A Statistical Study on Test Scores in US College Admissions

Tsirmpas Dimitris
Athens University of Economics and Business
Department of Informatics

June 5, 2023

Professors: I. Ntzoufras, X. Penteli
Athens University of Economics and Business
Department of Statistics
Greece

Contents

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. In this study we investigate how test scores can be explained by various candidate traits and prior performance. We discover a relationship between the candidate's gender and previous program and their overall test scores. We also identify a positive correlation between test scores, which we attribute to a confounding variable which we name "Competence".

2 Introduction

The aim of this study is to investigate possible links and relationships between a candidate's characteristics and their performance in multiple standardized tests. We employ 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. An overview of the data contained can be found in Table 1.

The study is structured as follows: In Section 3 we make general observations about our data, identify key relationships and form our first hypotheses. We follow up these hypotheses in Section 4 with robust analyses and in Section 5 by employing various regression models. Section 6 features an overview and discussion about our findings. Finally, we include graphs, tables and supporting documents in the report's Addendum (Section 8).

The data and replication code can be found in our GitHub repository [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 (and as such are considered independent samples).

For the sake of brevity we will refer to the following statistical tests with the following acronyms: Shapiro Wilk (S-W) and Lilliefors (Kolmogorov-Smirnov) (K-S) normality tests, Bartlett (Bart) and Levene's (Lev) tests of homogeneity of variances, Durbin Watson (D-W) autocorrelation test, Tukeys Honest Significance Test (Tukey), Welch Two Sample t-test (Welch) and the Wilcoxon (Wil) and Kruskal-Wallis (K-W) rank sum tests.

All the test, images and graphs were executed and built using `R 4.11` and the `haven, nortest, car, psych, sjPlot, gplot` and `stargazer` libraries.

3 Exploratory Analysis

The numerical variables contained in the dataset are described in Table 2 and their distributions can be seen in Figure 1. We observe that they are all almost symmetrical ($-0.5 \leq skew \leq 0.5$), and feature

[1] `https://github.com/dimits-exe/collegeanalysis`

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.

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	10.74	52	52.99	26	71	-0.3	-0.5	0.76

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

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.

We next study relationships between the candidates' characteristics (not including test scores). We run χ^2 tests on *Gender, Race, Program* and *School Type*. The only statistically significant relationship in our dataset is between *School Type* and *Program* ($p = 0.015$), which can be seen in the Addendum.

Finally, we study the relationships between the three subjects. There is a very statistically significant ($p < 0.0001$), positive (Pearson's $r > 0.5$) relationship between all three subjects. This could be indicative of either one of the variables influencing the other, or an unknown, confounding 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 confounding variable as *"Competence"* in this report. Since the three tests are strongly correlated we can extrapolate with some certainty that a relationship between one characteristic and one of the tests implies a similar relationship between that characteristic and the other tests as well.

We notice a probable correlation between gender and writing scores, as shown in Figure 2, as well as between the student's program and writing scores, as shown in Figure 3.

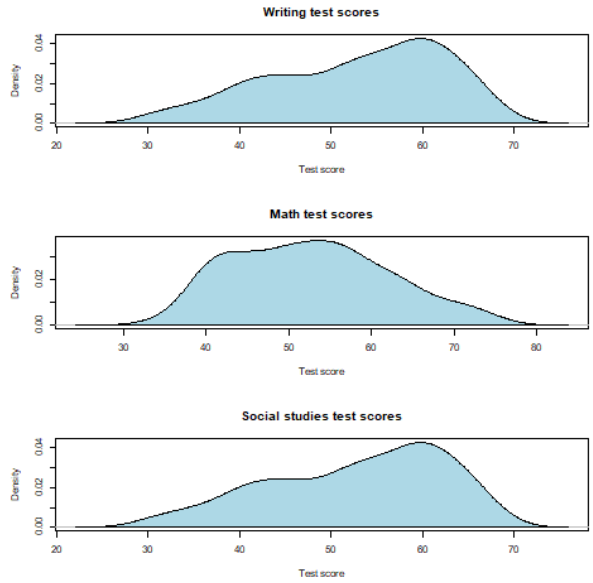

Figure 1: Plots displaying the test score's distributions in our dataset.

