Midterm	Homework
Final Scores	0.16745	0.06859	0.02673	0.66610	0.21405

10. Here is the ANOVA table reflecting the contribution of SSR and RSS as shown below;

Variables	**Degrees of Freedom**	**Sum squared**	**Mean squared**
Regression	5	13490.4	2698.08
Residuals	89	9723.1	109.24
Total	94	23213.5	

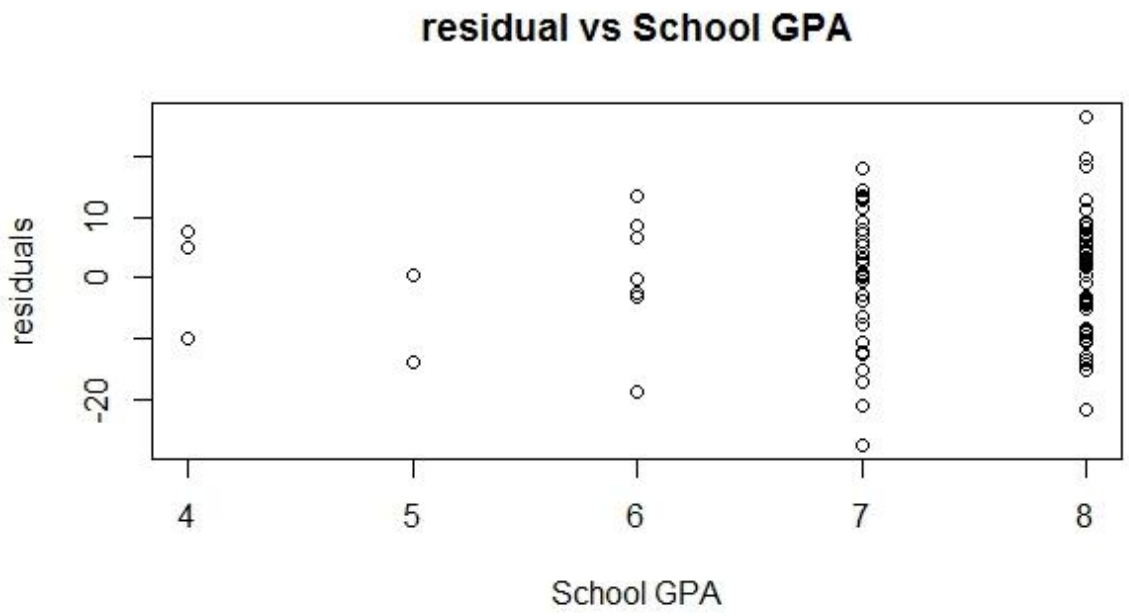
11. The residual plot between the fitted values and the residuals are given below;

