Problemset 1 - Marketing Analytics

Institute of Information Systems and Marketing (IISM)

Julius Korch, Marco Schneider, Stefan Stumpf, Zhaotai Liu

Last compiled on November 17, 2023

Contents

Task 1

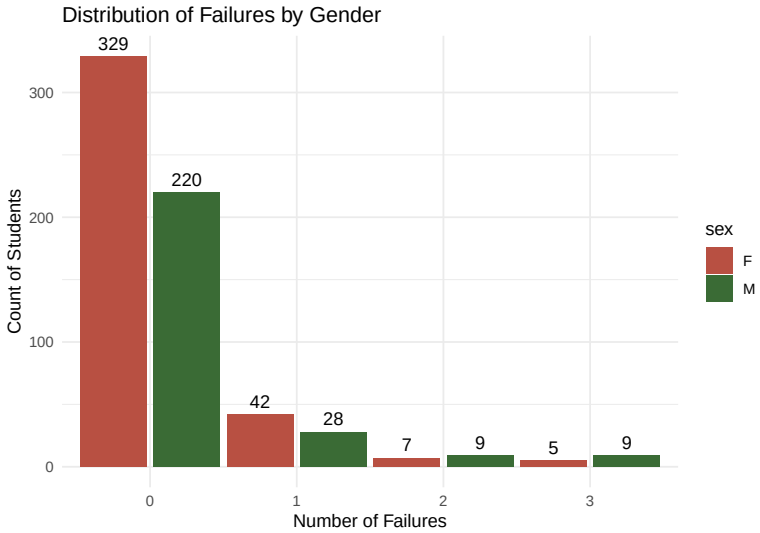
Disclaimer:
It should be noted that we used the Portuguese class dataset for our analysis in Task 1 because 382 students from the Portuguese class are also in the Maths class (395 students). Therefore, there should be 382 duplicates. In reality, only 39 out of 382 duplicates could be identified, so we decided to use only the Portuguese class dataset for the descriptive statistics analysis. This is not a problem when looking at grade-independent variables, as only 13 students who are present in the Maths classroom dataset are not also present in the Portuguese classroom dataset. A possible reason for not finding duplicates is that the survey took place at different times and students changed their opinions and information. We only used both data sets when analysing the grades (except for the trend analysis).

In order to analyse the key statistics and trends related to student demographics, behavior and grades, we assign the different variables to the three clusters. Variables are assigned to the cluster to which they contribute to.
Demographics:
age, Medu, Fedu, traveltime, famrel, school, sex, address, famsize, Pstatus, Mjob, Fjob, guardian, nursery, internet
Behavior:
studytime, freetime, goout, Dalc, Walc, health, absences, schoolsup, famsup, paid, activities, romantic, higher
Grades:
failures, G1, G2, G3

Key statistics of the data set

First, we want to give a general overview by analysing and visualising interesting categorical variables of the dataset with the help of different graphs. In addition, we will provide insights by visualising the numerical variables in the form of box plots to get a deeper insight into the data set. We then look at interesting assumptions and findings from the first parts of the analysis using a correlation heatmap. This allows us to highlight trends within the data in terms of student demographics, behavior and grades.

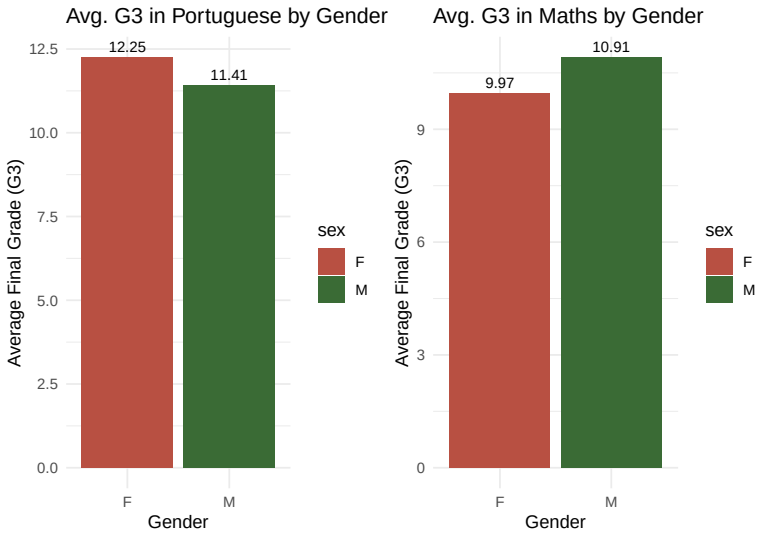
Failure Distribution by Gender:



A total of 662 students were interviewed. 13 students were not included in the analysis, as explained above. In the overall sample of 649 students, there are 383 female and 266 male students. Looking at the distribution of failures by gender, we can see that most students, regardless of gender (549 out of 649), have never failed a course. Here the proportion of female students is higher than the proportion of male students, which is logical as there are more female students than male students overall. It is interesting to note that although the proportion of female students in the dataset is higher, more male students have already failed 2 or more

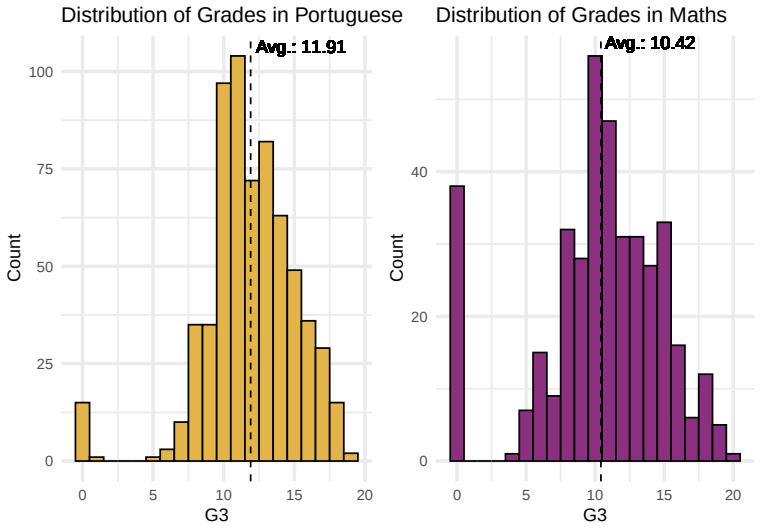
classes. It is therefore interesting to see whether male students actually do worse at school than female students.

