Data Analysis for Students' Portuguese Grades Project Report

xinyuan chen

12/1/2021

Abstract

This data was collected by school reports and questionnaires which approached student achievements in secondary education of two Portuguese schools. Their Portuguese language course's final grades are shown in column G3. This project examines what features could be important predictors of their final grades. I choose OLS linear models and make model selections.

As a result, I found that in these two schools, these students' self wishes to take higher education, and their numbers of school absences could influence the final grade very much. And, their study environment could also influence grades, especially their mother's education level and whether they pay extra educational support. And finally, female students do significantly better than males and weekend alcohol consumption could influence the grades.

Summary of methods section

The first I do is to find the relationship between G3 and G1, G2. This is a multiply regression shown G3 has a strong correlation with G1, G2. By doing this, we can find some outliers that their final grades are zero. Even some of them their G1 or G2 are not zero, I think we can delete these poor unlucky students because maybe some mystery reasons lead to it.

The second part is to check whether the past failures could lead to their present final grades. The answer is yes. This is a simple OLS regression. More failures in the past tend to get lower grades present.

The third part is doing model selection for other regressors. There are too many regressors and factors so that I decided to use Forward Selection. I decided to divide the total data into train set sample and test set sample. Get the model from the train set and predict it on the test set. Also, test this model on the total data set and find outliers. Finally, do the heteroskedasticity test, normality test, and collinearity test. The result is the model seems to fit the data well. R-squared is 0.971.

Data Analysis

```
library(ISLR2)
library(leaps)
library(alr4)
```

```
## Loading required package: car
```

```
## Loading required package: carData
```

```
## Loading required package: effects
```

```
## lattice theme set by effectsTheme()
## See ?effectsTheme for details.
```

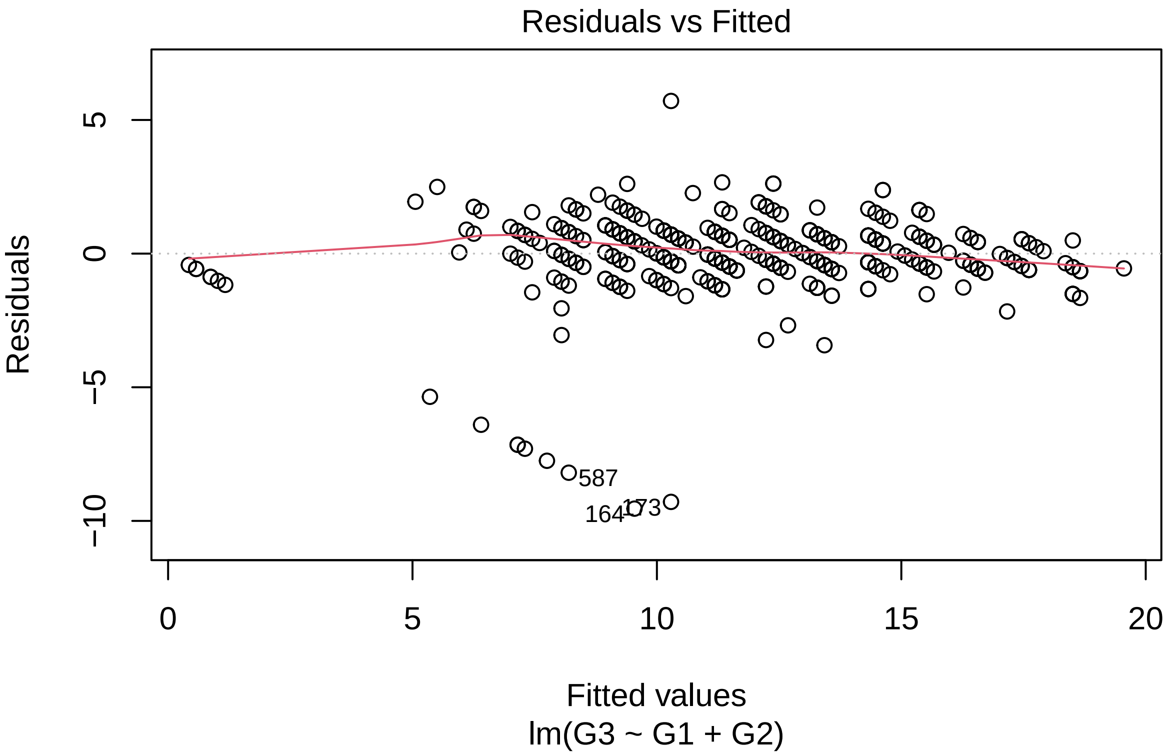
```
student_por = read.csv(file="student-por.csv",sep = ";")
factor_id = c("school","sex","address","famsize","Pstatus","Mjob","Fjob","reason","guardian","traveltim
student_por[,factor_id] <- lapply(student_por[,factor_id],factor)
```

The relation between G1,G2 and G3

```
lm_1 <- lm(G3~G1+G2,data=student_por)
summary(lm_1)
```

```
##
## Call:
## lm(formula = G3 ~ G1 + G2, data = student_por)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.5408 -0.4380 -0.0942  0.6296  5.7109
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.17128    0.21510  -0.796    0.426
## G1           0.14890    0.03600   4.136    4e-05 ***
## G2           0.89714    0.03392  26.448   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.262 on 646 degrees of freedom
## Multiple R-squared:  0.8478, Adjusted R-squared:  0.8473
## F-statistic:  1799 on 2 and 646 DF,  p-value: < 2.2e-16
```

```
plot(lm_1,1)
```



```
outlierTest(lm_1)
```

```
##      rstudent unadjusted p-value Bonferroni p
## 164 -7.931205         9.6002e-15   6.2305e-12
## 173 -7.689406         5.5393e-14   3.5950e-11
## 587 -6.725002         3.8733e-11   2.5138e-08
## 640 -6.368075         3.6376e-10   2.3608e-07
## 520 -5.950313         4.3914e-09   2.8500e-06
## 638 -5.823369         9.0995e-09   5.9056e-06
## 641 -5.823369         9.0995e-09   5.9056e-06
## 584 -5.193958         2.7641e-07   1.7939e-04
## 62   4.598158         5.1308e-06   3.3299e-03
## 627 -4.322633         1.7855e-05   1.1588e-02
```

The coefficient is 0.14890 and 0.89714, the intercept is -0.17. It seems that 10%G1 and 90%G2 and other little grades compose the final grade, G3.

For these 15 students who got 0 in G2 or G3, (some even got G1 and G2), but still zero points, just delete these unlucky students.

