AN ANALYSIS OF FACTORS INFLUENCING SAT SCORES

Alok Sheth and Samuel Lehr

Systems 302 Project
Fall, 2001
# TABLE OF CONTENTS

Abstract_____________________________________________________ 1

Motivation for Research________________________________________ 1

Sampling Method______________________________________________ 2
  Predictors______________________________________________ 2
  Method of Data Collection_________________________________ 8

Our Predictions________________________________________________ 10

Analysis of Data________________________________________________ 12

Multiple Regression Analysis_____________________________________ 19
  Stepwise Regression_______________________________________ 19
  Predicted Regression______________________________________ 21
  Regressing All Factors Together_____________________________ 23

Final Thoughts__________________________________________________ 24

Interesting Side Notes____________________________________________ 25

Possibilities for Further Study_____________________________________ 28

Appendix- Survey________________________________________________ 30
ABSTRACT:

We looked for ways to predict Penn Sophomores’ SAT scores by examining 17 different explanatory variables that we believed could be significant. The explanatory variables pertained to the student academic and social lives both in high school and during their freshman year of college. All data was gathered through a survey of 79 students. We found that none of the variables had a correlation above .5 with the SATs. In addition, by performing multiple regressions, we found that it is very difficult to predict SAT scores with accuracy; the step wise regression, which yielded our highest adjusted $R^2$ value, only yielded a value of .2254. Therefore, either there are explanatory variables that were missing from our survey, or SAT score really is primarily a test of aptitude as ETS suggests.

MOTIVATION FOR RESEARCH:

There have been many studies in the past decade about whether or not the SAT is a valid test. People question the fairness of the test on several grounds; some people say that the test is racially or socio-economically biased, others question whether or not the test is indicative of any future academic success. It is this second question that was of interest to us. Many studies have been done to see whether or not the SAT could be used as a predictor of success in college or life; most of these tests found that the SAT was in fact not a good predictor of people’s college record. The question we set out to answer was slightly different however; we wanted to know if there were some aspects of someone’s high school or college record that could be used to effectively predict their SAT score.
If in fact a relationship exists between some facet of academic or social life and people’s SAT score, it would lend credence to the argument that the SAT does not simply measure aptitude. For example if the size of the graduating class in a student’s high school, or where their high school was located in relationship to a big city was a good predictor of SAT scores, then perhaps the SAT is socio-economically biased. Or if it turns out that SAT scores can be determined based on someone’s GPA in high school or in college, that would also speak to the validity of the test as a pure measure of aptitude since grades are not meant to measure aptitude alone but rather some combination of aptitude and effort.

The SAT has come under a lot of criticism through its life; we hope that our study will either exonerate the test by demonstrating the difficulty to predict peoples SAT scores even given lots of information about their high school and college lives, or that our study will yield some sort of predictive formula which will strengthen and hone the argument against the SAT.

**SAMPLING METHOD**

We attained our data through a survey. We conducted a 19 question survey that examined the many variables we thought would or at least could affect high school SAT score. There is a copy of a survey attached as appendix 1.

*Predictors:*

Our primary concern was to make a comprehensive survey. Because we knew that we would not be able to go back and ask questions that we missed again, we dedicated time and effort to brainstorming and discussing possible variables involved.
We had to create a survey that balanced including the most important possible predictors without making the survey too long (students will not stop and fill out a survey that is too long). We also spent time finding ways to quantify each variable, so that it would be possible to perform regressions on them. The following is a discussion of the 17 different predictors selected, along with our hypotheses on the relationship between SAT score and the predictor.

**Male versus Female:** Accusations that the SATs are gender biased have often been encountered in public forums. Therefore we put this question on the survey to account for the variance that may be caused only by the gender of the individual. As part of our aim to quantify all the variables, we made the variable a Bernoulli, with males being a 1 and females being a 0. Our hypothesis is that males will have slightly higher SAT scores, so the variable will have a positive Beta with SAT scores.

**High School GPA on 4.0 scale:** Clearly, this should have some correlation with high school SAT score. A person’s high school GPA should reflect on their intelligence and work ethic in high school, which should in theory also reflect on their SAT score. The question was phrased to invoke unweighted GPA as the response. Because many high schools give additional weight to honors and advanced placement classes, many students have weighted GPAs of over 4.0. Our goal was to eliminate this irregular weighting and put each high school on an equal scale. However, this did not really happen in practice (one individual actually claimed 4.86 GPA on a 4.0 scale). This was one of the areas of our survey that we wish we could have corrected after receiving the surveys back. Our solution to this problem was to curve down there GPA to where we thought it should be; clearly, there is error in this method of correction, but there was no better way to expedite
the matter. Our prediction for this predictor was that as GPA increases, so does SAT score.

*Were you tutored for the SAT’s:* In our minds, there should be a clear relationship in our minds here. We made a Bernoulli variable with yes being 1 and no being 0. Error is introduced because there is only a yes and no; the variable does not take into account magnitude. A three-week course at the Princeton review has the same value as a year long private tutor. In addition, buying a book and studying on your own is not considered tutoring while it may be far better than going to one Princeton review course. These factors may negatively affect the variable’s ability to predict SAT scores. However, we still believe it will be a helpful predictor and that a value of one will equate to a higher SAT score.

