A Comprehensive Study on US College Admissions

Tsirmpas Dimitris
Athens University of Economics and Business
Department of Informatics

May 20, 2023

Athens University of Economics and Business
Department of Statistics
Greece

Contents

1 Abstract **2**

2 Introduction **2**

3 Results **4**
3.1 Writing scores influenced by gender 4
3.2 Writing scores influenced by previous program 4
3.3 Building a predictive model 5
3.4 Verifying test score correlations 9

4 Conclusions & Discussion **12**

5 Bibliography **13**

6 Addendum **13**
6.1 Exploratory Analysis 13
6.2 OLS model preconditions 13

Name	Type	Description	Range
Id	Nominal	The student's ID	[1-200]
Gender	Binary	The student's gender	{male, female}
Race	Nominal	The student's race	{white, latin-american, asian, african-american}
Schtype	Binary	The type of the student's secondary education institution	{public, private}
Prog	Nominal	The student's previous study cycle	{general, vocation, academic }
Write	Numeric	The grade on the writing test	[0-100]
Math	Numeric	The grade on the mathematics test	[0-100]
Socst	Numeric	The grade on the social studies test	[0-100]

Table 1: An overview of the data used in this study.

1 Abstract

US College Admissions have and continue to be a subject of great debate among scholars and analysts. Such educational institutions have an interest in selecting the most qualified applicants using limited data, while the applicants themselves often protest admission requirements, especially those deemed discriminatory in nature. This study aims to analyze the factors that contribute to successful admissions in US colleges and universities by comparing their performance on standardized tests.

2 Introduction

The study uses a random sample of 200 students who applied to continue their studies in their respective universities. Their application consisted of three standardized tests testing their skills and knowledge in mathematics, social studies and creative writing. The dataset we used is available at LINK. An overview of the data contained can be found in Table 1.

We make the assumption that the dataset has been acquired through random, unbiased sampling. We also make the assumption that the records using different IDs represent different students.

The study is structured as follows. In this Section we make general observations about our data and we form our first hypotheses. In Section 3 we follow up with robust analyses and regression models to prove/disprove these hypotheses. Section 4 follows with an overview and discussion about our findings. Finally, in Section 6 we include graphs, tables and supporting documents.

The numerical variables contained in the dataset are described in Table 2. We observe that they are all almost symmetrical ($-0.5 \leq skew \leq 0.5$), and feature moderate negative (right) skewness with a mean/median hovering just above a score of 50. This indicates most students score around the baseline, most of which pass the exams with a mediocre grade.

Var.	Obs.	Mean	Std	Median	Trim	Min	Max	Skew	Kurt	SE
Write	200	52.77	9.48	54	53.36	31	67	-0.4	-0.7	0.67
Math	200	52.65	9.37	52	52.33	33	75	0.28	-0.6	0.66
Socst	200	52.41	10.7[illegible]	52	52.99	26	71	-0.3	-0.5	0.76

Table 2: Summary statistics on the numerical data used in the study.

	write	math	socst
write	1 0.000 ****	-	-
math	0.62 0.000 ****	1 0.00 ****	-
socst	0.60 0.000 ****	0.54 0.000****	1 0.000****

Table 3: Pearson's correlation coefficient (Holm's correction) between the tests and their p_values. Stars indicate significance scores: $>$ 1:", 0.1:'*', 0.01: '**', 0.001: '***', $<$ 0.0001: '****'

We next study the relationships between the three subjects. As shown in Table 3, there is a very statistically significant (p_value = 0.0000), positive ($r > 0.5$) relationship between all three subjects. This could be indicative of either one of the variables influencing the other, or an unknown, interfering variable which positively affects the three test scores. We hypothesize the latter, as the existence of such a variable indicating the student's general competence in tests makes intuitive sense. We will refer to this potential, unknown variable as "Competence" in this report.

Since the three subjects are strongly correlated we can explore correlations between the rest of the factors and any of the tests, assuming that a correlation with one strongly indicates with the other two as well.

We notice a probable correlation between gender and writing scores, as shown in Figure 1, as well as between the student's program and writing scores, as shown in Figure 2.

