ADVANCED STATISTICS PROJECT

Using Poisson regression to predict secondary school students’ performance

Team: Fabiana Caccavale, Marco Amadori, Bruno Lenderink, Lisa Aita

The dataset used to conduct the analysis, and the reference paper “Using data mining to predict secondary school stutents’ performance” by Paulo Cortez and Alice Silva, can be found at: https://github.com/MarcoAmadori1/PORTOGUESE-PROJECT.git

Introduction

The present work intends to build a model which can effectively predict the students’ final grade in the Portuguese subject.

Modeling students’ performance is an important tool for both educators and students, since it can help a better understanding of this phenomenon and ultimately improve it.

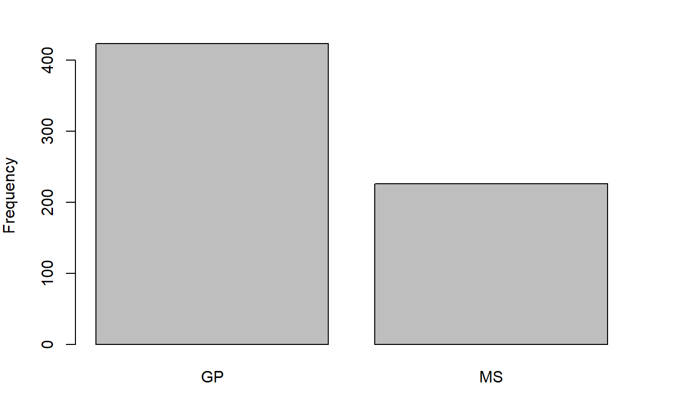
Variables Description

Attribute	Description (Domain)
sex	student’s sex (binary; female or male)
age	student’s age (numeric: form 15 to 22)
school	student’s school (binary: Gabriel Pereira(GP) or Mousinho da Silveira(MS)
address	student’s home address type (binary: urban or rural)
Pstatus	parent’s cohabitation status (binary: living together or apart)
Medu	mother’s education (numeric: from 0 to 4a)
Mjob	mother’s job (nominal[b])
Fedu	father’s education (numeric: from 0 to 4a)
Fjob	father’s job (nominal [b])
guardian	student’s guardian (nominal: mother, father or other)
famsize	family size (binary: = 3 or > 3)
famrel	quality of family relationships (numeric: from 1 – very bad to 5 – excellent)
reason	reason to choose this school (nominal: close to home, school reputation, course preference or other)
traveltime	home to school travel time (numeric: 1 – < 15 min., 2 – 15 to 30 min., 3 – 30 min. to 1 hour or 4 – > 1 hour)
studytime	weekly study time (numeric: 1 – < 2 hours, 2 – 2 to 5 hours, 3 – 5 to 10 hours or 4 – > 10 hours)
failures	number of past class failures (numeric: n if 1 = n < 3, else 4)
schoolsup	extra educational school support (binary: yes or no)
famsup	family educational support (binary: yes or no)
activities	extra-curricular activities (binary: yes or no)
paidclass	extra paid classes (binary: yes or no)
internet	Internet access at home (binary: yes or no)
nursery	attended nursery school (binary: yes or no)
higher	wants to take higher education (binary: yes or no)
romantic	with a romantic relationship (binary: yes or no)
freetime	free time after school (numeric: from 1 – very low to 5 – very high)
goout	going out with friends (numeric: from 1 – very low to 5 – very high)
Walc	weekend alcohol consumption (numeric: from 1 – very low to 5 – very high)
Dalc	workday alcohol consumption (numeric: from 1 – very low to 5 – very high)
health	current health status (numeric: from 1 – very bad to 5 – very good)
absences	number of school absences (numeric: from 0 to 93)
G1	first period grade (numeric: from 0 to 20)
G2	second period grade (numeric: from 0 to 20)
G3	final grade (numeric: from 0 to 20)

