Student Performance Analysis

Richard Hou, Ahnaf Ayub,
Shihao Tong, Siqi Wang

Problem Context & Exploratory Data Analysis

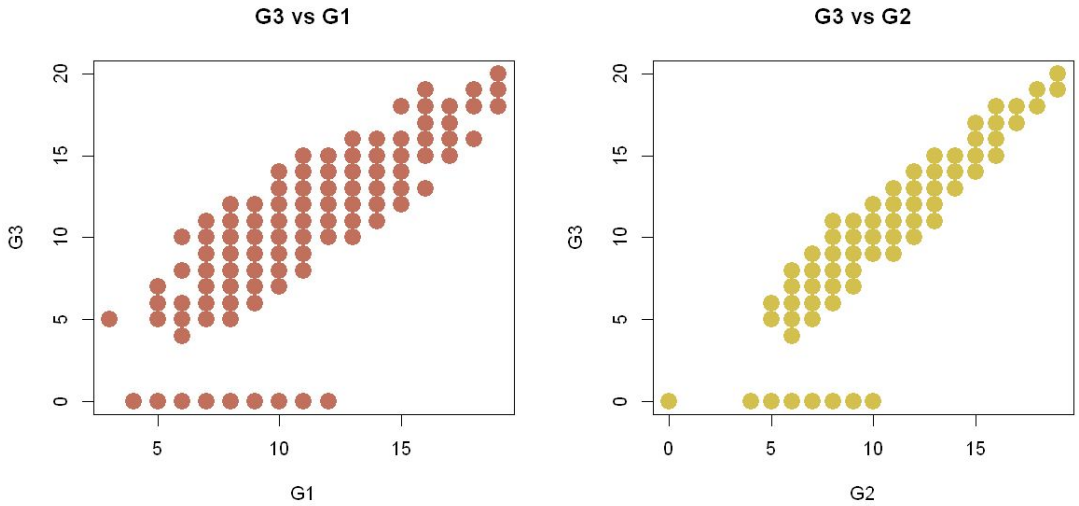
Problem Context

- Observational Study- Surveys and Questionnaires were used

- Null Hypothesis: All Regression Coefficients are Zero

Problem Context

Can we predict a student's **Final Course Grade** with Linear Regression?



The Data

Response Variable: G3 (ie. Final Grade)

Explanatory Variables:

- First/Second period grades (G1/G2)
- Age of Student
- Sex(F/M)
- Time spent studying(Weekly)
- Daily commute time

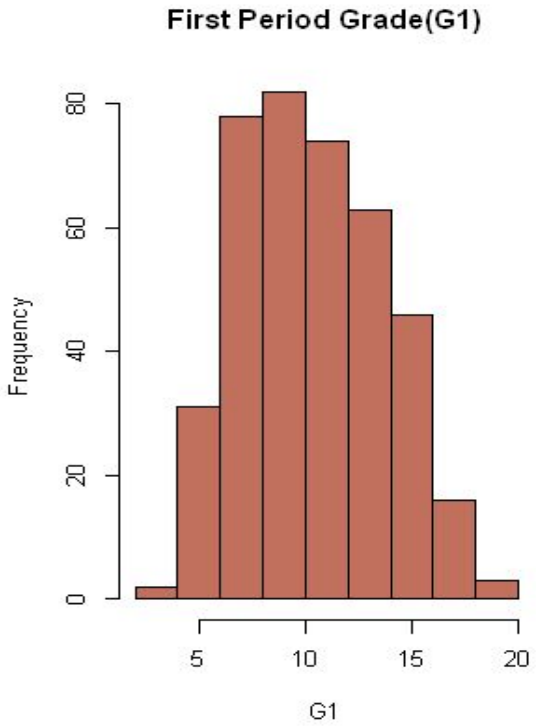
... (and 26 others)