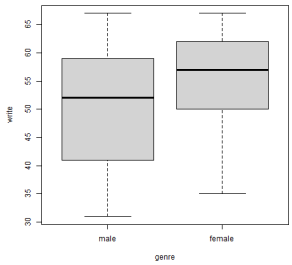

Figure 1: Boxplot displaying the writing score by gender.

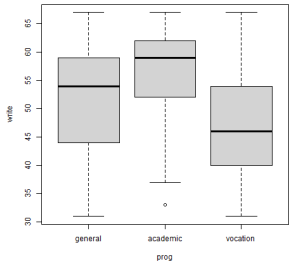

Figure 2: Boxplot displaying the writing score by current program. Notice the significant differences in means.

3 Results

3.1 Writing scores influenced by gender

Our exploratory analysis indicated a possible discrepancy between the results of writing tests between men and women, as well as between different programs. We thus investigate whether gender and the candidate's current program play a role in writing test score.

We begin by verifying the preconditions necessary for the standard t-test in order to compare the genders' scores. We compute the mean differences of the samples by subtracting the global mean by the women's scores [2], and conclude they are not normal (Shapiro-Wilk normality test, $p\_value = 0.0024$). The variances are not homogeneous (Levene's test $p\_value = 0.0022$), although this should not dissuade us from using a parametric t-test since the relatively large sample size ($N = 200$) and balanced groups ($N_{women} = 109, N_{men} = 91$) means the precondition's violation is not significant [3]. We will use a parametric test since the distribution of the differences has a normal kurtosis (Jarque-Bera Normality Test [1], $p\_value = 0.026$).

We conclude there is a statistically significant difference between the writing scores of men and women (Two-sided Welch Two Sample t-test, $p\_value = 0.0003$) with women having on average 5 more score than men (Welch Two Sample t-test with $H_a = less$, $p\_value = 0.00017$).

3.2 Writing scores influenced by previous program

We will now verify the preconditions for the parametric ANOVA test in order to test which programs are correlated with the writing tests' scores. The variances of the residuals are homogeneous (Levene's test $p\_value = 0.0022$), but not normal (Shapiro-Wilk normality test, $p\_value = 0.002$). We will thus use a non-parametric test.

We discover there is a statistically significant difference between the groups (Kruskall-Wallis rank sum test, $p\_value = 4e-08$). We can consult the boxplots in Figure 2, where we observe significant differences between all three groups.

3.3 Building a predictive model

Having confirmed our hypotheses regarding the relationships between the various variables and the writing test scores, we attempt to build a model which will estimate a candidate's math and social study scores, testing our hypothesis that the three scores are influenced by the same external factors.

Since there seem to be strong correlations between most independent variables and the writing scores, and since we already established a strong correlation between the scores themselves (attributed to the candidate's *"Competence"*), we will be using a simple, Ordinary Least Squares (OLS) model.

We initially build an OLS model estimating the writing scores which involves all the available variables. This model exhibits a good fit ($R^2_{adj} = 0.4753, F = 37.06 p\_value = 2.2e-16, df = 194$). We also verify all the preconditions for the regression model; the residuals appear to be normal (Shapiro-Wilk normality test $p\_value = 0.2099$), homogeneous (Levene's Test, $p\_value = 0.1057$) and non-correlated (Durbin-Watson test, $p\_value = 0.236$) and there are no significant outliers (see Addendum).

Our current model exhibits high confidence about the gender ($p\_value_{female} = 0.0056$), previous program ($p\_value_{progacademic} = 0.0025$), writing scores ($p\_value_{write} = 3.57e-08$), social study scores ($p\_value_{socst} = 0.0053$) and the constant ($p\_value_{intercept} = 8.29e-09$). However, it exhibits low confidence about the candidate's race ($p\_value_{raceasian} = 0.05$) and no confidence about his school type ($p\_value = 0.6502$). In order to build an optimal model, we consider dropping these two variables. Dropping the candidate's race along with the school type leads to a slight decrease of $R^2_{adj} = 0.4753$. Dropping only the school type on the other hand leads to an improved $R^2_{adj} = 0.4882$. Since the model maintains its confidence about the other variables, we keep this model as the optimal one. A summary of the optimal model can be found in Table 4.