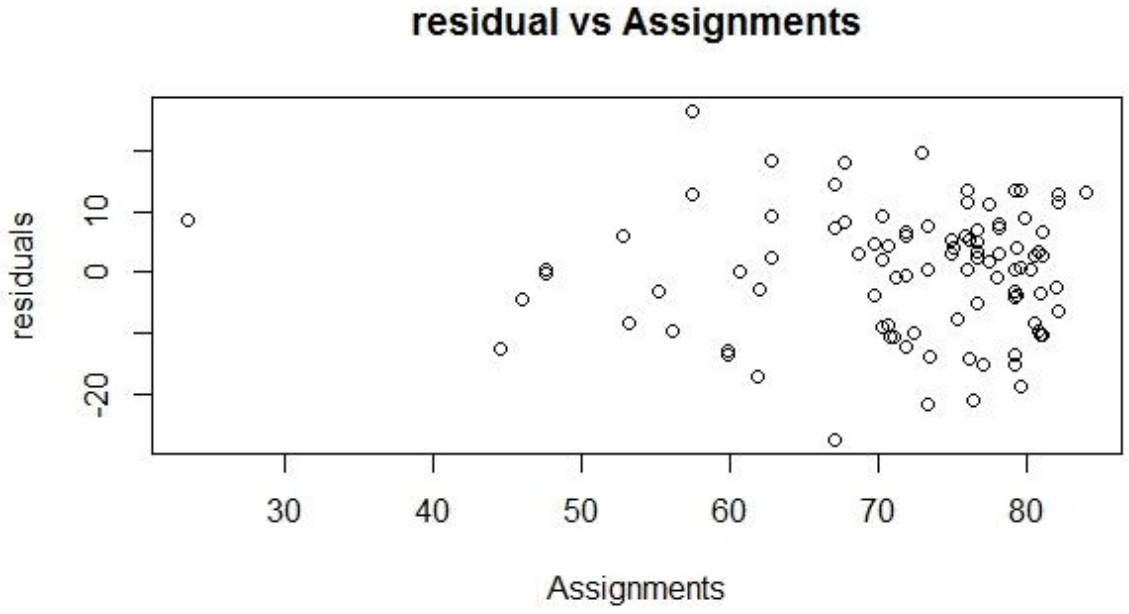


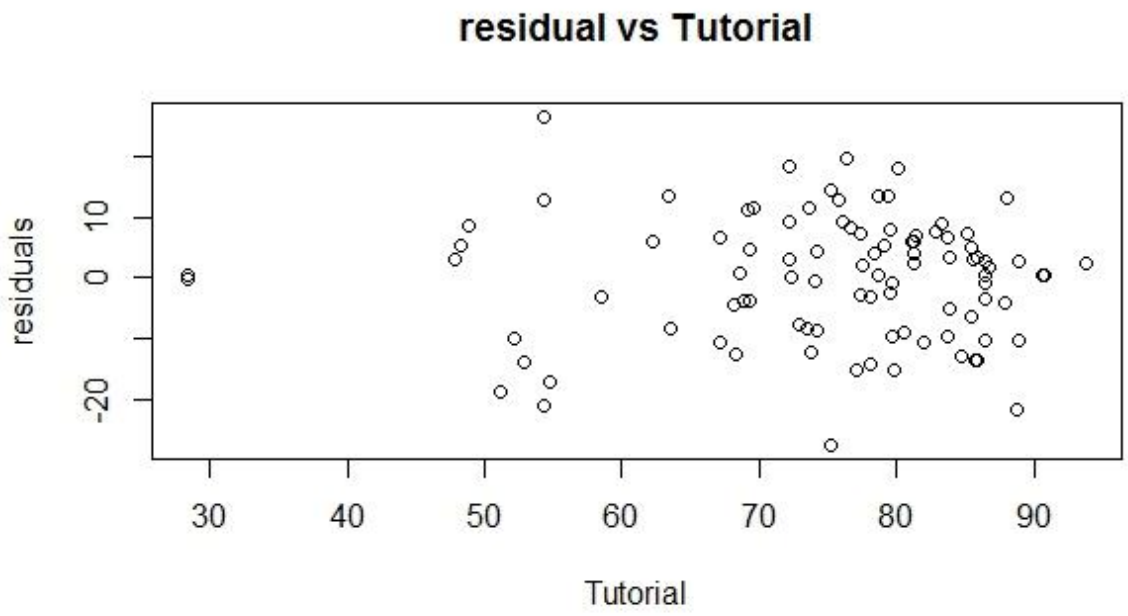
where the two horizontal lines are the bounds for the interval $(-1.5 * \sigma, 1.5 * \sigma)$, where σ is the residual standard error.

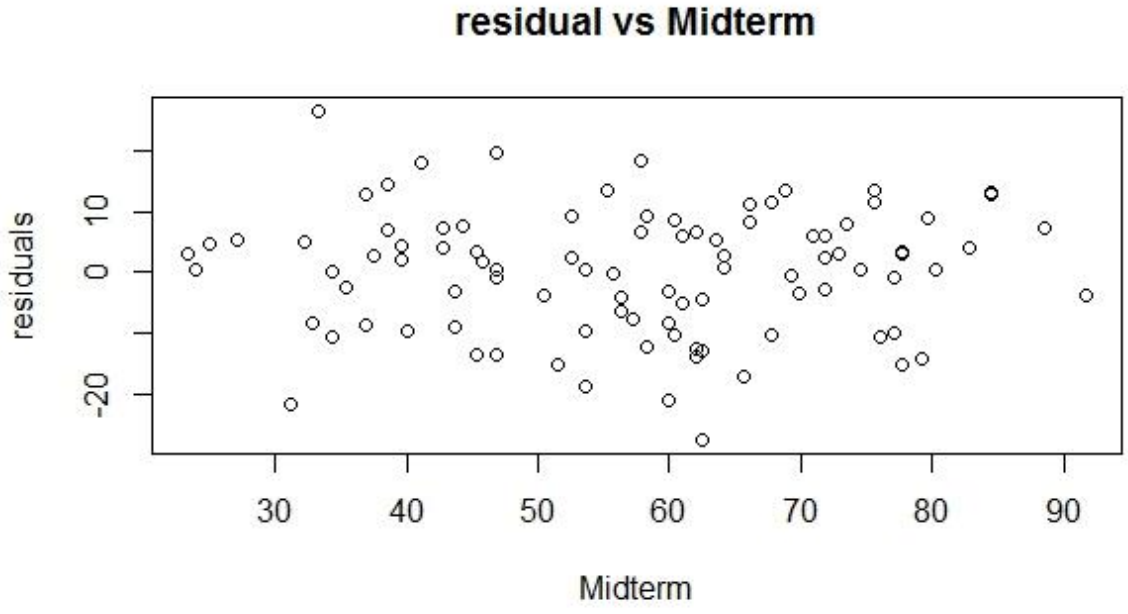
Here is the residuals plot for each predictors;

