**1.Executive summary**

This analysis report uses students alcohol consumption dataset to find out what kind of factor will infect the average grade of students. I use ANOVA and T-test to find out whether there are significant differences between numerical variables. Next, I use linear regression and put in all variables to find out which will significantly affect the average grade. Then I use backward and forward stepwise model selection to find out the best model. The adjusted R-Square of backward stepwise model is the highest, which means it can explain most amount of the data. The results of my data analyzed show that the study time, failures, getting school supply, extra paid classes, wanting to take higher education, and absences will affect the average grade. And surprisingly, alcohol consumption has little to do with average grade.

**2.Exploratory data analysis**

Do the correlation of all numerical variables using heatmap.

corrstudent=student[,c('age', 'studytime', 'failures', 'famrel', 'freetime', 'goout', 'Dalc', 'Walc', 'health', 'absences', 'G1', 'G2', 'G3' )]

corrstudent=round(cor(corrstudent),2)

library(“reshape2”)

melted\_corrstudent=melt(corrstudent)

library("ggplot2")

ggplot(data=melted\_corrstudent, aes(x=Var1, y=Var2, fill=value))+geom\_tile()



first period grade(G1), second period grade (G2), final grade (G3) are more correlated, so I create a new numerical variable “Gavg” which is the average of G1, G2, and G3. Weekday alcohol consumption (Dalc) and weekend alcohol consumption (Walc) are more correlated, so I create a new factor “alc” by sum Dalc and Walc. If the sum <=2 then define the level to be 1, if the sum<=4 then the level will be 2 and so on. We can see from this heatmap that most of the variables are not correlated. But maybe there are some correlation between gout and alc.

Since the data of several variables (studytime, failures, famrel, freetime, gout, health, alc) were break by degree, I transform them into factor.

#transfor numeric into factors

student$studytime=factor(student$studytime)

student$failures=factor(student$failures)

student$famrel=factor(student$famrel)

student$freetime=factor(student$freetime)

student$goout=factor(student$goout)

student$health=factor(student$health)

student$alc=factor(student$alc)

Do the summary of the dataset.

summary(student)

一張含有 文字, 收據 的圖片

自動產生的描述

Do x^2 test or ANOVA between the variables that I think there are logically interact to each other’s.

I try to find out several variables (failures, freetime, romantic, paid, schoolsup, alc, gout, internet) which would affect study time, and surprisingly found out that internet accessible at don’t really affect study time since I thought student would study less due to the use of internet.

Next, I want to find out whether alcohol consumption interact with several variables (failures, goout, health, schoolsup). As I thought, when the alcohol consumption goes up, the failures go up and less people get extra education support. And when people go out more, the alcohol consumption will also go up. The healthier they are, the less alcohol they consume.

Also, I do the ANOVA to test the differences of absences in different failures. And I find out that the significant difference only show up between students without failures and students with one failures.

一張含有 瓶, 桌 的圖片

自動產生的描述

**3.Methods:**

Model building:

First, I put all variables into a linear model to find out which variable will significantly affect average grade, and find out that significant variables are : address, studytime, failures, schoolsup, paid, higher, internet, romantic, goout, health.

mod\_all=lm(Gavg~sex+age+address+studytime+failures+schoolsup+famsup+paid+nursery+higher+internet+romantic+famrel+freetime+goout+health+absences+alc, data=student)

summary(mod\_all)

一張含有 文字 的圖片

自動產生的描述一張含有 發現 的圖片

自動產生的描述



And then I put all these significant variables into another linear regression, but the adjusted r square(0.2558) decrease. So, I use my first linear regression model with forward selection, backward selection, stepwise-forward selection and stepwise-backward selection. I found out that in all these models are significant and have the same adjusted r square(0.2669) (higher than the original model).

#backward

back=step(mod\_all, direction='backward')

summary(back)

#forward

fwd=step(mod\_n, scope=list(upper=mod\_all), direction='forward')

summary(fwd)

#stepwise\_forward

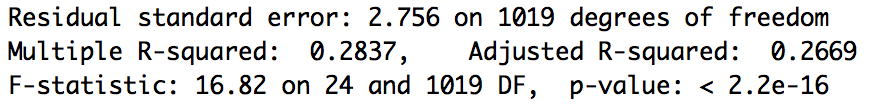
stepwisef=step(mod\_n, scope=list(lower=mod\_n, upper=mod\_all), direction='both')

summary(stepwisef)

#stepwise\_backward

stepwiseb=step(mod\_all, direction='both')

summary(stepwiseb)





I try to add some interaction into the original model. Since I don’t think age will affect the average grade, so I use the other numerical variable “absences” to interact with the significant variables (address, studytime, failures, schoolsup, paid, higher, internet, romantic, goout, health), the adjusted r square is better than all of the models above. Some of the interaction do work.

mod\_6=lm(Gavg~sex+age+address+studytime+failures+schoolsup+famsup+paid+nursery+higher+internet+romantic+famrel+freetime+goout+health+absences+alc+romantic\*absences+paid\*absences+internet\*absences+failures\*absences+alc\*absences+studytime\*absences+goout\*absences+health\*absences+schoolsup\*absences+higher\*absences, data=student)

summary(mod\_6)

一張含有 文字 的圖片

自動產生的描述一張含有 發現, 天空 的圖片

自動產生的描述



Again, I use this model (with interaction) to do the forward selection, backward selection, stepwise-forward selection and stepwise-backward selection, and find out that the model using stepwise-backward selection is significant and has the highest adjusted r square (0.2814).