Average Grades by Gender:



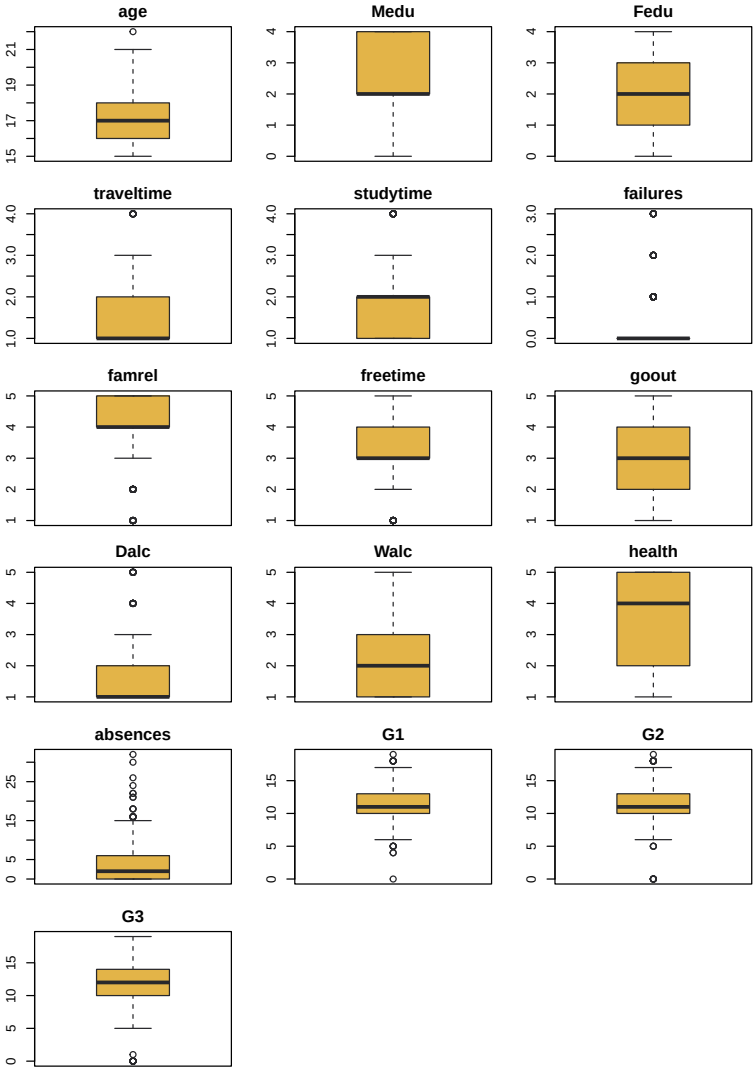
We can see that the female students are actually better on average in Portuguese with an average G3 of 12.3 compared to the male students with an average G3 of 11.4. On the other hand, male students are on average better in Maths with an average G3 of 10.9 compared to female students with an average G3 of 10. Looking at this graph, we think that the stereotype that girls are better at languages and boys at maths might be true. However, in order to make a valid statement about this stereotype, we would need to carry out further statistical tests.

Distribution of Grades by Class:



Looking at the distribution of the final grades in both classes, we can generally see that the final grades are normally distributed around the mean, except for the students who have 0 points in their final grade. It is interesting to note that the number of students with 0 points is much higher in the Maths class than in the Portuguese class, even though there are more students in the Portuguese class (649) than in the Maths class (395). We can also see that the average final grade of the students is higher in the Portuguese class (11.9) than in the Maths class (10.4).

Analysis **of numerical variables by using box plots:**



Some interesting statistics can be seen by looking at the box plots and the summary values of the numerical variables:

Age: The average age of a student is 16.7 years. Half of the students are aged 17 or over and half are aged 17 or under. The age of the students ranges from 15 to 22 years. The students who are 22 years old are outliers in this data set. It might be interesting to see if there is a correlation between a student's age and their grades, as well as the number of previous classes they have failed, since students who have failed a class are usually the oldest in the class.

Mother's & Father's education: The average education of both parents is close to 2.5. This means that the average parent has between 5th and 9th grade education. Interestingly, 75% of all mothers have an education level equal to or higher than 2, but only 50% of all fathers have an education level equal to or higher than 2. Supported by the means, we see that the education of the mothers is slightly higher than the education of the fathers. It would be interesting to see whether the parents' education has a positive influence on the students' grades.

Study time: The average study time is 1.9, which means that the average student studies either less than 2 or between 2 and 5 hours per week. 75% of students study 2 to 5 hours or less per week. It would be interesting to see the correlation between the amount of time students spend studying and their grades, as we assume a high correlation here.

Family relationship: The average quality of family relationships is 3.9, which indicates a good relationship

with the student's family. Intuitively, we would assume that the family relationship plays a crucial role in a student's performance at school.
Free time: The average amount of free time after school is 3.2. This means that the average student has a moderate amount of free time after school. 75% of all students ranked their free time between 3 and 4.
Going out: The average number of times students go out with friends is 3.2. This means that the average student goes out with friends a moderate amount. It could be interesting to see if a lower amount of going out has a positive effect on a student's grades and a higher amount of going out has a negative effect on a student's grades. Another assumption is that a student who goes out more has a better social life and therefore a better relationship with their family.
Workday alcohol consumption: The average alcohol consumption per workday is 1.5. This means that the average student doesn't drink much alcohol on a working day. Students who drink more than 3 on a working day are outliers. It might be interesting to see how the alcohol consumption is correlated with the final grade and with the student's behavior.
Weekend alcohol consumption: The average weekend alcohol consumption is 2.3. This means that the average student drinks more alcohol at the weekend than on a workday. In contrast to weekday drinking, 25% of students have a weekend drinking score of 3 or higher. Again, it is interesting to see the correlation between alcohol consumption and grades. And also the correlation with student behavior.
Current health status: The average current health status is 3.5. This is a good indicator that the average student feels healthy. It would be interesting to know whether a higher health status has a positive effect on grades and whether health status is influenced by alcohol consumption.
Number of school absences: The average number of absences is 3.7. This means that the average student is absent for 3 to 4 days per school year. As we feel that this is a low number of absences, we don't expect the number of absences to have much influence on the other variables except health.
Final Grade: The average final grade (G3) is 11.9. Interestingly, the average grade seems to improve slightly as the school year progresses.

Interesting connections within and between the demographic, behavioural and grades clusters, which are reviewed using a correlation heat map:

Behavior:
- Workday/weekend alcohol consumption and other behavior
- Health and workday/weekend alcohol consumption
- Absences and health

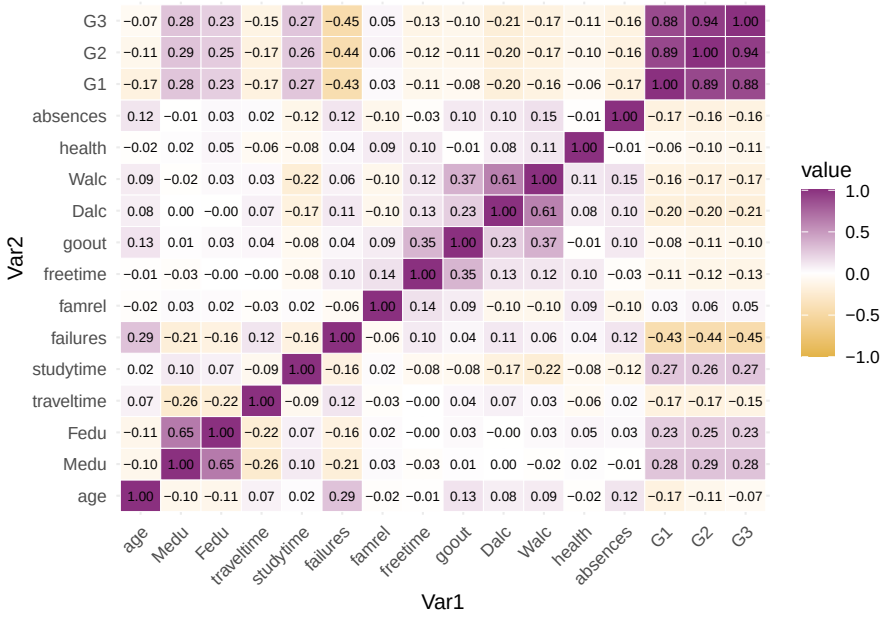
Behavior & Grades:
- Study time and G3
- Workday/weekend alcohol consumption and G3
- Health and G3
- Going out and G3

Demographics & Behavior:
- Going out and family relationship

Demographics & Grades:
- Age and G3
- Age and failures
- Educational level of parents and G3
- Family relationship and G3

Trends in the data set

In this heatmap we have used the Spearman correlation coefficient because we are analysing ordinal scaled variables. The correlations given here do not necessarily show a causal relationship between the variables, even if there is a relatively high correlation. Further research would be needed to prove relationships between the variables. Here we simply want to show possible relationships that we have identified as interesting in the section above.



