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 interpretable results.
4) Summarize your results (in one or two sentences). None of our three models – a step-wise regression model, a complete model with interaction terms, and a simplified model with interaction terms – performed well with respect to adjusted R-squared values. However, the simplified model yielded more interpretable results.

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)

[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.

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 socioeconomic 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 high 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.

Because interpretability of our results was key, we decided to use linear regression. However, because we had 33 possible variables to examine (plus an even greater number of interactions), We decided to start our analysis with the a stepwise regression, optimizing the Akaike information criterion (AIC) with a function stepAIC from the MASS library. This function starts off with a model, in this case, a linear model with all the dependent variables with no interations, and iteratively adds and removes predictors as it tries to improve the AIC.

Once we identified key variables in the linear model with all dependent variables but no interactions, we added interaction terms based on significant variables. The goal throughout this step was two-fold: improve adjusted R^2 while also making sure our results were still interpretable. Our models' inference performance was being evaluated based adjusted R^ as well as subsequent residual graphs (to test things like the presence of heteroscedascity).

Results

Results (about 3 pages): Describe the results you found. Make sure you quantify your findings (if inference, provide interpretation of resulting model.

Evolution of a Linear Model

Regression with Hybrid Step-Wise Approach

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

```
summary(step)
```

```
##
## 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
```

The result of the step-wise regression was a model with 14 predictors, and an adjusted R-squared of 0.3192322, the majority of these predicts were statistically significant. Attending Mousinho da Silveira over Gabriel Pereira, being male, failing a class previously, having extra educational support, and coming alcohol on a weekday all had statistically significant negative relationships with final grades. Meanwhile, desiring a higher education and having a mother with a good education had statistically significant positive relationships with final grade.

Comparison with Complete Model

While an adjusted R-squared of 0.3192322 seems bad, fitting a model that accounts for every possible interaction between the predictors doesn't result in a significantly better adjusted R-squared.

```
fitall <- lm(final_grade~school*sex*mom_edu*studytime*failures*
               extra_school_support*wants_higher_education*
               weekday_alcohol_cons*health, data=data)
summary(fitall)[c("call", "r.squared", "adj.r.squared", "df", "fstatistic")]
```

```
## $call
## lm(formula = final_grade ~ school * sex * mom_edu * studytime *
##     failures * extra_school_support * wants_higher_education *
##     weekday_alcohol_cons * health, data = data)
##
## $r.squared
## [1] 0.5797327
##
## $adj.r.squared
## [1] 0.3768119
##
## $df
## [1] 212 437 512
##
## $fstatistic
##      value      numdf      dendf
##   2.856941 211.000000 437.000000
```

A linear model that accounts for every single possible interaction between predictors has an adjusted-square of 0.3768119. The additional multitude of variables only increased the adjusted percent variation explained by the variables by 0.0575797.

While not a huge gain, it is also non-negligible.

Compromise between Number of Terms and R-Squared

```
fit2 = lm(formula = final_grade ~ school * mom_edu *
            failures * extra_school_support *
            wants_higher_education * weekday_alcohol_cons, data = data)
summary(fit2)[c("call", "r.squared", "adj.r.squared", "df", "fstatistic")]
```

```
## $call
## lm(formula = final_grade ~ school * mom_edu * failures * extra_school_support *
##     wants_higher_education * weekday_alcohol_cons, data = data)
##
## $r.squared
## [1] 0.3885079
##
## $adj.r.squared
## [1] 0.3428742
##
## $df
## [1]  46 603  64
##
## $fstatistic
##      value      numdf      dendf
##   8.513611  45.000000 603.000000
```

Accounting only for the 5 most statistically significant variables from the step-wise no-interaction regression and their interactions, we still manage to capture 0.3192322 of the variation. That's a 0.0339377 difference with only a fraction of the variables.

