Antecedents to Academic Performance of High school Students: An analytical approach

Anannya Chatterjee, Kashif Saeed

Abstract

Many practical studies are carried out to investigate factors affecting high school student's performance. The technology revolution and digitization of 21st century has brought in new challenges for the teachers, parents, and mentors to understand and quantify multiple factors that influence a student's conduct in various subjects which becomes a deciding factor for their grades/marks in the final year of graduation. The focus of this research is to evaluate antecedents to student performance, measured by high school percentage, while considering factors like gender, parental qualification, test preparation coursework, and attendance. The research is based on student profile developed on the bases of information and data collected through assessment from a group of high school students. Educational institutions can use the results of this study for policy formulation, whereas parents can use them as a guide for their children's success. Moreover, we argue that students can come across some factors about which they themselves are unaware of (hidden factors) that influence their performance.

1. Introduction

> "Dropping out of high school is no longer an option. It's not just quitting on yourself; it's quitting on your country, and this country needs and values the talents of every American."
> - Barack Obama, 44th President of the United States

Fifty years ago, the nation could afford to lose large numbers of students before graduation because high school dropouts could still land well-paying jobs and support their families. But times have changed. Today, jobs that require relatively little education are increasingly done by machines or shipped overseas, and individuals who fail to earn a high school diploma are at a great disadvantage.

Based on recent data from the U.S. Bureau of Labor Statistics, high school dropouts are nearly three times more likely to be unemployed than college graduates. Even when employed, high school dropouts earn about $8,000 a year less than high school graduates and approximately $26,500 a year less than college graduates, based on calculations by the Alliance for Excellent Education. According to research by the Georgetown Center on Education and the Workforce, 65 percent of all jobs in 2020 will require some form of education after high school.

As research by the Alliance for Excellent Education has found, high school dropouts also influence a community's economic, social, and civic health. While data on the education attainment of inmates is sparse, a 2004 survey of inmates in state and federal correctional facilities by the U.S. Bureau of Justice Statistics finds that 67 percent of inmates in America's state prisons, 56 percent of federal inmates, and 69 percent of inmates in local jails are high school dropouts. High school dropouts are also generally less healthy, require more medical care, and die earlier. In fact, cutting the number of

high school dropouts in half nationally would save $7.3 billion in annual Medicaid spending, according to an Alliance report on the subject.

Academic failure is one of the crucial causes that increases the high school dropout rates. Struggling in school daily is the biggest reason most students choose to drop out of high school. According to the Anne E. Casey Foundation by America's Promise, children who are not reading proficiently by fourth grade are four times more likely to quit high school than their peers. Systematic reforms can only be established by carrying out extensive research on the high school performance.

This vast investigation has striven to identify antecedents to academic performance of high school students and their impacts on it. Farooq, M.S., Chaudhry, A.H., Shafiq, M., Behranu, G. (2011) emphasize the factors affecting students' quality of academic performance. Caviglia Harris and Jill L. (2004) summarize how attendance policies and class size effects student performance. Luca, S. (2006) describes the effects of attendance on academic performance of a student. Musarat Azhar, Sundus Nadeem, Faqiha Naz, Fozia Perveen & Ayesha Sameen (2014) abridge the impact of parental education and socio – economic status on student's academic achievements. Yadav, V.S., Ansari, M.R., Savant, P.A. (1999) synopsize a critical analysis of study habits and academic achievement of high school / college students. Therefore, educational institutions may be expected to place considerable emphasis on the attendance, parental education background, study habits of the students for monitoring their academic performances.

Despite considerable research on identifying antecedents to academic performance of high school students, attention to the relationship between test preparation practices and academic performance of high school graduates has been very scant. Moreover, results that suggest that differences exist in the cognitive-motivational functioning of boys and girls in the academic environment, with the girls have a more adaptive approach to learning tasks are exiguous.

This research is conducted to address the gap highlighted above and to investigate the impact of the various antecedents spanning through test preparation practices, gender, attendance, parental education, and daily study hours on the high schooler's graduation percentages. This paper also seeks to determine the most influential contributor towards the graduation percentage.

The remainder of the paper is organized as follows. The next section introduces the related literature informing this study. The research model and research methods are described in the subsequent two sections. These are followed by the results and discussion sections, and the conclusion section wraps up the paper.

2. Literature Review

Students are the pioneers of every nation. The transition from high school to college is the most significant phase in their lives. Research about the students throughout their high school to unlock key issues that lead to success or failure of post-middle school education have the potential to not only help the national discussion regarding post-middle school success, but in this age of spiraling college tuition costs, this will assist families in making more thoughtful decisions about how best to prepare their children for productive and successful educational experiences after high school graduation.

Factors and determinants of academic success or achievement of high school graduates have been a subject of ongoing debate among researchers.

Farooq, M.S., Chaudhry, A.H., Shafiq, M., Behranu, G. (2011) elaborated on the factors affecting students' quality of academic performance. A survey was conducted by them using a questionnaire for information gathering about different factors relating to academic performance of students. The results of their study revealed that socioeconomic status (SES) and parents' education have a significant effect on students' overall academic achievement.

Caviglia-Harris, Jill L. (2004) summarized how Attendance rates and academic achievements are related. He explained how attendance policies and class size impact student performance.
He finds that a student that missed class was 9-14 percent more likely to respond incorrectly to a related exam question. This result suggests that student motivation, captured by attendance rates, actually impact grades.

Luca, S. (2006) investigated the effects of attendance on academic performance. The author presents new evidence on the effects of attendance on academic performance.
He used a large panel data set for introductory microeconomics students to explicitly take into account the effect of unobservable factors correlated with attendance, such as ability, effort, and motivation. Panel estimators indicate that attendance has a significant impact on academic performance.

