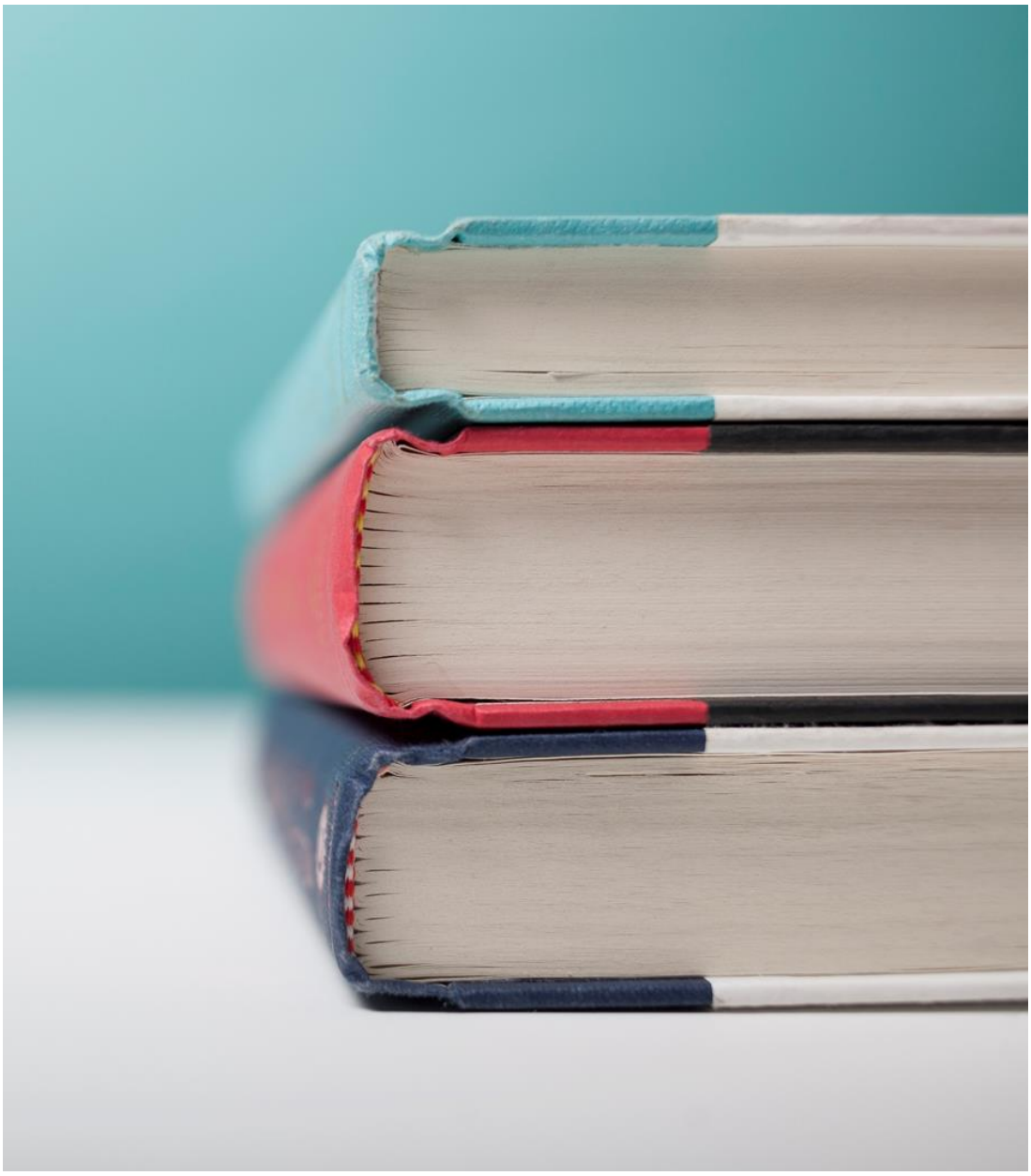


Predicting Students Grades

By Harutyun Hakobyan

Overview

I am a junior at Rutgers, majoring in Stat/Math. I have always wondered what influences students' grades, and if there is any way of predicting students' grades given some information about students' lives? For example, does a guardian's employment status affect students' grades? Do students do good when one of their parents stays home? Today we have a fascinating dataset with which we can explore many questions. These two are just an example of what kind of questions we will deal with today. Today's dataset has 30 predictors and three response variables with which we can do as many as 90 pairwise t-tests. We can use statistical tools to see which predictors affect students' grades. We can use different regression methods to find the best model to predict students' grades. This is just a tiny portion of what we can do with this dataset.

Data Set

This section will briefly describe all the predictors to understand the data set better. This is a dataset from the UCI datasets repository. It contains the final scores of students at the end of a math program with several features that might or might not impact the future outcome of these students. As I said, we have 30 predictors listed below.

