Statistic Model to Analyze Student's Performance - Group 8

Xuchuan Zheng, Sungki Park, Prashant Sharma

27 March, 2024

1. *Introduction*

In out final project for Data 603 - Statistical Modelling with Data, we have tried to develop a model to analyze the impact of various demographic and social factors on the performance of students. Academic performance, though it is not the only factor but is one of the crucial factors in shaping a student's future. To get into a good collage/university, student must score grades in school, a good college ca lead a better future and economic stability. So in order to secure good grades, getting into a great school is enough? Is there something more than a great school that can help a student to perform better? Do the social and demographic factors plays any role in student's performance? In our project we are trying to answer these question.

To answer these questions we are working with a dataset that is collected at 2 Portuguese schools for Mathematics and Portuguese subject. This data is collected by using school reports and questionnaires. The data attribute include students grades, family size information, education level of parents, free time of student, any many other factors. By working on this project we are hoping to develop more understanding about the factors which can impact the performance of a student.

2. *Data*

This data is from UC Irvine Machine Learning Repository. There are 649 rows instances and 30 features in the dataset. Below are details of each feature

1. school - student's school (binary: 'GP' - Gabriel Pereira or 'MS' - Mousinho da Silveira) [Qualitative]
2. sex - student's sex (binary: 'F' - female or 'M' - male) [Qualitative]
3. age - student's age (numeric: from 15 to 22)
4. address - student's home address type (binary: 'U' - urban or 'R' - rural) [Qualitative]
5. famsize - family size (binary: 'LE3' - less or equal to 3 or 'GT3' - greater than 3) [Qualitative]
6. Pstatus - parent's cohabitation status (binary: 'T' - living together or 'A' - apart) [Qualitative]
7. Medu - mother's education (numeric: 0 - none, 1 - primary education (4th grade), 2 - 5th to 9th grade, 3 - secondary education or 4 - higher education) [Qualitative]
8. Fedu - father's education (numeric: 0 - none, 1 - primary education (4th grade), 2 - 5th to 9th grade, 3 - secondary education or 4 - higher education) [Qualitative]
9. Mjob - mother's job (nominal: 'teacher', 'health' care related, civil 'services' (e.g. administrative or police), 'at_home' or 'other') [Qualitative]

10. Fjob - father's job (nominal: 'teacher', 'health' care related, civil 'services' (e.g. administrative or police), 'at_home' or 'other') [Qualitative]
11. reason - reason to choose this school (nominal: close to 'home', school 'reputation', 'course' preference or 'other') [Qualitative]
12. guardian - student's guardian (nominal: 'mother', 'father' or 'other') [Qualitative]
13. traveltime - home to school travel time (numeric: 1 - <15 min., 2 - 15 to 30 min., 3 - 30 min. to 1 hour, or 4 - >1 hour) [Qualitative]
14. studytime - weekly study time (numeric: 1 - <2 hours, 2 - 2 to 5 hours, 3 - 5 to 10 hours, or 4 - >10 hours) [Qualitative]
15. failures - number of past class failures (numeric: n if 1<=n<3, else 4) [Qualitative]
16. schoolsup - extra educational support (binary: yes or no) [Qualitative]
17. famsup - family educational support (binary: yes or no) [Qualitative]
18. paid - extra paid classes within the course subject (Math or Portuguese) (binary: yes or no) [Qualitative]
19. activities - extra-curricular activities (binary: yes or no) [Qualitative]
20. nursery - attended nursery school (binary: yes or no) [Qualitative]
21. higher - wants to take higher education (binary: yes or no) [Qualitative]
22. internet - Internet access at home (binary: yes or no) [Qualitative]
23. romantic - with a romantic relationship (binary: yes or no) [Qualitative]
24. famrel - quality of family relationships (numeric: from 1 - very bad to 5 - excellent) [Qualitative]
25. freetime - free time after school (numeric: from 1 - very low to 5 - very high) [Qualitative]
26. goout - going out with friends (numeric: from 1 - very low to 5 - very high) [Qualitative]
27. Dalc - workday alcohol consumption (numeric: from 1 - very low to 5 - very high) [Qualitative]
28. Walc - weekend alcohol consumption (numeric: from 1 - very low to 5 - very high) [Qualitative]
29. health - current health status (numeric: from 1 - very bad to 5 - very good) [Qualitative]
30. absences - number of school absences (numeric: from 0 to 93) [Quantitative]
31. G1 - first period grade (numeric: from 0 to 20) [Quantitative]
32. G2 - second period grade (numeric: from 0 to 20) [Quantitative]
33. G3 - final grade (numeric: from 0 to 20, output target) [Quantitative]