Musarat Azhar, Sundus Nadeem, Faqiha Naz, Fozia Perveen & Ayesha Sameen (2014) condensed the impact of parental education and socio – economic status on student's academic achievements. Parental education and Socio-Economic factors are of vital importance in effecting students' educational achievements also. They are like backbone in providing financial and mental confidence to students. Explicit difference can be observed between those students who belong to different financial status and different parental educational level. They stated that relation of parent's education to their children's academic performances rests upon quite specific beliefs & behaviors. Parent's educational qualification is linked with their language competence, which has a significant influence in which parent's communicate with their children. Consequently, parental education does have a major influence on children's academic achievements.

Yadav, V.S., Ansari, M.R., Savant, P.A. (1999) synopsize a critical analysis of study habits and academic achievement of high school / college students. Students should have a habit of asking questions which facilitates better understanding of the concepts. Also, they should strongly follow time management skills that will allow them to utilize the time duration efficiently. These factors have a positive relationship with high school graduation performance.

Finally, Ayodele, C.S., Adebiyi, D.R. (2013) describe study habits as influential factor of academic performance. The descriptive analysis revealed that self-concept was very strong determinant of study habit, so also was method of study, family background, socio-economic status, peer group and course of study. The outcome of this study immensely helped high schooler and undergraduates to improve their study habits skills and in turn facilitate students' performance.
Improvement in students' academic performance will therefore lead to national development as qualitative manpower will be produced. Also, the school, government and all stakeholders should make facilities and materials that facilitate studying available to students.

3. Research methods

3.1 Data collection

Data used in the analysis is collected from a secondary dataset named “High School Graduation Performance” in Kaggle website. This dataset includes the various factors that might have an impact on the final graduation percentage of the high schoolers (grade 12). There are all total 9 variables (Gender, Race, Parental_Education, TestPrep_Course, Special_Coaching, Attendance, DailyStudy_Hours and Result) and 298 observations. As part of this research project, this dataset is introspected for determining the factors that possibly influence the performance of a student. Result is considered as the dependent variable and all others are considered independent variables.

3.2 Data preparation and cleansing

The data is introspected by running descriptive statistics (mean, median, standard deviation, variance) on the content present in the numerical columns of Attendance, DailyStudy_Hours and Result.

	SL No.	Attendance	DailyStudy_Hours	Result
Mean	149.500000	71.323232	13.314815	70.380872
Median	149.500000	70.000000	14.000000	71.000000
Std. dev	86.169407	9.391084	2.368012	12.676634
Variance	7425.166667	88.192465	5.607482	160.697040

The mode is described as:

Gender	Race	Parental_Education	TestPrep_Course	Special_Coaching	Attendance	DailyStudy_Hours	Result
Male	African American	High school	None	Yes	59.0	15.0	74.0

The dataset contained missing/NULL values in certain rows. Those are imputed using the mean/median/mode values of the respective columns as appropriate. Python is used in Google colab platform for achieving this null handling in the dataset. There are no NULL values after imputation.

Then, the various categorical variables (Gender, Race, Parental_Education, TestPrep_Course and Special_Coaching) are transformed. Parental_Higher_Education’ column is added to the dataset. The various Education levels in the Parental_Education column are grouped into Yes and No. The Parental Education levels that are either ‘Associate Degree’, ‘Bachelor Degree’ or ‘Master Degree’ are categorized as ‘Yes’. The Parental Education levels that are either ‘Attended college’ or ‘High school’

are categorized as ‘No’. And the categorical values for the ‘Parent_Higher_Education’ variable are transformed to numerical values. Another new column PHEdu_c is created where ‘Yes’ is transformed to 1 and ‘No’ to 0.

Gender variable is transformed to a new numerical variable Gender_c where ‘Male’ is transformed to 1 and ‘Female’ to 0.

TestPrep_Course variable is transformed to a new numerical variable TPC_c where ‘Completed’ is transformed to 1 and ‘None’ to 0.

Special_Coaching variable is transformed to a new numerical variable SC_c where ‘Yes’ is transformed to 1 and ‘No’ to 0.

Since the performance of the Asian population is analyzed here, hence new column Race_c is created and ‘Asian’ is transformed to 1 and all other categories in Race variable are transformed to 0.

Now the cleaned data is exported from Google colab environment and made ready for further Regression analysis in MS Excel tools.

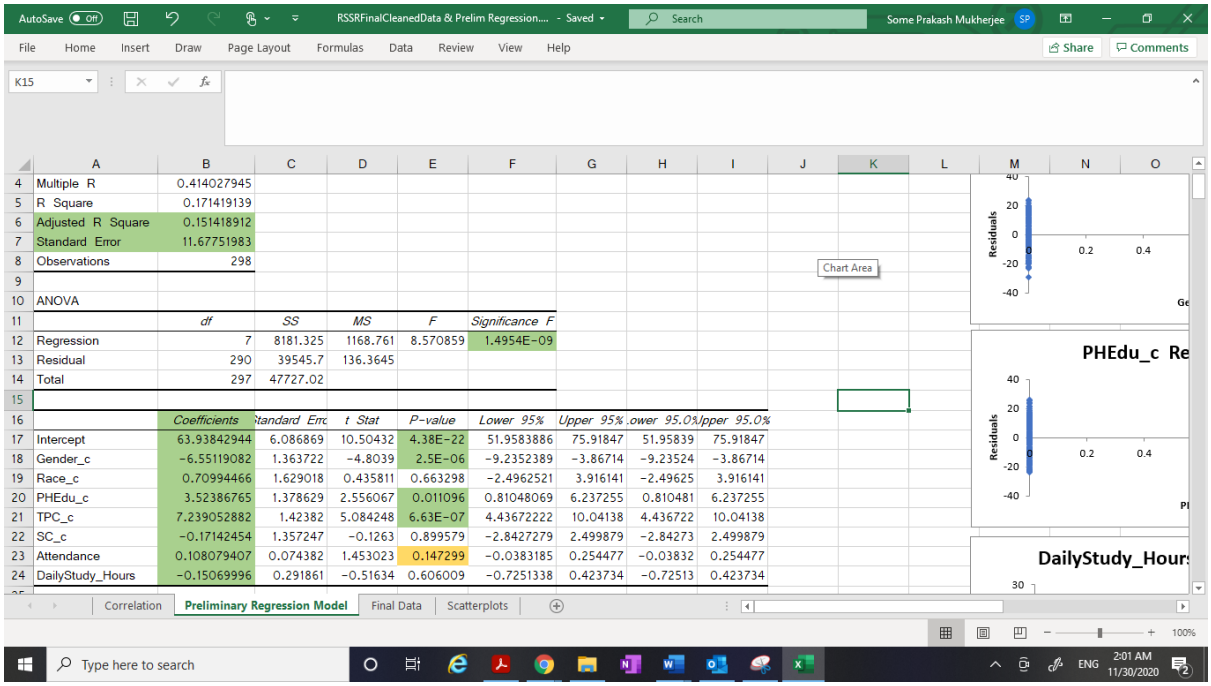
3.3 Pre-Regression testing and Preliminary Regression Analysis