Behavior:
Workday/weekend alcohol consumption and other behavior: The only interesting associations with other behavioral variables are a weak negative correlation with study time and a moderate positive correlation with goout.
Health and workday/weekend alcohol consumption: Surprisingly, these variables are only very weakly correlated, and it is also surprising that the correlation is positive. We would have expected a stronger negative correlation here, as alcohol is harmful to the human body.
Absences and health: These variables are almost completely uncorrelated. This is surprising, as students are usually absent from school for health reasons.

Behavior & Grades:
Study time and G3: There is a low moderate positive correlation between the two variables. We expected a stronger correlation here because we believe there is a causal relationship between the variables.
Workday/weekend alcohol consumption and G3: The correlations are weakly negative, which does not indicate a strong relationship.
Health and G3: Again, we would have expected a stronger correlation, but the variables do not seem to have a strong association.
Going out and G3: For this applies the same as above. The weak negative correlation does not indicate a significant connection between these variables.

Demographics & Behavior:
Going out and family relationship: We find only a very weak positive correlation of 0.09 between the two variables. We therefore assume that there is no significant relationship between the amount of going out and a student's family relationship.

Demographics & Grades:
Age and G3: Interestingly, the correlation between a student's age and final grade (G3) is weakly negative. However, as the correlation is very weak, we do not assume that there is a relationship between the two variables.
Age and failures: There is a moderate positive correlation of 0.32 between the variables. There could be a causal relationship, which would be logical because students who have to repeat a class are usually older than their new classmates.
Educational level of parents and G3: It can be seen that both (Medu and Fedu) are weakly positively correlated with a student's final grade (0.24 and 0.21). Again, the correlation is low, so we assume that there is only a weak relationship between these variables.
Mother's Educational Level and Father's Educational Level: We can see a moderately strong correlation. This

could mean that parents often have a similar level of education.
Family relationship and G3: Surprisingly, the correlation between family relationship and final grade is almost 0. This is interesting because we assumed that a very good/very bad relationship with one's family would have a strong positive/negative association with the student's performance at school.

Most surprising findings:
We think there are some surprises in the data. For example, it is interesting that alcohol consumption does not seem to affect students' health. It is also surprising that the health situation is not correlated with the number of days of absence. We would have expected a higher correlation, because our initial thought was that absenteeism is usually due to illness. One explanation for the low correlation could be that people are usually absent because of minor illnesses and not because of serious health problems. A less surprising but very interesting finding is that this dataset fulfils the stereotype that boys are better at maths and girls at languages (Portuguese). This hypothesis would have to be statistically tested to show statistical significance.

Task 2

In this task the goal is to develop a model that predicts the students performance.

Approach

Our approach for this task was to first program a modular machine learning pipeline which has the following features:

- **Data loading** The pipeline loads the two datasets, creates a "is_mat" and "is_por" column and merges the two datasets into one.
- **Data Preprocessing:** The pipeline preprocesses the data by removing the "G1" and "G2" columns because our target variable is the "G3" variable. One hot encoding the categorical variables is automatically done by the train function of the caret package and therefore not needed here.
- **Model Training:** The pipeline trains a linear regression model. It uses a parameter that defines which kind of training method should be used. By default the training function "lm" from caret is used which includes a cross validation that is set to 5 by default.
- **Model Evaluation:** The pipeline evaluates the model by calculating the RMSE, MAE, MSE and the R2 score. Additionally more model selection criteria that are mentioned in the lectures slides are used for the evaluation. Namely the adjusted R2 score, the AIC and the BIC. Finally the pipeline is capable of creating a plot comparing the metrics of different models with each other.

This approach makes our results reproducible and allows for a consistent evaluation of the models. Additionally it allows us to easily compare different models with each other.

In the following section we will train the models. By looking at the p-values of the variables we will decide which variables to remove from the model. Following that we will explain why a non linear-regression model doesn't fit our data and in the section after, we will compare the trained models with each other.

Training

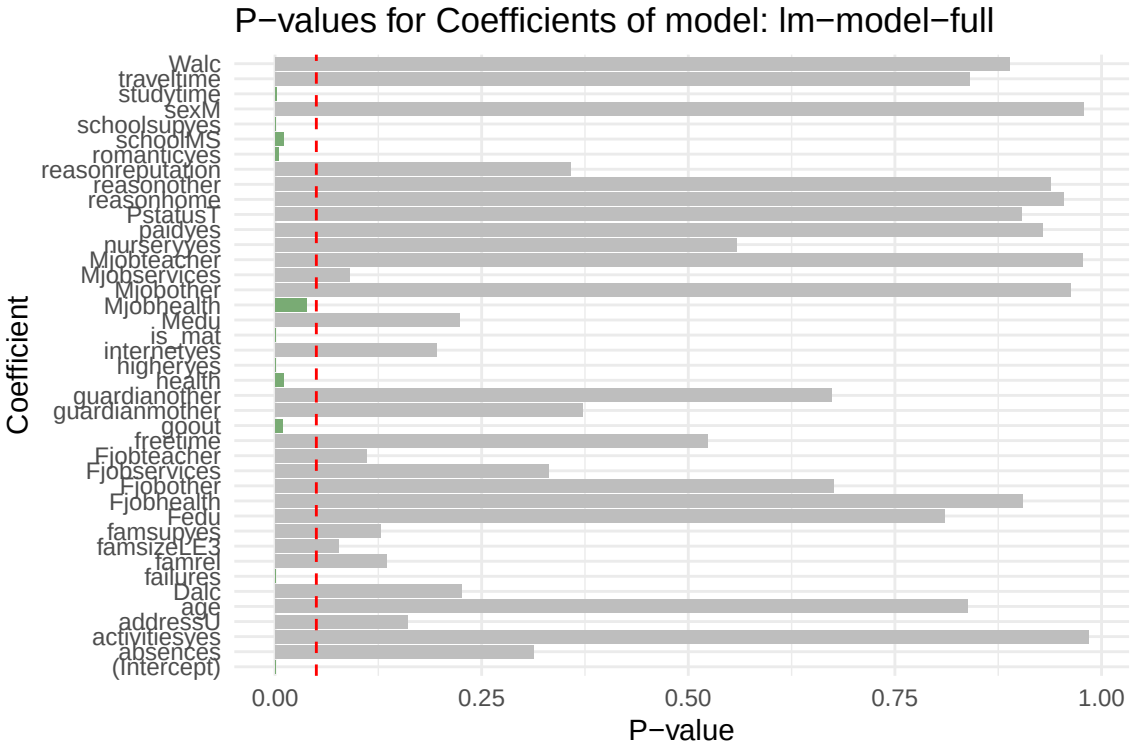
In this section we will be training different models using the pipeline structure mentioned before. The goal is to create a model with great performance in predicting the performance of a students final grade (G3). In the following

