Project

BI430: Limassol October 2023
Lanin Gleb
Study time and Travel time

Introduction

Abstract

This report is a companion piece to a research project on the relationship of student grades to time spent studying at home and commuting to school. The project aims to find out how dependent students' grades are on different values (such as absences, age of students, failures and of course the main values in this study are time spent studying and travelling to school).

Background

Education is a critical facet of societal development, and understanding the factors influencing student performance is imperative for educational policymakers. In this context, we have been commissioned by the Portuguese government to conduct a comprehensive study on student achievement. Our focus is on two distinct subjects: Mathematics and Portuguese language. But for convinience, only one dataset will be chosen.

Data Overview

Our analysis is grounded in rich datasets comprising student grades, demographic information, social attributes, and school-related features. These datasets were meticulously collected through school reports and questionnaires, resulting in 33 variables.

The target attribute, $G3$, represents the final year grade, and we acknowledge its strong correlation with $G2$ and $G1$, corresponding to grades from the 2nd and 1st periods, respectively. Predicting $G3$ without $G2$ and $G1$ poses a challenge, but such predictions hold significant utility.

Research plan

This project will utilize various techniques such as common sense, logic and of course statistical analysis. The report will include several main sections:

- Tests for 1 sample
- Tests for 2 samples
- Tests for 3+ Samples
- Regrssion

The tests will provide additional information about the different variables, helping us to better understand the data so that we can confidently run the regression at the end and complete the study.
The regression, in turn, will help us answer the most important question: Is it possible to predict a student's grade using only the time spent studying and traveling to school ?
For each test, I will formulate a precise question in order to avoid deviating from the main flow of the study.

Explanatory analysis

In this section, a general descriptive analysis will be presented to help better understand the nature of the data, obtain possible insights, and potentially make some decisions regarding the overall approach and use of the data.

First, I'd like to see general information about the datasets for our main values: time to study and time to commute to school.

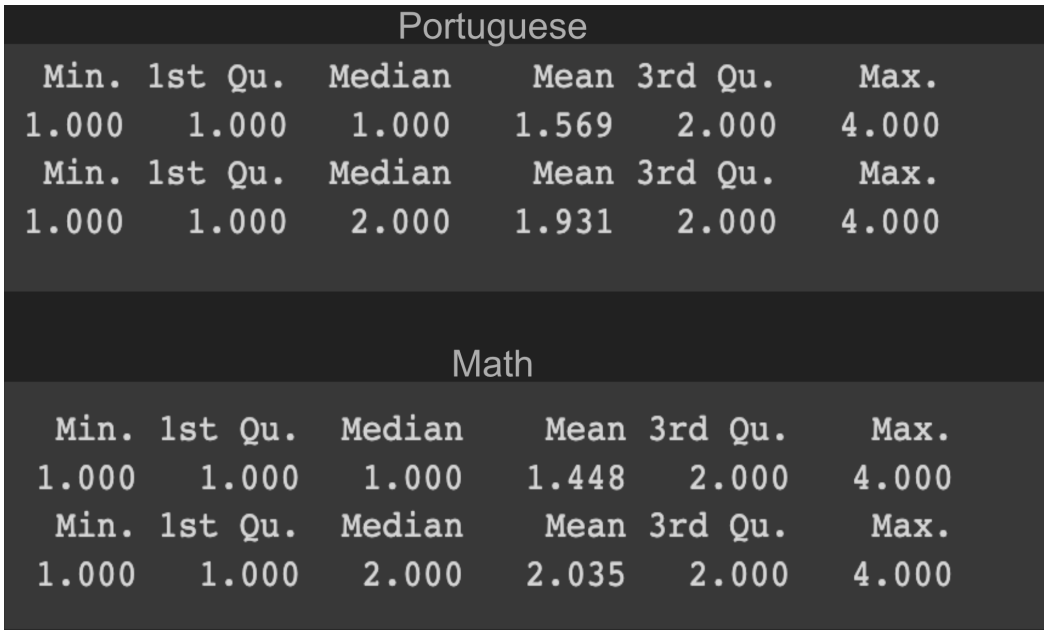

fig. 1

Fig. 1 shows the statistics for travel time (top row in each output) and study time (bottom row in each output). We can see that the extreme groups (1 and 4) in the Portuguese

and math datasets coincide, and that the corresponding categories have approximately the same median and mean values.

