Project 2

Vinh Van

4/16/2022

rm(list=ls())

grades <- read.csv(  
 file = "C:\\Users\\Duyet\\Downloads\\School\\Spring 2022\\STA4164\\Project 2\\student-mat.csv"  
)

##Comment out when knitting

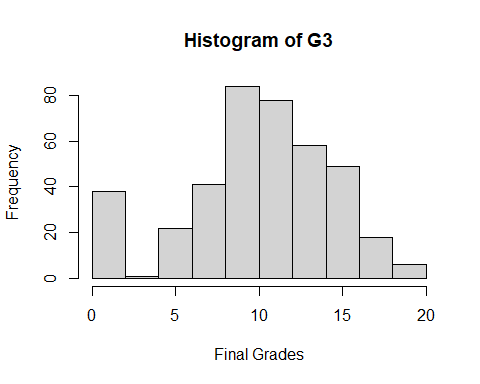
##grades

Summary Statistics and Histogram of G3

summary(grades[c('G3')])

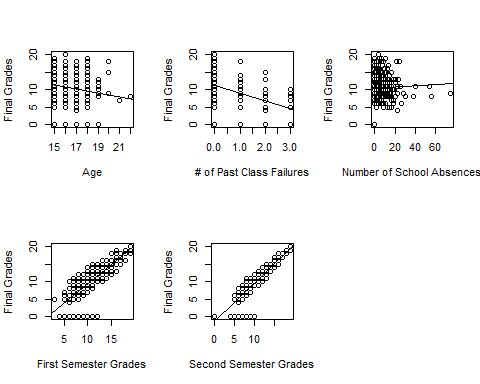
## G3   
## Min. : 0.00   
## 1st Qu.: 8.00   
## Median :11.00   
## Mean :10.42   
## 3rd Qu.:14.00   
## Max. :20.00

hist(grades[,'G3'],xlab='Final Grades',main='Histogram of G3')



Creating the scatter plots

par(  
 mfrow=c(2,3)  
)  
  
plot(  
 x = grades$age,  
 y = grades$G3,  
 xlab = "Age",  
 ylab = "Final Grades",  
)  
abline(  
 reg = lsfit(  
 x = grades$age,  
 y = grades$G3  
 )  
)  
  
plot(  
 x = grades$failures,  
 y = grades$G3,  
 xlab = "# of Past Class Failures",  
 ylab = "Final Grades",  
)  
abline(  
 reg = lsfit(  
 x = grades$failures,  
 y = grades$G3  
 )  
)  
  
plot(  
 x = grades$absences,  
 y = grades$G3,  
 xlab = "Number of School Absences",  
 ylab = "Final Grades",  
)  
abline(  
 reg = lsfit(  
 x = grades$absences,  
 y = grades$G3  
 )  
)  
  
plot(  
 x = grades$G1,  
 y = grades$G3,  
 xlab = "First Semester Grades",  
 ylab = "Final Grades",  
)  
abline(  
 reg = lsfit(  
 x = grades$G1,  
 y = grades$G3  
 )  
)  
  
plot(  
 x = grades$G2,  
 y = grades$G3,  
 xlab = "Second Semester Grades",  
 ylab = "Final Grades",  
)  
abline(  
 reg = lsfit(  
 x = grades$G2,  
 y = grades$G3  
 )  
)



Creating a model of just the quantatative variables

lm\_initial <- lm(  
 formula = G3 ~ G1 + G2 + age + failures + absences,  
 data = grades,  
)  
  
lm\_initialsquared <- lm(  
 formula = G3^2 ~ G1 + G2 + age + failures + absences,  
 data = grades,  
)  
  
summary(lm\_initial)

##   
## Call:  
## lm(formula = G3 ~ G1 + G2 + age + failures + absences, data = grades)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -9.0336 -0.4191 0.2659 0.9710 3.5596   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 0.92705 1.35308 0.685 0.493664   
## G1 0.16002 0.05607 2.854 0.004546 \*\*   
## G2 0.96300 0.04958 19.425 < 2e-16 \*\*\*  
## age -0.16465 0.07950 -2.071 0.039012 \*   
## failures -0.19332 0.14235 -1.358 0.175231   
## absences 0.04179 0.01218 3.432 0.000664 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 1.903 on 389 degrees of freedom  
## Multiple R-squared: 0.8296, Adjusted R-squared: 0.8274   
## F-statistic: 378.7 on 5 and 389 DF, p-value: < 2.2e-16

summary(lm\_initialsquared)

