An Investigation of the Fit of Linear Regression Models to Data from an SAT® Validity Study

By Jennifer L. Kobrin, Sandip Sinharay, Shelby J. Haberman, and Michael Chajewski
College Board Research Report 2011-3  
ETS Research Report RR-11-19

Jennifer Kobrin is a research scientist at the College Board.

Sandip Sinharay is a principal research scientist at ETS.

Shelby J. Haberman is a Distinguished Presidential Appointee at ETS.

Michael Chajewski is an assistant research scientist at the College Board.

Researchers are encouraged to freely express their professional judgment. Therefore, points of view or opinions stated in College Board reports do not necessarily represent official College Board position or policy.

The College Board

The College Board is a mission-driven not-for-profit organization that connects students to college success and opportunity. Founded in 1900, the College Board was created to expand access to higher education. Today, the membership association is made up of more than 5,900 of the world’s leading educational institutions and is dedicated to promoting excellence and equity in education. Each year, the College Board helps more than seven million students prepare for a successful transition to college through programs and services in college readiness and college success — including the SAT® and the Advanced Placement Program®. The organization also serves the education community through research and advocacy on behalf of students, educators and schools.

Acknowledgments

The authors would like to acknowledge Brian Patterson, Jeff Wyatt, Tom Proctor, and Sandra Barbuti for their assistance with this study. The authors also appreciate the review and helpful comments from Michalis Michaelides, Jerry Melican, Neil Dorans, and Skip Livingston.

© 2011 The College Board. College Board, Advanced Placement Program, AP, SAT and the acorn logo are registered trademarks of the College Board. inspiring minds is a trademark owned by the College Board. All other products and services may be trademarks of their respective owners. Printed in the United States of America.

For more information on College Board research and data, visit www.collegeboard.org/research.
4. Mean and Median Standardized Residuals of the Regression of FYGPA on HSGPA, SAT®-CR, SAT-M, and SAT-W by Levels of HSGPA and SAT-W ........................................................................................................................................ 12

5. Mean and Standard Deviation of Standardized Residuals of the Regression of FYGPA on HSGPA, SAT-CR, SAT-M, and SAT-W by Levels of the Predicted Values (\(\hat{Y}\)) ......................................................................................................................... 15

6. ANOVA for the Linear Regression Along with the Lack of Fit ..................... 16

7. Comparison of Regression Models to Test the Significance of the Addition of Nonlinear Terms .................................................................................................................................................. 17

**Figures**

1. Example of a hypothetical nonlinear relationship between SAT and FYGPA .................................................................................................................. 5

2. Boxplots of FYGPA by HSGPA ........................................................................ 10

3. Plot of SAT-W and FYGPA ............................................................................... 10

4. Boxplots of standardized residuals by HSGPA ............................................. 11

5. Plot of SAT-W and standardized residuals ...................................................... 11

6. Normal probability plot of the residuals ......................................................... 12

7. Histogram of standardized residuals............................................................. 13

8. Partial regression plot for HSGPA and FYGPA ............................................. 13

9. Partial regression plot for SAT-W and FYGPA ............................................. 14

10. Plot of predicted values and standardized residuals ..................................... 15
Abstract
This study examined the adequacy of a multiple linear regression model for predicting first-year college grade point average (FYGPA) using SAT® scores and high school grade point average (HSGPA). A variety of techniques, both graphical and statistical, were used to examine if it is possible to improve on the linear regression model. The results suggest that the linear regression model mostly provides an adequate fit to the data and that more complicated models do not significantly improve the prediction of FYGPA from SAT scores and HSGPA.

Introduction
Since the SAT® was first administered in 1926, hundreds of studies have examined the validity of SAT scores for predicting college performance. The vast majority of these studies report correlations between SAT scores and college grades and/or apply multiple linear regression using SAT scores and other measures such as high school grades to predict college grades. In a recent national SAT validity study, Kobrin, Patterson, Shaw, Mattern and Barbuti (2008) investigated the validity of SAT scores for predicting first-year college grade point average (FYGPA) by computing the simple and multiple correlations of SAT scores and high school grade point average (HSGPA) with first-year college grade point average (FYGPA).

Among the recent studies using multiple regression to examine the validity of SAT scores for predicting college performance are Geiser and Studley (2002), Cornwell, Mustard and Van Parys (2008), Norris, Oppler, Kuang, Day and Adams (2006), and Agronow and Studley (2007). In these studies, the authors used SAT scores along with other variables (e.g., HSGPA, demographic characteristics of the students) to predict FYGPA using a linear regression model. In this type of study, it is rarely assessed whether a more complicated model would explain the data better than a linear regression model.

Arneson and Sackett (under review) examined the linearity of ability-performance relationships in organizational selection and higher education admission at the upper region of the ability distribution. Using four large-scale data sets, including one containing SAT scores and first-year college grades from entering cohorts from 1995 to 1997, the authors used three different methods to investigate the shape of the relationship between test scores and criterion measures. These included fitting a smooth curve (i.e., lowess curve) to the scatterplots of the relationships, conducting hierarchical polynomial regression analyses with quadratic terms added to the linear regression model, and conducting separate regression analyses at different test score ranges and comparing the unstandardized regression coefficients across the regions. The results showed that in all four data sets, linearity was at least approximately maintained at the upper region of the distribution.