#backward

backii=step(mod\_6, direction='backward')

summary(backii)

#forward

fwdii=step(mod\_n, scope=list(upper=mod\_6), direction='forward')

summary(fwdii)

#stepwise\_forward

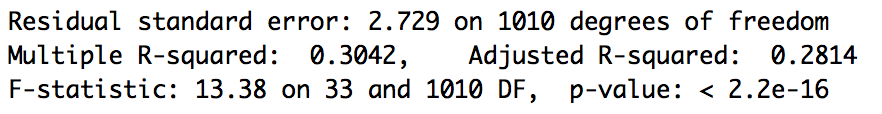
stepwisefii=step(mod\_n, scope=list(lower=mod\_n, upper=mod\_6), direction='both')

summary(stepwisefii)

#stepwise\_backward

stepwisebii=step(mod\_6, direction='both')

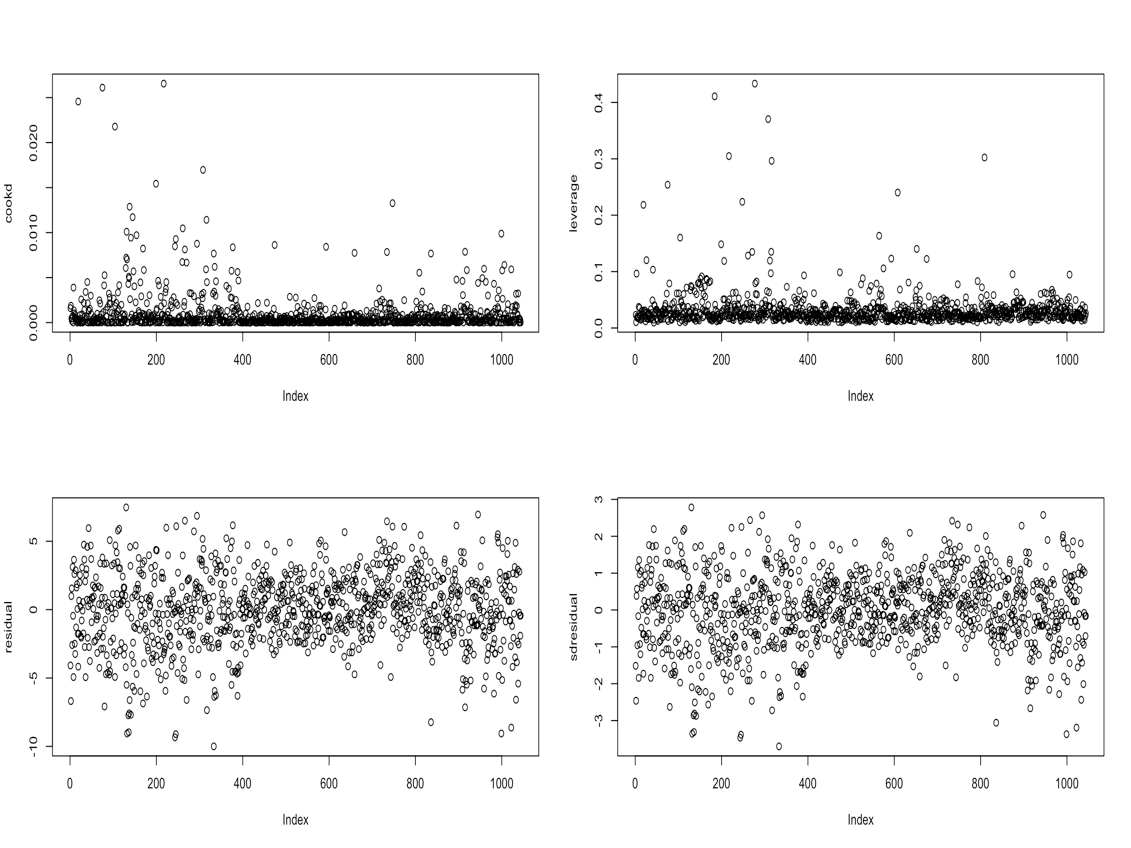
summary(stepwisebii)





Model fit:

I calculate and plot the cook-d, leverage, residual, and standardized residuals to find out the outliers.