##   
## Call:  
## lm(formula = G3^2 ~ G1 + G2 + age + failures + absences, data = grades)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -109.00 -17.24 -1.27 19.02 79.11   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -54.2431 20.9166 -2.593 0.00986 \*\*   
## G1 9.0809 0.8667 10.478 < 2e-16 \*\*\*  
## G2 14.0159 0.7664 18.289 < 2e-16 \*\*\*  
## age -3.9477 1.2289 -3.212 0.00143 \*\*   
## failures 3.3350 2.2006 1.516 0.13046   
## absences -0.1357 0.1882 -0.721 0.47143   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 29.42 on 389 degrees of freedom  
## Multiple R-squared: 0.8818, Adjusted R-squared: 0.8803   
## F-statistic: 580.6 on 5 and 389 DF, p-value: < 2.2e-16

anova(lm\_initial)

## Analysis of Variance Table  
##   
## Response: G3  
## Df Sum Sq Mean Sq F value Pr(>F)   
## G1 1 5312.2 5312.2 1466.1552 < 2.2e-16 \*\*\*  
## G2 1 1487.0 1487.0 410.4188 < 2.2e-16 \*\*\*  
## age 1 12.5 12.5 3.4467 0.0641337 .   
## failures 1 6.1 6.1 1.6855 0.1949603   
## absences 1 42.7 42.7 11.7773 0.0006639 \*\*\*  
## Residuals 389 1409.4 3.6   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

anova(lm\_initialsquared)

## Analysis of Variance Table  
##   
## Response: G3^2  
## Df Sum Sq Mean Sq F value Pr(>F)   
## G1 1 2189292 2189292 2528.5897 < 2.2e-16 \*\*\*  
## G2 1 313407 313407 361.9795 < 2.2e-16 \*\*\*  
## age 1 8446 8446 9.7546 0.001923 \*\*   
## failures 1 1956 1956 2.2594 0.133615   
## absences 1 450 450 0.5196 0.471433   
## Residuals 389 336802 866   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Correlation matrix

mcor <- cor(grades[c('G3','G1','G2','age','failures','absences')])  
mcor

## G3 G1 G2 age failures absences  
## G3 1.00000000 0.8014679 0.9048680 -0.1615794 -0.36041494 0.03424732  
## G1 0.80146793 1.0000000 0.8521181 -0.0640815 -0.35471761 -0.03100290  
## G2 0.90486799 0.8521181 1.0000000 -0.1434740 -0.35589563 -0.03177670  
## age -0.16157944 -0.0640815 -0.1434740 1.0000000 0.24366538 0.17523008  
## failures -0.36041494 -0.3547176 -0.3558956 0.2436654 1.00000000 0.06372583  
## absences 0.03424732 -0.0310029 -0.0317767 0.1752301 0.06372583 1.00000000

Variance Inflation Factor

library(car)