First of all we will train a linear regression model using all the columns of the dataset. A quick summary about the results from the model can be seen below.

```
# run a pipeline including all columns but G1 and G2
lm.full <- run_pipeline_cv(data=load_problemset1_data(), random_seed = 1996)
```

```
## RMSE: 3.438325
## MAE: 2.498063
## R-squared: 0.2204252
## MSE: 11.82208
## BIC: 5727.541
## AIC: 5519.607
## Adjusted R-squared: 0.2555508
```

In summary we can see that the model has some explanatory power which is indicated by the r-squared and adjusted r-squared but the errors (RMSE, MAE, MSE) are relatively high. We will have a more in depth comparison at the bottom of this section where we will compare the results of all the models we trained.



The plot above shows the p-values of the coefficients of the linear regression model. The p-values are a measure of the significance of the coefficients. The lower the p-value the more significant is the coefficient. The plot shows that most of the features are not significant.

One additional finding we had, was that using different kinds of random seeds drastically influenced the p-values of the model. To solve this issue we applied cross validation which stabilized the values.

Features with a significance level of 0.05 or lower are:

- **studytime**: Which is a feature we expected to be of importance for the final grade. It indicates the weekly study time of the student and therefore is expected to impact the final grade.
- **schoolsup**: This feature also makes sense, as it indicates whether the student received extra educational support or not.
- **school**: This is a interesting and unexpected find. The significance of the school feature seems to suggest to have an impact on the final grade. Therefore the performance of the schools seem to vary.
- **romantic**: The significance of this feature is less than the other features mentioned. However there is a indication that the romantic relationship of the student is a indicator for the final grade.
- **Mjob**: This is another unexpected feature to have a significance. Especially whether the father is working in a health related field or not seems to impact the prediction performance of the model

regarding the final grade.

- **is_mat**: Whether the student is enrolled in the mathematics course or not seems to be a good indicator for the final grade. This is a interesting find as it would suggest that a good portion of variance is explained by the course the student is enrolled in. This could be due to the fact that the mathematics course is more difficult than the portuguese course.
- **higher**: Whether the student wants to take higher education or not is also a good indicator for the final grade. Logically this makes sense as students that want to take higher education are more likely to be motivated to get good grades.
- **health**: The health of the student is also a good indicator for the final grade. This is also a logical find as students that are in good health are more likely to have a higher final grade.
- **goout**: The going out habits of the student is also a good indicator for the final grade. We already expected this feature to have an impact on the performance grades of the students.
- **failures**: The number of past class failures is also a good indicator for the final grade. This is also a logical find as students that failed in the past are more likely to have a lower final grade.

In the next step we will remove the columns that have a p-value > 0.05 and run the pipeline again.

```
columns_to_keep = c("school", "studytime", "failures", "schoolsup",
                    "higher", "romantic", "health", "is_mat", "G3",
                    "Mjob", "goout")
lm.lower.05 <- run_pipeline_cv(data=load_problemset1_data(),
                               columns_to_keep = columns_to_keep)
```

```
## RMSE: 3.384669
## MAE: 2.45573
## R-squared: 0.2395161
## MSE: 11.45599
## BIC: 5573.921
## AIC: 5499.659
## Adjusted R-squared: 0.2510308
```

Having a quick look at the results we can see similar results to the full model. As mentioned already, we will have a more in depth look comparing all models following below.

Testing out the performance of single features In this section we wanted to try out how models are performing when only one feature is used. We will use the same pipeline as before but only use one feature at a time. The results will be interpreted in the evaluation section below.

```
lm.studytime <- run_pipeline_cv(show_output=FALSE,
                                data=load_problemset1_data(),
                                columns_to_keep = c("studytime", "G3"))
lm.higher <- run_pipeline_cv(show_output=FALSE,
                             data=load_problemset1_data(),
                             columns_to_keep = c("higher", "G3"))
lm.failures <- run_pipeline_cv(show_output=FALSE,
                               data=load_problemset1_data(),
                               columns_to_keep = c("failures", "G3"))
lm.schoolsup <- run_pipeline_cv(show_output=FALSE,
                                data=load_problemset1_data(),
                                columns_to_keep = c("schoolsup", "G3"))
lm.is_mat <- run_pipeline_cv(show_output=FALSE,
                             data=load_problemset1_data(),
                             columns_to_keep = c("is_mat", "G3"))
```

About non linear regression The two models above were linear regression models. Another possibility for a regression model that was mentioned in the lecture is a non linear regression model. But looking at

our data, we can see that we have no continuous variables. All features are either binary or categorical. As mentioned in the lecture, looking for a non linear relation in a binary case doesn't make much sense. Because categorically encoded features are also binary encoded through one-hot encoding we are not expecting a non linear regression model to perform better than a linear regression model.

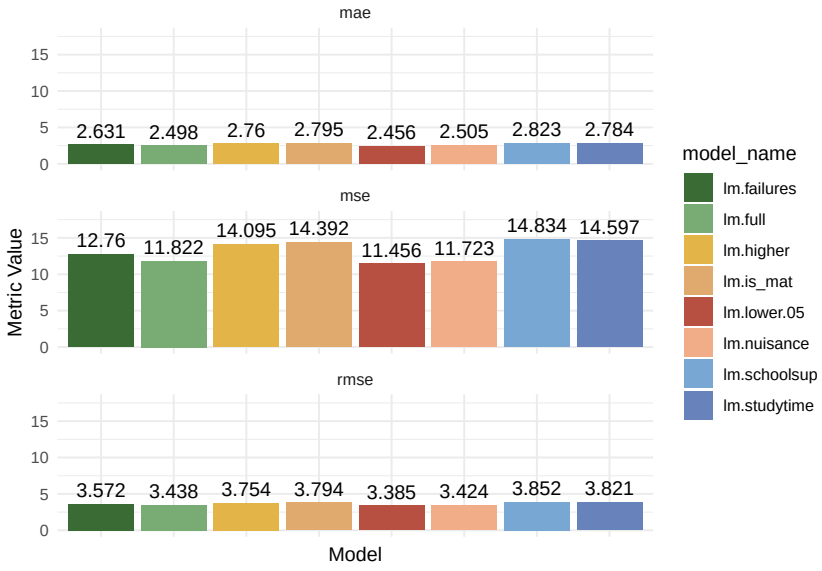
Nuisance variables Next we wanted to try out the effects of nuisance variables on the model. For this we added 20 random variables to the data and ran the pipeline.

```
data <- load_problemset1_data()
seed = 1996
amount_to_add = 20
columns_to_keep_nuisance <- columns_to_keep
# we add nuisance variables to the data
for (i in 1:amount_to_add) {
  set.seed(seed + i)

  # Create a new column name dynamically
  new_col_name <- paste("Nuisance", i, sep = ".")

  # Add the new column with random values to the existing data
  data[[new_col_name]] <- rnorm(nrow(data))
  columns_to_keep_nuisance <- c(columns_to_keep_nuisance, new_col_name)
}
lm.nuisance <- run_pipeline_cv(data=data, columns_to_keep = columns_to_keep_nuisance)
```

