REGRESSION ANALYSIS AND EXPERIMENTAL DESIGN

INSTRUCTED BY DR. JAVED IQBAL

CONTENTS

INTRODUCTION

OVERVIEW

Modern society has put a lot of stress on youth to excel in their careers. Academic success is important because, it is strongly linked to the positive outcomes we value. Adults who are academically successful and with high levels of education are more likely to be employed, they have more employment opportunities than those with less education and earns higher salaries, they are less likely to engage in criminal activity, are more active as citizens and charitable volunteers and are healthier and happier. Thus, students who do well in school are better able to make the transition into adulthood and to achieve occupational and economic success. In their efforts to survive in such a competitive environment to attainable such level of position, there are considerable factors that affects a student's performance.

DATA SET INFORMATION:

This data approaches student achievement in secondary education. The data attributes include student grades, demographic, social and school related features) and it was collected by using school reports and questionnaires. This dataset is provided regarding the performance in a subject.

OBJECTIVES OF THE CURRENT STUDY

- To access students' performance
- To determine the specific predators associated that affects students' performance

OUTLINE OF THE STUDY

We were assigned the task to develop a statistical report for which we chose the topic "Students' Performance". We have picked this dataset from an online website named, UCI-Machine Learning Repository. There were all together 31 factors available that were claimed to affect a student's performance. Hence, we did few sample t-tests to check the significance that each factor holds and finalized 14 factors that we think are more impactful. After the completion of this process, we applied few statistical tests and obtained results.

COLLECTING THE DATA FOR REGRESSION

DATA ATTRIBUTES:

1. Gender - student's sex (binary: 'F' - female or 'M' - male)
2. age - student's age (numeric: from 15 to 22)
3. Familysize (binary: 'LE3' - less or equal to 3 or 'GT3' - greater than 3)
4. ParentsStatus - parent's cohabitation status (binary: ‘living together or ' apart’)
5. Mothereducation - mother's education (binary: ‘educated’ or ‘uneducated’)
6. Fathereducation - father's education (binary: ‘educated’ or ‘uneducated’)
7. Studytime - weekly study time (nominal: - <2 hours, 2-10hours, >10 hours)
8. activities - extra-curricular activities (binary: yes or no)
9. internet - Internet access at home (binary: yes or no)
10. familyrelation - quality of family relationships (nominal: from poor-moderate-Good)
11. health - current health status (nominal: from 1 -poor-moderate-Good)
12. absences number of school absences (numeric: from 0 to 93)

-These grades are related with the course subject

13. First Term grade (numeric: from 0 to 20)
14. Second Term grade (numeric: from 0 to 20)
15. Final grade (numeric: from 0 to 20, output target)

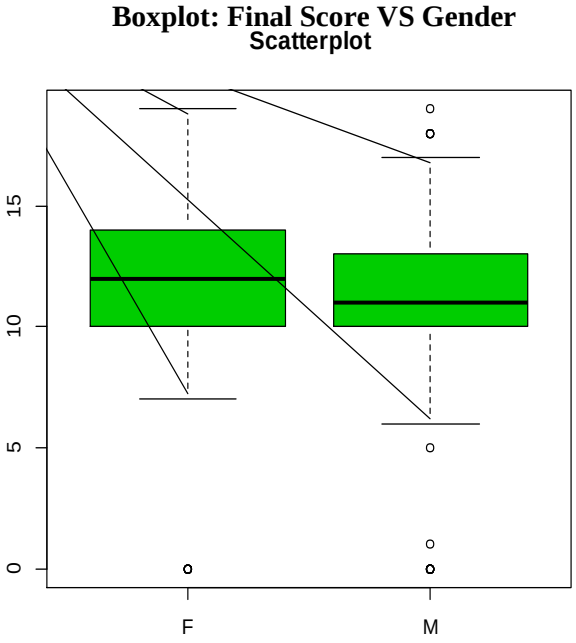
DATA INTERPRETATION

We have used R-statistical software to interpret the data. Our data contains sample size of 649 students and 14 factors to determine final marks of each student.

