ACM Final Report

Alyster Alcudia, Cynthia Gan, Mihai Sirbu

Monday, May 18, 2015

Read in Data

```
data = as.tbl(read.csv("student-por.csv", sep=";", stringsAsFactors=TRUE))
```

Rename the Variables

```
rename_cols <- c(4:12,16:21,23:24,26:28,31:33)

colnames(data)[rename_cols] = c("urban_rural",
                                "fam_size",
                                "parents_cohabit",
                                "mom_edu",
                                "dad_edu",
                                "mom_job",
                                "dad_job",
                                "school_reason",
                                "student_guardian",
                                "extra_school_support",
                                "fam_edu_support",
                                "paid_extra_classes",
                                "extracurricular_activities",
                                "attended_nursery",
                                "wants_higher_education",
                                "has_romantic_partner",
                                "family_relationship",
                                "going_out_amount",
                                "weekday_alcohol_cons",
                                "weekend_alcohol_cons",
                                "grade_one",
                                "grade_two",
                                "final_grade")
```

Meaningful Variable Values

```
data <-  data %>%
  mutate(school = ifelse(school=='GP', 'Gabriel Pereira', 'Mousinho da Silveira')) %>%
  mutate(urban_rural = ifelse(urban_rural=='R', 'Rural', 'Urban')) %>%
  mutate(fam_size = ifelse(fam_size=='GT3', '>3', '<=3')) %>%
  mutate(parents_cohabit = ifelse(parents_cohabit=='A', 'Apart', 'Together')) %>%
  mutate_each_(funs(ifelse(.=='yes', TRUE, FALSE)),
               c("extra_school_support", "fam_edu_support", "paid_extra_classes",
                 "extracurricular_activities", "attended_nursery",
                 "wants_higher_education", "internet", "has_romantic_partner"))
```

Summary

1) Describe the task you performed (in one sentence). The task of this project was to understand what student demographics are associated with positive academic achievement in a language class in Portugal.
2) Describe the data you used (in one sentence). The dataset we used was from the UCI Machine Learning Repository and consisted of socioeconomic traits and language class grades of students in two Portuguese secondary schools.
3) Describe the analysis method you used (in one sentence). For this project, we used linear regression to examine the relationship between student demographics and final grade (our operationalization of academic achievement) because we believed it would give us interpretative results.
4) Summarize your results (in one or two sentences).

Introduction

(1/2 page): Describe the task you performed in more detail. Include motivation for performing this task.

The goal of our project was to better understand what student demographics are associated with positive student academic performance. Understanding why some kids do well, while others fail, is an extremely important topic; schools and states across the country are all grappling with different proposals to help all students succeed regardless of their background. In our view, understanding the relationship between student characteristics and positive outcomes would

One obvious source for finding education data is the education.gov website. Unfortunately, the datasets we found (or our ability to extract them) on the website were simply too burdensome and unwieldy given the time frame of this project. Nevertheless, given that we were still interested in an education project we decided to use (clean) education data from the U.C. Irvine Machine Learning Repository. This dataset consisted of socioeconomic traits and language class grades of students in two Portuguese secondary schools.

Given this dataset was tailored to a specific class, the goal and task of this project morphed from understanding "what student demographics are associated with positive student academic performance?" to "What can we infer about student demographics and academic achievement in a language class in Portugal?". Moreover, the goal of this project was inference rather than any sort of specific prediction. Although the generalizability of our results is reduced in this case, we believe that trying to understand what variables are important in this dataset may provide insights to future education projects.

Researchers Paulo Cortez and Alice Silva from the University of Minho had previously used this dataset for modeling[1]. Their main goals were to build binary (pass/fail) classification, 5-level classification, and regression models for prediction purposes. Their methods included Decision Trees, Random Forest, Neural Networks and Support Vector Machines.

Although they were primarily focused on prediction, their Random Forest analysis yielded relative importance of variables used in their models. Our focus, however, is on the interpretability of model estimates.

Materials and Methods

(1.5-2 pages): Describe the dataset you used in more detail (you can get this from Step 3) Describe analysis methods you used, and how you are evaluating performance (from Step 3)

As mentioned above the dataset was the "Student Performance Data Set" from the UCI Machine Learning Repository (https://archive.ics.uci.edu/ml/datasets/Student+Performance). The dataset consisted of socioe-

[1]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.

conomic traits and language class grades of students in two Portuguese secondary schools and was compiled from school reports and questionnaires. All in all we had 649 observations and 33 variables.