1 school - student's school (binary: 'GP' - Gabriel Pereira or 'MS' - Mousinho da Silveira)

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 'R' - rural)

5 famsize - family size (binary: 'LE3' - less or equal to 3 or 'GT3' - greater than 3)

6 Pstatus - parent's cohabitation status (binary: 'T' - living together or 'A' - apart)

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)

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)

9 Mjob - mother's job (nominal: 'teacher', 'health' care related, civil 'services' (e.g., administrative or police), 'at home' *or 'other')*

***10 Fjob - father's job (nominal: 'teacher', 'health' care related, civil 'services' (e.g., administrative or police), 'at home'* or 'other')**

11 reason - reason to choose this school (nominal: close to 'home', school 'reputation', 'course' preference or 'other')

12 guardian - student's guardian (nominal: 'mother', 'father' or 'other')

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)

14 studytime - weekly study time (numeric: 1 - <2 hours, 2 - 2 to 5 hours, 3 - 5 to 10 hours, or 4 - >10 hours)

15 failures - number of past class failures (numeric: n if 1<=n<3, else 4)

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 Portuguese) (binary: yes or no)

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 bad to 5 - excellent)

25 freetime - free time after school (numeric: from 1 - very low to 5 - very high)

26 goout - going out with friends (numeric: from 1 - very low to 5 - very high)

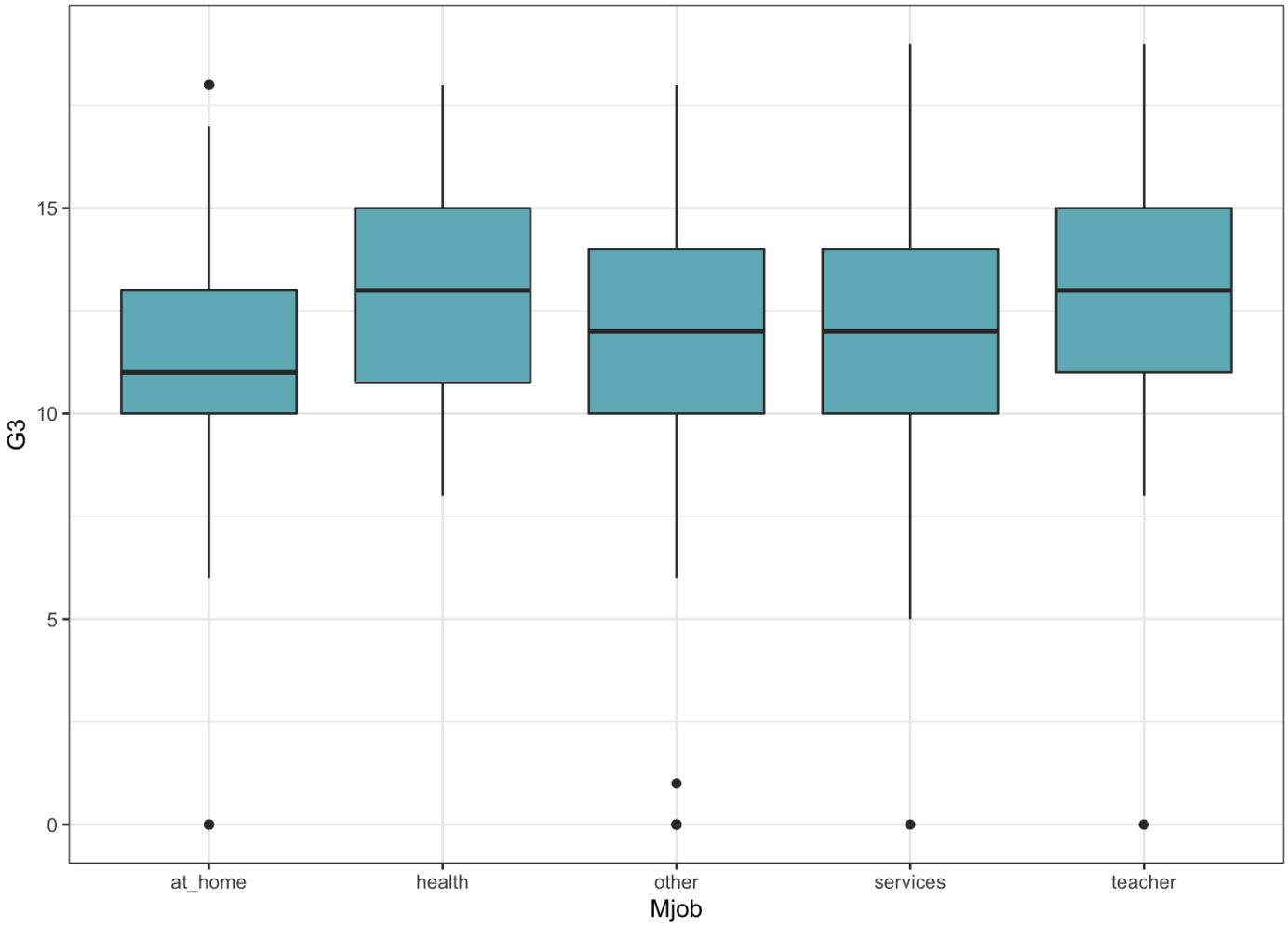
27 Dalc - workday alcohol consumption (numeric: from 1 - very low to 5 - very high)