Next, I decided to check how many records for the various schools we have in our datasets:

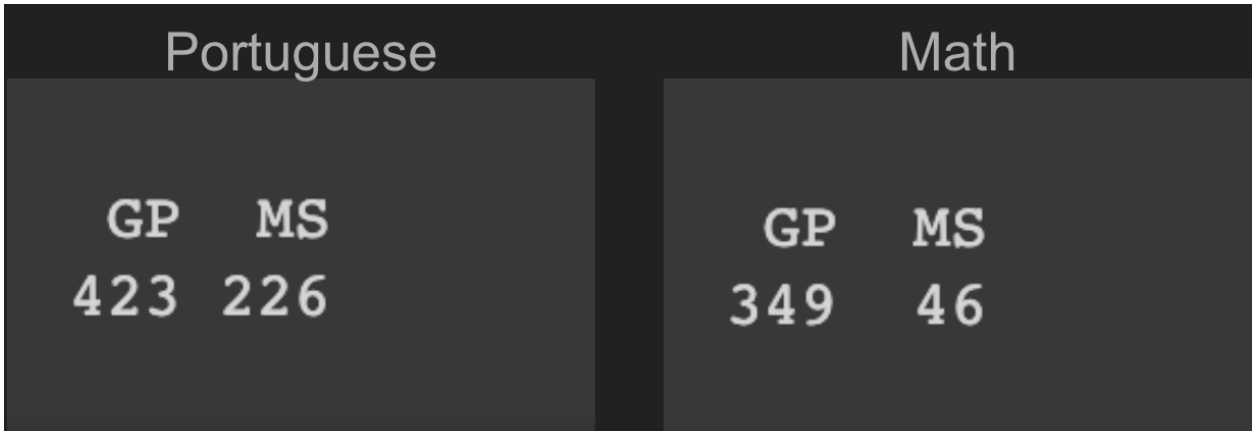

fig. 2

In *fig.2* we see that in the Portuguese dataset, the ratio of students from different schools is better than in the math dataset. Therefore, in my research, my main dataset will be the one that contains records related to Portuguese.

Next, since we are now only dealing with the Portuguese dataset, I would like to see the distribution of scores:

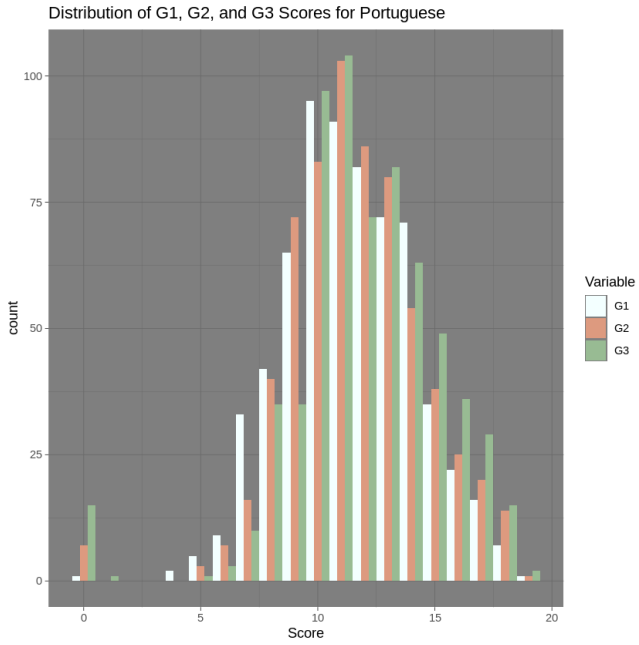

fig.3

Of course, on *fig.3* there are some things that prevent us from saying that the distribution is perfectly normal, as there are certain entries in the left tail far from the rest, as well as sags among the estimates, but in general terms, I don't see anything serious here. In addition, we will try not to deal with $G1$ and $G2$, as using them to predict $G3$ is logical, but not as effective because of the high correlation and therefore high explanatory power.

And so as not to drag on with descriptive analysis, I suggest lastly to look at the scatter plots of grades in relation to time to study and commute to school.