To further simplify the model, we looked at only the most statistically significant interactions, and then manually factored the variables to produce a more legible formula describing the model. This resulted in our final model.

```
fit3 = lm(final_grade ~ failures + wants_higher_education + school +
            extra_school_support + mom_edu + weekday_alcohol_cons +
            (school:extra_school_support)*(mom_edu*weekday_alcohol_cons), data=data)
summary(fit3)
```

```
##
## Call:
## lm(formula = final_grade ~ failures + wants_higher_education +
##     school + extra_school_support + mom_edu + weekday_alcohol_cons +
##     (school:extra_school_support) * (mom_edu * weekday_alcohol_cons),
##     data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -13.0559  -1.4302  -0.0425   1.5622   8.0115
##
## Coefficients: (3 not defined because of singularities)
```

```
##                                                                      Estimate
## (Intercept)                                                           9.4265
## failures                                                             -1.5947
## wants_higher_educationTRUE                                            1.9179
## schoolMousinho da Silveira                                           -1.2948
## extra_school_supportTRUE                                             -1.7651
## mom_edu                                                             -11.8258
## weekday_alcohol_cons                                                -15.3233
## schoolMousinho da Silveira:extra_school_supportTRUE                  23.3572
## mom_edu:weekday_alcohol_cons                                          8.0385
## schoolGabriel Pereira:extra_school_supportFALSE:mom_edu              12.7674
## schoolMousinho da Silveira:extra_school_supportFALSE:mom_edu         12.9608
## schoolGabriel Pereira:extra_school_supportTRUE:mom_edu               12.6816
## schoolMousinho da Silveira:extra_school_supportTRUE:mom_edu              NA
## schoolGabriel Pereira:extra_school_supportFALSE:weekday_alcohol_cons  15.9976
## schoolMousinho da Silveira:extra_school_supportFALSE:weekday_alcohol_cons  15.7116
## schoolGabriel Pereira:extra_school_supportTRUE:weekday_alcohol_cons   16.8428
## schoolMousinho da Silveira:extra_school_supportTRUE:weekday_alcohol_cons      NA
## schoolGabriel Pereira:extra_school_supportFALSE:mom_edu:weekday_alcohol_cons  -8.4682
## schoolMousinho da Silveira:extra_school_supportFALSE:mom_edu:weekday_alcohol_cons  -8.5190
## schoolGabriel Pereira:extra_school_supportTRUE:mom_edu:weekday_alcohol_cons  -8.6301
## schoolMousinho da Silveira:extra_school_supportTRUE:mom_edu:weekday_alcohol_cons      NA
##                                                                     Std. Error
## (Intercept)                                                             0.8160
## failures                                                                0.1887
## wants_higher_educationTRUE                                              0.3664
## schoolMousinho da Silveira                                              1.0406
## extra_school_supportTRUE                                                1.7731
## mom_edu                                                                 2.4286
## weekday_alcohol_cons                                                    3.9592
## schoolMousinho da Silveira:extra_school_supportTRUE                     5.1862
## mom_edu:weekday_alcohol_cons                                            