We now follow the same procedure for the social studies test. We fit a model involving all available variables, which explains our sample to a decent degree ($R^2_{adj} = 0.4535, F = 17.7 p\_value = 2.2e-16, df = 188$). We detect an outlier in our precondition verification, a student with an above-average grade in writing but an abysmal one in social studies. Since this individual exerts a significant influence in our model (as it is largely dependent on the writing scores for its estimations) we remove them from the sample. The other preconditions are met, the residuals are sufficiently normal (Shapiro-Wilk normality test $p\_value = 0.0125$), homogeneous (Levene's Test, $p\_value = 0.3394$) and non-correlated (Durbin-Watson test, $p\_value = 0.678$).

As mentioned above, this model is reliant on the math and especially on the writing scores, while the rest of the variables are mostly non-statistically significant. We thus employ a stepwise procedure in order to eliminate variables deemed statistically insignificant while retaining, or increasing our models goodness of fit. A summary of the resulting model can be found in Table 5.

The results seem to verify our main hypothesis, that the test scores are predominately caused by the unknown "Competence" value. While other variables such as the candidate's past program, have a statistically significant influence in our model, we can see that the model's estimations are consistently based on the writing and other-lesson's test scores. We can additionally rule out this relationship being a result of multi-co-linearity between the test variables, as seen in the previous auto-correlations

tests.

Table 4: Linear regression model predicting math test scores, taking into account other test scores.

	Dependent variable:
	math
genrefemale	-2.828^{***}
	$(-4.799, -0.858)$
progacademic	3.788^{***}
	(1.346, 6.229)
progvocation	-0.375
	$(-3.161, 2.410)$
write	0.404^{***}
	(0.266, 0.542)
socst	0.167^{***}
	(0.052, 0.283)
raceasian	4.978^{*}
	$(-0.017, 9.972)$
raceafrican-amer	-1.103
	$(-5.104, 2.898)$
racewhite	2.324
	$(-0.686, 5.334)$
Constant	20.334^{***}
	(13.730, 26.938)
Observations	200
R^2	0.509
Adjusted R^2	0.488
Residual Std. Error	6.702 (df = 191)
F Statistic	24.729^{***} (df = 8; 191)
Note:	*p<0.1; **p<0.05; ***p<0.01

Table 5: Linear regression model predicting social study test scores, taking into account other test scores.

	Dependent variable:
	socst
raceasian	-5.642^{*}
	p = 0.060
raceafrican-amer	0.778
	p = 0.745
racewhite	0.308
	p = 0.865
progacademic	2.264
	p = 0.131
progvocation	-2.939^{*}
	p = 0.077
write	0.479^{***}
	p = 0.000
math	0.233^{***}
	p = 0.005
Constant	14.585^{***}
	p = 0.001
Observations	199
R^2	0.477
Adjusted R^2	0.458
Residual Std. Error	7.844 (df = 191)
F Statistic	24.899^{***} (df = 7; 191)
Note:	*p<0.1; **p<0.05; ***p<0.01

3.4 Verifying test score correlations

The results above, although encouraging, do not rule out the alternative hypothesis we posed in the Introduction of this report, that the definite correlation between the test scores is caused by one of the test scores themselves influencing the others. We thus repeat the experiments of the previous subsection while omitting the writing scores. If our alternative hypothesis was correct we would expect our models to no longer have a good performance, while having a much smaller statistical significance on the other lesson's model.

We begin by constructing an OLS regression model which tries to estimate the math score by considering the candidate's characteristics and his scores in the social studies test. Our initial model including all the variables scores a $R^2_{adj} = 0.3999$ score. We next verify all the necessary preconditions; the residuals appear to be normal (Shapiro-Wilk normality test $p\_value = 0.8403$), homogeneous (Levene's Test, $p\_value = 0.1641$) and non-correlated (Durbin-Watson test, $p\_value = 0.534$) and there are no significant outliers.

This model again displays high confidence in the following variables: Race ($p\_value_{raceasian} = 0.002$, $p\_value_{racewhite} = 0.012$), Program ($p\_value_{progacademic} = 0.0005$), Social Study Test Scores ($p\_value_{socst} = 4.25e-09$) and the Constant ($p\_value_{socst} = 2e-16$). By employing a backwards procedure we end up with our final model with a total $R^2_{adj} = 0.4022$, found in Table 6.