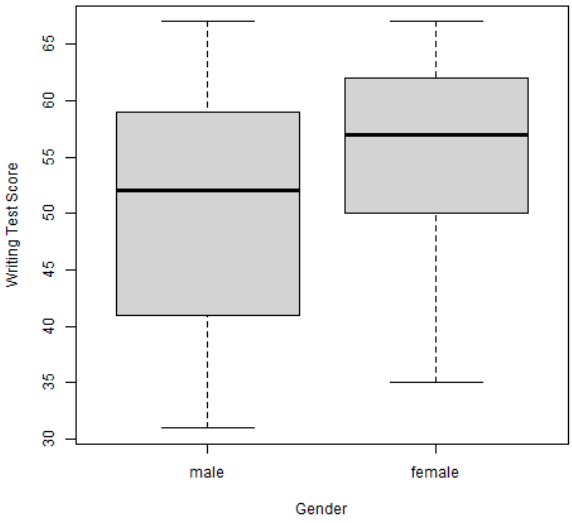

Figure 2: Boxplot displaying the writing score by gender.

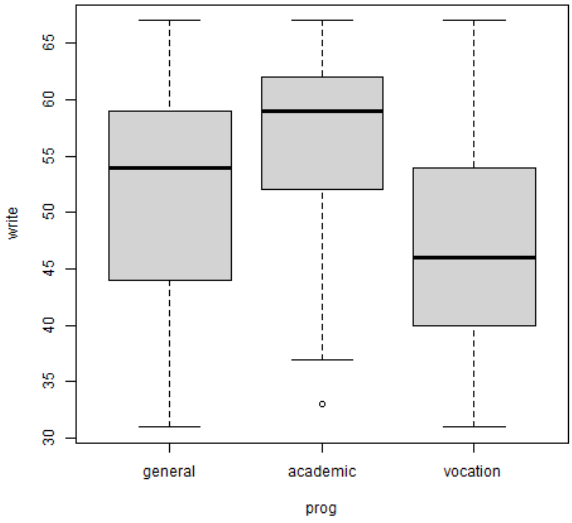

Figure 3: Boxplot displaying the writing score by current program. Notice the significant differences in means.

4 Variable correlations

4.1 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 scores.

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 (Loftus and Masson [1]), and conclude they are not normal (S-W, $p = 0.0024$). The variances are not homogeneous (Lev $p = 0.0022 < 0.05$, Bart $p = 0.0019 < 0.05$). These 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 violation of the normality and homogeneity preconditions is not significant (Zimmerman [3]). Since the mean is incredibly close to our median in our sample, and the writing score distribution we saw in Figure 1 seems well behaved, we can use a parametric t-test to compare the differences between the test scores of men and women.

We conclude there is a statistically significant difference between the writing scores of men and women (Welch $p = 0.0003$) with women having on average 5 more score than men (Welch with $H_a = less$, $p = 0.0002$). The differences with 95% confidence intervals can be seen in Figure 4.

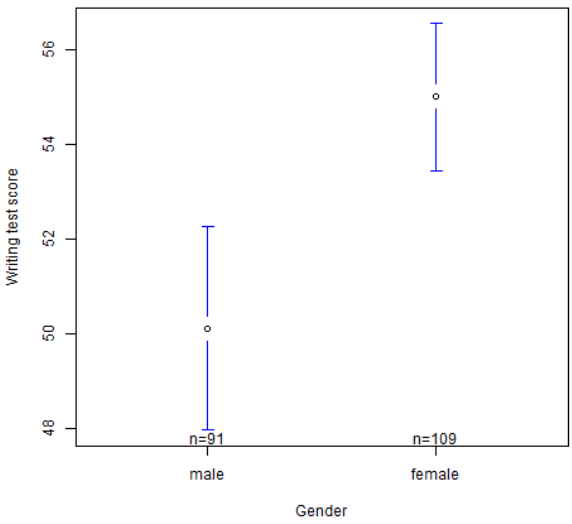

Figure 4: Error bars displaying writing score by gender. The blue lines indicate the 95% confidence interval, meaning the range of values we expect 95% of the observations existing in, for each sample.

4.2 Writing scores influenced by previous program

We will now verify the preconditions for the parametric ANOVA test in order to test which past programs are correlated with the writing tests' scores. The variances of the residuals are homogeneous (Lev $p = 0.1873$, Bart $p = 0.27$), but not normal (S-W $p = 0.002$, K-S $p = 0.024$). Since we have a large sample size, and since the mean seems to be suitable for comparing differences between the groups (see section above), we can use a parametric ANOVA test.