G3 (Final grade) is the dependent variable for our model.

NOTE: We need to convert the qualitative variable from numeric to string

3. *Methodology*

Below is the outline of the steps we are going to perform in our analysis:

1. First Build a full additive model.
2. We will apply some model selection technique to come up with the best additive model.
3. Based on p-value (assuming $\alpha = 0.05$) we will drop variable which are non-significant.

4. Perform partial F-test to verify that dropped variables are indeed non-significant.
5. Provide interpretation for the best additive model to predict our dependent variable (G3 - Final Grades).
6. Based on our best additive model, we will check for interaction between the variables.
7. Using p-value (assuming $\alpha = 0.05$), we will drop the non-significant interaction terms.
8. Use partial F-test and analysis of Variance to verify the usability of our best interaction model.
9. Provide interpretation of our best interaction model.
10. Then we will check if we can include any higher order term in our model (Moving towards Higher order multiple regression model).
11. Verify the significance of higher order terms using p-value (assuming $\alpha = 0.05$).
12. Once we have done all our analysis we will try to define our best regression model (liner or higher order) to predict our dependent variable (G3 - Final Grades).
13. Using our final regression model, we will start checking the regression assumptions.
14. Linearity Assumption.
15. Independence Assumption.
16. Normality Assumption.
17. Multi-collinearity.
18. Outliers.

Starting our analysis with building additive model, then we will try to include interaction terms and higher order terms in our model.