1.7812
## schoolGabriel Pereira:extra_school_supportFALSE:mom_edu                 2.4414
## schoolMousinho da Silveira:extra_school_supportFALSE:mom_edu            2.4477
## schoolGabriel Pereira:extra_school_supportTRUE:mom_edu                  2.4967
## schoolMousinho da Silveira:extra_school_supportTRUE:mom_edu                 NA
## schoolGabriel Pereira:extra_school_supportFALSE:weekday_alcohol_cons    3.9842
## schoolMousinho da Silveira:extra_school_supportFALSE:weekday_alcohol_cons  3.9772
## schoolGabriel Pereira:extra_school_supportTRUE:weekday_alcohol_cons     4.0534
## schoolMousinho da Silveira:extra_school_supportTRUE:weekday_alcohol_cons     NA
## schoolGabriel Pereira:extra_school_supportFALSE:mom_edu:weekday_alcohol_cons  1.7876
## schoolMousinho da Silveira:extra_school_supportFALSE:mom_edu:weekday_alcohol_cons  1.7879
## schoolGabriel Pereira:extra_school_supportTRUE:mom_edu:weekday_alcohol_cons  1.8080
## schoolMousinho da Silveira:extra_school_supportTRUE:mom_edu:weekday_alcohol_cons     NA
##                                                                     t value
## (Intercept)                                                          11.552
## failures                                                             -8.452
## wants_higher_educationTRUE                                            5.234
## schoolMousinho da Silveira                                           -1.244
## extra_school_supportTRUE                                             -0.996
## mom_edu                                                              -4.869
## weekday_alcohol_cons                                                 -3.870
## schoolMousinho da Silveira:extra_school_supportTRUE                   4.504
## mom_edu:weekday_alcohol_cons                                          4.513
```

```
## schoolGabriel Pereira:extra_school_supportFALSE:mom_edu                                5.229
## schoolMousinho da Silveira:extra_school_supportFALSE:mom_edu                           5.295
## schoolGabriel Pereira:extra_school_supportTRUE:mom_edu                                 5.079
## schoolMousinho da Silveira:extra_school_supportTRUE:mom_edu                               NA
## schoolGabriel Pereira:extra_school_supportFALSE:weekday_alcohol_cons                   4.015
## schoolMousinho da Silveira:extra_school_supportFALSE:weekday_alcohol_cons              3.950
## schoolGabriel Pereira:extra_school_supportTRUE:weekday_alcohol_cons                    4.155
## schoolMousinho da Silveira:extra_school_supportTRUE:weekday_alcohol_cons                  NA
## schoolGabriel Pereira:extra_school_supportFALSE:mom_edu:weekday_alcohol_cons       -4.737
## schoolMousinho da Silveira:extra_school_supportFALSE:mom_edu:weekday_alcohol_cons  -4.765
## schoolGabriel Pereira:extra_school_supportTRUE:mom_edu:weekday_alcohol_cons        -4.773
## schoolMousinho da Silveira:extra_school_supportTRUE:mom_edu:weekday_alcohol_cons       NA
##                                                                                  Pr(>|t|)
## (Intercept)                                                                       < 2e-16
## failures                                                                          < 2e-16
## wants_higher_educationTRUE                                                       2.26e-07
## schoolMousinho da Silveira                                                        0.21387