As part of pre regression testing, correlation test among all the independent variables (Gender_c, Race_c, PHEdu_c, TPC_c, SC_c, Attendance and DailyStudy_Hours) is carried out for checking possible multicollinearity.

	A	B	C	D	E	F	G	H
1		*Gender_c*	*Race_c*	*PHEdu_c*	*TPC_c*	*SC_c*	*Attendance*	*DailyStudy_Hours*
2	Gender_c	1						
3	Race_c	-0.06193	1					
4	PHEdu_c	-0.05425	-0.01068	1				
5	TPC_c	0.053454	-0.02144	-0.00818	1			
6	SC_c	0.011553	-0.01728	-0.00531	0.034427	1		
7	Attendance	-0.04816	0.076149	-0.00323	0.131339	0.052099	1	
8	DailyStudy_Hours	0.034205	0.006496	0.055141	0.053243	0.020508	0.171473	1
9								
10								

As the value of 0.8 or above indicates possible multicollinearity, so from the findings, it is stated that Multicollinearity does not exist between the independent variables here.

Now, as part of running a preliminary regression model, all the independent variables are considered.



Multiple R	0.414027945	
R Square	0.171419139	
Adjusted R Square	0.151418912	
Standard Error	11.67751983	
Observations	298	

ANOVA

	df	SS	MS	F	Significance F
Regression	7	8181.325	1168.761	8.570859	1.4954E-09
Residual	290	39545.7	136.3645		
Total	297	47727.02			

	Coefficients	Standard Erro	t Stat	P-value	Lower 95%	Upper 95%	Lower 95.0%	Upper 95.0%
Intercept	63.93842944	6.086869	10.50432	4.38E-22	51.9583886	75.91847	51.95839	75.91847
Gender_c	-6.55119082	1.363722	-4.8039	2.5E-06	-9.2352389	-3.86714	-9.23524	-3.86714
Race_c	0.70994466	1.629018	0.435811	0.663298	-2.4962521	3.916141	-2.49625	3.916141
PHEdu_c	3.52386765	1.378629	2.556067	0.011096	0.81048069	6.237255	0.810481	6.237255
TPC_c	7.239052882	1.42382	5.084248	6.63E-07	4.43672222	10.04138	4.436722	10.04138
SC_c	-0.17142454	1.357247	-0.1263	0.899579	-2.8427279	2.499879	-2.84273	2.499879
Attendance	0.108079407	0.074382	1.453023	0.147299	-0.0383185	0.254477	-0.03832	0.254477
DailyStudy_Hours	-0.15069996	0.291861	-0.51634	0.606009	-0.7251338	0.423734	-0.72513	0.423734

Key points that are considered after this are:

This preliminary regression is significant as the ‘p’ value for the overall F test is 1.4954E-09 which is much lower than 0.05

Race_c, SC_c and DailyStudy_Hours are statistically insignificant variables because those variables are with p values much higher than 0.05 (significance level, Alpha = 0.05)

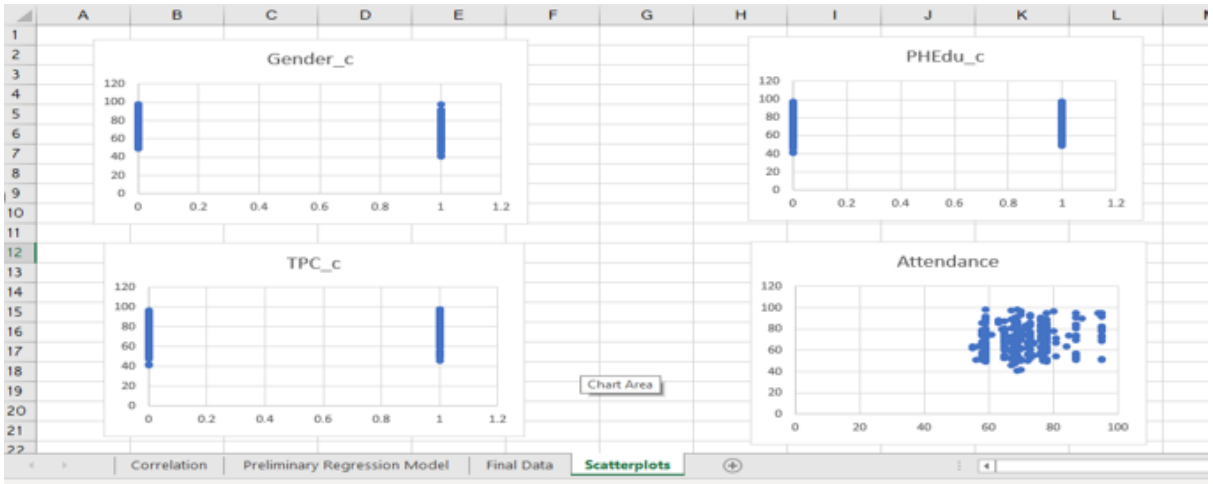
Gender_c, PHEdu_c and TPC_c are practically and statistically significant variables because those variables are with p values much lower than 0.05 and they are having higher beta weights.