... it is rarely assessed whether a more complicated model would explain the data better than a linear regression model.
Current Study

Using the data from Kobrin et al. (2008), the linear regression of FYGPA on SAT critical reading, mathematics, and writing scores as well as HSGPA (where all of these variables have been standardized with a mean of zero and a standard deviation of one) is:

\[
FYGPA = \beta_0 + \beta_1 \times SAT_M + \beta_2 \times SAT_{CR} + \beta_3 \times SAT_W + \beta_4 \times HSGPA
\]

\[
\overline{FYGPA} = 0 + .06 \times SAT_M + .07 \times SAT_{CR} + .18 \times SAT_W + .29 \times HSGPA
\]

The sample on which this model was calculated is described in the Methods section of this paper. Let us refer to the linear regression model in Equation (2) as Model 1. The purpose of this study is to investigate whether a regression model that is more general than Model 1 performs substantially better than Model 1 for the data from the National SAT Validity Study (Kobrin et al., 2008). For example, Figure 1 shows a hypothetical plot of SAT score against FYGPA, where SAT score is the sum of the three SAT sections (critical reading, mathematics, and writing). In the plot, the relationship between FYGPA and SAT scores is linear for students earning SAT scores from 600 through 2000. However, for SAT scores higher than 2000, students reach the ceiling of FYGPA, thus producing a nonlinear trend at the upper end of the SAT scale. For such a situation, a regression model including squares of the SAT scores in addition to the terms in Model 1 might perform better than Model 1. Arneson and Sackett (under review) did not find evidence for such a trend for high SAT scores and FYGPA; however, the current study uses a larger data set, including current data on the revised SAT that includes a writing section and fits regression models for the whole score range. Thus, the current study is more general than Arneson and Sackett (under review), and it uses several methods not previously considered.
Evaluating Linear Regression

The adequacy of a linear regression model can be tested using a variety of visual and statistical tools. Regression diagnostics (e.g., Atkinson & Riani, 2000; Belsley, Kuh & Welsch, 1980; Berry, 1993; Chatterjee & Hadi, 1988) are techniques for exploring problems that compromise a regression analysis (Fox, 1991). Many of the most commonly used regression diagnostic methods involve analysis of the residuals from the regression, which are the values representing the difference between each predicted and actual value on the response variable. Analyses of the residuals provide an indication of possible improvements on a linear regression model fitted to the data. If the residuals show a systematic — as opposed to a random — pattern, the addition of terms, for example, the square of an independent variable, may be required.

The most common methods for examining residuals include plotting the standardized residuals against the predicted values, and constructing normal probability or partial-regression plots. A plot of the standardized residuals against the predicted values allows one to evaluate the extent to which the linear regression model is adequate for the data. For example, if the model is adequate, this plot will show a random scatter of points. A curvilinear pattern in the plot would signal that a nonlinear regression model would perform better than the linear regression model, and if the plot does not show an even scatter of points around the mean, it would signal heteroscedasticity.

Normal probability plots show a variable’s cumulative proportions against the cumulative proportion of a normal distribution, where deviations from a linear pattern indicate non-normality. Partial regression plots, also called added-variables plots, allow one to examine whether and how a predictor variable should be included in the model in addition to the other predictor variables. These plots show the residuals of each independent variable on the $x$-axis against the residuals of the dependent variable on the $y$-axis. The residuals in these plots are obtained when the rest of the independent variables are used to predict the dependent and then the given independent variable separately. No pattern (or a random scatter) in such a plot would indicate that the predictor is not needed in the model in addition to the other predictors. A nonlinear pattern in such a plot would suggest that the model should include nonlinear terms involving the predictor.

FIGURE 1. Example of a hypothetical nonlinear relationship between SAT and FYGPA.
The above-mentioned plots usually suggest a necessary remediation of model specifications. For example, a curvilinear pattern in a plot of the residuals against a predictor variable indicates that the investigator should include polynomial terms (squares, cubes, etc.) of the predictor variable in the model in addition to the existing terms. Such models are then compared to the original model using $F$-tests to examine if they provide a significantly better fit and account for the variance in the dependant variable above and beyond the base model.

When the data set has more than one observed value of the response variable for at least some combinations of values of the predictor variables (for example, five individuals, all with different $y$-values, with $x_1 = 1, x_2 = 0.5, ..., x_p = 0$; two individuals, all with different $y$-values, with $x_1 = 1, x_2 = 0.0, ..., x_p = 0.3$ etc.), the lack-of-fit (LOF) test (e.g., Draper & Smith, 1998, p. 47) can be employed to test the hypothesis that a regression model fits the data adequately. In this test, the total unexplained variability in the regression model (the sum of squared errors, or SSE) is divided into two sources of variability known as the pure error variability and the sum of squares (SS) due to lack of fit:

$$
\sum \sum (y_{ij} - \hat{y}_{ij})^2 = \sum \sum (y_{ij} - \bar{y}_i)^2 + \sum \sum (\bar{y}_i - \hat{y}_{ij})^2
$$

The pure error variability is the error due to natural variation in a variable, that is, the differences in the values of the response or criterion variable when they are measured at the same level of the predictor or explanatory variable. The lack of fit sum of squares is the remaining portion of the SSE and is the unexplained error not due to natural variation. The resulting LOF test examines how much of the SSE is a result of lack of fit and determines whether this portion of the SSE is statistically significant (Draper & Smith, 1998; Ott & Longnecker, 2001). If the LOF $F$-statistic is not significant, the regression model can be considered adequate. Otherwise, the model is inadequate and should be refined (for example, by including squares of the predictors in the model). See the appendix for further details on the LOF test.