## Warning: package 'car' was built under R version 4.1.3

## Loading required package: carData

## Warning: package 'carData' was built under R version 4.1.3

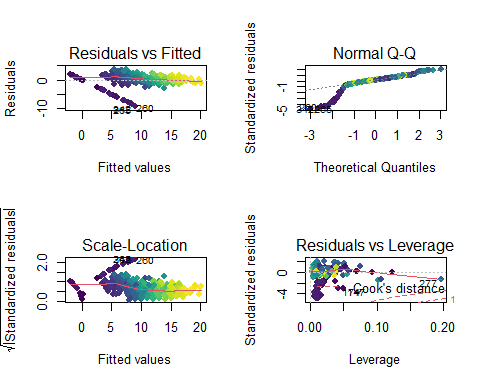
vif <- vif(lm\_initial)  
vif

## G1 G2 age failures absences   
## 3.765914 3.781400 1.119064 1.218642 1.032821

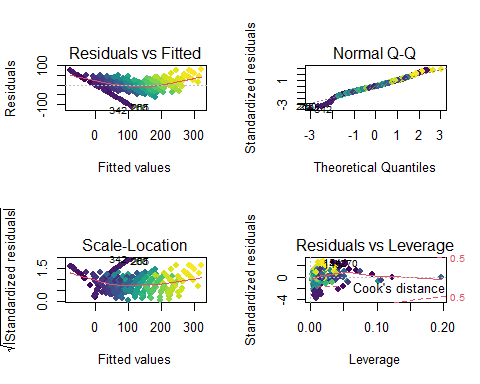
Diagnostic Plots

v\_color <- viridis::viridis(  
 n = nrow(  
 x = grades  
 )  
)  
grades$color <- v\_color[Matrix::invPerm(  
 p = order(  
 x = grades$G3  
 )  
)]

par(mfrow=c(2,2))  
plot(  
 x = lm\_initial,  
 pch = 19,  
 col = grades$color  
)



par(mfrow = c(2,2))  
plot(  
 x = lm\_initialsquared,  
 pch = 19,  
 col = grades$color  
)



Installing the fast dummies package to handle the ludacrous amount of dummy variables in the dataset

library('fastDummies')

## Warning: package 'fastDummies' was built under R version 4.1.3

Creating the dummy variables

gradedummy <- dummy\_cols(grades, select\_columns = c(  
 'school', 'sex','address','famsize','Medu','Fedu','Mjob','Fjob','traveltime','studytime','schoolsup','famsup','activities','nursery','higher','internet','romantic','famrel','freetime','goout','Dalc','Walc','health'))

Building the linear Model with dummy variables

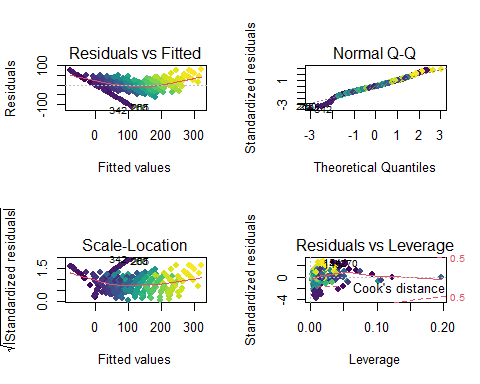
lm\_full <- lm(  
 formula = G3^2 ~ G1 + G2 + age + failures + absences + school\_GP + school\_MS + sex\_F + sex\_M + address\_R + address\_U  
 + famsize\_GT3 + famsize\_LE3 + Medu\_0 + Medu\_0 + Medu\_1 + Medu\_2 +Medu\_3 + Medu\_4 +Fedu\_0 + Fedu\_1 + Fedu\_2 + Fedu\_3 + Fedu\_4 + Mjob\_at\_home + Mjob\_health + Mjob\_other + Mjob\_services + Mjob\_teacher + Fjob\_at\_home + Fjob\_health + Fjob\_other + Fjob\_services + Fjob\_teacher + traveltime\_1 + traveltime\_2 + traveltime\_3 + traveltime\_4 + studytime\_1 + studytime\_2 + studytime\_3 + studytime\_4 + schoolsup\_no + schoolsup\_yes + famsup\_no + famsup\_yes + activities\_no + activities\_yes + nursery\_no + nursery\_yes + higher\_no + higher\_yes + internet\_no + internet\_yes + romantic\_no + romantic\_yes + famrel\_1 + famrel\_2 + famrel\_3 + famrel\_4 + famrel\_5 + freetime\_1 + freetime\_2 + freetime\_3 + freetime\_4 + freetime\_5 + goout\_1 + goout\_2 + goout\_3 + goout\_4 + goout\_5 + Dalc\_1 + Dalc\_2 + Dalc\_3 + Dalc\_4 + Dalc\_5 + Walc\_1 + Walc\_2 + Walc\_3 + Walc\_4 + Walc\_5 + health\_1 + health\_2 + health\_3 + health\_4 + health\_5,  
 data = gradedummy  
)  
  
summary(lm\_full)

