Analysis of College Reputation in Relation to Structural Characteristics

Eric Huang

01-13-2023

Executive Summary

We investigate the relationship between structural characteristics of US colleges and their prestige. We find that the most important of the variables that we considered is age, with a higher age being significantly associated with better rankings on a consistent basis. In addition, when considering school type, in which a higher school type denotes awarding higher-level degrees such as doctorates and a lower school type denotes awarding lower-level degrees such as associates, we also find that higher age is significantly associated with better rankings on a consistent basis.

We find that having a Special Purpose, most notably having HBCU status, seems at first significant and negatively correlated with prestige, but is actually positively correlated with higher prestige after controlling for SAT scores. We find that being religious is a consistent, negative factor for the most prominent colleges. We find that being a public school is significantly associated with being more developed in school type. However, whereas public vs private seems to matter very little for most school rankings, among the very oldest universities, private schools far outrank public schools. These oldest universities also associate small city locales with better ranks above all other locales, contrasting the rest of the schools which

favor large cities, then small cities, then rural areas. These oldest universities also ignore the correlations with state population and religious affiliation that the rest of the universities have. We note that while age is significant for school type, it hardly affects the school type for private schools. Finally, we note that a school being at the state capital and a school being off the mainland USA are inconsistent factors to associate with prestige, while state population is moderately consistent.

Introduction

Today, America's most prestigious colleges wield an enormous amount of power and influence. This small subset of schools possess billions in endowment funds, graduate alumni who create billion dollar companies and win Nobel Prizes, and lead the world in cutting-edge research. They are a coveted destination not just for the most ambitious American high schoolers, but also for world leaders and their children.

Yet these schools were founded from humble beginnings. For example, Harvard University was established in order to educate the clergy, and graduated only nine students during its first commencement ceremony. Its founders would have been hard-pressed to imagine that their schools would become synonymous with prestige and reputation.

Motivated by these observations, we wish to explore the following: out of the thousands of American colleges, why do some achieve prestige and status, while other do not? Is it random chance, or can we systematically identify structural differences between colleges who reached the top, and those who did not?

We believe this question to be both intellectually and practically important.

On an intellectual level, understanding why certain certain institutions have become highly competitive, scarce commodities is crucial to understanding why the current educational landscape is the way that it is.

On a practical level, if such characteristics can be identified, then perhaps this phenomenon

can be recreated. Technology and society is constantly evolving, and there may come a time when the next set of institutions will need to be created. Countries and companies will want to know what matters, to understand the systems of building up soft power, and fight to come out on top. Can you buy your way to the top? Or does it become largely out of your hands after a certain point?

There has been some literature investigating the most significant factors in educational prestige, such as Volkwein and Sweitzer 2006, but they have all used factors that we believe act as effects of prestige itself. For example, average SAT score may be a strong predictor of college ranking, but we are fairly certain that prestigious colleges attract more high-performance students. Thus, high SAT scores tell us little about the nature of how the school become so prestigious in the first place. Other examples of factors that likely act as both causes and effects of prestige are faculty pay, research output, graduation rate, post-graduation outcomes, and institutional funding.

The chart below made by Volkwein and Sweitzer is a good examination of these factors. They propose their concept of how school reputation is created.

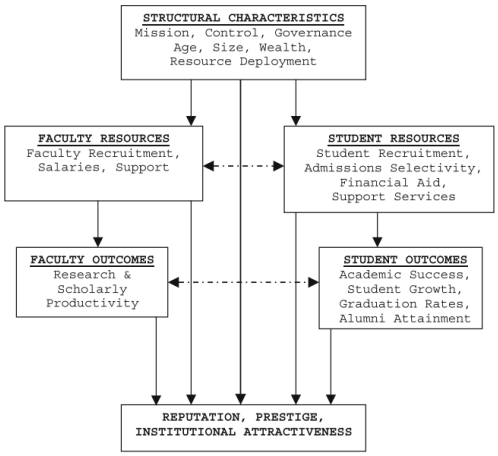


FIG. 1. Conceptual model.

Motivated by our reasoning above, we propose an alternate concept. In this new concept, not only do the non-structural factors feed into the outcome, but the outcome also feeds into these factors.

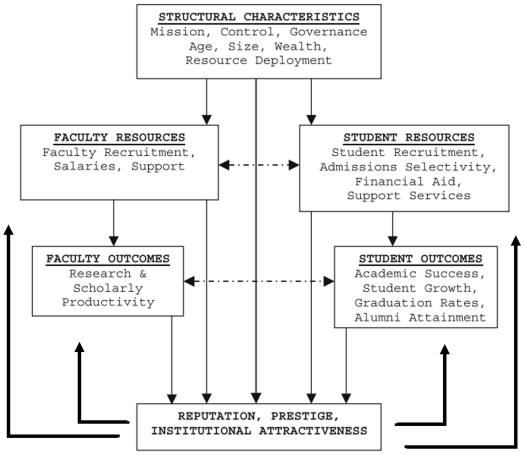


FIG. 1. Conceptual model.

This alternative concept of how school prestige works motivates us to use only structural factors, and not factors such as student outcomes or faculty resources, in order to find evidence of root causes of prestige. Considering factors such as faculty productivity or graduation rate would be similar to trying to predict predict based on an effect of prestige. (One interesting thing to note is that US News College Rankings use these factors in their ranking formula. However, they have historically used opinion surveys of academics. The rankings have changed little since then, and we can view the usage of these factors as a proxy for reputation with a more objective justification with less elitism.)

We wish to investigate factors that are structural characteristics and are unlikely to have been caused by prestige itself. Thus, these are much stronger evidence for being causal factors of educational prestige.

If these structural characteristics have little statistical significance, then perhaps universities gain prestige in a random fashion. For example, a certain college might, through chance, welcome in a stronger than average class, which increases their reputation, which attracts a stronger incoming class, and continues to snowball in that fashion.

We develop some initial hypotheses that we could like to investigate based on casual observation. The most primary is the factor of age. When we examine the highest ranked schools

in the country, we notice that many of them were founded in the 1700's. Looking at a list of the 10 oldest colleges, we notice that half of them are prestigious, Ivy League universities.

Another hypothesis is that a higher state population might provide schools a larger pool of students that apply, thus allowing schools to select the best in a large pool, or perhaps providing the schools with more funding from tuition. Additionally, we hypothesize that public schools would benefit more from a larger state population, as public schools draw from their local state population more than private schools do.

In addition, I would like to investigate the effects of "mission", as stated by the chart by Volkwein and Sweitzer. Specifically, I hypothesize that having a niche special mission would lower a school's appeal more than a broadly appealing school. For this mission, we look at the characteristics of being a religious school, being a gender-exclusive school, and being a school with a mission and history of serving an ethnic minority group. Perhaps the most relevant of these categories is that of Historically Black Colleges and Universities (HBCU), as they have received close attention in politics in recent years. In 2019, Senator Warren proposed $50 billion in HBCU funding and in 2022, President Biden says "HBCUs don't have the endowments others have, but guess what? You're just as smart. You're just as bright. You're just as good as any college in America." I have not found literature on this either. From glancing at rankings, one can see that there are few HBCUs in the very top ranks, but a formal statistical analysis would be able to compare HBCUs with similar characteristics, as most of them were founded later in American history.

Another casual observation, since the author grew up in Texas, is the flagship University of Texas at Austin. Is there an advantage at being located at the capital? If so, we might also expect that public schools benefit more from this effect than do private schools.

In addition, we note that the top of the college rankings is dominated by private schools, so we wish to see the effect of a school's control (i.e whether a school is private or public).

We wish to investigate whether these factors are associated with greater college reputation,

and whether their correlation suggests a positive or negative effect. After identifying such factors, do their correlation suggest a positive or negative effect on reputation, and how large of an effect is it?

Through this, we would like to gather evidence that certain factors might be causes of college prestige. This would allow social scientists and policymakers to understand why the current college landscape is the way it is currently. It would allow possible applications of these findings to schools in other countries (for example, countries that are starting or have recently started to develop higher education). It would have possible applications for the establishment of new institutions in domains outside of education. For example, if the future comes to a Metaverse race or an AI race, what factors would cause certain institutions to come out on top? Could the findings presented here be used to get ahead?

In addition to rankings and reputation, we can general think of college classification as a kind of outcome of success as well. That is, are there factors that can be used to predict whether a school will end up as a doctorate-granting research university, or a school that primarily gives out bachelor's degrees, or a community college?

Datasets

The first dataset we use is the US News College Ranking 2022-2023 (note that around September 2023, this link will direct to the US News College Ranking 2023-2024, and so on every year). This is the most popular rankings list of colleges and universities. The college rankings are divided into multiple categories. There are National Universities (research universities) and Liberal Arts Colleges, which are the two most prominent lists. There are also Regional Universities and Regional Colleges for the four regions of North, South, Midwest, and West. We only use these ten datasets for this analysis, although there are other lists on the website, such as Best Value Schools, Top Law Schools, and Best Colleges for Veterans.

We collect the data from the US News Ranking using the RSelenium package in R. From the US News Rankings, we extract, for each school, the college rankings (as ranked by US News), the founding date, and a unique numeric ID from hidden metadata. Around 191 schools had missing founding dates on the US News Ranking website, so these founding dates were searched for using Google. The unique numeric ID is the Integrated Postsecondary Education Data System's (IPEDS) Unit ID. For schools ranked in the bottom quarter, US News publishes a range instead of exact ranks. Schools with a range instead of an exact rank are assigned the rank that is the average of this range (e.g. if a school's rank is the range 300-400, their rank would be converted to 350). In total, there are 1631 schools in all datasets.