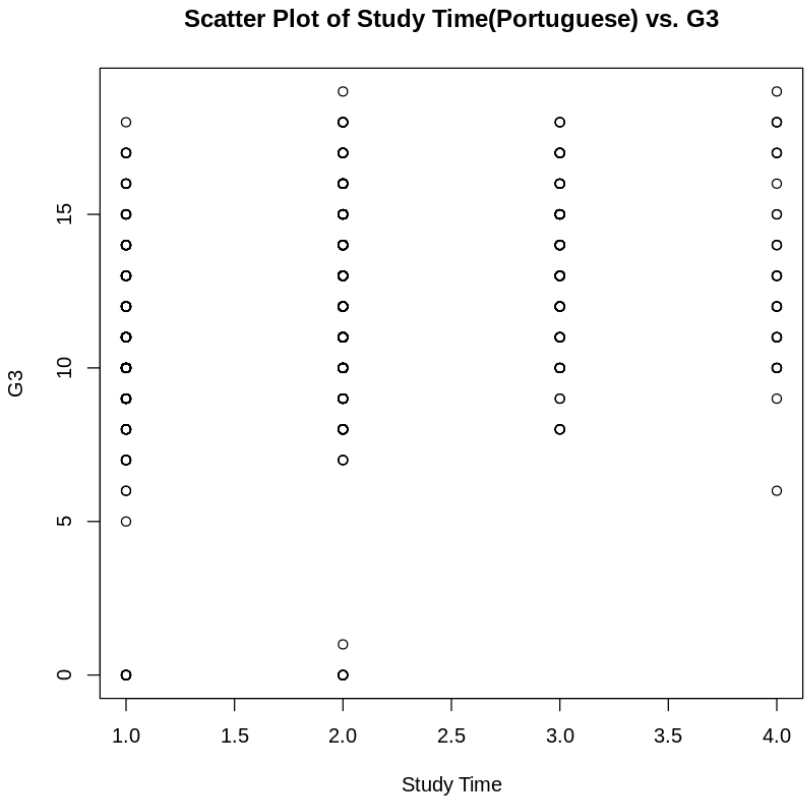

fig.4

And I think this is one of the most important graphs in this report, because essentially, we see that there are high scores in absolutely every category. Based on the visual component of *fig.4*, we can roughly say that it doesn't matter how much a student does in terms of grades.

In final, let's take a look at the same schedule, which is already broken down into groups with different travel times:

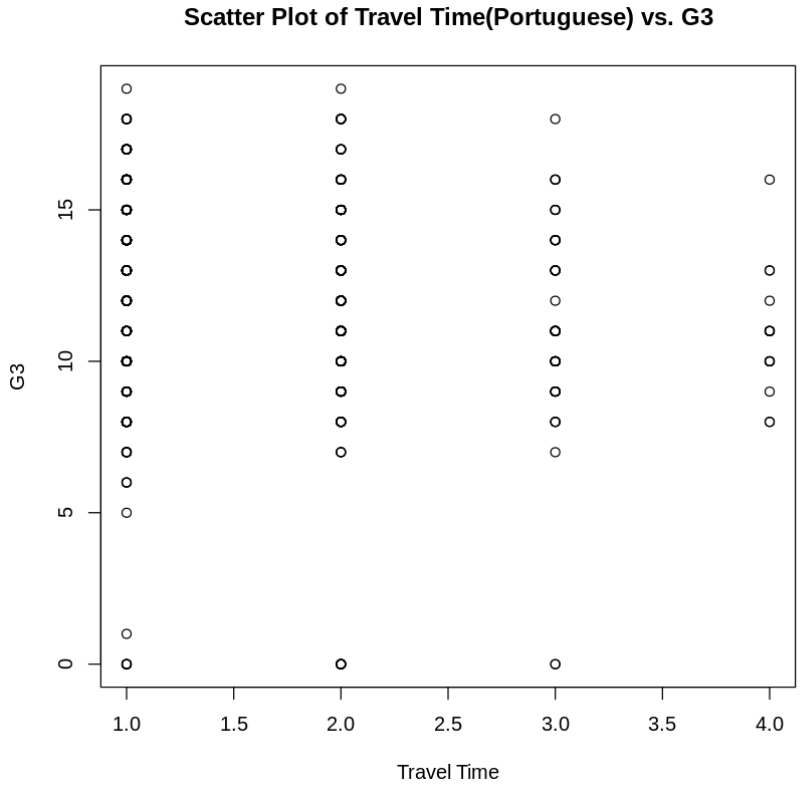

fig.5

And that's where things get interesting. It feels like the longer it takes to get to school (group 4), the fewer high grades there are. But this is just a visual conclusion, in reality it may be that there are simply very few students in group 4 and therefore fewer chances to get a high grade.

This is where I will end with the descriptive analysis, it gave me the thoughts and insights I needed regarding the data. In the next sections we will look at specific tests.

1 Sample

In this section, I would immediately want to check the study time, specifically whether the average study time is less than I would recommend. Unfortunately, I haven't found a clear recommended study time in Portugal, but according to google, you should spend about 1-2 hours studying at home, which gives us about 5-10 hours per week and that is the third group.

Question 1

- Is the mean study time less than recommended ?(3rd group)

So, we are dealing with μ and we are doing a test on a single sample, in that case, we first need to check the distribution (normal or not). For this purpose, we will perform the Shapiro test.

Output of the test:

```
Shapiro-Wilk normality test
data:  Por$studytime
W = 0.82508, p-value < 2.2e-16
```

As we can see, the p-value is extremely low, hence we reject the normal distribution. Next, it is necessary to check the symmetry of the data. For this purpose, we use Symmetry test.

Output of the test:

```
data:  Por$studytime
Test statistic = -3.1944, p-value = 0.01
alternative hypothesis: the distribution is asymmetric.
```

The P-value is less than 0.05, hence we reject H_0, and accept the alternative hypothesis that the data is asymmetric.