##   
## Call:  
## lm(formula = G3^2 ~ G1 + G2 + age + failures + absences + school\_GP +   
## school\_MS + sex\_F + sex\_M + address\_R + address\_U + famsize\_GT3 +   
## famsize\_LE3 + Medu\_0 + Medu\_0 + Medu\_1 + Medu\_2 + Medu\_3 +   
## Medu\_4 + Fedu\_0 + Fedu\_1 + Fedu\_2 + Fedu\_3 + Fedu\_4 + Mjob\_at\_home +   
## Mjob\_health + Mjob\_other + Mjob\_services + Mjob\_teacher +   
## Fjob\_at\_home + Fjob\_health + Fjob\_other + Fjob\_services +   
## Fjob\_teacher + traveltime\_1 + traveltime\_2 + traveltime\_3 +   
## traveltime\_4 + studytime\_1 + studytime\_2 + studytime\_3 +   
## studytime\_4 + schoolsup\_no + schoolsup\_yes + famsup\_no +   
## famsup\_yes + activities\_no + activities\_yes + nursery\_no +   
## nursery\_yes + higher\_no + higher\_yes + internet\_no + internet\_yes +   
## romantic\_no + romantic\_yes + famrel\_1 + famrel\_2 + famrel\_3 +   
## famrel\_4 + famrel\_5 + freetime\_1 + freetime\_2 + freetime\_3 +   
## freetime\_4 + freetime\_5 + goout\_1 + goout\_2 + goout\_3 + goout\_4 +   
## goout\_5 + Dalc\_1 + Dalc\_2 + Dalc\_3 + Dalc\_4 + Dalc\_5 + Walc\_1 +   
## Walc\_2 + Walc\_3 + Walc\_4 + Walc\_5 + health\_1 + health\_2 +   
## health\_3 + health\_4 + health\_5, data = gradedummy)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -82.766 -14.630 0.023 15.476 88.777   
##   
## Coefficients: (23 not defined because of singularities)  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -51.0344 31.6859 -1.611 0.10821   
## G1 9.3959 0.9616 9.771 < 2e-16 \*\*\*  
## G2 13.5253 0.8215 16.465 < 2e-16 \*\*\*  
## age -3.8748 1.4892 -2.602 0.00968 \*\*   
## failures 2.1440 2.4290 0.883 0.37806   
## absences -0.1175 0.2059 -0.570 0.56874   
## school\_GP -3.0240 5.7687 -0.524 0.60049   
## school\_MS NA NA NA NA   
## sex\_F -0.4231 3.6813 -0.115 0.90856   
## sex\_M NA NA NA NA   
## address\_R 0.9532 4.2030 0.227 0.82072   
## address\_U NA NA NA NA   
## famsize\_GT3 -2.6262 3.5515 -0.739 0.46015   
## famsize\_LE3 NA NA NA NA   
## Medu\_0 3.7385 18.8560 0.198 0.84296   
## Medu\_1 -14.5039 7.5668 -1.917 0.05612 .   
## Medu\_2 -12.5269 5.8925 -2.126 0.03425 \*   
## Medu\_3 -2.2361 5.1953 -0.430 0.66718   
## Medu\_4 NA NA NA NA   
## Fedu\_0 15.9588 22.2350 0.718 0.47343   
## Fedu\_1 13.3307 6.1891 2.154 0.03197 \*   
## Fedu\_2 0.6407 5.3175 0.120 0.90417   
## Fedu\_3 5.8309 4.9891 1.169 0.24335   
## Fedu\_4 NA NA NA NA   
## Mjob\_at\_home 6.0008 7.7183 0.777 0.43743   
## Mjob\_health 3.4504 6.8743 0.502 0.61605   
## Mjob\_other 7.1494 6.2507 1.144 0.25354   
## Mjob\_services 9.5168 6.0110 1.583 0.11432   
## Mjob\_teacher NA NA NA NA   
## Fjob\_at\_home -9.2823 9.5541 -0.972 0.33198   
## Fjob\_health -8.5288 9.4705 -0.901 0.36847   
## Fjob\_other -7.8617 6.9826 -1.126 0.26102   
## Fjob\_services -12.5343 7.2482 -1.729 0.08469 .   