New data

```
student_por <- student_por[-c(164,441,520,564,568,583,587,584,598,604,606,611,627,638,640,641,173),]
```

Test outliners again in G3

```
lm_2 <- lm(G3 ~ G1+G2,data = student_por)
outlierTest(lm_2)
```

```
##      rstudent unadjusted p-value Bonferroni p
## 62   6.624344         7.5035e-11   4.7422e-08
## 63  -4.124874         4.2088e-05   2.6599e-02
## 280 -4.072726         5.2412e-05   3.3124e-02
```

Student 62,63,280’s G3 score is also very strange, they didn’t match G1 and G2. It is very hard to imagine they got higher 6 points or lower 4 points than other students should got. Unreasonable. Just delete them.

```
student_por <- student_por[-c(62,63,280),]
```

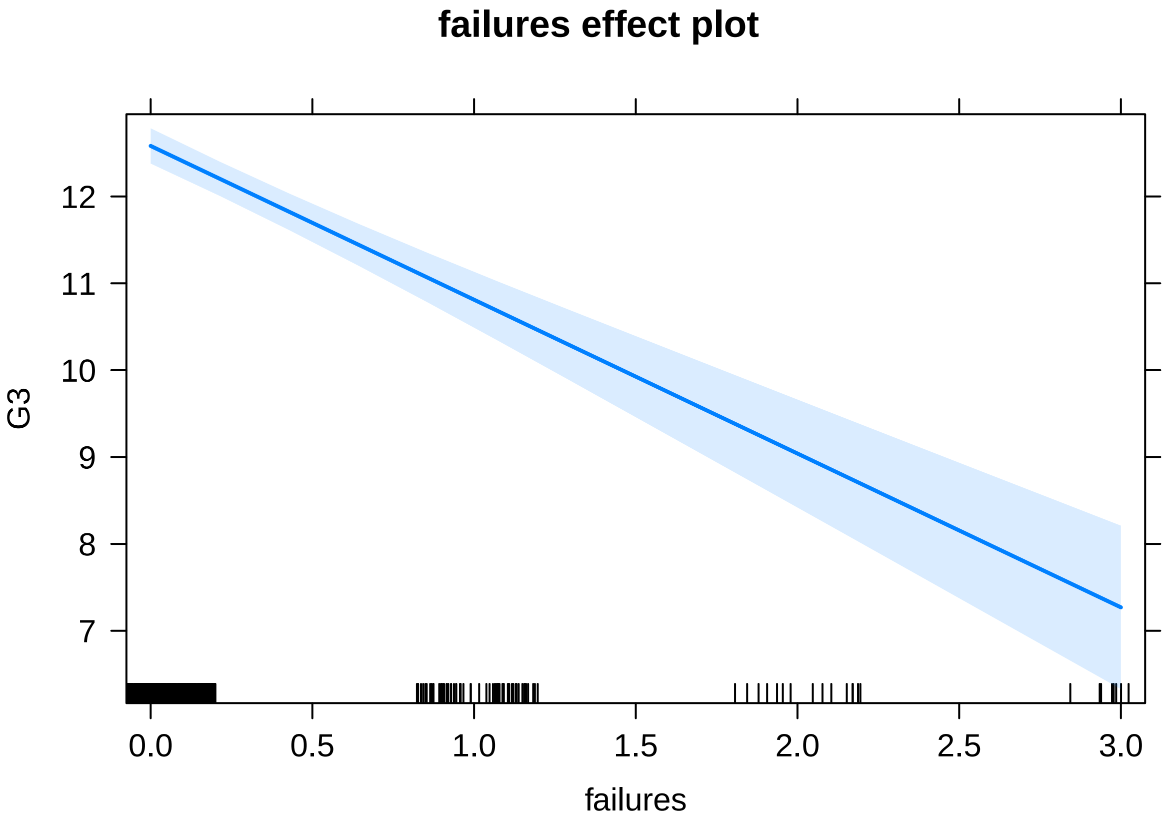
How does the past failures contribute to the final grade?

I notice that the failure column seems has a strong correlation with the grades. Check it.

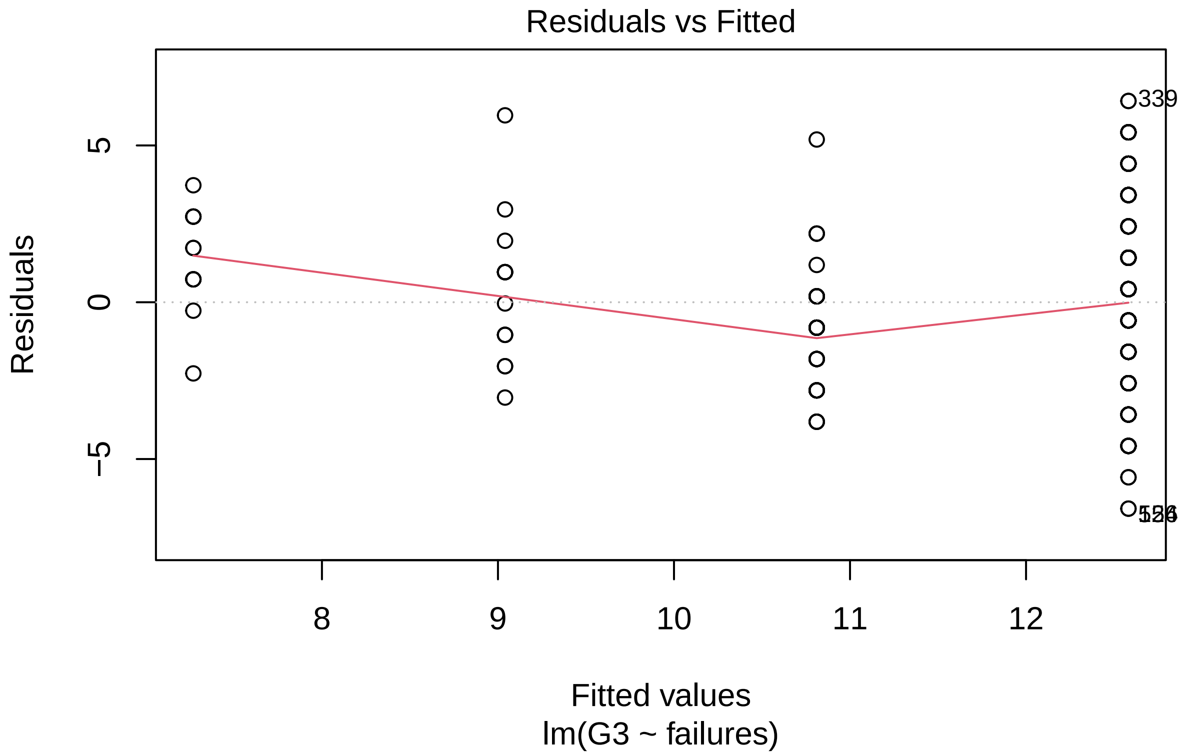
```
# Simple Linear Regression
lm_failure <- lm(G3~ failures,student_por)
summary(lm_failure)
```

```
##
## Call:
## lm(formula = G3 ~ failures, data = student_por)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.5818 -1.5818 -0.5818  1.4182  6.4182
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  12.5818     0.1036  121.45   <2e-16 ***
## failures     -1.7708     0.1681  -10.54   <2e-16 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 2.445 on 627 degrees of freedom
## Multiple R-squared:  0.1504, Adjusted R-squared:  0.1491
## F-statistic:   111 on 1 and 627 DF,  p-value: < 2.2e-16
```

```
plot(allEffects(lm_failure))
```



```
plot(lm_failure,1)
```