Intercept is statistically significant because it is with p values much lower than 0.05, but do not have practical significance.

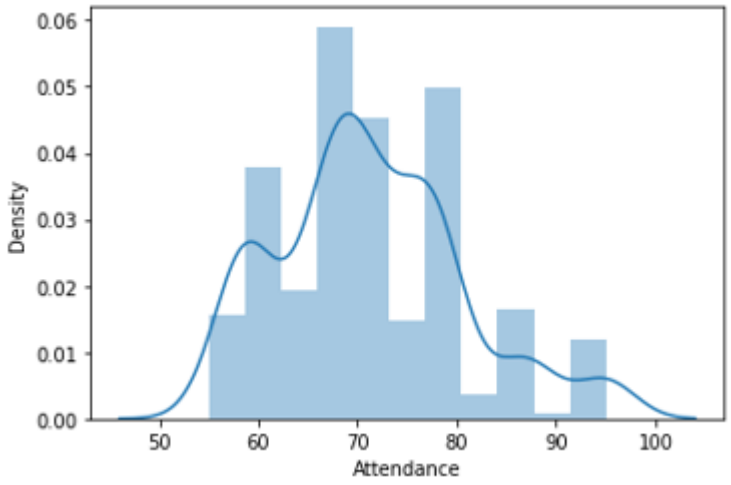
The p value of the Attendance variable is not exceptionally higher than 0.05. So, although it may seem statistically insignificant, but it can have practical significance.

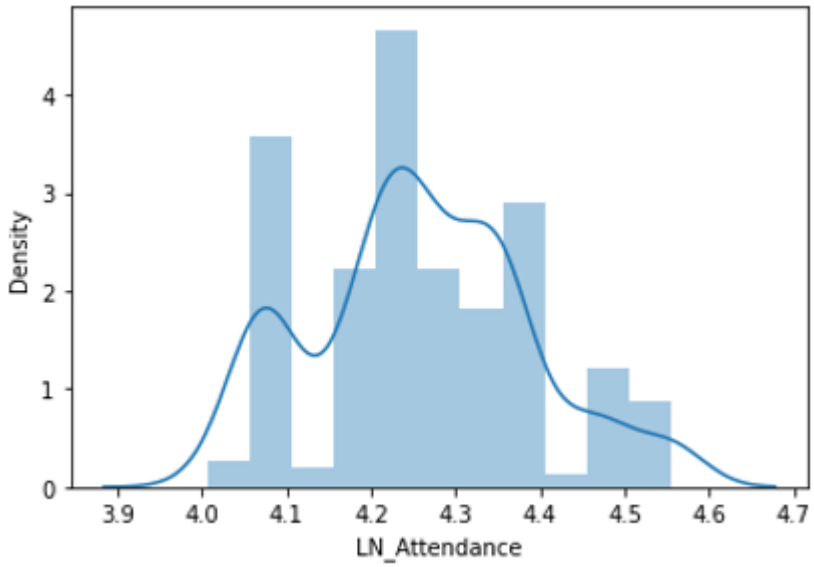
Hence Race_c, SC_c and DailyStudy_Hours variables are removed from being considered for creating further regression models.

The scatterplots between the Result (dependent variable) with individual independent variables respectively are completed in Excel.



Also, the distribution plot of Attendance variable seemed to be skewed towards the right side and hence Logarithmic transformation of the Attendance variable is done. This new variable added is LN_Attendance which will be taken into consideration in further analysis.

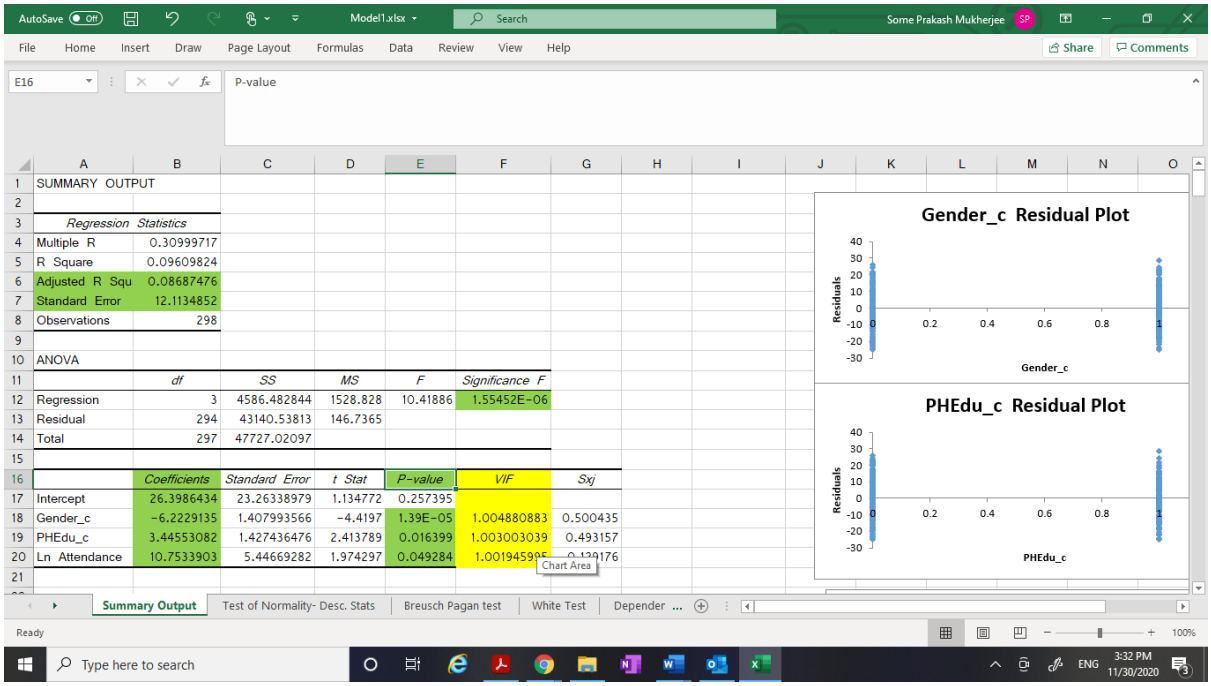




3.4 Creation of 3 different Regression Models

Model#1:

Here Gender, PHEdu_c, Ln_Attendance are considered as the independent variables and Result as the dependent variable and linear Regression model is as follows.