We discover there is a statistically significant difference between the groups (ANOVA, $p = 4.3e-09$). We specifically discover significant differences between the academic and general programs (Paired t-test $p = 0.032$, Tukey $p = 0.005$), academic and vocational (Paired t-test $p = 3.4e-09$, Tukey $p < 0.0001$) and between general and vocational programs (Paired t-test $p = 0.107$, Tukey $p = 0.029$). The mean differences between the different past programs can be seen in Figure 5.

5 Predictive / descriptive models

5.1 Building the base models

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,

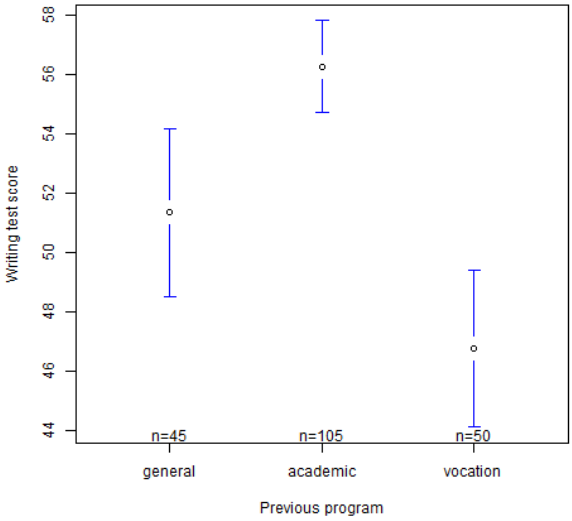

Figure 5: Error bars displaying writing score by gender. The blue lines indicate the 95% confidence interval, meaning the range of values we expect 95% of the observations existing in, for each sample.

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, being able to explain almost half of the variance in our data ($R^2_{adj} = 0.4861, BIC = 1377$). However, it exhibits low confidence about the candidate's race and no confidence about the school type ($p = 0.6502$). 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 $R^2_{adj} = 0.488, BIC = 1372, F = 24.729 p =< 0.001$. We use BIC to compare our models as they are descriptive, not predictive.

We briefly check the pre-conditions for linear regression. The residuals are not sufficiently normal (S-W $p = 0.0125$, K-S $p = 0.0241$), although this does not discourage us, since linear models have been robust against normality assumption violations for large samples ($N \geq 100$) [2]. The residuals are homogeneous given a 95% confidence level (Lev $p = 0.081$, Bart $p = 0.05368$) and non-correlated (D-W, $p = 0.26$). There are also no sings of multicollinearity. While some outliers exist, none of them are considered anomalies.

An issue with our model is that there doesn't seem to be a strong linear relationship (see Addendum). Our data however are exclusively either categorical variables, or numeric variables following the same scale (0-100 score) and distribution (see Section 3). Thus, our model performs worse when mathematical transformations such as logarithms are applied to our data.

Table 3: Linear regression model predicting math test scores, taking into account other test scores.

	Dependent variable:
	math
genderfemale	-2.828^{***}
	p = 0.006
progacademic	3.788^{***}
	p = 0.003
progvocation	-0.375
	p = 0.793
write	0.404^{***}
	p = 0.00000
socst	0.167^{***}
	p = 0.005
raceasian	4.978^{*}
	p = 0.053
raceafrican-amer	-1.103
	p = 0.590
racewhite	2.324
	p = 0.132
Constant	20.334^{***}
	p = 0.000
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

The resulting model can be found in Table 3 and can be interpreted as such, assuming all other variables remain constant:

- If a candidate is female, Hispanic, has a generic background, and completely failed her other tests (writing score = social studies score = 0), her score in the math test would be 20.334.
- Each point scored in the writing test means the math score will be *higher* for an average of 0.4 points.
- Each point scored in the social studies test means the math score will be on average 0.167 points *higher*.
- If the candidate is male, he will score an average of 2.8 *less* points.
- Depending on his background the candidate will either on average score 3.788 *higher* (academic) or 0.375 (vocational) *lower* than average.
- Finally, depending on his race a candidate will score an average of 5 *higher* (Asian), 0.308 *higher* (White) or 1.1 *lower* (African American) than average.