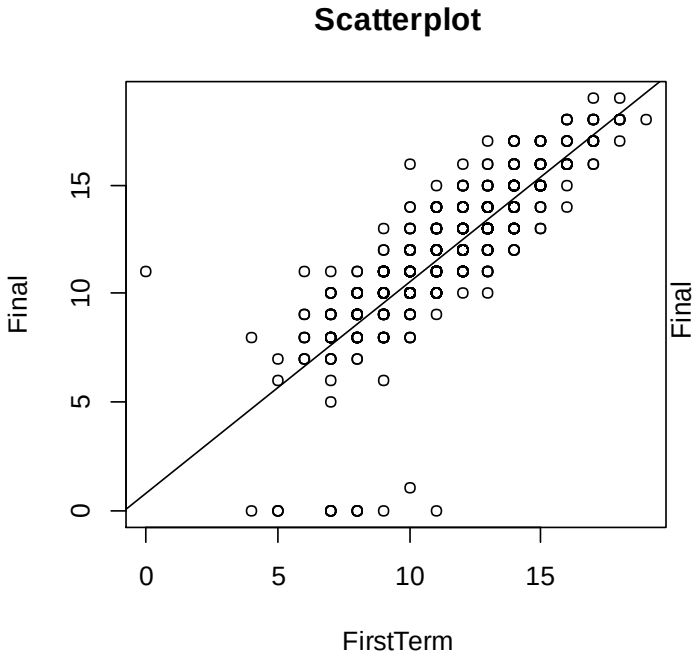
TYPE OF REGRESSION MODEL

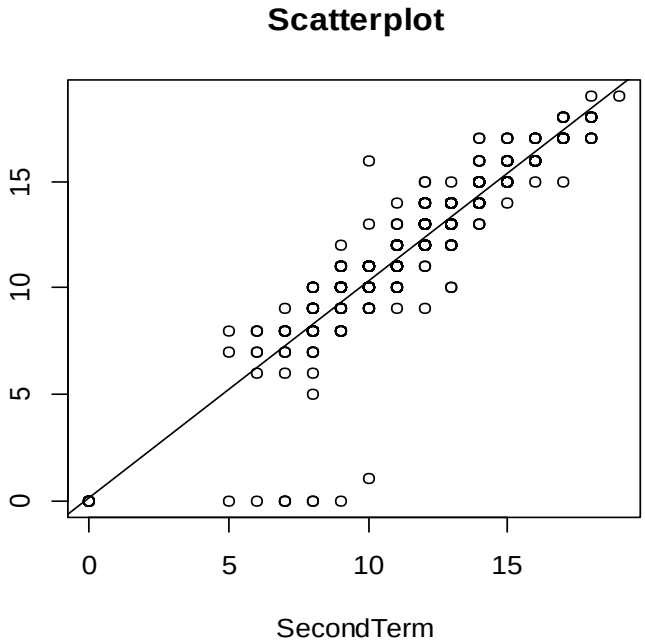
Observational (values of x’s(attributes) are uncontrolled.

DATA REPRESENTATION

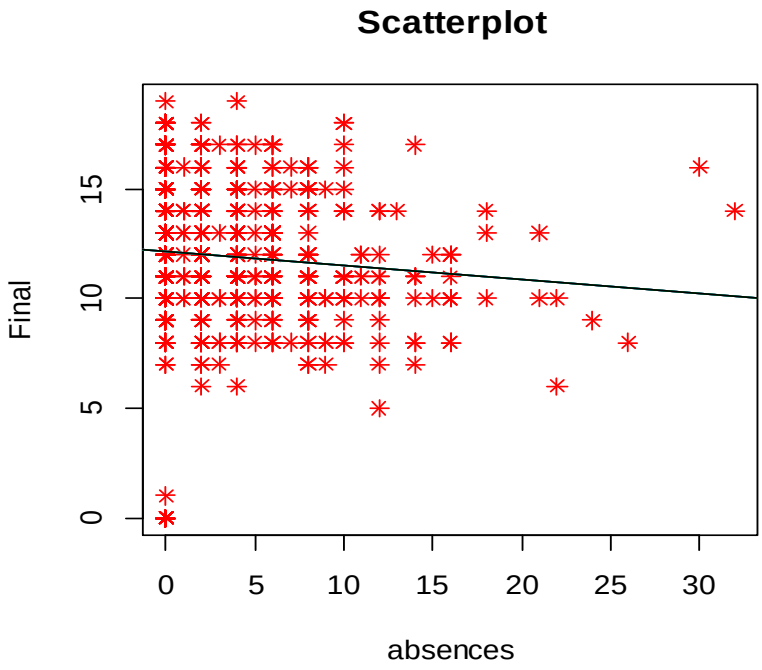


The above boxplot shows that the mean score of a female is higher than male. The average final score for female is 12.5 and for male is 11. It is noted that female score varies from 7 to 18, score 0 is an outlier. On the other hand, Male score varies from 6 to 16 and 0,1,17,18 are outliers.