Since the data is not symmetric, we will use a permutation test in order to check if μ is less than 3.

$H_0: \ \mu = 3$

$H_A: \ \mu < 3$

Output of the test:

```
Test Statistic:          Sum(x - 3) = -694
P-value:                 0
```

So the p-value is zero, so TISSE to say, that mean study time is less than recommended.

Question 2

After this result, I decided to see what percentage of students were going to pursue higher education. In my understanding, in an ideal world, the amount of time spent on studying and the desire for higher education are strongly correlated. Thus, by conducting this test, I want to better understand the nature of the data I am dealing with. So the question is defined as follows:

- Is the proportion of students searching for higher education less than 0.75 ?

$H_0 : P = 0.75$

$H_A : P < 0.75$

Since my test will be about proportion, I can use the exact test or approximation test. Actually, I'm going to conduct both of them just to double check.

Output of the exact test:

```
Exact binomial test

data:  successes and total_trials
number of successes = 580, number of trials = 649, p-value = 1
```

Output of the approximation test:

```
Exact binomial test

data:  successes and total_trials
number of successes = 580, number of trials = 649, p-value = 1
```

So, we see that in both tests p-value is 1, meaning TISSE to say that more than 75% of students are willing to get higher education.

Which does not resonate with my world understanding, but ok. That basically means that they study less than recommended, but most of them are interested in getting higher education.

2 Samples

Question 1

Here I want to compare study time in two different schools, namely I want to check if the mean differs from one to another. I know that I said that we have different numbers of records from different schools, but after this test, I'm not going to consider schools as groups anymore. It may turn out that study time is different, so then I'd consider usage only of 1 school in regression or schools as a dummy variable. Anyway, let's see what we have here.
The question will be formed as follows:

- Is the mean studytime differs significantly between two schools ?

$H_0 : \mu_1 =' mu_2$
$H_A : \mu$ differs significantly.

Firstly, we need to check the distribution of both groups, so I will use the Shapiro test for each school group.
Output for GP school:

```
W = 0.83567, p-value < 2.2e-16
```

Output for MS school:

```
W = 0.79535, p-value < 2.2e-16
```

So, we reject normality in both groups. Therefore, we need to check if they are symmetric, so I will conduct a symmetry test.
Output for GP school:
```
Test statistic = 0.53411, p-value = 0.688
```

Output of MS school:
```
Test statistic = -5.9978, p-value = 0.002
```

That means, that for the first one, we cannot reject symmetry (there is not enough statistical evidence to say that data is assymetric). But for the second one, we can reject symmetry.
Anyway, I cannot run the permutation test on my machine, so I will use the Wilcoxon test.
Output of the test:
```
data:  studytime by school
W = 55438, p-value = 0.0002901
```

So, there is enough statistical evidence to reject H_0 : mean studytime is the same for both schools. Therefore, we know that studytime in two schools differs significantly, and this test will help me a lot, because I may not include records of one school in regression or I will use school as a dummy variable.

Question 2

Now let's move to travel time. I want to see if the proportion of good grades is the same between groups with small and long travel times. So it will help me better understand the relationship of these variables and its behaviour.
Therefore the question will be stated as follows:

- Is the proportion of good grades the same between groups with small and long travel times?

$H_0 : P_0 = P_1 = P_2$

$H_A : P$ is not equal

Here I need to specify what is a good grade. In my opinion it is more than 75% of the possible grade (like an upper quartile). Therefore, for this test a good grade is defined as 15 and higher out of 20.
Additionally, I have to specify groups for small and large travel time. So I consider large travel time to be group 4, so in this case a threshold is 3. Actually, I tested additionally with threshold = 2 and the conclusion was the same.

So, I will run a proportion test, converting to binaries with good grades and small travel time.
Output of the test:
```
X-squared = 1.1899, df = 1, p-value = 0.2754
```

There is a significant difference in the proportion of good grades between groups with small and long travel times.
Since the p-value is greater than 0.05, we do not have sufficient evidence to reject the null hypothesis.
Therefore, we can conclude that travel time doesn't really matter in our research. But still, I will include it in the model.

3+ Samples

In this section, the main purpose of the tests is to find auxiliary values for more advanced linear regression. By running tests on new values, I can gain the necessary knowledge about other values that may potentially benefit my regression model.

Question 1

While researching the dataset values, my attention was drawn to the variable 'goout', which denotes how often a student goes out with friends, etc. It is expected that the more a student goes out, the less busy he/she is with studying and the less attention he/she pays to the whole process. Thus, I hypothesize that this variable has a lot of explanatory power. But, from now, I think it is better to consider records only from GP school (according to the knowledge I got from ***Question 1*** from **2 Samples** section). Yes, we reduce the total number of records from about 600 to 400, but we still have a bigger dataset compared to the size of the Math dataset. I think
So the question is:

- Is mean go out the same for groups with different grades ?