*Number of High School Clubs with Major Involvement:* This was one of the variables we believed should make a difference but we are not sure if it actually does. We believe that as number of clubs with major involvement goes up, the SAT score will decrease. Our reasons are twofold. First, as number of clubs increases, time to study for the SATs should decrease. Second, SAT score plays less of a factor in their admittance to Penn if they are very involved people. The unavoidable error in this question is the word “major”; every person answering this question probably interprets this word differently and answers accordingly.

*Number of Hours of Community Service per Month:* Our hypothesis on this is similar to that of number of clubs. We believe that as number of hours of community service goes up, your time to study for the SATs goes down, and the importance of the SATs to the admissions committee goes down as well.
Number of Varsity Sports in High School: This is similar to the last two predictors. Following the same logic, as number of varsity sports goes up, SAT score should go down.

Instrument Played in High School: This variable helps us to gain further insight into what kind of person they were in high school and what kind of application package they presented. This variable was made into a Bernoulli variable with Yes being 1 and No being 0. This variable should have a negative Beta with SAT scores.

Public or Private High School: There also have been many debates on campus over public versus private schooling. The prevailing belief is that private high schools give kids the personalized attention and resources to succeed. Therefore they should have higher SAT scores. We are going to disagree with this statement; there should be no significant difference between public and private high schools, and the Beta and correlation value should be close to zero. We made this predictor a Bernoulli with private being a 1 and public being a zero.

Number of Students in Graduating Class: Although the relationship is not obvious between this predictor and SAT scores, we felt that there might be a relationship and thus decided to include it. Our forecast for this predictor is that as the number of kids in a graduating class increase, the SAT score goes down. This is because we feel a more personalized education will yield higher SAT scores.

Number of Miles from Large City: We share the general belief that location makes a difference. In general, more urban areas are wealthier or at least are more educationally focused. Although we realize this is not necessarily true (there are certainly terrible education systems in cities and close to cities), we felt it was an interesting theory to
consider. We also believe that people that are very far from the city will be more unique in the applicant pool (considering Penn is composed mainly of suburban and city students), and thus will require lower SATs to be admitted. We took special care to make this question quantitative by asking number of miles. This question is coupled with the next one; we wanted to be able to make a judgement on if the given city is actually large with respect to the rest of the sampling pool. Our notion is that as miles from a large city increases, the SAT score will go down.

*The Big City:* This variable did not enter our data sheet since it is not a quantitative variable. We used this information on a case by case basis. If we felt the city given was not large enough, we adjusted the individual’s miles to closest city using Mapquest.com. Although this introduced some estimation and subjectivity, we thought it improved our data overall.

*Number of Hours Spent on Homework Per Week:* This predictor should have an interesting relationship on SAT score. On one hand, the number of hours worked per week is an indication on how hard working a student really is, and thus will reflect how hard working and prepared they were for their SATs. Conversely, the number of hours spent on homework could also mean the student had more time to study and work for the SATs. Further, if the SATs are truly a measure of aptitude, then a person who did less work in high school may be smarter and thus would have gotten a better SAT score. We, however, do not believe that the SATs are strictly an aptitude test or that people who worked harder couldn’t find time to study for the SATs; therefore, our proposition is the first one: Number of hours of work will have a positive correlation with SAT score.
Cumulative GPA after Freshman Year: The College Board has sold the SAT score to the country as a valid predictor of college academic success. There is no better measure of academic success in college than your GPA. There was much more success with this variable than high school GPA, since everyone at Penn has the same weighting on their GPA’s here. Our obvious prediction is that as college GPA rises, so will SAT score.

Home School: We deemed this important factor to consider because of the different selectivity of each school. There is little argument that Wharton is a lot harder to get into than Engineering or College of Arts and Science. But the distinction is harder to make between engineering and college. Quantifying these differences was a daunting and subjective task, especially if nursing was added to the equation. We decided to set up a Bernoulli variable, with Wharton being equal to 1, and the college and engineering school being equal to 0. We didn’t feel the difference between engineering and college SAT scores would be that great. We decided not to interview nursing students. This was because the variable did not seem accurate to scale if they were given a 0 or a -1 value; moreover, there are so few nursing kids that we would not have enough data for it to matter. We believe our predictor will support that Wharton has a higher SAT score than the college or engineering.

Varsity Sport at Penn: The Daily Pennsylvanian among other sources has reported that Varsity athletes have more lax standards for getting into school. Agreeing with this statement, Varsity sports involvement should lower the predicted GPA. Unlike the high school varsity sports question, we did not believe it was possible to play more than one Varsity sport at Penn. This assumption did not create any problems while surveying. A Bernoulli variable was used for this predictor with 1 being yes and 0 being no.
Member of the Greek System: This predictor provides a look into the character of a student. It is a safe generalization that Greek members party more than their non-Greek counterparts. We believe that the more focused non-Greek member will have been more studious in high school, and thus would have had a higher SAT score. This predictor also fit nicely as a Bernoulli.

Hours of work a week: Similar to the homework per week in high school, this variable captures the same hard work ethic that we believe should be reflected in SAT scores. We believe that the more hours of work a week, the greater the persons SAT score. Examining the survey results, it appears we have some error in this question. It looks like some people thought this question asked number for hours spent at a job, and not hours spent on homework.