	A	B	C	D	E	F	G
1	SUMMARY OUTPUT						
2							
3	*Regression Statistics*						
4	Multiple R	0.30999717					
5	R Square	0.09609824					
6	Adjusted R Squ	0.08687476					
7	Standard Error	12.1134852					
8	Observations	298					
9							
10	ANOVA						
11		*df*	*SS*	*MS*	*F*	*Significance F*	
12	Regression	3	4586.482844	1528.828	10.41886	1.55452E-06	
13	Residual	294	43140.53813	146.7365			
14	Total	297	47727.02097				
15							
16		*Coefficients*	*Standard Error*	*t Stat*	*P-value*	*VIF*	*Sxj*
17	Intercept	26.3986434	23.26338979	1.134772	0.257395		
18	Gender_c	-6.2229135	1.407993566	-4.4197	1.39E-05	1.004880883	0.500435
19	PHEdu_c	3.44553082	1.427436476	2.413789	0.016399	1.003003039	0.493157
20	Ln Attendance	10.7533903	5.44669282	1.974297	0.049284	1.0019459	0.499176
21							

Key findings from Model#1 are:

a. R^2 = 0.096. The model#1 explains 9.6% of the total variance of the dependent variable.
b. The Adjusted R Square is '0.0868'
c. The Standard Error is 12.11
d. The 'p' value for the overall F test is 1.5545183E-06 which is significant.
e. The various coefficients below explain about practical significance:

	Coefficients
Intercept	26.3986434
Gender_c	-6.2229135
PHEdu_c	3.44553082
Ln Attendance	10.7533903

f. The 'p' value for the partial slopes that are used in partial slope's 't' test are:

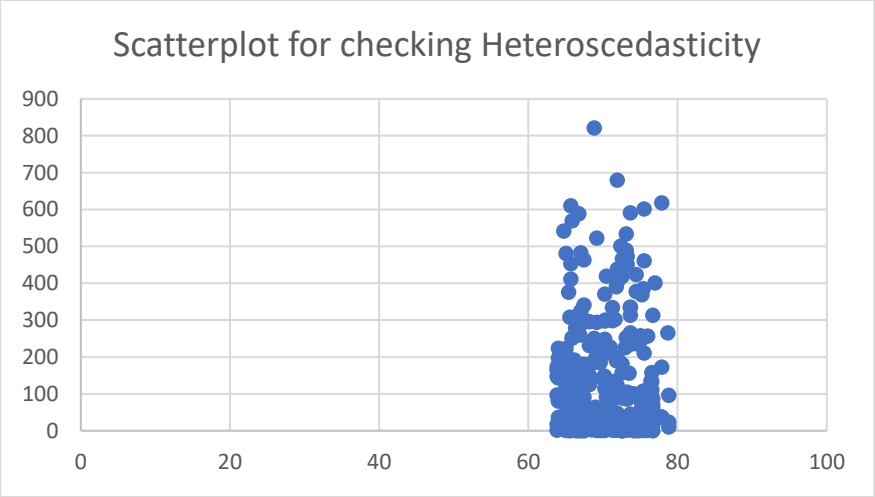
P-value
0.257395
1.39E-05
0.016399
0.049284

g. The p-value for Intercept is more than 0.05 making it statistically insignificant. But p-value for Gender_c, PHEdu_c and Ln Attendance are lesser than 0.05 (Alpha = 0.05) which makes these significant.

h. From all these findings, the predicted equation for Model#1 is:
Y(hat) = B0 + B1*X1 + B2*X2 +B3*Ln(X3)
Where B0 is the intercept and B1, B2 and B3 are the various respective coefficients. After filling the equation with the actual coefficients values the equation becomes:
Result(hat) = 26.398 - 6.222*Gender_c + 3.445*PHEdu_c + 10.7533*Ln_Attendance

i. The 'test for multicollinearity' is conducted by finding out the VIF (Variance Inflation Factor) for the independent variables that were considered in Model#1 in Excel. As we know that values of VIF that exceed 10 are often regarded as indicating multicollinearity. So, it clearly indicates that there does not exist any multicollinearity in Model 1.

j. 'Test of heteroscedasticity' is performed for Model# 1. Please find the attached snapshot. The Scatter plot, for the 'Predicted Result' in x-axis and 'Residual^2' in the y-axis shows that there is not much increase in Residual^2 as the Result value was going higher. This suggests that Heteroscedasticity might not be there in model 1. Here goes the excel snapshot for this.



k. For more confirmation Brusch Pagan test is applied to the Model# 1. This is essentially where the 'Residual^2' is regressed on all the independent variables that are there in the model 1.

Model#2:

Here PHEdu_c, TPC_c, and Gender_c are considered as the independent variables and Result as the dependent variable and linear Regression model is as follows.