We present each US News list an overview of the data, with a selection of schools at different rankings and counts for the ranked and unranked schools. Note that "75th percentile" corresponds with the school ranked above 75% of schools in its respective dataset, and "50th percentile" corresponds with the school ranked above 50% of schools in its respective dataset.

list_name	n (ranked)	n (unranked)	top ranked	75th percentile	50th percentile
natlonal	440	3	Princeton University	William Carey University	Illinois State University
liberal	201	9	Williams College	University of Mary Washington	Cornell College
regcol_north	45	9	United States Coast Guard Academy	Vermont Technical College	Thiel College
regcol_south	99	33	High Point University	Central Baptist College	University of Puerto Rico--Aguadilla
regcol_west	46	56	Embry-Riddle Aeronautical University--Prescott	Paul Quinn College	Lewis-Clark State College
regcol_midwest	76	10	Cottey College	Hannibal-LaGrange University	Olivet College
reguni_north	175	6	Providence College	La Roche University	Regis College
reguni_south	135	1	Rollins College	Faulkner University	Auburn University at Montgomery
reguni_west	117	3	University of Portland	Humphreys University	Texas A&M International University
reguni_midwest	166	1	Butler University	Indiana University--Kokomo	Northern Michigan University

The second dataset we use is the U.S. Department of Education College Scorecard from 2022. This is a dataset compiled by the US government on institutional characteristics, enrollment, student aid, costs, and student outcomes, with the purpose of aiding students in making financial decisions about college. From this dataset, we extract, for each school, the state the school is located in, control type (i.e. public, private, or for profit), Carnegie Classification (e.g. Doctoral, Master's, Baccalaureate, or Associate's), locale (e.g. city with

population of 250,000 or more, suburb outside city with population between 100,000 and 250,000), HBCU and tribal status, religious affiliation, latitude/longitude, and IPEDS Unit ID.

The third dataset we use is the US State Demographics csv, compiled by Whitcomb, Choi, and Guan of the CORGIS Dataset Project, using data collected by the US Census. From this dataset, we extract, for each state, total population, and population for the ethnicities White, Black, American Indian and Alaska Native, Asian, Native Hawaiian and Other Pacific Islander, Two or More Races, and Hispanic.

The fourth dataset we use is the US State Capitals csv, compiled by Jochen. From this dataset, for each state, we extract the latitude/longitude coordinates of the state capital. We then manually added capital information for US territories (e.g. Puerto Rico and the Virgin Islands).

We join together the US News dataset and the Government Scorecard dataset using the IPEDS Unit ID. We join the State Demographics and State Capitals datasets to this dataset using the state that the schools are located in.

One of our assumptions about the data is that the change in real prestige as rank changes is higher for better ranked schools than it is for lower ranked schools. For example, the difference in reputation between the rank #1 and rank #30 schools is larger than the difference in reputation between the rank #301 and rank #330 schools. To encode this assumption, we will use the log of the rank in our visualizations below, because the change in log(x) for large values of x is smaller than for small values of x.

Because there are very few observations of For_profit control, and because this lack of observations complicates the Hessian matrix of when fitting future models, we change all instances of For_profit to Private. Each dataset has relatively very few instances of For_profit.

The three levels of variable Locale are Large City, Small City, and Rural/Town. Large City corresponds to being inside or in the suburbs of a city with a population of 250,000 or more.

Small City corresponds to being inside or in the suburbs of a city with a population of less than 250,000. Rural/Town corresponds with being in a rural area or being in an urban cluster separated from an urbanized area.

Special Purpose is a true or false variable. Special Purpose is true if a school is a Historically Black College or University (HBCU), a men's only college, a women's only college, or a tribal college.

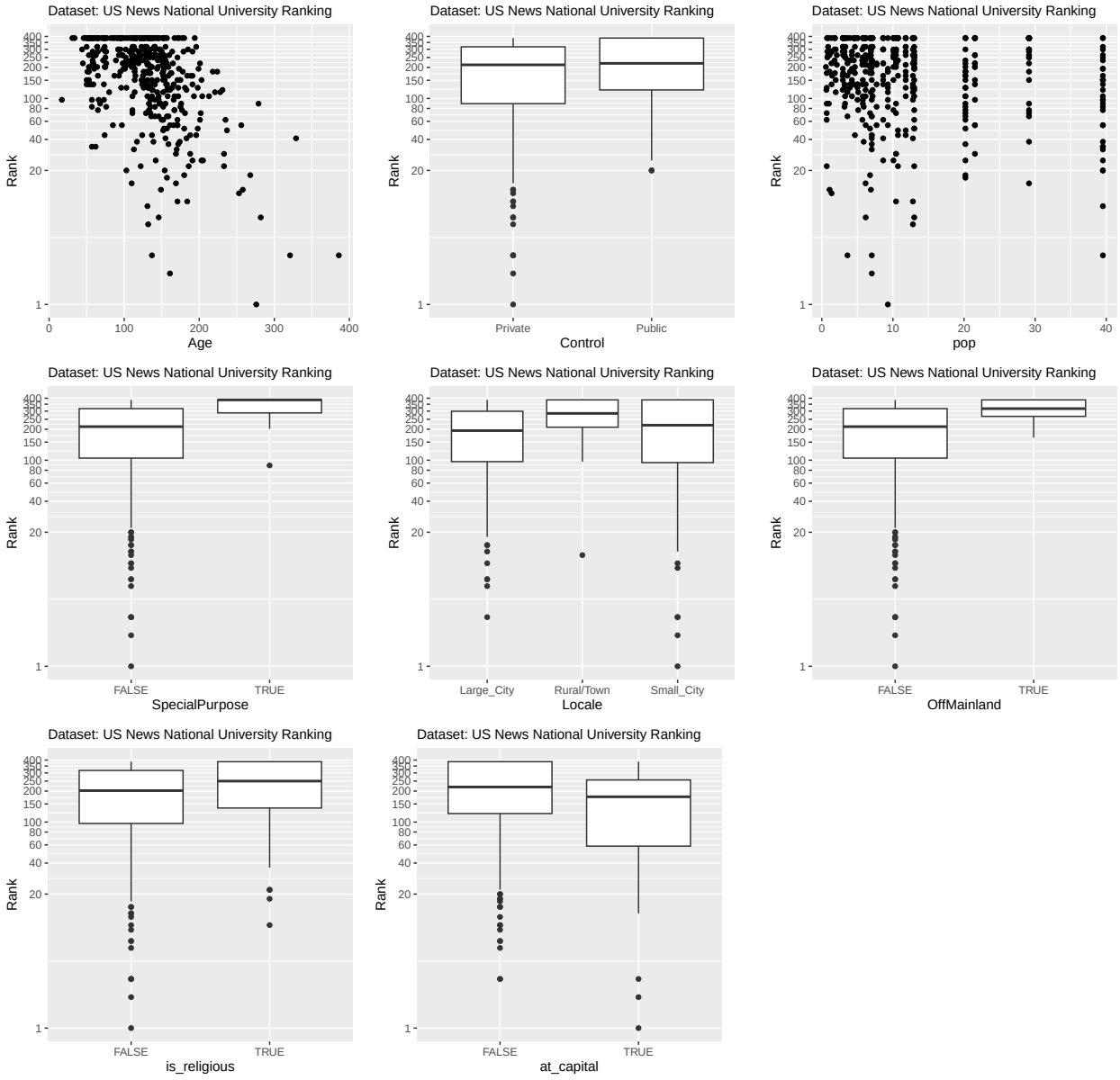
Off Mainland is a true or false variable. Off Mainland is true if a school is located off the mainland of the United States, which would be in the states of Alaska and Hawaii, as well as Puerto Rico and other U.S. territories.

At Capital is a true or false variable. At Capital is true if the school's coordinates put it within 15 miles of the coordinates of the state capital (note that for schools in Washington, DC, DC itself acts as the capital).

The population variable is the population of the state that the school is located in. It has been divided by one million for easier interpretation of coefficients. We hypothesize that, since there are drastic differences between some states in population, the square root of the population might be a more effective variable. Our reasoning is that it gives a more qualitative look at the differences in population, as the differences will be less drastic between states like California (39 million) and Rhode Island (1 million).

Using 10-fold cross validation on a linear model on the national dataset between a model using population and a model using the square root of population, we find that the population model has an average mean squared error of 0.6054 and the square root population model has an average mean squared error of 0.606. We thus opt to use the raw population.

We create visualizations to gain an understanding of the patterns in the data. Note that these are only for the National University dataset, and that although the data points are along the y-axis have their position decided by the log of the rank, the y-axis label is the untransformed rank, for easy readability.



Models and Analysis

We first fit a linear regression on log rank for each dataset. For our base model, our response is the log rank, and our variables are Age, Control, pop, SpecialPurpose, Locale, OffMainland, is_religious, and at_capital. The table of coefficients and p-values is below. (Note that for this table only, a negative coefficient for a variable means that an increase in the variable corresponds with a better rank. For all other coefficient tables, a positive coefficient for a

variable means that an increase in the variable corresponds with a better rank.)

list_name	n	(Intercept)_coef	(Intercept)_pval	Age_coef	Age_pval	at_capitalTRUE_coef	at
national	443	6.27319420926213	0	-0.0104656486619065	0	-0.31279761706443	
liberal	210	4.80858700603325	0	-0.00543587982558825	0	-0.144032938192194	
regcol_north	54	3.97871027183149	0	-0.00868921339902355	0.01	0.563126717280471	
regcol_south	132	4.62256488851793	0	-0.00604565530091642	0.02	0.240440243306814	
regcol_west	102	3.35249072429918	0	-0.00754457894928866	0.09	-0.183885244759886	
regcol_midwest	86	3.79241106563857	0	-0.00484766811770001	0.14	0.121815876337284	
reguni_north	181	5.038770128392	0	-0.00407270324313738	0.01	-0.704374796339356	
reguni_south	136	4.47931170675654	0	-0.00768044678352523	0	0.353800633087278	
reguni_west	120	4.98872339988202	0	-0.00904970921294362	0	0.113804158702038	
reguni_midwest	167	5.26009222079952	0	-0.00706788423015502	0	-0.525250983641681	

We also fit a cumulative link model regression on each dataset. A cumulative link model is a regression for ordered, categorical response variables. In this model, the log odds probability of being less than or equal to a certain category is modeled as a linear function of the variables. Each category has a unique intercept that either increases or decreases monotonically as you move up or down the categories. The model is shown below, where k is one of the outcome categories.

$$\log \frac{P(Y_i \leq k)}{1 - P(Y_i \leq k)} = \beta_{k,0} + \beta_1 X_{i,1} + \beta_2 X_{i,2} + \dots$$

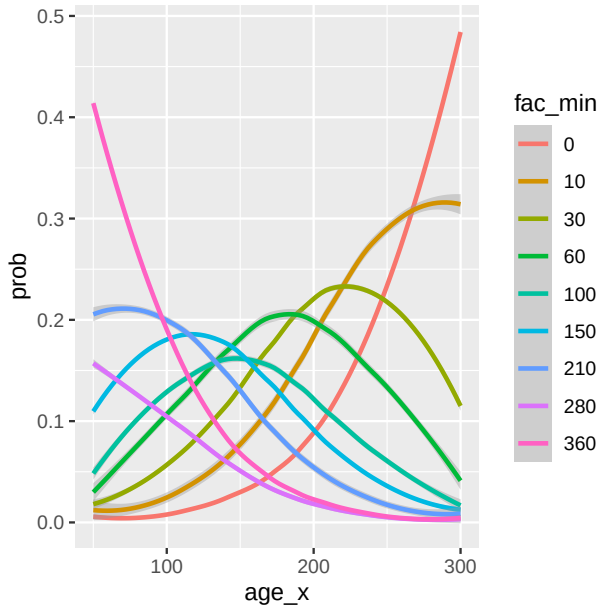
The simple model makes the assumption that the effect of the covariates on the log odds are the same for each category. This uniform effect of the covariates and the monotonic change in intercept ensures logical interpretations of the probabilities. As we move towards better ranked categories, our probability of being better than or equal to that category must decrease (e.g. the probability of of a school being ranked better than or equal to #50 is greater than the probability of of a school being ranked better than or equal to #40). The parameters of the model are fitted using maximum likelihood. The models are implemented through the R package "ordinal".

