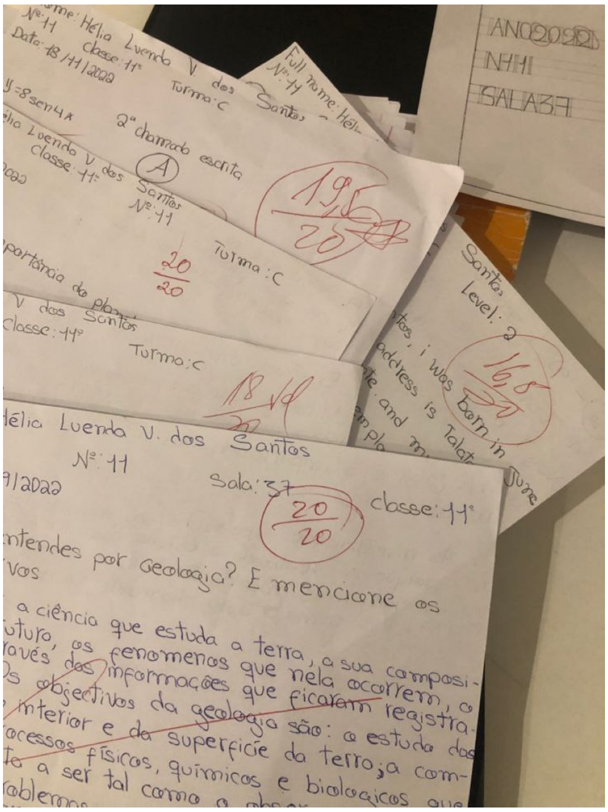


Students' performance model implementing multiple linear regression

Lamri Moahmed Yassine

Table of Contents

Contents

Abstract

Abstract

In this work, we present a data-set that concerns students in an academic Establishment. It contains most variables one can think of to effect the overall academic performance. We use the Linear regression Model in order to predict and draw conclusions, constructed by selecting appropriate variables. We use software results, consisting of statistical indexes, guiding us to the final model.

1 Introduction

Academic performance is a non-negligeable factor that determines a student success in his high-study project. Without doubt, there are many variables in a student's life, behaviour and envirement that largely contribute to his overall academic performance. In this peoject, we consider a few major ones, and we seek their effect on a students final mark, specefically language and mathematics, such as free time, study hours etc.. which will be presented in detail in the data section. We wish to establish and actual relation between the statistics related to these variables and the final marks.

2 The Data

This data approach student achievement in secondary education of two Portuguese schools. The data attributes include student grades, demographic, social and school related features) and it was collected by using school reports and questionnaires. Two datasets are provided regarding the performance in two distinct subjects: Mathematics (mat) and Portuguese language (por). We had chosen only numerical and seemingly significant features. That is : studytime, absences, age, freetime, goout, Father's education, health, work-day alchohol consumption, failures, G1 G2. G1, G2 being the respective marks for each trimestre, and G3 being the final score of the year. All data is of numerical values, each ranges in adequate intervals, which are presented in depth in the data associated file. The svc file was taken from a paper By P. Cortez, A. M. G. Silva. 2008, Published in Proceedings of 5th Annual Future Business Technology Conference1.This dataset is licensed under a Creative Commons Attribution 4.0 International (CC BY 4.0) license. 2

3 The Model

We chose in order to have predictions about our data, to use the multiple linear regression model. it is well mathematically established, and we have all sorts of tests that will confirm that indeed the model is able to predict of good accuracy. Our models consists of one dependnet variable Y_i and 4 independent variables (regressor) ${X_i}^j$. Where we want to explain Y_i, i.e G3, as the independent variable by the other ones, mentioned in order in the data section. In order to construct the linear regression model these variables need to verify the conditions of the model, all given by :

- (y_i) is a sequence of independent variables.
- $\forall i \in N$ we have : $y_i = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_{11} X_{11} + \epsilon_i$, supposed to be the distribution of Y knowing X. As vectors. Our goal as usually done in the linear regression model, is to estimate each of the unknown coefficients β_i.
- ϵ are normally distributed, of mean 0, and variance σ^2
- σ^2 the variance is constant for all errors ϵ_i

The model is the multiple linear regression model, using the usual least-square method, to estimate such coefficients and then provide predictions. i.e, the estimated value of Y, denoted $\hat{y}$.

