**University of Warsaw**

**Faculty of Economic Sciences**

Mikołaj Gronowski

361377

Dawid Siwicki

348887

**„Impact of the situation and time spent on students’ performance”**

The work was carried out under

the supervision of Rafał Woźniak

from the Department of

Statistics and Econometrics WNE UW

Warsaw, May 2018

**Abstract**

The main purpose of the paper is to examine the influence of situation of the students and their spent time on overall year grades. For this research we used 395 observations of students achievement in secondary education in math subject of two Portuguese schools. Because grades are natural and have values between 0 and 20 we performed count data models.

**Introduction**

As we mentioned we use 395 observations of students from two portuguese schools. The final grade was from math subject. It is important topic because it will help students to take better decision regarding better grades. For modeling students grades econometricians usually use ordinal logit model. In our case we used count models because we can measure differences between values in the dependent variable which are natural numbers from 0 to 20. In the next section we present literature and hypothesis based on it and our thinking.

**Literature Review**

Students’ performance is common topic conducted in studies on advanced econometrics methods. Generally, authors try to find determinants of students’ grades. Below are presented key findings in literature.

Alija (2015) examined probability of passing the external testing to students’ in Macedonian Economics’ Secondary School in Tetovo. He used ordinal logistic regression to examine impact of determinants like Gender, Ethnicity, School year and Subjects of the course.External testing is a part of a reform of Macedonian educational system, introduced after civil conflict in 2001. Reforms caused many positive changes like boost of secondary school enrollment, opportunity to train for teachers and higher achievements of students. Students are assessed based on internal school-based assessed and external national examinations. Alija used data of 2331 surveys in External Testing in Academic Year 2013-2014. He got Cox & Snell R Square statistic equal to 0,29, which means 29% of variation of dependent variables is explained by the model. NagelkerkeR Square value is equal to 0,57, which means there is strong relationship between dependent variable and independent variables. The results of Alija’s model verified that females have higher achievement than males, also Macedonian students have better achievement than Albanian. Based on p-value, Ethnicity is the only variable with significant impact on students’ achievement in this model. The first-year students’ achievements are better than second-year students’, which is better than for the third-year students. Achievement of Vocational Course is better than the General Course.

Sharmin and Rahman (2015) examined impact of some socio-economic, demographic, familial, individual students’ scholastic and institutional factors on achievement in the 1st semester of undergraduate students. Authors based their study questionnaires of 140 students of International University of Business, Agriculture and Technology in Bangladesh. They examined determinants of students’ CGPA on their 1st semester of studies. This variable has categories: A (which is the best), B, C, D and F (which is the worst).The CGPA means Cumulative Grade Points Average and is the average of Grade Points obtained in all subjects during the semester. Students get 10 points for each received A grade, 8 points for each B, and so on. As independent variables, the authors used students average CGPA during studies, their exactly previous semester’s CGPA, percentage of attendance during 1st semester, average study hour, father’s educational qualification and residence place. As a result, the authors got three variables with significant impact on dependent variable. Estimated model predicted 53% of the students’ grades.Students who were serious about their studies on 1st semester were more likely to get A, B, C or D rather than F. Students who did not have problems with understanding the courses were more likely to get A, B, C or D.

AdejumoandAdetunji (2013) used ordinal logistic regression to impact on performance of students of Faculty of Science, University of Ilorin, Nigeria in academic year 2011-2012. The Nigerian society expect education can lead to technological advancement of the nation. They think it is only way for national development. Unfortunately, students’ performance is declining. The authors tried to find factors which affect on academic performance, because it’s measurement can be useful for universities and help in understanding how to improve students’ performance. The authors used data about 623 students which included factors like department, age at entry, sex and mode of entry. As a result, only sex was not significant variable. Mode of entry gave the highest explanation of variation of independent variable. Students who were under 20 at the time of entry are more likely to graduate with First Class than students who were above 20. The students who entered onthe university directly are more likely to graduate with First Class than students who were admitted through REM (Remedial) and also more likely than students who passed JAMB Exam. The students who were admitted through UME are more likely to graduate with First Class than students who were admitted through REM.

