OLS Regression Project

David Purucker

Introduction

What is the relationship between popularity and academic achievement among adolescents? The question is an important one to answer because popularity seems to correlate with both positive and negative outcomes for adolescents. For example, higher teenage popularity is associated with a higher propensity to use drugs and alcohol and to engage in petty crime (Allen et al. 2014). Other research, however, finds a positive correlation between popularity and higher future earnings (Conti et al. 2013), though that result has failed to replicate in larger-dataset research (Fletcher 2013). Boosting academic achievement among young people is a perennial suggestion for improving the life chances of Americans and reducing social inequality. Given the importantce attached to academic success, and the ambiguous findings of research on popularity effects, research is needed to clarify the relationship between popularity and academic achievement. It may be the case that the observed strongly positive association between high school GPA and adult income (French et al. 2015) is reduced in magnitude when controlling for the popularity of studens. If academic achievement is strongly associated with popularity, and popularity is itself causally related to future earnings (e.g. via social capital gained in high school, and/or the development of networking skills valuable in the job market), then a policy shift towards supporting a more even distribution of social engagement among adolescents could be warranted. On the other hand, given the observed associations between popularity and certain negative health outcomes, at least in the short term, an association between popularity and academic achievement might warrant a policy approach that attempts to detach this 'unhealthy popularity' from academic success (if this is possible).

Using data from the National Longitudinal Study of Adolescent to Adult Health (Add Health), it is possible to analyze the relationship between various social and academic attributes of adolescents. Add Health is a nationally-representative survey of high school students in grades 7-12, collected in waves. This analysis uses data from the first wave, conducted in 1994-95. This analysis uses a simple measure of popularity, in-degree scores, which measures the number of times a student was nominated as a friend by other students in the school. In-degree scores are modeled using simple OLS regression for association with variables representing academic achievement and socioeconomic status measured in Add Health. To test the specific effects of academic achievement on student popularity, a number of potentially confounding variables are controlled for in the models. Independent variables used in this analysis are student membership in school honor societies (a binary categorical variable), parental income (measured in thousands of dollars), extracurricular sports participation (measured by number of sports), sex (male or female), and pseudoGPA. PseudoGPA is assessed by calculating GPA from student reports of their most recent grades in math, language arts, and science courses. These potential confounders were chosen becasue they represent a mix of academic, social, and demographic contributors to student popularity, and can therefore capture more of the complex identity characteristics that could contribute to popularity, which is unlikely to be purely an academic, extracurricular, or family-history (class) phenomenon.

Popularity could plausbily be associated with higher or lower academic achievement among adolescents. It may be the case that being more popular confers social capital that contributes to higher academic achievement - providing, for example, easy access to tutoring or mentoring, or eliciting more grading lenience from instructors. Academic achievement could itself contribute to popularity, for example, by eliciting respect from other students. On the other hand, popularity could create social pressures to avoid investing time in academics or reduce the short-term costs of low academic investment, or deviant behaviors associated with popularity could carry over into academic sanctions. Finally, it may be the case that social popularity and academic achievement are only weakly related but are both strongly associated with a third variable, such as parental income.

Testing for principles of OLS regression

OLS regression models and the variables that incorporate are subject to five common problems: heteroscedasticity, multicollinearity, missing values, design effects, and poor model selection.

Testing for heteroscedasticity

Heteroscedasticity means that the variance of the residuals in a linear model is not constant as x increases. A model testing the association of GPA with in-degree shows that the dispersion of residuals increases with the fitted values, producing a cone shape. At higher predicted levels of GPA, the variance of residuals increases. A model testing the association of the binary honor society membership variable with in-degree also shows some heteroscedasticity. Left unmodified, heteroscedasticity will produce a somewhat inefficient estimate of regression coefficients. A log transformation is the standard way to solve a probelm with heteroscedasticity. A test model with only the GPA variable logged does not seem to reduce heteroscedasticity compared to the un-logged model.

Testing for multicollinearity

Multicollinearity occurs when there is a moderate to high degree of correlation between the independent variables in a model. This increases standard errors and makes coefficient estimates unstable among different combinations of independent variables in a model. Multicollinearity between independent variables can be tested with a VIF (variance inflation factor) measure. VIF applied to a test model with all variables shows low values for all the variables in this dataset. All VIF values are well below 4, indicating that there is no problem with multicollinearity between these variables.

Testing for missing values, multiple imputations used to remedy

There are 1027 missing values in the parent income variable, and 1235 total missing values in the dataset (the remaining missing values are found in the student grade variable). Income measures are notorious for missing value problems. To correct for this, I use multiple imputation chained equations, a non-parametric matching method which generates and imputes plausible values for all variables with missing values. Five rounds of imputation are used to minimize imputation variability issues.

Testing for survey design effects

Survey design effects need to be adjusted for to ensure representative statistics and accurate statistical inference. To mitigate design effects, this analysis uses the 'cluster' and 'weight' values included in the Add Health dataset as arguments to the 'svyglm' command in R.