From the summary, we can see that students who got zero failure in the past tend to get a higher score in the final test, almost 12 points. The coefficient of failure is -1.77, which means other students, if failure 1 more time, their score tend to lower 2 points than those who did not fail this time. The "Failures" has a strong influence on the "Final Score".

Using "failure" could better predict the final grades, however, we are not here to predict the final grades. This is a questionnaire that focuses on students' grades and their study environments. I think we want is to find the relations between students' features and their final grades. So I delete the "failures" column, just focus on other regressors.

What would influence the final grade G3 ?

Build Train.set and Test.set

```
set.seed(1234)
train = sample(1:dim(student_por)[1], dim(student_por)[1] / 2)
test <- -train
student_por.train <- student_por[train, ]
student_por.test <- student_por[test, ]
```

Model Selection

```
#Forward Selection
m0 = lm(G3 ~0-G1-G2-failures, student_por.train)
m1 = lm(G3 ~.-G1-G2-failures, student_por.train)
A = step(object = m0, scope = list(lower=m0, upper=m1), direction = "forward", trace = F)
#B = step(object = m1, scope = list(lower=m0, upper=m1), direction = "backward", trace = F)
A # A B are same
```

```
##
## Call:
## lm(formula = G3 ~ school + higher + absences + schoolsup + Walc +
##     Medu + sex + paid + reason + studytime + activities - 1,
##     data = student_por.train)
##
## Coefficients:
##         schoolGP          schoolMS         higheryes          absences
##           9.9879            9.2200            1.9680           -0.1161
##     schoolsupyes              Walc              Medu              sexM
##          -1.5654           -0.2730            0.4195           -0.7308
##          paidyes        reasonhome       reasonother  reasonreputation
##          -1.1875            0.6638            0.5254            0.7729
##       studytime2        studytime3        studytime4     activitiesyes
##           0.7065            0.9089            0.5702            0.4499
```

we could include these regressors in OLS model: G3 ~ school + higher + absences + schoolsup + Walc + Medu + sex + paid + reason + studytime + activities - 1

OLS linear model in Train

```
lm_G3 <- lm(G3 ~ school + higher + absences + schoolsup + Walc +
    Medu + sex + paid + reason + studytime + activities - 1, data = student_por.train)
summary(lm_G3)
```

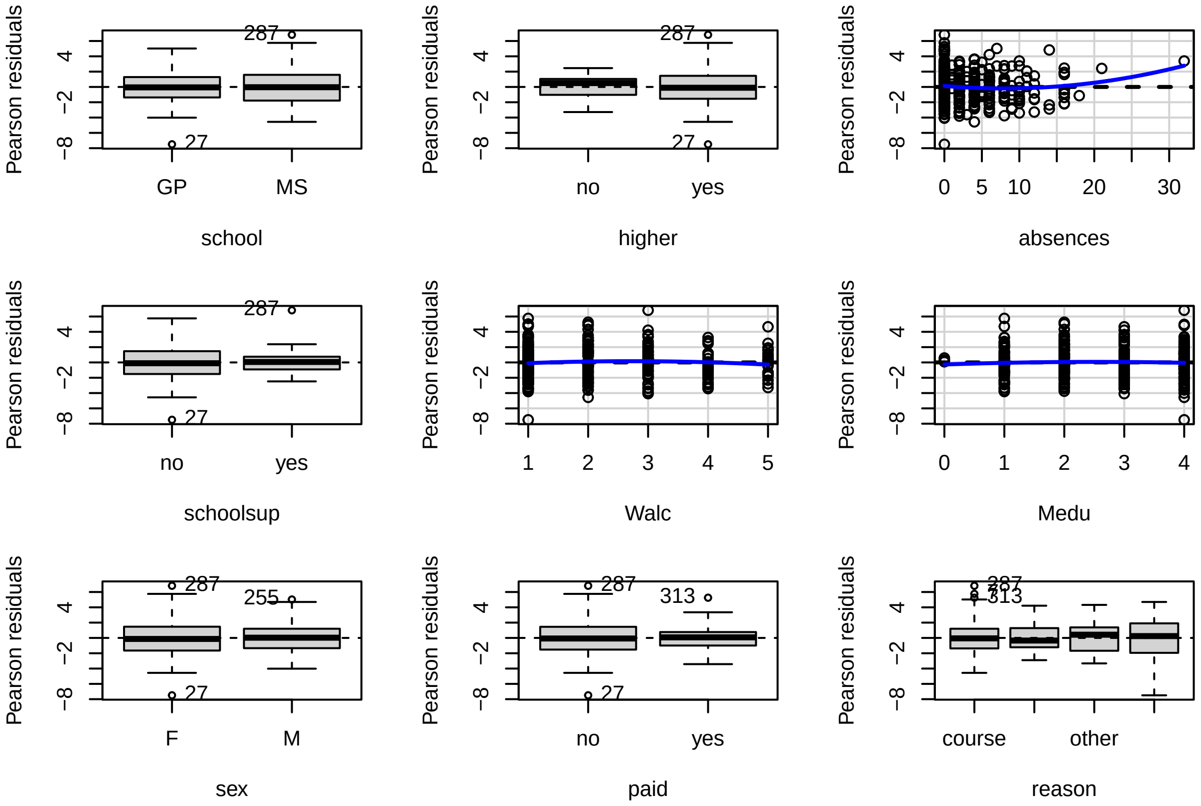
```
##
## Call:
## lm(formula = G3 ~ school + higher + absences + schoolsup + Walc +
##     Medu + sex + paid + reason + studytime + activities - 1,
##     data = student_por.train)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7.4925 -1.4771 -0.0585  1.3466  6.8120
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## schoolGP          9.98785    0.61056  16.358  < 2e-16 ***
## schoolMS          9.22005    0.54359  16.962  < 2e-16 ***
## higheryes         1.96804    0.44673   4.405 1.47e-05 ***
## absences         -0.11609    0.02808  -4.135 4.63e-05 ***
## schoolsupyes     -1.56542    0.38994  -4.015 7.54e-05 ***
## Walc             -0.27304    0.10357  -2.636 0.008824 **
## Medu              0.41949    0.11879   3.531 0.000479 ***
## sexM             -0.73081    0.28257  -2.586 0.010176 *
## paidyes          -1.18748    0.48741  -2.436 0.015424 *
## reasonhome        0.66383    0.32642   2.034 0.042874 *
## reasonother       0.52538    0.40460   1.299 0.195114
## reasonreputation  0.77285    0.33563   2.303 0.021984 *
## studytime2        0.70647    0.30195   2.340 0.019960 *
## studytime3        0.90895    0.40505   2.244 0.025565 *
## studytime4        0.57015    0.55969   1.019 0.309175
## activitiesyes     0.44990    0.25868   1.739 0.083037 .
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 2.166 on 298 degrees of freedom
## Multiple R-squared:  0.971,  Adjusted R-squared:  0.9694
## F-statistic: 622.6 on 16 and 298 DF,  p-value: < 2.2e-16
```