3.1 *Full Additive Model*

Creating full additive model:

```
studentPerformance_fm = lm(G3 ~(school+sex+age+address+famsize+Pstatus+Medu+
                                  Fedu+Mjob+Fjob+reason+guardian+traveltime+studytime+
                                  failures+schoolsup+famsup+activities+nursery+higher+internet+
                                  romantic+famrel+freetime+
                                  goout+Dalc+Walc+health+absences),
                           data = studentDataset)

summary(studentPerformance_fm)
```

```
##
## Call:
## lm(formula = G3 ~ (school + sex + age + address + famsize + Pstatus +
##     Medu + Fedu + Mjob + Fjob + reason + guardian + traveltime +
##     studytime + failures + schoolsup + famsup + activities +
##     nursery + higher + internet + romantic + famrel + freetime +
##     goout + Dalc + Walc + health + absences), data = studentDataset)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -12.8073  -1.3732  -0.0209   1.5406   7.7138
```

```
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)        8.68811    1.98474   4.377 1.41e-05 ***
## schoolMS          -1.20316    0.26722  -4.502 8.05e-06 ***
## sexM              -0.64629    0.24941  -2.591 0.009791 **
## age                0.15741    0.10215   1.541 0.123838
## addressU           0.33409    0.26136   1.278 0.201629
## famsizeLE3         0.31094    0.24473   1.271 0.204381
## PstatusT           0.17592    0.34659   0.508 0.611934
## Medu               0.02475    0.15073   0.164 0.869617
## Fedu               0.16082    0.13752   1.169 0.242689
## Mjobhealth         0.92045    0.53684   1.715 0.086930 .
## Mjobother          0.05760    0.30271   0.190 0.849160
## Mjobservices       0.42725    0.37289   1.146 0.252341
## Mjobteacher        0.54186    0.50038   1.083 0.279289
## Fjobhealth        -0.58529    0.75139  -0.779 0.436317
## Fjobother         -0.17803    0.45599  -0.390 0.696354
## Fjobservices      -0.63314    0.47892  -1.322 0.186663
## Fjobteacher        0.61522    0.67059   0.917 0.359284
## reasonhome         0.04053    0.28456   0.142 0.886777
## reasonother       -0.44517    0.36731  -1.212 0.225985
## reasonreputation   0.23506    0.29713   0.791 0.429186
## guardianmother    -0.35229    0.26453  -1.332 0.183432
## guardianother      0.07699    0.53039   0.145 0.884642
## traveltime         0.06776    0.15897   0.426 0.670075
## studytime          0.40587    0.13990   2.901 0.003852 **
## failures          -1.42486    0.20384  -6.990 7.23e-12 ***
## schoolsupyes      -1.32197    0.36370  -3.635 0.000302 ***
## famsupyes         -0.03548    0.22745  -0.156 0.876076
## activitiesyes      0.20670    0.22281   0.928 0.353930
## nurseryyes        -0.22177    0.27122  -0.818 0.413875
## higheryes          1.72741    0.38257   4.515 7.59e-06 ***
## internetyes        0.24860    0.27618   0.900 0.368401
## romanticyes       -0.42852    0.22913  -1.870 0.061928 .
## famrel             0.15906    0.11604   1.371 0.170972
## freetime          -0.13026    0.11192  -1.164 0.244911
## goout             -0.06546    0.10745  -0.609 0.542577
## Dalc              -0.20751    0.15298  -1.356 0.175470
## Walc              -0.08284    0.11842  -0.700 0.484443
## health            -0.18997    0.07711  -2.463 0.014034 *
## absences          -0.03685    0.02481  -1.485 0.137997
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.665 on 610 degrees of freedom
## Multiple R-squared:  0.3597, Adjusted R-squared:  0.3198
## F-statistic: 9.016 on 38 and 610 DF,  p-value: < 2.2e-16
```

From the summary of the full model we can see that many of the variables are non-significant. We can apply some techniques for model selection to get significant parameters.

3.2 *Model Selection*

Using Stepwise model selection procedure to get the significant parameters:

```
studentPerformance_stepwise=ols_step_both_p(studentPerformance_fm,p_enter = 0.1, p_remove = 0.3, details
summary(studentPerformance_stepwise$model)
```

```
##
## Call:
## lm(formula = paste(response, "~", paste(preds, collapse = " + ")),
##     data = l)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -11.7284  -1.3971  -0.0786   1.5821   8.0247
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  11.32132    0.60084  18.842  < 2e-16 ***
## failures     -1.39179    0.19048  -7.307 8.20e-13 ***
## schoolMS     -1.54984    0.23766  -6.521 1.42e-10 ***
## higheryes     1.77979    0.37142   4.792 2.06e-06 ***
## studytime     0.44664    0.13458   3.319 0.000956 ***
## schoolsupyes -1.42034    0.35101  -4.046 5.84e-05 ***
## Dalc         -0.33270    0.12177  -2.732 0.006464 **
## health       -0.19232    0.07382  -2.605 0.009397 **
## Fedu          0.28905    0.10023   2.884 0.004062 **
## sexM         -0.51686    0.23277  -2.221 0.026734 *
## absences     -0.03986    0.02400  -1.661 0.097220 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.677 on 638 degrees of freedom
## Multiple R-squared:  0.3238, Adjusted R-squared:  0.3132
## F-statistic: 30.56 on 10 and 638 DF,  p-value: < 2.2e-16
```

Getting best subset

```
#ExecSubsets=ols_step_best_subset(studentPerformance_fm, details=TRUE)
#ExecSubsets$metrics
```