## Fjob\_teacher NA NA NA NA   
## traveltime\_1 -8.9608 11.9649 -0.749 0.45443   
## traveltime\_2 -11.4222 12.1678 -0.939 0.34856   
## traveltime\_3 -11.9468 13.3160 -0.897 0.37028   
## traveltime\_4 NA NA NA NA   
## studytime\_1 14.1225 7.1472 1.976 0.04899 \*   
## studytime\_2 5.4969 6.4795 0.848 0.39686   
## studytime\_3 6.8513 7.1450 0.959 0.33831   
## studytime\_4 NA NA NA NA   
## schoolsup\_no 4.5112 4.9705 0.908 0.36476   
## schoolsup\_yes NA NA NA NA   
## famsup\_no -4.1839 3.3528 -1.248 0.21295   
## famsup\_yes NA NA NA NA   
## activities\_no 4.9804 3.1858 1.563 0.11893   
## activities\_yes NA NA NA NA   
## nursery\_no 1.1461 3.9413 0.291 0.77139   
## nursery\_yes NA NA NA NA   
## higher\_no 4.5720 7.7711 0.588 0.55671   
## higher\_yes NA NA NA NA   
## internet\_no -3.3247 4.4760 -0.743 0.45813   
## internet\_yes NA NA NA NA   
## romantic\_no 5.6396 3.4724 1.624 0.10530   
## romantic\_yes NA NA NA NA   
## famrel\_1 -12.1585 11.4918 -1.058 0.29082   
## famrel\_2 -21.2262 7.8441 -2.706 0.00716 \*\*   
## famrel\_3 -9.2420 4.9421 -1.870 0.06236 .   
## famrel\_4 -9.0060 3.6906 -2.440 0.01520 \*   
## famrel\_5 NA NA NA NA   
## freetime\_1 1.5939 9.1366 0.174 0.86162   
## freetime\_2 -1.9551 6.5712 -0.298 0.76625   
## freetime\_3 -2.5485 5.9368 -0.429 0.66801   
## freetime\_4 -1.2874 5.9130 -0.218 0.82778   
## freetime\_5 NA NA NA NA   
## goout\_1 -11.7248 8.0544 -1.456 0.14642   
## goout\_2 0.0856 6.0345 0.014 0.98869   
## goout\_3 0.2012 5.7013 0.035 0.97187   
## goout\_4 -2.4013 5.7055 -0.421 0.67412   
## goout\_5 NA NA NA NA   
## Dalc\_1 20.9113 13.3679 1.564 0.11870   
## Dalc\_2 13.0369 13.3330 0.978 0.32889   
## Dalc\_3 11.5578 13.5828 0.851 0.39543   
## Dalc\_4 9.3077 14.9946 0.621 0.53520   
## Dalc\_5 NA NA NA NA   
## Walc\_1 -7.4305 9.6458 -0.770 0.44165   
## Walc\_2 -16.5081 9.6093 -1.718 0.08674 .   
## Walc\_3 -12.1104 9.2325 -1.312 0.19052   
## Walc\_4 -7.7665 9.0671 -0.857 0.39231   
## Walc\_5 NA NA NA NA   
## health\_1 5.5274 5.3494 1.033 0.30222   
## health\_2 -1.4869 5.2407 -0.284 0.77681   
## health\_3 0.8419 4.2990 0.196 0.84485   
## health\_4 4.9722 4.7278 1.052 0.29370   
## health\_5 NA NA NA NA   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 28.76 on 332 degrees of freedom  
## Multiple R-squared: 0.9036, Adjusted R-squared: 0.8856   
## F-statistic: 50.21 on 62 and 332 DF, p-value: < 2.2e-16