Hours of Community Service per Month: Hours of community service in college probably should not have all that high a correlation with SAT score, but we felt we should include it to be thorough. It also provides an interesting separate study that relates community service hours in high school to college. As we expected, hours of community service in college significantly decreased from the average high school level.

Method of Data Collection

We realized early on that collecting data was going to be a critical aspect of our study. Knowing that the type of data desired was not easily attainable, we had to gather it ourselves. The first thing that we needed to decide was what type of students our sample was going to consist of. We decided we did not want to deal with the variation of different classes, so we chose to focus on sophomores because they were not too far removed from high school but far enough removed that they had a year of college grades
to use in the analysis. They also had a year of other college activities such as sports and Greek life.

Once it was decided that sophomores were to be used, a method of picking sophomores to use in the study needed to be determined. We ended up choosing to survey people as they walked down locust walk, we felt that this was the method that would result in the least sampling bias. We had considered two other possibilities. First we thought that we could select students at random from some list of all the sophomores at the university. Initially this seemed like a good idea, however we soon realized that although this simple random sample of the class of 2004 was the most randomized way to get the people, it would be very hard to get the chosen people to fill out a survey. This technique would require us to track these students down and would probably lead to a situation in which they were unwilling to fill out the survey and we would have to find some other random student to fill it out instead. Therefore, although the simple random sample approach would have probably been the best had we been able to accomplish it, we decided that it was not a feasible sampling method. Another possible way to get the surveys filled out was just to ask people we knew to fill them out. Since it would not involve approaching any strangers and asking them to fill out some fairly personal information, this would clearly be the easiest way for us to do it. Unlike the simple random sample method, however, the resulting responses will not be random at all.

The chosen sampling method was to survey people as they were walking down locust walk. We thought that this would be the most feasible way to get a relatively random sample of Penn Sophomores. We chose locust walk because almost all students walk it. Even thought the library, gym, or a dining hall might have been an easier and
warmer way to sample, it would not have been a good idea because only certain types of student frequent these locations. We printed up 120 surveys, bought two clipboards, and headed down the walk. As we mentioned earlier, we knew that some of the information that we were asking for was somewhat personal, so we stressed the confidentiality of the survey. All students were asked to put their responses into either of our two envelopes themselves, so we could not associate any information with any individual. This was particularly important because we worried about an insecurity bias. If students felt we might read their information, they might be encouraged to lie and inflate their numbers. Realizing incomplete surveys might be a problem; we stressed the importance of completing the entire survey. As an additional precaution, we also moved the more important questions to the top of each section. We feel that our sample was fairly well done. There was some bias in that we ended up asking some people who we knew to fill out the survey just because they were the ones that were most likely to stop and actually do it. In general, the sampling biases and errors were minimized.

**OUR PREDICTIONS**

We have stated in the previous sections our predictions for the relationship of each individual variable to SAT score. In this section, we discuss our hypotheses for which variables will be most important to the regression.

The factors we believe will be most likely be essential are Tutoring for the SAT’s, Number of Varsity Sports, Instruments, Number of Miles from Large City, Cumulative GPA after Freshman Year, Home School, and Varsity Sport. Tutoring for the SATs was picked for obvious reasons (or Princeton Review is a gigantic scam). Number of Varsity Sports and Instrument were picked as the best gauge of how busy a student was in high
school and strength of the rest his college application. Number of Miles from a Large City was picked to estimate the rural effect on SAT scores. Cumulative GPA was picked because it is a relatively accurate prediction of intelligence and hard work ethic; we perceive these as the two most important factors in determining SAT score. Home School was picked because the entrance requirements for the different schools are different; thus, the SAT scores should also be affected. Finally, Varsity athletes are believed to have different standards of entrance than regular students, and therefore, there should be a strong relationship between these two factors.

The rest of the factors we do not believe will be as important in the predicting equation. Some of the predictors were omitted because we believe that they are not as related or other predictors measure the same general concept better (Community Service Hours, Male versus Female, Public versus Private, Students in Graduating Class, Hours of Homework in High School, Greek system). Alternatively, there are other variables that will have a higher sampling error and thus the relationship will suffer (High school GPA, Number of Activities, and Number of Hours of Work per Week in College).

As far as the overall multiple regressions go, we do not expect very high adjusted $R^2$. This is because the SATs are supposed to be an aptitude test, and hence a person’s score should essentially come down to intelligence and general knowledge. None of the factors on our survey can perfectly measure these two factors. But although it will be impossible to perfectly predict the SAT score of Penn students, we believe that we can at least make a decent prediction from the data collected in this survey.
ANALYSIS OF DATA

SAT SCORES

Median – 1430
Mean – 1443.7
Std Dev - 77.3

SAT scores is the clearly the most important of our variable; it can be seen from this histogram that the gathered SAT are rather normal. They seem to be unimodal and not particularly skewed in one direction or another. Interestingly, our median SAT score is not far from the published median of 1420 for incoming freshman this fall. This suggests that as far as SAT scores are concerned our sample is representative of the population.