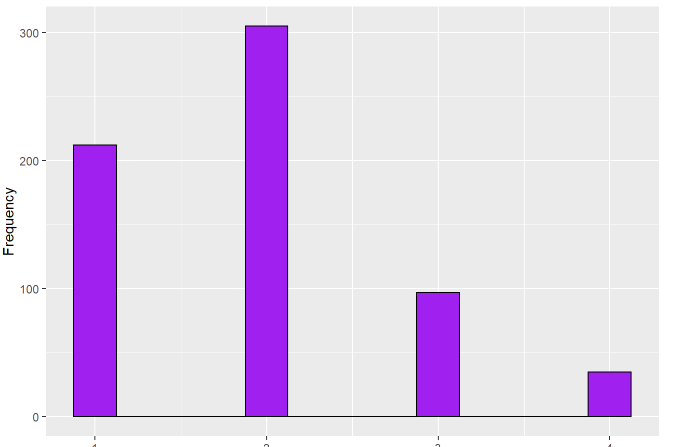
Exploratory Data Analysis

Conducting a preliminary exploration of the data can be useful to better understand the context of inquiry and ease the modeling process.

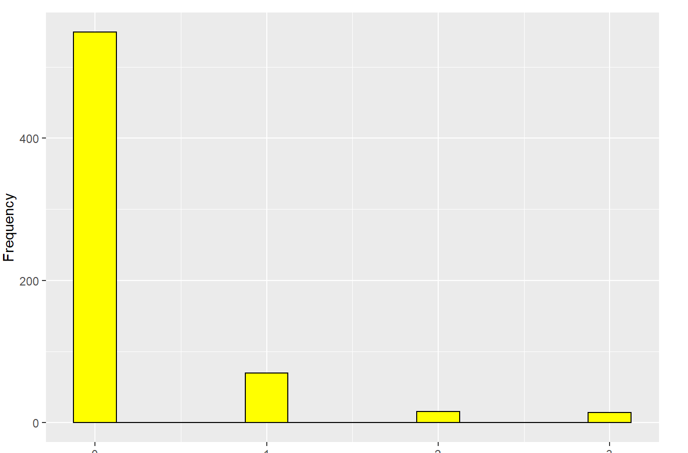
To begin with, given the fact that the dataset is composed by students coming from two different schools, the following barplot shows the different frequencies. The frequency detected in the GP school corresponds to 423 students, while in the MS school corresponds to 226 students.



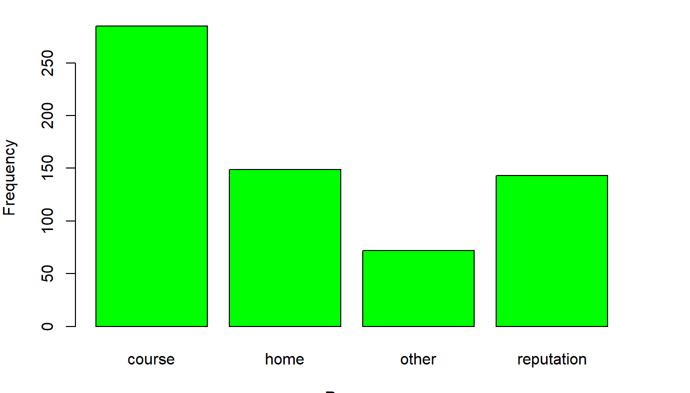
In the next barplot it is investigated how the students can be divided on the basis of the hours of study. What emerges is that the highest frequency of students study two hours, while the lowest number study four hours.