The variables that we had available to us were as follows:

Variables in Table 1:

- school - student's school ("Gabriel Pereira" or "Mousinho da Silveira")
- sex - student's sex ("F" - female or "M" - male)
- age - student's age (from 15 to 22)
- address - student's home address type ("Urban" or "Rural")
- fam_size - family size ("<=3" - less than or equal to three or ">3" - greater than 3)
- parent_status - parent's cohabitation status ("together" or "apart")
- mom_edu - mother's education (0 - none, 1 - primary education (4th grade), 2 – 5th to 9th grade, 3 – secondary education or 4 – higher education)
- dad_edu - father's education (0 - none, 1 - primary education (4th grade), 2 – 5th to 9th grade, 3 – secondary education or 4 – higher education)
- mom_job - mother's job ("teacher", "health" care related, civil "services" (e.g. administrative or police), "at_home" or "other")
- dad_job - father's job ("teacher", "health" care related, civil "services" (e.g. administrative or police), "at_home" or "other")
- reason - reason to choose this school (close to "home", school "reputation", "course" preference or "other")
- guardian - student's guardian ("mother", "father" or "other")
- traveltime - home to school travel time (1 - <15 min., 2 - 15 to 30 min., 3 - 30 min. to 1 hour, or 4 - >1 hour)
- studytime - weekly study time (1 - <2 hours, 2 - 2 to 5 hours, 3 - 5 to 10 hours, or 4 - >10 hours)
- failures - number of past class failures (1, 2, 3, or 4 (with 4 representing 4 or more))
- extra_school_support - extra educational support (TRUE or FALSE)
- fam_edu_sup - family educational support (TRUE or FALSE)
- paid_extra_classes - extra paid classes within the course subject (Portuguese) (TRUE or FALSE)
- extracurricular_activities - extra-curricular activities (TRUE or FALSE)
- nursery - attended nursery school (TRUE or FALSE)
- higher - wants to take higher education (TRUE or FALSE)
- internet - Internet access at home (TRUE or FALSE)
- romantic - with a romantic relationship (TRUE or FALSE)
- family_relationship - quality of family relationships (from 1 - very bad to 5 - excellent)
- freetime - free time after school (from 1 - very low to 5 - very high)
- going_out_amount - amount going out with friends (from 1 - very low to 5 - very high)
- weekday_alcohol_cons - workday alcohol consumption (from 1 - very low to 5 - very high)
- weekend_alcohol_cons - weekend alcohol consumption (from 1 - very low to 5 - very high)
- health - current health status (from 1 - very bad to 5 - very good)
- absences - number of school absences (from 0 to 93)
- **grade_one** - first period Portuguese grade (from 0 to 20)
- **grade_two** - second period Portuguese grade (from 0 to 20)
- **final_grade** - final Portuguese grade (from 0 to 20)

As you can see, there are actually three target variables we could use as our response measure (grade one, grade two, and final grade). In our exploratory data analysis section of the project however, we found that these three response variables had very correlations between each other (all above 0.82). Consequently, we

decided to only use final grade as our response AND made sure not to use grade one or grade two in our subsequent model.

Cynthia could you describe StepAIC here? (Of course we need to add more details about the analysis method too haha)

Results

Describe the results you found. Make sure you quantify your findings (if inference, provide interpretation of resulting model. E.g., we found that house prices increase by $10,000 (P<0.05) on average when they are within 5 miles of a university campus; if prediction, we found we could predict house prices with mean squared error of $10,000.

Regression with Hybrid Step-Wise Abroach

```
fit <- lm(final_grade~.-grade_one-grade_two,data=data)
step2 <- stepAIC(fit, direction="both")
```

```
summary(step2)
```

```
##
## Call:
## lm(formula = final_grade ~ school + sex + age + mom_edu + student_guardian +
##     studytime + failures + extra_school_support + wants_higher_education +
##     has_romantic_partner + weekday_alcohol_cons + health + absences,
##     data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -12.1548  -1.3687   0.0072   1.5292   7.2845
##
## Coefficients:
##                              Estimate Std. Error t value Pr(>|t|)
## (Intercept)                   8.90516    1.75710   5.068 5.28e-07 ***
## schoolMousinho da Silveira -1.51318    0.24021  -6.299 5.59e-10 ***
## sexM                          -0.57091    0.23574  -2.422 0.015726 *
## age                            0.16711    0.09910   1.686 0.092231 .
## mom_edu                        0.30127    0.09906   3.041 0.002454 **
## student_guardianmother        -0.45308    0.25282  -1.792 0.073592 .
## student_guardianother          0.03407    0.51153   0.067 0.946911
## studytime                      0.40872    0.13508   3.026 0.002580 **
## failures                      -1.48437    0.19764  -7.511 2.01e-13 ***
## extra_school_supportTRUE   -1.33575    0.35655  -3.746 0.000196 ***
## wants_higher_educationTRUE  1.86377    0.37726   4.940 9.99e-07 ***
## has_romantic_partnerTRUE   -0.42199    0.22456  -1.879 0.060679 .
## weekday_alcohol_cons       -0.35842    0.12260  -2.924 0.003584 **
## health                        -0.17961    0.07351  -2.443 0.014826 *
## absences                      -0.03687    0.02412  -1.529 0.126848
```