Appendix

END

```
studentPerformance_p1 = lm(G3 ~ (sex+age+address+famsize+Pstatus+Medu+Fedu+traveltime+studytime+
                                   failures+schoolsup+famsup+activities+higher+internet+romantic+
                                   famrel+freetime+goout+Dalc+Walc+health+absences),
                           data = studentDataset)

summary(studentPerformance_p1)
```

```
##
```

```
## Call:
## lm(formula = G3 ~ (sex + age + address + famsize + Pstatus +
##     Medu + Fedu + traveltime + studytime + failures + schoolsup +
##     famsup + activities + higher + internet + romantic + famrel +
##     freetime + goout + Dalc + Walc + health + absences), data = studentDataset)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -13.1329  -1.3258   0.0203   1.6039   6.9956
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)    6.64666    1.91340   3.474 0.000549 ***
## sexM          -0.47948    0.25147  -1.907 0.057024 .
## age            0.16025    0.09985   1.605 0.109021
## addressU       0.68264    0.25423   2.685 0.007443 **
## famsizeLE3     0.25918    0.24642   1.052 0.293321
## PstatusT       0.13849    0.34673   0.399 0.689714
## Medu           0.18486    0.13159   1.405 0.160562
## Fedu           0.21878    0.13016   1.681 0.093301 .
## traveltime     0.01709    0.15903   0.107 0.914449
## studytime      0.47928    0.13947   3.437 0.000628 ***
## failures      -1.46053    0.20397  -7.160 2.27e-12 ***
## schoolsupyes  -1.06895    0.36391  -2.937 0.003432 **
## famsupyes      0.03129    0.22952   0.136 0.891593
## activitiesyes  0.31105    0.22303   1.395 0.163620
## higheryes      1.81839    0.38663   4.703 3.16e-06 ***
## internetyes    0.50635    0.27251   1.858 0.063620 .
## romanticyes   -0.48830    0.23139  -2.110 0.035229 *
## famrel         0.12601    0.11720   1.075 0.282713
## freetime      -0.11299    0.11308  -0.999 0.318067
## goout         -0.09618    0.10797  -0.891 0.373380
## Dalc          -0.28026    0.15203  -1.843 0.065737 .
## Walc          -0.08747    0.11838  -0.739 0.460237
## health        -0.15695    0.07644  -2.053 0.040478 *
## absences      -0.01217    0.02447  -0.497 0.619151
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.726 on 625 degrees of freedom
## Multiple R-squared:  0.3132, Adjusted R-squared:  0.2879
## F-statistic: 12.39 on 23 and 625 DF,  p-value: < 2.2e-16
```

```
subsets = ols_step_forward_p(studentPerformance_fm,p_val = 0.05, details = FALSE)
forwardMdl = subsets$model
summary(forwardMdl)
```

```
##
## Call:
## lm(formula = paste(response, "~", paste(preds, collapse = " + ")),
##     data = l)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -11.6636  -1.4016  -0.0964   1.5758   8.0791
```

```
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  11.08092    0.58395  18.976  < 2e-16 ***
## failures     -1.42279    0.18982  -7.495 2.21e-13 ***
## schoolMS     -1.46345    0.23222  -6.302 5.48e-10 ***
## higheryes     1.83412    0.37049   4.951 9.48e-07 ***
## studytime     0.47100    0.13396   3.516 0.000469 ***
## schoolsupyes -1.37587    0.35047  -3.926 9.58e-05 ***
## Dalc         -0.36685    0.12018  -3.052 0.002364 **
## health       -0.18599    0.07382  -2.519 0.012002 *
## Fedu          0.28340    0.10031   2.825 0.004872 **
## sexM         -0.48505    0.23230  -2.088 0.037189 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.681 on 639 degrees of freedom
## Multiple R-squared:  0.3209, Adjusted R-squared:  0.3113
## F-statistic: 33.55 on 9 and 639 DF,  p-value: < 2.2e-16
```