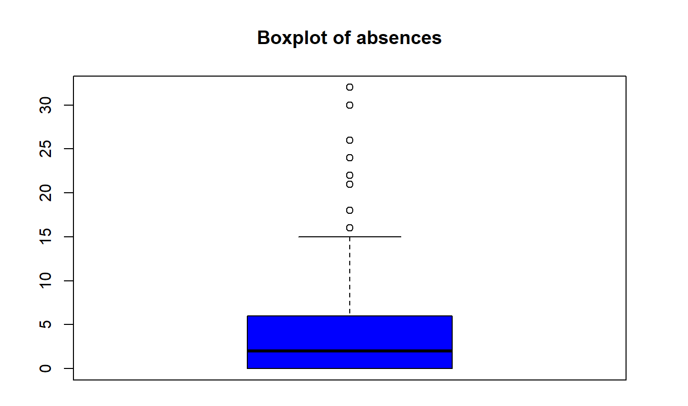
Moreover, the following barplot helps understand the frequency of failures among the students. The vast majority accounts for zero failures, followed by one failure. Only few of the students have failed 2 or 3 times in the past.



From the exploratory data analysis, it is also possible to see that the majority of students chooses the school based on the course. It is then possible to calculate the relative percentages of the choices and the results are that the 44% of the students has chosen the school based on the course offered, 23% based on the proximity to home, 22% for the reputation, and the remaining 11% for non specified reasons.



Lastly, through the absences boxplot, it can be clearly seen that there are some outliers, the maximum observation being 32 and it is possible to observe that the interquartile range of the values related to absences lies in the interval between 0 and 6. The median value for absences is, in fact, equal to two.



Contingency tables

To better understand the relationship between the dataset’s predictors, it is useful to inspect the conditional probabilities.

Table 2. School vs Studytime

	1	2	3	4
GP	0.2813239	0.4869976	0.1678487	0.0638298
MS	0.4115044	0.4380531	0.1150442	0.0353982

Table 3. Sex vs Failures

	0	1	2	3
F	0.8590078	0.1096606	0.0182768	0.0130548
M	0.8270677	0.1052632	0.0338346	0.0338346

Table 4. Higher vs Sex

	F	M
no	0.5072464	0.4927536
yes	0.6000000	0.4000000

Table 5. Medu vs Fedu

	0	1	2	3	4
0	0.1666667	0.3333333	0.5000000	0.0000000	0.0000000
1	0.0349650	0.6433566	0.2517483	0.0559441	0.0139860
2	0.0000000	0.2795699	0.5430108	0.1344086	0.0430108
3	0.0000000	0.1438849	0.3021583	0.3884892	0.1654676
4	0.0057143	0.0457143	0.1542857	0.2514286	0.5428571

Contingency tables interpretation:

• From the first contingency table it is possible to understand that the probability to find a student studying one hour in the MS school is of 41%, while in the School GP is 28%. For higher number of hours studied, the probability to find student studying more than one hour is slightly but consistently higher in the latter school.

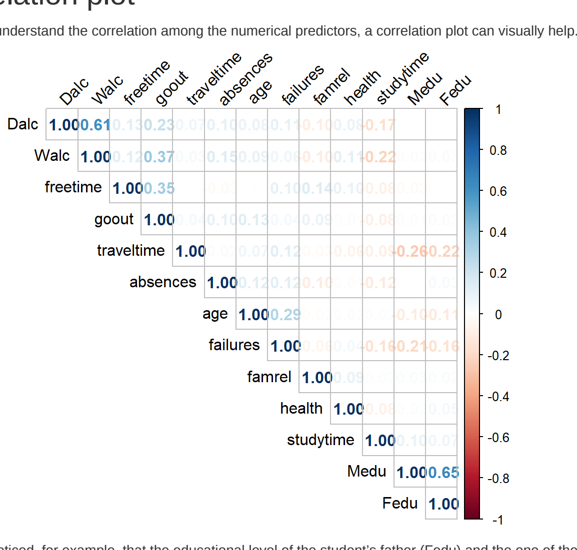
• Concerning the relation between the sex and the number of failures, it is found that there is no significant correlation among the two conditions.

• An important result has been found in the relation between the student’s sex and whether he/she wants to continue his/her academic career, in fact the probability of finding a female who wants to take higher education is 60% while for a male is 40%. However, since these two are categorical variables, it is possible to perform a chi-squared test whose result suggests that the variables are independent.