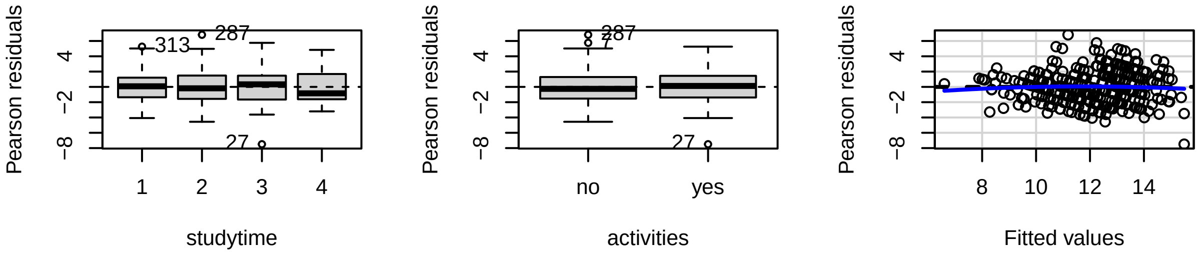
```
# checking for collinearity
vif(lm_G3) # No collinearity
```

```
## Warning in vif.default(lm_G3): No intercept: vifs may not be sensible.
```

```
##                 GVIF Df GVIF^(1/(2*Df))
## school     26.187326  2        2.262157
## higher     12.043312  1        3.470348
## absences    1.865041  1        1.365665
## schoolsup   1.199663  1        1.095291
## Walc        4.831331  1        2.198029
## Medu        7.519240  1        2.742123
## sex         2.060172  1        1.435330
## paid        1.114511  1        1.055704
## reason      2.945725  3        1.197288
## studytime   4.156657  3        1.268014
## activities  2.183222  1        1.477573
```

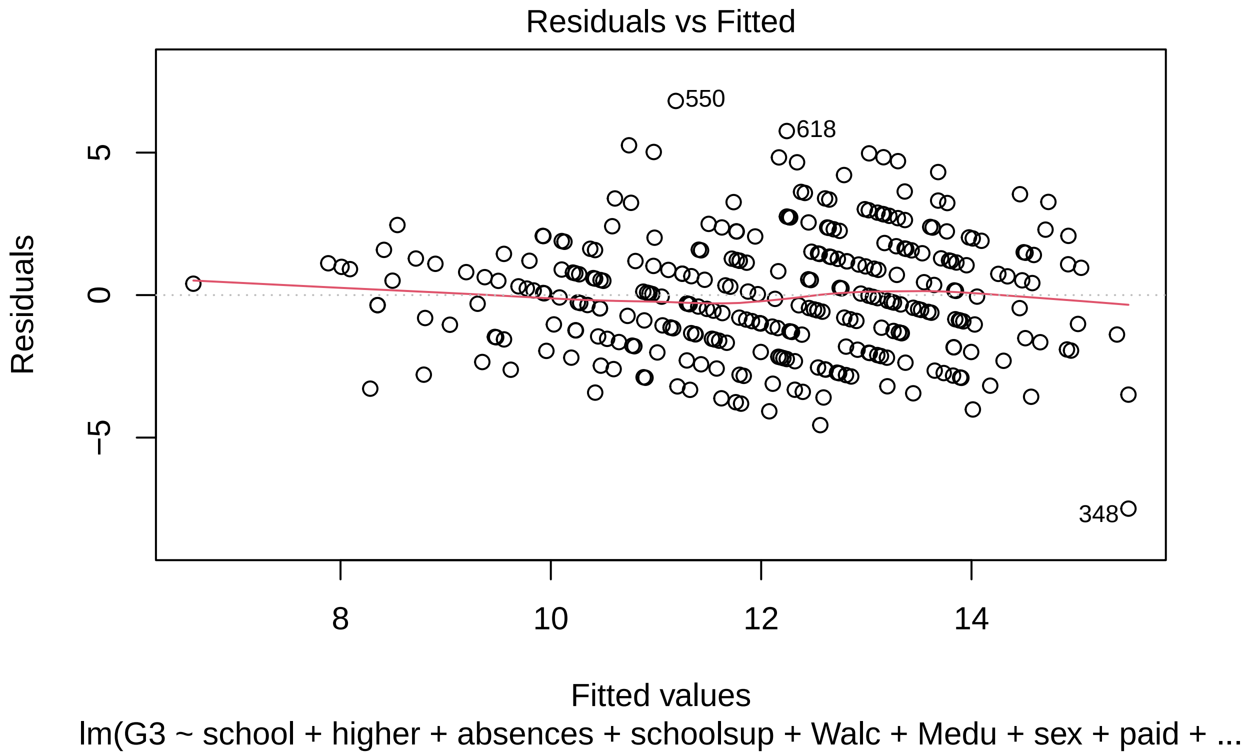
```
# check resudual plots
residualPlots(lm_G3)
```