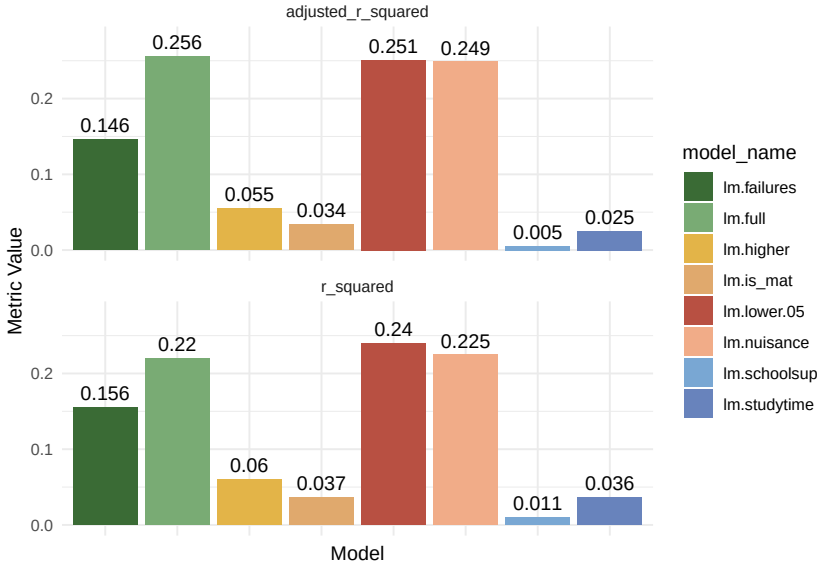
```
## RMSE: 3.423898
## MAE: 2.504734
## R-squared: 0.2246636
## MSE: 11.72308
## BIC: 5695.932
## AIC: 5522.654
## Adjusted R-squared: 0.2485399
```

Evaluation

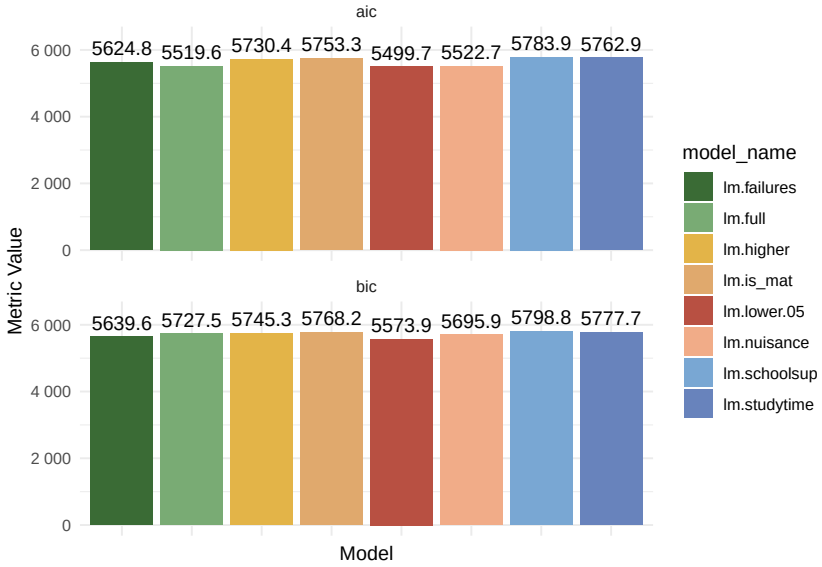
In this section we will evaluate the models that were trained in the training section.



Looking at the first plot we can see the metrics MAE, MSE and RMSE. For all those metrics the lower the value the better the model. We can see that for each of the models above the values are relatively similar. The full model (lm.full) which included all the features performed almost as good as the model with only features with a significant p-value (lm.lower.05). Even the models which only contained single features had a comparable performance when looking at the scores above.



When looking at the plot showing the R-squared and adjusted R-squared values the distribution changes. Here the higher the value the better the model. The best performing model is the lm.lower.05 model which only contains features with a significant p-value. Still the performance is only slightly better than the full model (lm.full) which contains all the features. The nuisance model contains additional features that are randomly generated. Still the performance stays almost the same suggesting a robust model. The single feature models perform worse than the other models. But interestingly the lm.failures model which only contains the failures feature has a relatively good performance, even though it contains only one value. This suggests that the past failures of a student has a great impact in the prediction of the final grade.



The last plot shows the AIC and BIC values. The lower the value the better the model. Here we can see that the lm.lower.05 model has the lowest AIC and BIC but only by a small margin. Overall the values for every model is very similar. Comparing the AIC and BIC values to the R-squared and adjusted R-squared values

we can see the same trend. The lm.lower.05 is the best performing model overall while the lm.failures model has a good performance for a single feature model.

Performance prediction on fictive data

In this section we will create two student profiles and predict their final grade using the best performing model (lm.lower.05).

Looking at the results from above we can see which of the features have a high impact on the final grade. Furthermore we created four single feature models which showed that the most important feature is the failures feature. To confirm this we will create two nearly identical student profiles. The only difference between the two profiles is that one student has failed in the past while the other student has not. We hypothesize that the student who has failed in the past will have a significantly lower final grade predicted than the other nearly identical student.

In the following we describe the student profiles:

Student Best: This student we expect to have the best performance of all the student profiles we create. We try to give him the best possible conditions for a good grade. By the results we learned in this task the best school is "MS". Obviously studytime should be high so we set it to "4". Failures should have the highest impact on the final grade. This student should have "0" past failures. We also set schoolsup to "no" as this should indicate that the student doesn't need additional support from the school. Higher should be set to "yes" as this should indicate that the student wants to pursue higher eduction and therefore give him motivation for getting better grades. Romantic should be set to "no". This was the most interesting for us as we thought that the relationship should have no impact on the final grade. The health of the student should be good so we set it to "4". The grades of the math course are lower than the grades of the portuguese course so we set is_mat to "0". The goout value should be low as this should indicate that the student is not distracted by going out with friends.

Student Best Alternative: This student is an alternative to the best student with only one difference. This student has failed in the past. With this profile we want to verify our hypothesis that the failures feature has the highest impact on the final grade. We expect this student to have a lower predicted final grade than the best student but still better than the worst student

Student Worst: This student we expect to have the worst performance of all the student profiles we create. We try to give him the worst possible conditions for a good grade. ALl the features are set to the opposite of the best student.

```
##   school studytime failures schoolsup higher romantic health is_mat     Mjob
## 1     MS         4        0        no    yes       no      4      0   health
## 2     MS         4        4        no    yes       no      4      0   health
## 3     GP         1        4        no     no      yes      1      1 services
##   goout predictions     profile_name
## 1     1   14.559765     Student Best
## 2     1    7.781357 Student Best.alt
## 3     5    2.580040    Student Worst
```

Looking at the results we can see that our hypothesis was confirmed. The student who has failed in the past has a lower predicted final grade than the best student by a large margin. Also every student profile has a prediction that we expected.