## extra_school_supportTRUE                                                          0.31987
## mom_edu                                                                          1.42e-06
## weekday_alcohol_cons                                                              0.00012
## schoolMousinho da Silveira:extra_school_supportTRUE                              7.95e-06
## mom_edu:weekday_alcohol_cons                                                     7.63e-06
## schoolGabriel Pereira:extra_school_supportFALSE:mom_edu                          2.32e-07
## schoolMousinho da Silveira:extra_school_supportFALSE:mom_edu                     1.64e-07
## schoolGabriel Pereira:extra_school_supportTRUE:mom_edu                           4.99e-07
## schoolMousinho da Silveira:extra_school_supportTRUE:mom_edu                            NA
## schoolGabriel Pereira:extra_school_supportFALSE:weekday_alcohol_cons             6.65e-05
## schoolMousinho da Silveira:extra_school_supportFALSE:weekday_alcohol_cons        8.68e-05
## schoolGabriel Pereira:extra_school_supportTRUE:weekday_alcohol_cons              3.70e-05
## schoolMousinho da Silveira:extra_school_supportTRUE:weekday_alcohol_cons               NA
## schoolGabriel Pereira:extra_school_supportFALSE:mom_edu:weekday_alcohol_cons     2.68e-06
## schoolMousinho da Silveira:extra_school_supportFALSE:mom_edu:weekday_alcohol_cons 2.35e-06
## schoolGabriel Pereira:extra_school_supportTRUE:mom_edu:weekday_alcohol_cons      2.25e-06
## schoolMousinho da Silveira:extra_school_supportTRUE:mom_edu:weekday_alcohol_cons        NA
##
## (Intercept)                                                                      ***
## failures                                                                         ***
## wants_higher_educationTRUE                                                       ***
## schoolMousinho da Silveira
## extra_school_supportTRUE
## mom_edu                                                                          ***
## weekday_alcohol_cons                                                             ***
## schoolMousinho da Silveira:extra_school_supportTRUE                              ***
## mom_edu:weekday_alcohol_cons                                                     ***
## schoolGabriel Pereira:extra_school_supportFALSE:mom_edu                          ***
## schoolMousinho da Silveira:extra_school_supportFALSE:mom_edu                     ***
## schoolGabriel Pereira:extra_school_supportTRUE:mom_edu                           ***
## schoolMousinho da Silveira:extra_school_supportTRUE:mom_edu
## schoolGabriel Pereira:extra_school_supportFALSE:weekday_alcohol_cons             ***
## schoolMousinho da Silveira:extra_school_supportFALSE:weekday_alcohol_cons        ***
## schoolGabriel Pereira:extra_school_supportTRUE:weekday_alcohol_cons              ***
## schoolMousinho da Silveira:extra_school_supportTRUE:weekday_alcohol_cons
## schoolGabriel Pereira:extra_school_supportFALSE:mom_edu:weekday_alcohol_cons     ***
## schoolMousinho da Silveira:extra_school_supportFALSE:mom_edu:weekday_alcohol_cons ***
```

```
## schoolGabriel Pereira:extra_school_supportTRUE:mom_edu:weekday_alcohol_cons       ***
## schoolMousinho da Silveira:extra_school_supportTRUE:mom_edu:weekday_alcohol_cons
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.647 on 631 degrees of freedom
## Multiple R-squared:  0.3464, Adjusted R-squared:  0.3288
## F-statistic: 19.67 on 17 and 631 DF,  p-value: < 2.2e-16
```