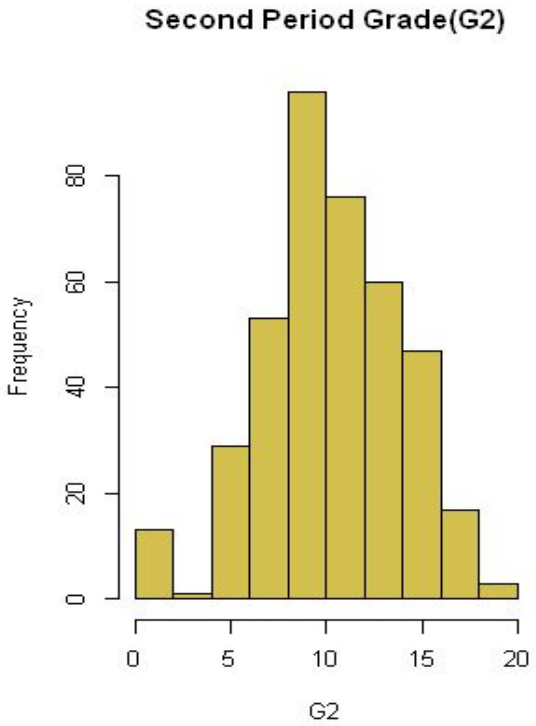
Attribute Information:

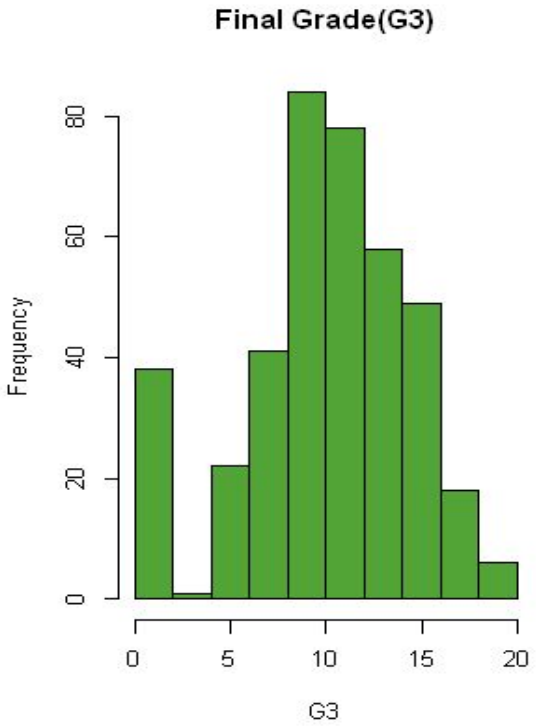
Attributes for both student-mat.csv (Math course) and student-
1 school - student's school (binary: 'GP' - Gabriel Pereira or 'MS
2 sex - student's sex (binary: 'F' - female or 'M' - male)
3 age - student's age (numeric: from 15 to 22)
4 address - student's home address type (binary: 'U' - urban or '
5 famsize - family size (binary: 'LE3' - less or equal to 3 or 'GT3'
6 Pstatus - parent's cohabitation status (binary: 'T' - living togeth
7 Medu - mother's education (numeric: 0 - none, 1 - primary edu
8 Fedu - father's education (numeric: 0 - none, 1 - primary educ
9 Mjob - mother's job (nominal: 'teacher', 'health' care related, ci
10 Fjob - father's job (nominal: 'teacher', 'health' care related, civ
11 reason - reason to choose this school (nominal: close to 'hom
12 guardian - student's guardian (nominal: 'mother', 'father' or 'o
13 traveltime - home to school travel time (numeric: 1 - <15 min.
14 studytime - weekly study time (numeric: 1 - <2 hours, 2 - 2 to
15 failures - number of past class failures (numeric: n if 1<=n<3,
16 schoolsup - extra educational support (binary: yes or no)
17 famsup - family educational support (binary: yes or no)
18 paid - extra paid classes within the course subject (Math or P
19 activities - extra-curricular activities (binary: yes or no)
20 nursery - attended nursery school (binary: yes or no)
21 higher - wants to take higher education (binary: yes or no)
22 internet - Internet access at home (binary: yes or no)
23 romantic - with a romantic relationship (binary: yes or no)
24 famrel - quality of family relationships (numeric: from 1 - very
25 freetime - free time after school (numeric: from 1 - very low to
26 goout - going out with friends (numeric: from 1 - very low to 5
27 Dalc - workday alcohol consumption (numeric: from 1 - very l
28 Walc - weekend alcohol consumption (numeric: from 1 - very
29 health - current health status (numeric: from 1 - very bad to 5
30 absences - number of school absences (numeric: from 0 to 9