$H_0 : \mu_0 = mu_1 = \mu_2 = \mu_3$
$H_A : \mu$ is not equal

Now, I need to specify groups of grades. In order to do that I will use some statistical techniques such as histograms and quartiles, because grades is not a categorical variable.

So, let's have a look at the histogram of the G_3:

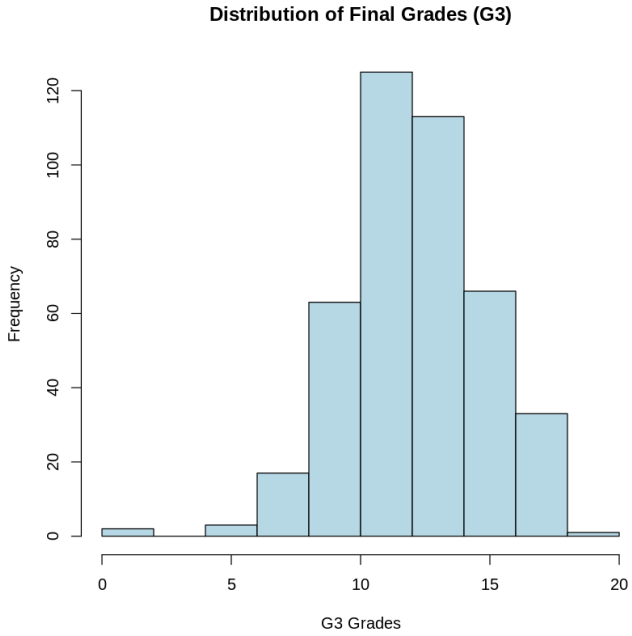

fig.6

I am going to consider *fig.6* only as a supportive graph, as i'm also want to do some quartiles checking for the grades.
Output of the quartiles for GP school records of Portuguese dataset:

```
0%  25%  50%  75% 100%
0   11   13   14   19
```

After considering both outputs, i decided to split grades in the following groups:

- low: 0-9
- satisfactory: 10 - 12
- good: 13-16
- excellent: 17-20

This division is based on the information from *fig.6,* quartiles info and my personal experience of grading in schools and universities.

So now, we need to check the normality of each group of grades.
Output of test:

```
data:  Por_GP$goout[Por_GP$G3_groups == "low"]
W = 0.8306, p-value = 0.0001621

data:  Por_GP$goout[Por_GP$G3_groups == "satisfactory"]
W = 0.9129, p-value = 8.796e-09

data:  Por_GP$goout[Por_GP$G3_groups == "good"]
W = 0.91138, p-value = 6.486e-09

data:  Por_GP$goout[Por_GP$G3_groups == "excellent"]
W = 0.80168, p-value = 2.763e-05
```

We can see that the p-value is small for all of them, therefore, we reject normality, meaning we need to use Kruskal-Wallis test.
Result of the test:

```
Kruskal-Wallis chi-squared = 9.3342, df = 3, p-value = 0.02516
```

We can reject H_0 that mean of goout is the same, meaning that it is not the same for different grades group, therefore, i expect that it will affect the final grade. So I'd like to include that variable later in my more advanced regression model.

Question 2

Now I want to see if we have different variances of failures in different groups of absences. I have a feeling that absence will improve my advanced regression model, so by checking variances I can justify its impact on grades.
Therefore the question is stated as follows:

- Is the variance of failures the same for different groups of absences ?

$H_0 : \sigma_0 = \sigma_1 = \sigma_2$

$H_A : \sigma$ is not equal among groups

Since we are dealing with groups of numerical variable, I will again start with a histogram of absences and then calculate quartiles for it.