• Another interesting result has been found inspecting the relationship of the level of education of the parents. It is in fact more probable to find two parents with the same level of education looking at educational levels higher than zero, in the latter case it doesn’t apply. The probability of finding two parents both with the highest level of education (4) is 54%, with a level of education (3) the probability is 38%, with a level of education (2) is 54% and with a level of education (1) is 64%.

Correlation plot

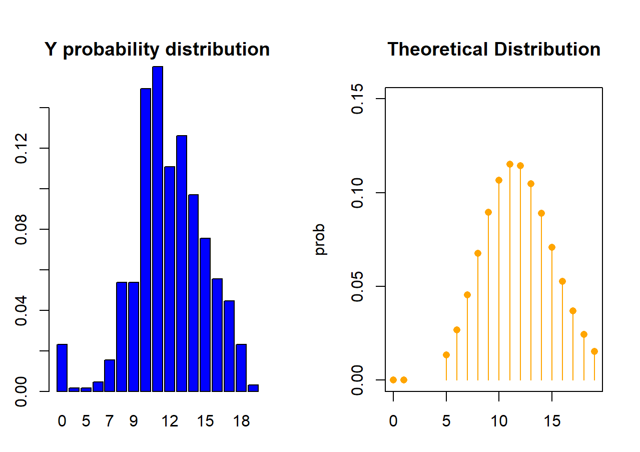
To further understand the correlation among the numerical predictors, a correlation plot can visually help.



It can be noticed, for example, that the educational level of the student’s father (Fedu) and the one of the mother (Medu) have a rather high correlation of 0.65, also noticed in the contingency table. In addition, the daily consumption of alcohol (Dalc) and the weekly one (Walc) are quite correlated.

Probability distribution

Since the target variable G3 (students’ final grade) is a count variable, it is expected to be distributed as a Poisson distribution. In order to check this assumption, a theoretical Poisson distribution having the same range (from 0 to 19) and λ of the target variable was compared with the actual distribution of the target variable.



As it can be noticed from this comparison, the two distributions look quite similar except for the values of G3 equals to 0. The probability distribution of Y, as a matter of fact, has a higher number of zero values, which may be caused by typing errors in the collection of the data since, in their correspondence, the first and second periods’ grades have values that differ from zero.

In general, however, it can still be stated that the target variable G3 approximately follows a Poisson distribution.

Additionally, if we compare the mean and the variance of Y, we find that the mean is equal to 11.9, while the variance is equal to 10.44. Since the mean is higher than the variance, there is a situation of underdispersion. The Poisson distribution condition, according to which the mean of Y should be equal to its variance is not exactly met.

Regression model

In order to estimate a model which can be useful to predict students’ final grade in the Portuguese subject, a generalized linear model was built. It is a generalization of ordinary linear regression that allows for response variables that have error distribution models other than a Gaussian distribution.