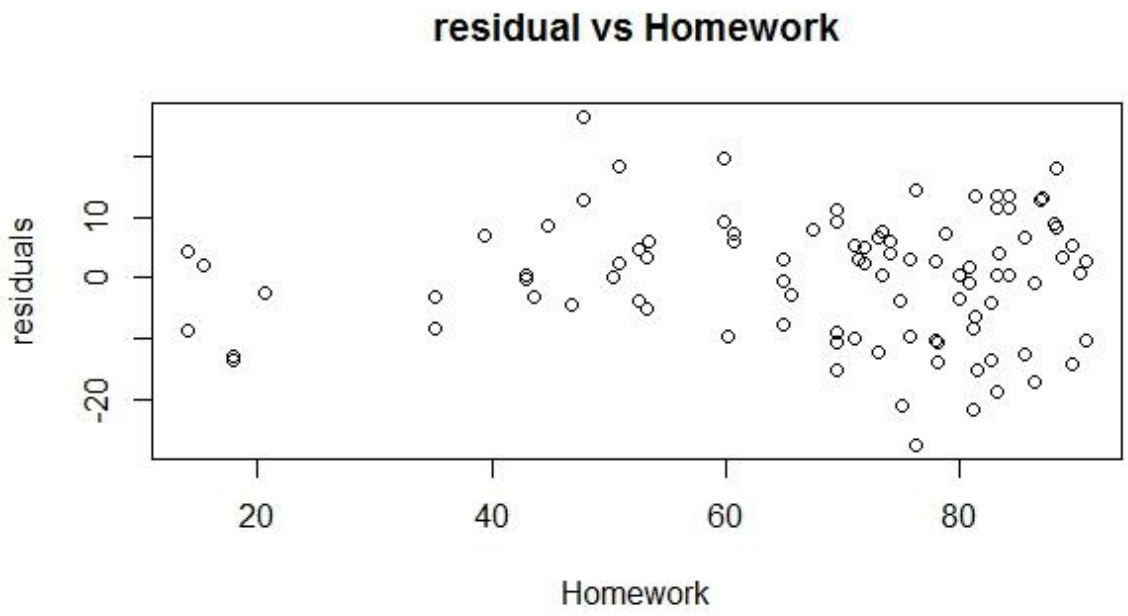
As seen there most of the points clustered to one side for each of the variables, while in together the residual plot shows no pattern. It may indicate one could try weighted least squares for multiple linear regression and perform the analysis to see the effect of residuals over the whole range.


residual vs School GPA
residuals
-20
0
10
4
5
6
7
8
School GPA


residual vs Assignments
residuals
-20
0
10
30
40
50
60
70
80
Assignments


residual vs Tutorial
residuals
-20
0
10
30
40
50
60
70
80
90
Tutorial


residual vs Midterm
residuals
-20
0
10
30
40
50
60
70
80
90
Midterm



Conclusion

The following conclusions can be made from the multivariate analysis done on the dataset and by comparing the ideal situation (simulation) with the practical dataset.

- Residual standard error can be expected to be close to the actual errors assumed in the linear model.
- The concentration of points in residual plot should be close to the line “residual = 0”, and the concentration should decrease as one goes away from the line (effectively showing the presence of normally distributed errors).
- Only bivariate analysis fails to give a correct picture of the relationship between two variables as they may be correlated to some other variables (confounding factors). As in the practical dataset, we see that by simple bivariate analysis School GPA was an insignificant factor to predict Final scores, but as it turned out from the partial correlation that it actually has more contribution that it was expected to give. Rather, it was surprising that the other factors that were seemed somehow significant to predict the final scores, have low partial correlation with it.
- School GPA is found to have negative correlation with midterm marks and Homework grades. Obviously, this is not the case as their relationship is being influenced by other several confounding variables.
- School GPA is not a good measure of final scores or their performance, as it is given in terms of whole numbers (or positive integers), hence the close performance in high school gives as lot of students exactly same GPA and this causes unnatural variation along the same line.
- As seen from the regression plots for individual variables, a weighted least square approach to find the regression equation would be more accurate in explaining the model.