Task 3

First of all we decided to only use the data of the portuguese language course, since 382 students are in both courses and the math courses only has data for 395 students. So by choosing only the language course data we only lose the information from 13 students.

First we create a measurement for overall alcohol consumption: Oalc = Walc + Dalc We just decided to add these two values for the final measurement.

Create a basic model

First we thought about which variables are probably influencing the alcohol consumption of the students. After that we came up with the following variables we wanted to test.

Number of absences: Because this could be a measure how serious a student takes his studies, so we would expect that with a raising number of absences, the alcohol consumption increases.
How often the students go out with their friends: Since a lot of young people like to drink alcohol when going out with friends, our idea is that when students go out more often they drink more alcohol.
How much they study: This shows how much effort a student puts into his studies and again could be a sign how invested the student is. So again we would hypothesize that the alcohol consumption decreases when study time goes up.
Their gender: Drinking habits are often different between woman and man, so we would expect that male students in general drink more alcohol than female students.
The relationship with the family: Alcohol is often abused when people are frustrated. So we would expect that alcohol consumption increases when the relationship to the family is not that close.
Address: Often drinking habits differ between rural and urban area, for example because of the different possibilities when going out.
Age: We also want to include the age variable since some students a to young to drink legally. So we would expect that alcohol consumption increases with age.
Freetime: We would also expect that freetime has a effect on alcohol consumption because simply when someone has more freetime the person also has more time for drinking alcohol.
Guardian: We would expect a difference of drinking habits when the father is the guardian, compared to when the mother is the guardian. This relies on the idea that there are different drinking habits between woman and man. And since students look up to their guardian, their behavior could maybe also influence the student.

For testing this model first we have to create 4 binary variables:

- male (1, when the student is a male, else 0)
- addressurban (1, when the student lives in an urban area, else 0)
- guardianmother (1, when the mother is the guardian, else 0)
- guardianfather (1, when the father is the guardian, else 0)

When we run this model, we get the following results:

```
##
## Call:
## lm(formula = Oalc ~ absences + goout + studytime + male + famrel +
##     addressurban + guardianmother + guardianfather + age, data = dataporg)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.6646 -1.1505 -0.2458  0.8437  6.2039
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)     0.822426   1.142949   0.720 0.472056
## absences        0.050893   0.014800   3.439 0.000623 ***
## goout           0.570651   0.057345   9.951  < 2e-16 ***
## studytime      -0.232460   0.082560  -2.816 0.005018 **
## male            1.256196   0.138466   9.072  < 2e-16 ***
```

```
## famrel         -0.297624   0.070277  -4.235 2.62e-05 ***
## addressurban   -0.211298   0.145080  -1.456 0.145766
## guardianmother -0.005523   0.292294  -0.019 0.984929
## guardianfather  0.259813   0.318548   0.816 0.415025
## age             0.128414   0.058527   2.194 0.028587 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.685 on 639 degrees of freedom
## Multiple R-squared:  0.2943, Adjusted R-squared:  0.2844
## F-statistic: 29.61 on 9 and 639 DF,  p-value: < 2.2e-16
```

By looking at the p-values we can derive that only absences, goout, studytime, male & famrel are significant. Because we want to test a interaction effect which includes the address component (Hypothesis 2), we also keep the address variable and only delete the two variables referring to the guardian of the student.

Because we also got a hypothesis for an interaction effect which includes if a student wants to take higher education (Hypothesis 3) we also added a binary variable to the model which takes the value 1, if the student aims for higher education and else 0.

The final regression model than has the following results:

```
##
## Call:
## lm(formula = Dalc ~ absences + goout + studytime + male + famrel +
##     addressurban + age + higherbinary, data = dataporg)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.6757 -1.1492 -0.2485  0.7993  6.4003
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   1.21387    1.07098   1.133 0.257463
## absences      0.04878    0.01474   3.311 0.000983 ***
## goout         0.56624    0.05738   9.869  < 2e-16 ***
## studytime    -0.22578    0.08380  -2.694 0.007242 **
## male          1.26303    0.13863   9.111  < 2e-16 ***
## famrel       -0.29448    0.07034  -4.187 3.23e-05 ***
## addressurban -0.20627    0.14552  -1.417 0.156846
## age           0.11239    0.05731   1.961 0.050327 .
## higherbinary -0.08448    0.22881  -0.369 0.712099
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.688 on 640 degrees of freedom
## Multiple R-squared:  0.2914, Adjusted R-squared:  0.2825
## F-statistic: 32.89 on 8 and 640 DF,  p-value: < 2.2e-16
```

Now the age variable is not significant on the 5% level anymore, so we also deleted it from the model. So the results for the final model for testing the interaction effects than looks like this:

```
##
## Call:
## lm(formula = Dalc ~ absences + goout + studytime + male + famrel +
##     addressurban + higherbinary, data = dataporg)
##
```

```
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.5343 -1.1662 -0.2699  0.8362  6.4767
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   3.14498    0.42179   7.456 2.90e-13 ***
## absences      0.05224    0.01466   3.563 0.000394 ***
## goout         0.57702    0.05724  10.081  < 2e-16 ***
## studytime    -0.21750    0.08388  -2.593 0.009733 **
## male          1.24762    0.13872   8.994  < 2e-16 ***
## famrel       -0.29474    0.07050  -4.181 3.31e-05 ***
## addressurban -0.21166    0.14582  -1.451 0.147131
## higherbinary -0.19767    0.22190  -0.891 0.373369
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.691 on 641 degrees of freedom
## Multiple R-squared:  0.2871, Adjusted R-squared:  0.2793
## F-statistic: 36.88 on 7 and 641 DF,  p-value: < 2.2e-16
```

So our final model includes 7 variables and the value for the adjusted R-squared is 27,93%.

1. Hypothesis

The effect of going out on the alcohol consumption depends on the gender of the student. Directly speaking men tend to drink more alcohol when going out with friends compared to women. So we assume there is an interaction effect, between the goout variable and the gender variable.

Step 1 - Mean centering

Since for the gender variable the zero value is easily interpretable (just means the student is a woman), we only need to mean center the variable which indicates how often the student goes out with friends.

Step 2 - Calculating the interaction effect

Not required when using R.

Step 3 - Analyze basic model

As seen above in the basic model 5 variables had a significant effect on the drinking behavior and the value for the adjusted R-squared is 27,93%.