Fielding, Yang and Goldstein (2002) examined A-levels results in England and Wales in 1997 using multilevel ordinal models. They chose outcomes for Chemistry and Geography because of popularity of these two exams (the largest number of students chose these subjects) and they have distinct distributions of grades. As independent variables, the authors used students’ gender, age and 11 categories of institutions, which students belong to, according to their admission policy and type of funding, and examination boards. Then, authors compared results between the Normal point score and ordinal models. The authors observed that predictions of normal models are difficult to relate to grades, on the other hand ordinal models directly predict probability that students with given characteristics will achieve certain grade.

Park and Kerr (1990) in their paper examined determinants of academic performance using multinomial logit model. They used sample of 97 students’ questionnaire responses associated with money and banking course between spring of 1984 and spring of 1987. Used variables covers 5 main topics: intelligence (percentile ranking on the ACT, cumulative GPA), preparation (number of hours of economics, total credit hours, and whether student had completed courses like math, accounting or finance), attitude (student’s perception of the value of the course for his major and overall education), effort (hours spent each week at a job, number of hours in which student was concurrently enrolled, hours spent studying for the course) and demography (age, sex, commuting distance, living arrangements). Based on correlation coefficients authors chose 3 variables (cumulative GPA – dependent variable, percentile rank on the ACT, absences) which were used in prior regression. Firstly, they estimated OLS model with cumulative GPA as a dependent variable. Then, using estimated GPA, they have estimated logit model. The results suggested that the key determinants are cumulative GPA and percentile ranking on the ACT (college entrance exam). There was no impact of demographical variables on students’ performance. The most important were effort and intelligence.

**Methodology**

As it was said in our literature review and because of specifics of our dependent variablewe decided to estimate count data model instead of ordinal logistic regression. Usually researchers examine determinants of students’ performance using students’ grades (generally in 2-5 or A-E scale) as outcome. We will compare Poisson, Negative Binomial and Zero-Inflated models and choose the most fitting.

**Hypothesis**

As studentswe would like to know how our situation and spent time affect our overall grades. Our main hypothesis is that students study time will be the most significant variable which affects final grade positively and the absences negatively. From our experience we know that learning couple of hours is enough to get good grades. We think that learning less than 2 hours is not enough and more than 10 hours is too much. On the other hand too many absences make it almost impossible to pass.

One of our secondary hypothesis is that small consumption of alcohol during weekends frees students from stress and will affect grades positively. We also assume that very low consumption during weekend will have similar effect.

The next hypothesis is that high number of past failures indicates that the students will have low number of points at the end of year. There will be no suprise if we say that students with no failures will have top grades.

The last hypothesis is about health. If student is in bad health it will result in bad grades because he will not have time or strength to learn. We also think that people in romantic relationships can have slightly lower points.

**Data Description**

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 subject Portuguese (por).

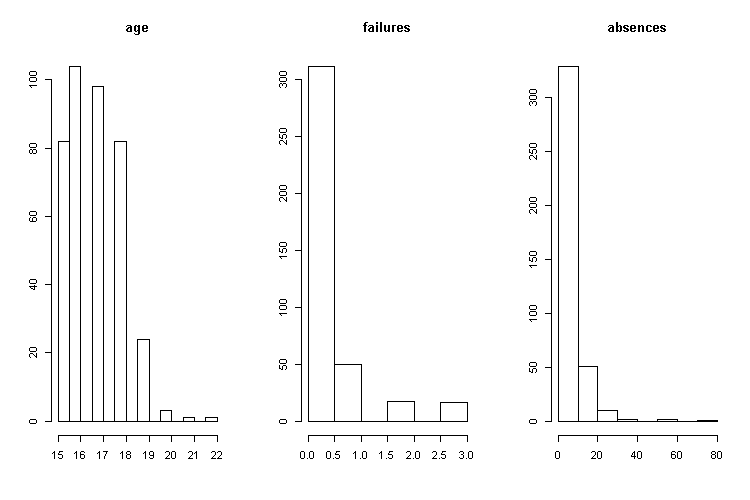
**Attributes**

Table 1. Variablesdescription.