We now follow the same procedure for the social studies test, by employing a stepwise procedure in order to eliminate variables deemed statistically insignificant while retaining our models goodness of fit. The residuals are not sufficiently normal (S-W $p = 0.008$, K-S $p = 0.061$), but are homogeneous (Lev $p = 0.452$, Bart $p = 0.635$) and non-correlated (D-W, $p = 0.848$). There appears to be no multicollinearity, the residuals are marginally not linear and there are no anomalies. Since the normality assumption can be waived because of our large sample size, we assume all preconditions are met, other than linearity.

Our model features a score of ($R^2_{adj} = 0.488, BIC = 1452$) which is an improvement over the full model ($R^2_{adj} = 0.4581, BIC = 1377$). The reason for this is, as mentioned above, that the model is reliant on the math and especially on the writing scores, while the rest of the variables are mostly not statistically significant. A summary of the resulting model can be found in Table 4 and can be interpreted in a similar manner to the model above.

The results seem to verify our main hypothesis, that the test scores are predominately caused by the unknown *"Competence"* variable. 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 correlation between the test variables, as seen in the previous correlation tests.

5.2 Identifying test score causation

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

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

	Dependent variable:
	socst
progacademic	2.276
	p = 0.137
progvocation	-2.669
	p = 0.118
write	0.446^{***}
	p = 0.00000
math	0.242^{***}
	p = 0.004
Constant	15.610^{***}
	p = 0.0003
Observations	200
R^2	0.441
Adjusted R^2	0.429
Residual Std. Error	8.111 (df = 195)
F Statistic	38.416^{***} (df = 4; 195)
Note:	*p<0.1; **p<0.05; ***p<0.01

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 their score in the social studies test. Our initial model including all the variables performs adequately $R^2_{adj} = 0.3999, BIC = 1404$. By employing a backwards procedure we end up with our final model with a total $R^2_{adj} = 0.4022, BIC = 1395$, found in Table 5.

We verify all the necessary preconditions; the residuals appear to be normal (S-W $p = 0.863$, K-S $p = 0.885$), homogeneous (Lev $p = 0.431$, Bart $p = 0.42$) and non-correlated (D-W, $p = 0.57$). There appears to be no multicollinearity, the residuals are marginally not linear and there are no anomalies.

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 $R^2_{adj} = 0.3393, BIC = 1478$, 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 ($p_{math} = 4.25e-09$) and the Constant ($p_{Intercept} = 5.27e-08$) is the candidate's Program ($p_{progvocation} = 0.0283$).

Because of the many possible variables that are candidates for removal we can again employ a stepwise model selection algorithm. The best model by BIC (Table 6) keeps only the math and program variables (as expected) but overall explains less of the data ($R^2_{adj} = 0.3383, BIC = 1457$).

We again verify the preconditions necessary for the linear regression model. The residuals appear to

Table 5: 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

not be normal (S-W $p = 0.03985$, K-S $p = 0.02452$), homogeneous (Lev $p = 0.1413$, Bart $p = 0.2216$) and the variables are non-correlated (D-W, $p = 0.984$). There appears to be no multicollinearity, the residuals are linear and there are no anomalies. We consider the model sound despite the non-normality of the residuals because of our large sample size.

Table 6: 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

These findings appear 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.

5.3 Ruling out overfitting

We briefly consider the possibility that our descriptive models overfit on the test scores, as they are the only numerical values in our dataset. To rule this possibility out, we re-run the tests from Section 5.1 where all test variables are replaced with a binary variable denoting whether the candidate passed the respective test ($score > 50$). Besides practical and symbolic significance, the threshold of 50 was picked because the distribution of the test scores follows the normal distribution around a mean a little over 50, so we expect most observations to be meaningfully differentiated across the two bins.

If our models utilized the information efficiently, we would expect the new models' performance to not be substantially impacted. Indeed the respective binary math model performs with a respectable R^2_{adj} of 0.4255 ($BIC = 1395$) and the social studies model with an R^2_{adj} of 0.3442 ($BIC = 1460$). This further disproves the hypothesis that the models just overfitted on the other test scores.