HIGH SCHOOL GPA

Median – 3.9
Mean – 3.8
Std Dev - .21

Correlation with SAT Score: .1410

From this histogram of High School GPA it can be seen that the data is fairly skewed to the left. There are a couple possible explanations for this. First of all most people at Penn just had really high GPAs in high school. Another reason is that some people’s GPAs were not done on a 4.0 scale so that might have thrown off the results also since they might not have known how to translate their GPA onto a 4.0 scale. We realized that this was a skewed variable and tried running a logarithmic transformation on it. This did not really help, however, because the extreme values were still the extreme values, and the data was still skewed. The correlation with the SAT Score was actually lower (.1301) when we ran the regression using the logarithmically transformed data. We also tried to change the data into a categorical variable but to no avail as the data remained skewed to the left as almost everyone fell into the highest category. The correlation between High School GPA and SAT Scores is .1410 which is not very high at all. This might be due to our error in rounding down GPAs that were above 4.0 as discussed earlier in the paper. This correlation might also be low because GPA in high school is so related to the difficulty of the high school, which is clearly not constant.
TUTORED FOR THE SAT

Correlation with SAT Score: .0191

Of the people surveyed, 43 were not tutored in any way for the SATs while 36 of them were. The correlation between this variable and the SAT score is only .0191 which is somewhat surprising giving its ostensibly linked nature. Our predictions for the strong relationship appear wrong in this case.

NUMBER OF CLUBS IN HIGH SCHOOL

<table>
<thead>
<tr>
<th>Median</th>
<th>Mean</th>
<th>Std Dev</th>
</tr>
</thead>
<tbody>
<tr>
<td>4</td>
<td>4.9</td>
<td>5.4</td>
</tr>
</tbody>
</table>

Correlation with SAT Score: .1925

The data for the Number of Clubs that the students were a member of in high school is relatively normal as can be seen from the histogram at right. There is one extreme outlier at 47, which might be attributed to false information given on the survey. The correlation between the number of clubs and the SAT Score is .1925 which is relatively high compared with some of the other correlations in the study. This correlation actually goes down when the outlier is removed from the data set, so we decided to leave it in. Our prediction appears to be wrong since the positive correlation suggests that as number of clubs increases so does SAT score.

NUMBER OF VARSITY SPORTS PLAYED IN HIGH SCHOOL

<table>
<thead>
<tr>
<th>Median</th>
<th>Mean</th>
<th>Std Dev</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1.3</td>
<td>1.1</td>
</tr>
</tbody>
</table>

Correlation with SAT Score: -.2568

The data for the Number of Sports played in High School by the surveyed students appears to be relatively uniform. One is the most popular answer by a slim margin. Perhaps the most interesting part about this data is that it is negatively correlated with the students SAT Scores. That is, as the number of Varsity sports increases the SAT Score tends to decrease. This might be because athletics is factored into the admissions decision at Penn. The varsity sports at Penn explanatory variable will probably shed more light on this result.
NUMBER OF HOURS OF COMMUNITY SERVICE PER MONTH IN HIGH SCHOOL

Median – 10
Mean – 11.2
Std Dev – 9.9

Correlation with SAT Scores: -.1114

The data for the Number of Hours of Community Service that the students surveyed did per month in high school appears to be fairly normal with one mode around 10. There is a large number of people at the extreme (0) similar to some of our other data. The correlation between hours of community service and SAT Scores is also negative implying an inverse relationship. These inverse relationships might be due to the fact that people can get into Penn with lower SAT scores if they have other stuff going for them such as varsity sports or community service. The correlation is weak however, implying that people that do lots of community service by no means do significantly worse than people that do not do lots of service.

PLAY AN INSTRUMENT IN HIGH SCHOOL

Correlation with SAT Score: .1458

The data for whether or not the studentsPlayed an Instrument in High School is a Bernoulli variable similar to the SAT tutor variable. The numbers are exactly the same as well. 43 people did not play an instrument and 36 people did. The correlation between instrument playing and the SAT Score is also rather small at .1458. This correlation contradicts our prediction for the explanatory variable.
PUBLIC OR PRIVATE HIGH SCHOOL

Correlation with SAT Score: -0.2860

The data for Public and Private high school attendance is also a Bernoulli variable. Private high school is represented by a 1 and public by a 0. Of the 79 people surveyed. There were 53 people that attended public school and 26 people that attended private school. The most interesting aspect of this data is that the correlation with SAT scores is negative implying that people that went to public school tended to do better on the SATs.

NUMBER OF PEOPLE IN HIGH SCHOOL GRADUATING CLASS

Median – 300  
Mean – 342.8  
Std Dev – 268.7

Correlation with SAT Score: 0.0748

The data for Number of People in the student’s Graduating Class from high school appears to be rather normal however it is hardly correlated at all to the students SAT Scores. In fact if the outlier (1600) is removed and the regression is run again the r value drops to 0.018. This correlation value seems to contradict our prediction for the variable. However since the correlation is minutely positive, it does not definitively disprove our prediction.