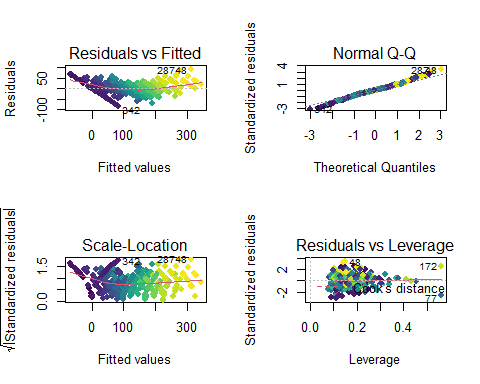
Diagnostic Plots

v\_color1 <- viridis::viridis(  
 n = nrow(  
 x = gradedummy  
 )  
)  
gradedummy$color <- v\_color[Matrix::invPerm(  
 p = order(  
 x = gradedummy$G3  
 )  
)]

par(mfrow = c(2,2))  
plot(  
 x = lm\_initialsquared,  
 pch = 19,  
 col = grades$color  
)



par(mfrow=c(2,2))  
plot(  
 x = lm\_full,  
 pch = 19,  
 col = gradedummy$color  
)



Stepwise Regression

library(olsrr)

## Warning: package 'olsrr' was built under R version 4.1.3

##   
## Attaching package: 'olsrr'

## The following object is masked from 'package:datasets':  
##   
## rivers

Stepwise with p values

stepwise.p <- ols\_step\_both\_p(lm\_full,prem = 0.05, details = FALSE)  
stepwise.p

##   
## Stepwise Selection Summary   
## ---------------------------------------------------------------------------------------------  
## Added/ Adj.   
## Step Variable Removed R-Square R-Square C(p) AIC RMSE   
## ---------------------------------------------------------------------------------------------  
## 1 G2 addition 0.847 0.847 136.0170 3894.5829 33.3100   
## 2 G1 addition 0.878 0.877 31.1760 3807.0920 29.7804   
## 3 famrel\_5 addition 0.882 0.881 19.0530 3795.5866 29.3130   
## 4 age addition 0.885 0.884 10.2080 3786.8933 28.9560   
## 5 studytime\_1 addition 0.887 0.885 6.4430 3783.0888 28.7809   
## 6 Fjob\_services addition 0.889 0.887 2.5310 3779.0465 28.5984   
## 7 Fedu\_2 addition 0.890 0.888 -0.5640 3775.7647 28.4445   
## 8 Walc\_1 addition 0.892 0.889 -3.5330 3772.5431 28.2937   
## 9 Medu\_3 addition 0.893 0.890 -4.9950 3770.8647 28.1988   
## 10 Medu\_3 removal 0.892 0.889 -3.5330 3772.5431 28.2937   
## 11 Medu\_2 addition 0.893 0.890 -4.9310 3770.9334 28.2013   
## 12 Fedu\_2 removal 0.892 0.889 -4.0490 3771.9978 28.2742   
## 13 Fedu\_1 addition 0.893 0.890 -5.5810 3770.2388 28.1765   
## 14 Fedu\_1 removal 0.892 0.889 -4.0490 3771.9978 28.2742   
## 15 goout\_1 addition 0.893 0.890 -4.7200 3771.1586 28.2093   
## 16 goout\_1 removal 0.892 0.889 -4.0490 3771.9978 28.2742   
## ---------------------------------------------------------------------------------------------

Stepwise with AIC values #Comment out when Knitting

##stepwise.aic <- ols\_step\_both\_aic(lm\_full, details = FALSE)  
##stepwise.aic

Reduced model post Stepwise p-value

lm\_reduced <- lm(  
 formula = G3^2 ~ G2 + G1 + famrel\_5 + age + studytime\_1 + Fjob\_services + Walc\_1 + Medu\_2,  
 data = gradedummy  
)  
  
summary(lm\_reduced)