4 Model Building and Model Fitness

As we said, we have to estimates the coefficients of the model, i.e β_i. We give the equation of the estimated y, denoted $\hat{y}$ by : $\hat{y} = \hat{\beta_0} + \hat{\beta_1} X_1 + \cdots + \hat{\beta_{11}} X_{11}$, where $\hat{\beta_i}$ are the estimated coefficients. The tests values are all provided in the software output section below. We remind that we need to split our data set into training values and ones that are to be predicted, so can concretly check wether ot not our model was valid.

4.1 Model for all regressor variables

We have first to select the significant features that we will use to train the regression model, since it isn't optimal to use all variables in order to obtain a fit model.

4.1.1 T-test

As we can see in the software output, the T-statistics is too high (172), indicating rejecting the null hypothesis. Indicates strong evidence against the null hypothesis, suggesting a highly significant model.

4.1.2 F-test

We will do the requested fitness tests on the whole set of features, then we test on a couple subsets that have desirable test results and take these in order to get trustable and valuable conclusions. In the next section ,we are going to observe the obtained indexed that determine the fitness of the model, and besed on it we will select the features that will eventually be significant, after cleaning the data-set and detecting outliers.

4.1.3 The R^2 and adj R^2 coefficients:

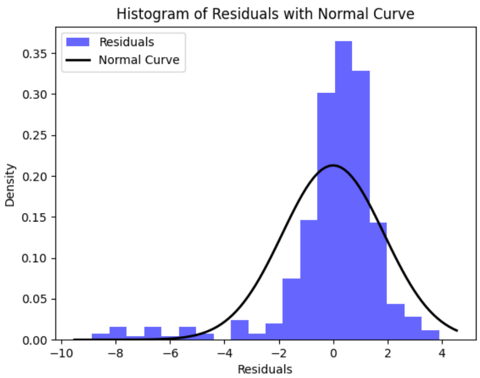
What we have, for 11 regressor variables, was $R^2 = 0.83$, which shows strong colinearity between Y and X, so its reasonable to use the multi-linear regression model, as far as we are going. As for the adjusted coefficient, its given by 0.827, wich is a very good value, since we have 11 regressor variables, that is relatively high, and we still have good colinearity conditions verified for our data-set.

4.2 Normality tests

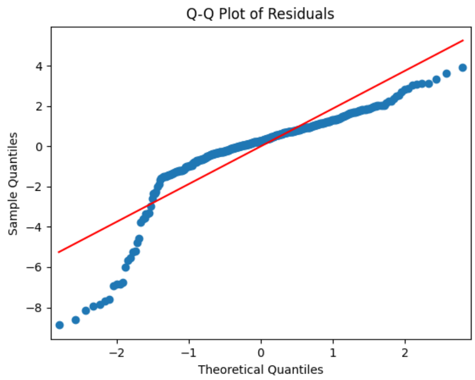
To check the normality of the errors $e = y - \hat{y}$ We first showcase the histogram of the residual errors approaching a normal curve, in the figure below. Even tho, it is still not quite adjusted. We will check it again after chosing the appropraite selected variables, and we will find a way to make it fit the best way possible. In order to do that, we speceify a section in the model building to detect outliers and remove them, to adjust variabce, since the mean approaches 0 with a very acceptable error. along with the QQ plot to check wether the quantiles of the errors match those of a normal distribution figure 1. However, by better cleaning our data and removing the outliers, we can get a better sample, that furthermore allignes with the normality hypothesis.:

4.3 p values

All p values are provided by the software. We notice some of them are high, some are very low, which counts in taking an appropriate subset for the final model. The lowest are those of G1, G2, failures and absences. We can show our values in the following plot, that visualizes the sigficance gicen by the t-test. The last bar is related to the constand added y the software, so shouldn't be taking into consideration :



(a) normality plot



(b) QQ plot

Figure 1: The plots to check the normality hypothesis

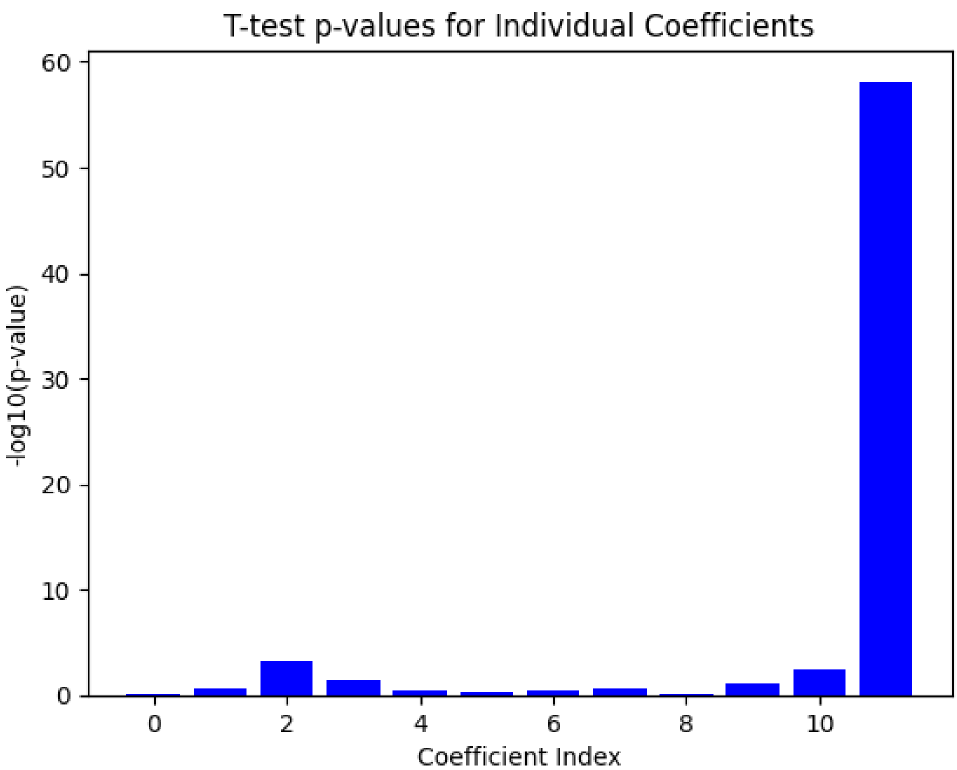


Figure 2: Caption

4.4 corelation matrix

The corelation matrix is also given by the software, allowing us to even analyse the linear relation between all given variables. We present the corelation matrix associated to all features as a heatmap. The matrix is of course symmetric, and we notice that the regressor variables X_i have very low linear corelation, which is a good indicator of indepnedence. We can also see that the variables that had low p values, have also strong corelation with $G3$, the regressed variable. Which makes our judegement of taking a candidate subset easier, based on all the previour indicators.