|  |  |  |  |
| --- | --- | --- | --- |
| **Name** | **Full Name** | **Type** | **Values** |
| school | student'sschool | Binary | 'GP' - Gabriel Pereira or 'MS' - Mousinho da Silveira |
| sex | student's sex | Binary | 'F' - female or 'M' - male |
| age | student'sage | Numeric | from 15 to 22 |
| address | student'shomeaddresstype | Binary | 'U' - urban or 'R' - rural |
| famsize | family size | Binary | 'LE3' - less or equal to 3 or 'GT3' - greater than 3 |
| Pstatus | parent'scohabitation status | Binary | 'T' - living together or 'A' - apart |
| Medu | mother'seducation | Categorical | 0 - none, 1 - primary education (4th grade), 2 - 5th to 9th grade, 3 - secondary education or 4 - higher education |
| Fedu | father'seducation | Categorical | 0 - none, 1 - primary education (4th grade), 2 - 5th to 9th grade, 3 - secondary education or 4 - higher education |
| Mjob | mother'sjob | Categorical | 'teacher', 'health' care related, civil 'services' (e.g. administrative or police), 'at\_home' or 'other' |
| Fjob | father'sjob | Categorical | 'teacher', 'health' care related, civil 'services' (e.g. administrative or police), 'at\_home' or 'other' |
| reason | reason to choose this school | Categorical | lose to 'home', school 'reputation', 'course' preference or 'other' |
| guardian | student'sguardian | Categorical | 'mother', 'father' or 'other' |
| traveltime | home to school travel time | Categorical | 1 - <15 min., 2 - 15 to 30 min., 3 - 30 min. to 1 hour, or 4 - >1 hour |
| Studytime | weeklystudytime | Categorical | 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 educationalsupport | Categorical | yesor no |
| Famsup | family educationalsupport | Categorical | yesor no |
| Paid | extra paid classes within the course subject | Categorical | yesor no |
| Activites | extra-curricularactivities | Binary | yesor no |
| Nursery | attendednurseryschool | Binary | yesor no |
| Higher | wants to take higher education | Binary | yesor no |
| Internet | Internet accessathome | Binary | yesor no |
| Romantic | with a romanticrelationship | Binary | yesor no |
| Famrel | quality of family relationships | Categorical | from 1 - very bad to 5 - excellent |
| Freetime | freetimeafterschool | Categorical | from 1 - very low to 5 - very high |
| Goout | going out with friends | Categorical | from 1 - very low to 5 - very high |
| Dalc | workdayalcoholconsumption | Categorical | from 1 - very low to 5 - very high |
| Walc | weekend alcoholconsumption | Categorical | from 1 - very low to 5 - very high |
| Health | currenthealth status | Categorical | from 1 - very bad to 5 - very good |
| Absences | number of schoolabsences | Numeric | from 0 to 93 |
| G3 | finalgrade | Numeric | from 0 to 20, output target |

Source: Ownelaboration.

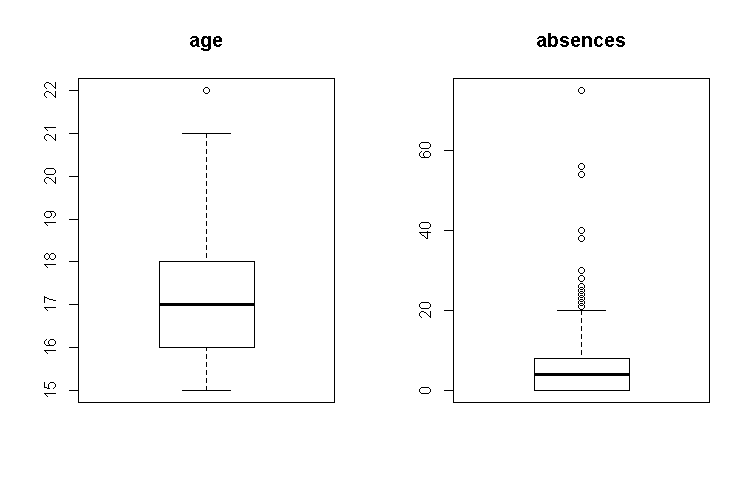
Figure 1. Histograms for numerical variables.



Source: Own elaboration.

These three variables are right skewed. Logarithm of these variables does not help much in the case of normal distribution. The right skewness indicates that students are pretty young and did not fail too many times in the past and they try to be present in the lessons.