Models

```
##
## ====================================================================================================
##                          Model 1      Model 2      Model 3      Model 4      Model 5      Model 6
## ----------------------------------------------------------------------------------------------------
## Intercept                   2.26 ***     2.52 ***     2.27 ***     1.77 ***     2.36 ***     4.83 **
##                            (0.32)       (0.32)       (0.33)       (0.31)       (0.31)       (0.13)
## pseudo-GPA                  0.83 ***     0.70 ***     0.60 ***     0.66 ***     0.50 ***
##                            (0.11)       (0.11)       (0.11)       (0.10)       (0.10)
## member of honor society                  1.30 ***     1.20 ***                  1.06 ***
##                                         (0.32)       (0.31)                    (0.31)
## parental income                                       0.01 ***     0.01 ***     0.01 ***
##                                                      (0.00)       (0.00)       (0.00)
## number of sports played                                            0.46 ***     0.49 ***
##                                                                   (0.08)       (0.07)
## male                                                                            -0.51 ***    -0.46 **
##                                                                                (0.15)       (0.15)
## ----------------------------------------------------------------------------------------------------
## R2                            0.03         0.04         0.05         0.06         0.07         0.00
## BIC (null)                 -117         -147         -179         -248         -275           -4
## N                          4397         4397         4397         4397         4397         4397
## ====================================================================================================
## *** p < 0.001, ** p < 0.01, * p < 0.05

##
## % Table created by stargazer v.5.2 by Marek Hlavac, Harvard University. E-mail: hlavac at fas.harvar
## % Date and time: Wed, Apr 03, 2019 - 16:08:01
## \begin{table}[!htbp] \centering
##   \caption{}
##   \label{}
## \begin{tabular}{@{\extracolsep{5pt}}lccccc}
## \\[-1.8ex]\hline
## \hline \\[-1.8ex]
## Statistic & \multicolumn{1}{c}{N} & \multicolumn{1}{c}{Mean} & \multicolumn{1}{c}{St. Dev.} & \multic
## \hline \\[-1.8ex]
## b.pool & 2 & 1.545 & 1.013 & 0.829 & 2.261 \\
## se.pool & 2 & 0.216 & 0.151 & 0.109 & 0.322 \\
## t.pool & 2 & 7.312 & 0.420 & 7.015 & 7.609 \\
## pvalue.pool & 2 & 0.000 & 0.000 & 0 & 0 \\
## \hline \\[-1.8ex]
## \end{tabular}
## \end{table}
##
## % Table created by stargazer v.5.2 by Marek Hlavac, Harvard University. E-mail: hlavac at fas.harvar
## % Date and time: Wed, Apr 03, 2019 - 16:08:01
## \begin{table}[!htbp] \centering
##   \caption{}
##   \label{}
## \begin{tabular}{@{\extracolsep{5pt}} c}
## \\[-1.8ex]\hline
## \hline \\[-1.8ex]
## $4,397$ \\
## \hline \\[-1.8ex]
## \end{tabular}
## \end{table}
##
## % Table created by stargazer v.5.2 by Marek Hlavac, Harvard University. E-mail: hlavac at fas.harvar
## % Date and time: Wed, Apr 03, 2019 - 16:08:01
## \begin{table}[!htbp] \centering
##   \caption{}
##   \label{}
## \begin{tabular}{@{\extracolsep{5pt}} c}
## \\[-1.8ex]\hline
## \hline \\[-1.8ex]
## $0.028$ \\
## \hline \\[-1.8ex]
## \end{tabular}
## \end{table}
##
## % Table created by stargazer v.5.2 by Marek Hlavac, Harvard University. E-mail: hlavac at fas.harvar
## % Date and time: Wed, Apr 03, 2019 - 16:08:01
## \begin{table}[!htbp] \centering
##   \caption{}
##   \label{}
## \begin{tabular}{@{\extracolsep{5pt}} c}
## \\[-1.8ex]\hline
## \hline \\[-1.8ex]
## $$-$117.126$ \\
## \hline \\[-1.8ex]
## \end{tabular}
## \end{table}
```

Findings

The OLS regression models incorporate a nested structure, progressively controlling for potentially confounding variables to isolate the effect of academic achievement (measured in pseudoGPA and honor society membership) on student popularity scores. Model 1 assesses the correlation between pseudoGPA and popularity without controlling for any potential confounders. Model 2 adds a control for the other academic achievement measure, honor society membership. Models 3, 4, and 5 incorporate potential confounding variables with increasing parsimony, controlling progressively for parental income, number of sports played, and student sex (male or female). Model 6, which is not nested, removes the academic achievement measures and directly assesses the impact of student sex on popularity scores, without controlling for any other variables. All effects in the models are statistically significant, indicating that the association observed in the sample is statistically distinguishable from zero.

Model results

Model 1 predicts that the average student receiving a zero GPA score will have an in-degree score of 2.29. For each additional 1-unit increase in GPA, a student will receive 0.82 more friend nominations.

Model 2 predicts that a student receiving a zero GPA score and is not an honor society member [FALSE is the reference category for honor society] will receive 2.51 friend nominations. For each additional 1-unit increase in GPA, an average student will receive 0.70 more friend nominations, controlling for honor society membership. Students in honor societies receive 1.30 more friend nominations than students not in honor societies, on average, controlling for GPA.