these grades are related with the course subject, Math or Port
31 G1 - first period grade (numeric: from 0 to 20)
31 G2 - second period grade (numeric: from 0 to 20)
32 G3 - final grade (numeric: from 0 to 20, output target)

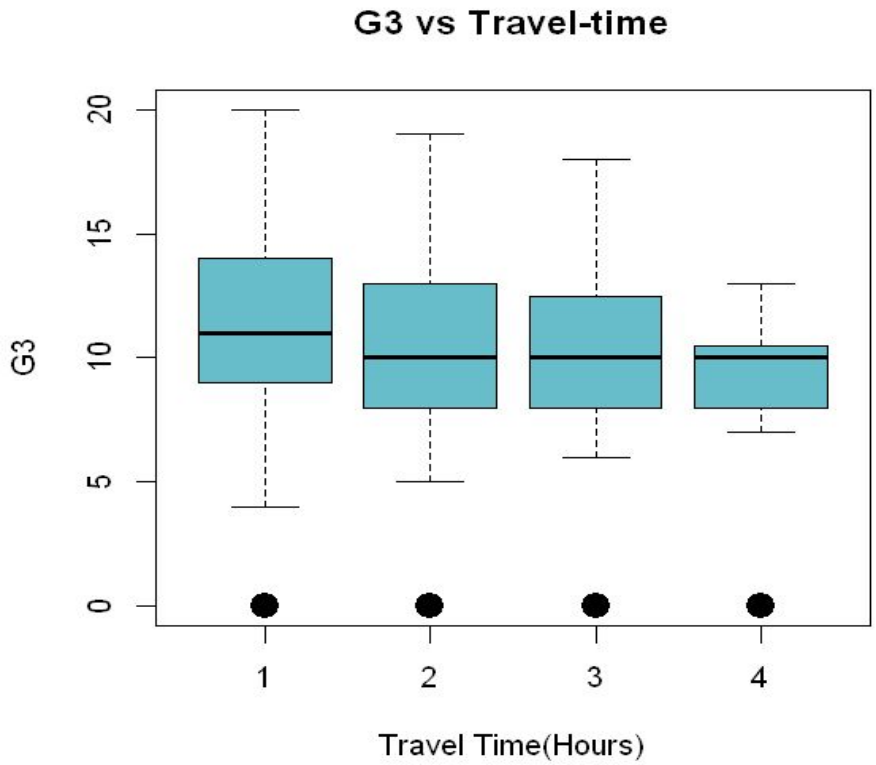
Underlying distributions

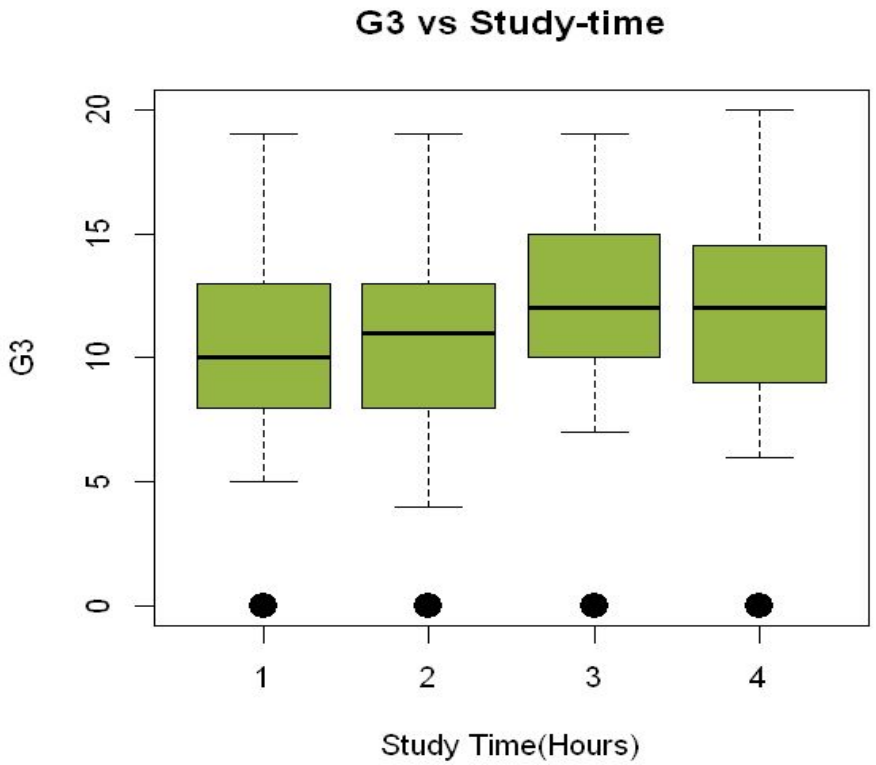




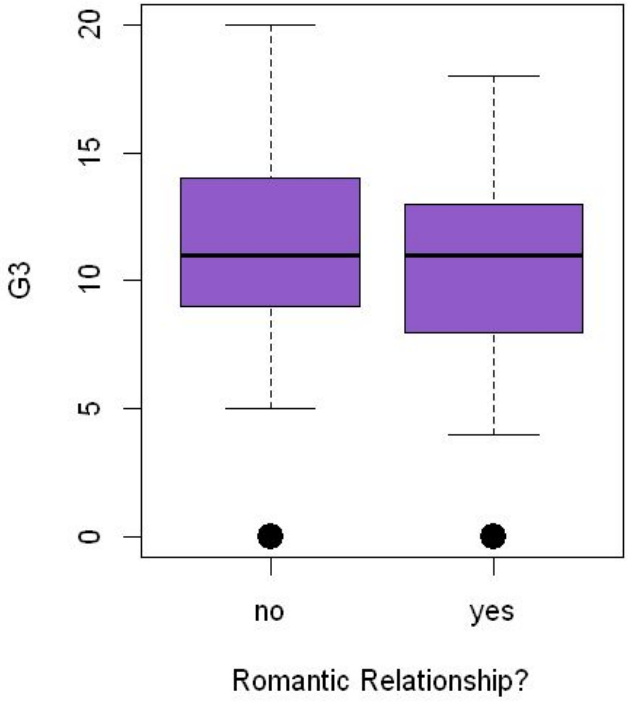


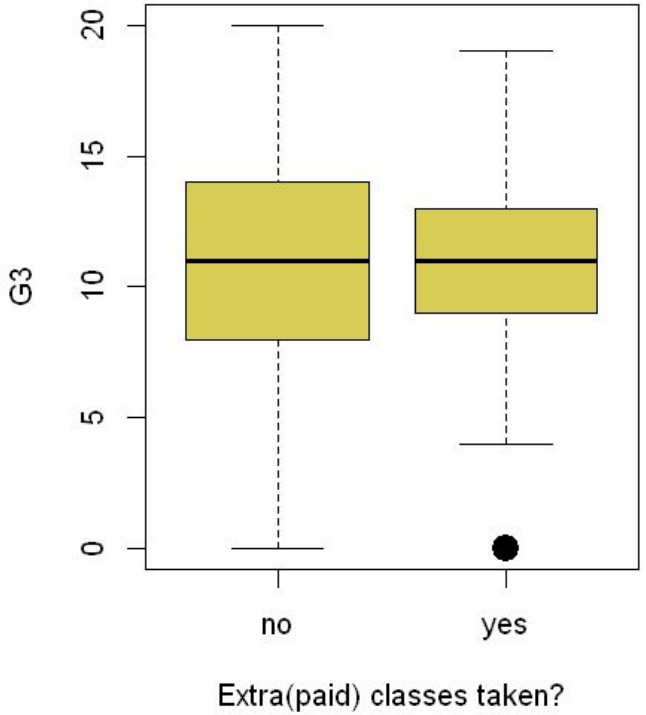
Travel-time and Study-time





More features...





Models and Methodology

Models and Methodology

Backward Elimination

- All the independent variables are entered into the equation first and each one is deleted one at a time if they do not contribute to the regression equation.

- Calculate the AIC(Akaike information criterion) and find the lowest score.

AIC

- AIC:an estimator of out-of-sample prediction error and estimates the quality of each model, relative to each of the other models

- AIC provides a means for model selection.

$$\mathrm{AIC} = 2k - 2\ln(\hat{L})$$

Backward Elimination

Code:

Backward Elimination

```
reg_ob <- lm(G3~., data=df)

step(reg_ob, direction="backward")

#Find the final set of features

reg_2 <- lm( G3 ~ school + age + activities + romantic + famrel + Walc +

absences + G1 + G2, data = df)

summary(reg_2)
```

```
Call:
lm(formula = G3 ~ school + age + activities + romantic + famrel +
    Walc + absences + G1 + G2, data = df)

Residuals:
    Min      1Q  Median      3Q     Max
-8.8416 -0.4534  0.2645  1.0247  4.0315

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   0.65012    1.45806   0.446 0.655933
schoolMS      0.45495    0.32293   1.409 0.159693
age          -0.26017    0.08405  -3.095 0.002109 **
activitiesyes -0.32049    0.18989  -1.688 0.092278 .
romanticyes  -0.32668    0.20622  -1.584 0.113989
famrel        0.38361    0.10668   3.596 0.000365 ***
Walc          0.11979    0.07521   1.593 0.112050
absences      0.04714    0.01232   3.826 0.000152 ***
G1            0.18058    0.05517   3.273 0.001159 **
G2            0.96153    0.04904  19.607  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.864 on 385 degrees of freedom
Multiple R-squared:  0.8382,    Adjusted R-squared:  0.8344
F-statistic: 221.6 on 9 and 385 DF,  p-value: < 2.2e-16
```

Models and Methodology

Forward Selection

Start with a NULL model with no variable and gonging forward to chose the one with lowest AIC.

```
basemodel <- lm(G3~NULL, data = d1)

model.forward <- step(basemodel, direction = "forward", trace = F, scope = ~ school + sex + age +
                        Fedu + Mjob + Fjob + reason + guardian + traveltime + studytime +