Final Model

Our final model has an adjusted R-squared of 0.3287862. While it is lower than the all-encompassing model's adjusted R-squared of 0.3768119, it only depends on 18 terms instead of 212. It captures 0.8725472 of the variation explained by the complete model, and is far more interpretable.

The final model depends on 5 predictors to predict the students final grade. While the model performs too poorly to actually predict anything, which variables and interactions were important can be meaningful. The 5 variables that we used were: * school - student's school ("Gabriel Pereira" or "Mousinho da Silveira") * mom_edu - mother's education (0 - none, 1 - primary education (4th grade), 2 – 5th to 9th grade, 3 – secondary education or 4 – higher education) * 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) * weekday_alcohol_cons - workday alcohol consumption (from 1 - very low to 5 - very high)

For how variables interacted with each other, we found that in statistically significant interaction terms in earlier models, which school and extra school support always influenced the same things, so could be grouped together. Meanwhile, failure rate and the desire for higher education did not significantly interact with any other variable, so could be isolated. Therefore, we only had 3 terms interacting with each other, and 8 interaction terms: the school/extra school support pair, mother's education, and weekday alcohol consumption.

A few notable relationships are:

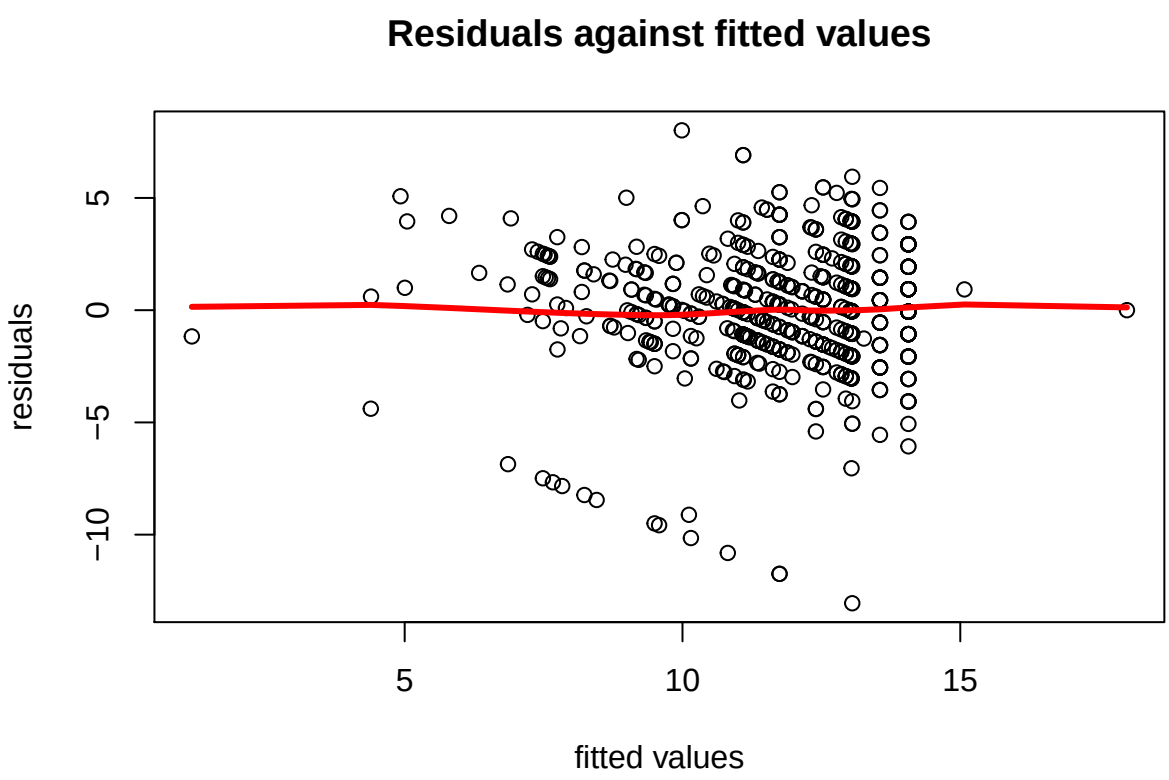
- The most statistically significant relationships:
  - Each past failure of a class results in a 1.59 point drop in predicted final grade
  - Desiring higher education results in an almost 2 point rise in predict final grade
- The most dramatic relationships:
  - Receiving extra school support from Mousinho da Silveira results in an approximately 23 rise in predicted final grade
  - Each increase in tier of amount of alcohol consumed on a weekday results in a fifteen point decrease in predicted final grade