We repeat the procedure for estimating the social study test scores without relying on the writing tests. Our base model, which considers all the available variables, displays a $R^2_{adj} = 0.3393$, which is considerably worse than the respective math model. This may be because of the previously mentioned reliance on the writing scores, which are no longer available to the model. Additionally, similarly to the previous models, the model lacks confidence in almost all other variables; the only statistical significant variable other than the math scores ($p\_value_{math} = 4.25e-09$) and the Constant ($p\_value_{Intercept} = 5.27e-08$) is the candidate's Program ($p\_value_{progvocation} = 0.0283$).

We again verify the preconditions necessary for the linear regression model. The residuals appear to be normal (Shapiro-Wilk normality test $p\_value = 0.7449$), homogeneous (Levene's Test, $p\_value = 0.2314$) and non-correlated (Durbin-Watson test, $p\_value = 0.952$) and there are no significant outliers.

Because of the many possible variables that are candidates for removal we can again employ a stepwise model selection algorithm. The best model by AIC keeps only the math and program variables (as expected) but explains less of the data ($R^2_{adj} = 0.3383$). Since the reduction in R^2_{adj} is minimal, and since the new model can reach this score by discarding almost all other variables (which as described above are essentially considered as noise by our model), we will be using this model as optimal (Table 7).

These findings seem to contradict our alternative hypothesis. While our models' performance certainly degraded, they still achieve comparable results, with their performance loss being explained by the degree to which *"Competence"* can be measured. In other words, this unknown variable can be approximated more accurately by considering both other tests, instead of just one. This is further proof that our initial hypothesis appears to be correct.

Table 6: Linear regression model predicting math test scores, without relying on the writing tests.

	Dependent variable:
	math
raceasian	8.074^{***}
	p = 0.003
raceafrican-amer	-1.261
	p = 0.567
racewhite	4.026^{**}
	p = 0.015
progacademic	4.726^{***}
	p = 0.0005
progvocation	-1.039
	p = 0.499
socst	0.335^{***}
	p = 0.000
Constant	29.617^{***}
	p = 0.000
Observations	200
R^2	0.420
Adjusted R^2	0.402
Residual Std. Error	7.243 (df = 193)
F Statistic	23.318^{***} (df = 6; 193)
Note:	*p<0.1; **p<0.05; ***p<0.01

Table 7: Linear regression model predicting social study test scores, without relying on the writing tests.

	Dependent variable:
	socst
progacademic	2.833*
	p = 0.085
progvocation	−3.829**
	p = 0.037
math	0.486***
	p = 0.000
Constant	26.285***
	p = 0.000
Observations	200
R^2	0.348
Adjusted R^2	0.338
Residual Std. Error	8.733 (df = 196)
F Statistic	34.908*** (df = 3; 196)
Note:	*p<0.1; **p<0.05; ***p<0.01

4 Conclusions & Discussion

Donec et nisl id sapien blandit mattis. Aenean dictum odio sit amet risus. Morbi purus. Nulla a est sit amet purus venenatis iaculis. Vivamus viverra purus vel magna. Donec in justo sed odio malesuada dapibus. Nunc ultrices aliquam nunc. Vivamus facilisis pellentesque velit. Nulla nunc velit, vulputate dapibus, vulputate id, mattis ac, justo. Nam mattis elit dapibus purus. Quisque enim risus, congue non, elementum ut, mattis quis, sem. Quisque elit.

Maecenas non massa. Vestibulum pharetra nulla at lorem. Duis quis quam id lacus dapibus interdum. Nulla lorem. Donec ut ante quis dolor bibendum condimentum. Etiam egestas tortor vitae lacus. Praesent cursus. Mauris bibendum pede at elit. Morbi et felis a lectus interdum facilisis. Sed suscipit gravida turpis. Nulla at lectus. Vestibulum ante ipsum primis in faucibus orci luctus et ultrices posuere cubilia Curae; Praesent nonummy luctus nibh. Proin turpis nunc, congue eu, egestas ut, fringilla at, tellus. In hac habitasse platea dictumst.