                        failures + schoolsup + famsup + paid + activities + nursery +
                        higher + internet + romantic + famrel + freetime + goout +
                        Dalc + Walc + health + absences + G1 + G2)
summary(model.forward)
```

Forward Selection

```
Call:
lm(formula = G3 ~ G2 + famrel + absences + G1 + age + activities +
    Walc + romantic + school, data = d1)

Residuals:
    Min      1Q  Median      3Q     Max
-8.8416 -0.4534  0.2645  1.0247  4.0315

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   0.65012    1.45806   0.446 0.655933
G2            0.96153    0.04904  19.607  < 2e-16 ***
famrel        0.38361    0.10668   3.596 0.000365 ***
absences      0.04714    0.01232   3.826 0.000152 ***
G1            0.18058    0.05517   3.273 0.001159 **
age          -0.26017    0.08405  -3.095 0.002109 **
activitiesyes -0.32049    0.18989  -1.688 0.092278 .
Walc          0.11979    0.07521   1.593 0.112050
romanticyes  -0.32668    0.20622  -1.584 0.113989
schoolMS      0.45495    0.32293   1.409 0.159693
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.864 on 385 degrees of freedom
Multiple R-squared:  0.8382,    Adjusted R-squared:  0.8344
F-statistic: 221.6 on 9 and 385 DF,  p-value: < 2.2e-16
```