To do that, the glm() function was used in R.

```
## 
## Call:
## glm(formula = Y ~ ., family = poisson, data = X)
## 
## Deviance Residuals: 
##     Min       1Q   Median       3Q      Max  
## -5.0472  -0.3941   0.0102   0.4373   2.1373  
## 
## Coefficients:
##                   Estimate Std. Error z value Pr(>|z|)    
## (Intercept)       2.200366   0.218376  10.076  < 2e-16 ***
## schoolMS         -0.100681   0.029724  -3.387 0.000706 ***
## sexM             -0.054827   0.027100  -2.023 0.043055 *  
## age               0.012108   0.011113   1.090 0.275930    
## addressU          0.027353   0.029006   0.943 0.345676    
## famsizeLE3        0.026878   0.026468   1.015 0.309870    
## PstatusT          0.011226   0.037577   0.299 0.765126    
## Medu              0.001663   0.016671   0.100 0.920554    
## Fedu              0.015802   0.014924   1.059 0.289658    
## Mjobhealth        0.072957   0.057782   1.263 0.206724    
## Mjobother         0.004305   0.033988   0.127 0.899199    
## Mjobservices      0.033251   0.041079   0.809 0.418266    
## Mjobteacher       0.040910   0.054189   0.755 0.450274    
## Fjobhealth       -0.052801   0.080907  -0.653 0.514005    
## Fjobother        -0.017513   0.050475  -0.347 0.728615    
## Fjobservices     -0.054036   0.052938  -1.021 0.307377    
## Fjobteacher       0.040343   0.071362   0.565 0.571850    
## reasonhome        0.003244   0.030793   0.105 0.916100    
## reasonother      -0.043619   0.041678  -1.047 0.295305    
## reasonreputation  0.015179   0.031690   0.479 0.631954    
## guardianmother   -0.028408   0.028656  -0.991 0.321528    
## guardianother     0.013515   0.059216   0.228 0.819459    
## traveltime        0.006514   0.017566   0.371 0.710774    
## studytime         0.032331   0.015082   2.144 0.032061 *  
## failures         -0.148940   0.026150  -5.696 1.23e-08 ***
## schoolsupyes     -0.109094   0.040417  -2.699 0.006951 ** 
## famsupyes        -0.003441   0.024865  -0.138 0.889935    
## paidyes          -0.024491   0.051594  -0.475 0.635008    
## activitiesyes     0.015655   0.024281   0.645 0.519088    
## nurseryyes       -0.018370   0.029755  -0.617 0.536990    
## higheryes         0.173776   0.046629   3.727 0.000194 ***
## internetyes       0.021649   0.030772   0.704 0.481718    
## romanticyes      -0.035293   0.025087  -1.407 0.159472    
## famrel            0.012315   0.012909   0.954 0.340094    
## freetime         -0.011816   0.012297  -0.961 0.336605    
## goout            -0.004999   0.011737  -0.426 0.670194    
## Dalc             -0.018388   0.017463  -1.053 0.292362    
## Walc             -0.007880   0.013007  -0.606 0.544630    
## health           -0.014370   0.008365  -1.718 0.085805 .  
## absences         -0.002727   0.002745  -0.993 0.320470    
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## (Dispersion parameter for poisson family taken to be 1)
## 
##     Null deviance: 745.61  on 648  degrees of freedom
## Residual deviance: 529.31  on 609  degrees of freedom
## AIC: 3352
## 
## Number of Fisher Scoring iterations: 4
```

The model that is obtained using all predictors has several variables with a p-value > 0.05 and, therefore, not significant.

To improve the model, the step() function was used, which gives as a result the best model according to the Akaike Information Criterion. However, as it can be seen in the summary below, the model that is obtained has still some variables whose significance level and, consequently, their effects on the target variable, are not very relevant.

```
## 
## Call:
## glm(formula = Y ~ school + sex + Fedu + studytime + failures + 
##     schoolsup + higher + Dalc + health, family = poisson, data = X)
## 
## Deviance Residuals: 
##     Min       1Q   Median       3Q      Max  
## -4.8298  -0.4079  -0.0174   0.4508   2.2538  
## 
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)    
## (Intercept)   2.378652   0.066350  35.850  < 2e-16 ***
## schoolMS     -0.121977   0.025793  -4.729 2.25e-06 ***
## sexM         -0.043068   0.025087  -1.717  0.08603 .  
## Fedu          0.024226   0.010755   2.253  0.02429 *  
## studytime     0.037387   0.014268   2.620  0.00879 ** 
## failures     -0.149565   0.024409  -6.127 8.93e-10 ***
## schoolsupyes -0.115969   0.038696  -2.997  0.00273 ** 
## higheryes     0.182953   0.045147   4.052 5.07e-05 ***
## Dalc         -0.032828   0.013745  -2.388  0.01692 *  
## health       -0.014715   0.007938  -1.854  0.06376 .  
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## (Dispersion parameter for poisson family taken to be 1)
## 
##     Null deviance: 745.61  on 648  degrees of freedom
## Residual deviance: 550.42  on 639  degrees of freedom
## AIC: 3313.1
## 
## Number of Fisher Scoring iterations: 4
```

