Analyzing Large U.S. Universities Prestige and its Relation to Student Privilege Using Statistical Regression

Social science is concerned largely with the understanding of human behavior in a social system. By social system, is meant any type of connection between people. A social system can be as small as two people, or as large as all the entire human population, with the latter constituting the whole reservoir of potential research in social science.

Science is concerned with the human study of the world as a whole. Some areas of science have discovered law relationships between matter, others have described laws of human commerce. Social science is often unlike others in that its discoveries are dynamic, or at least specific to time. Saying “the United States has the largest military” is saying a fact. But saying that 1000 years ago would be nonsense. In this way, social science is in a sense a scribe for the future, recording the present for posterity. Social science is only observational, with any professed experiments being only accurate observations. All science is observational at its core, except fields assume different thing about what they are observing. When a chemist observes how matter behaves in a specific condition, he or she assumes it to constantly behave that way under the same conditions. Social scientists don't hold this assumption about human systems over time.

This paper is intended to give insight into the current nature of United States universities. Specifically, what is sought is to what extent does the prestige of a University explain the financial characteristics of students who are attending it. In other words, I am seeking if “good” and “bad” schools equally represent poor and wealthy students. It is my hypothesis that more prestigious schools have a higher proportion of privileged students. This analysis looks only at large schools (20,000 or more students) that primarily award baccalaureate degrees or higher, are Title IV participating which allows their students to participate in federal financial aid programs, and educate primarily offline. This totaled to 154 universities. All data is from the 2012-2013 academic year.

The National Center for Education Statistics (NCES) collects information about all universities in the United States. Each school is legally required to fill out a form giving data about its characteristics and to some extent its students' characteristics.

Prestige is a variable impossible to measure with complete accuracy. However, I have characterized prestige with three variables observed by the NCES. These variables are: the 75th percentile of ACT math scores at each university (referred to as ACT), the percent of students admitted to each university (referred to as percent admitted), and the total cost of attendance for out of state students living off campus, not with family, for each university (referred to as cost). Out of state cost was used to reconcile the fact that public universities charge vastly different tuition to in-state and out of state students, while private universities charge a constant amount. To measure privilege, I use the variable percent of students receiving Pell grants (referred to as Pell), where a Pell grant is a need based financial award given to students. Larger ACT, smaller percent admitted and larger cost are relationships that indicate an increase in prestige, while a smaller Pell value indicates an increase in the privilege of students.
Some universities omitted ACT values, because either they accepted only an equivalent SAT score from students or they did not require test scores for enrollment. There were six universities in the former category, and using regression imputation, I converted these math SAT 75th percentiles into math ACT 75th percentiles. This I assume to be valid, given the high correlation between these variables: the R squared value, describing the percent of variability one variable explains the other, was over 90 percent. Twelve universities still remained uncharacterized by the math ACT 75th percentile. These universities had to be omitted in computing the statistical model.

Still other universities omitted percent of students admitted, for, according to the NCES website, these universities had an open admissions policy. For these omissions, I concluded their percent admitted value to be 100 percent.

In examining the relationship between the predictor variables and the response, I first look at each relationship individually. The strongest predictor is the ACT score. However, the relationship has curvature, as seen in the residual plot between the variables (Figure 1).

To account for this, I performed a Box-Cox scaled power transformation. Figure 2 shows a power of -2 as the best choice. Although not ideal, this transformation removes most of the curvature. On the y-axis is the Pell variable, and on the x-axis is the ACT variable.
The percent admitted and cost variables show a more variable relationship, but are informative and linear nonetheless. Six universities omitted the cost variable, and where hence excluded in computing the model. These variables do not suggest a non-linear relationship.

Provided below is statistical output from R of the multiple linear model with response Pell and predictors ACT, percent admitted and cost. It suggests ACT being the most important variable, followed by percent admitted and lastly cost.

```
Call: lm(formula = PercentPell ~ ACT + percent.Admitted + cost)

Residuals:
    Min     1Q    Median     3Q    Max
-18.199 -4.489  -1.142   3.582  23.640

Coefficients:         Estimate Std. Error t value  Pr(>|t|)
(Intercept)       2.979e+04  2.765e+03   10.774  < 2e-16 ***
ACT              -5.957e+04  5.540e+03  -10.754  < 2e-16 ***
percent.Admitted -1.316e-01  4.363e-02  -3.016   0.00307 **

---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 7.61 on 132 degrees of freedom
(18 observations deleted due to missingness)
Multiple R-squared:  0.5606,    Adjusted R-squared:  0.5506
F-statistic: 56.14 on 3 and 132 DF,  p-value: < 2.2e-16
```

Interestingly this model gives a non-linear relationship, as shown by its residual vs. fitted plot (figure 3). Even more interesting is when removing the transformation shown above on ACT, the curvature nearly vanishes (see figure 4). I cannot explain this other than pointing to variable interactions.
Due to this comparison I elected to choose the second model, without transforming any variables. Below is the R output of this model.