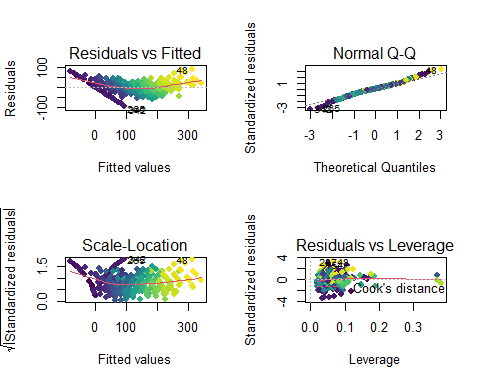
##   
## Call:  
## lm(formula = G3^2 ~ G2 + G1 + famrel\_5 + age + studytime\_1 +   
## Fjob\_services + Walc\_1 + Medu\_2, data = gradedummy)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -95.501 -16.601 -1.086 16.851 85.753   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -68.4566 20.4057 -3.355 0.000873 \*\*\*  
## G2 14.0896 0.7373 19.109 < 2e-16 \*\*\*  
## G1 8.7084 0.8294 10.499 < 2e-16 \*\*\*  
## famrel\_5 12.2177 3.2349 3.777 0.000184 \*\*\*  
## age -3.1481 1.1511 -2.735 0.006530 \*\*   
## studytime\_1 9.5868 3.2982 2.907 0.003864 \*\*   
## Fjob\_services -7.3567 3.1998 -2.299 0.022032 \*   
## Walc\_1 7.2681 3.0500 2.383 0.017656 \*   
## Medu\_2 -7.9047 3.2525 -2.430 0.015541 \*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 28.27 on 386 degrees of freedom  
## Multiple R-squared: 0.8917, Adjusted R-squared: 0.8895   
## F-statistic: 397.4 on 8 and 386 DF, p-value: < 2.2e-16

lm\_reduced\_FINAL <- lm(  
 formula = G3^2 ~ G2 + G1 + famrel\_1 + famrel\_2 + famrel\_3 + famrel\_4 + famrel\_5 + age + studytime\_1 + studytime\_2 + studytime\_3 + studytime\_4 + Fjob\_services + Fjob\_at\_home + Fjob\_health + Fjob\_other + Fjob\_teacher + Walc\_1 + Walc\_2 + Walc\_3 + Walc\_4 + Walc\_5 + Medu\_1 + Medu\_2 + Medu\_3 + Medu\_4,  
 data = gradedummy  
)  
  
summary(lm\_reduced\_FINAL)

##   
## Call:  
## lm(formula = G3^2 ~ G2 + G1 + famrel\_1 + famrel\_2 + famrel\_3 +   
## famrel\_4 + famrel\_5 + age + studytime\_1 + studytime\_2 + studytime\_3 +   
## studytime\_4 + Fjob\_services + Fjob\_at\_home + Fjob\_health +   
## Fjob\_other + Fjob\_teacher + Walc\_1 + Walc\_2 + Walc\_3 + Walc\_4 +   
## Walc\_5 + Medu\_1 + Medu\_2 + Medu\_3 + Medu\_4, data = gradedummy)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -93.175 -15.393 0.517 15.110 88.131   
##   
## Coefficients: (4 not defined because of singularities)  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -32.7200 29.7202 -1.101 0.27164   
## G2 13.9773 0.7641 18.292 < 2e-16 \*\*\*  
## G1 8.8067 0.8711 10.109 < 2e-16 \*\*\*  
## famrel\_1 -16.2970 10.6639 -1.528 0.12730   
## famrel\_2 -22.1800 7.3580 -3.014 0.00275 \*\*   
## famrel\_3 -11.3358 4.5630 -2.484 0.01342 \*   
## famrel\_4 -10.9686 3.4826 -3.150 0.00177 \*\*   
## famrel\_5 NA NA NA NA   
## age -3.3380 1.1929 -2.798 0.00540 \*\*   
## studytime\_1 11.9759 6.4381 1.860 0.06365 .   
## studytime\_2 2.0922 6.0313 0.347 0.72887   
## studytime\_3 5.1610 6.7018 0.770 0.44173   
## studytime\_4 NA NA NA NA   
## Fjob\_services -10.5974 6.4184 -1.651 0.09957 .   
## Fjob\_at\_home -7.3147 8.6219 -0.848 0.39677   
## Fjob\_health -6.4034 8.8226 -0.726 0.46842   
## Fjob\_other -3.2370 6.1466 -0.527 0.59876   
## Fjob\_teacher NA NA NA NA   
## Walc\_1 -0.6622 6.1213 -0.108 0.91391   
## Walc\_2 -9.7159 6.3476 -1.531 0.12671   
## Walc\_3 -8.0028 6.3983 -1.251 0.21180   
## Walc\_4 -8.5763 6.8749 -1.247 0.21300   
## Walc\_5 NA NA NA NA   
## Medu\_1 -15.1549 17.1881 -0.882 0.37850   
## Medu\_2 -19.7615 17.0168 -1.161 0.24627   
## Medu\_3 -8.8689 17.0358 -0.521 0.60295   
## Medu\_4 -12.6703 16.9833 -0.746 0.45611   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 28.47 on 372 degrees of freedom  
## Multiple R-squared: 0.8943, Adjusted R-squared: 0.888   
## F-statistic: 143 on 22 and 372 DF, p-value: < 2.2e-16