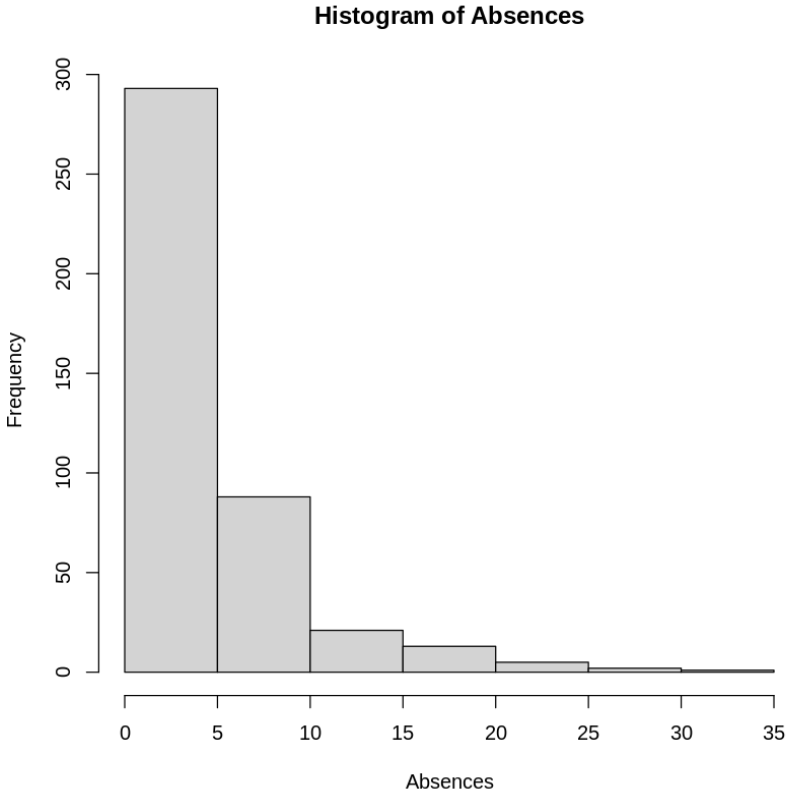

fig.7

So we see that most of the absences are located between 0 and 5 (*fig.7*). But this is not enough to come up with different groups.

Now, let's consider the suggested quartiles of absences:

```
0%  25%  50%  75% 100%
0    0    2    6   32
```

Using this output and *fig.7* I splitted absences in the following groups:

- low: 0-2
- medium: 3-4
- high: 4+

We are ready to make the Levene test. Let me share the output:

```
Df F value  Pr(>F)
group   2  4.1494 0.01642 *
      420
```

So, since p value is smaller than α = 0.05, there is enough statistical significant evidence to say that there is a significant difference in the variances of the failures variable among the absences groups or that variances of failures are not the same among the groups. So, we can conclude that depending on number of absences, we will have different failures, which directly affects our target variable $G3$

Regression

Question

Let's move on to the final part, which is the regression. In this part I will describe my approach to creating linear regression, followed by a section to discuss the results and possible future work.
I had a bit of a hard time formulating a clear question for the regression, so I think the best option sounds like this:

- "is it possible to predict student grades with reasonable accuracy based only on study time and travel time to school? If it is not possible, how can the predictions be improved ?"

by saying reasonable accuracy, i mean R^2 > 0.5

Approach

First, I would like to use the brute force method and immediately build a model where G3 is the target variable and traveltime and studytime are explanatoryб without data preprocessing or cleaning/transformation.
R^2 of the base model:

```
Adjusted R-squared:  0.07579
```

And this is really low. Let's add some more features (explanatory variables) in our model and see what will happen.

Base model + absences:
Adjusted R-squared: 0.09795

We see that R^2 has increased, but insignificantly. Ok, let's try adding more variables step by step.

Base model + absences + goout:
Adjusted R-squared: 0.1069

Adding information about how frequently students go out with friends also increased model accuracy, but it is slightly more than 0.1, when I was aiming at 0.5 at least.

Base model + absences + age + failures + goout:
Adjusted R-squared: 0.2206

Here I also tried to add also the number of failures and age of the students. And I'm not going to add more variables, because we have a lot of them in the dataset. I wanted to select the most interesting to me and those which have high explanatory power in my opinion and those that must increase model performance according to the tests from previous sections. We have R^2 = 0.22, which is much more than we had in the beginning, but it is still not enough. But I didn't make any data transformation before that, so can it help in this case ?
Well, I don't think that we need to do that. Let's discuss why.