Our usage of the clm model is motivated by the fact that the linear regression model has the assumption that the output is a real number. When we use raw ranks as the output, we can get fitted values/predictions of ranks in the negatives, and when we use log ranks as the output, we can get fitted values/predictions numerically higher than the number of schools

in the dataset. Additionally, numeric outputs come with the interpretation that rank #20 is "twice as much" as rank #10 (or in the case of log ranks, rank #27 is "twice as much" as rank #10).

In order to turn our ranks into an effective categorical variable, I implement a custom ordered factor maker function. It takes as input ranks from 1 to n, as well as a positive integer "interval" (let us assume "interval" is its default value, 10). Categories are made starting from the best rank. The first category has a range of "interval", so all schools ranked #1-10 are in this interval. The second category has a range of 2 times "interval", so all schools ranked #11-30 are in this interval, and so on. Like the log ranks, this encodes our assumption that the real difference in prestige decreases as ranks decrease.

To illustrate the probabilities distributions that the cumulative link model gives as a variable changes, consider the graph below. It shows the probability distribution of being in a certain rank range as Age increases for a dummy school. We can see that the probability of being in a better ranked category increase as age increases, and decrease as age decreases, and vice versa for worse ranked categories.



We show our coefficients and p-values below, removing variables if a dataset has insufficient

counts for a good Hessian matrix for model fitting.

```
## Removing OffMainland from liberal
## Removing is_religious from regcol_north
## Removing at_capital from regcol_north
## Removing SpecialPurpose from regcol_midwest
## Removing SpecialPurpose from reguni_west
```

list_name	n	Age_coef	Age_pval	at_capitalTRUE_coef	at_capitalTRUE_pval	ControlPublic_c
national	443	0.0227017478013429	0	0.460052673078857	0.06	-0.52623782283
liberal	210	0.0126010293562944	0	0.363978140552131	0.49	-1.0224124929
regcol_north	54	0.0235693256484131	0			0.93367931504
regcol_south	132	0.016347447053018	0.01	-0.635136810936112	0.32	0.34846588978
regcol_west	102	0.0298892685779917	0.01	0.185878009368889	0.85	-0.46975870704
regcol_midwest	86	-0.000440593597669256	0.95	0.0312806583360247	0.97	-1.89071336
reguni_north	181	0.0107093472891747	0	0.933938378222402	0.03	-0.48445991622
reguni_south	136	0.0161350755669374	0	-1.62286331804217	0.02	0.060361889984
reguni_west	120	0.0202830066103756	0	0.117797922586428	0.83	0.90835808358
reguni_midwest	167	0.0181807426239567	0	0.971340165154261	0.08	0.7740901274

We seek to test if the cumulative linked model has any objective advantages over the linear regression model. To do so, we split each dataset into an 80-20 training/testing split, then compare the predictions with the observations, and find the error rate across all data sets. The predictions of the linear regression model are "rounded" to their nearest categorical counterpart.

```
## Removing at_capital

## Removing at_capital
```

The cumulative link model had a 0.6439546 error rate, while the linear regression model had a 0.7044088 error rate. We define the "residual" of a categorical ranking prediction as the number of steps it takes to arrive at the correct category in order. The cumulative link model had a 1.1215765 absolute mean residual and 2.660655 mean squared residual, while the linear regression model had a 1.0875084 absolute mean residual and 2.188377 mean squared residual.

In the comparison of the two models, since the clm made moderately less mistakes than the linear model, and the linear model made moderately smaller mistakes than the clm, we conclude that the two are roughly equally viable for use. We choose to favor the cumulative link model for its better fit to the data type of the response.

We consider not only school rank, which can be viewed as the social standing of a school, but also school classification, which can be viewed as the capabilities of the school. We consider a simplified version of the Carnegie Classification, which classifies schools based upon their highest degree awarded. We also consider the US News classification, which is based on the Carnegie Classification, and which US News Report uses to categorize schools into rankings lists.

The Carnegie Classification has levels such as "R1: Doctoral Universities – Very high research activity", "M3: Master's Colleges and Universities – Smaller programs", and "Associate's Colleges - High Transfer-High Traditional". The full list can be found here. We simplify these categories into an ordered factor with four categories. They are, in order from best ranked to worst ranked: Doctoral, Master, Bachelor, and Associate. We then combine all US News datasets, and run a cumulative link model on this as a response, with the variables that we used before.

Our results are as follows:

```
## formula:
## Type ~ Age + Control + pop + SpecialPurpose + Locale + OffMainland + is_religious + a
## data:    data %>% as.data.frame()
##
##  link  threshold nobs logLik   AIC     niter max.grad cond.H
##  logit flexible  1627 -1916.36 3856.72 6(0)  1.34e-10 1.1e+06
##
## Coefficients:
```

```
##                      Estimate Std. Error z value Pr(>|z|)
## Age                  0.006330   0.001036   6.109 1.00e-09 ***
## ControlPublic        0.689962   0.124635   5.536 3.10e-08 ***
## pop                 -0.003500   0.005058  -0.692  0.48889
## SpecialPurposeTRUE -0.917335   0.212076  -4.325 1.52e-05 ***
## LocaleRural/Town   -1.287182   0.125275 -10.275  < 2e-16 ***
## LocaleSmall_City   -0.315256   0.114195  -2.761  0.00577 **
## OffMainlandTRUE    -0.268238   0.288697  -0.929  0.35282
## is_religiousTRUE    0.036389   0.120336   0.302  0.76235
## at_capitalTRUE      0.172184   0.143178   1.203  0.22914
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Threshold coefficients:
##                    Estimate Std. Error z value
## Associate|Bachelor  -2.6259     0.2202 -11.927
## Bachelor|Master     -0.2230     0.1947  -1.145
## Master|Doctoral      1.5474     0.1985   7.794
## (4 observations deleted due to missingness)
```

Next, we create a model for the US News classification. The four regions for Regional Universities and Regional Colleges are consolidated into their respective categories. Our response, listed from best ranked to worse ranked, is the ordered factor consisting of the levels National, Liberal, Regional University, Regional College. This ordering is similar to that of the Carnegie Classification above: National corresponds to Doctoral, Regional University corresponds to Master, and Regional College corresponds to Bachelor and Associate. This information can be found in the US News College Rankings methodlogy.

An exception is the more subjective ordering of Liberal, which corresponds to the Carnegie Classification "Baccalaureate Colleges: Arts & Sciences Focus". This ordering is motivated by the author's understanding of the national prominence of each category. We believe that it is very likely that this is the same ordering of the number of views each ranking receives by prospective students, as well as the general excitement and interest each category generates upon update each year. It is thus a measure of social prestige. The results of our cumulative link model are shown below.

```
## formula:
## usnews_type ~ Age + Control + pop + SpecialPurpose + Locale + OffMainland + is_religi
## data:    data %>% as.data.frame()
##
##  link  threshold nobs logLik   AIC     niter max.grad cond.H
##  logit flexible  1627 -2012.90 4049.79 6(0)  1.48e-10 1.1e+06
##
## Coefficients:
##                     Estimate Std. Error z value Pr(>|z|)
## Age                 0.012286   0.001058  11.616  < 2e-16 ***
## ControlPublic       0.206969   0.122356   1.692  0.09073 .
## pop                 0.003902   0.005030   0.776  0.43787
## SpecialPurposeTRUE -0.722246   0.220278  -3.279  0.00104 **
## LocaleRural/Town   -1.380082   0.126176 -10.938  < 2e-16 ***
## LocaleSmall_City   -0.268263   0.112624  -2.382  0.01722 *
## OffMainlandTRUE    -0.757461   0.307495  -2.463  0.01377 *
## is_religiousTRUE   -0.273001   0.118821  -2.298  0.02159 *
## at_capitalTRUE      0.040956   0.142744   0.287  0.77417
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Threshold coefficients:
##                   Estimate Std. Error z value
## regcol|reguni      -0.4053     0.1966  -2.061
## reguni|liberal      1.4303     0.1992   7.181
## liberal|national    2.0913     0.2031  10.298
## (4 observations deleted due to missingness)
```