Model Evaluation

Residuals

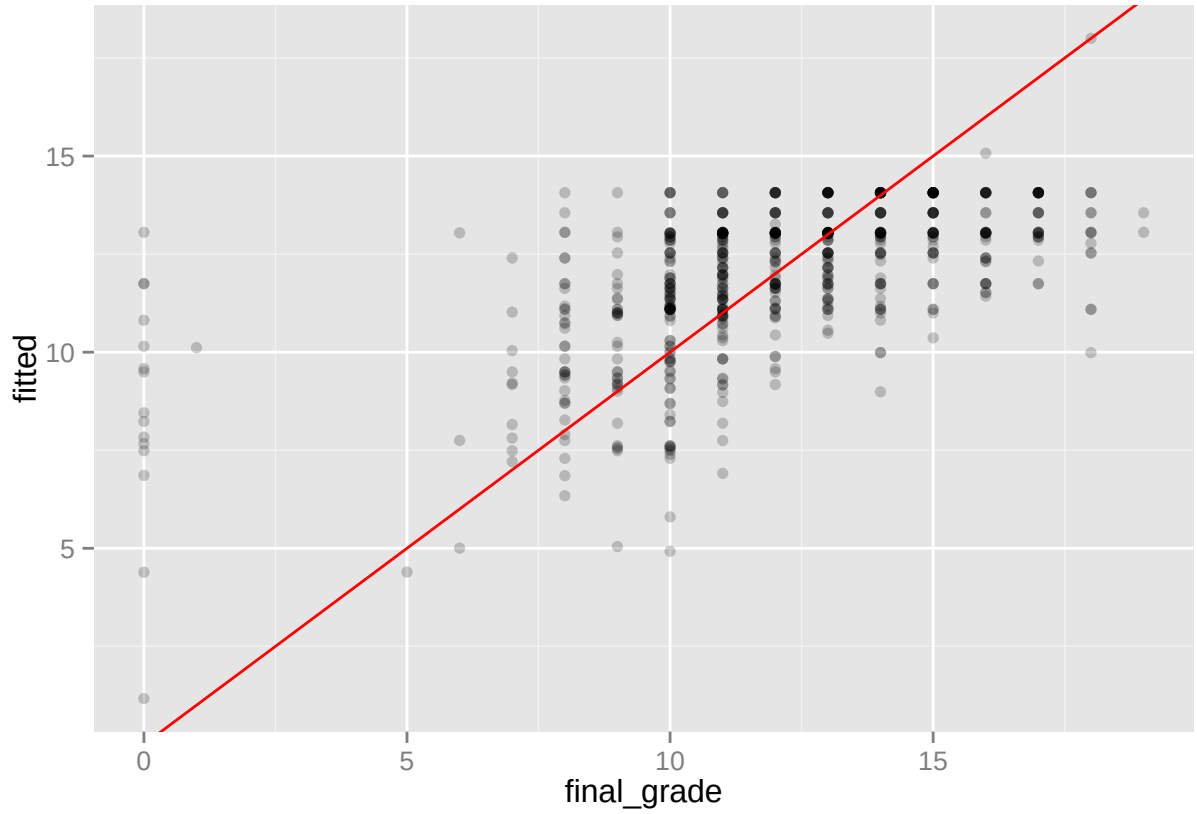
```
fitted_res <- as.tbl(data.frame(fitted = fitted(fit3), res = residuals(fit3)))

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)
```



Though the points are centered around 0, the residuals exhibit a strange pattern. There is clear heteroscedasticity, the spread of the residuals increases dramatically as the predicted value increases. There is also the presence of diagonal streaks in the residuals. Both of these phenomena become more clear if you examine a plot of fitted value vs true value, another view of the same information as a residual plot, and the density plot of final grades.