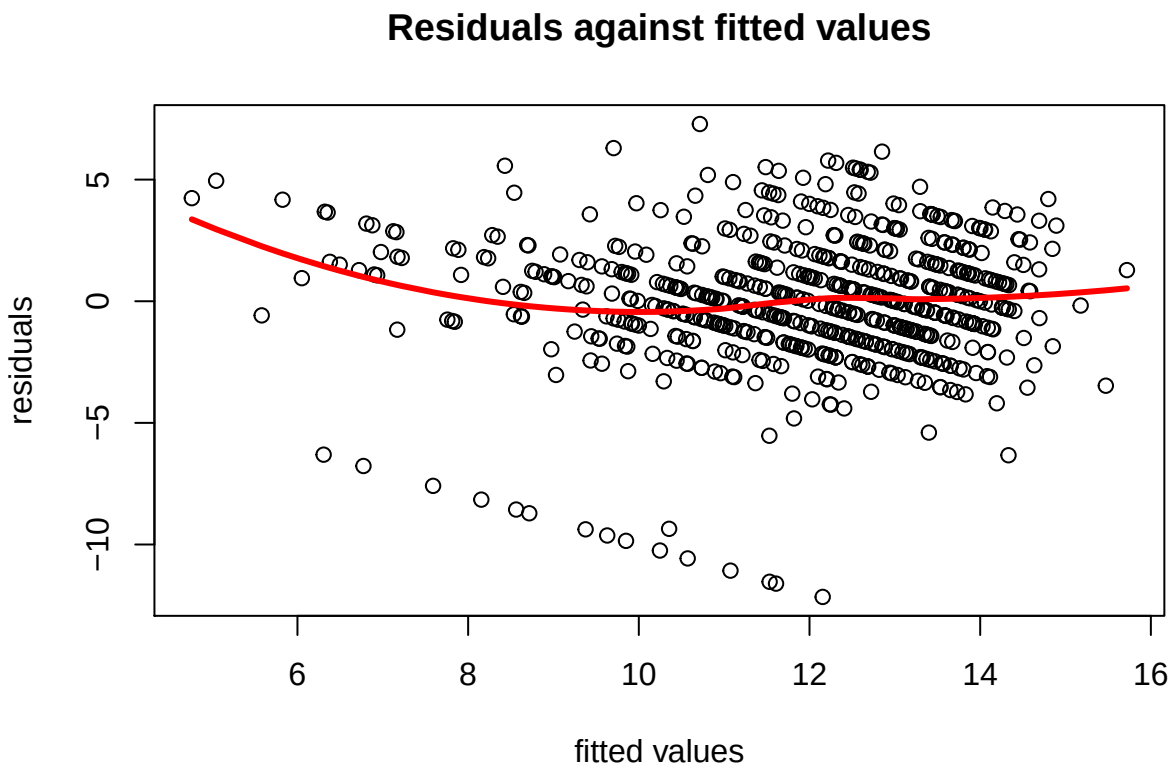
```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.666 on 634 degrees of freedom
## Multiple R-squared:  0.3339, Adjusted R-squared:  0.3192
## F-statistic:  22.7 on 14 and 634 DF,  p-value: < 2.2e-16
```

Someone should take a few moments to interpret these results...

The banding pattern of the residuals is WEIRD. We need to investigate why this is. (See other document to see how I've been temporarily trying to do that)

```
fitted_res <- as.tbl(data.frame(fitted = fitted(step2), res = residuals(step2)))

lw1 <- loess(res ~ fitted, data=fitted_res)
plot(res~fitted, data=fitted_res, main = "Residuals against fitted values",
     xlab="fitted values", ylab="residuals")
j <- order(fitted_res$fitted)
lines(fitted_res$fitted[j], lw1$fitted[j], col="red",lwd=3)
```



This thing below takes FOREVER to run! (JUST FYI)

```
# fit <- lm(final_grade~school*sex*mom_edu*studytime*failures*extra_school_support*
#             wants_higher_education*weekday_alcohol_cons*health, data=data)
# step <- stepAIC(fit, direction="both")
#
#
# summary(step)
```

Conclusion

(1/2 page): Concluding remarks summarizing findings and discussing possible improvements to your analysis.