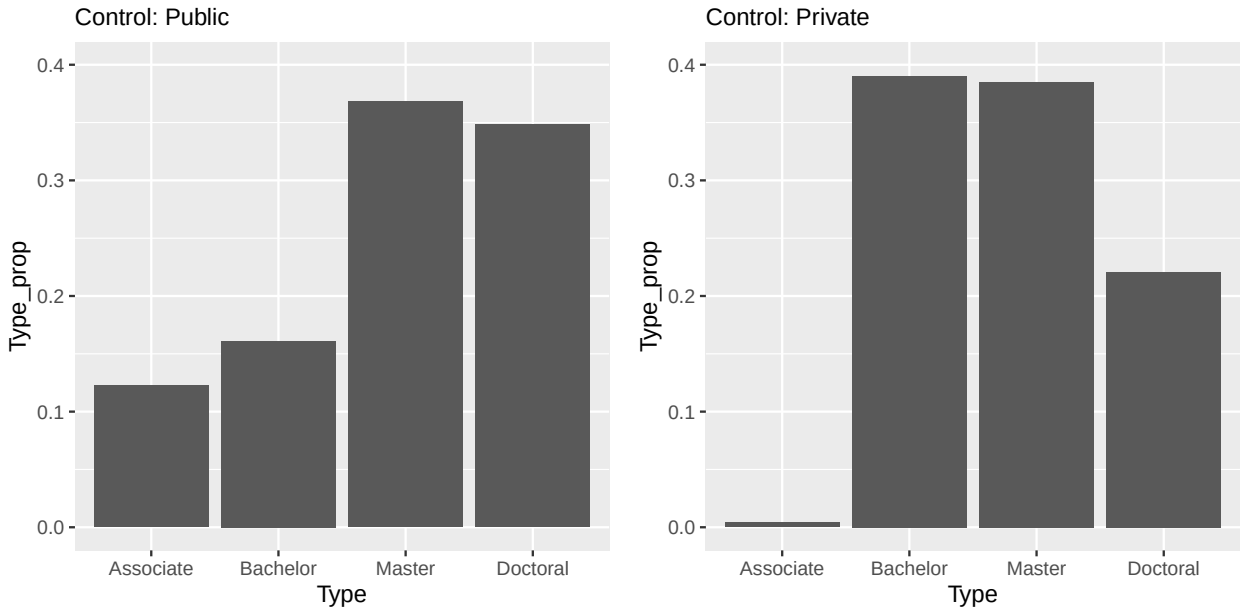
Control in Rankings vs School Type Yet for the Carnegie Classification model, this Control=Public is significant and positively correlated with a higher level classification. This is an unexpected reversal of the effect of Control = Public.

We examine the reasons why this occurs.

We examine the link between Control and Carnegie Classification type in the combined data set. We can see in the first plot that Public schools have a higher proportion of Doctoral schools, a similar proportion of Master schools, and a lower proportion of Master schools than Private schools. One can hypothesize that Public schools are more likely to receive significant government funding for research activities, or that many private individuals/organizations use limited resources to create small-scale Bachelor schools.

```
## # A tibble: 8 x 3
##   Control Type          n
##   <fct>   <ord>     <int>
## 1 Private Associate     4
## 2 Private Bachelor    382
## 3 Private Master      377
## 4 Private Doctoral    216
## 5 Public  Associate    80
```

```
## 6 Public  Bachelor    105
## 7 Public  Master      240
## 8 Public  Doctoral    227
```



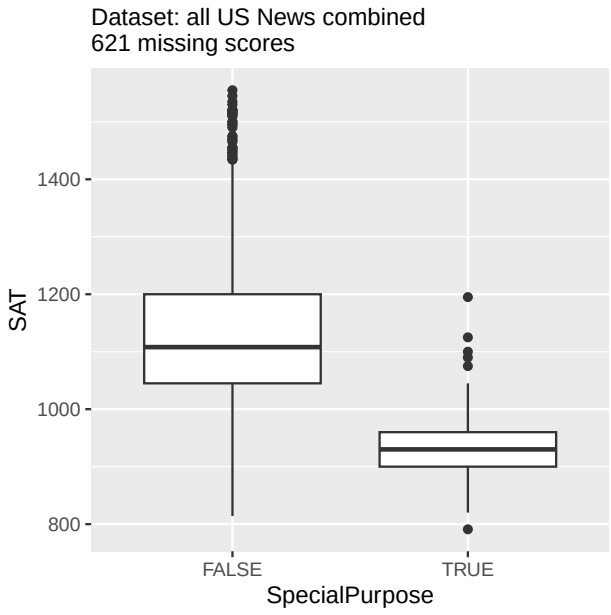
Special Purpose Analysis We would like to analyze the variable SpecialPurpose more closely. We notice that SpecialPurpose has a very consistent and significant negative correlation with rank. Note that when we say a variable is "negatively correlated" with rank or has a "negative effect" on rank, we mean that a higher value or a True value of the variable is correlated with worse ranks, and vice versa for the terms "positively correlated" and "positive effect". As mentioned earlier, SpecialPurpose is True if the school is an HBCU, or tribal college, or men's only or women's only college. The counts are as follows.

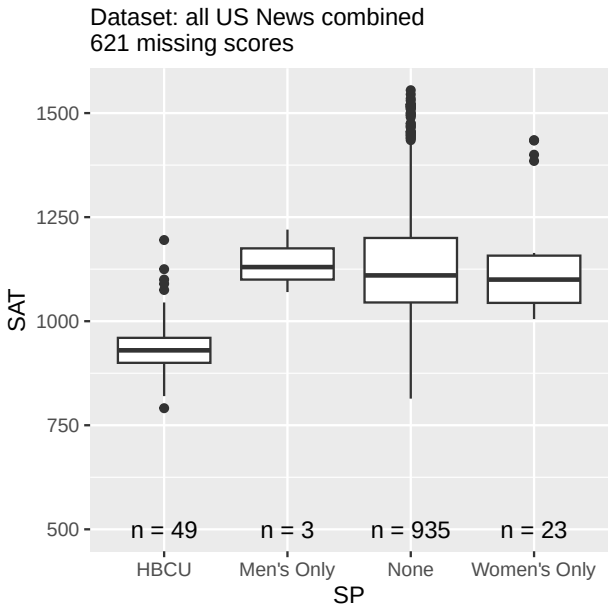
```
## # A tibble: 10 x 5
##    list_name      HBCU_n Womens_n Mens_n Tribal_n
##    <chr>           <dbl>    <dbl>  <dbl>    <dbl>
##  1 national           14        4      0        0
##  2 liberal            16       15      5        0
```

```
##  3 regcol_north        0        0      0        0
##  4 regcol_south       18        1      0        0
##  5 regcol_west         6        0      0        0
##  6 regcol_midwest      4        1      0        0
##  7 reguni_north        5        4      0        0
##  8 reguni_south       17        1      0        0
##  9 reguni_west         1        1      0        0
## 10 reguni_midwest      0        5      0        0
```

We know from SAT demographic information that African American students have lower average test scores. Thus, we would like to investigate whether the significance and negative effect of SpecialPurpose is simply due to HBCU being a proxy for lower SAT scores.

We see that the SpecialPurpose=True subset has a significantly lower average SAT score than the non-Special Purpose subset, with HBCU schools having significantly lower average SAT scores, and gender exclusive schools having similar average SAT scores compared to non-Special Purpose schools. This matches the general demographic patterns reported by Collegeboard.





We now run the cumulative link model on rank again, this time controlling for SAT score. Normally we would like to see whether factors associated with prestige, which is tightly linked with student academic preparedness (SAT scores). However, since we introduce SAT scores into the model, the goal of this particular model is no longer to predict and find correlation between root factors and school prestige. We now simply wish to see whether SpecialPurpose is still significant or still has a negative association with rank, even after controlling for the fact that they have lower average SAT scores.

```
## Removing OffMainland from liberal
## Removing is_religious from regcol_north
## Removing at_capital from regcol_north
## Removing SpecialPurpose from regcol_west
## Removing SpecialPurpose from regcol_midwest
```

We see that for every dataset in which there are sufficient counts for both SAT and SpecialPurpose, the correlation of SpecialPurpose has reversed. We find that, controlling for SAT score, SpecialPurpose's correlation implies that it has a positive effect on rank.

Notably, we find that for National University, Regional College South, Regional University South, Regional University West, and Regional University Midwest, Age is still significant and positively correlated, even after controlling for SAT score. This means that even controlling for the fact that older schools might attract students with higher academic ability, older schools still have an advantage over younger ones.

Age/Control Interaction Analysis One possible interaction we would like to investigate is interaction between Age and Control. Specifically, we hypothesize that private colleges receive a greater benefit from an older Age than public colleges do. We fit the base model with Age/Control interaction on all datasets.

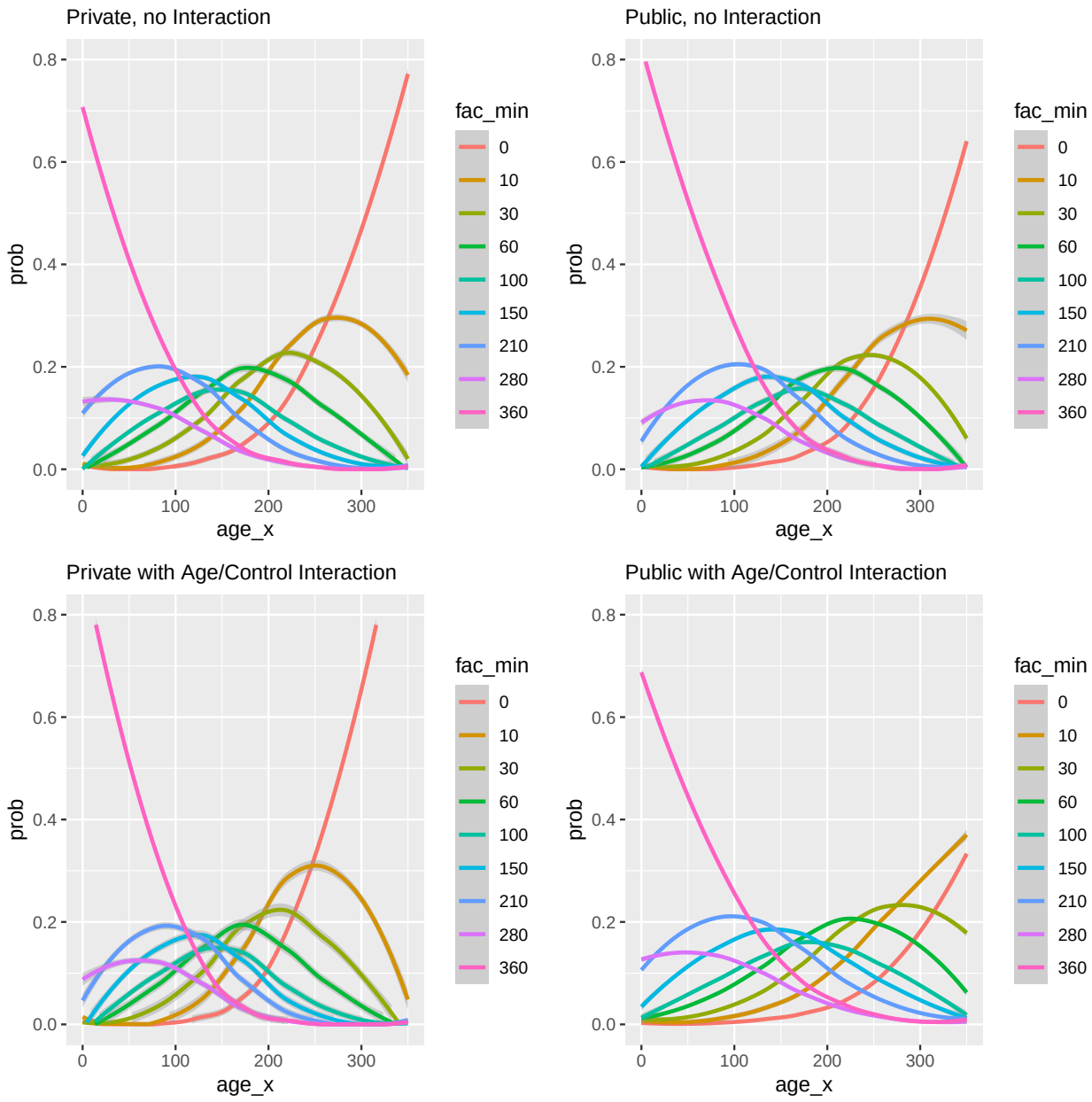
```
## Removing OffMainland from liberal
## Removing is_religious from regcol_north
## Removing at_capital from regcol_north
## Removing SpecialPurpose from regcol_midwest
## Removing SpecialPurpose from reguni_west
```