The math model seems to satisfy all preconditions (normal residuals - S-W $p = 0.904$, K-S $p = 0.72$, homogeneous - Lev $p = 0.1805$, Bart $p = 0.1028$ and not auto-correlated - D-W $p = 0.794$). The social studies model is more unstable (residuals not normal - S-W $p = 0.009$, K-S $p = 0.0008$, not homogeneous - Lev $p = 0.001$, Bart $p = 3.395e-07$ and auto-correlated with a significance level of 10% - D-W $p = 0.61$), but curiously the factorization of the variables seems to have fixed the linearity issues (see Addendum). The variables remain non-correlated with no anomalies.

6 Conclusions & Discussion

In this report we studied extensively the relationships between different characteristics of US university applicants. We observed that characteristics such as race, gender, schooling and previous programs are generally uncorrelated, with the exception of a positive relationship between private schooling and academic background. We verified that female candidates score on average higher than males in writing tests, as well as that the candidate's previous program influences their writing test scores.

Additionally, we observed a definite positive correlation between candidate test scores. We hypothesized this was the product of a confounding variable we called *"Competence"* which similarly influences all test scores, and posed an alternative hypothesis stating that one of the test scores was the cause of the correlation. We show evidence of this variable's existence by constructing linear regression models and disprove the alternative hypothesis by constructing such models without access to the common variable *Writing Score*.

This study highlights that other test scores are consistently robust and important variables when attempting to assess future test scores, even if these test scores are on different disciplines (such as social studies and mathematics). Further research is warranted to check whether past test scores can be used to consistently predict candidate performance, as well as from how far in the past, and between which disciplines these observations would be useful.

We note that our statistical models were used exclusively as explanatory models and should not be used for prediction. We also note that the relationships presented in this report are only representative of our sample, and should not be generalized for the general population. Finally, we warn against extrapolating any relationships from our tests, since they only imply a correlation, not necessarily causation.

7 References

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

[2] Thomas Lumley et al. "The importance of the normality assumption in large public health data sets". In: *Annual review of public health* 23.1 (2002), pp. 151–169.

[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.

	Writing	Math	Social Studies
Writing			
Math	0.62 0.000 ****		
Social Studies	0.60 0.000 ****	0.54 0.000****	

Table 7: Pearson's correlation coefficient (Holm's correction) between the tests and their ps. Stars indicate significance scores: > 1:", 0.1:'*', 0.01: '**', 0.001: '***', < 0.0001: '****'.

Previous Program				
School Type	**general**	**academic**	**vocation**	**Total**
public	39 23.2%	81 48.2%	48 28.6%	168 100%
private	6 18.8%	24 75%	2 6.2%	32 100%
Total	45 22.5%	105 52.5%	50 25%	200 100%

Table 8: χ^2 test between *School Type* and *Program*. Notice the overwhelming majority of candidates who attended private schools having an academic background prior to applying.

8 Addendum

8.1 Exploratory Analysis

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

8.2 OLS model preconditions

In order to check for the existence of a linear relationship between the model's residuals, we plot them against their values. A model is considered to have a linear relationship, if the conditional mean (red) line deviates from the horizontal reference line ($y = 0$). Figures 8, 9 show the two math models whose conditional means show varying deviations from the reference line.

In order to visually confirm the homogeneity of variances 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 10, 11 show models including the writing scores. Figures 12, 13 show models including only the other lesson's scores.

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 14, 15 show models including the writing scores. Figures 16, 17 show models including only the other

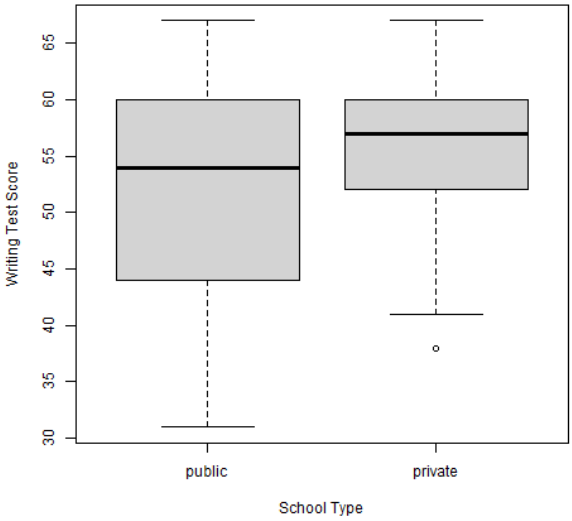

Figure 6: Boxplot displaying the writing score by school type.

lesson's scores.