Vivamus eu tellus sed tellus consequat suscipit. Nam orci orci, malesuada id, gravida nec, ultricies vitae, erat. Donec risus turpis, luctus sit amet, interdum quis, porta sed, ipsum. Suspendisse condimentum, tortor at egestas posuere, neque metus tempor orci, et tincidunt urna nunc a purus. Sed facilisis blandit tellus. Nunc risus sem, suscipit nec, eleifend quis, cursus quis, libero. Curabitur et dolor. Sed vitae sem. Cum sociis natoque penatibus et magnis dis parturient montes, nascetur ridiculus mus. Maecenas ante. Duis ullamcorper enim. Donec tristique enim eu leo. Nullam molestie elit eu dolor. Nullam bibendum, turpis vitae tristique gravida, quam sapien tempor lectus, quis pretium tellus purus ac quam. Nulla facilisi.

Duis aliquet dui in est. Donec eget est. Nunc lectus odio, varius at, fermentum in, accumsan non, enim. Aliquam erat volutpat. Proin sit amet nulla ut eros consectetuer cursus. Phasellus dapibus aliquam justo. Nunc laoreet. Donec consequat placerat magna. Duis pretium tincidunt justo. Sed sollicitudin vestibulum quam. Nam quis ligula. Vivamus at metus. Etiam imperdiet imperdiet pede. Aenean turpis. Fusce augue velit, scelerisque sollicitudin, dictum vitae, tempor et, pede. Donec wisi sapien, feugiat in, fermentum ut, sollicitudin adipiscing, metus.

Donec vel nibh ut felis consectetuer laoreet. Donec pede. Sed id quam id wisi laoreet suscipit. Nulla lectus dolor, aliquam ac, fringilla eget, mollis ut, orci. In pellentesque justo in ligula. Maecenas turpis. Donec eleifend leo at felis tincidunt consequat. Aenean turpis metus, malesuada sed, condimentum sit amet, auctor a, wisi. Pellentesque sapien elit, bibendum ac, posuere et, congue eu, felis. Vestibulum mattis libero quis metus scelerisque ultrices. Sed purus.

Donec molestie, magna ut luctus ultrices, tellus arcu nonummy velit, sit amet pulvinar elit justo et mauris. In pede. Maecenas euismod elit eu erat. Aliquam augue wisi, facilisis congue, suscipit in, adipiscing et, ante. In justo. Cras lobortis neque ac ipsum. Nunc fermentum massa at ante. Donec orci tortor, egestas sit amet, ultrices eget, venenatis eget, mi. Maecenas vehicula leo semper est. Mauris vel metus. Aliquam erat volutpat. In rhoncus sapien ac tellus. Pellentesque ligula.

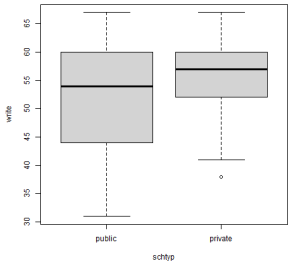

Figure 3: Boxplot displaying the writing score by school type.

5 Bibliography

References

[1] Carlos M. Jarque and Anil K. Bera. "Efficient tests for normality, homoscedasticity and serial independence of regression residuals". In: *Economics Letters* 6.3 (1980), pp. 255–259. ISSN: 0165-1765. DOI: https://doi.org/10.1016/0165-1765(80)90024-5. URL: https://www.sciencedirect.com/science/article/pii/0165176580900245.

[2] Geoffrey R Loftus and Michael EJ Masson. "Using confidence intervals in within-subject designs". In: *Psychonomic bulletin & review* 1.4 (1994), pp. 476–490.

[3] Donald W Zimmerman. "A note on preliminary tests of equality of variances". In: *British Journal of Mathematical and Statistical Psychology* 57.1 (2004), pp. 173–181.

6 Addendum

6.1 Exploratory Analysis

In this section we include tables and figures which were used in the exploratory analysis of the data in the Introduction.

6.2 OLS model preconditions

In order to visually confirm the normal distribution of the model's residuals, we plot their boxplots for each of the 4 quantiles. We expect these boxplots to resemble those of the normal distribution, centered on $y = 0$ and with 95% of their values not going above/below the $y = 1.95$ and $y = -1.95$ respectively. Figures 5, 6 show models including the writing scores. Figures 7, 8 show models including only the other lesson's scores.