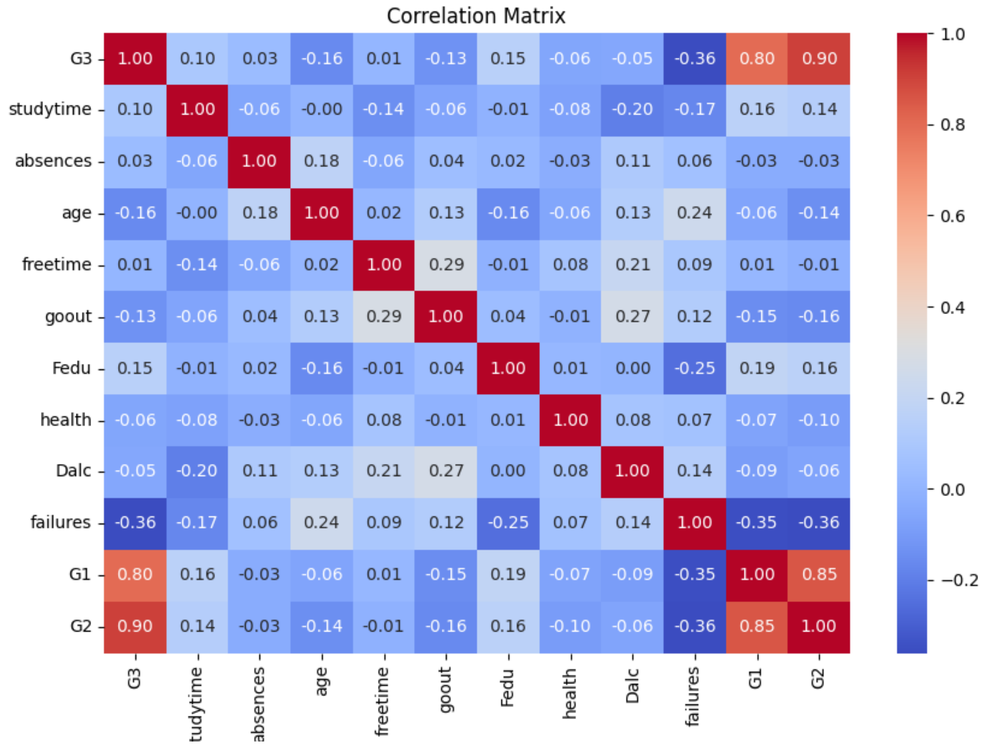


Figure 3: Heat map of the corelation matrix

4.5 VIF values

VIF values, as we know, . VIF measures how much the variance of the estimated regression coefficients are inflated due to multicollinearity in the predictors. Our VIF values obtained, as we can see in the software output, excluded from that of the constant (necessary to add as a new feature so statmodels operate correctly). Our favorable variables in the previous test happen to have very acceptable VIF values, the highest not exceeding

5, which is statistically sufficient to accept that the varibales arent very related. Which means, they provide us more information in the process of prediction, and add solidity to the model.

4.6 Model of different subsets of features

A provided document (html) shows diffirent R^2 values for all possible subsets taken from the regressor variables. We can see that the best ones match the variables that we are favouring up to this point. Hence, the others are automatically eliminated. Our final Model will contain these variables, that we will state, and we will provide a full check for the significance of the model, since other tests arent much useful for the primary data-set with all variables. Indeed, it is high likely that the overall significance check will give satisfying results.

4.7 Chosing an appropriate subset of features

According to all previous results, we can select an appropriate subset of regressor variables (features) .

5 Software Output

We start by stating the general summary of the software's output. As we can see, after plugging the data set in Python's statsmodel library :