Thus, the model is improved manually by removing these variables. They are: sex and health

```
## 
## Call:
## glm(formula = Y ~ school + Fedu + studytime + failures + schoolsup + 
##     higher + Dalc, family = poisson, data = X)
## 
## Deviance Residuals: 
##     Min       1Q   Median       3Q      Max  
## -4.8644  -0.4123   0.0077   0.4322   2.2086  
## 
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)    
## (Intercept)   2.31329    0.06048  38.249  < 2e-16 ***
## schoolMS     -0.11398    0.02557  -4.457 8.31e-06 ***
## Fedu          0.02189    0.01071   2.045  0.04088 *  
## studytime     0.04312    0.01401   3.078  0.00208 ** 
## failures     -0.15292    0.02435  -6.279 3.40e-10 ***
## schoolsupyes -0.11027    0.03844  -2.868  0.00413 ** 
## higheryes     0.18141    0.04512   4.021 5.80e-05 ***
## Dalc         -0.04003    0.01333  -3.002  0.00268 ** 
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## (Dispersion parameter for poisson family taken to be 1)
## 
##     Null deviance: 745.61  on 648  degrees of freedom
## Residual deviance: 557.59  on 641  degrees of freedom
## AIC: 3316.3
## 
## Number of Fisher Scoring iterations: 4
```

In this model the regression coefficients represent the expected change in the log of the mean per unit change in the predictors, keeping all other predictors fixed.

In order to interpret them, the exponent of the coefficients should be taken. The obtained results are the following:

```
##  (Intercept)     schoolMS         Fedu    studytime     failures schoolsupyes 
##   10.1076406    0.8922735    1.0221308    1.0440676    0.8582006    0.8955949 
##    higheryes         Dalc 
##    1.1989040    0.9607653
```

Examples of interpretation

Keeping fixed all other independent variables:

• A student attending the “MS” school, is expected to have a final grade 10.8% lower with respect to a student attending the “GP” school (baseline).

• One-unit increase of a student’s father education level (“Fedu”) leads to an increase of 2.2% of the average final grade.

• The number of hours of studytime impact the final grade positively, one additional hour of studying is expected to increase the average final grade by 4.4%.

• The number of failures impact negatively the final grade, one additional failure is expected to reduce the average final grade by 14.2%.

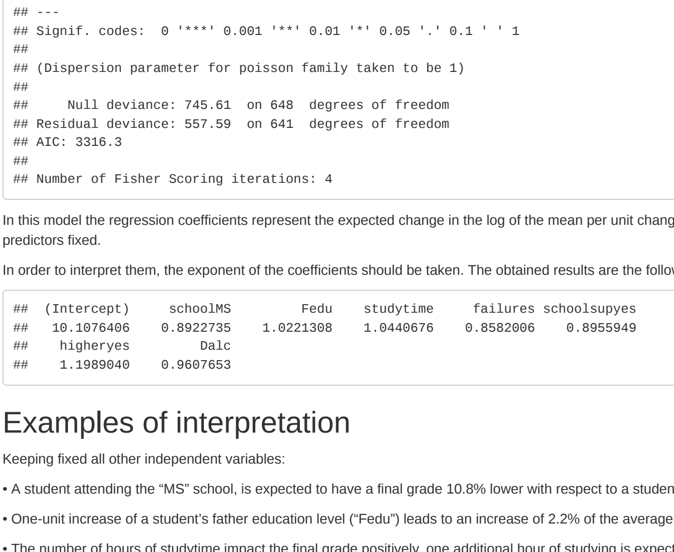
• A student who needs extra educational support (“schoolsupyes”), is expected to have a final grade 10.4% lower than a student without this need.

• A student who wants to take higher education (“higheryes”) is expected to have a final grade 19.9% higher than a students who do not want to continue his/her academic career.

