Regression Project

**Introduction:**

In the era of data-driven decision-making, regression analysis is a powerful tool for indetifying patterns and relationships within datasets and enables us to make informed predictions of complex models. Through this project, we will be focusing on the use of regression analysis to predict a target variable from the dataset at hand. The dataset we have at hand approaches student performance of two Portuguese Schools, with two subsets: mathematics and Portuguese language. For the sake of our investigation, we will use the mathematics subset to model our target variable G3, the final year grade, using different predictors in the dataset. Our objective is to develop a robust predictive model that can discern patterns and trends within the data, ultimately contributing valuable insights to the field of education. We will adhere to established guidelines and criteria of regression analysis, to ensure the efficiecny and effective of our model. This will be achieved by comparing different combinations of the regression model formulated using different variable selection methods, and then formulating a conclusion with the best model. Through this project, we aim not only to build an accurate regression model but also to unravel the intricate relationships within the data, providing the schools with valuable insights that will help them in improving their academics.

**Description of Dataset:**

As mentioned earlier, the dataset at hand focuses on various aspects that are related to mathematics education, which encompasses several data points relating to student’s performance and demographics. The dataset will help us in finding insights into what factors lead to academic achievement in the field of mathematics. The dataset is composed of several variables which can be grouped into two groups. The first group represents our target variable which is the variable G3, the student’s final year grade. On the other hand, all the other variables in the dataset would be considered as our potential predictors, 32 variables, which will be used during our comparison process to find the best combination of these predictors in predicting the target variable. The dataset also contains two variables, G1 and G2, which represent the grades for both semesters. The target variable has strong correlation with these variables, making it easier to predict G3 using them. However, for our analysis, we will be excluding the two variables as the G3 variable is a more insightful target variable and will perhaps provide us with better results. Therefore, there is a total of 30 predictors that we will be using to define the best model to predict the final year grade, G3. Moreoever, the are 395 observations in the dataset which might limit our regression analysis.

**Data Preprocessing:**

As we glance through the dataset, we can see that it is highly made up of different categorical variables such as sex, school, famsize, etc. To correctly interpret the effects of these categorical variables, we need to handle them to be able to use them for our analysis. This means that we need to change all of these variable from type “chr” to “fctr” using the as.factor() method in r. This would lead to the presence of more coefficients in the model that account for each category. Moreover, we will check that there are no null values and remove any row with null values. Finally, as G1 and G2 will not be used in our analysis, we will be dropping them from the dataset. At this point, we have a clean dataset that we can use for our regression analysis.

**Analysis**

Model Definition:

To start our analysis phase, we should first start by defining our starting model predicting the target variable, G3. We will be defining a multiple linear regression model for predicting G3 based on the various predictors. Therefore, we would need to define our null and full models.

***Null Model:***

*G3 = 1*

***Full model***

*G3 = β0​ + β1​⋅Age + β2​⋅Gender……. β30​⋅Attendance + ε*

These models will represent the basis of our analysis process as these are multiple linear regression models that do not consider the different factors that can affect the model such as the interaction between the predictors. Hence, the model will be further refined through our investigation based on the characteristics of the dataset and the nature of the relationships between the predictors.

Variable Selection Process

To perfectly select the predictors that best predict the target variable, we will be applying several selection methods to select these variables. As we are comparing different models and trying to find the best model, we can apply the Akaike information criterion (AIC) to estimate the quality of each model relative to the other models. We can also use the Bayesian information criterion (BIC) as a measure for comparison between the different models. To apply these criteria, we will use backward elimination which will start at the full model and eliminate predictors until we reach the null model. After we get the two models from BIC and AIC, we would use Analysis of Variance (ANOVA) to test which of the two models would be the better choice to predict the target variable.

Interactions:

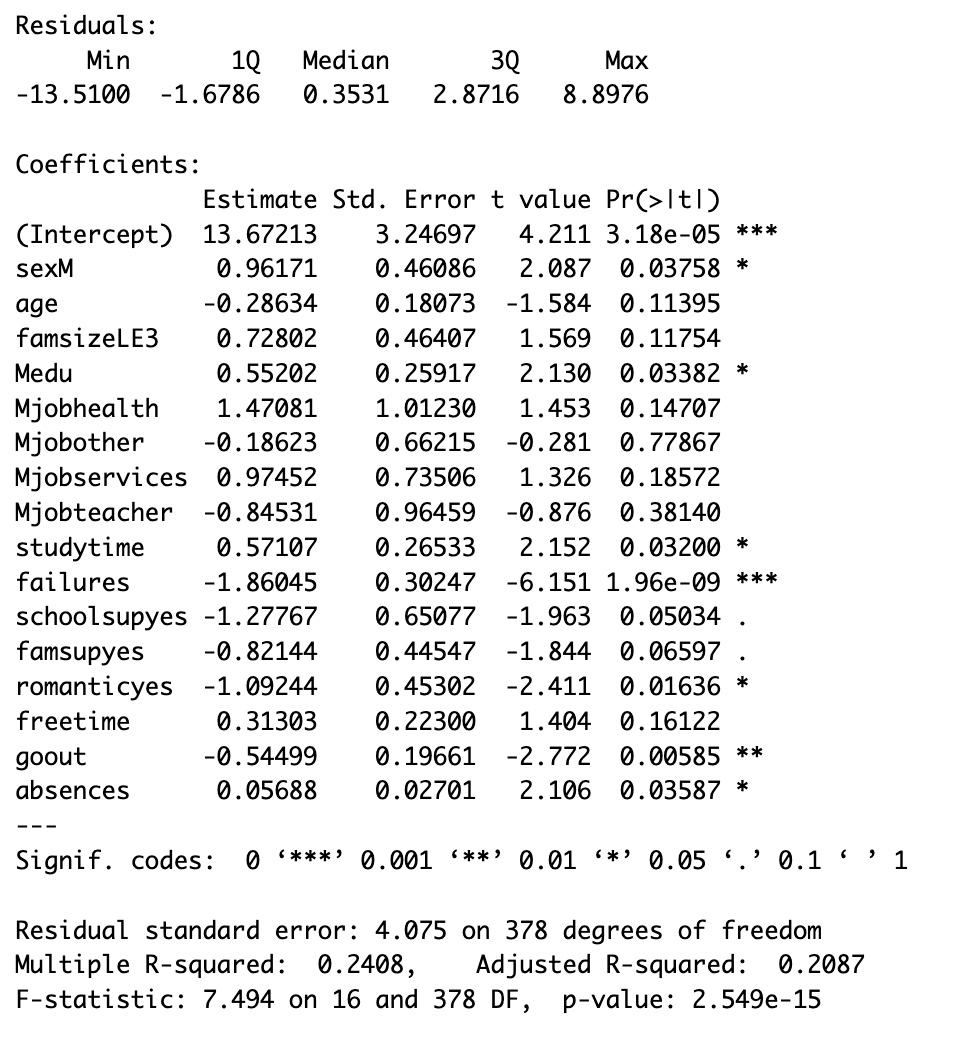
In the analysis of the mathematics dataset, interactions between variables need to be carefully considered during the model development process. Interactions occur when the effect of one predictor on the target variable is dependent the effect of another predictor. This means that we would have coefficients, βn, that would account for the effect of X1 by X2. When we find the best model through our variable selection process, we would test to see the interactions between the different predictors that are part of our model. If any of the interactions are significant, we would add them to our model and compare them to the main model and see if the effect is significant.

Checking model assumptions:

After finding the best model that predicts the target variable using certain predictors, we need to check that the model follows the regression assumptions linearity, equal variance and normality. These assumptions can be tested using a residual plot and a normal q-q plot. If any of the assumptions are violated, we can apply certain transformation to our target variable such as log-transformation to adhere to any violations.