---

This output suggests that ACT is still the most important predictor, again followed by percent admitted and lastly cost. The coefficient estimate for ACT is -3.291. This suggests that, on average, for every 1 point increase in the score of ACT values, the percent of students receiving Pell grants decreases by just over 3 percent. Figure 2 shows the plot of Pell and ACT, notice the distinct, though variable, relationship. Similarly, the coefficient estimate for percent admitted is -.1646, a much weaker relationship than ACT, but still significant. This coefficient suggests that, on average, universities that admit more students have a lower proportion of students receiving Pell grants.

This last result seems counter-intuitive, so I compared this coefficient to the estimate given by simple linear regression of Pell against percent admitted. The simple coefficient is +.1477, small but still significant, and in the opposite direction of the more complex model. This discrepancy must be attributed to variable interaction.

The output suggests cost is not significant at all. Despite this, I shall continue to include it due to its seemingly logical relationship with prestige.

Furthering the line of inquiry into university prestige and privilege, we see that there is, as expected, a strong relationship between the two. But care must be taken in describing this relationship. Firstly, the data used is purely observational, and tells very little about student privilege. The Pell grant proportion of each university is the closest measure to privilege we can study. A better variable would be, for example, median income of students' families for each university. Prestige too is hard to measure and a non-quantitative variable. Better variables for measuring prestige would be, for example, number of research articles published by the university annually or percent of graduates becoming experts in their
respective field.

The fact that we cannot measure the variables in question with high accuracy does little to diminish the importance of the statistical model we can create. Again, care must be taken in its interpretation.

Firstly, let us look at what the model says. The coefficients, with the minor exception of percent admitted, are positively correlated, suggesting an increase in prestige results in increase proportion of privileged students. The adjusted R squared value is estimated to be .5507, meaning prestige explains away 55 percent of the variability inherent in the Pell variable. In other words, given a prestige estimate, we can estimate with 55 percent more accuracy what the corresponding Pell value will be.

These figures are only as valid as the model is. To test the validity of our model, we check against the standard regression assumptions. As shown in figure 4, our model does not provide unquestionable certainty of linearity, but it also does not break the assumption. Figure 4 also provides no evidence against the assumption of equivariance of errors, instead it provides great support towards it.

Figure 5 presents a Q-Q plot of the residuals. The residuals look well except for the right hand side. This I attribute to an unexpected density distribution in the residuals between values +15 to +22 (shown in figure 6). This “hump” is due to a group of observations that do not fit well with the others. Specifically, a group universities have a much higher rate of Pell recipients than average. The distribution of the errors appears bimodal because of this, but is otherwise normal.

Investigating further, I examine how the differences from average are related spatially. Each university under study was provided with a postal Zipcode. I converted these Zipcodes to latitude and longitude values and plotted them over a map of the United States with the size of the residual given by the size of the point. Figure 7, with the black dots, refers to positive residuals over space, and figure 8 with the red dots to negative residuals over space.

The maps are extremely informative. Firstly, they provide strong evidence against the assumption of independence of errors. We can see from figure 7 that the California and Florida universities under study have mostly big, positive residuals. This means that California and Florida have a greater than
expected proportion of students receiving Pell grants, given their prestige. Conversely, from figures 7 and 8 jointly, we can observe that near the East coast of the United States, almost all universities under study have moderate to large negative residuals, meaning that this geographic area has a less than expected proportion of students receiving Pell grants, given their prestige.

Although informative, this information also suggests that the statistical errors are correlated with geography, making our current model less reliable. This means that Pell can be explained by more than prestige alone. Our objective, however, is not to explain Pell, but to discern its relation to prestige, that is, the current predictors. Why it is that Florida, California and the East coast fail to regress should be the question of further investigation.

**Figure 7**

This plot shows locations of positive residuals in black, with the size of the dot representing the difference from the mean function.

**Figure 8**

This plot shows the locations of negative residuals in red, with the size of the dot representing the difference from the mean function.
We can nonetheless confidently infer that more prestigious schools are more likely to have a greater proportion of privileged students. This relationship is less interesting than what it implies. Students at more prestigious universities are more likely to be privileged, indicating that prestigious schools do not equally represent poor and wealthy students. This is probably a truisum. But, it means either, that prestigious universities discriminate between wealthy and poor students, or that wealthy and poor students attend universities according to factors related to richness and poorness.

I believe the first option to be false, and the second to go hand in hand with a manipulation of the first. It is not that schools discriminate between wealthy and poor students, but that they discriminate according to factors that richness or poorness are associated with. It is understandable that wealthy parents will be more financially stable, providing, for example, a better education for their child. If wealthier, or equivalently, more privileged students receive a better pre-university education, they will fare better in attending a better university. This is because universities discriminate upon student aptitude and knowledge, characteristics we assume to be in large part determined by pre-university education. This assumption, of course, deserves further investigation with its own research.