28 Walc - weekend alcohol consumption (numeric: from 1 - very low to 5 - very high)

29 health - current health status (numeric: from 1 - very bad to 5 - very good)

30 absences - number of school absences (numeric: from 0 to 93)

Student grades by parent's jobs



As we can see in the graph above, the distribution of final student grades in accordance with mother's job does not appear to be a visual difference between them. Next, we will apply statistical analysis to back up our assumptions.

Analysis of Variance Table

Response: G3

	Df	Sum Sq	Mean Sq	F Value	Pr(>F)
Mjob	**4**	**296.1**	**74.014**	**7.3702**	**8.305e-06 *****
Residuals	**644**	**6467.2**	**10.042**		

Signif. codes: 0 '*' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1**

Here we can see the output from ANOVA, it appears that there is a significant difference between the means. Next, we would run the Scheffe test to see which notably differs.

```
Posthoc multiple comparisons of means: Scheffe Test
   95% family-wise confidence level

$Mjob
                        diff      lwr.ci    upr.ci    pval
health-at_home    2.01805556  0.37293555 3.6631756 0.00661 **
other-at_home     0.62609819 -0.41377188 1.6659683 0.48466
services-at_home  1.10261438 -0.08673061 2.2919594 0.08579.
teacher-at_home   2.09444444  0.66584137 3.5230475 0.00045 ***
other-health     -1.39195736 -2.93078431 0.1468696 0.10024
services-health -0.91544118 -2.55897399 0.7280916 0.56480
teacher-health    0.07638889 -1.74777266 1.9005504 0.99997
services-other    0.47651619 -0.56084106 1.5138734 0.73325
teacher-other     1.46834625 0.16355688 2.7731356 0.01741 *
teacher-services 0.99183007 -0.43494499 2.4186051 0.33059

---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

After running Scheffe test, we can see more clearly which groups have significantly different means, students whose have mom who works as a teacher on average score higher, than all other students. Here we go. If you have a mother who works as a teacher, you are more likely to have a high grade. We will take a quick look at the student who has a father as a teacher.

```
Posthoc multiple comparisons of means: Scheffe Test
   95% family-wise confidence level