```
bestSubsetMdl = ols_step_best_subset(forwardMdl, details = FALSE)
(bestSubsetMdl$metrics)
```

```
##   mindex n                                                    predictors
## 1      1 1                                                      failures
## 2      2 2                                              failures school
## 3      3 3                                      failures school higher
## 4      4 4                              failures school higher studytime
## 5      5 5                     failures school higher studytime schoolsup
## 6      6 6                failures school higher studytime schoolsup Dalc
## 7      7 7           failures school higher studytime schoolsup Dalc health
## 8      8 8      failures school higher studytime schoolsup Dalc health Fedu
## 9      9 9 failures school higher studytime schoolsup Dalc health Fedu sex
##     rsquare      adjr   predrsq        cp      aic     sbic      sbc     msep
## 1 0.1546971 0.1533906 0.1485985 150.40381 3259.856 1417.270 3273.282 5734.681
## 2 0.2128290 0.2103919 0.2041741  97.70354 3215.615 1373.059 3233.517 5348.584
## 3 0.2515254 0.2480441 0.2414928  63.29143 3184.900 1342.477 3207.277 5093.551
## 4 0.2709123 0.2663838 0.2589857  47.04889 3169.868 1327.540 3196.721 4969.334
## 5 0.2858874 0.2803344 0.2720338  34.95784 3158.399 1316.205 3189.727 4874.849
## 6 0.3010296 0.2944971 0.2843035  22.70951 3146.490 1304.522 3182.293 4778.925
## 7 0.3086828 0.3011333 0.2898122  17.50804 3141.344 1299.525 3181.623 4733.984
## 8 0.3162794 0.3077329 0.2950988  12.35992 3136.173 1294.549 3180.928 4689.292
## 9 0.3209128 0.3113482 0.2974530  10.00000 3133.760 1292.291 3182.990 4664.814
##        fpe       apc        hsp
## 1 8.863411 0.8505289 0.01367830
## 2 8.279345 0.7944822 0.01277713
## 3 7.896640 0.7577580 0.01218675
## 4 7.715844 0.7404089 0.01190802
## 5 7.580692 0.7274398 0.01169977
## 6 7.442853 0.7142128 0.01148741
## 7 7.384082 0.7085732 0.01139714
## 8 7.325485 0.7029503 0.01130718
## 9 7.298303 0.7003419 0.01126576
```

```r
forwardMdl_int = lm(G3 ~ (failures+school+higher+studytime+schoolsup+Dalc+health+Fedu+sex)^2, data = stu
summary(forwardMdl_int)
```

```
##
## Call:
## lm(formula = G3 ~ (failures + school + higher + studytime + schoolsup +
##     Dalc + health + Fedu + sex)^2, data = studentDataset)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -11.6996  -1.3645  -0.0729   1.5228   6.6223
##
## Coefficients:
##                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)                6.994081   2.303352   3.036  0.00250 **
## failures                  -1.429746   0.949960  -1.505  0.13283
## schoolMS                  -0.459683   1.234799  -0.372  0.70982
## higheryes                  3.757566   1.817193   2.068  0.03909 *
## studytime                  1.040572   0.691466   1.505  0.13288
## schoolsupyes              -0.199247   3.359452  -0.059  0.95273
## Dalc                       1.010457   0.646689   1.563  0.11869
## health                     0.055507   0.397180   0.140  0.88890
## Fedu                       0.865558   0.645584   1.341  0.18051
## sexM                      -0.500809   1.320527  -0.379  0.70464
## failures:schoolMS         -0.625788   0.427210  -1.465  0.14349
## failures:higheryes        -0.203330   0.486690  -0.418  0.67626
## failures:studytime         0.027837   0.346027   0.080  0.93591
## failures:schoolsupyes      1.190662   0.797121   1.494  0.13578
## failures:Dalc             -0.116986   0.196514  -0.595  0.55186
## failures:health            