We find that Age/Control interaction is significant for National University, Liberal Arts, Regional College Midwest, and is close to our significance cutoff for Regional University West.

We see that the general pattern is what we hypothesized: Public schools reap less benefit from a higher Age than Private schools do. However, we see that after accounting for the interaction between Age and Control, the effect of Control is flipped. Whereas before, Control=Public had a negative effect on rank, it now has a positive effect. (Except for Regional University West, where the Control=Public coefficient already had a positive effect, but it doubled in magnitude). So, for National Universities, initially Public schools are stronger, but Private schools overtake them in 85 years in the model.

The Liberal Arts dataset, however, has a total reversal of this trend. The regression coefficients imply that Public schools reap larger benefits from older Age, but the effect of Control=Public stays negative, as it was before the interaction was added. One thing to note, however, is that the effect of Control=Public has tripled in magnitude. So, for Liberal Arts, initially Private schools are stronger, but Public schools overtake them in 175 in the model.

We illustrate the difference with and without interaction for the National Universities dataset. The plot below shows the Private vs Public model without interaction. The Private schools have higher predicted probabilities for the higher ranks and lower predicted probabilities for the lower ranks at every level of Age.



Perhaps this relationship only exists because of the influence of extremely old, Private, very well-ranked schools, such as Harvard and Princeton. However, we find that this relationship still holds even when not considering the extremely old schools in the dataset.

```
## formula:
## Rank_fac ~ Age + Control + pop + SpecialPurpose + Locale + OffMainland + is_religious
## data:    data %>% as.data.frame()
```

```
##
##  link  threshold nobs logLik  AIC     niter max.grad cond.H
##  logit flexible  413  -755.86 1547.72 6(0)  9.77e-12 6.9e+06
##
## Coefficients:
##                     Estimate Std. Error z value Pr(>|z|)
## Age                 0.026270   0.003497   7.512 5.82e-14 ***
## ControlPublic       0.720751   0.628380   1.147 0.251382
## pop                 0.040204   0.009318   4.314 1.60e-05 ***
## SpecialPurposeTRUE -1.991697   0.533040  -3.736 0.000187 ***
## LocaleRural/Town   -0.991882   0.305214  -3.250 0.001155 **
## LocaleSmall_City   -0.409139   0.201190  -2.034 0.041993 *
## OffMainlandTRUE    -0.225432   0.690119  -0.327 0.743927
## is_religiousTRUE   -0.914988   0.264112  -3.464 0.000531 ***
## at_capitalTRUE      0.422385   0.252107   1.675 0.093852 .
## Age:ControlPublic  -0.009388   0.004758  -1.973 0.048498 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Threshold coefficients:
##                     Estimate Std. Error z value
## (360,450]|(280,360]   1.6914     0.5000   3.382
## (280,360]|(210,280]   2.2923     0.5024   4.563
## (210,280]|(150,210]   3.1766     0.5116   6.209
## (150,210]|(100,150]   3.9326     0.5239   7.507
## (100,150]|(60,100]    4.5627     0.5354   8.522
## (60,100]|(30,60]      5.4148     0.5559   9.741
```

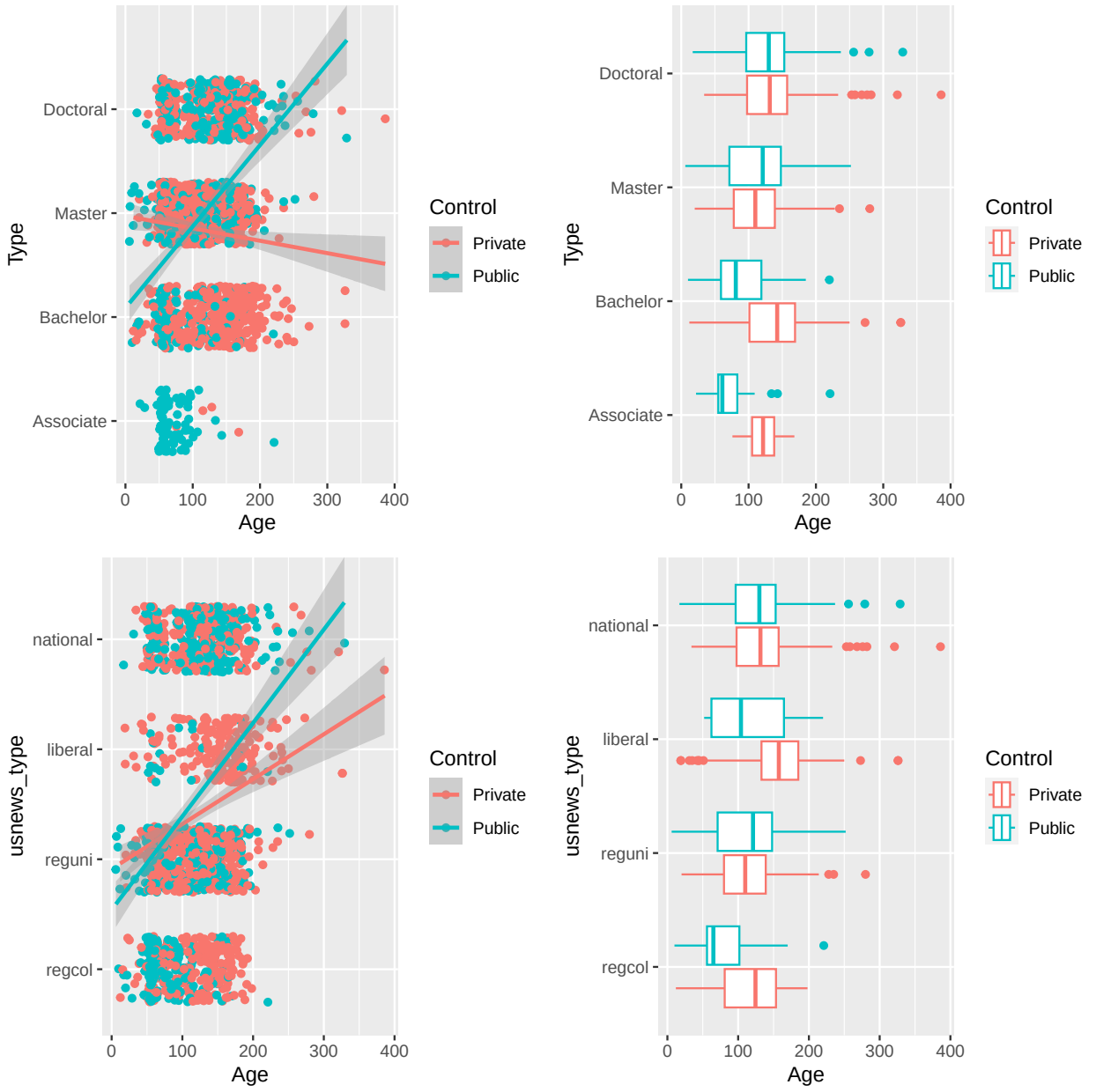
```
## (30,60]|(10,30]       6.4830     0.5938  10.917
## (10,30]|(0,10]        7.6320     0.6715  11.366
## (3 observations deleted due to missingness)
```

When testing the Type and US News Type datasets for Age/Control interaction, we find something surprising.

Previously, we found that both Age and Control=Public were significant and had positive correlations with Type. We now find that the effect of Age for Private schools ceases to be significant, with an estimated coefficient close to zero. Instead, Public schools start off as more likely to be a lower-level Type, but they overtake Private schools after 82 years.

We try to create an interpretation for these results. The small coefficient of Age for Private schools could be interpreted as Private schools don't "accumulate" and build up towards Doctoral level status, or that the level of recent Private schools is similar to the levels of older Private schools. If older Private schools do "accumulate" or develop, then newer Private schools are somehow better equipped and getting to that level in a much shorter time. The first interpretation seems more likely, because Public schools overtake their chance of being a higher-level type fairly soon, so it is less likely that the Type levels of old Private schools are extremely high. In addition, a possible interpretation for the Age coefficient of Public schools is that Public schools do "accumulate", and that perhaps government bodies build upon their oldest and existing schools to create schools with strong research and education capabilities.

We can see a rough visualization of this phenomenon below (of course, this fit line of graph makes the simplifying assumption that the distance between each category is equal).