Pearson residuals
−8
−2
4
287
27
GP
MS
school
no
yes
higher
0
5
10
20
30
absences
schoolsup
1
2
3
4
5
Walc
0
1
2
3
4
Medu
F
M
255
sex
313
paid
course
other
reason

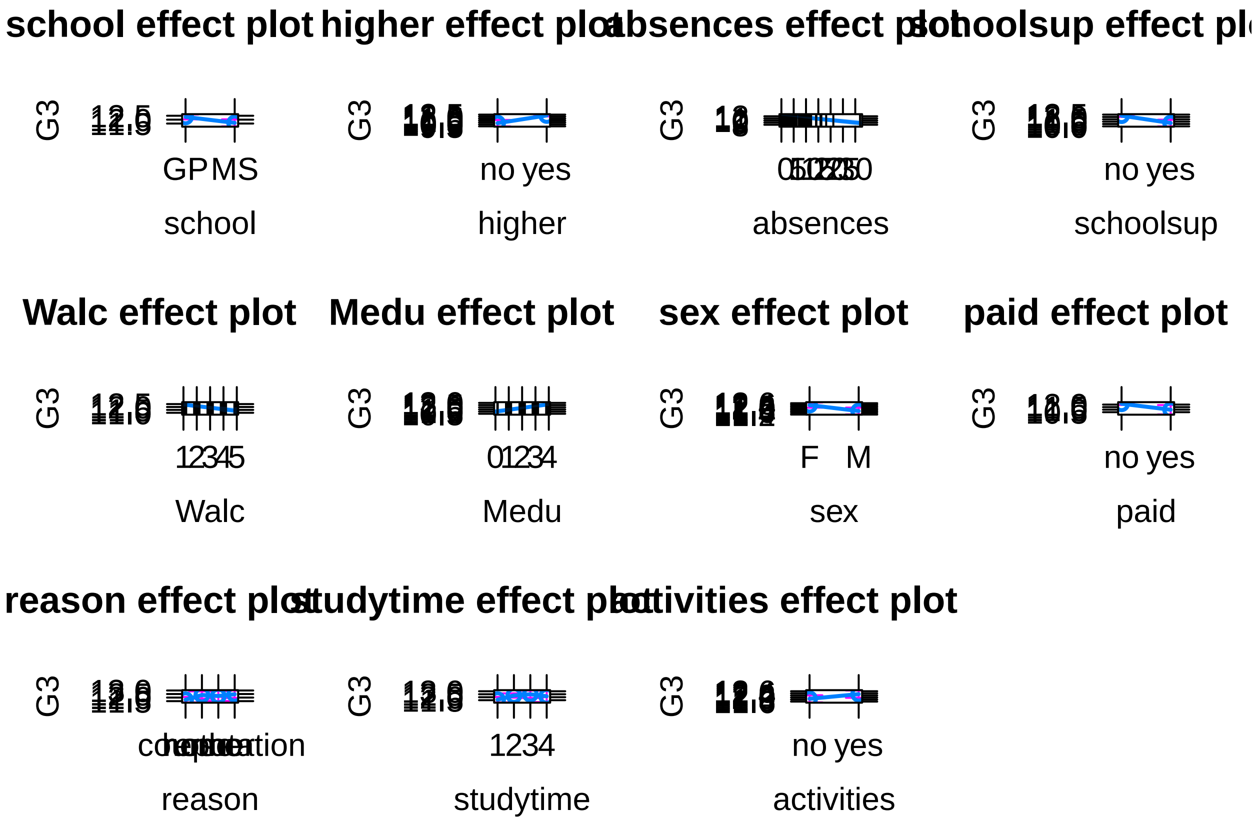


```
##              Test stat Pr(>|Test stat|)
## school
## higher
## absences        1.7242          0.08571 .
## schoolsup
## Walc          -1.1477          0.25201
## Medu          -0.5037          0.61487
## sex
## paid
## reason
## studytime
## activities
## Tukey test    -0.7694          0.44165
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
```

```
plot(lm_G3,1) # seems a null plot, no signficant patten, unbiased and homoscedastic
```



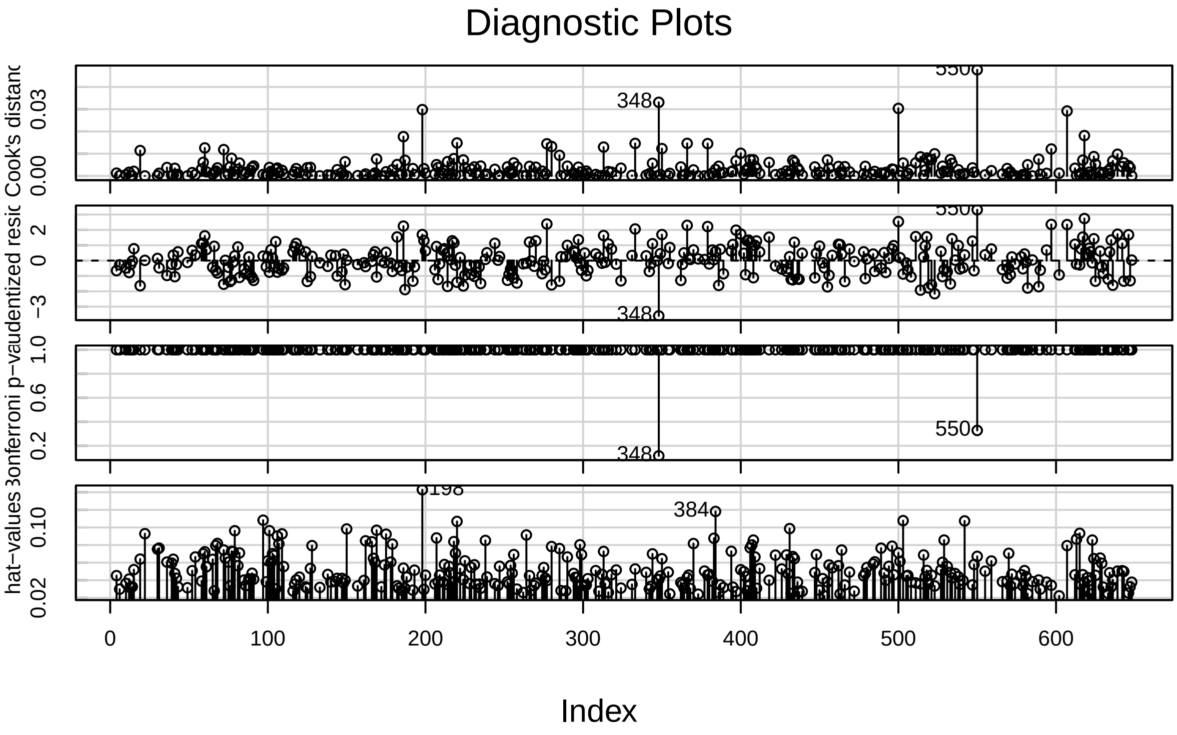
```
            #  the outliner seems including student 348

plot(allEffects(lm_G3))
```



```
# heteroskedasticity test:
ncvTest(lm_G3)# p is significant,implying that the variance is non-constant#
```

```
## Non-constant Variance Score Test
## Variance formula: ~ fitted.values
## Chisquare = 7.837596, Df = 1, p = 0.0051171
```