MILES FROM HIGH SCHOOL TO LARGE CITY

Median – 20  
Mean – 29.2  
Std Dev – 28.5

Correlation with SAT Score: 0.1417

The data for Distance from a Large City is similar to some of our other data in that it is rather normal but has a collection of people at the extreme, in this case 0. The correlation between miles from a city and SAT scores is 0.1417; the fact that it is positive implies that contrary to our predictions the further the students were away from a city the better they did on the SATs. This correlation is weak but breaking the data down into a Bernoulli variable can strengthen it. We realized that an area can only be considered urban/suburban if it is within a distance from a large city. So, we decided that a more
indicative variable might be a Bernoulli that is 1 if the students went to High school within 30 miles of large city and 0 if they did not. We also thought a Bernoulli categorical variable would be best because the difference between living 10 and 20 miles away from a city is far less significant than whether it is an urban/suburban or rural setting. We ran a regression on this new Bernoulli variable and found that in fact the correlation increased to -.2192. The reason why this correlation is negative is because 1 represents close to a city and 0 represents far from a city in the transformed data. The opposite is the case in the original so the correlation is now negative. Since the correlation was stronger on the transformed variable, we decided to use that in our multiple regression analysis.

**HOURS OF HOMEWORK DONE A WEEK IN HIGH SCHOOL**

- **Median** – 10
- **Mean** – 10.6
- **Std Dev** – 7.2

**Correlation with SAT Score:** -.0199

The data for Amount of Time Spent on Homework in High School appears to be rather normal with a mode at 10. The correlation for this data, however, is so low that no conclusions can really be drawn about SAT scores from it at all.

**FRESHMAN YEAR GPA**

- **Median** – 3.6
- **Mean** – 3.5
- **Std Dev** – .34

**Correlation with SAT Score:** .4104

The data for Freshman Year GPA is rather normal although it does seem to have two modes at 3.6 and 3.25. The correlation between freshman year GPA and SAT Score is by far the highest of any of the variables that we tested. The correlation is positive which means that the better you did your freshman year in college the better you probably did on your SATs. Interestingly the correlation between freshman year GPA and SAT scores is higher than the correlation between High School GPA and SAT scores.
VARSITY SPORT IN COLLEGE

Correlation with SAT Score: -.1072

The data for whether or not the students played a varsity sport their freshman year in college is not very useful because we did not survey enough people to get enough people that actually played varsity sports. Only 5 of the 79 people surveyed played a varsity sport which makes the data pretty meaningless. For the minimal data we have there was a negative correlation, so this explanatory variable should probably be explored further.

HOME SCHOOL AT PENN

Correlation with SAT Score: .1517

The data for Home School is also a Bernoulli variable. 25 of the students surveyed were in Wharton and 54 were in either the college or the school of engineering and applied science. The fact that the correlation is positive suggests that Wharton students tended to do better on the SATs then non-Wharton students. This correlation is not very high however.

MEMBER OF THE GREEK SYSTEM

Correlation with SAT Score: -.1285

The data for whether or not the students are a member of the Greek system at Penn is a Bernoulli variable with 1 being membership and 0 being non-members. Of the 79 students surveyed 38 were in the Greek system and 41 were not. The slightly negative correlation hints that perhaps students in the Greek system did not do as well on their SATs.
**HOURS OF WORK DONE A WEEK FRESHMAN YEAR**

<table>
<thead>
<tr>
<th>Median</th>
<th>Mean</th>
<th>Std Dev</th>
</tr>
</thead>
<tbody>
<tr>
<td>15</td>
<td>15.6</td>
<td>10.0</td>
</tr>
</tbody>
</table>

**Correlation with SAT Score**: -.1145

The data for Hours of Work Done a Week Freshman Year is rather normal with a mode at 20. The most interesting aspect of this data is that the correlation is negative, which implies our predictor was erroneous. This implies that people that did more work freshman year tended to do worse on the SATs. Perhaps this is indicative of the SATs ability to test for aptitude and not preparation. However, this correlation is not very high. Also, as discussed in the Sampling Method section, we believe that there may have been sampling error caused by people interpreting the question wrong. Therefore, the correlation should not be taken too seriously.

**NUMBER OF HOURS OF COMMUNITY SERVICE PER MONTH IN COLLEGE**

<table>
<thead>
<tr>
<th>Median</th>
<th>Mean</th>
<th>Std Dev</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>3.6</td>
<td>5.1</td>
</tr>
</tbody>
</table>

**Correlation with SAT Score**: .0037

The data for hours of community is heavily skewed because of the large number of people that do zero hours a month. The correlation for this data indicates that it has no bearing at all on SAT scores.

**MALE/FEMALE**

**Correlation with SAT Score**: -.1049

The data for male and female is a Bernoulli variable with males being assigned a 1 and females assigned a 0. Of the 79 people surveyed 45 of them were male and 34 were female. The slight negative correlation implies that females are slightly more likely to outperform males on the SAT. This contradicts feminist’s claims and our predictions.
MULTIPLE REGRESSION ANALYSIS

STEPWISE REGRESSION

Since we have so many explanatory variables, we decided to perform a stepwise regression as a way to determine which predictors are most significant, and should therefore be included in the regression. The P-value to enter was set at .15. And the P-value to leave was set at .15. Four predictors were selected. The first predictor to be selected was Freshman Year GPA. Second was Public or Private high school. The third predictor selected was Number of Varsity Sports in High School. The final predictor with under a .15 p-value was Number of High School Clubs. The amount variation in SAT scores that can be accounted for by these explanatory variables is .2254 (adjusted R²). Because our adjusted R² is low, we know that there is either something else creating the variability, or SAT scores are truly random (as ETS wants you to believe); this idea will be discussed in greater detail in our conclusion section.