**Method**

**Data and Sample**

The data were obtained from 110 colleges and universities that participated in the College Board’s National SAT Validity Study (Kobrin et al., 2008). These institutions provided first-year college grade point average (FYGPA) for a total of 195,099 students who finished their first year of college in spring 2007. Official scores from the SAT were obtained from the 2006 College-Bound Senior Cohort database that is comprised of the students who participated in the SAT program and reported plans to graduate from high school in 2006. The most recent critical reading (SAT-CR), mathematics (SAT-M), and writing (SAT-W) scores from a single administration of the SAT were used in the analyses for students with multiple testing results. Students’ self-reported HSGPA was obtained from the SAT Questionnaire that the students completed when they registered to take the SAT. A total of 150,377 students had complete data on all of the variables used in this study: SAT scores, HSGPA, and FYGPA.
Kobrin et al. (2008) provided a comparison of the sample from the SAT Validity Study to the population of students from four-year institutions that received at least 200 SAT score reports in 2005. The sample is fairly representative of the target population. The gender and racial/ethnic composition of the sample is similar to the 2006 College-Bound Seniors cohort (College Board, 2006). On average, the sample for this study performed better on the SAT and had higher mean HSGPA than the 2006 College-Bound Seniors cohort. A higher level of performance for the sample compared to the national population of SAT takers was expected because all of the students in the sample were enrolled in four-year colleges.

**Analyzes**

As a first step in examining the adequacy of the linear regression Model 1, contour plots and/or boxplots showing the relationship between the predictor variables and FYGPA were visually inspected. The contour plots show the density of cases in the bivariate distribution of the two variables, which were rescaled to be on a percentage scale (i.e., density divided by the sum of the densities, multiplied by 100 percent). In the plots, darker shading is associated with a higher density, and no shading is associated with the lowest densities (less than 0.1). As a result, the observed cases at the extreme ends of the distribution (e.g., SAT-W scores less than 350) are not visible on the contour plots. The scale of each contour plot (shown to the right of each plot) is determined separately based on the observed bivariate densities.

Next, the standardized residuals from Model 1 were plotted against each predictor variable, against FYGPA and against the predicted values. To further evaluate the distribution of the residuals, a normal probability plot and a histogram of the residuals were constructed. Finally, the means, medians, and distributions of the residuals were examined for different levels of the predictor variables.

After the visual inspection of the plots described above, four different statistical tests were used to determine the fit of the linear regression model and to determine whether a nonlinear model fits the data better. First, a lack-of-fit (LOF) test (e.g., Draper & Smith, 1998, p. 47) was performed. It is appropriate to apply this test because several examinees can be found for each of several combinations of SAT scores and HSGPA (an example of such a combination is SAT-CR = 600, SAT-M = 540, SAT-W = 520, and HSGPA = 4.0, which has eight examinees). The LOF test was performed in two steps. First, Model 1 was fitted to the data, the SSE was computed for the regression and the residuals were saved. Second, to compute the pure error SS, the residuals were used as the dependent variable in an ANOVA, with the four predictor variables and their interactions as the factors.

Because each of the three SAT sections has 61 possible scores (i.e., 200, 210, 220...800), and HSGPA has 12 possible values (i.e., 0, 1.00, 1.33, 1.67...4.33), a full factorial ANOVA would involve (61 x 61 x 61 x 12), or 2,723,772 cells. Due to the computational difficulty of running a full factorial ANOVA, a new variable V was created that combined the information from the four predictors. This variable took on a unique value for every possible combination of the
four predictors. For example, an individual who scored 500 on SAT-CR, 550 on SAT-M, 600 on SAT-W, and earned a 3.0 HSGPA had the value 5005506003 on $V$, and an individual who scored 670 on SAT-CR, 670 on SAT-M, 680 on SAT-W, and earned a 4.0 HSGPA had the value 6706706804. A One-Way ANOVA was performed on the residuals using $V$ as a single factor. The Error SS from this ANOVA is the pure error sum of squares. The LOF SS was computed as the pure error SS from SSE.

Second, in the LOF test, the number of combinations of predictor variables (SAT scores and HSGPA) was large. In such a case, the LOF test statistic may not have the theorized $F$-distribution. So, as an additional test for lack of fit, an ANOVA was performed on FYGPA using the original variables (SAT-CR, SAT-M, SAT-W, and HSGPA) as main effects and no interaction terms. The percent of variance of FYGPA accounted for by this model ($R^2$) was compared with the $R^2$ from the simple linear regression described in Model 1. A substantial difference in these two $R^2$ values would indicate that a model more general than the linear regression model should fit the data better. (A plot of the average response variable for different levels of a predictor variable is helpful in finding the more general model.)