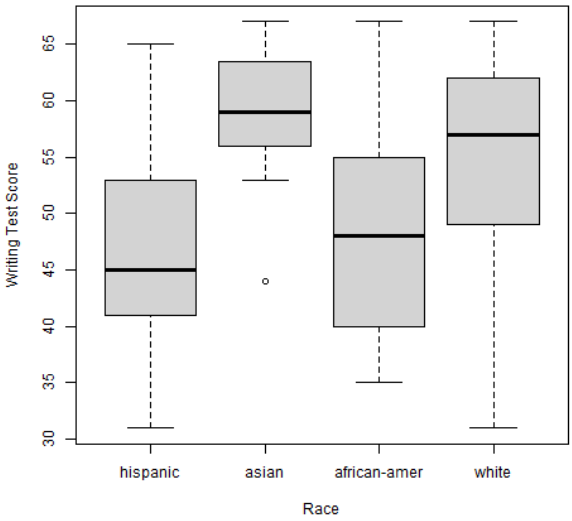

Figure 7: Boxplot displaying the writing score by candidate race.

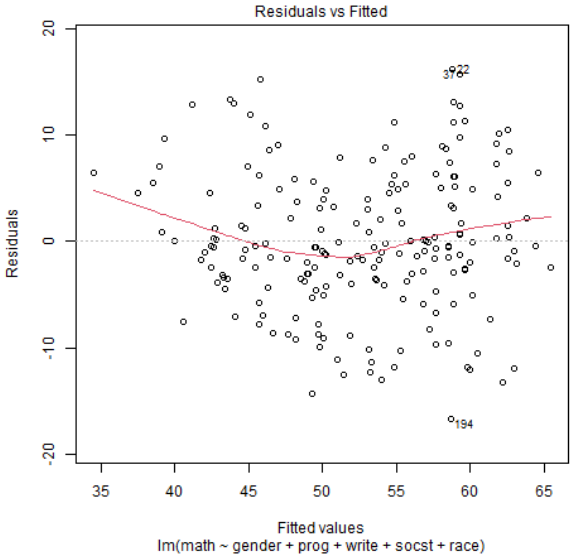

Figure 8: Linear reference plot for the original (optimal) math model. Notice the conditional mean (red) line shape deviating from $y = 0$. This is evidence of something exerting influence over our data other than the model's variables.

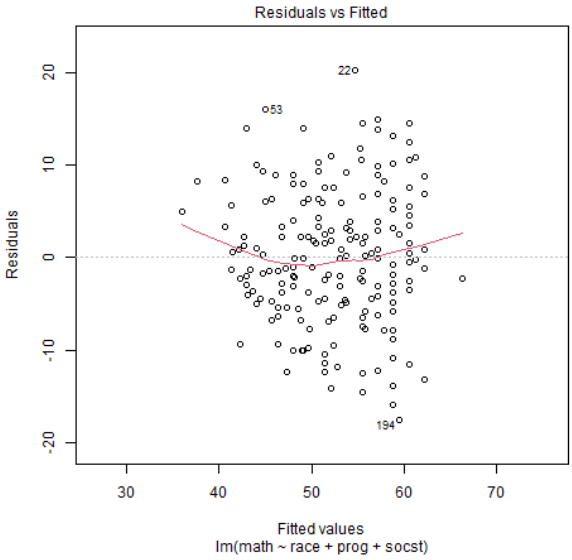

Figure 9: Linear reference plot for the second (optimal) math model. The conditional mean (red) line has an identical shape to the one in Figure 8.

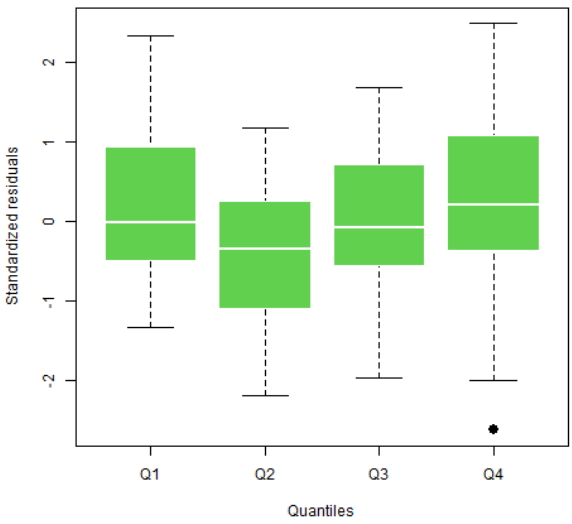

Figure 10: The normalized residuals of the math model plotted for each of the 4 quantiles.