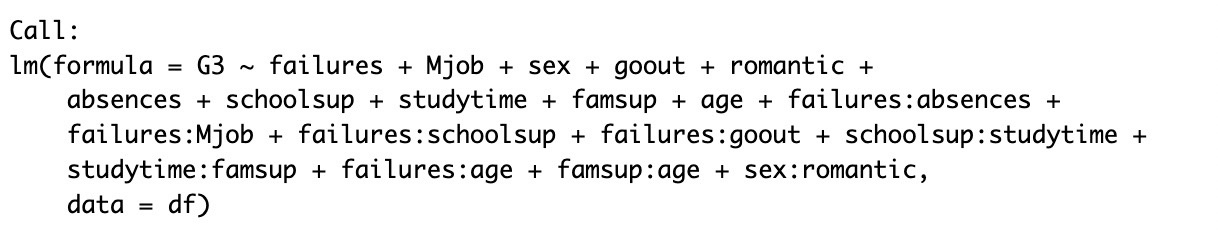
**Summary of Output and Results:**

To start off, we tested to see if any of the predictors are significant using an ANOVA table which yielded a low p-value(3.317e10) suggesting significance from some of the predictors.Then we applied backward elimination from the full model until we reached the null model and compared both the BIC and AIC scores. The model with the lowest AIC score was (formula = G3 ~ sex + age + famsize + Medu + Mjob + studytime failures + schoolsup + famsup + romantic + freetime + goout + absences), while the model with the lowest BIC score was (formula = G3 ~Medu + failures). When comparing the two using an ANOVA table, we see that the p-value is less than 0.05 which means that the presence of more predictors is significant. Moreover, the adjusted R-squared is higher for the AIC model which means that it better fits the data. Therefore, we now have the AIC model which has the following coefficients:



As we can see, this is the best model that we got from the AIC and BIC tests which has yielded an adjusted R-squared of 0.2087 which is considerably low. Perhaps we can see if the interactions between these predictors would improve the model.

To identify the interactions between the models, we will use a new formula for the model (formula = G3 ~ sex \* age \* famsize \* Medu \* Mjob \* studytime failures \* schoolsup \* famsup \* romantic \* freetime \* goout \* absences) which considers the interaction between the predictors. We will then use the bidirectional method to compare the AIC score of the models as it is much quicker and efficient that the backward method especially with interactions. Through this process, we get a new model which has the lowest AIC score of 1100.39, which is generally considered a high score, with the following formula:



This model is considered the best model that we were able to find as it has the lowest AIC score and the highest adjusted R-squared at 0.2752. Final step is to check the regression assumptions. This can be done by looking at the residual and normal q-q plots below:

A graph of a normal q-q plot

Description automatically generatedA graph showing residual plot

Description automatically generated

Looking at both plots we can see that all the assumptions are violated by the model. This means that we need to apply transformations to address these violations. We have applied the cubic root transformation of the target variable G3 which have yielded the following plots:

A graph of a normal q-q plot

Description automatically generatedA graph showing the residual plot

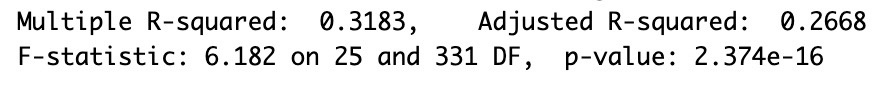
Description automatically generated

As we can see from the plots, the linearity assumption holds as shown on the residual plot. Also, the BP test has yielded a value greater than 0.05 indicating that the assumption of equal variance also holds. Finally, we can see from the normal q-q plot that the normality assumption holds as well.

Therefore, we now have our final model which is summarized below:

A screenshot of a computer

Description automatically generated



From the above summary, we can outline effect of several individual variables as well as interactions. For instance, a one-unit increase in Mjobhealth is associated with a 0.14 increase in the cube root of the target variable. Higher goout is associated with a decrease in the cube root of the target variable of (-0.3) which makes sense as when a student would go out more that might lead to lower grades. Moreover, the interaction failures:Mjobservice has a negative impact on the cube root of the target variable leading to a decrease of -0.11. The R-squared value indicates that the model explains 31.83% of the target variable, while when adjusted based on the number of predictors, the adjusted R-squared explains 26.68% of the data. So, in conclusion, the regression model seems to have significant effect in explaining the target variable which is supported by the low p-value.

**Interpreting the findings:**

Through our analysis, we were able gain several insights into the different factors that influence the mathematics final year grade. For instance, the predictor Mjobhealth has a positive impact on the target variable which possibly indicates that students with a mother working in the services sector tend to perform better, suggesting the potential influence of parental occupation on student’s performance. The negative effect of the goout predictor is expected. It can be predicted that more social interaction would lead to lower grades possibly due to less studying time. On the other hand, a predictor like schoolsupp has a low significant effect in modeling the target variable. This perhaps indicates that school support is showing no progress in helping them study and improve their grades. An unexpected finding is the negative effect that absences has on G3. Intuitively, we would expect that more absences would lead to a decrease in grades which is not the case in the model but at a small magnitude. The interactions that are present in the model are mostly unexpected. However, they have a significant effect in modeling the target variable. In conclusion, the analysis and findings have yielded unexpected insights into the different factors that influence a student’s mathematics final year grade. The findings open exploration into how different social and economic factors that might seem to be irrelevant in explaining a student’s final grade, however, our investigation has revealed that perhaps that there some underlying effects that can even be further explored.