Third, quadratic terms (i.e., the square of each predictor variable) and interaction terms were added to Model 1 to determine if the extended model performed better. To prevent collinearity, the predictor variables were standardized before creating the squared terms and before entering into the regression. The change in $R^2$ as well as the significance level of the $F$-test were examined.

Finally, because the range of the dependent variable in linear regression is minus to plus infinity, and the dependent variable in the current study (FYGPA) has a range from 0 to 4, we fit the following nonlinear model:

$$f(FYGPA) = 4 \times \frac{e^{FYGPA}}{1 + e^{FYGPA}}$$

(4)

$$FYGPA = \beta_0 + \beta_1 \times SAT_M + \beta_2 \times SAT_{CR} + \beta_3 \times SAT_W + \beta_4 \times HSGPA + \epsilon$$

(5)

This model restricts the range of the predicted response to between 0 and 4. The nonlinear model is equivalent to a linear regression of $\log \left(\frac{FYGPA}{4-FYGPA}\right)$ on SAT-M, SAT-W, SAT-CR, and HSGPA. Before fitting the model to our data, cases with FYGPA > 4 were removed from the data set. (A total of 118 such cases were found, representing about 0.1 percent of the sample.)

**Results**

The summary statistics for FYGPA, SAT scores, and HSGPA are given in Table 1, and the correlations among variables as provided in Kobrin et al. (2008) are shown in Table 2. The linear regression model equation for predicting FYGPA from the other variables was given earlier as Model 1. The value of the square of the multiple correlation coefficient ($R^2$) for the regression model is 0.24. The ANOVA table for the linear regression is given in Table 3.
Table 1
Descriptive Statistics for FYGPA and Predictor Variables

<table>
<thead>
<tr>
<th></th>
<th>Minimum</th>
<th>Maximum</th>
<th>Mean</th>
<th>SD</th>
<th>Skew</th>
<th>Kurtosis</th>
</tr>
</thead>
<tbody>
<tr>
<td>FYGPA</td>
<td>0.00</td>
<td>4.27</td>
<td>2.97</td>
<td>0.71</td>
<td>-0.99</td>
<td>1.15</td>
</tr>
<tr>
<td>HSGPA</td>
<td>0.00</td>
<td>4.33</td>
<td>3.60</td>
<td>0.50</td>
<td>-0.66</td>
<td>0.25</td>
</tr>
<tr>
<td>SAT-CR</td>
<td>200</td>
<td>800</td>
<td>560</td>
<td>96</td>
<td>0.06</td>
<td>-0.19</td>
</tr>
<tr>
<td>SAT-M</td>
<td>200</td>
<td>800</td>
<td>579</td>
<td>97</td>
<td>-0.09</td>
<td>-0.31</td>
</tr>
<tr>
<td>SAT-W</td>
<td>200</td>
<td>800</td>
<td>554</td>
<td>94</td>
<td>0.07</td>
<td>-0.23</td>
</tr>
</tbody>
</table>

Note: N = 150,377

Table 2
Correlation Matrix of FYGPA and Predictor Variables

<table>
<thead>
<tr>
<th></th>
<th>HSPGA</th>
<th>SAT-CR</th>
<th>SAT-M</th>
<th>SAT-W</th>
</tr>
</thead>
<tbody>
<tr>
<td>FYGPA</td>
<td>0.36</td>
<td>0.29</td>
<td>0.26</td>
<td>0.33</td>
</tr>
<tr>
<td>HSGPA</td>
<td>--</td>
<td>0.21</td>
<td>0.23</td>
<td>0.25</td>
</tr>
<tr>
<td>SAT-CR</td>
<td>0.45</td>
<td>--</td>
<td>0.50</td>
<td>0.71</td>
</tr>
<tr>
<td>SAT-M</td>
<td>0.49</td>
<td>0.72</td>
<td>--</td>
<td>0.50</td>
</tr>
<tr>
<td>SAT-W</td>
<td>0.49</td>
<td>0.84</td>
<td>0.72</td>
<td>--</td>
</tr>
<tr>
<td>FYGPA</td>
<td>0.54</td>
<td>0.48</td>
<td>0.47</td>
<td>0.51</td>
</tr>
</tbody>
</table>

Note: From Kobrin et al. (2008). The correlations above the diagonal are unadjusted, and those below the diagonal are adjusted for range restriction using the 2006 SAT cohort of college-bound seniors as the reference group.

Table 3
ANOVA for the Linear Regression

<table>
<thead>
<tr>
<th></th>
<th>Sum of Squares</th>
<th>df</th>
<th>Mean Sum of Squares</th>
<th>F</th>
<th>Sig.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Regression</td>
<td>17,747.05</td>
<td>4</td>
<td>4,436.76</td>
<td>11,630.82</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>Residual</td>
<td>57,360.64</td>
<td>150,369</td>
<td>0.38</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>75,107.68</td>
<td>150,373</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Graphical Plots

Figure 2 shows the boxplots of FYGPA by levels of HSGPA, and Figure 3 shows the contour plot of FYGPA by SAT-W. (The plots for SAT-CR and SAT-M with FYGPA are very similar to the SAT-W plot and therefore are not shown.) Figure 2 shows an overall linear relationship between the two variables with the exception of the two lowest levels of HSGPA, but a large number of cases deviate from it; these cases have a large positive or negative residual. Figure 3 shows the bivariate density of SAT-W and FYGPA, and it indicates a roughly linear relationship among the highest concentration (darkest shading) of cases. Yet, again, there are many cases that deviate from the linear relationship of SAT-W and FYGPA.
Figure 1. Example of a hypothetical nonlinear relationship between SAT and FYGPA.