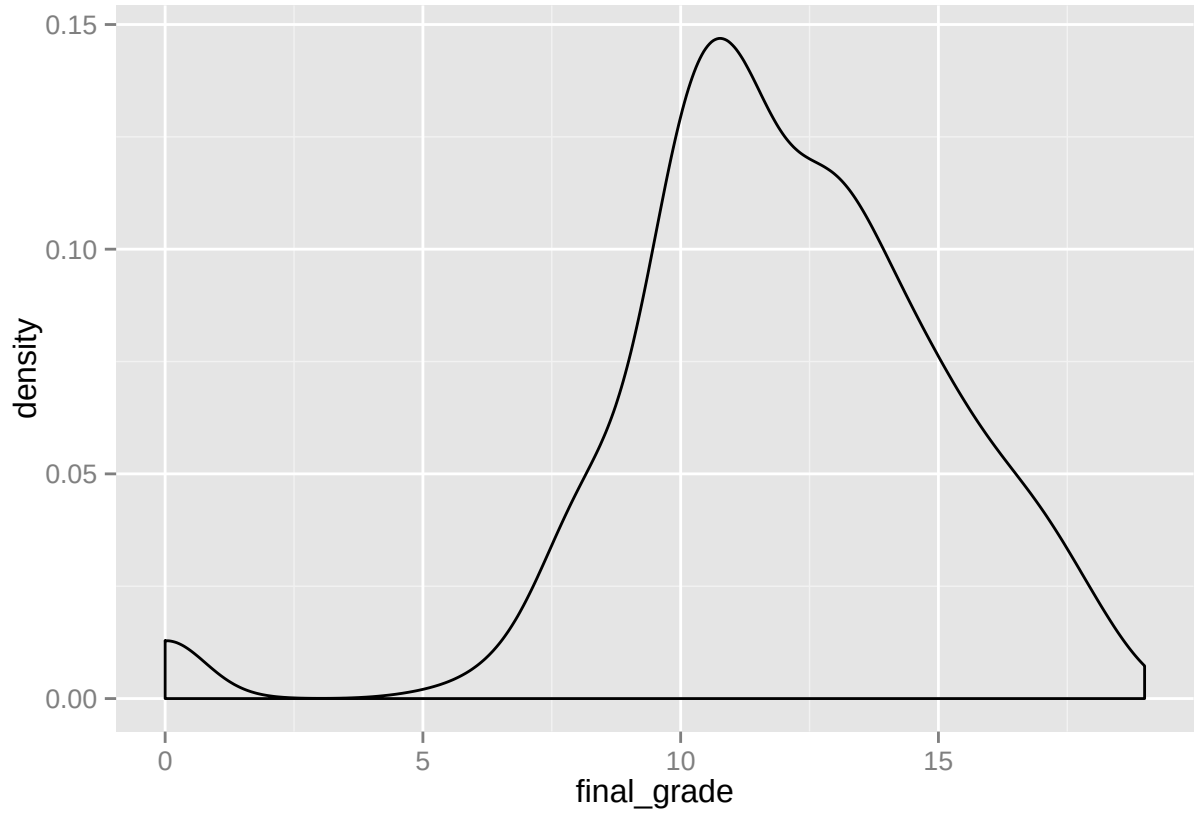
```
fitted3_res <- as.tbl(data.frame(fitted = fitted(fit3), res = residuals(fit3)))
data_fitted3_res <- as.tbl(cbind(data, fitted3_res))
ggplot(data_fitted3_res, aes(x = final_grade, y = fitted)) +
  geom_point(alpha=0.2) +
  geom_abline(intercept=0, slope=1, colour="red")
```



The red line represents the ideal y=x, where the fitted value matches the actual value.

This plot is a rotation of the previous residual plot. With these axes, it becomes clear that the streaks result from actual result, i.e. the final grade, being limited to integer values. It is also more clear that the model mysteriously experiences a ceiling of a score of 14, resulting in the clean cut off on the right side of the residual plot.

```
ggplot(data) + geom_density(aes(x=final_grade))
```



The distribution of scores is bears a resemblance to the distribution of residuals. Both are vaguely triangular, with a peak at about 10 points, and decreasing linearly past that. Our model performs quite poorly, and appears to lack sufficient data. One possible hypothesis for the varying spread is that the plethora of students receiving 10 points had a wide variety of socioeconomic backgrounds. Therefore, there were different point predictions due to the poor model, causing a wide range of residuals. As fewer students get a particular score, they grow more similar to each other in socioeconomic terms, resulting in more accurate predictions.

Conclusion

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

The three models used in our analysis each had low adjusted R-squared values. Even the complete model had an adjusted R-squared value of only 0.3768119. This might suggest an insufficient amount of data or a more complicated relationship between the predictors and responses.

We found that a simpler model can capture much (about 0.8725472) of the variation explained by the complete model. Furthermore, this simpler model has the benefit of more interpretable results.

The heteroscedasticity of residuals suggests that we could try to transform the `final_grade` response using a concave function such as `sqrt()` or `log()`. However, such a transformation would make comparisons between other models more difficult.

Non-linear transformations of the data variables themselves could possibly improve our models. This approach would require much more time simply because of the number of variables involved. A possible benefit of this approach would be a deeper understanding of the relationship between variables and responses. Since our main focus is on inference, non-linear transformations could help further reveal important relationships. A related approach would be to derive new features from our current data; for example, Bill James's Pythagorean

expectation formula for baseball teams, which is derived from the numbers of runs scored and runs allowed. However, this would require more research to find relationships between predictors and possibly a lot of formula tuning.

Gathering additional data might also be useful, but finding good data to cross reference with our current data would be difficult. Perhaps a more detailed and more comprehensive survey might have revealed more relationships.