ACM Preliminary Analysis

Alyster Alcudia, Cynthia Gan, Mihai Sirbu

Wednesday, May 06, 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"))
```

Regression Using Sex as a Variable

```r
lm_sex <- lm(final_grade ~ sex, data = data)

summary(lm_sex)
```

```
##
## Call:
## lm(formula = final_grade ~ sex, data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -12.2533  -1.4060  -0.2533   1.7467   7.5940
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  12.2533     0.1638  74.795  < 2e-16 ***
## sexM         -0.8472     0.2559  -3.311 0.000982 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.206 on 647 degrees of freedom
## Multiple R-squared:  0.01666,    Adjusted R-squared:  0.01514
## F-statistic: 10.96 on 1 and 647 DF,  p-value: 0.0009815
```

The average grade for females is 12.25 whereas the average grade for Males is 11.40. The difference between the two is significant ($p < 0.001$); the R-squared is 0.015.

Regression Using Family/Home Attributes

```r
lm_family <- lm(final_grade ~ fam_size + parents_cohabit + fam_edu_support + family_relationship + mom_

summary(lm_family)
```

```
##
## Call:
## lm(formula = final_grade ~ fam_size + parents_cohabit + fam_edu_support +
##     family_relationship + mom_edu + mom_job, data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -13.3384  -1.6538   0.0319   2.0203   7.4870
##
## Coefficients:
##                           Estimate Std. Error t value Pr(>|t|)
## (Intercept)                 9.3701     0.7137  13.129  < 2e-16 ***
## fam_size>3                 -0.3588     0.2796  -1.283 0.199911
## parents_cohabitTogether     0.1837     0.3892   0.472 0.637060
## fam_edu_supportTRUE         0.2188     0.2558   0.855 0.392604
## family_relationship         0.1953     0.1294   1.510 0.131569
## mom_edu                   0.5132     0.1416   3.624 0.000314 ***
## mom_jobhealth             1.0470     0.5903   1.774 0.076597 .
## mom_jobother              0.3878     0.3418   1.134 0.257024
## mom_jobservices           0.4969     0.4124   1.205 0.228669
## mom_jobteacher            0.8951     0.5599   1.599 0.110407
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.137 on 639 degrees of freedom
## Multiple R-squared:  0.07044,    Adjusted R-squared:  0.05734
## F-statistic:  5.38 on 9 and 639 DF,  p-value: 4.05e-07
```

Note that `dad_edu` and `dad_job` were not used since they are highly correlated to `mom_edu` and `mom_job`, respectively.

Out of the variables used, it looks like `mom_edu` has the only significant estimate. The estimate of `0.5132` suggests that each additional level of education reached by the mother corresponds with an additional `0.5132` points in the student's final grade in Portuguese.

Regression with Student Behaviors

```r
lm_habits <- lm(final_grade ~ paid_extra_classes + extracurricular_activities + attended_nursery + wants

summary(lm_habits)
```

```
##
## Call:
## lm(formula = final_grade ~ paid_extra_classes + extracurricular_activities +
##     attended_nursery + wants_higher_education + has_romantic_partner +
##     going_out_amount + weekday_alcohol_cons + weekend_alcohol_cons,
##     data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -13.1748  -1.6646   0.1527   1.8600   6.5574
##
## Coefficients:
##                                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)                     10.231070   0.588339  17.390  < 2e-16 ***
## paid_extra_classesTRUE        -0.800661   0.497388  -1.610   0.1079
## extracurricular_activitiesTRUE  0.386926   0.237758   1.627   0.1041
## attended_nurseryTRUE          -0.003662   0.297699  -0.012   0.9902
## wants_higher_educationTRUE      3.174917   0.387856   8.186 1.47e-15 ***
## has_romantic_partnerTRUE      -0.401965   0.246684  -1.629   0.1037
## going_out_amount              -0.041095   0.109327  -0.376   0.7071
## weekday_alcohol_cons          -0.369190   0.163366  -2.260   0.0242 *
## weekend_alcohol_cons          -0.204184   0.122994  -1.660   0.0974 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.997 on 640 degrees of freedom
## Multiple R-squared:  0.1502, Adjusted R-squared:  0.1396
## F-statistic: 14.14 on 8 and 640 DF,  p-value: < 2.2e-16
```

Out of the variables used, it seems that only wanting a higher education (i.e. `wants_higher_education`) and weekday alcohol consumption (i.e. `weekday_alchol_cons`) were significant.

```r
lm_habits2 <- lm(final_grade ~ wants_higher_education + weekday_alcohol_cons, data = data)

summary(lm_habits2)
```

```
##
## Call:
## lm(formula = final_grade ~ wants_higher_education + weekday_alcohol_cons,
##     data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -12.5393  -1.5393   0.2858   1.7498   6.6052
##
## Coefficients:
##                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)                  9.8587     0.4337  22.732  < 2e-16 ***
## wants_higher_educationTRUE   3.2529     0.3862   8.423 2.38e-16 ***
## weekday_alcohol_cons        -0.5723     0.1288  -4.442 1.05e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.006 on 646 degrees of freedom
## Multiple R-squared:  0.1367, Adjusted R-squared:  0.134
## F-statistic: 51.15 on 2 and 646 DF,  p-value: < 2.2e-16
```

The combined adjusted R-Squared id about 0.134 which is very low. However, the p-values for the two variables are both very low, suggesting that it is very unlikely that there is no relationship and the final score and either of those two variables. The strength of wanting higher education both has a larger slope estimate and a more significant p-value compared to weekday alcohol consumption. It is heartening to see that academic ambition corresponds strongly to academic achievement.

Regression with Hybrid Step-Wise Abroach

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

```r
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 hybrid step-wise linear regression model builder results in 14 different variables. There are 4 variables with higher statistical significance than the others, the school the student attended, the number of previous failures the student has, whether or not the student had extra educational support, and whether or not the student wants a higher education. Among these, the one that has the highest estimate on the influence is whether or not the student wants higher education, although all 4 variables have fairly similar estimates. That is also the only predictor that has a positive relationship with the final score. All the other predictors, including, surprisingly whether a student has extra educational support, have negative relationships.

Sadly, the adjusted R-squared is 0.3192. This is low enough that we cannot make good predictions, even if we can identify influential factors.