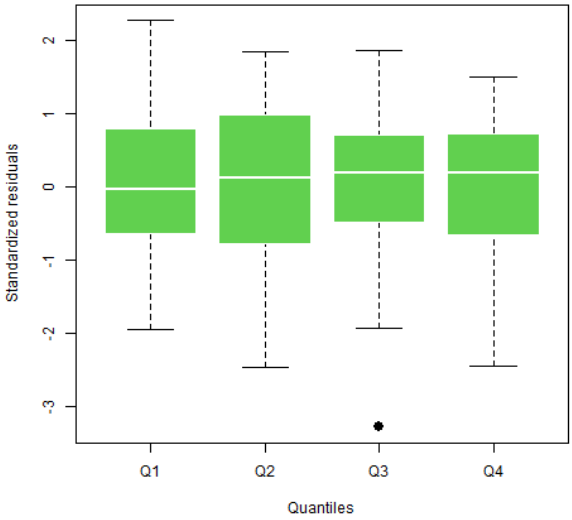

Figure 11: The normalized residuals of the social studies model plotted for each of the 4 quantiles.

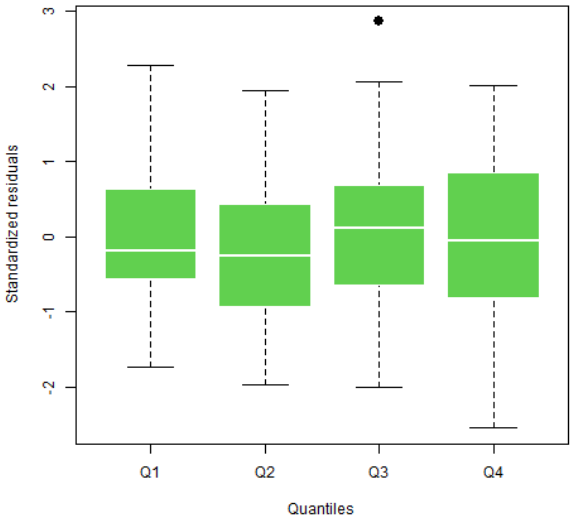

Figure 12: The normalized residuals of the math model plotted for each of the 4 quantiles.

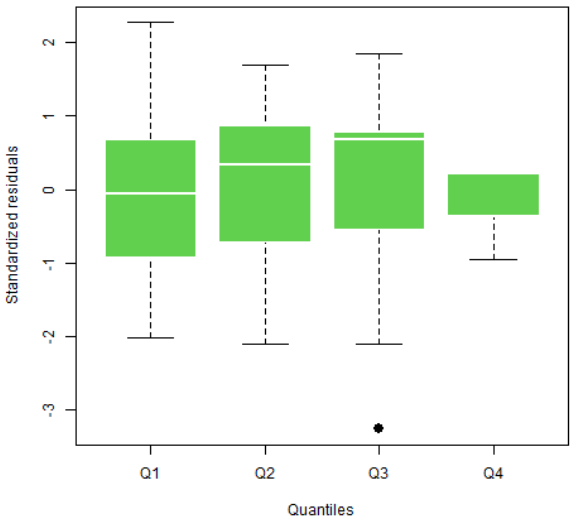

Figure 13: The normalized residuals of the math model plotted for each of the 4 quantiles.