We notice that the majority of Associate schools are young, Public schools. Perhaps the interpretation for this currently is that as Age gets younger, Public is worse due to the influence of young, public Associate schools. We run the regression without Associate schools to see if this interaction between Age and Control still holds.

```
## formula:
## Type ~ Age + Control + pop + SpecialPurpose + Locale + OffMainland + is_religious + a
```

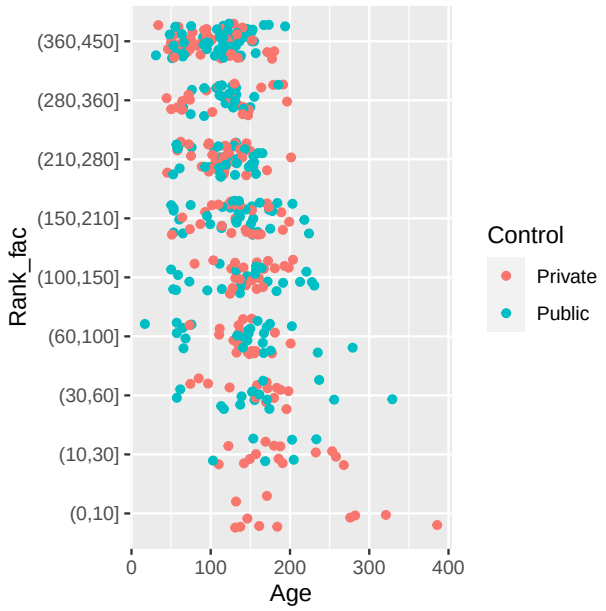
```
## data:    data %>% as.data.frame()
##
##  link  threshold nobs logLik   AIC     niter max.grad cond.H
##  logit flexible  1544 -1512.04 3048.08 5(0)  1.03e-11 1.2e+06
##
## Coefficients:
##                     Estimate Std. Error z value Pr(>|z|)
## Age                -0.001619   0.001359  -1.191   0.2336
## ControlPublic      -0.453100   0.286496  -1.582   0.1138
## pop                 0.005673   0.005381   1.054   0.2917
## SpecialPurposeTRUE -1.456049   0.241733  -6.023 1.71e-09 ***
## LocaleRural/Town   -1.503878   0.135506 -11.098  < 2e-16 ***
## LocaleSmall_City   -0.258724   0.120147  -2.153   0.0313 *
## OffMainlandTRUE    -0.608226   0.304154  -2.000   0.0455 *
## is_religiousTRUE    0.049330   0.128191   0.385   0.7004
## at_capitalTRUE      0.265867   0.151592   1.754   0.0795 .
## Age:ControlPublic   0.014851   0.002240   6.628 3.39e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Threshold coefficients:
##                 Estimate Std. Error z value
## Bachelor|Master  -1.1445     0.2282  -5.015
## Master|Doctoral   0.8492     0.2276   3.730
## (3 observations deleted due to missingness)

## Likelihood ratio tests of cumulative link models:
##
```

```
##                               formula:      link: threshold:
## type_noAssociate              model_formula logit flexible
## type_AgeControl_noAssociate model_formula logit flexible
##
##                               no.par    AIC logLik LR.stat df Pr(>Chisq)
## type_noAssociate                  11 3091.9  -1535
## type_AgeControl_noAssociate       12 3048.1  -1512  45.862  1  1.269e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

We find that the Age/Control interaction still holds the same pattern and is still significant. Thus, the interaction was not just the result of these young, public Associate schools.

For the US News Type dataset, we see that Age/Control interaction is significant as well. We find a similar pattern. Public schools start out with higher probabilities for lower-level types, but overtake Private schools in 96 years.

We would like to further investigate the nature of this Age/Control interaction. Consider the graph below. We can see that in the area at which Age is greater than 200 years, there is a clear divide between better ranked Private schools and worse ranked Public schools. We perform a regression on data, splitting the data between schools with Age greater than or less than 200 years, in order to examine whether these two subsets of the data behave by different rules.




```
## Regression on National University for Age < 200

## formula:
## Rank_fac ~ Age + Control + pop + SpecialPurpose + Locale + OffMainland + is_religious
## data:    data %>% as.data.frame()
##
##  link  threshold nobs logLik  AIC     niter max.grad cond.H
##  logit flexible  413  -757.81 1549.62 6(0)  2.92e-12 3.4e+06
##
## Coefficients:
##                      Estimate Std. Error z value Pr(>|z|)
## Age                  0.021852   0.002635   8.293  < 2e-16 ***
## ControlPublic       -0.427824   0.238154  -1.796 0.072428 .
## pop                  0.041876   0.009309   4.498 6.85e-06 ***
## SpecialPurposeTRUE -2.033386   0.533856  -3.809 0.000140 ***
## LocaleRural/Town   -1.033818   0.303704  -3.404 0.000664 ***
## LocaleSmall_City   -0.443150   0.200232  -2.213 0.026885 *
```

```
## OffMainlandTRUE    -0.241047   0.691020  -0.349 0.727219
## is_religiousTRUE   -0.913499   0.262770  -3.476 0.000508 ***
## at_capitalTRUE      0.403746   0.251258   1.607 0.108077
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Threshold coefficients:
##                     Estimate Std. Error z value
## (360,450]|(280,360]   1.1417     0.4101   2.784
## (280,360]|(210,280]   1.7429     0.4127   4.223
## (210,280]|(150,210]   2.6259     0.4229   6.209
## (150,210]|(100,150]   3.3782     0.4362   7.744
## (100,150]|(60,100]    4.0054     0.4489   8.922
## (60,100]|(30,60]      4.8509     0.4704  10.312
## (30,60]|(10,30]       5.9046     0.5091  11.598
## (10,30]|(0,10]        7.0470     0.5958  11.829
## (3 observations deleted due to missingness)

## Regression on National University for Age >= 200

## formula: Rank_fac ~ Age + Control + pop + Locale + is_religious + at_capital
## data:    data %>% as.data.frame()
##
##  link  threshold nobs logLik AIC   niter max.grad cond.H
##  logit flexible  27   -36.25 98.50 6(0)  4.87e-07 4.4e+07
##
## Coefficients:
##                  Estimate Std. Error z value Pr(>|z|)
```

```
## Age                0.03325    0.01508   2.205   0.0274 *
## ControlPublic     -3.22911    1.37062  -2.356   0.0185 *
## pop                0.08751    0.08619   1.015   0.3099
## LocaleRural/Town   0.31478    1.20927   0.260   0.7946
## LocaleSmall_City   2.33940    0.95807   2.442   0.0146 *
## is_religiousTRUE  -0.07556    1.71750  -0.044   0.9649
## at_capitalTRUE     0.20770    1.25059   0.166   0.8681
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Threshold coefficients:
##                       Estimate Std. Error z value
## (210,280]|(150,210]    3.012      4.074   0.739
## (150,210]|(100,150]    4.859      4.023   1.208
## (100,150]|(60,100]     6.466      4.040   1.600
## (60,100]|(30,60]       7.580      4.105   1.846
## (30,60]|(10,30]        8.463      4.168   2.031
## (10,30]|(0,10]        11.557      4.525   2.554
```

We find that the negative correlation of Control=Public that we found in the base model is no longer significant in the Age >= 200 subset, but is even more significant and strong in the Age < 200 subset. We also find that in the Age >= 200 subset, population is no longer significant, and neither is religious status.

In addition, whereas the Age < 200 subset displays a significant and clear ordering that Large City is better than Small City, which is better than Rural/Town, the Age >= 200 subset finds that Large City and Rural/Town don't have a significant difference, but Small City is significantly better. A possible interpretation of the change in association of Locale is

the that oldest schools benefit from the image of a small-town, high-status college campus, while the rest of the schools follow the traditional hierarchy of favoring more resources, opportunities, etc.

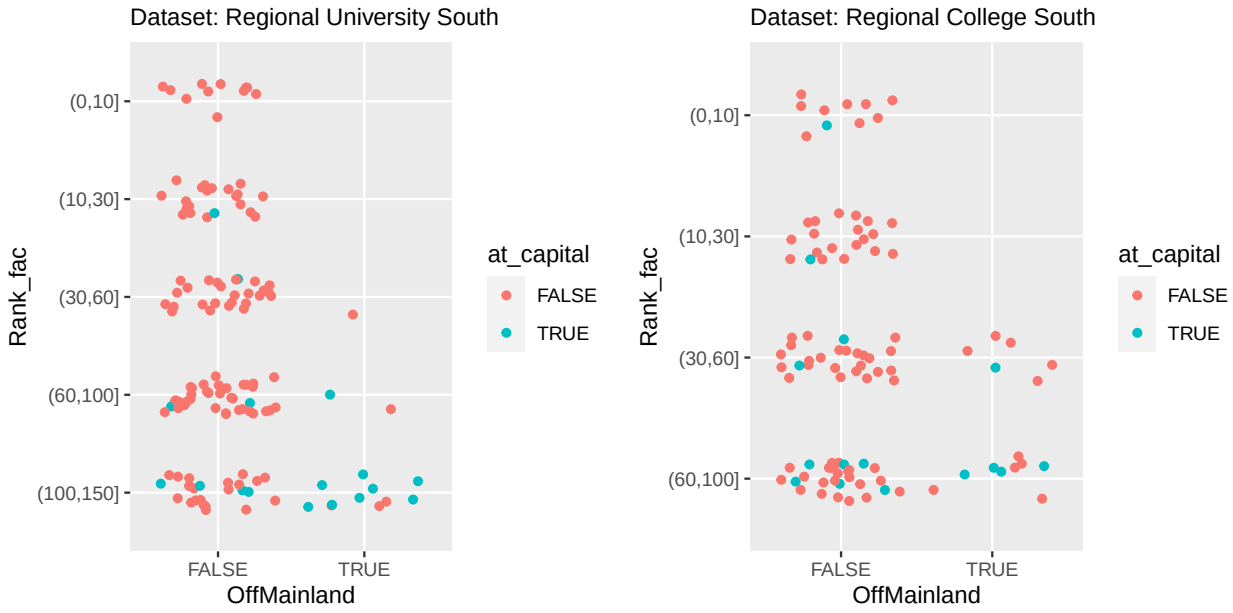
We see that Control = Public ceases to become significant in the Age <= 200 subset, just as it is in all other datasets. However, it becomes even more significant in the Age > 200 subset. Thus, we hypothesize that Control has a strong interaction with Age. Specifically, Control becomes more relevant as Age increases, in the direction of Control = Positive having a negative effect on rank.

Control/At Capital Interaction Analysis We also wish to test our hypothesis that Public schools benefit more from being at the capital than do Private schools. We add the interaction effect between Control and at_capital to the base model, and fit to all datasets. We find that this interaction is not significant for all datasets. In fact, under this model at_capital itself is only significant for one of these datasets, Regional University North, suggesting that not only is there no difference in the benefit of being at the capital between Private and Public schools, but there is not significant evidence of a benefit at all.