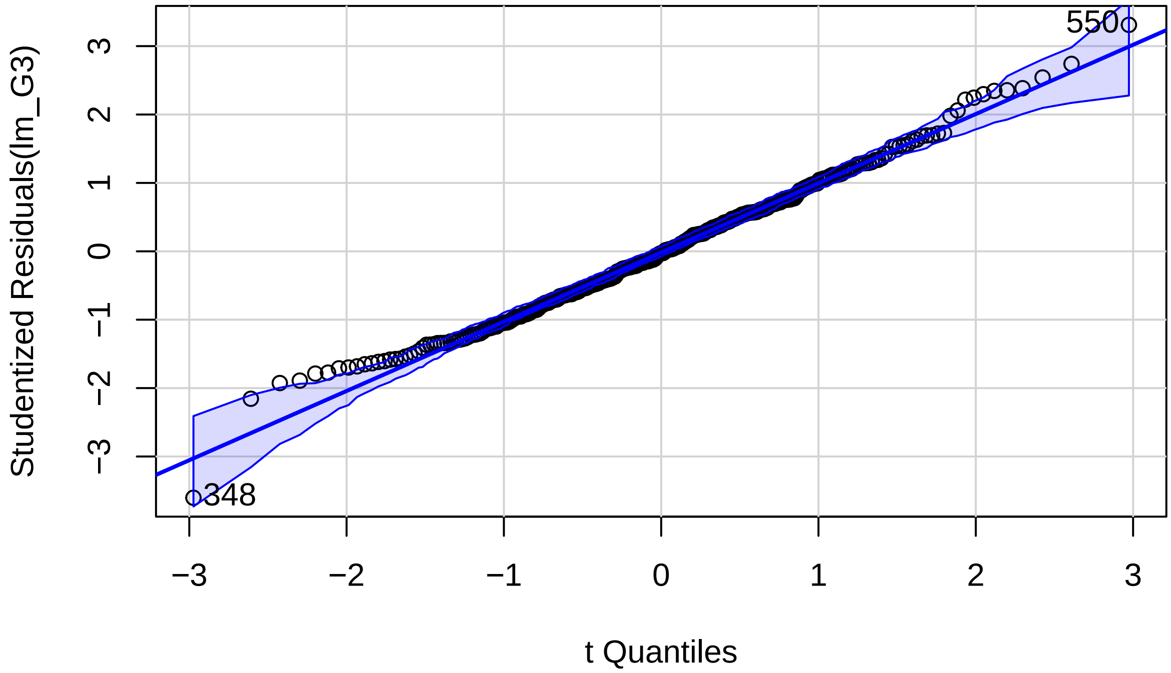
```
influenceIndexPlot(lm_G3) # seems no strong influence point, no high leveage point.
```



```
outlierTest(lm_G3)#348,(550)
```

```
## No Studentized residuals with Bonferroni p < 0.05
## Largest |rstudent|:
##      rstudent unadjusted p-value Bonferroni p
## 348 -3.602843         0.00036889      0.11583
```

```
# Normality test
qqPlot(lm_G3) #seems good
```



```
## 348 550
##  27 287
```

```
#cor(model.matrix(G3 ~ school + higher + absences + schoolsup + Walc + Medu + sex + paid + reason + stu
```

The vif test shows no collinearity. The residual plot seems a null plot, but ncvtest implies that the variance is non-constant. The outlier is 348,550. However, this is just a sample from the total students' data set, whether they are outliers, we should check the model on the whole data set.

Compare the predict with Test.set

```
pred.lm_G3 <- predict(lm_G3, student_por.test)
mean((pred.lm_G3 - student_por.test$G3)^2)
```

```
## [1] 5.453149
```

The predicts seems ok. It means the final predicts on test about bias 2 points higher or lower on average.