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                     G3   R-squared:                       0.832
Model:                            OLS   Adj. R-squared:                  0.827
Method:                 Least Squares   F-statistic:                     172.8
Date:                Wed, 01 May 2024   Prob (F-statistic):          5.38e-141
Time:                        21:36:05   Log-Likelihood:                -808.55
No. Observations:                 395   AIC:                             1641.
Df Residuals:                     383   BIC:                             1689.
Df Model:                          11
```

Covariance Type: nonrobust

	coef	std err	t	P>|t|	[0.025	0.975]
const	0.6430	1.497	0.430	0.668	-2.300	3.586
studytime	-0.1528	0.120	-1.275	0.203	-0.388	0.083
absences	0.0427	0.012	3.471	0.001	0.019	0.067
age	-0.1666	0.081	-2.050	0.041	-0.326	-0.007
freetime	0.0907	0.103	0.879	0.380	-0.112	0.294
goout	0.0711	0.094	0.754	0.451	-0.114	0.257
Fedu	-0.0969	0.093	-1.040	0.299	-0.280	0.086
health	0.0813	0.070	1.159	0.247	-0.057	0.219
Dalc	-0.0245	0.116	-0.210	0.834	-0.254	0.205
failures	-0.2695	0.147	-1.835	0.067	-0.558	0.019
G1	0.1654	0.057	2.909	0.004	0.054	0.277
G2	0.9691	0.050	19.396	0.000	0.871	1.067

Omnibus:	204.531	Durbin-Watson:	1.866
Prob(Omnibus):	0.000	Jarque-Bera (JB):	1110.916
Skew:	-2.235	Prob(JB):	5.86e-242
Kurtosis:	9.894	Cond. No.	385.

Standard Errors of Coefficients: const 1.496741

studytime	0.119808
absences	0.012306
age	0.081247
freetime	0.103167
goout	0.094303
Fedu	0.093155
health	0.070171
Dalc	0.116499
failures	0.146838
G1	0.056856

```
G2           0.049966
dtype: float64
```

The VIF values :

```
        Features  VIF Factor
0          const  244.342230
1      studytime    1.099891
2       absences    1.055257
3            age    1.169356
4       freetime    1.155297
5          goout    1.199131
6           Fedu    1.117979
7         health    1.035472
8           Dalc    1.171527
9       failures    1.297239
10            G1    3.874501
11            G2    3.843050
```

```
F-statistic: 172.77659235531846
```

6 Presenting the final model

Considering all the above, we can safely pick our final moding that consititues of 4 regressor variables. G1, G2, absences and failures. We plug everything again into the software, while splitting our data-set into a training one to find the eventual model. We find all teh results compact in the following pages. Notice how all the indices are satisfying, in the way we presented earlier. We double check the significance of the model when predicting the actual results, in the rest of the data-set:

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                     G3   R-squared:                       0.862
```

Model:	OLS	Adj. R-squared:	0.861
Method:	Least Squares	F-statistic:	462.2
Date:	Sun, 12 May 2024	Prob (F-statistic):	1.13e-125
Time:	00:31:37	Log-Likelihood:	-580.91
No. Observations:	300	AIC:	1172.
Df Residuals:	295	BIC:	1190.
Df Model:	4		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-1.0771	0.385	-2.798	0.005	-1.835	-0.319
absences	0.0235	0.012	1.957	0.051	-0.000	0.047
failures	-0.2841	0.140	-2.026	0.044	-0.560	-0.008
G1	0.1084	0.055	1.984	0.048	0.001	0.216
G2	0.9731	0.047	20.922	0.000	0.882	1.065

Omnibus:	203.672	Durbin-Watson:	1.925
Prob(Omnibus):	0.000	Jarque-Bera (JB):	1971.699
Skew:	-2.741	Prob(JB):	0.00
Kurtosis:	14.300	Cond. No.	68.7

Correlation Matrix:

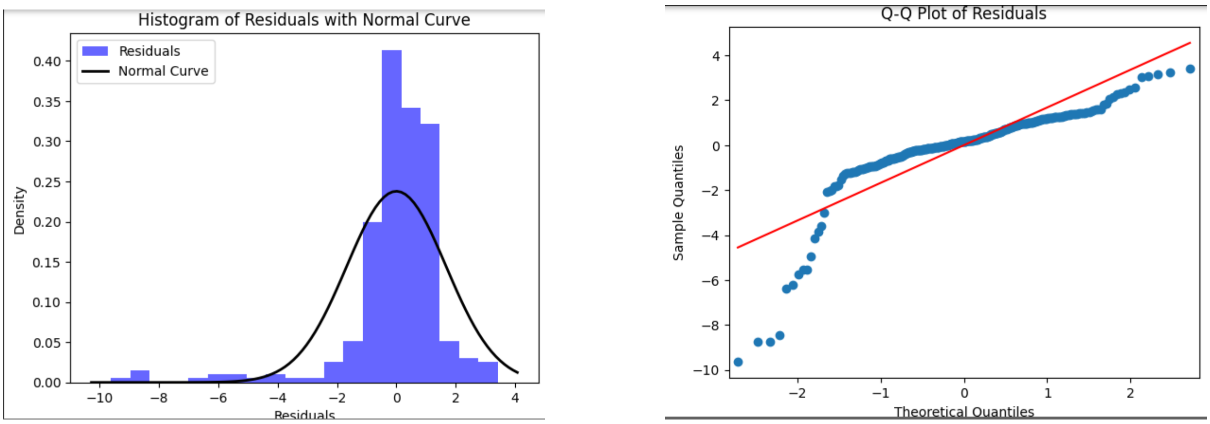
	G3	absences	failures	G1	G2
G3	1.000000	0.004344	-0.399095	0.805894	0.925362
absences	0.004344	1.000000	0.026455	-0.054242	-0.038455
failures	-0.399095	0.026455	1.000000	-0.404299	-0.380104
G1	0.805894	-0.054242	-0.404299	1.000000	0.842181
G2	0.925362	-0.038455	-0.380104	0.842181	1.000000

Features VIF Factor

```
0      const   15.533044
1   absences    1.003168
2   failures    1.203203
3         G1    3.545869
4         G2    3.462609
```

```
  F-statistic: 468.34849923637285
```

```
Standard Errors of Coefficients: const       0.384978
absences     0.012024
failures     0.140175
G1           0.054664
G2           0.046510
```



(a) normality plot

(b) QQ plot

Figure 4: The plots to check the normality hypothesis

We can see that indeed, our model is well-fit.

7 Predictions

We use our trained model, to show it actually does valid predictions. Indeed, the values predicted are very close to the real ones. we give the following table :

```
         Actual  Predicted
```

300	11	10.284105
301	10	10.819358
302	14	12.226177
303	18	17.308347
304	13	14.358755
305	12	12.021887
306	18	18.281405
307	8	9.157853
308	12	11.942118
309	10	10.094151
310	0	8.372304
311	13	12.588258
312	11	10.822758
313	11	10.296695
314	13	12.960482
315	11	11.693222
316	0	7.574864
317	9	9.841155
318	10	10.819358
319	11	10.866410
320	13	13.523454
321	9	9.155555
322	11	10.889937
323	15	13.870498
324	15	15.253791
325	11	11.646114
326	16	15.107489
327	10	10.034509
328	9	8.929485
329	14	14.157957

8 Managerial Report

Our model is able to predict, up to a very good accuracy, an idea of the academic performance of students, giving their previous mark, absences and number of failures. These were the most important variables, impacting in a strong linear manner the explained variables in question. We can hence us the model in order the final results of students, according to a few variables known in advance. These could be taken is consideration for students to improve, such as absences and failures, in order to expect higher overall grades.

9 Conclusion

As a conclusion, we can see that the main factors, linearly effecting a students performance, is his previous grades, absences and failures. A good remark is a lot a variables, which seem at first to be significant, have low to very low significance to grades. Such as going out, parents education, age and distance from school and health. Overall our model is very good to predict with, giving a relatively small set of regressor variables. Hence, the behaviour of the overall performance is quite linear with respect to the chosen variables. We can safely assume our linear regression model is a good fit, and quite significant for this data-set in order to predicts students performance in the big scheme of things.

10 References

1. https://archive.ics.uci.edu/dataset/320/student+performance

2. https://creativecommons.org/licenses/by/4.0/legalcode