Step 4 - Analyze full model

When we run the same model with the interaction effect included, we get the following regression analysis:

```
##
## Call:
## lm(formula = Dalc ~ absences + gooutMC + studytime + male + famrel +
##     higherbinary + addressurban + gooutMC:male, data = dataporg)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.1211 -1.0737 -0.3379  0.8654  6.8837
##
```

```
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   4.96582    0.37986  13.073  < 2e-16 ***
## absences      0.04266    0.01445   2.952  0.00327 **
## gooutMC       0.31528    0.07375   4.275 2.20e-05 ***
## studytime    -0.25270    0.08231  -3.070  0.00223 **
## male          1.22201    0.13579   8.999  < 2e-16 ***
## famrel       -0.28495    0.06899  -4.130 4.10e-05 ***
## higherbinary -0.15206    0.21724  -0.700  0.48420
## addressurban -0.16804    0.14288  -1.176  0.23999
## gooutMC:male  0.61835    0.11338   5.454 7.05e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.655 on 640 degrees of freedom
## Multiple R-squared:  0.3188, Adjusted R-squared:  0.3102
## F-statistic: 37.43 on 8 and 640 DF,  p-value: < 2.2e-16
```

When we run the analysis, we can derive from the p-value that the interaction term between going out and living in an urban area us significant at the 0,1% level. The value for the adjusted R-squared is now 31,02%.

Step 5 - Testing the interaction

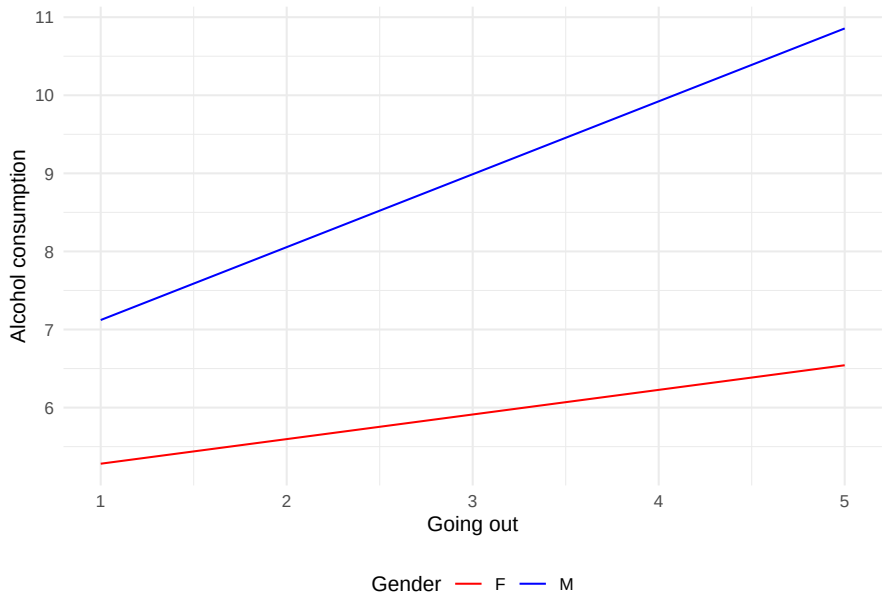
First of all we can see by looking at the p-value that the interaction term of going out and being a male is significant to the 0,1% level.

F-Test for R-squared-differences:

H0: no difference between the models
H1: models are different

```
## Analysis of Variance Table
##
## Model 1: Dalc ~ absences + gooutMC + studytime + male + famrel + higherbinary +
##     addressurban + gooutMC:male
## Model 2: Dalc ~ absences + goout + studytime + male + famrel + addressurban +
##     higherbinary
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1    640 1752.4
## 2    641 1833.8 -1   -81.446 29.745 7.046e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

This analysis leads in a F-value of 29,745. The corresponding p-value shows that there is a significant difference between the two models. This can also be seen by visualizing the interaction effect:



Interpretation

The results implicate that there is an ordinal interaction effect between gender and going out. This means that a male student drinks more alcohol when going out often than a female student. Even though male students seem to have a general higher alcohol consumption than female students, this effect is amplified when the student often goes out.

So one suggestion for policymakers could be that they should focus male students when doing a alcohol prevention program. To target them when they go out an idea could be that they create information desk at places students like to go out. This desk should be more attractive for male students, for example by giving small rewards to male students who visit the desk.

Another suggestion could be an information event for male students at which they are educated about drinking habits when going out and responsable drinking in general. This could maybe also include something like a alcohol tracker which is given to the students so that they are able to monitor their own alcohol consumption when going out. This could maybe help them to decide whether to get another drink when going out or not.

2. Hypothesis

The 2nd hypothesis we had is that the effect of going out on alcohol consumption depends on whether the student lives in a urban or rural area. We would expect that in rural areas the effect of going out on alcohol consumption is lower, simply because of the different nightlife options, which should be higher in urban areas.

Step 1 - Mean centering

For that we also only have to mean center the variable “goout” which we already done when testing the 1. Hypothesis.

Step 2 - Computing the interaction term

Not necessary when using R.

Step 3 - Analyze basic model

As seen above in the basic model 5 variables had a significant effect on the drinking behavior and the value for the adjusted R-squared is 27,93%.

Step 4 - Analyze full model

```
##
## Call:
## lm(formula = Dalc ~ absences + gooutMC + studytime + male + famrel +
##     addressurban + higherbinary + gooutMC:addressurban, data = dataporg)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.2652 -1.1853 -0.2651  0.8391  6.5583
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)            5.01670    0.38785  12.935  < 2e-16 ***
## absences               0.05154    0.01464   3.522 0.000459 ***
## gooutMC                0.41719    0.10065   4.145 3.86e-05 ***
## studytime             -0.22388    0.08377  -2.673 0.007720 **
## male                   1.25730    0.13851   9.077  < 2e-16 ***
## famrel                -0.30237    0.07046  -4.292 2.05e-05 ***
## addressurban          -0.20785    0.14553  -1.428 0.153697
## higherbinary          -0.19508    0.22143  -0.881 0.378663
## gooutMC:addressurban   0.23449    0.12159   1.929 0.054226 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.688 on 640 degrees of freedom
## Multiple R-squared:  0.2912, Adjusted R-squared:  0.2824
## F-statistic: 32.87 on 8 and 640 DF,  p-value: < 2.2e-16
```

When we run the analysis, we can derive from the p-value that the interaction term between going out and living in an urban area is only significant at the 10% level. The value for the adjusted R-squared is now 28,24%.

Step 5 - Testing the interaction

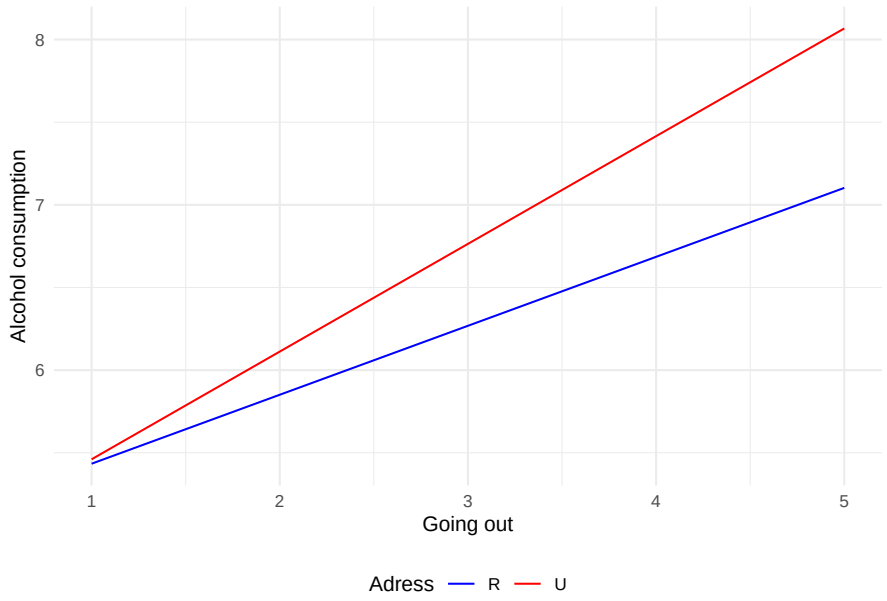
First of all we can see by looking at the p-value that the interaction term of going out and living in an urban area is only significant to the 10% level.

F-Test for R-squared-differences:

H0: no difference between the models
H1: models are different