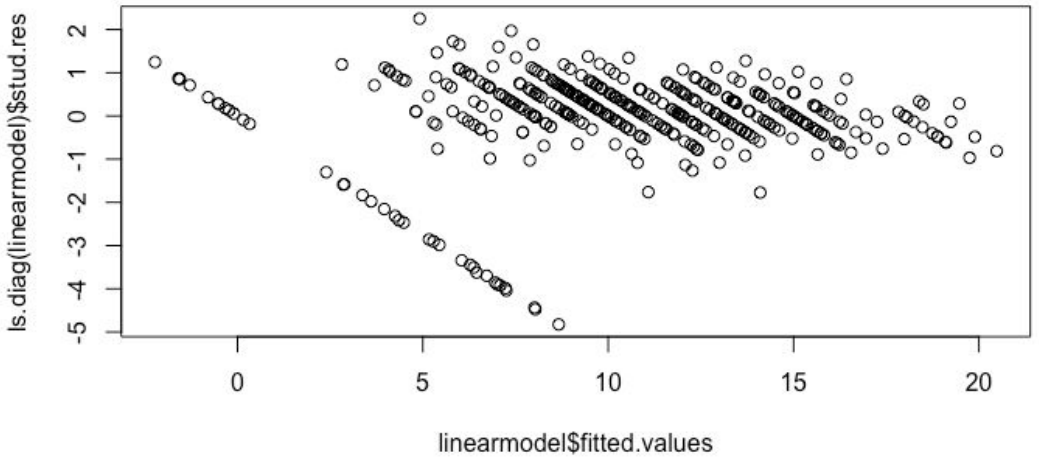
Adjusted R-square: 0.8344

Interpretation of Results & Conclusion

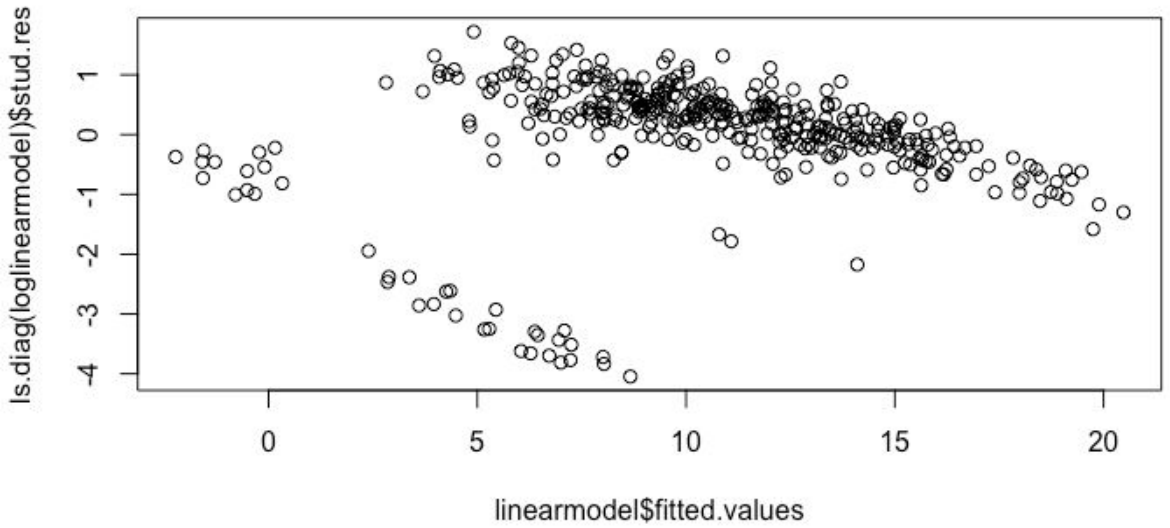
Interpretation of Results

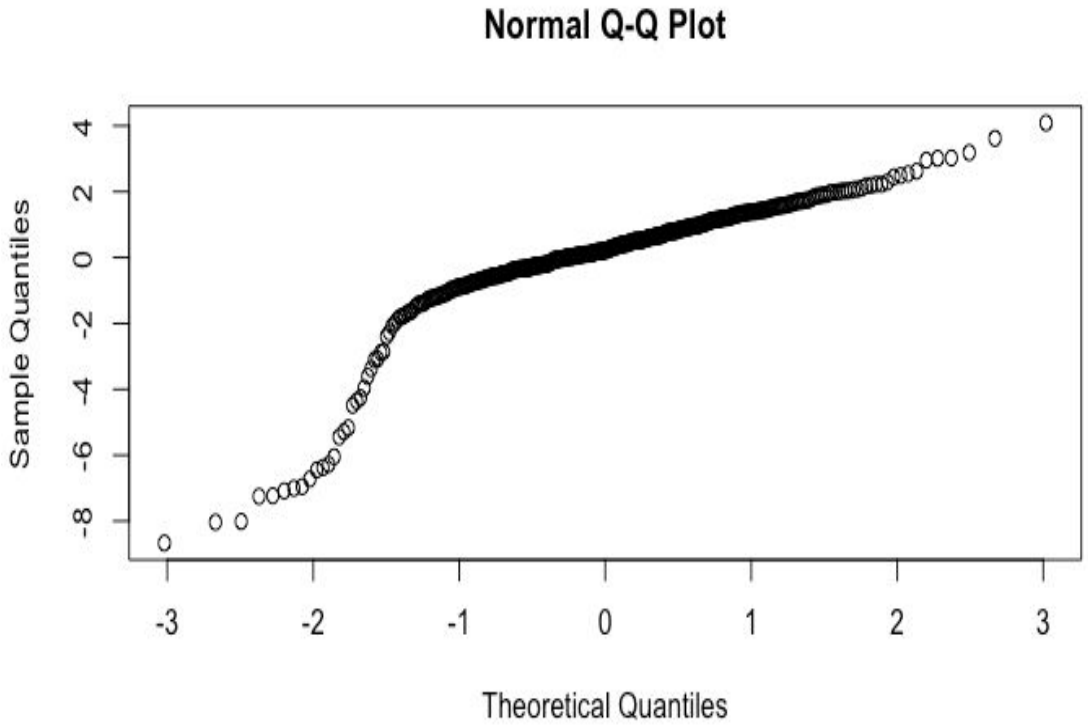
$$\hat{G3} = 0.65012 + 0.45495 * schoolMS - 0.26017 * age - 0.32049 * activity - 0.32668 * romantic + 0.38361 * famrel + 0.11979 * Walc + 0.04714 * absences + 0.18058 * G1 + 0.96153 * G2$$

adjR^2 = 0.8344



Log-linear model:
adjR^2 = 0.6099





The effect of a p-unit increase in explanatory variable X_j is to multiply the mean of y by exp(pβj).

Conclusion

We started with 32 features, went down to ... 9 features

We only did this for Mathematics grades

A similar approach can be taken for Portuguese data

Reference

Citation: P. Cortez and A. Silva. Using Data Mining to Predict Secondary School Student Performance. In A. Brito and J. Teixeira Eds., Proceedings of 5th FUture BUsiness TEChnology Conference (FUBUTEC 2008) pp. 5-12, Porto, Portugal, April, 2008, EUROSIS, ISBN 978-9077381-39-7.