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.405342673
R Square	0.164302683
Adjusted R Sqı	0.155775159
Standard Error	11.6475076
Observations	298

ANOVA

	df	SS	MS	F	Significance F
Regression	3	7841.67759	2613.893	19.26734	1.96574E-11
Residual	294	39885.3434	135.6644		
Total	297	47727.021			

	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%	VIF	Sxj
Intercept	69.74405452	1.24095856	56.20176	2.7E-159	67.30176655	72.18634		
PHEdu_c	3.46404573	1.37251061	2.523875	0.012134	0.762854681	6.165237	1.00298	0.493157
TPC_c	7.465944226	1.40565878	5.311349	2.15E-07	4.699515407	10.23237	1.002894	0.481507
Gender_c	-6.726631276	1.35444145	-4.96635	1.16E-06	-9.39226101	-4.061	1.005787	0.500435

Key findings from Model#2 are:

a. R^2 = 0.164. The model#1 explains 16.4% of the total variance of the dependent variable.
b. The Adjusted R Square is '0.1557'
c. The Standard Error is 11.64
d. The 'p' value for the overall F test is 1.96574E-11 which is much significant.
e. The various coefficients below explain about practical significance:

	Coefficients
Intercept	69.74405452
PHEdu_c	3.46404573
TPC_c	7.465944226
Gender_c	-6.726631276

f. The 'p' value for the partial slopes that are used in partial slope's 't' test are:

P-value
2.7E-159
0.012134
2.15E-07
1.16E-06

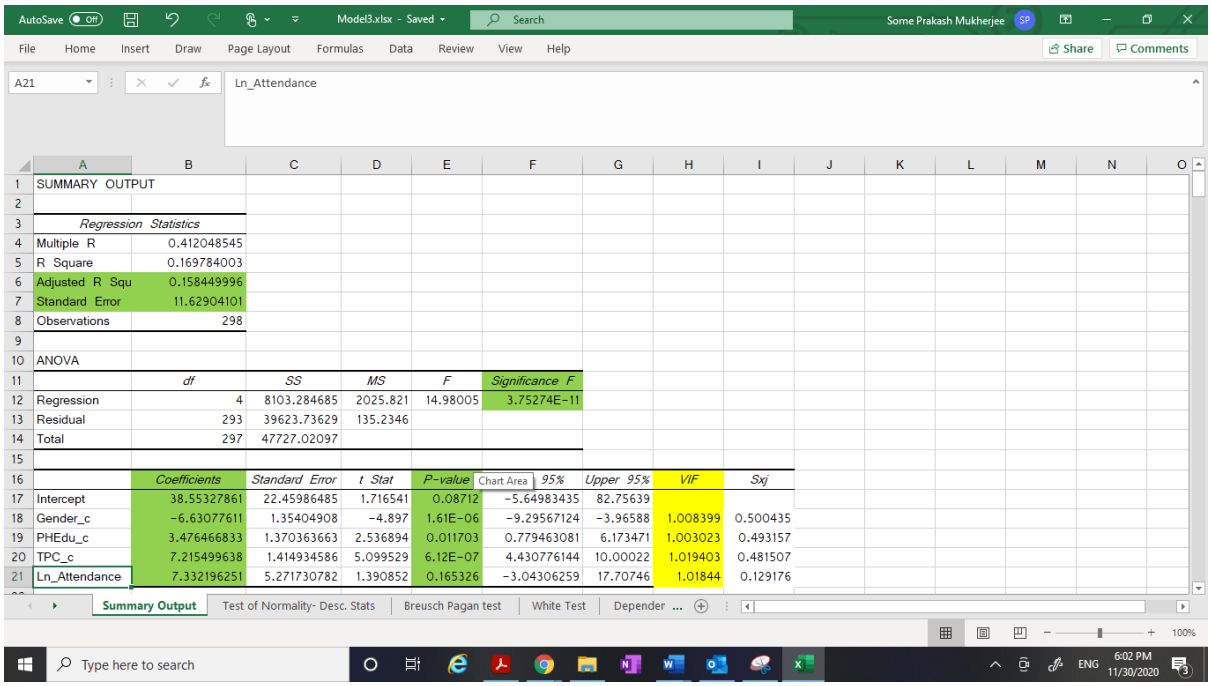
g. Here the p-value for Intercept, PHEdu_c, TPC_c and Gender_c and are much lesser than 0.05 (Alpha = 0.05) which make these highly significant.

h. From all these findings, the predicted equation for Model#2.
Result(hat) = 69.744 + 3.464*PHEdu_c + 7.465*TPC_c - 6.726*Gender_c

i. The 'test for multicollinearity' and the test of heteroscedasticity' are performed for Model# 2 which confirms that there does not exist any multicollinearity in Model 2 and the model is homoscedastic.

Model#3:

Here Gender_c, PHEdu_c, TPC_c , and Ln_Attendance are considered as the independent variables and Result as the dependent variable and linear Regression model is as follows.



SUMMARY OUTPUT									
Regression Statistics									
Multiple R	0.412048545								
R Square	0.169784003								
Adjusted R Squ	0.158449996								
Standard Error	11.62904101								
Observations	298								
ANOVA									
	df	*SS*	*MS*	*F*	*Significance F*				
Regression	4	8103.284685	2025.821	14.98005	3.75274E-11				
Residual	293	39623.73629	135.2346						
Total	297	47727.02097							
	Coefficients	*Standard Error*	*t Stat*	*P-value*	*95%*	*Upper 95%*	*VIF*	*Sxj*	
Intercept	38.55327861	22.45986485	1.716541	0.08712	-5.64983435	82.75639			
Gender_c	-6.63077611	1.35404908	-4.897	1.61E-06	-9.29567124	-3.96588	1.008399	0.500435	
PHEdu_c	3.476466833	1.370363663	2.536894	0.011703	0.779463081	6.173471	1.003023	0.493157	
TPC_c	7.215499638	1.414934586	5.099529	6.12E-07	4.430776144	10.00022	1.019403	0.481507	
Ln_Attendance	7.332196251	5.271730782	1.390852	0.165326	-3.04306259	17.70746	1.01844	0.129176	