As we can see by the graph on the right, a good portion of the data does not fall in the 95 percent confidence interval. But the regression still has some predictive value. The predictive equation yielded by this regression is:

\[
\text{SAT SCORE} = \ldots
\]
As we can see from the table below, the Betas for the different explanatory variables are quite significant. The P-values are all significant to the .13 level. The Betas for Number of Varsity Sports in High School, Public and Private High School, and Freshman Year GPA are all significant to the .07 level. With the Freshman Year GPA’s Beta is actually significant to the .04 level.

| Term                  | Estimate | Std Error | t Ratio | Prob>|t| | Lower 95% | Upper 95% |
|-----------------------|----------|-----------|---------|------|---------|-----------|
| Intercept             | 1214.53481 | 87.3937875 | 13.8972671 | 2.5511e-22 | 1040.39889 | 1388.67073 |
| # of High School clubs| 2.2734652 | 1.46060511 | 1.55652283 | 0.12385056 | -0.636854 | 5.18378441 |
| # of Varsity Sports   | -13.73654 | 7.34997512 | -1.8689233 | 0.06559085 | -28.381685 | 0.90860527 |
| Public or Private HS  | -32.121935 | 16.9095439 | -1.8996334 | 0.06137813 | -65.814937 | 1.57106659 |
| Freshman year GPA     | 70.5068841 | 23.6743107 | 2.97820219 | 0.00391842 | 23.3347903 | 117.678978 |

There are also some interesting conclusions that can be drawn about SAT scores looking at the 95 percent confidence region of the Betas. We can be 95 percent certain that the Beta for freshman year GPA is positive. Since the Beta is essentially the relationship between the explanatory variable and SAT score, we can be almost certain that there is a relationship when the 95 percent confidence interval is either completely positive or negative. We can almost be 95 percent confident that the Betas for the other three are either positive or negative. We say this because the 95 percent confidence interval barely crosses the zero for each of these explanatory variables.

Finally, we examine the residual plot (see top of next page) for the data to determine if there is any heteroscedasticity. There does not appear to be any pattern.
showing up in the residuals (for example higher SAT scores having a larger variance), so
the variance is relatively constant and heteroscedasticity does not appear to be a problem.
The residuals do not lead us to believe that a non-linear model would be a better fit.
Looking at the residual distribution, the residuals appear to be normally distributed. The
Normal Quantile plot (below) further confirms this. Independence appears to be a
reasonable assumption. All Gauss Markov assumptions hold in the model.

PREDICTED REGRESSION

In our predictions section, we thought the following explanatory variables would
be significant to predicting SAT score:

- Tutoring for the SATs
- Number of Varsity Sports in High School
- Instrument Playing
- Number of Miles from a Large City (We will use the Bernoulli transformed data discussed earlier because this has a higher correlation with SAT scores)
- Cumulative GPA after Freshman Year
- Home School
- Varsity Sport

Running a regression with the aforementioned predictors yields an adjusted $R^2$ of .1723. The stepwise regression yielded a higher adjusted $R^2$ of .2254. Therefore, more of the SAT score variation is accounted for by the variation in the explanatory variables chosen by the stepwise method. Even though it is a less affective predictor of SAT scores, we will still examine it because it was our prediction.

Again, we can see that a lot of the data points do not fall within the 95 percent confidence lines. Looking at the different Betas, only the Bernoulli distance and the Freshman year GPA is significant to the .15 level. And there are a bunch of Betas that do not appear to be significant at all. Tutoring for the SATs, Playing and Instrument, and

| Term               | Estimate | Std Error | t Ratio | Prob>|t| | Lower 95% | Upper 95% |
|--------------------|----------|-----------|---------|-------|-----------|-----------|
| Intercept          | 1220.6353| 92.12905  | 13.25   | <.0001| 1036.9352 | 1404.3354 |
| Tutor for the SAT  | 12.293091| 18.12265  | 0.68    | 0.4998| -23.84244 | 48.428625 |
| # of Varsity Sports| -10.89808| 7.856366  | -1.39   | 0.1697| -26.56323 | 4.767026  |
| Play an Instrument | 12.954792| 16.6306   | 0.78    | 0.4386| -20.20568 | 46.115269 |
| Bernoulli distance | -29.70375| 17.45991  | -1.70   | 0.0933| -64.51782 | 5.1103164 |
| Freshman year GPA  | 68.778852| 24.74238  | 2.78    | 0.0070| 19.443941 | 118.11376 |
| Home School        | 20.252941| 18.0855   | 1.12    | 0.2666| -15.80851 | 56.314397 |
| Varsity Sport      | -23.89013| 32.81391  | -0.73   | 0.4690| -89.3192  | 41.53894  |
Varsity Sport in College are among these in that they all have P-values greater than .4. As far as being 95 percent confident that the true value of Beta is either positive or negative, Freshman Year GPA is the only explanatory variable which has this property. The Bernoulli distance is close, but zero and some positive values are included in the confidence interval. Surprisingly, Number of Varsity Sports in High School (one of the predictors chosen in the stepwise regression) has become less significant. But it is still the third most significant explanatory variable in this regression.

We will not do further analysis into this regression, since it is not as significant a regression as the one created by the stepwise function.