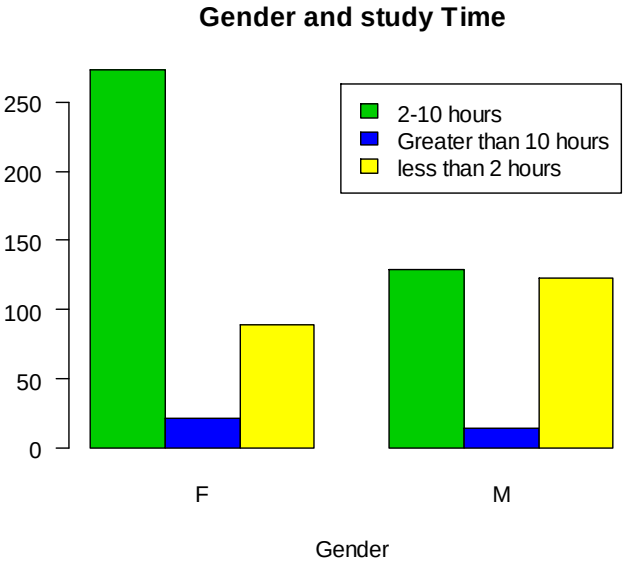




The two graphs generated by the software, shows the linear relationship between the Final Score-First term and Final-Second term score obtained by the students.



The plot on the left shows a decreasing linear relation, as the number of absences of a student increases, the marks scored decreases. However, as illustrated in the diagram, the relation seems weak.



Study time	Gender	
	Female	Male
2-10 hours	273	129
Greater than 10 hours	21	14
Less than 2 hours	89	123

The result depicts that, in todays' time very few students denote more than 10 hours to their studies this is merely because other factors like internet and friends/family gatherings have occupied some of their time. However, it is observed that majority of females (273) still serves approximately between 2-10 hours to their studies.

REGRESSION MODEL

Modeling a response:

a) **Multiple Regression**: A Model Relating E(y) for Qualitative Independent Variables

$E(y)=\beta 0+ \beta 1age+ \beta 2SecondTerm+ \beta 3FirstTerm+ \beta 4absences+ \beta 5Gender+ \beta 6$ familysize$+ \beta 7$ FatherEducation$+ \beta 8$MotherEducation$+ \beta 9$studytime1$+\beta 10$studytime2$+ \beta 11$internet$+ \beta 12$familyrelation1$+ \beta 13$familyrelation2$+ \beta 14$health1$+ \beta 15$health2$+ \beta 16$ParentsStatus$+ \beta 17$activities $+ \varepsilon$

Where:

y=Final exam marks of a student (score ranging from 0 to 20)

Dummy variables

Gender:	familysize	Father Education	Mother Education:	studytime1	studytime2	internet
1-If Male 0-If not	1-If less than 3 0-If not	1-If uneducated 0-If not	1-If uneducated 0-If not	1-If Greater than 10 hours 0-If not	1-If Less than 2 hours 0-If not	1-If yes 0-If not
Base case: Female	Base case: Greater than 3	Base case: Educated	Base case: Educated	Base case: 2-10 hours	Base case: 2-10 hours	Base Case: No

Familyrelation1	Familyrelation2	Health1	Health2	ParentsStatus	activities
1-If Moderate 0-If not	1-If Poor 0-If not	1-If Moderate 0-If not	1-If Poor 0-If not	1-If Not Together 0-If not	1-If yes 0-If not
Base case: Good	Base case: Good	Base case: Good	Base case: Good	Base case: Living together	Base Case: No