Key findings from Model#3 are:

a. R^2 = 0.169. The model#1 explains 16.9% of the total variance of the dependent variable.
b. The Adjusted R Square is ‘0.158’
c. The Standard Error is 11.629
d. The ‘p’ value for the overall F test is 3.75274E-11 which is much significant.
e. The various coefficients below explain about practical significance:

	Coefficients
Intercept	38.55327861
Gender_c	-6.63077611
PHEdu_c	3.476466833
TPC_c	7.215499638
Ln_Attendance	7.332196251

f. The ‘p’ value for the partial slopes that are used in partial slope’s ‘t’ test are:

P-value
0.08712
1.61E-06
0.011703
6.12E-07
0.165326

g. Here the p-value for Intercept and Ln_Attendance are more than 0.05 making it statistically insignificant. But p-value for Gender_c, PHEdu_c and TPC_c is much lesser than 0.05 (Alpha = 0.05) which makes these highly significant.

h. From all these findings, the predicted equation for Model#3 is:
Y(hat) = B0 + B1*X1 + B2*X2 +B3*X3 + B4*Ln(X4)
Where B0 is the intercept and B1, B2, B3 and B4 are the various respective coefficients. The actual coefficients values from the Regression generates the equation as: Result(hat) = 38.553 - 6.630*Gender_c + 3.476*PHEdu_c + 7.215*TPC_c + 7.332*Ln_Attendance

i. The 'test for multicollinearity' and the test of heteroscedasticity' are performed for Model# 3 which confirms that there does not exist any multicollinearity in Model 3 and the model is homoscedastic.

3.5 Best Model detection

The values of "Adjusted R Square "and "Standard Error" among all the three different models are compared. Adjusted R-squared increases only when independent variable is significant and affects dependent variable. Since the "Adjusted R Square" is highest in Model 3 (compared to model 1 and 2) and "Standard Error" is least in "Model 3" compared to Model 1 and 2, so it is predicted that Model 3 is the best among all the three models.

	Adjusted R Square	Standard Error
Model 1	0.086874755	12.11348522
Model 2	0.155775159	11.6475076
Model 3	0.158449996	11.62904101

4 Discussion

4.1 Findings from the best model.

1. There was total 298 observations.
2. The NULL hypothesis or H0 for regression states that:

B0 = B1 = B2 =B3=B4 =0. It states that there is no relationship between the output and input variables which can be described with any model.
The alternate hypothesis H(a) states that at least one of the 'B's is not equal to zero and there exists some form of relationship between the output and input variables and hence, the model exists that explains this relationship.
Now for proving that there exists some relationship between the variables, we had to reject Null hypothesis (H0).

3. ANOVA is the Analysis of Variance of the dependent variable. It is especially important because it tells whether the model is useful or not. It is breaking the variance into what is explained by the model and what is not explained by the model.

   The total variance of the 'Result' variable is 47727.02 and out of that, Model 3 explains 8103.28 and the unexplained part is 39623.7 given by Residual.

a. R^2 = SSR/SST = 0.169

b. Degree of Freedom is n-1 = 298 – 1 = 297

c. Mean Square (MS) = SS/ df

d. The probability of significance of F is 3.75274e-11. Alpha is the level of significance. At 95% confidence interval, value of alpha is 0.05.

So here the p-value of F test is much lesser than 0.05 which means that we are in the rejection region, and we are confident to reject H0.
This concludes that the model 3 is better than the H0 and there is at least one coefficient which is not 0.

4. The Standard error of the regression Model3 is 11.629

5. Adjusted R-squared increases only when independent variable is significant and affects the dependent variable. Here the Adjusted R-squared is 0.158

6. The coefficients section helped in describing about the individual variables also. We have seen that:

a. For Intercept (B0): the value ranges from -5.649 to 82.75 which means it contains 0 in the range. Also, the p-value for t test here is 0.08 which is more than 0.05 which is the threshold value. So, intercept is statistically not significant.
b. For Gender_c(B1): the value ranges from -9.29 to -3.9 which means it does not include 0 in the range. Also, the p-value for t test here is 1.61e-06 which is much less than 0.05 which is the threshold value. So, Gender_c is statistically significant.
c. For PHEdu_c (B3): the value ranges from 0.77 to 6.17 which means it does not contain 0 in the range. Also, the p-value for t test here is 0.011 which is very lesser than 0.05 which is the threshold value. So, PHEdu_c is statistically significant.
d. For TPC_c (B3): the value ranges from 4.43 to 10.00 which means it does not contain 0 in the range. Also, the p-value for t test here is 6.12e-07 which is very lesser that 0.05 which is the threshold value. So, TPC_c is statistically significant.
e. For Ln_Attendance (B4): the value ranges from -3.0 to 17.70 which means it includes 0 in the range. Also, the p-value for t test here is 0.16 which is higher than 0.05 which is the threshold value. So, Ln_Attendance is statistically not significant.

From the Model# 3, it can be stated that H0 is rejected and Model# 3 is better than H0. Hence, it is proved that that there is a relationship between the variables that is explained by Model 3.

The model#3 states the relationship among the variables as follows:

Result(hat) = 38.553 - 6.630*Gender_c + 3.476*PHEdu_c + 7.215*TPC_c + 7.332*Ln_Attendance

Where ‘Result’ denotes the Graduation Result Percentage, ‘Gender_c’ tells about the gender, ‘PHEdu_c’ denotes whether parent is highly educated or not, ‘TPC_c’ denotes whether Test Preparation Course was completed or not and ‘Attendance’ denotes the percentage of attendance.So, parent higher education, and completion of test preparation courses indeed had a positive impact on the student’s graduation performance.

4.2 Interpretation of Result

As compared to Female student, the Graduation Result percentage decreases by 6.63 for Male student, while everything else is constant.

As compared to the student whose parents are not highly qualified, the Graduation Result percentage increases by 3.476 for the students whose parents are highly qualified, while everything else is constant.