Figure 2. Boxplots of FYGPA by HSGPA.

Figure 3. Plot of SAT-W and FYGPA.

Figure 4 displays boxplots of the standardized residuals by levels of HSGPA, and Figure 5 displays a contour plot of the standardized residuals against SAT-W. These plots show that most of the cases have residuals near zero; however, there are a greater number of negative residuals than positive residuals, likely due to a ceiling effect. Table 4 gives the mean and median standardized residuals for each level of HSGPA and for each 50-point score range on SAT-W. There are larger positive mean residuals for those scoring at the lower end of the scale, but these means are based on a smaller number of examinees than those at the middle and upper ends of the scale.
FIGURE 4. Boxplots of standardized residuals by HSGPA.

FIGURE 5. Plot of SAT-W and standardized residuals.
Table 4
Mean and Median Standardized Residuals of the Regression of FYGPA on HSGPA, SAT-CR, SAT-M, and SAT-W by Levels of HSGPA and SAT-W

<table>
<thead>
<tr>
<th>HSGPA</th>
<th>N</th>
<th>Mean (SD)</th>
<th>Median</th>
<th>SAT-W</th>
<th>N</th>
<th>Mean (SD)</th>
<th>Median</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>3</td>
<td>---</td>
<td>---</td>
<td>200-249</td>
<td>87</td>
<td>0.65 (1.40)</td>
<td>0.91</td>
</tr>
<tr>
<td>1.0</td>
<td>12</td>
<td>---</td>
<td>---</td>
<td>250-299</td>
<td>219</td>
<td>0.31 (1.26)</td>
<td>0.43</td>
</tr>
<tr>
<td>1.33</td>
<td>17</td>
<td>0.53 (1.54)</td>
<td>0.30</td>
<td>300-349</td>
<td>1,103</td>
<td>0.98 (1.25)</td>
<td>0.21</td>
</tr>
<tr>
<td>1.67</td>
<td>187</td>
<td>0.50 (1.46)</td>
<td>0.51</td>
<td>350-399</td>
<td>4,715</td>
<td>0.03 (1.18)</td>
<td>0.15</td>
</tr>
<tr>
<td>2.0</td>
<td>1,027</td>
<td>0.26 (1.31)</td>
<td>0.36</td>
<td>400-449</td>
<td>12,443</td>
<td>-0.01 (1.14)</td>
<td>0.12</td>
</tr>
<tr>
<td>2.33</td>
<td>2,432</td>
<td>0.10 (1.29)</td>
<td>0.25</td>
<td>450-499</td>
<td>22,602</td>
<td>-0.11 (1.09)</td>
<td>0.13</td>
</tr>
<tr>
<td>2.67</td>
<td>6,484</td>
<td>-0.14 (1.24)</td>
<td>0.13</td>
<td>500-549</td>
<td>30,363</td>
<td>-0.15 (1.06)</td>
<td>0.14</td>
</tr>
<tr>
<td>3.0</td>
<td>19,642</td>
<td>-0.02 (1.13)</td>
<td>0.12</td>
<td>550-599</td>
<td>29,032</td>
<td>-0.00 (0.99)</td>
<td>0.16</td>
</tr>
<tr>
<td>3.33</td>
<td>27,113</td>
<td>-0.03 (1.07)</td>
<td>0.12</td>
<td>600-649</td>
<td>23,573</td>
<td>0.01 (0.90)</td>
<td>0.17</td>
</tr>
<tr>
<td>3.67</td>
<td>36,237</td>
<td>-0.01 (0.98)</td>
<td>0.14</td>
<td>650-699</td>
<td>15,187</td>
<td>0.03 (0.83)</td>
<td>0.20</td>
</tr>
<tr>
<td>4.0</td>
<td>40,275</td>
<td>0.02 (0.89)</td>
<td>0.18</td>
<td>700-749</td>
<td>7,697</td>
<td>0.00 (0.74)</td>
<td>0.16</td>
</tr>
<tr>
<td>4.33</td>
<td>16,948</td>
<td>0.01 (0.80)</td>
<td>0.17</td>
<td>750-800</td>
<td>3,356</td>
<td>-0.05 (0.64)</td>
<td>0.09</td>
</tr>
</tbody>
</table>

The medians are much more similar across levels of HSGPA and SAT-W, due to a slight skew in the distribution of the residuals. The normal probability plot and histogram of the residuals from the regression of FYGPA on the four predictors (Figures 6 and 7) show a slight deviation in the distribution of the residuals as compared to the expected.

FIGURE 6. Normal probability plot of the residuals.
Figure 7. Histogram of standardized residuals.