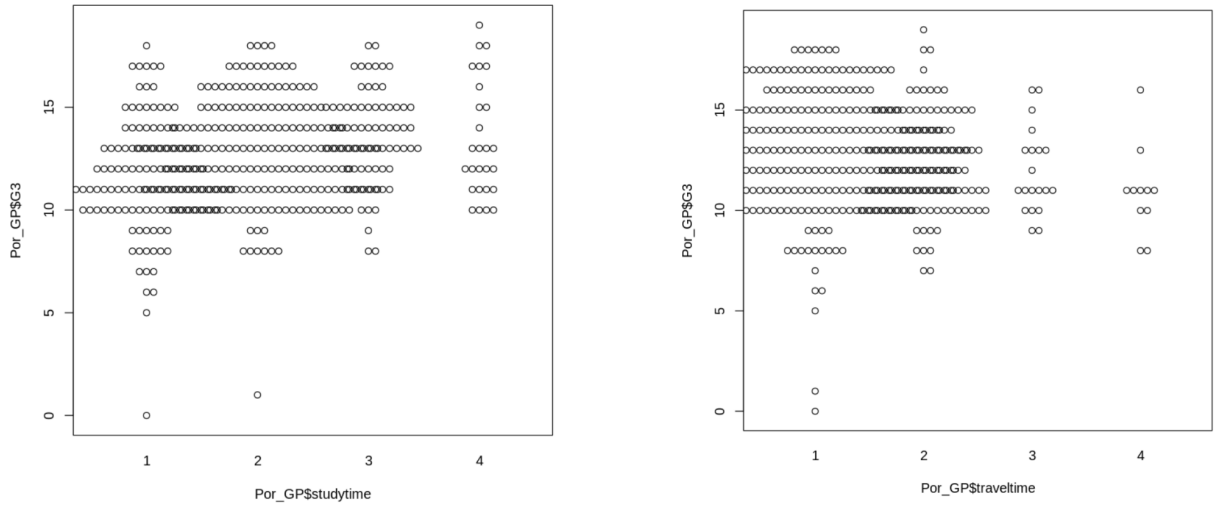

fig.8

On *fig.8* we can see two graphs for $G3$ according to 'traveltime' and 'studytime'. And we see that one of the main assumptions for linear regression about linearity of the data is violated. Basically, *fig.8* explains why in the base model with traveltime and studytime

only we had such a small R^2. We've been trying to predict $G3$ using only categorical variables, which is hard. So later in expanded versions of the base model I tried to add some more numerical data (age and absences for instance).

So let's check what type of dependency we have between other variables and $G3$.

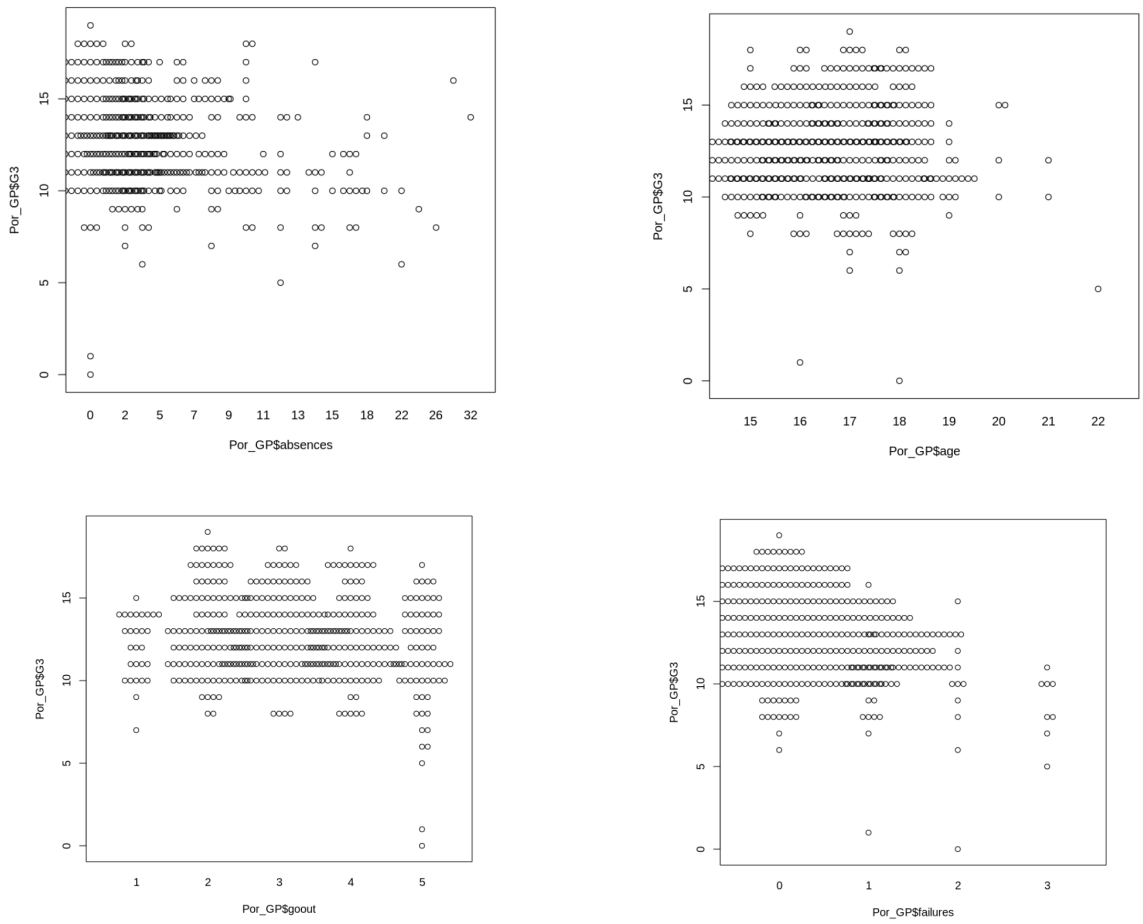

fig.9

In the graphs shown in *fig.9*, we can observe that linear dependence is also not present for the other variables. We could try to transform the data, but it did not give any result. Probably, it is not necessary to attach all the graphs with transformation attempts, as there will be a lot of them. I can say that I tried several methods, such as: logarithmization, raising to different degrees and taking square roots.