As compared to the students who did not complete Test preparation courses, the Graduation Result percentage increases by 7.215 for the students who completed Test preparation courses, while everything else is constant.

This Model #3 is a semi log model as we have one logarithmic form of independent variable (Attendance) and our dependent variable is not in logarithmic form. As we know that whenever we are using Logarithmic form for any of the variable in the regression model, the coefficient (B value) tells the percentage change for that variable. That is the increase or decrease is measured as B%. Here Attendance variable had been logged and we had Ln(Attendance) in our Model.

So, the model speaks the below statement about the Attendance variable: For 1 percent increase in Attendance percentage, Graduation Result percentage increases by 0.07332 (i.e 7.332/100), while everything else is constant.

4.3 Social Implication

Based on the data analysis, it is quite imperative that teenager female students remain more focused on studies compared to male counterparts who get easily distracted on a wide array of things ranging from sports, outdoor activities, gaming, music and other extra-curricular activities. The thread which ties together the attention in students easily loosens up in male student compared to female and hence we see a drop in their performance during graduation.
Similarly, parents who are highly educated and earned higher degrees have the natural tendency to pass on the value of education to their off springs. Additionally, highly qualified parents tend to provide their kids with an atmosphere to learn and grow and they set it all by example which makes it easy for kids to interpret.
Also, the test preparatory courses provide an edge in the success of the students through giving them a look and feel of the actual exam pattern and environment. Students who complete those get equipped with the techniques required to set themselves apart from rest of the students in high school graduation exam.

Last but not the least, attending class sessions helps students to stay on track, understand expectations, foster important peer social interactions, and generally promote a sense of connectedness. Increasingly, attendance is being understood as a precursor and leading indicator for student success. Attendance improves performance.

4.4 Forecast future values

The estimated equation could be better understood by using that in forecasting a future value of the Result variable. In this case, the estimated equation is:

Result(hat) = 38.553 - 6.630*Gender_c + 3.476*PHEdu_c + 7.215*TPC_c + 7.332*Ln_Attendance

Suppose high schooler Tom who maintains a 90% attendance record has reasons to believe that Test preparation Courses hardly impacts his performance. His highly educated parents are really worried about his graduation results.
The regression equation can be used to predict the result percentage of Tom in both the scenarios (with and without completing Test preparation Courses).

Case 1: Tom considers NOT to complete Test Preparation Courses.

Plugging in the values of Gender_c = 1, PHEdu_c = 1, TPC_c = 0 and Attendance = 90 in the above equation, I get:

Result = 38.553 – 6.630*1 + 3.476*1 + 7.215*0 + 7.332*ln(90)
= 68.33

Case 2: Tom considers completing Test Preparation Courses.

Plugging in the values of Gender_c = 1, PHEdu_c = 1, TPC_c = 1 and Attendance = 90 in the above equation, I get:

Result = 38.553 – 6.630*1 + 3.476*1 + 7.215*1 + 7.332*ln(90)
= 75.55

So, the regression equation could really help in estimating the result of students that is impacted by multiple variables.

4.5 Limitations

Impact of Social desirability biasness on the regression analysis:
Social-desirability bias is a type of response bias that is the tendency of survey respondents to answer questions in a manner that will be viewed favorably by others. It can take the form of over-reporting "good behavior" or undesirable behavior. The dataset that I worked on to determine validity of relationships between the Student Performance and its various deciding factors (using Regression Analysis) is social desirability biased as respondents often hide their true attitudes - to impress the researcher or interviewer or to preserve one’s self-esteem. That is why the coefficient of variation (also known as R2) that is used to determine how closely a regression model “fits” or explains the relationship between all the independent variables (Parent higher education, Gender, Test preparation course and Attendance) and the dependent variable

(student Performance) has emerged low. Moreover, some other factors like inherent talent, genetic aptitude also determine the performance of a student that are very unlikely determined by any mathematical or statistical models.

5. Conclusion

This research examined the antecedents to student performance, measured by high school percentage, while considering factors like gender, parental qualification, test preparation coursework, and attendance into consideration. The dataset determined the validity of relationships between the Student Performance and its various deciding factors using Multiple Linear Regression Analysis. Educational institutions, parents, and students themselves would definitely find this research highly beneficial in framing the next course of action in their respective domains.

References

Farooq, M.S., Chaudhry, A.H., Shafiq, M., Behranu, G. (2011). Factors affecting students' quality of academic performance: A case of secondary school level. Journal of Quality and Technology Management, 7(2), 1–14.

Google Scholar

Caviglia-Harris, Jill L., Attendance Rates and Academic Achievement: Do Attendance Policies and Class Size Effects Impact Student Performance? (September 2004). Available at SSRN: https://ssrn.com/abstract=605462 or http://dx.doi.org/10.2139/ssrn.605462

Luca, S. (2006). The effects of attendance on academic performance: Panel data evidence for introductory microeconomics. The Journal of Economic Education, 37(3), 251–266.

Google Scholar | Crossref | ISI

Musarat Azhar, Sundus Nadeem, Faqiha Naz, Fozia Perveen, Ayesha Sameen (2014) Impact of parental education and socio-economic status on academic achievements of university students

http://www.idpublications.org/wp-content/uploads/2014/11/IMPACT.pdf

Yadav, V.S., Ansari, M.R., Savant, P.A. (1999). A critical analysis of study habits and academic achievement of college students. Karnataka Journal of Agricultural Sciences, 13(4), 914–918.

Google Scholar

Ayodele, C.S., Adebiyi, D.R. (2013). Study habits as influence of academic performance of university undergraduates in Nigeria. Research Journal in Organizational Psychology and Educational Studies, 2(3), 72–75.

Google Scholar

Dornbusch, S. M., Ritter, P. L., Mont-Reynaud, R., & Chen, Z. (1990). Family Decision Making and Academic Performance in a Diverse High School Population. Journal of Adolescent Research, 5(2), 143–160. https://doi.org/10.1177/074355489052003