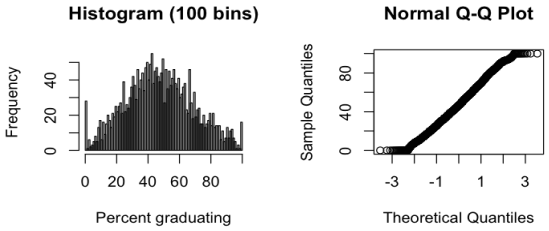
Correlates of Undergraduate Completion Rate in American Universities

Sigfried Gold
INST627 HW4
November 13, 2018
sigfried@sigfried.org

INTRODUCTION

We analyzed an observational data set from the National Center for Education Statistics [1] and compared the rate of undergraduate completion against six other variables from the NCES 2013 survey of 7,804 American universities. The distribution of graduation rates across universities is roughly normal, with peaks at 0% and 100%, as can be seen in the histogram and normal qq plot below. (It might have been reasonable to exclude the extreme values, but the assignment did not hint that this would be appropriate.)



ANALYSIS

Of the six variables we tested for relationship to graduation rate (see Table 1), all showed statistical significance (p-value < 0.0001) in single linear regression models, but the goodness-of-fit (represented by adjusted R^2 in Table 1) ranges widely between variables. Average SAT score is most predictive of graduation rate while school size is least predictive.

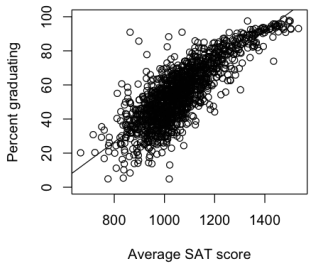
Without attempting to remove outliers or suspect data points, the six variables can be characterized as follows:

Student count is slightly correlated (R^2=0.015) with graduation rate, with an increase in undergraduate population of roughly 3,000 student being associated with a 1% increase in graduation rate, but accounting for less than 2% of variation in graduation rate.

Admission rate is somewhat more (negatively) correlated (R^2=0.104) with graduation rate, however, the residuals for this association are not normally distributed. The linear model for this relationship indicates that school with extremely low admissions rates should have a graduation around 73%, but in actuality, they are very close to 100%. For each 10% increase in the admissions rate, graduation rate falls by about 3%.

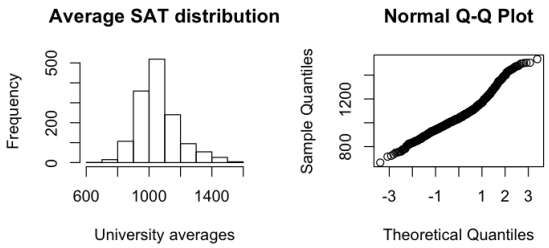
Average faculty salary, **pct w/ PELL grants**, and **average annual cost** are each moderately correlated with graduation rate (R^2 close to 0.3). Faculty salary and cost are positively associated with graduation rate, and percent with PELL grants, negatively.

Finally, **average SAT score**, with an R^2 of 0.662, is considerably more correlated (positively) with graduation rate than the other variables. As can be seen in these plots of graduation rate by average SAT score, their relationship has a discernible curve and the variability is not consistent.



Nevertheless, the observations cluster tightly around the least squares regression line, so a linear approximation appears suitable within the range of actual SAT score averages. At the low end of that range, an average score around 600 is associated with a 10% graduation rate, increasing by about 1% for each 10 point increase in average SAT score, up to rates in the high nineties for schools with average SAT scores around 1600.

The plots below show that the distribution of average SAT scores across universities in the sample are mostly normal, except at the high end, where they may fit a log normal pattern. This likely explains the curvature apparent in the scatter plot above.



Variable	Adj R^2	DF	Missing	Est. Intercept	Est. Slope	p-val Intercept	p-val Slope
Student count	0.0150	2446	5356	46	0.0004	0.00E+00	7.63E-10
Admission rate	0.1035	1794	5374	73	-31.4810	0.00E+00	9.46E-45
Average faculty salary	0.2734	2428	5417	17	0.0047	7.07E-55	7.42E-171
Pct w/ PELL grants	0.2839	2444	6427	73	-55.7200	0.00E+00	1.12E-179
Average annual cost	0.2894	2385	5358	21	0.0009	2.74E-98	2.04E-179
Average SAT score	0.6621	1375	6008	-61	0.1088	1.82E-129	0.00E+00

Table 1. Statistics for candidate model variables

Variable	Adj R^2	DF	Missing	Est. Intercept	Est. Slope	p-val Intercept	p-val Slope
Student count	0.0257	1370	0	53	0.0004	0.00E+00	1.45E-09
Admission rate	0.0855	1370	0	72	-27.4594	1.86E-270	1.12E-28
Average annual cost	0.3485	1370	0	29	0.0008	5.53E-136	7.18E-130
Average faculty salary	0.3672	1370	0	17	0.0050	2.04E-33	1.64E-138
Pct w/ PELL grants	0.5005	1370	0	86	-85.9499	0.00E+00	5.62E-209
Average SAT score	0.6747	1370	0	-61	0.1095	1.15E-136	0.00E+00

Table 2. Statistics for candidate model variables, omitting missing before analysis

DISCUSSION

Degrees of freedom ranges from 1375 to 2446 amongst the variables tested, due to differences in missing values. Table 2 shows the same analysis, but with rows containing missing values removed first. The degrees of freedom are identical in this case because the analysis is performed on the same set of rows for all variables.

REFERENCES

1. Susan Aud, Sidney Wilkinson-Flicker, Paul Kristapovich, Amy Rathbun, Xiaolei Wang, and Jijun Zhang. 2013. The Condition of Education 2013. NCES 2013-037. *National Center for Education Statistics* (2013).

https://www.overleaf.com/project/5be6c4cc94ef363ab610370d