The outliers are: (cook-d>0.01), (leverage>0.2), (residual<-7)

Also, I print out the outliers by cook-d, leverage, and residual. I found that the influential points is “absences”. Since it is the only significant numerical variable in this model. To test it, I make another model with predictors only “Obs” and “absences” and found that the outliers are the same.

I delete these outliers and do a new model, found out the adjusted r square increase (0.3207), which means this new model can fit more data and I am satisfied with it.

outliers = cbind(student,cookd, leverage, residual, sdresidual)

moveout = outliers %>% filter(leverage<0.2)

moveout = moveout %>% filter(cookd<0.01)

moveout = moveout %>% filter(residual>(-7))

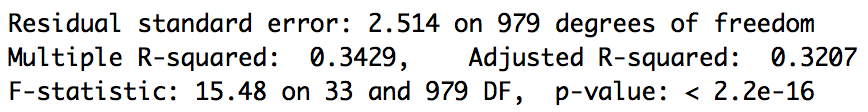
mod\_7=lm(Gavg~sex+age+address+studytime+failures+schoolsup+famsup+paid+nursery+higher+internet+romantic+famrel+freetime+goout+health+absences+alc+romantic\*absences+paid\*absences+internet\*absences+failures\*absences+alc\*absences+studytime\*absences+goout\*absences+health\*absences+schoolsup\*absences+higher\*absences, data=moveout)

summary(mod\_7)

stepwisebout=step(mod\_7, direction='both')

summary(stepwisebout)

一張含有 文字 的圖片

自動產生的描述



Model interpretation:

Model (rank by importance):

Gavg =

9.923 + (-3.15\* failures \_1) + (-3.9\* failures \_2) + (-4.34\* failures \_3)

+ (-1.19\*schoolsup\_yes) + (-0.84\*paid\_yes) + (1.57\*higher\_yes)

+ (0.57\*address\_U) + (-0.62\*romantic\_yes) + (-0.13\*absences)

+ (0.56\*internet\_yes)

+ (-1.19\*health \_2) + (-1.35\*health \_3) + (-0.76\*health \_4) + (-1.04\*health \_5)

+ (0.41\*studytime\_2) + (1.19\*studytime\_3) + (1.11\*studytime\_4)

+ (0.2\*health\_2\*absences) + (0.1\*health\_3\*absences)

+ (0.06\*health\_4\*absences) + (0.08\*health\_5\*absences)

+ (0.87\*freetime\_2) +(-0.07\*freetime\_3) + (0.3\*freetime\_4) + (0.91\*freetime\_5)

+ (0.96\* goout\_2) + (0.56\* goout\_3) + (0.13\* goout\_4) + (-0.24\* goout\_5)

+ (0.11\*failures\_1\*absences) + (0.13\*failures\_2\*absences)

+ (0.15\*failures\_3\*absences)+ (0.04\*romantic\*absences)

We can see when there are more failures, the average will decrease.

If students have extra education support or extra paid classes, the average grade will also decrease which is an interesting discover because most people belief that if we pay more money and take extra classes to get more practice, students’ grade will increase. If students want to get higher education, they will get a higher average grade. Students live in urban will get higher average grade than students live in rural. Student with internet access at home will get a higher average grade which might because they can use it to find some useful information related to their academic work. And not surprisingly, having a romantic relationship and absences will decrease the average grade. Students with very good health status will get higher average grade (the health\_4 is not so significant). The more time students spent on learning will help them to get a higher average grade. The absences might due to their health status, as the result, student with very good health status their absences would only do a little affect to their average grade than students with very bad health status. Students with more free time will get a higher average grade (freetime\_3 and freetime\_4 are not so significant). Student don’t go out often will get a higher average grade, and student go out often will decrease the average grade. Since the coefficient of “failures” were negative and more failures will cause lower average grade, more absences with more failures will decrease more than fewer failures. Students with romantic relationship and absence the class will decrease more than student without romantic relationship.

**4.Conclusions:​**

According to my final model, I found out there are several variables won’t affect average grade, which are sex, age, famsup, nursery, famrel, alc. What surprised me the most is that the alcohol consumption won’t affect the average grade. Since I thought students with highly alcohol consumption would be the one who hangs out more and not spending their time on studying. Another surprising discovery is that living area do significantly affect average grade. It is surprise because i wanted to delete this variable at first and I believe the average grade will only affected by some self-control factors. We can dig in more to find out whether there is different education level between urban and rural. From the model, we can see that the extra payment for study will decrease the average grade. I think it is important to know why this happened because usually when people spending more money and time on something, they will get a higher reward. Maybe these extra studies make students feel too tires and can’t be at a healthy status? As the model shows, student with good health status will gain more than bad health status. Also, I learned that students with more free time will gain higher average grade, so I think maybe students can use the free time to do some sports which can also improve their health status. I believe focusing on students’ health will be an effective way to improve their grade.