MINITAB FOR MULTIPLE REGRESSION

Coefficients	Estimate	Std.Error	t-value	Pr(>|t|)
(Intercept)	0.16011	0.78307	0.204	0.0838055
age	-0.01598	0.04225	-0.378	0.705391
SecondTerm	0.89067	0.03448	25.83	< 2e-16
FirstTerm	0.14449	0.03707	3.898	0.000107
absences	0.02102	0.01115	1.886	0.059811.
GenderM	-0.16615	0.10755	-1.545	0.122881
familysizeLE3	0.06446	0.1132	0.569	0.569245
FatherEducationUneducated	0.33412	0.49062	0.681	0.496105
ParentsStatusNotTogether	-0.08103	0.15945	0.508	0.00611518
studytimeGreaterThan10hours	0.06139	0.22635	-0.271	0.0786297
studytimeLessThan2Hours	-0.2204	0.11448	-1.925	0.054645.
internetyes	0.102	0.12098	0.843	0.03995
familyrelationModerate	-0.0325	0.24396	-0.133	0.00894071
familyrelationPoor	-0.04889	0.28328	-0.173	0.0086304
healthModerate	-0.03769	0.15525	0.243	0.00808239
healthpoor	-0.13852	0.14852	0.933	0.0351372
MotherEducationUneducated	0.30417	0.53087	0.573	0.056687.
activitiesyes	-0.0136	0.10167	-0.134	0.0893596

The regression Equation is

Final= 0.16011-0.01598age+ 0.89067SecondTerm+ 0.14449FirstTerm+ 0.02102absences-0.16615Gender+ 0.06446 familysize+ 0.33412FatherEducation+ 0.30417MotherEducation-0.06139studytime1-0.2204studytime2+ 0.102internet
-0.0325familyrelation1+-0.04889familyrelation2+0.03769health1+ 0.13852health2+ 0.08103ParentsStatus-0.0136activities + ε

COEFFICIENT INTERPRETATIONS

- **Coefficient of Age:** Keeping all other variables fixed, when age is increased by 1 year, the final score is decreased by 0.01598 on average.
- **Coefficient of Second Term:** Keeping all other variables fixed, when Second term score is increased by 1 mark, the final score is increased by 0.89067 on average.
- **Coefficient of First Term:** Keeping all other variables fixed, when First term score is increased by 1 mark, the final score is increased by 0.14449 on average.
- **Coefficient of Gender**: Keeping all other variables fixed, Male student will score 0.16615 less than the Female.
- **Coefficient of Family Size:** Keeping all other variables fixed, a student with a family of less than 3 persons will have 0.06446 marks more than a student belonging to a family of more than 3 person.
- **Coefficient of Father education**: Keeping all other variables fixed, a student whose father is uneducated will score 0.33412 more than a student whose father is educated.
- **Coefficient of Mother education:** Keeping all other variables fixed, a student whose mother is uneducated will score 0.30417 more than a student whose mother is educated.
- **Coefficient of StudyTime1:** Keeping all other variables fixed, a student who gives more than 10 hours to studies will score 0.06139 more than a student who puts in 2-10 hours.
- **Coefficient of StudyTime2:** Keeping all other variables fixed, a student who gives more than 10 hours to studies will score 0.224 less than a student who puts in 2-10 hours.
- **Coefficient of internet:** Keeping all other variables fixed, a student who uses internet will score 0.102 more than a student who doesn't use internet.
- **Coefficient of FamilyRelation1:** Keeping all other variables fixed, a student with a moderate family relation will score 0.0325 less than a student with a good family relation.
- **Coefficient of FamilyRelation2**: Keeping all other variables fixed, a student with a poor family relation will score 0.04889 less than a student with a good family relation.
- **Coefficient of health1:** Keeping all other variables fixed, a student with a moderate health will score 0.03769 less than a student with a good health.
- **Coefficient of health2:** Keeping all other variables fixed, a student with a poor health will score 0.13852 less than a student with a good health.
- **Coefficient of Parents Status:** Keeping all other variables fixed, a student whose parents are not together will score 0.08103 less than a student whose parents are together.
- **Coefficient of Activities**: Keeping all other variables fixed, a student who is indulged in extra-curricular activities will score 0.0136 marks less than a student who's not indulged.