```
## Removing OffMainland from liberal
## Removing SpecialPurpose from liberal
## Removing Control from liberal
## Removing is_religious from regcol_north
## Removing at_capital from regcol_north
## Removing SpecialPurpose from regcol_midwest
## Removing SpecialPurpose from reguni_west
```

In the model without interaction, at_capital is significant for the National University, Regional University North, and Regional University South. The variable has a positive corre-

lation with rank for the first of these two datasets, but Regional University South reverses this trend with a negative correlation with rank. In fact, the coefficients follow a pattern: all the datasets have at_capital positive correlated with rank, except for the Southern region schools. One thing to note about the Southern datasets is that they include the island territory schools. Regional University South contains 13 Puerto Rican schools (and one from the Virgin Islands), and Regional College South contains 15 Puerto Rican schools. Additionally, note that these are the only schools in their dataset for which OffMainland is true, and that these OffMainland schools are disproportionately at the capital (likely due to their smaller area or perhaps the capital is much more developed than the other areas). For Regional University South, 9/14 = 64% of OffMainland schools are at the capital, while 17/136 = 13% of schools in the dataset are at the capital. Similarly, for Regional College South, 5/15 = 33% of OffMainland schools are at the capital, while 17/132 = 13% of schools in the dataset are at the capital. Since, these off mainland schools are worse ranked than average, this might explain the negative correlation with being at the capital.



Thus, there is possible confounding between being off the mainland and being at the capital, since territories have smaller areas. We then seek to find whether at_capital does have a negative correlation with rank for the Southern datasets, or if it is due to confounding with

being off the mainland. After we run regressions for the two datasets on only schools that are on the mainland, and on only schools that are off the mainland, we find that at_capital is still negatively correlated with rank.

```
## Regional University South, off mainland

## formula: Rank_fac ~ at_capital + Age
## data:    reguni_south %>% filter(OffMainland == T)
##
##  link  threshold nobs logLik AIC   niter max.grad cond.H
##  logit flexible  14   -6.32  20.65 6(0)  3.91e-13 7.0e+05
##
## Coefficients:
##                Estimate Std. Error z value Pr(>|z|)
## at_capitalTRUE -0.23716    1.84989  -0.128   0.8980
## Age             0.08784    0.05139   1.709   0.0874 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Threshold coefficients:
##                    Estimate Std. Error z value
## (100,150]|(60,100]    7.976      4.761   1.675
## (60,100]|(30,60]     10.209      5.419   1.884

## Regional University South, on mainland

## formula:
## Rank_fac ~ Age + Control + pop + SpecialPurpose + Locale + is_religious + at_capital
```

```
## data:    data %>% as.data.frame()
##
##  link  threshold nobs logLik  AIC    niter max.grad cond.H
##  logit flexible  121  -174.53 373.06 5(0)  4.08e-11 2.2e+06
##
## Coefficients:
##                     Estimate Std. Error z value Pr(>|z|)
## Age                 0.015013   0.004470   3.359 0.000783 ***
## ControlPublic      -0.072734   0.583868  -0.125 0.900862
## pop                 0.050797   0.039481   1.287 0.198224
## SpecialPurposeTRUE -1.344331   0.564356  -2.382 0.017216 *
## LocaleRural/Town   -0.849613   0.508842  -1.670 0.094979 .
## LocaleSmall_City   -0.007951   0.485357  -0.016 0.986930
## is_religiousTRUE   -0.232065   0.564289  -0.411 0.680887
## at_capitalTRUE     -1.397479   0.714969  -1.955 0.050630 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Threshold coefficients:
##                     Estimate Std. Error z value
## (100,150]|(60,100]  -0.1970     0.9471  -0.208
## (60,100]|(30,60]     1.4346     0.9448   1.518
## (30,60]|(10,30]      2.4868     0.9632   2.582
## (10,30]|(0,10]       3.9829     1.0263   3.881

## Regional College South, off mainland

## Removing Rural/Town
```

```
## formula: Rank_fac ~ Age + Control + Locale + at_capital
## data:    data %>% as.data.frame()
##
##  link  threshold nobs logLik AIC   niter max.grad cond.H
##  logit flexible  14   -5.74  21.48 8(0)  3.12e-09 1.7e+06
##
## Coefficients:
##                  Estimate Std. Error z value Pr(>|z|)
## Age               -0.1232     0.2660  -0.463    0.643
## ControlPublic      1.1542     2.2500   0.513    0.608
## LocaleSmall_City  -2.0027     1.6159  -1.239    0.215
## at_capitalTRUE    -2.9331     2.4191  -1.212    0.225
##
## Threshold coefficients:
##                  Estimate Std. Error z value
## (60,100]|(30,60]   -8.311     16.533  -0.503

## Regional College South, on mainland

## formula:
## Rank_fac ~ Age + Control + pop + SpecialPurpose + Locale + is_religious + at_capital
## data:    data %>% as.data.frame()
##
##  link  threshold nobs logLik AIC    niter max.grad cond.H
##  logit flexible  84   -97.73 217.46 5(0)  8.22e-12 2.2e+06
##
## Coefficients:
##                     Estimate Std. Error z value Pr(>|z|)
```

```
## Age                 0.017157   0.006268   2.737 0.006195 **
## ControlPublic      -0.587409   0.851645  -0.690 0.490362
## pop                 0.038628   0.044877   0.861 0.389381
## SpecialPurposeTRUE -2.143438   0.639766  -3.350 0.000807 ***
## LocaleRural/Town    0.848580   0.736234   1.153 0.249076
## LocaleSmall_City    1.286481   0.734679   1.751 0.079932 .
## is_religiousTRUE   -0.193396   0.733231  -0.264 0.791966
## at_capitalTRUE     -0.357775   0.776457  -0.461 0.644958
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Threshold coefficients:
##                   Estimate Std. Error z value
## (60,100]|(30,60]    1.556      1.282   1.213
## (30,60]|(10,30]     3.202      1.311   2.442
## (10,30]|(0,10]      4.949      1.388   3.566
```

We wish to pursue this phenomenon further. We take the subset of national schools that are in the South (note: this is Florida, South Carolina, North Carolina, Georgia, Virginia, West Virginia, Alabama, Tennessee, Kentucky, Louisiana, Arkansas, Missouri, Puerto Rico, the Virgin Islands). We find that, whereas at_capital was previously positively correlated with rank, it is now negatively correlated, albeit not significant.

This trend, however, does not continue with the Liberal Arts dataset. After subsetting the Southern schools and performing a regression, we find that at_capital is positively correlated with rank and not significant, as it was before the subset for Liberal Arts.

```
## formula:
## Rank_fac ~ Age + Control + pop + SpecialPurpose + Locale + OffMainland + is_religious
```

```
## data:    data %>% as.data.frame()
##
##  link  threshold nobs logLik  AIC    niter max.grad cond.H
##  logit flexible  120  -194.43 422.86 6(0)  3.95e-10 4.8e+06
##
## Coefficients:
##                     Estimate Std. Error z value Pr(>|z|)
## Age                 0.025552   0.004504   5.673  1.4e-08 ***
## ControlPublic      -0.256943   0.509744  -0.504 0.614218
## pop                 0.114525   0.033721   3.396 0.000683 ***
## SpecialPurposeTRUE -1.828412   0.697612  -2.621 0.008768 **
## LocaleRural/Town   -1.570860   0.604560  -2.598 0.009367 **
## LocaleSmall_City   -1.060689   0.406646  -2.608 0.009097 **
## OffMainlandTRUE     0.294151   1.066982   0.276 0.782790
## is_religiousTRUE   -1.338674   0.599482  -2.233 0.025545 *
## at_capitalTRUE     -0.161841   0.468991  -0.345 0.730032
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Threshold coefficients:
##                     Estimate Std. Error z value
## (360,450]|(280,360]   1.8481     0.8494   2.176
## (280,360]|(210,280]   2.6015     0.8591   3.028
## (210,280]|(150,210]   3.4860     0.8874   3.928
## (150,210]|(100,150]   4.3964     0.9287   4.734
## (100,150]|(60,100]    5.0265     0.9602   5.235
## (60,100]|(30,60]      5.6774     1.0042   5.653
```

```
## (30,60]|(10,30]       6.5968     1.0808   6.104
## (10,30]|(0,10]        8.7122     1.4370   6.063

## Removing OffMainland

## formula:
## Rank_fac ~ Age + Control + pop + SpecialPurpose + Locale + is_religious + at_capital
## data:    data %>% as.data.frame()
##
##  link  threshold nobs logLik AIC    niter max.grad cond.H
##  logit flexible  59   -63.71 151.41 5(0)  2.16e-09 4.7e+06
##
## Coefficients:
##                     Estimate Std. Error z value Pr(>|z|)
## Age                 0.019703   0.006409   3.074  0.00211 **
## ControlPublic       0.399086   0.980646   0.407  0.68404
## pop                 0.124199   0.070575   1.760  0.07844 .
## SpecialPurposeTRUE -2.900979   0.998234  -2.906  0.00366 **
## LocaleRural/Town    0.332964   0.763761   0.436  0.66287
## LocaleSmall_City    0.151230   0.672040   0.225  0.82195
## is_religiousTRUE   -0.773467   0.647613  -1.194  0.23235
## at_capitalTRUE      1.538764   0.860678   1.788  0.07380 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Threshold coefficients:
##                     Estimate Std. Error z value
## (150,210]|(100,150]    3.302      1.676   1.970
```

```
## (100,150]|(60,100]     4.990      1.767   2.824
## (60,100]|(30,60]       5.660      1.800   3.144
## (30,60]|(10,30]        6.728      1.869   3.600
## (1 observation deleted due to missingness)
```

Analysis of Public/Private and being at the capital for other countries would be an interesting area for further research. We note that in Asian, the top universities (Beijing U, Tokyo U, Seoul U) are public and located in the capital.

Outliers We now examine the schools which have the highest residual under the base model with Age/Control interaction, in order to see if there any are patterns we can find in the type of school that our model does not do well on.

School Rank_fac pred Age Control pop is_religious at_capital OffMainland SpecialPurpose Locale

Upon manual inspection, we can identify certain types of schools that the model fails to explain. One group is a set of very young schools that are part of a strong public university system, and thus perform much better than expected given their age. These include UC Santa Cruz and UC Merced, as well as SUNY Stony Brook and SUNY Binghamton. Since these are two of the highest population states in the country, perhaps a larger state means a stronger public school system. We test this by testing for interaction between population and control.