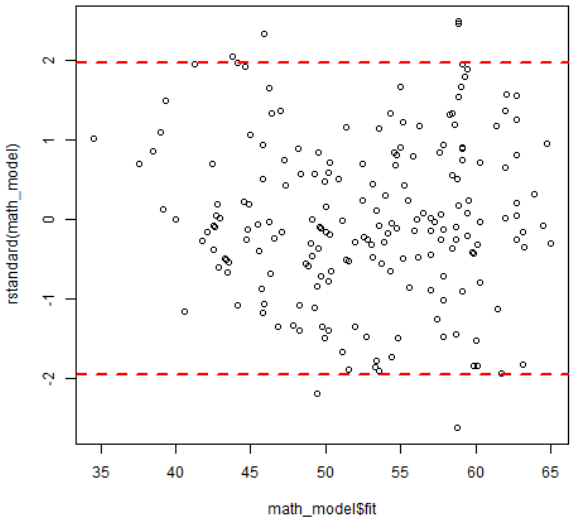

Figure 14: The normalized residuals of the math model plotted against the model's estimations.

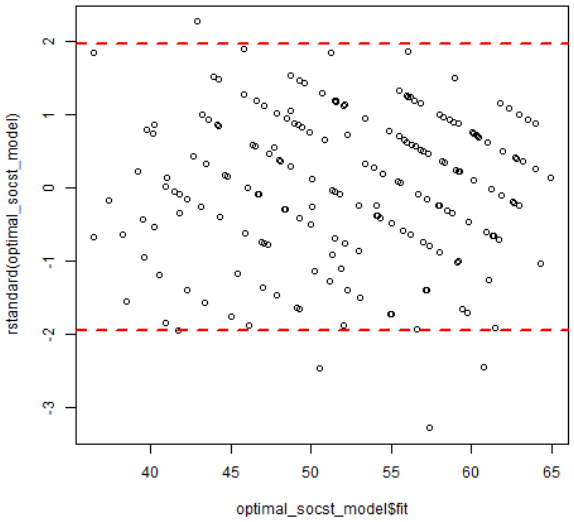

Figure 15: The normalized residuals of the social studies model plotted against the model's estimations. Notice the one outlier above the $y = 3$ line, which represents the data point that was investigated and removed.

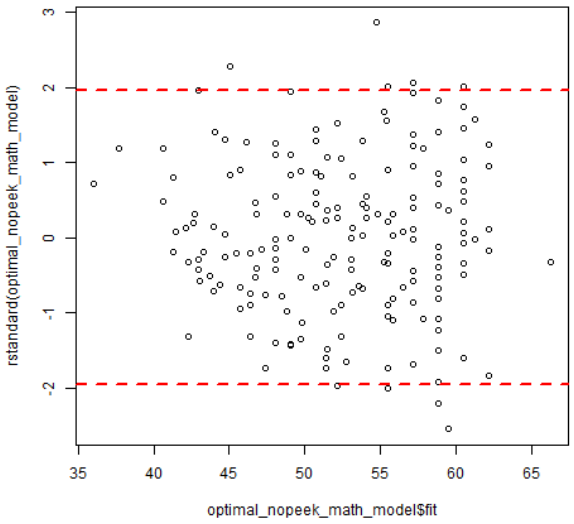

Figure 16: The normalized residuals of the math model plotted against the model's estimations. We notice three potential outliers. These values are considered non-anomalous, as they stray sufficiently away from the $y = 3$ and $y = -3$ brackets, and are to be expected in a sample of 200 values.

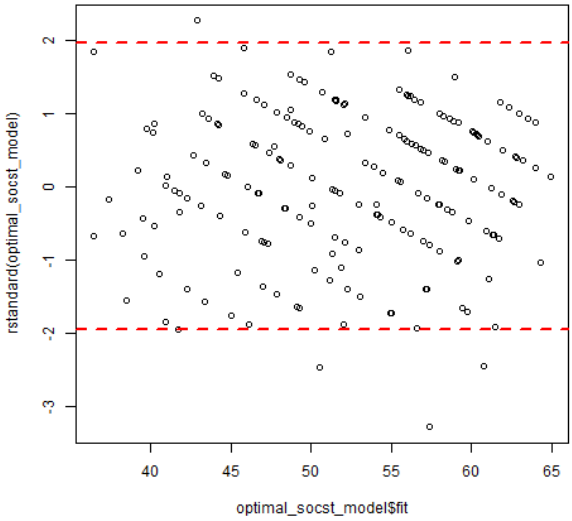

Figure 17: The normalized residuals of the social studies model (without the writing scores variable) plotted against the model's estimations. We don't consider the values above and below the $y = 3$ and $y = -3$ brackets respectively for the same reasons as in Figure 16.