**Conclusion:**

To conclude, our regression analysis of the mathematics dataset introduces the intricate relationships between several predictors and academic performance. The results obtained were able to provide insights that can help improve educational performance for the student and obtain better final year grades. For instance, the influence of parental occupation, particularly the positive impact of a service-related job (Mjobservices), indicates the significance of family demographics in shaping academic outcomes. The interaction between failures:Mjobother and failures:Mjobservices suggest that there is a deeper relationship between failures and a mother’s job which can be further explored to provide a more nuanced understanding of the effects they have on a student’s performance. Surprisingly, absences has shown a negative impact on a student’s final year grade. This perhaps is an indication for educational institutes to start enforcing more strict attendance policies to further improve student’s performance. The identified factors, such as socializing habits (goout) and the presence of school support (schoolsup), outline how several factors seem to be influencing the final year grade. This further entails how complex and intricate the model is. The non-significant impact of schoolsupyes prompts schools to improve their academic support program so that they help students in excelling. Overall, the regression analyses will act as reference for the educators to improve their student’s overall final year grade in mathematics. By outlining the several relationships that exists between the different social and economic factors in the model, the school can tailor their curriculum and support services to the address the needs of the students, therefore leading to a better educational environment.

**Appendix**

df <- read.csv("student-mat.csv", sep = ";")

print(dim(df))

head(df)

categorical\_vars <- c("school", "sex", "address","famsize","Pstatus","Mjob","Fjob","reason","guardian",

"schoolsup","famsup","paid","activities","nursery", "higher","internet","romantic")

df[, categorical\_vars] <- lapply(df[, categorical\_vars], as.factor)

df <- df[, !(names(df) %in% c("g1", "g2"))]

df

null\_model <- lm(G3 ~ 1 , data = df)

full\_model <- lm(G3 ~ ., data = df)

aic <- step(full\_model, direction = "backward")

bic <- step(full\_model, direction = "backward", k=log(n))

aic

bic

aic\_model <- lm(formula = G3 ~ sex + age + famsize + Medu + Mjob + studytime +

failures + schoolsup + famsup + romantic + freetime + goout +

absences, data = df)

bic\_model <-lm(formula = G3 ~ Medu + failures, data = df)

anova(bic\_model,aic\_model)

adj\_r\_squared\_aicmodel <- summary(aic\_model)$adj.r.squared

adj\_r\_squared\_bicmodel <- summary(bic\_model)$adj.r.squared

cat(adj\_r\_squared\_aicmodel,adj\_r\_squared\_bicmodel)

first\_final\_model <-aic\_model

summary(first\_final\_model)

full\_model <- lm(G3 ~ sex \* age \* famsize \* Mjob \* romantic \* failures \* schoolsup

\* famsup \* freetime \* goout \*

studytime \* absences, data = df)

aic\_full\_model <- step(null\_model, scope = formula(full\_model), direction = "both")

summary(aic\_full\_model)

final\_model <- aic\_full\_model

plot(fitted(final\_model), resid(final\_model), col = "grey", pch = 20,

xlab = "Fitted", ylab = "Residual",cex=2,

main = "Residual plot")

abline(h = 0, col = "darkorange", lwd = 2)

qqnorm(resid(final\_model), col = "grey",pch=20,cex=2)

qqline(resid(final\_model), col = "dodgerblue", lwd = 2)

library(lmtest)

bptest(final\_model)

shapiro.test(resid(final\_model))

final\_model <- lm((G3-0.00001)^(1/3) ~ failures + Mjob + sex + goout + romantic +

absences + schoolsup + studytime +studytime + famsup + age + failures:absences +

failures:Mjob + failures:schoolsup + failures:goout + schoolsup:studytime^2 +

studytime:famsup + failures:age + famsup:age + sex:romantic,

data = df)

plot(fitted(final\_model), resid(final\_model), col = "grey", pch = 20,

xlab = "Fitted", ylab = "Residual",cex=2,

main = "Residual plot")

abline(h = 0, col = "darkorange", lwd = 2)

qqnorm(resid(final\_model), col = "grey",pch=20,cex=2)

qqline(resid(final\_model), col = "dodgerblue", lwd = 2)

bptest(final\_model)

shapiro.test(resid(final\_model))

summary(final\_model)