Two of the partial regression contour plots appear in Figures 8 (for HSPGA) and 9 (for SAT-W). The plots were very similar for SAT-CR and SAT-M and are not shown. Figures 8 and 9 show a random pattern centered around zero, suggesting that the variance of the residuals is homogeneous. The plots do identify a few outlying cases. Because the sample size in this study is so large, it is expected that some outliers will exist by chance, and these outliers do not exert much influence on the regression model estimates.

Figure 8. Partial regression plot for HSGPA and FYGPA.
Figure 10 shows the contour plot of the predicted values and standardized residuals. Figures 8, 9, and 10 do not reveal any departures from linearity, but Figure 10 does indicate that there may be some heteroscedasticity, as the scatter of points appears less variable in the upper end of the scale of the predicted values. Table 5 shows the mean and standard deviation of the residuals by levels of the predicted values. As the predicted values increase, the standard deviations of the residuals decrease. The correlation of the predicted values and squared standardized residuals was -0.128, which is slightly different from 0. Overall, the graphical analyses show some indication of misfit of the linear regression model. However, the extent of misfit is not severe, as will be clear from the analyses that follow. Next, several statistical tests were performed to examine if the misfit observed in the graphical analyses is statistically significant.
Table 5
Mean and Standard Deviation of Standardized Residuals of the Regression of FYGPA on HSGPA, SAT-CR, SAT-M, and SAT-W by Levels of the Predicted Values ($\hat{Y}$)

<table>
<thead>
<tr>
<th>$\hat{Y}$</th>
<th>N</th>
<th>Mean</th>
<th>SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\leq$ 1.8</td>
<td>35</td>
<td>0.96</td>
<td>1.55</td>
</tr>
<tr>
<td>1.8 – 2.0</td>
<td>180</td>
<td>0.37</td>
<td>1.35</td>
</tr>
<tr>
<td>2.2 – 2.4</td>
<td>773</td>
<td>0.10</td>
<td>1.23</td>
</tr>
<tr>
<td>2.4 – 2.6</td>
<td>3,503</td>
<td>0.01</td>
<td>1.21</td>
</tr>
<tr>
<td>2.6 – 2.8</td>
<td>11,325</td>
<td>0.00</td>
<td>1.14</td>
</tr>
<tr>
<td>2.8 – 3.0</td>
<td>24,837</td>
<td>-0.01</td>
<td>1.12</td>
</tr>
<tr>
<td>3.0 – 3.2</td>
<td>36,120</td>
<td>-0.00</td>
<td>1.04</td>
</tr>
<tr>
<td>3.2 – 3.4</td>
<td>38,352</td>
<td>-0.01</td>
<td>0.97</td>
</tr>
<tr>
<td>3.4 – 3.6</td>
<td>27,603</td>
<td>0.02</td>
<td>0.83</td>
</tr>
<tr>
<td>$&gt;3.6$</td>
<td>7,649</td>
<td>-0.02</td>
<td>0.68</td>
</tr>
</tbody>
</table>

Statistical Tests

Lack-of-fit test. The ANOVA for the LOF test is shown in Table 6. The first two rows of numbers in the table reproduce the results from the linear regression of FYGPA on SAT-CR, SAT-M, SAT-W, and HSGPA (see Table 3). The next two rows show the results of the LOF test. Note that the degrees of freedom (df) and sum of squares of the second row is the sum of those on the third and fourth rows. The resulting LOF $F$-statistic is 1.11, which is statistically significant at $p < .0001$, but it is not much larger than the median (of an $F$-distribution with degrees of freedom equal to 80,470 and 69,903) of 1 and is not practically significant. The adjusted $R^2$ for Model 1 is 0.24, and the adjusted $R^2$ corresponding to the LOF test is 0.28, so there does not seem to be much room for improvement of the regression model for these data.

FIGURE 10. Plot of predicted values and standardized residuals.
Table 6
ANOVA for the Linear Regression Along with the Lack of Fit Calculations

<table>
<thead>
<tr>
<th></th>
<th>Sum of Squares</th>
<th>df</th>
<th>Mean Sum of Squares</th>
<th>F</th>
<th>Sig.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Regression</td>
<td>17,747.05</td>
<td>4</td>
<td>4,436.76</td>
<td>11,630.82</td>
<td>&lt; 0.0001</td>
</tr>
<tr>
<td>Residual</td>
<td>57,360.64</td>
<td>150,369</td>
<td>0.38</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lack of Fit</td>
<td>32,202.34</td>
<td>80,466</td>
<td>0.40</td>
<td>1.11</td>
<td>&lt; 0.0001</td>
</tr>
<tr>
<td>Pure Error</td>
<td>25,158.30</td>
<td>69,903</td>
<td>0.36</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>75,107.68</td>
<td>150,373</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

ANOVA with no interaction terms. The ANOVA with only main effects produced an adjusted $R^2$ of 0.24, which is identical to that produced by Model 1. This result indicates that there is no model that is expected to predict FYGPA from SAT scores and HSGPA much better than the linear regression model for this data set.