MODEL SUMMARY

Residual standard error: 1.263 on 631 degrees of freedom	
Multiple R-squared: 0.8513	Adjusted R-squared: 0.8473

The value R^2 = .8513 is highlighted on the printout. This relatively high value of R^2 implies that using the independent variables, the model explains 85.13% of the total sample variation in Final marks, y. Thus, $R^2$is a sample statistic that tells how well the model fits the data and thereby represents a measure of the usefulness of the entire model. The value of Ra^2 is also highlighted. Note that Ra^2 = .84735, a value only slightly smaller than R^2.

ANOVA

F-statistic: 212.4 on 17 and 631 DF, p-value: < 2.2e-16

TESTING THE UTILITY OF A MODEL: THE ANALYSIS OF VARIANCE F-TEST

Null hypothesis: $\beta 1 = \beta 2 = \beta 3 = \cdots = \beta 17 = 0$

Alternate hypotheses: At least one of the coefficients is non-zero

The test statistic used to test this hypothesis is an F statistic, the statistical software calculates the F statistic):

Test statistic: $F = (SSyy - SSE)/k \; SSE/ [n - (k + 1)] = 212.4$

Conclusion: Since p-value< level of significance=0.05 we will reject null hypothesis and conclude that at least one of coefficient is non-zero. (conclusion based on p-value given in the table)

95 % CONFIDENCE INTERVAL

We are 95% confident for βk is (in the box below) using the formula:

$$\hat{\beta}_k \pm (t_{\alpha/2})\, s_{\hat{\beta}_k} \qquad \text{where } k=1,2,3\ldots17$$

lies in the following range,

coefficients	2.50%	97.50%
age	-0.09895	0.066991945
SecondTerm	0.822952	0.958380085
FirstTerm	0.071706	0.217283504
absences	-0.00087	0.042909565
GenderM	-0.37735	0.045049213
familysizeLE3	-0.15783	0.286755274
FatherEducationUneducated	-0.62932	1.297561524
ParentsStatusNotTogether	-0.23209	0.394146958
studytimeGreaterThan10hours	-0.50588	0.383088709
studytimeLessThan2Hours	-0.44521	0.004403302
internetyes	-0.13558	0.339569119
familyrelationModerate	-0.51157	0.446577405
familyrelationPoor	-0.60518	0.50740431
healthModerate	-0.26718	0.342565487
healthpoor	-0.15315	0.430180112
MotherEducationUneducated	-0.73832	1.346661605
activitiesyes	-0.21326	0.186051007

b) Interaction Model: An interaction model with qualitative predictors

E(y)= β0+ β1age+ β2SecondTerm+ β3FirstTerm+ β4absences+ β5Gender+ β6 familysize+ β7 FatherEducation+ β8MotherEducation+ β9studytime1+β10studytime2+ β11internet+ β12familyrelation1+ β13familyrelation2+ β14health1+ β15health2+ β16ParentsStatus+ β17activities+ β18Gender * SecondTerm + β19 Gender * FirstTerm + β20 ParentsStatus *FirstTerm + β21 activities * FirstTerm + β22 activities * SecondTerm + β23 ParentsStatus * SecondTerm+ ε

Coefficients	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	0.250205	0.826232	0.303	0.0076212
age	-0.006245	0.042224	-0.148	0.088247
SecondTerm	0.883679	0.063718	13.869	2.00E-16
FirstTerm	0.131307	0.068386	1.92	0.0553
absences	0.021141	0.011143	1.897	0.05825
GenderM	-1.298864	0.467828	-2.776	0.00566
familysizeLE3	0.055439	0.112922	0.491	0.062364
FatherEducationUneducated	0.240753	0.49142	0.49	0.062437
ParentsStatusNotTogether	-0.522992	0.657805	0.795	0.042688
studytimeGreaterThan10hours	-0.019055	0.227117	-0.084	0.093316
studytimeLessThan2Hours	-0.20536	0.114322	-1.796	0.07292
internetyes	0.086737	0.121575	0.713	0.047583
familyrelationModerate	-0.026526	0.244411	-0.109	0.0091361
familyrelationPoor	-0.153871	0.284607	-0.541	0.058894
healthModerate	-0.051579	0.155044	0.333	0.73949
healthpoor	-0.147266	0.149085	0.988	0.32363
MotherEducationUneducated	0.30307	0.529807	0.572	0.05675
activitiesyes	0.292553	0.44908	0.651	0.0515
SecondTerm:GenderM	-0.096614	0.075864	-1.274	0.20331
FirstTerm:GenderM	0.198868	0.081234	2.448	0.01464
FirstTerm:ParentsStatusNotTogether	-0.06045	0.096908	-0.624	0.053299
FirstTerm:activitiesyes	-0.09698	0.077236	-1.256	0.020972
SecondTerm:activitiesyes	0.069035	0.071309	0.968	0.033336
SecondTerm:ParentsStatusNotTogether	0.019075	0.087477	0.218	0.082745