**REGRESSING ALL THE FACTORS TOGETHER**

One final interesting regression is to try to incorporate all the data we collected into one comprehensive model that predicts SAT scores. The adjusted $R^2$ is .15937, which accounts for the least amount of variance.

It is interesting to see that the explanatory variables picked by the stepwise regression function are the only variables that are significant in this regression. Number of Varsity Sports Played in High School, Public or Private High School, and Freshman Year GPA are all significant to the .10 level. However, the Number of High School Clubs is now only significant to the .2933 level. The fact this predictor is similar to other predictors include in this regression led us
to believe that multicolinearity could be a problem. We believed this predictor was probably highly correlated with one or some of the other predictors, and that was driving down the T ratio. In order to test this belief, we regressed the other predictors against number of high school clubs. We found that none of these correlations were particularly high (none above .25). Although there is not a textbook case of multicolinearity (finding a correlation around .9), multicolinearity might still be a factor when you consider the low correlations to SAT score. But, we can safely conclude that multicolinearity is not severe in this case, and although it might be a factor, there very well could be something else.

| Term                               | Estimate | Std Error | t Ratio | Prob>|t| | Lower 95%   | Upper 95%   |
|------------------------------------|----------|-----------|---------|---------|-------------|-------------|
| Intercept                          | 1291.0064| 190.0057  | 6.79    | <.0001  | 911.06661   | 1670.9462   |
| High School GPA                    | 7.8775327| 42.91064  | 0.18    | 0.8550  | -77.9276    | 93.682669   |
| Tutor for the SAT                  | 20.735396| 19.85762  | 1.04    | 0.3005  | -18.97237   | 60.44316    |
| # of High School clubs             | 1.8681832| 1.762346  | 1.06    | 0.2933  | -1.655845   | 5.3922113   |
| # of Varsity Sports                | -15.43002| 8.558267  | -1.80   | 0.0763  | -32.54334   | 1.6832887   |
| Hours of Community                 | -0.889325| 0.88318   | -1.01   | 0.3179  | -2.655352   | 0.8767031   |
| Service/Month                      | 16.686732| 17.92245  | 0.93    | 0.3555  | -19.15141   | 52.524877   |
| Play an Instrument                 | -36.9993 | 20.40013  | -1.81   | 0.0746  | -77.79189   | 3.7932804   |
| Public or Private HS               | -0.031566| 0.037391  | -0.84   | 0.4019  | -0.106334   | 0.0432026   |
| Students in Graduating Class       | -19.23788| 19.15171  | -1.00   | 0.3191  | -57.53409   | 19.058329   |
| Bernoulli distance                 | -0.076553| 1.302589  | -0.06   | 0.9533  | -2.681242   | 2.5281353   |
| Hours a week spent on HW in HS     | 49.230314| 28.43143  | 1.73    | 0.0884  | -7.621843   | 106.08247   |
| Freshman year GPA                  | 24.30192 | 18.9875   | 1.28    | 0.2054  | -13.66594   | 62.269782   |
| Home School                        | -16.64436| 35.7283   | -0.47   | 0.6430  | -88.08751   | 54.798786   |
| Varsity Sport                      | -7.792842| 17.69621  | -0.44   | 0.6612  | -43.17859   | 27.592908   |
| Greek System                       | -0.672015| 0.914458  | -0.73   | 0.4652  | -2.500587   | 1.1565576   |
| Hours of work per week             | 1.6074104| 1.77732   | 0.90    | 0.3693  | -1.946561   | 5.1613817   |
| freshman year                      | MALE/FEMALE| 17.14247 | -0.79   | 0.4298  | -47.9025    | 20.654493   |
FINAL THOUGHTS

The ETS, the company that administers the SATs would probably be happy with our findings. They have always claimed that the SAT is a pure measure of aptitude. Because we were not able to find a multiple or single regression with a very high adjusted \( R^2 \), our results support this claim. No combinations of our explanatory variables were able to predict SAT score all that significantly. It is possible that there are other explanatory variables not included in our survey that better predict SAT score. But, we feel this is unlikely given the number and variety of factors we tested.

Furthermore, the SAT was created as a measure of success in college. This statement is also supported by our results in that Freshman Year GPA had by far the highest correlation with SAT score.

INTERESTING SIDE NOTES

There were some other interesting questions that we had going into the project. We have decided to discuss them briefly in this section. We will test the following using hypothesis testing.

First we wanted to test the validity of the Princeton review and other tutoring services. We wanted to see if they really made a difference in SAT scores in the Penn population. Our null hypothesis is that tutoring does not make a difference. And our alternate hypothesis is that the mean of tutored scores will higher than the mean of untutored scores \( (u_1 = \text{Tutored SAT score and } u_2 \text{ is equal to untutored SAT score}) \).

\[ H_0: u_1 - u_2 = 0 \]
\[ H_a: u_1 - u_2 > 0 \]
We chose these hypotheses because we wanted to minimize this type of Type I error. The type I error in this case is falsely concluding that tutoring is beneficial when it is in fact not significant. Because we do not know the population standard deviation and the N’s for both samples are not greater than 40, we will use the T statistic and a one sided test.