Test of the addition of nonlinear terms. Because the LOF statistic was statistically significant, we proceeded to examine if a refinement of the linear regression model would lead to a better model. Statistical tests were performed to determine whether adding squared terms and interaction terms to the linear regression model would result in statistically significant improvements. Model 1 was compared with six additional regression models that included a quadratic (squared) term for each predictor (M2 to M7), with Model 6 including squared terms for all four predictors, and Model 7 including squared terms and first-order interaction terms involving all four predictors:

- [M2] $\text{FYGPA} = \beta_1(SAT-M) + \beta_2(SAT-CR) + \beta_3(SAT-W) + \beta_4(HSGPA) + \beta_5(SAT-M^2) + \epsilon.$
- [M3] $\text{FYGPA} = \beta_1(SAT-M) + \beta_2(SAT-CR) + \beta_3(SAT-W) + \beta_4(HSGPA) + \beta_5(SAT-CR^2) + \epsilon.$
- [M4] $\text{FYGPA} = \beta_1(SAT-M) + \beta_2(SAT-CR) + \beta_3(SAT-W) + \beta_4(HSGPA) + \beta_5(SAT-W^2) + \epsilon.$
- [M5] $\text{FYGPA} = \beta_1(SAT-M) + \beta_2(SAT-CR) + \beta_3(SAT-W) + \beta_4(HSGPA) + \beta_5(HSGPA^2) + \epsilon.$
- [M6] $\text{FYGPA} = \beta_1(SAT-M) + \beta_2(SAT-CR) + \beta_3(SAT-W) + \beta_4(HSGPA) + \beta_5(SAT-M^2) + \beta_6(SAT-CR^2) + \beta_7(SAT-W^2) + \beta_8(HSGPA^2) + \epsilon.$
- [M7] $\text{FYGPA} = \beta_1(SAT-M) + \beta_2(SAT-CR) + \beta_3(SAT-W) + \beta_4(HSGPA) + \beta_5(SAT-M^2) + \beta_6(SAT-CR^2) + \beta_7(SAT-W^2) + \beta_8(HSGPA^2) + \beta_9(SAT-M \times SAT-CR) + \beta_{10}(SAT-M \times SAT-CR) + \beta_{11}(SAT-CR \times SAT-W) + \beta_{12}(SAT-CR \times SAT-W) + \beta_{13}(SAT-CR \times HSGPA) + \beta_{14}(SAT-W \times HSGPA) + \epsilon.$

The fit of each model including a quadratic term was compared with the fit of Model 1 using $F$-tests to compare nested regression models (see, for example, Draper & Smith, 1998, pp. 149–150 for a description of the $F$-test). Table 7 displays these results. The addition of the squared term was statistically significant in all cases, and each of Models 2–7 performed better than Model 1 according to the $F$-tests. However, there was hardly any increase in $R^2$ in these models above that of Model 1 so that the addition of the quadratic terms or the interaction terms did not cause any practically significant improvement of the model.
Table 7
Comparison of Regression Models to Test the Significance of the Addition of Nonlinear Terms

<table>
<thead>
<tr>
<th>Model</th>
<th>R Square</th>
<th>R Square Change</th>
<th>F</th>
<th>Sig.</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>.236</td>
<td>---</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>2</td>
<td>.236</td>
<td>.000</td>
<td>7.395</td>
<td>.007</td>
</tr>
<tr>
<td>3</td>
<td>.236</td>
<td>.000</td>
<td>18.651</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>4</td>
<td>.237</td>
<td>.000</td>
<td>12.732</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>5</td>
<td>.237</td>
<td>.001</td>
<td>158.326</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>6</td>
<td>.237</td>
<td>.001</td>
<td>42.048</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>7</td>
<td>.239</td>
<td>.003</td>
<td>58.426</td>
<td>&lt; .0001</td>
</tr>
</tbody>
</table>

Note: R Square Change is the difference in R Square between Model 1 and each other model.

Model using transformed FYGPA. The ANOVA for this model produced an $R^2$ of 0.23, which, again, is very similar to that produced by the other tested models, especially Model 1. Thus the transformation of FYGPA did not lead to a better-fitting model.

Conclusions

An important aspect of any study involving statistical analysis is ensuring that the model resulting from that analysis is the optimum model that can be fitted to the data. This study examined the adequacy of a linear regression model to predict FYGPA from SAT scores and HSGPA using the data analyzed by Kobrin et al. (2008). A variety of techniques, both visual and statistical, were used to examine if it is possible to improve on the linear regression for modeling the validity of SAT scores for predicting FYGPA in the National SAT Validity Study.

Although the effect of the addition of nonlinear terms to the linear regression model and the LOF test were both statistically significant, the change in $R^2$ in the regression was minimal, and the $F$-value for the LOF test was only 1.11, leading to the conclusion that the statistical significance of the results can be largely attributed to the very large sample size. The results of this study suggest that a linear regression model is adequate to predict FYGPA from SAT scores and HSGPA.

This study indicates that there is some degree of heteroscedasticity in the data. The estimates obtained from linear regression are still unbiased in the presence of heteroscedasticity, but they are estimated variances, and hence associated confidence intervals are not accurate. Even the effect on the confidence intervals will be modest, given that the absolute magnitude of the correlation between the predicted values and the squared residual is rather small ($r = -0.128$). It may be possible to apply weighed least squares or apply a transformation (such as the Box-Cox transformation; see, for example, Draper & Smith, 1998) to overcome the heteroscedasticity problem, but these techniques were not explored in the current study.