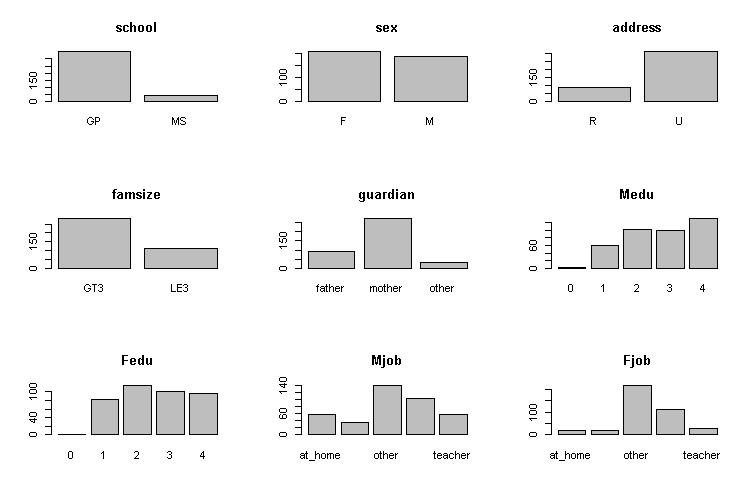
Figure 2. Boxplots for age and absences.



Source: Own elaboration.

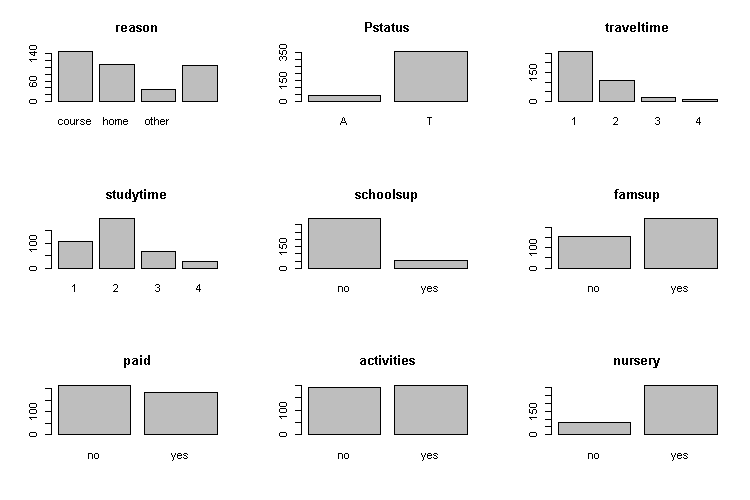
We see that we have couple of outliers. The highest number of outliers is in the variable absences. Maybe the main reason for that situation is that the students were sick.

Figure 3. Histograms for categorical variables (1).

  
Source: Own elaboration.

The distributions of mother’s and father’s education are very similar. The same goes to categories of parents job.We also see that most of the students are from Urban area and from GP school.

Figure 4. Histograms for categorical variables (2).

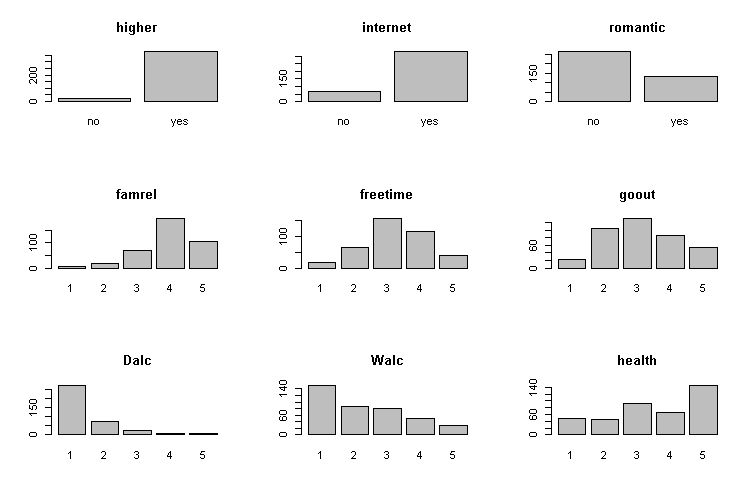


Source: Own elaboration.

We see that the biggest difference between categories are in variables: schoolsup and Pstatus. There is no suprise here because students often do not go for extra classes and parents live together.

Most of the students study for 0-5 hours weekly and do not spend much time in the transport because the big part of them live in the Urban area.