```
## Likelihood ratio tests of cumulative link models:
##
##                     formula:      link: threshold:
```

```
## nat_clm             model_formula logit flexible
## nat_clm_pop_control model_formula logit flexible
##
##                     no.par    AIC  logLik LR.stat df Pr(>Chisq)
## nat_clm                 17 1642.4 -804.19
## nat_clm_pop_control     18 1643.8 -803.89  0.6014  1      0.438
```

Another group is the set of schools that are religiously affiliated and high-ranked. This includes Pepperdine University, American University, and Gonzaga University. Perhaps there is interaction between Age and religious status, wherein among older schools, the correlation between being religious and lower ranks is lessened. However, we find that there is not good evidence for this possibility.

```
## Likelihood ratio tests of cumulative link models:
##
##                 formula:      link: threshold:
## nat_clm         model_formula logit flexible
## nat_clm_age_rel model_formula logit flexible
##
##                 no.par    AIC  logLik LR.stat df Pr(>Chisq)
## nat_clm             17 1642.4 -804.19
## nat_clm_age_rel     18 1644.4 -804.19  0.0151  1     0.9023
```

An area of possible future research would be investigating why some schools seem to so strongly buck the trend of being religious correlating with a drastic decrease in ranking.

Conclusions

Age is consistently a very significant and positive predictor of college ranking. It has a negative coefficient and a p-value less than or equal to 0.01 for all datasets except for Regional Colleges Midwest. Additionally, it is a significant and positive predictor of college type. However, we note the interesting phenomenon in which private schools vary little in school type as age changes. We thus conclude Age to be the single most important factor correlating with the prestige of a school. What can a hypothetical person who wishes to found a highly ranked university learn from this? This person will be at a severe disadvantage. The best strategy seems to be having started 400 years ago. The author hypothesizes that this applies to other institutions in which prestige and social standing play a large factor as well. As the proverb says: "The best time to plant a tree was 20 years ago. The second best time is now."

It is possible that more successful schools have have a smaller chance of shutting down, and thus prestige does in fact feed into the factor of age (i.e. survivor bias). Investigating school open/closed status as an outcome would be an interesting line of future research to address this consideration. However, if the only interaction between age and prestige was that schools below a certain threshold of some measure of success shut down over time, then we would expect to only see that older ages have less worse ranked schools. However, we see that not only do the older ages have less worst ranked schools, but they also disproportionately contain the outlier high ranked schools. Thus, while we must consider the fact that prestige could affect age, we can be reasonably sure that it is not the sole explanation of the effects of age we have found.

We find that at_capital is significant and has a positive effect on rank for our most prominent dataset, National University. We also find that there is an interesting trend in which being at the capital in a Southern state can have a negative correlation. Overall, however, at_capital it is not a consistently important factor.

We find that for all datasets except National University, Control varies in direction and is not significant. Control = Public is significant and negatively correlated with rank in the National University dataset, but we find that almost all of this correlation is contained in the National University subset where Age >= 200. That is, for school rankings, Public vs Private seems not to matter, except for the very oldest of universities.

Yet for the Carnegie Classification model, this Control=Public is significant and positively correlated with a higher level classification. This is an unexpected reversal of the effect of Control = Public.

Religious status is extremely significant and negatively correlated with rank for the two most prominent datasets, National University and Liberal Arts. It is also significant and negatively correlated with US News type. One possible consideration is: do the best universities become secular as they grow in reputation, while the less prestigious universities stay religious? In other words, does ranking cause religious status?

We see that for the National University dataset, Locale=Rural (the comparison level is Locale=LargeCity) has a very significant, negative correlation with rank. The only other dataset with this is Regional University West. LocaleSmallCity is insignificant for most of the datasets, suggesting that the difference between SmallCity and LargeCity is not incredibly big. LocaleRural is a good predictor for the national dataset, but otherwise is not very correlated with rank.

However, Locale is extremely important in regards to Carnegie Classification and US News type. For both of these models, they exhibit the pattern that LargeCity is better than SmallCity, and Small City is better than Rural, and both the SmallCity and Rural coefficients are extremely significant.

We may hypothesize that once a school reaches a certain classification or type, the locale makes little difference to their ranking or prestige. But the locale is extremely important in getting to that classification or type in the first place. Perhaps schools created in more

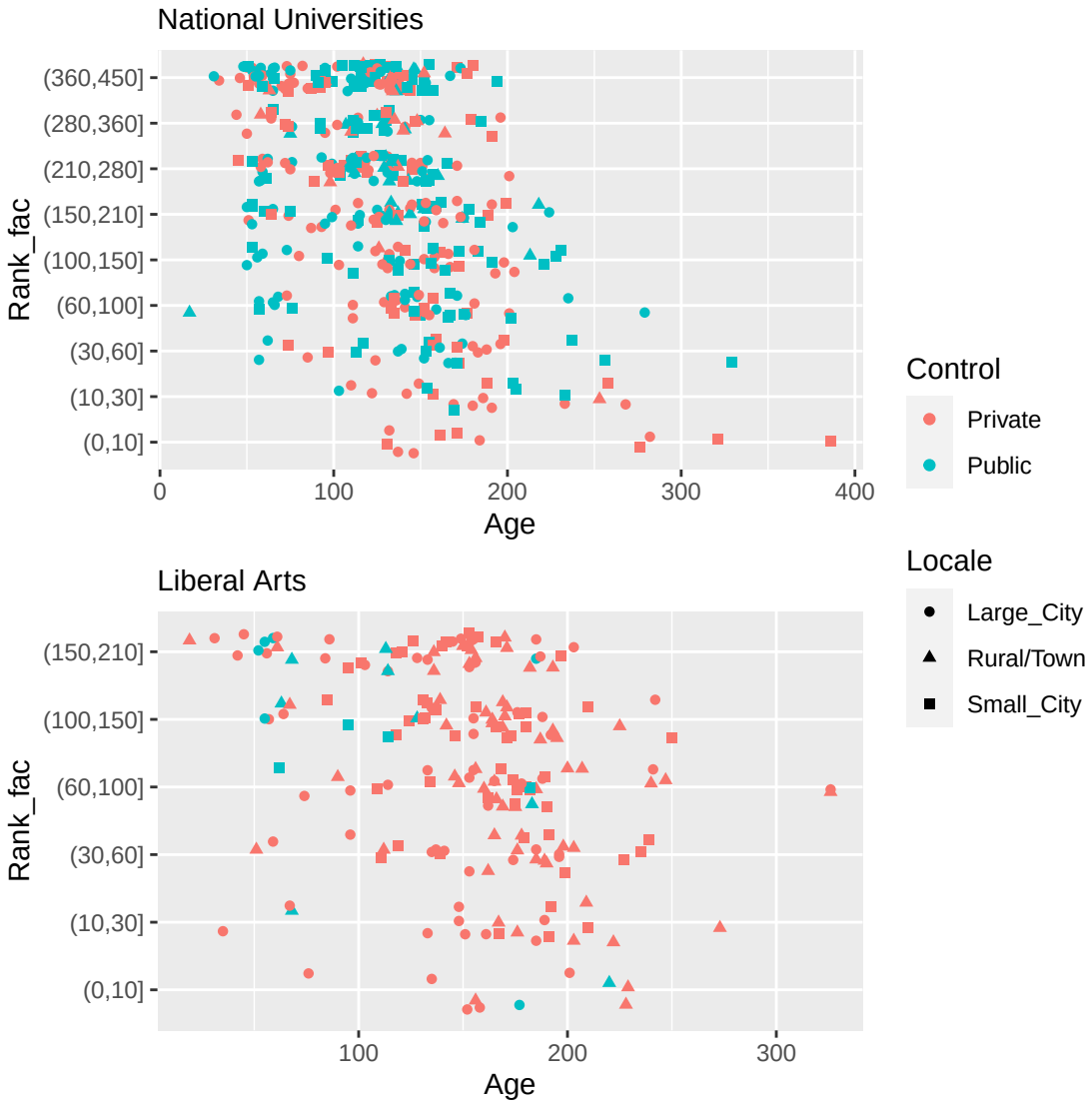
urban areas receive larger financial and structural support due to people in urban areas having more resources.

We find that OffMainland is significant only for Regional University South, where it is negatively correlated with rank, and for the US News type regression, where it is negatively correlated with rank. This stems from the presence of Puerto Rican schools in this dataset, as they are the only schools off the mainland, and they have much lower ranks than average. For all other datasets, we observe it is either not significant, or there are not enough observations.

SpecialPurpose is a consistent factor for almost all datasets that have sufficient observations. In these datasets, we observe that SpecialPurpose is negatively correlated with the response. This might indicate that it might be best to not give your school a special mission. However, we also observed that SpecialPurpose schools actually performed better when taking into account SAT scores. Thus, having a special mission might actually be an advantageous trait that sets you apart from the competition. This would contradict the possible narrative that HBCU or women's colleges are ranked unfairly low.

We observe that state population is significant and positively correlated with rank for our most prominent dataset, National University. We found that this significance was contained only in the subset Age < 200. Although it has similar effects for a few other datasets, it is not an overall very consistent variable.

We illustrate the relationship we have found between the most important variables below.


National Universities
Rank_fac
(360,450]
(280,360]
(210,280]
(150,210]
(100,150]
(60,100]
(30,60]
(10,30]
(0,10]
0
100
200
300
400
Age
Liberal Arts
Rank_fac
(150,210]
(100,150]
(60,100]
(30,60]
(10,30]
(0,10]
100
200
300
Age
Control
Private
Public
Locale
Large_City
Rural/Town
Small_City