Model 3 predicts that a student receiving a zero GPA score, is not an honor society member, and whose parents have zero income will receive 2.27 friend nominations, on average. For each additional 1-unit increase in GPA, an average student will receive 0.61 more friend nominations, controlling for honor society membership and parental income. Students in honor societies receive 1.19 more friend nominations than students not in honor societies, on average, controlling for GPA and parental income. For each additional thousand dollars in annual parental income, a student will receive just 0.01 more friend nominations on average, controlling for GPA and membership in honor societies.

Model 4 predicts that a student receiving a zero GPA score, plays no sports, and whose parents have zero income will receive 1.77 friend nominations, on average. For each additional 1-unit increase in GPA, a student will receive 0.66 more friend nominations, controlling for sports team participation and parental income. For each additional sport played, a student will receive 0.46 more friend nominations on average, controlling for parental income and GPA. For each additional thousand dollars in annual parental income, a student will receive 0.01 more friend nominations on average, controlling for GPA and sports participation.

Model 5 predicts that a female student receiving a zero GPA score, is not an honor society member, plays no sports, and whose parents have zero income will receive 2.36 friend nominations, on average. For each additional 1-unit increase in GPA, controlling for student sex, sports participation, honor society membership, and parent income, students will receive 0.50 more friend nominations. Honor society students receive 1.06 more friend nominations than non-honor society students on average, controlling for GPA, parental income, sports participation, and student sex. For each additional thousand dollars in annual parental income, a student will receive 0.01 more friend nominations on average, controlling for GPA, honor society membership, sports participation, and student sex. For each additional sport played, a student will receive 0.48 more friend nominations on average, controlling for GPA, honor society membership, parental income, and student sex. Male students in this model receive 0.51 fewer friend nominations than female students, on average, controlling for GPA, honor society membership, parental income, and sports participation.

Model 6 assesses the correlation between student sex and friend nominations without controlling for other variables. Model 6 predicts that direction and magnitude of correlation are similar to Model 5 for male students, who receive 0.46 fewer friend nominations than female students, on average. The model predicts that female students will receive 4.83 friend nominations, on average.

Discussion

The regression analysis indicates three primary findings. First, honor society membership is the strongest predictor of friend nominations. The effect of student participation in an honor society on popularity is about twice as strong as that for participation in additional sports or for GPA scores, though the effect of honor society participation decreases somewhat as additional variables are added to the models. GPA is also a fairly consistent and strong predictor of additional friend nominations, but the effect declines with the addition of honor society membership as an independent variable in models 2 and 3, and increases again when honor society membership is removed as a controlled variable in model 4. This leads me to hypothesize that academic achievement *per se* is not the main cause of the observed positive relationship between achievement and popularity. Rather, students who perform well academically gain access to academically-based social groups like honor societies, which offer opportunities for making friend relationships and increasing one's popularity. Students who perform well academically but who elect *not* to participate in honor societies would not have the same opportunities for making friends. The stronger popularity effect conferred by honor society membership compared to sports participation is a surprising finding that I am not sure how to explain.

Second, male students receive fewer friend nominations than female students, with similar magnitudes predicted by both controlled and uncontrolled models. I speculate that female students are likely to have larger social networks and more mutually-recognized friendship relations than male students. It may also be the case that male students commonly have an inflated sense of their social position vis-a-vis the networks of female students. Male students, then, may nominate female friends at a higher rate than those 'friends' reciprocate the nomination.

Third, parental income is very weakly correlated with student popularity across the models in which it is controlled. This is a surprising finding which contradicts commonsense understandings of the relation between class and social position, at least in schools. I assumed that students with higher class resources would have greater access to other popularity-conferring resources like extracurricular activities. I am not sure how to explain this finding.

Model selection

OLS regression models need to be assessed for goodness-of-fit - their balance between predictive accuracy and parsimony. The Bayesian Information Criterion (BIC) is an alternative goodness-of-fit measure. Preferred models score lower on BIC. Adjusted R-squared measures the accuracy of the model in predicting the dependent variable. Preferred models score higher on R-squared.

The regression table indicates that both adjusted R-squared and BIC favor model 5, which controls for the largest number of independent variables. Both model 1 and model 6 control for only one independent variable, but model 1 has a substantially stronger BIC score. Lower BIC for model 1 indicates that in-degree is much better-predicted by pseudoGPA than by being a male, as in model 6. The apparent 0.00 R-squared score for model 6 likely indicates that the score is in the hundredth or thousandth decimal and could not be rendered on the regression table.

Conclusion

The preceding analysis explored the relationship between adolescent academic achievemtn and popularity. A key finding is that a strong association between membership in academic honor societies and popularity. This effect was stronger than the effect of GPA score alone, indicating that academic achievement is connected to popularity indirectly through a social activity that depends on GPA achievement, rather than popularity deriving from academic achievement per se. Other findings include stronger popularity scores for female students than male students, and a very weak effect of student socioeconomic status on popularity. Future research should continue to explore the effects of popularity on student life chances, and should seek to clarify the relationship between the bundle of positive and negative effects of popularity on measures of academic achievement.