Figure 5. Histograms for categorical variables (3).

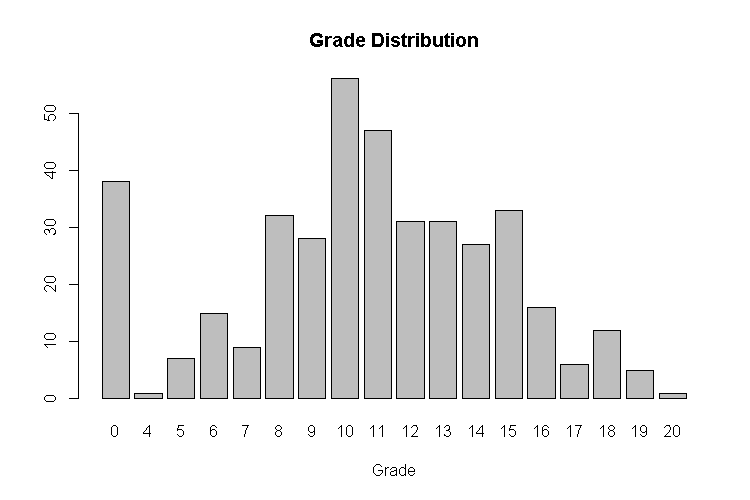


Source: Own elaboration.

Variables freetime and goout have very similar distributions maybe because student spend their free time by going out together.The suprising fact is that about 16.5% of students did not have Internet Access in Home maybe because the data was gathered in 2008.About alcochol consumption students do not drink much during workdays but the numbers go up on the weekends.

We check the distribution of student’s grades.

Figure 6. Histogram for dependent variables (G3).

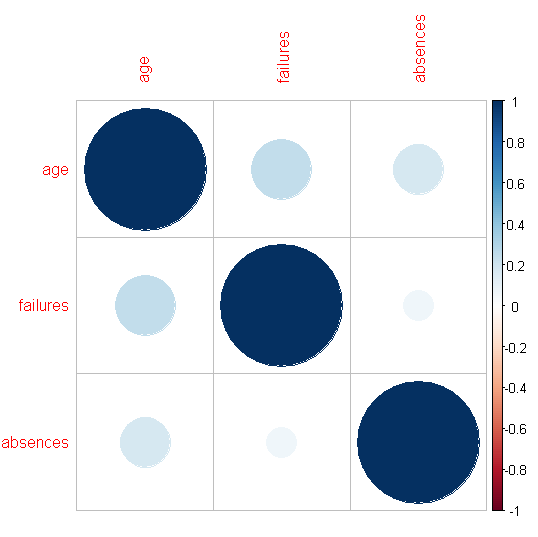


Source: Own elaboration.

We see that there are many students who have 0 points but if we assume that student has to have more than 50% of points than about 53% students pass.

The next step is checking correlation between numerical variables. In our dataset the only numerical variables are age, failures and absences.

Figure 7. Correlation between numerical variables.



Source: Own elaboration.

Our assumption was that number of failures could be correlated to the number of student’s absences. As we can see there is no significant correlation between these variables.

**General to specific model**

Firstly, we estimated 3 models for count data using all variables: Poisson regression, negative binomial regression and Zero-Inflated Poisson Model to choose the most appropriate model for our data. The results of each of these models will be compared with final model in a table below. General models were given by following formula:

The first step to choose the most appropriate model to our data is to check whether there is over-dispersion in data set. Over-dispersion means conditional variance is larger than conditional mean. If there is over-dispersion in data set, Poisson regression model should not be used. The Poisson regression and negative binomial model are nested, Poisson is a special case of negative binomial with parameter theta is equal to infinity.

To check whether there is over-dispersion in our dataset we made likelihood ratio test, where the null hypothesis is that theta is equal to infinity, and alternative hypothesis that theta is lower than infinity. If null hypothesis cannot be rejected, the Possion regression should be chosen and if null hypothesis is rejected, the negative binomial model should be chosen.