E(y)= 0.250205-0.006245age+ 0.883679SecondTerm+ 0.131307FirstTerm+ 0.021141absences-1.298864Gender+0.055439 familysize+0.240753 FatherEducation+ 0.30307MotherEducation-0.019055studytime1-0.20536studytime2+0.086737internet+ -0.026526familyrelation1-0.153871familyrelation2-0.051579health1+ −0.147266health2-0.06045ParentsStatus+0.292553activities-0.096614Gender * SecondTerm + 0.198868Gender * FirstTerm + -0.06045ParentsStatus *FirstTerm + -0.09698 activities * FirstTerm + 0.069035activities * SecondTerm + 0.019075ParentsStatus * SecondTerm+ ε

ANALYSIS OF VARIANCE

Residual standard error: 1.257 on 625 degrees of freedom Multiple R-squared: 0.854, Adjusted R-squared: 0.8486 F-statistic: 158.9 on 23 and 625 DF, p-value: < 2.2e-16

TESTING THE OVERALL UTILITY OF MODEL USING THE GLOBAL F-TEST AT α = .05

The global F-test is used to test the null hypothesis

Null hypothesis: $\beta 1 = \beta 2 = \beta 3 = \cdots = \beta 23 = 0$

Alternate hypotheses: At least one of the coefficients is non-zero

The test statistic and p-value of the test (highlighted on the MINITAB printout) are F = 158.9 and p = 2.2e-16, respectively. Since α = .05 exceeds the p-value, there is sufficient evidence to conclude that the model fit is a statistically useful predictor of Final marks, y. Reject null hypothesis.

TEST FOR INTERACTION

Null hypothesis: $\beta 18 = \beta 19 = \beta 20 = \cdots = \beta 23 = 0$

Alternate hypotheses: At least one of the coefficients is non-zero (βi is not zero)

Where: i=18,19,20….,23

Test statistic: $F = (SSyy - SSE)/k / SSE/[n - (k + 1)] = [(1.263-1.257)/6]/[1.257/649-(23+1)]$

$$F=0.4972$$

F (6,625) at 0.05 level of significance= 2.10.

Conclusion: We have insufficient evidence at 0.05 level of significant to reject null hypothesis. Thus, we conclude that interaction is meaningless.

The value R^2 is reported to be 0. 854.This implies that using the independent variables, the model explains 85.4% of the total sample variation in Final marks, y.

CONCLUSION AND DISCUSSION

After observing both the models:

when we looked at the adjusted R-squared value for both the regression models, they are relatively close, but since the interaction test is insignificant we may declare model A to be more suitable to predict students' performance.

Model A:

Final= 0.16011-0.01598age+ 0.89067SecondTerm+ 0.14449FirstTerm+ 0.02102absences-0.16615Gender+ 0.06446 familysize+ 0.33412FatherEducation+ 0.30417MotherEducation-0.06139studytime1-0.2204studytime2+ 0.102internet -0.0325familyrelation1+-0.04889familyrelation2+0.03769health1+ 0.13852health2+ 0.08103ParentsStatus-0.0136activities + ε

REFERENCES

P. Cortez and A. Silva. Using Data Mining to Predict Secondary School Student Performance. In A. Brito and J. Teixeira Eds., Proceedings of 5th FUture BUsiness TEChnology Conference (FUBUTEC 2008) pp. 5-12, Porto, Portugal, April 2008, EUROSIS, ISBN 978-9077381-39-7.

Breiman L., 2001. Random Forests. Machine Learning, 45, no. 1, 5–32.

Eurostat, 2007. Early school-leavers. http://epp.eurostat.ec.europa.eu/.

Kotsiantis S.; Pierrakeas C.; and Pintelas P., 2004. Predicting Students' Performance in Distance Learning Using Machine Learning Techniques. Applied Artificial Intelligence (AAI), 18, no. 5, 411–426.