• One-unit increase of a student’s daily alcohol intake decreases the average final grade by 4%.

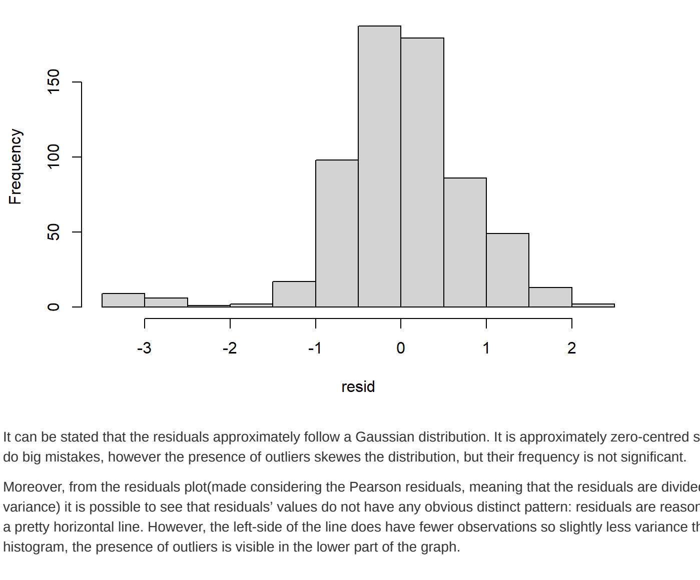
Residuals analysis

If the model correctly describes the variability of the data, then residuals are expected to be normally distributed and independent. In order to determine the shape of the residuals’ distribution, the hist() function was used.



It can be stated that the residuals approximately follow a Gaussian distribution. It is approximately zero-centred so on average the model will not do big mistakes, however the presence of outliers skewes the distribution, but their frequency is not significant.

Moreover, from the residuals plot(made considering the Pearson residuals, meaning that the residuals are divided by the square root of the variance) it is possible to see that residuals’ values do not have any obvious distinct pattern: residuals are reasonably well spread above and below a pretty horizontal line. However, the left-side of the line does have fewer observations so slightly less variance there and as noticed before in the histogram, the presence of outliers is visible in the lower part of the graph.



It is possible to deduce that the model is approximately correct.

Prediction

In order to assess the predictive performance of the model, the dataset is split into train set and test set and the seed is set. In particular, the train set consists of the 80% of the entire dataset, while the test set embarks the remaining 20% of the data. The model built in the previous steps is applied to the train test.

The measure that is used to quantify the error of the model in predicting quantitative data is the Root Mean Square Error (RMSE).

The RMSE of the train set is 2.68. If we apply the same regression model to the test set, the resulting RMSE is 2.66. The RMSE values of the train and test sets are not very different. This implies that there are no obvious problems of overfitting or underfitting in the data.

However, in order to state whether the value of the RMSE is good or not, it is needed to compare it with one of another model.

In this study, the benchmark that is taken is the empty model (model with only the intercept). Also in this case, the RMSE is computed for both the train and the test set. The same conclusion drawn from the comparison of the train and test sets’ RMSEs obtained using the estimated model, also holds here.

The RMSE of the benchmark’s test set is then compared with the RMSE that results by applying the estimated model to the test set. The former is 3.09, the latter as stated before is 2.66. This means that if the values of Y are predicted by applying the estimated model, the predicted values are more accurate than the ones that result by applying the empty model, thus by only knowing the mean value of Y and without having any information about students.

Conclusions

In conclusion, it is possible to state that the model that has been built through this study has a good predictive performance if compared to the one of the empty model. However, the difference between them is not remarkable.

Additionally, if we include in the dataset the predictors that, according to the reference paper, should be the most significant (G1 and G2 – first and second periods’ grades) and estimate a new model, we find out that only the “G2” variable is significant according to the Akaike Information Criterion. The resulting test set’s RMSE is 1.59, which is lower than the one obtained with the proposed model of this study and, thus, it would be possible to obtain better predictions.