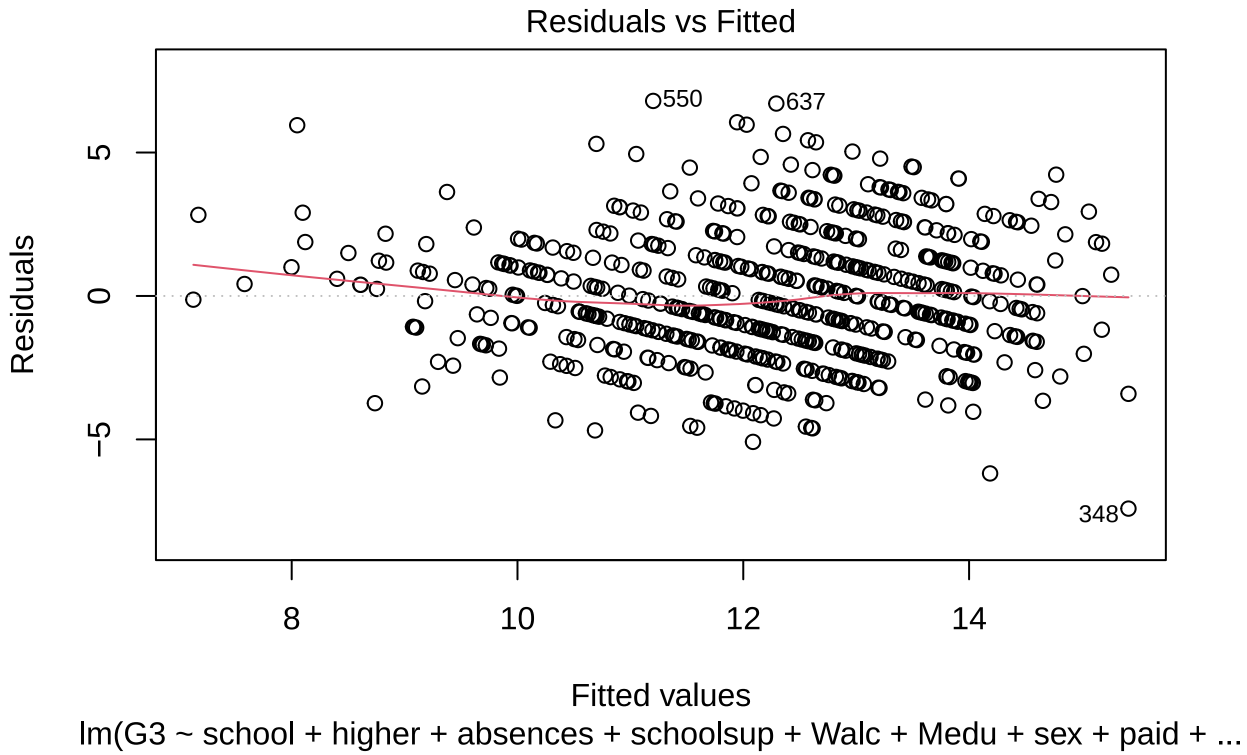
Use this model on the whole original dataset:

```r
lm_G3_all <- lm(G3 ~ school + higher + absences + schoolsup + Walc +
    Medu + sex + paid + reason + studytime + activities - 1, data = student_por)
compareCoefs(lm_G3, lm_G3_all, se = T)
```

```
## Calls:
## 1: lm(formula = G3 ~ school + higher + absences + schoolsup + Walc + Medu +
##   sex + paid + reason + studytime + activities - 1, data = student_por.train)
## 2: lm(formula = G3 ~ school + higher + absences + schoolsup + Walc + Medu +
##   sex + paid + reason + studytime + activities - 1, data = student_por)
##
##                   Model 1 Model 2
## schoolGP            9.988  10.266
## SE                  0.611   0.441
##
## schoolMS            9.220   9.368
## SE                  0.544   0.403
##
## higheryes           1.968   1.940
## SE                  0.447   0.314
##
## absences          -0.1161 -0.1031
## SE                 0.0281  0.0201
##
## schoolsupyes       -1.565  -1.568
## SE                  0.390   0.301
##
## Walc              -0.2730 -0.1975
## SE                 0.1036  0.0757
##
## Medu               0.4195  0.4064
## SE                 0.1188  0.0852
##
## sexM               -0.731  -0.573
## SE                  0.283   0.201
##
## paidyes            -1.187  -0.827
## SE                  0.487   0.380
##
## reasonhome          0.664   0.553
## SE                  0.326   0.234
##
## reasonother        0.5254  0.0504
## SE                 0.4046  0.3106
##
## reasonreputation    0.773   0.789
## SE                  0.336   0.242
##
## studytime2          0.706   0.429
## SE                  0.302   0.217
##
## studytime3          0.909   0.835
## SE                  0.405   0.294
##
```

```
## studytime4         0.570   0.755
## SE                 0.560   0.424
##
## activitiesyes      0.450   0.153
## SE                 0.259   0.184
##
```

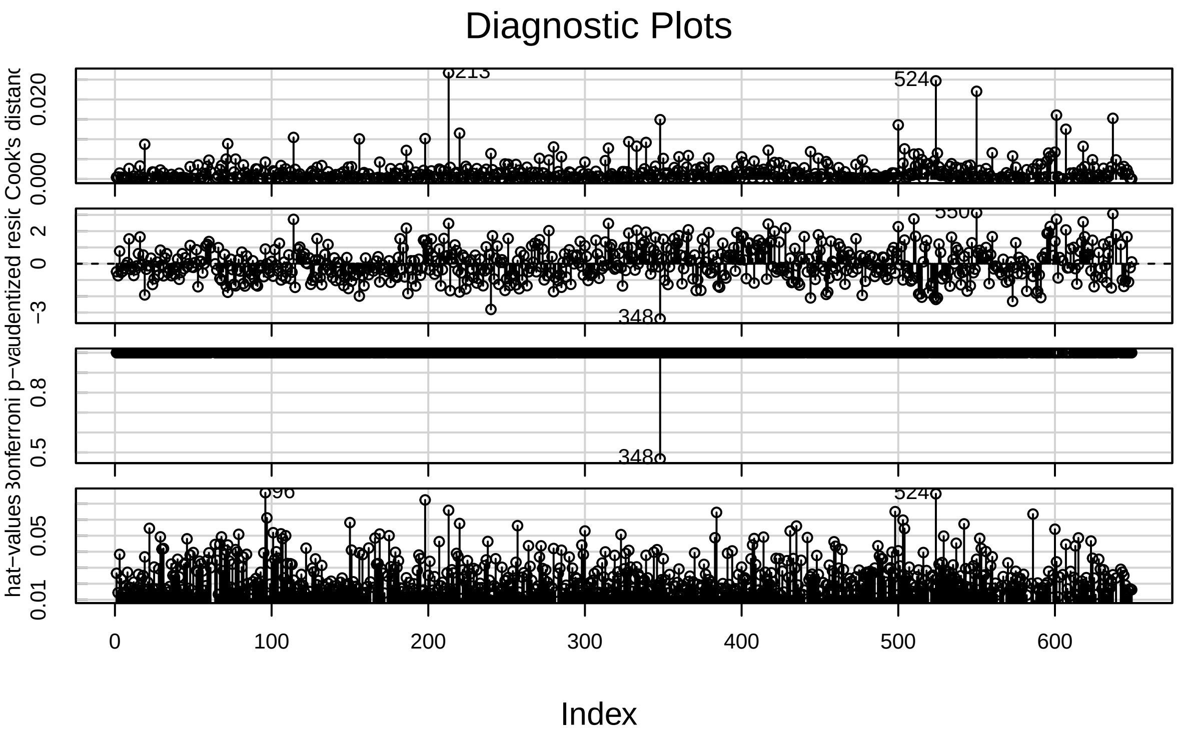
```
# check resudual plots
plot(lm_G3_all,1)
```



```
ncvTest(lm_G3_all) # non-constant
```

```
## Non-constant Variance Score Test
## Variance formula: ~ fitted.values
## Chisquare = 5.001101, Df = 1, p = 0.025331
```