Diagnostic Plots for Final Model

par(mfrow=c(2,2))  
plot(  
 x = lm\_reduced\_FINAL,  
 pch = 19,  
 col = gradedummy$color  
)



Comparing the models

anova(lm\_reduced,lm\_reduced\_FINAL)

## Analysis of Variance Table  
##   
## Model 1: G3^2 ~ G2 + G1 + famrel\_5 + age + studytime\_1 + Fjob\_services +   
## Walc\_1 + Medu\_2  
## Model 2: G3^2 ~ G2 + G1 + famrel\_1 + famrel\_2 + famrel\_3 + famrel\_4 +   
## famrel\_5 + age + studytime\_1 + studytime\_2 + studytime\_3 +   
## studytime\_4 + Fjob\_services + Fjob\_at\_home + Fjob\_health +   
## Fjob\_other + Fjob\_teacher + Walc\_1 + Walc\_2 + Walc\_3 + Walc\_4 +   
## Walc\_5 + Medu\_1 + Medu\_2 + Medu\_3 + Medu\_4  
## Res.Df RSS Df Sum of Sq F Pr(>F)  
## 1 386 308580   
## 2 372 301421 14 7159.5 0.6311 0.8391

anova(lm\_initialsquared,lm\_reduced\_FINAL)

## Analysis of Variance Table  
##   
## Model 1: G3^2 ~ G1 + G2 + age + failures + absences  
## Model 2: G3^2 ~ G2 + G1 + famrel\_1 + famrel\_2 + famrel\_3 + famrel\_4 +   
## famrel\_5 + age + studytime\_1 + studytime\_2 + studytime\_3 +   
## studytime\_4 + Fjob\_services + Fjob\_at\_home + Fjob\_health +   
## Fjob\_other + Fjob\_teacher + Walc\_1 + Walc\_2 + Walc\_3 + Walc\_4 +   
## Walc\_5 + Medu\_1 + Medu\_2 + Medu\_3 + Medu\_4  
## Res.Df RSS Df Sum of Sq F Pr(>F)   
## 1 389 336802   
## 2 372 301421 17 35381 2.5686 0.0006468 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Predicted vs Observed

pred <- predict(lm\_reduced\_FINAL)  
  
dt\_prediction = (data.frame((sqrt(pred)),(gradedummy$G3)))

## Warning in sqrt(pred): NaNs produced

head(dt\_prediction,30)

## X.sqrt.pred.. X.gradedummy.G3.  
## 1 3.589381 6  
## 2 2.737057 6  
## 3 7.443364 10  
## 4 14.659051 15  
## 5 8.705532 10  
## 6 15.234848 15  
## 7 12.439552 11  
## 8 3.324327 6  
## 9 16.972440 19  
## 10 15.475368 15  
## 11 8.891672 9  
## 12 12.429983 12  
## 13 14.349997 14  
## 14 11.023543 11  
## 15 15.319006 16  
## 16 14.430232 14  
## 17 13.597151 14  
## 18 10.649982 10  
## 19 4.146491 5  
## 20 10.041313 10  
## 21 14.210311 15  
## 22 14.992981 15  
## 23 14.927996 16  
## 24 13.438043 12  
## 25 9.921576 8  
## 26 7.058646 8  
## 27 12.605969 11  
## 28 15.195356 15  
## 29 12.402404 11  
## 30 12.170429 11

plot(x=sqrt(pred),y=gradedummy$G3,xlab='Predicted Values',ylab='Actual Values',main='Predicted vs Observed')

## Warning in sqrt(pred): NaNs produced

abline(  
 reg = lsfit(  
 x = sqrt(pred),  
 y = gradedummy$G3  
 )  
)

## Warning in sqrt(pred): NaNs produced

## Warning in lsfit(x = sqrt(pred), y = gradedummy$G3): 19 missing values deleted