For the sample that wasn’t tutored the distribution looks normal (see histogram on the right). The mean of the sample is 1442.3256. The N equals 43. The Standard Deviation of the sample is 78.219387.

For the sample that was tutored on the SATs, the data once again looks relatively normal (see histogram on left). The mean for this sample is 1445.2778. The N is 36, and the sample standard deviation is 77.256016.

Just looking at the sample means and standard deviations, we can see right away that the difference is not going to be significant. The difference between means is only about 3, and the standard deviations are much larger, so it will not be significant to the .05 level. But we will calculate the T statistic anyway. The T statistic for this sample is 0.16819622418269. There is really no point in calculating degrees of freedom and a 95 percent significant T value to compare it to because we know that .168 will lower. Therefore, we fail to reject the null hypothesis. Consequently, in the Penn population, getting tutored for the SATs does not significantly raise your score.
A second interesting side note and the final one we will examine is the Greek systems SAT score difference. There is has been a lot written up about Greek students being less intelligent and motivated than non-Greek students. The Daily Pennsylvanian even did a series of articles on this concept. But we have never seen statistical proof of this. We will attempt to statistically prove this statement using hypothesis testing. Our null hypothesis will be that Greek members will have the same mean SAT score. Our alternate hypothesis is that non-Greek students will have a higher mean SAT score. $u_1$ = non-Greek SAT score and $u_2$ is equal to Greek SAT score.

$H_0: u_1 - u_2 = 0$

$H_a: u_1 - u_2 > 0$

We chose these hypotheses because we wanted to minimize the error of saying Non Greeks have a higher SAT score; we figured we should give the Fraternities and Sororities benefit of the doubt. And thus we made it type I error. Again, we do not know the population standard deviation, and the Ns for both samples are not greater than 40 so we must use a T statistic. Again, The set of hypothesis lends it self to a one sided test.

For sample 1 (non Greek SAT score) shown on the right we see that the data is relatively normal. The mean value of the sample is 1451.4286 and the standard deviation of the sample is 84.811741. The N is equal to 42.

The second sample (Greek SAT
score) also looks relatively normal from the histogram on the left. The mean and
standard deviation of the sample are 1433.4211 and 67.514816 respectively. The N is 37.

The T statistic is 1.0494 from Jumpin. Even though this T statistic is higher than
the last one, we still do not feel it will be significant. Without calculating the actual T
statistic that we should compare it to, we know that we do not have to. The lowest
possible T statistic we could compare it to would the one attained with a Z (if degrees of
freedom were infinite). For the .05 level, this is 1.645. Because our calculated statistic is
less than this, we know that it cannot be significant to the .05 level.

Besides these two interesting side notes, there are others that would also be
interesting to investigate with our data set. For example, you could see if the difference
in means of SAT scores between the different home schools is significant, or if Varsity
Sports really have a significantly lower mean SAT score than non Varsity athletes. We
will discuss other opportunities in our further research section.

POSSIBILITIES FOR FURTHER STUDY

Although our data was collected specifically for use in this study it could be
applied to a wide variety of other interesting questions. The data we collected consisted
mostly of general academic information from sophomores at the University of
Pennsylvania and could therefore be used to draw other conclusions about the academics
of this group.

There are a couple of interesting applications for this data outside the scope of our
report. First, the data could be used to draw conclusions about how people’s activities in
High School affected their High School GPA. Such and analysis could answer questions
like “do people that play instruments in High School tend to have Higher GPAs” etc.
This application of these statistics is almost identical to the one outlined in this report except the variable trying to be predicted would be High School GPA. The same process could be applied to college GPA; that is “how does someone’s college life affect their college GPA.”

Perhaps more interestingly however, one could use the data to try to draw conclusions about how high school information affects college GPA. Many critics of the SAT have argued that GPA in college is not reflective of that person’s SAT scores, but maybe it is indicative of something else. Perhaps students that were more active in their community in High School tend to have higher GPAs in college for instance.

Another direction which our data could be taken which would require some additional information being collected would be to look at how the High School atmosphere that a person was in relates to how they did on their SATs or what their freshman year GPA was. In order to do a thorough analysis of this, we would have to collect more specific data about the high schools that people went to. Only a rudimentary analysis could be done, however, using the data we have already collected regarding the size of the people’s graduating class, the proximity of the school to a large city, and whether or not the school was public or private. With more data it is possible that SAT scores could in fact be predicted better.
APPENDIX:

Please Complete all Questions

Sex:  M  F

PRE PENN:

SAT Score (Highest Combined) ____________
High School GPA (on 4.0 scale) ____________
Were you in any way tutored for the SATs (Y/N) ____________
Number of High School clubs you had major involvement in ____________
Number of Varsity Sports played in high school ____________
Number of hours of community service per month in high school ____________
Did you play an instrument in High School (Y/N) ____________
Did you go to a public or private high school ____________
Number of students in your graduating class in high school (estimate is fine) ____________
Number of miles from your high school to a large city ____________
Which city ____________
Number of hours a week spent on homework in high school ____________

AT PENN:

Cumulative GPA after freshman year ____________
Which school are you a member of (Wharton, college, etc.) ____________
Do you play a varsity sport ____________
Are you a member of the Greek system ____________
How many hours of work do you do a week ____________
How many hours per month of community service do you do ____________

Thank you for your time