Additionally, i tested model, with $G1$ as an explanatory variable. Here is the result without data transformation and data cleaning:
Adjusted R-squared: 0.6427

So, the performance of the model is several times better with only $G1$, in contrast to model with several variables but without $G1$. So in this dataset, grades indeed have the highest explanatory power.

Interpretation

In this subsection, I want to give interpretation to the model, and explain the impact of the variables on final grade.
Firstly, I want to stress, that logically, my best model without $G1$ is correct:

```
Call:
lm(formula = G3 ~ traveltime + studytime + absences + age + failures +
    goout, data = Por_GP)

Residuals:
    Min      1Q  Median      3Q     Max
-9.8960 -1.5706 -0.1676  1.4850  7.2907

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  8.87209    1.60719   5.520 5.97e-08 ***
traveltime  -0.29491    0.16149  -1.826 0.068536 .
studytime    0.46469    0.13942   3.333 0.000936 ***
absences    -0.06481    0.02259  -2.869 0.004329 **
age          0.27209    0.09706   2.803 0.005296 **
failures    -1.73884    0.21986  -7.909 2.36e-14 ***
goout       -0.24503    0.09960  -2.460 0.014292 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.318 on 416 degrees of freedom
Multiple R-squared:  0.2317,    Adjusted R-squared:  0.2206
F-statistic: 20.91 on 6 and 416 DF,  p-value: < 2.2e-16
```

fig.10

So, basically, my equation will look like this:

$$G3 = 8.87209 - 0.29491 * traveltime + 0.46469 * studytime - 0.06481 * absences + 0.27209 * age - 1.73884 * failures - 0.24503 * goout$$

eq.1

From *eq.1* we can conclude that the more students spend on a road to school, the less grade he will get. In return, study time affects positively, meaning that the more a student dedicates time for stude, the greater grade he will get.
We also see negative dependency from the number of absences, failures and how frequently students go out with friends, which all logically negatively affects potential grades. And we see positive dependency from age, which is also logical, because usually the older the student , the better grades he will have. so these results support all of my assumptions during tests from previous sections. I didn’t describe here model coefficients in the way it is usually done for several reasons. Firstly, we have a lot of categorical variables, containing different groups, and secondly, our R^2 is really low, so

from this very model in my opinion it is better to gain the idea rather than exact coefficient interpretation.
At this point, I suggest we move on to the discussion and further work section.

Discussion and Further work

Perhaps I should say that I did a lot of work, I looked at various aspects of the dataset, ran the necessary tests that helped me choose variables for regression and I was also able to do a linear regression model, which turned out to be not particularly accurate. In summary, we have categorical variables dominating the entire dataset and only a small fraction being numerical. In my experience, categorical variables serve mainly as an additive in the model rather than the main explanatory resource. Thus, I can answer the question I posed for the Regression section:
"We cannot predict grades accurately enough based on the variables considered in this project (except $G1$ and $G2$ of course)".

Because of the lack of linear dependence, I did not proceed to data cleaning and transformation, as I intentionally consider linear regression invalid in this setting. However, let's discuss possible ways to solve this problem.

One of them is to measure students' learning time more accurately, for example, on a minute-by-minute basis. This might violate some ethical boundaries, but I think it would have a positive impact on the accuracy of the model.
The same is true for travel time to school.
Because students most likely took a special questionnaire, they could not accurately write their study time. Therefore, the method described above may work.

The second method is to solve the problem not by regression but by classification. This is of course a separate topic for the project in my opinion, but it can also work. One can try to change $G3$ from a numerical variable to a categorical one, for example "passed"/"failed". And using other models try to categorize the student's grade according to these two groups. In theory, this would also increase the accuracy of the model.

In any case, I don't think it is possible to get good enough results with the current data without using $G1$ and $G2$. Of course, you could clean the data and try to add new features to the model, but in my opinion this is a bit out of the original scope of the project, since I was interested in specific two values and possible use of additional variables, and adding all possible values to the model and cleaning them up would take an extremely long time. The code will be available in the appendix and also at the following link. The link leads to google colab, where all the code was written. I will also publish all the components of the project on my GitHub.

List of sources

http://www.sthda.com/english/wiki/kruskal-wallis-test-in-r

https://www.geeksforgeeks.org/levenes-test-in-r-programming/

http://www.sthda.com/english/wiki/unpaired-two-samples-wilcoxon-test-in-r

https://r-charts.com/distribution/beeswarm/