At first, we checked theta parameter for negative binomial model:

>model\_negative$theta

[1] 14.782

It meant that theta is lower than infinity and we made likelihood ratio test to confirm that:

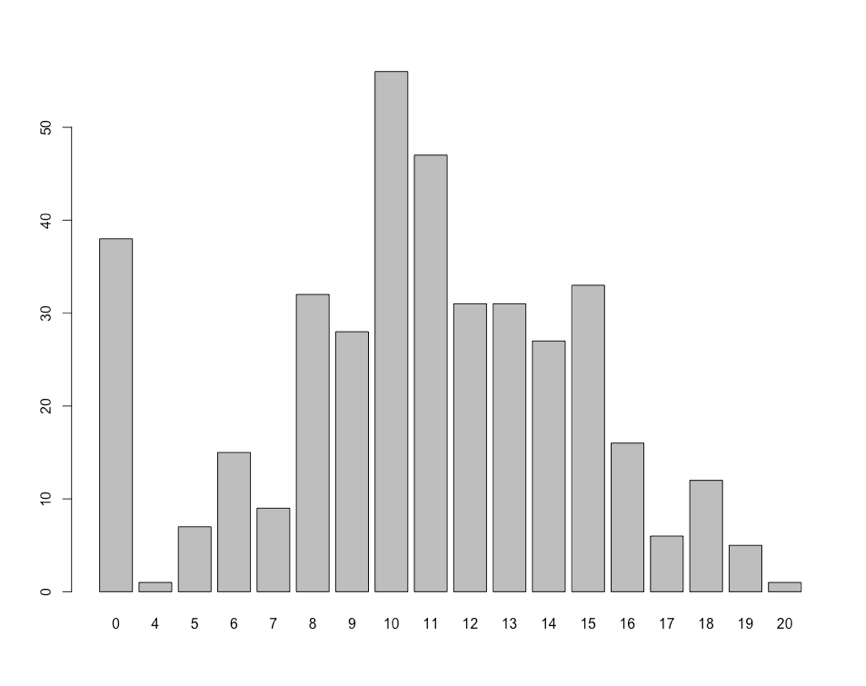
>pchisq(2\*(logLik(model\_negative)-logLik(model\_poisson)), df=1, lower.tail= FALSE)

'log Lik.' 0.000000009154813 (df=69)

The result of likelihood ratio test meant that null hypothesis should be rejected. P-value was lower than 5% level of significance, so we decided to move to next step – checking whether there are excess zeros to choose the most appropriate model between negative binomial model and Zero-Inflated Poisson Model.

Next step after examining over-dispersion is to check whether there are excess zeros in outcome variable. For this purpose, we drew distribution of our dependent variable.

Figure 8. Histogram for dependent variable (G3).



Source: Own elaboration.

Based on this plot we can tell that there is a lot of zeros, but it is hard to say that there are excess zeros.Now we know that there is over-dispersion in our dataset and there are no excess zeros in dependent variables, so basing on this knowledge we can tell that it will be best to estimate negative binomial regression on our dataset. Subsequently we had to find correct formula for our model. For this purpose, we made ANOVA for general negative-binomial model and model with only variables which were significant in general model. This procedure allowed us to check whether rejected variables were jointly insignificant. Based on p-values for variables in general model, we estimated simplified model which was given by following formula:

Next, we made ANOVA for general and simplified model. The result of ANOVA is given below:

Table 2. ANOVA results for General model and simpler model.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| *Model* | *theta* | *Resid. df* | *2 x log-lik.* | *Test* | *df* | *LR stat.* | *Pr(Chi)* |
| Simpler model | 8.149 | 386 | -2421.804 |  | NA | NA | NA |
| General model | 14.782 | 327 | -2348.255 | 1 vs 2 | 59 | 73.549 | 0.0963 |

Source: Ownelaboration.

Based on this result, we can tell that rejected variables are jointly significant and simplified model is appropriate model for our data.

In the next step we check p-values for variables in simpler model, we estimated model with p-values >0.05.

Table 3.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| *Model* | *theta* | *Resid. df* | *2 x log-lik.* | *Test* | *df* | *LR stat.* | *Pr(Chi)* |
| sex + studytime\_3 + failures + romantic + absences | 7.8481 | 389 | -2427.044 |  | NA | NA | NA |
| sex + Medu\_primary + studytime\_3 + failures + romantic + freetime\_5 + Walc\_5 + absences | 8.149 | 386 | -2421.804 | 1 vs 2 | 3 | 5.240 | 0.155 |

Source: Ownelaboration.