0.186855   0.154445   1.210  0.22681
## failures:Fedu             -0.312372   0.224290  -1.393  0.16422
## failures:sexM              0.556967   0.505512   1.102  0.27099
## schoolMS:higheryes        -1.584560   0.796134  -1.990  0.04701 *
## schoolMS:studytime         0.207736   0.312560   0.665  0.50654
## schoolMS:schoolsupyes      1.276426   1.027482   1.242  0.21461
## schoolMS:Dalc             -0.352666   0.274444  -1.285  0.19928
## schoolMS:health           -0.031410   0.162989  -0.193  0.84725
## schoolMS:Fedu              0.360982   0.224136   1.611  0.10780
## schoolMS:sexM              0.313861   0.530588   0.592  0.55438
## higheryes:studytime        0.019090   0.533220   0.036  0.97145
## higheryes:schoolsupyes    -1.361578   2.569781  -0.530  0.59642
## higheryes:Dalc            -0.137191   0.389437  -0.352  0.72475
## higheryes:health          -0.269271   0.291272  -0.924  0.35561
## higheryes:Fedu             0.225749   0.469491   0.481  0.63081
## higheryes:sexM            -0.459027   0.915415  -0.501  0.61624
## studytime:schoolsupyes    -0.714214   0.470425  -1.518  0.12948
## studytime:Dalc            -0.131844   0.156875  -0.840  0.40099
## studytime:health           0.001314   0.097092   0.014  0.98920
## studytime:Fedu            -0.134601   0.128938  -1.044  0.29694
## studytime:sexM            -0.154244   0.291727  -0.529  0.59719
## schoolsupyes:Dalc          0.275537   0.434904   0.634  0.52661
## schoolsupyes:health        0.316403   0.262646   1.205  0.22880
## schoolsupyes:Fedu         -0.175121   0.357993  -0.489  0.62490
## schoolsupyes:sexM         -0.273266   0.830377  -0.329  0.74220
```

```
## Dalc:health             0.026317   0.089638   0.294  0.76917
## Dalc:Fedu              -0.390208   0.118785  -3.285  0.00108 **
## Dalc:sexM              -0.167309   0.271844  -0.615  0.53848
## health:Fedu            -0.030415   0.069874  -0.435  0.66351
## health:sexM            -0.013722   0.161758  -0.085  0.93242
## Fedu:sexM               0.358688   0.213186   1.683  0.09299 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.636 on 603 degrees of freedom
## Multiple R-squared:  0.3803, Adjusted R-squared:  0.3341
## F-statistic: 8.224 on 45 and 603 DF,  p-value: < 2.2e-16
```

```r
forwardMdl_int1 = lm(G3 ~ (failures+school+higher+studytime+schoolsup+Dalc+health+Fedu+sex+Dalc:Fedu), d
summary(forwardMdl_int1)
```

```
##
## Call:
## lm(formula = G3 ~ (failures + school + higher + studytime + schoolsup +
##     Dalc + health + Fedu + sex + Dalc:Fedu), data = studentDataset)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -11.6622  -1.3997  -0.0194   1.6199   8.0217
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   9.70984    0.70543  13.764  < 2e-16 ***
## failures     -1.43731    0.18832  -7.632 8.43e-14 ***
## schoolMS     -1.45282    0.23034  -6.307 5.31e-10 ***
## higheryes     1.91295    0.36819   5.196 2.75e-07 ***
## studytime     0.49847    0.13311   3.745 0.000197 ***
## schoolsupyes -1.42977    0.34796  -4.109 4.49e-05 ***
## Dalc          0.44342    0.26619   1.666 0.096236 .
## health       -0.17308    0.07332  -2.361 0.018540 *
## Fedu          0.80743    0.18328   4.405 1.24e-05 ***
## sexM         -0.50997    0.23051  -2.212 0.027295 *
## Dalc:Fedu    -0.34879    0.10245  -3.404 0.000705 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.659 on 638 degrees of freedom
## Multiple R-squared:  0.333,  Adjusted R-squared:  0.3226
## F-statistic: 31.86 on 10 and 638 DF,  p-value: < 2.2e-16
```