```
## Analysis of Variance Table
##
## Model 1: Dalc ~ absences + gooutMC + studytime + male + famrel + addressurban +
##     higherbinary + gooutMC:addressurban
## Model 2: Dalc ~ absences + goout + studytime + male + famrel + addressurban +
##     higherbinary
##   Res.Df    RSS Df Sum of Sq      F  Pr(>F)
## 1    640 1823.3
## 2    641 1833.8 -1   -10.596 3.7194 0.05423 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

By looking at the p-value we cant reject H0 at the 5% level so this test would suggest that the interaction term does not bring significant value to the regression model.



Interpretation

In this case the interaction effect is not significant, so we would not derive any suggestions for policymakers based on this result.

3. Hypothesis

The effect of studytime on alcohol consumption is dependent on the fact if the student wants to take higher education. Our assumption would be that if a student wants to take higher education the more he or she studies the less alcohol he or she drinks.

Step 1 - Mean centering

We need to mean center the variable for studytime. For the variable which expresses if the student wants to take higher education, we don't need mean centering, since there is a natural interpretation when the variable is 0 (student does not want higher education).

Step 2 - Computing the interaction term

Not necessary when using R.

Step 3 - Analyzing the basic model

As seen above in the basic model 5 variables had a significant effect on the drinking behavior and the value for the adjusted R-squared is 27,93%.

Step 4 - Analyze the full model

Now the interaction term between higher education and studytime is added.

```
##
## Call:
## lm(formula = Dalc ~ absences + goout + studytimeMC + male + famrel +
##     higherbinary + addressurban + higherbinary:studytimeMC, data = dataporg)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
```

```
## -4.5669 -1.1580 -0.2472  0.8658  6.4698
##
## Coefficients:
##                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)                 2.95382    0.40873   7.227 1.41e-12 ***
## absences                    0.05105    0.01463   3.489 0.000518 ***
## goout                       0.57633    0.05709  10.096  < 2e-16 ***
## studytimeMC                 0.34009    0.27757   1.225 0.220930
## male                        1.24938    0.13835   9.031  < 2e-16 ***
## famrel                     -0.28717    0.07040  -4.079 5.09e-05 ***
## higherbinary               -0.45103    0.25187  -1.791 0.073808 .
## addressurban               -0.20722    0.14545  -1.425 0.154722
## studytimeMC:higherbinary -0.61097    0.29000  -2.107 0.035524 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.687 on 640 degrees of freedom
## Multiple R-squared:  0.292,  Adjusted R-squared:  0.2832
## F-statistic:    33 on 8 and 640 DF,  p-value: < 2.2e-16
```

The value of the adjusted R-squared increased to 28,32%. The interaction term is significant to the 5% level. Interestingly now the studytime isnt significant anymore. But the significance of the variable which represents if the student wants to take higher education is now significant at the 10% level.

Step 5 - Testing the interaction

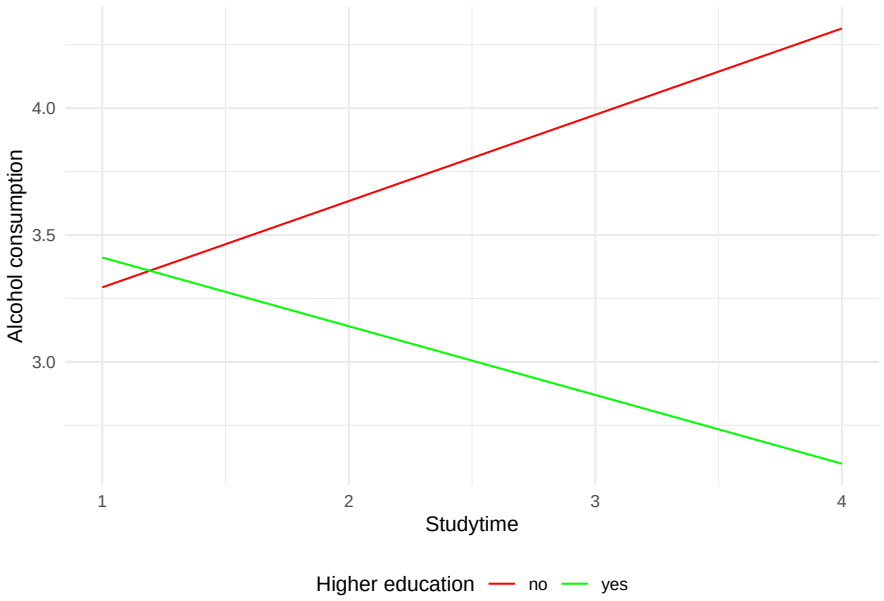
First of all we can see by looking at the p-value that the interaction term of studytime and if the student wants to take higher education is significant to the 5% level.

F-Test for R-squared-differences:

H0: no difference between the models
H1: models are different

```
## Analysis of Variance Table
##
## Model 1: Dalc ~ absences + goout + studytimeMC + male + famrel + higherbinary +
##     addressurban + higherbinary:studytimeMC
## Model 2: Dalc ~ absences + goout + studytime + male + famrel + addressurban +
##     higherbinary
##   Res.Df    RSS Df Sum of Sq      F  Pr(>F)
## 1    640 1821.2
## 2    641 1833.8 -1   -12.631 4.4386 0.03552 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

By looking at the p-value we can derive that the models are different from each other at the 5% significance level. So in this case we can reject H0. In other words the inclusion of the interaction term added value to the regression model.



Interpretation

The results imply that there is a significant disordinal interaction effect between studytime and whether a student wants higher education on alcohol consumption. Specifically the alcohol consumption goes up with studytime, when a student does not want to take higher education. But if the students aims for higher education the alcohol consumption goes down with studytime.

A reason behind that could be that students who want to take higher education set that as their goal. So if they study a lot that could maybe be a sign that they are really focused on this goal and maybe realized that alcohol consumption could lower their chances of higher education since then they have less time for studying and have more days where their brain capacity is restricted because of the alcohol intake the day before. This maybe could hurt their grade which is important to get into higher education. Students who don't aim for higher education don't have that goal. So one possible explanation why their alcohol consumption goes up with studytime could be that they want to reward themselves for the time they put into studying by going out with friends and drink alcohol.

A suggestion for policymakers could be to target the students who don't aim for higher education. So they could create a alcohol prevention program which focuses to prevent students from using alcohol as a reward. Another suggestion could be to target the students who aim for higher education and make them clear that their current grades are important if they want to have the chance for higher education. This could maybe then increase studytime and therefore decrease alcohol consumption.