According to p-value which is >0.05 we cannot reject null hypothesis that variable are jointly insignificant. So in our case we choose even simpler model.

After performing ANOVAtests we get our final model which is

Below is summary of our final model:

Table 4. Final model results.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  |  | Final model: negative-binomial | | |
|  |  | *Estimate* | *Conf. Int.* | *p-value* |
| (Intercept) |  | 2.33 | 2.24 – 2.42 | <.001 |
| sex (M) |  | 0.12 | 0.03 – 0.22 | .014 |
| studytime\_3 |  | 0.15 | 0.02 – 0.28 | .026 |
| failures |  | -0.26 | -0.34 – -0.19 | <.001 |
| romantic (yes) |  | -0.11 | -0.22 – -0.01 | .030 |
| absences |  | 0.01 | 0.00 – 0.01 | .019 |
| Observations |  | 395 | | |

Source: Ownelaboration.

**Interpretation of parameters**

Given the other variables are held constant in the model. The difference in the logs of expected counts is expected to be 0,122292 unit higher for males compared to females students, while holding the other variables constant in the model.

Given the other variables are held constant in the model. The difference in the logs of expected counts is expected to be 0,147019 unit higher for students studying between 5-10 hours compared to students studying less than 2 hours, while holding the other variables constant in the model.

If a student were to increase his number of failures by one, the difference in the logs of expected counts would be expected to decrease by 0,264784 unit while holding the other variables in the model constant.

Given the other variables are held constant in the model. The difference in the logs of expected counts is expected to be 0,113837 unit lower for students with romantic relationship compared to students without romantic relationship while holding the other variables constant in the model.

If a student were to increase his number of absences by one, the difference in the logs of expected counts would be expected to decrease by 0,006978 unit while holding the other variables in the model constant.

**R2 statistics**

In our case McKelvey-Zavoina is NA so At least 0 differentiation of the dependent variable could be explained in favorusing variables included in the model if the values ​​of the dependent variable were observed.

Count R2 equals to 0.134177215 so about 13.5% of the value of a dependent variable can be correctly predicted using the estimated values parameters of the model.

Our Count Adjuested R2 is -0.008849558 so we should not interpret this.

**Linktest**

If our model really is specified correctly, then if we were to regress G3 on the prediction and theprediction squared, the prediction squared would have no explanatory power.

Table 5. Linktest results.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Estimate | Std. Error | Z value | Pr(>|z|) |  |
| (Intercept) | 0.330092 | 0.431705 | 0.765 | 0.444497 |  |
| yhat | 0.297091 | 0.089736 | 3.311 | 0.000931 | \*\*\* |
| yhat2 | -0.009832 | 0.004617 | -2.129 | 0.033221 | \* |

Source: Ownelaboration.

When we performed link test for out model we found that the prediction squared does have explanatory power because p-value<0.05, so our specification is not good as we thought.

**All models comparison**

Table 6. All models’ comparison.

Source: Own elaboration.

**Verification of hypotheses**

Our main hypothesis was that students who study between 2 and 10 hours will have better grades than other students. Our model confirms that students who study between 5 and 10 hours have better grades.As we said studying for less than 2 hours is too less and studying more than 10 hours too much. To our surprise absences positively influence grades.

First secondary hypothesis said that alcohol consumption during weekend will have positive effect on grades. Alcohol consumption during week and weekend do not have affect students’ grades. We rejected both variables when we performed anova tests.

Next secondary hypothesis was that if students has failed many exams in the past he will have low grade at the end of the year. The failures variable in presented model is the most significant one and has the biggest influence. Students we bigger number of failures tend to have lower grades.

Because students’ current health status was varied we assumed that those with better health will have better grades. In this situation this variables also does not have significant effect on grades. If student was in relationship he would have lower grades than those who are alone. Reason for that is maybe student tend to focus on their relationship rather on studying.

**End notes**

Our literature suggested to use ordinal logit model to build a model about students’ grades but in our research we used count model. The best from poisson, negative binomial and zero inflated model was the second one. From almost 30 variables the significant were only 5: student gender, studytime\_3, failures, romanticyes and absences. Our adjusted count R2 was about 13,5% which says that our model performance is not the best. We also performed link test to check our model specification which turned out to be not the best one. We only confirmed about two our hypothesis about study time and number of failures.