$Fjob
                        diff      lwr.ci    upr.ci   pval
health-at_home    1.1366460 -1.43420932 3.7075012 0.7605
other-at_home     0.4624367 -1.15197196 2.0768455 0.9406
services-at_home  0.2012628 -1.49619117 1.8987168 0.9978
```

teacher-at_home 2.1547619 -0.09626622 4.4057900 0.0691.

other-health -0.6742092 -2.80452933 1.4561109 0.9163

services-health -0.9353831 -3.12930620 1.2585399 0.7843

teacher-health 1.0181159 -1.62746073 3.6636926 0.8418

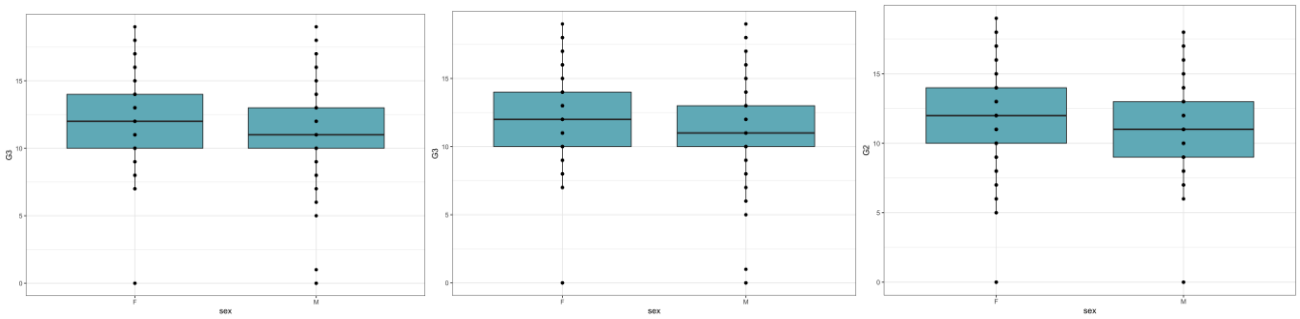
services-other -0.2611739 -1.16135001 0.6390022 0.9379

teacher-other 1.6923252 -0.03859787 3.4232482 0.0593.

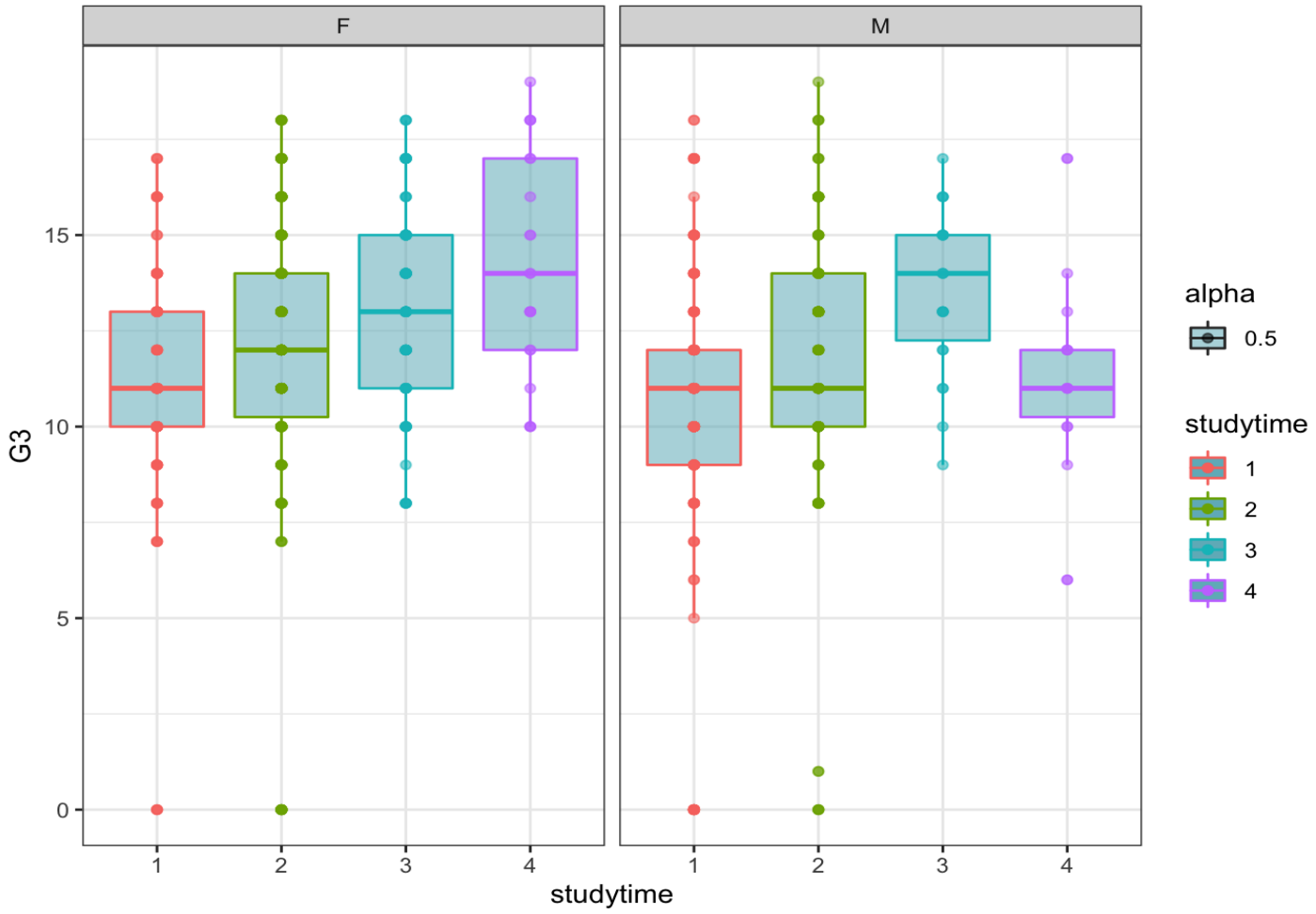
teacher-services 1.9534991 0.14487277 3.7621254 0.0260 *

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Surprisingly, fathers' occupation has no effect on students' grades.



Another interesting fact is that females in general score higher than males. The three graphs above show first quarter, second quarter, and final grades distributed by sex. The p-value for t-test are, p-value = 0.001125, p-value = 9.593e-06, p-value = 1.946e-07 respectfully.



Here we can see another interesting graph. By looking at this graph, we can see that females who studied 4 hours have higher mean grades than males who studied 4 hours. After running the t-test, we get a p-value = 0.01568, which means we can reject the null hypothesis; therefore, we can say for sure that they have significantly different mean grades. Apparently, the best study time for males is 3 hours and for female, it is 4 hours.

Conclusion

This paper observed what factors influence students' grades, and we found some interesting results. There is still a lot to learn from this data, but we will stop here. I hope you were intrigued by these findings.

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.