There are some limitations in the current study. Because no data are available from examinees who did not go to college after taking the SAT, the range of high school GPAs and test scores is narrower than the range found in the potential applicant pool. This "range
restriction” results in a reduction of the correlation between these measures and college performance. However, although the range restriction reduces the magnitude of the values of R² reported here, it influences Model 1 and any extensions of it in the same way so that it is unlikely that accounting for range restriction would change any of the conclusions of this study. Second, because the structure of the data is hierarchical, that is, the examinees attended 110 different institutions, a multilevel model may be more appropriate than the linear regression model that was examined in this paper. Kobrin and Patterson (2010) applied a multilevel model to the data from the National SAT Validity Study. This study examined the appropriateness of linear regression for predicting the most common criterion in college admission validity studies — FYGPA. The appropriateness of linear regression for predicting other relevant outcomes, such as cumulative grades or individual course grades, was not addressed. The methods described in this paper should be replicated using these other outcome measures.
Appendix

When the data set has more than one value of the response variable for at least some combinations of values of the predictor variables (for example, five individuals, all with different y-values, with \( x_1 = 1, x_2 = 0.5, \ldots, x_p = 0 \); two individuals, all with different y-values, with \( x_1 = 1, x_2 = 0.0, \ldots, x_p = 0.3 \) etc.), the lack-of-fit test can be employed to test the hypothesis that a regression model fits the data adequately. For example, consider a regression line

\[
y_y = \beta' x_i + e_y, \quad i = 1, 2, \ldots, I, \quad j = 1, 2, \ldots, n_i
\]

where \( x_i \) is the \( p \)-dimensional vector of predictor variables for the \( i \)-th combination of the predictor variables and \( y_{ij} \) is the response variable for the \( j \)-th individual with predictor variable vector \( x_i \), that was fitted by the method of least squares. One takes as estimates of \( \beta \) the values that minimize the sum of squares of residuals, that is, the sum of squares of the differences between the observed \( y \)-values and the fitted \( y \)-values. To have a lack-of-fit sum of squares, one observes more than one \( y \)-value for some values of \( x \). One then partitions the “sum of squares due to error,” that is, the sum of squares of residuals, into two components:

- **sum of squares due to error (or SSE)** = sum of squares due to “pure” error + sum of squares due to lack of fit.

The sum of squares due to “pure” error is the sum of squares of the differences between each observed \( y \)-value and the average of all \( y \)-values corresponding to the same \( x \)-value.

The sum of squares due to lack of fit is the weighted sum of squares of differences between each average of \( y \)-values corresponding to the same \( x \)-value and corresponding fitted \( y \)-value, the weight in each case being the number of observed \( y \)-values for that \( x \)-value.

\[
\sum \frac{(\text{observed value} – \text{fitted value})^2}{\text{weight}} = \sum \frac{(\text{observed value} – \text{local average})^2}{\text{weight}} + \sum \text{weight} \times (\text{local average} – \text{fitted value})^2.
\]

Suppose \( \hat{\beta} \) is the least square estimate of \( \beta \). Denote the fitted value from the regression model as \( \hat{y}_i = \hat{\beta}' x_i \). Suppose \( \hat{y}_i = \frac{1}{n_i} \sum y_j \). The error/residual sum of squares \( \sum_i \sum_j (y_{ij} - \hat{y}_j)^2 \) can then be partitioned as

\[
\sum_i \sum_j (y_{ij} - \hat{y}_j)^2 = \sum_i \sum_j (y_{ij} - \hat{y}_j)^2 + \sum (\hat{y}_j - \bar{y}_j)^2,
\]

where the first term on the right-hand side is the sum of squares due to pure error (SSPE), and the second term on the right-hand side is the sum of squares due to lack-of-fit (SSLOF). One then computes the \( F \)-statistic

\[
F = \frac{\text{SSLOF} / (I - p)}{\text{SSPE} / \left( \sum n_i - I \right)}
\]

that has an \( F \)-distribution with degrees of freedom \((I - p)\) and \( \left( \sum n_i - I \right) \).
$\sum (n_i - I)$ under the null hypothesis that the linear regression model fits the data. A significant value of this statistic indicates that the regression model appears to be inadequate and that an extension of the regression model, for example, by including square and interaction terms involving the predictors, might perform much better.

In addition, one often compares the adjusted $R^2 = 1 - \frac{(SSE/dfE)}{(SST/dfT)}$ from the linear regression, where $dfE$ and $dfT$ are the error and total degrees of freedom in the regression, and adjusted $R^2$ corresponding to the lack of fit test $= 1 - \frac{(SSPE/dfPE)}{(SST/dfT)}$, $dfPE = \sum n_i - I$. The latter adjusted $R^2$ can be considered an upper bound that one can reach by extending the regression model. If the later quantity is much larger than the former, that indicates that an extension of the regression model, for example, by including square and interaction terms involving the predictors, might perform much better than the regression model itself.
References