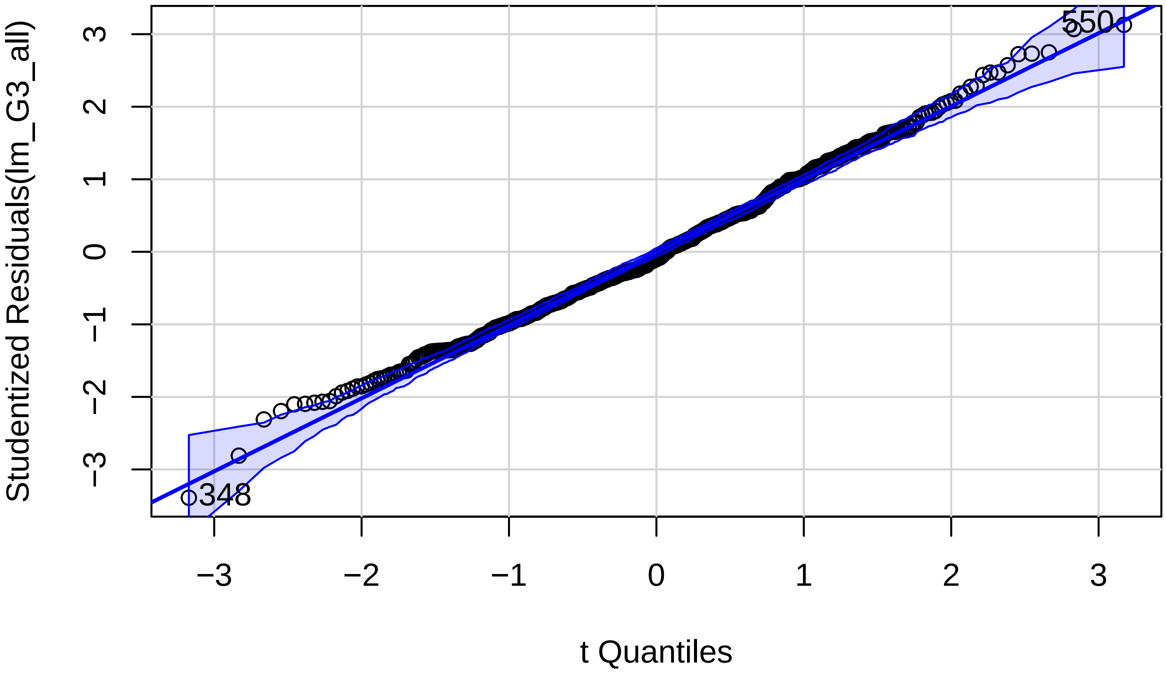
```
influenceIndexPlot(lm_G3_all) # seems no strong influence point, no high leveage point.
```



```
outlierTest(lm_G3_all) # 348, it is no reason to delete this student.
```

```
## No Studentized residuals with Bonferroni p < 0.05
## Largest |rstudent|:
##      rstudent unadjusted p-value Bonferroni p
## 348 -3.390128         0.00074355      0.46769
```

```
# Normality test
qqPlot(lm_G3_all)
```



```
## 348 550
## 343 543
```

From the residual plot, it seems a null plot, no significant pattern, unbiased and homoscedastic. But from the ncvTest, p is significant, implying that the variance is non-constant. It means those predicted grades are 12 tend to get a larger residual, this model is not so much perfect on predict when got 12 grades, the confident interval is about +- 2 points.

The outlier is still 348, and no strong influence point, no high leverage point. Their cook distance and hat values are small.

The Normality test is good. If not outlier 348, and 550, it may perform better. But unreasonable to delete these students.

What we conclude from this model?

```
lm_G3_all <- lm(formula = G3 ~ school + higher + absences + schoolsup + Walc + Medu + sex + paid + reas
summary(lm_G3_all)
```

```
##
## Call:
## lm(formula = G3 ~ school + higher + absences + schoolsup + Walc +
##     Medu + sex + paid + reason + studytime - 1, data = student_por)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7.3556 -1.5031 -0.2202  1.3358  6.7847
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## schoolGP         10.31835    0.43597  23.668  < 2e-16 ***
## schoolMS          9.41178    0.39917  23.578  < 2e-16 ***
## higheryes         1.93926    0.31368   6.182 1.15e-09 ***
## absences         -0.10334    0.02013  -5.133 3.82e-07 ***
## schoolsupyes     -1.57862    0.30017  -5.259 2.00e-07 ***
## Walc             -0.19698    0.07567  -2.603 0.009463 **
## Medu              0.41114    0.08500   4.837 1.67e-06 ***
## sexM             -0.55250    0.19964  -2.767 0.005820 **
## paidyes          -0.80355    0.37918  -2.119 0.034476 *
## reasonhome        0.53621    0.23285   2.303 0.021625 *
## reasonother       0.03356    0.30981   0.108 0.913770
## reasonreputation  0.80394    0.24116   3.334 0.000909 ***
## studytime2        0.43733    0.21636   2.021 0.043683 *
## studytime3        0.84649    0.29406   2.879 0.004133 **
## studytime4        0.77119    0.42348   1.821 0.069081 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.227 on 614 degrees of freedom
## Multiple R-squared:  0.969,  Adjusted R-squared:  0.9682
## F-statistic:  1279 on 15 and 614 DF,  p-value: < 2.2e-16
```

First, in those school-related features, we know school of GP tends to get a higher score than School MS, but in fact, they may have different score standards. These students' reason to choose school also significant on alpha = 0.01, those who choose school due to the "reputation" tend to get higher. So one of these schools may really do better than the other one, and deserve its reputation.

Second, in those school-related questionnaires, the students' self wishes to take higher education could influence the final grade very much. The wishes to receive education higher 1 point, the grade tend to get higher 2 points. And also, weekend alcohol consumption could influence the grades. The more drink every week, the fewer grades. I do not know why these teenagers under 21 could drink, any way, it could affect the final grade significantly.

Third, in those school reports, students' numbers of school absences influence the final grade significantly. The fewer absences, the higher grade. However, the "studytime" is not so much significant, which means not so much difference between the studytime is 5 - 10 and those 10+ hours. But study 5-10 hours perform well than those less than 2 hours, about 1 point score higher than them (if fit other values to an arbitrary value).

Fourth, in these students' demographic and environmental factors, their mother's education level and whether they pay extra educational support could influence the grade. And, female students do significantly better than males.

Fifth, in other factors, they are not so significant. For example, the activities may influence some students. In the train set, it is a little significant, but in the whole data set, it is not. So I exclude activities in the final model. Other factors, like romance, health, travel time, father's job, no evidence that they could influence the final grades.

R-squared is 0.969, it is very close to 1, means almost 97% of the variation in grades is explained by this model.