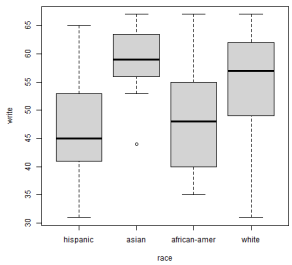

Figure 4: Boxplot displaying the writing score by candidate race.

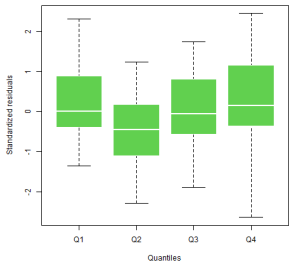

Figure 5: The normalized residuals of the math model plotted for each of the 4 quantiles.

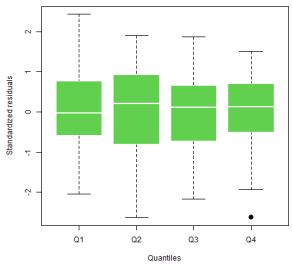

Figure 6: The normalized residuals of the social studies model plotted for each of the 4 quantiles.

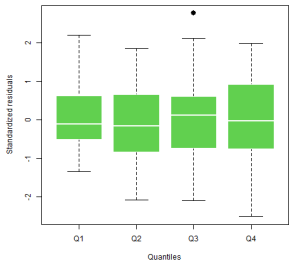

Figure 7: The normalized residuals of the math model plotted for each of the 4 quantiles.

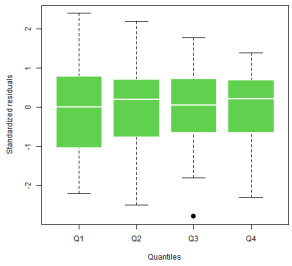

Figure 8: The normalized residuals of the math model plotted for each of the 4 quantiles.

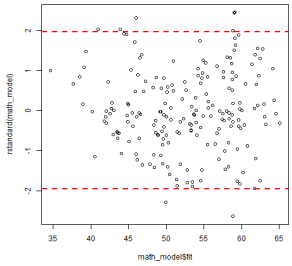

Figure 9: The normalized residuals of the math model plotted against the model's estimations.

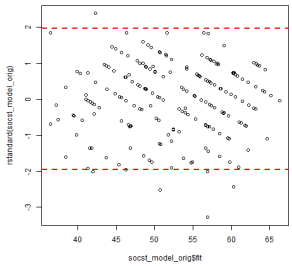

Figure 10: The normalized residuals of the social studies model plotted against the model's estimations. Notice the one outlier above the $y = 3$ line, which represents the data point that was investigated and removed.

We also check for outliers by plotting the normalized residuals against the model's estimations. The red lines denote the $y = 1.95$ and $y = -1.95$ values respectively and we expect 95% of the points to be within them. Any value outside of $[-3, 3]$ indicates a strong outlier which must be investigated. Figures 9, 10 show models including the writing scores. Figures 11, 12 show models including only the other lesson's scores.

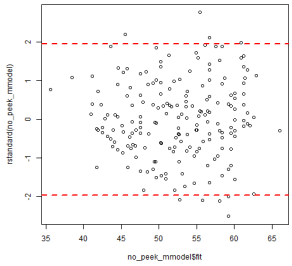

Figure 11: The normalized residuals of the math model plotted against the model's estimations. We notice three potential outliers. These values are considered non-anomalous, as they stray sufficiently away from the $y = 3$ and $y = -3$ brackets, and are to be expected in a sample of 200 values.

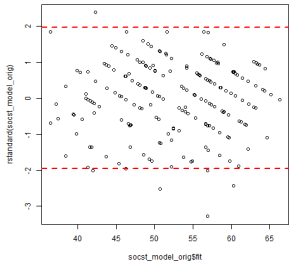

Figure 12: The normalized residuals of the social studies model (without the writing scores variable) plotted against the model's estimations. We don't consider the values above and below the $y = 3$ and $y = -3$ brackets respectively for the same reasons as in Figure 11.