Math 3080 – Final Project – Chun-hie Fox Group – Project 4

The Rising Cost of College Affects Student Debt

Kurt Fisher and Elijah Neilson with James Bosnell, Chun-hei Fok, and Elena Nazareuko

University of Utah

April 21, 2013

Abstract: *No statistically valid multi-linear regression model can be constructed that predicts average student indebtedness from 2001 data provided by 2001 data reported in Levine, Ramsey, and Smidt (2001). Initially, a computationally valid polynomial regression equation that associated total annual college cost to student indebtedness was constructed [$R^2 = 0.36$, sig. level $\alpha = 0.05$, $F(2,72)=20.55$, $p<0.001$], but on closer examination a grossly non-random pattern to residuals in the total cost data from only private universities was found. This hidden non-randomness of residuals led to the conclusion that the association predicted by the initial model was a spurious artifact. Other factors, not assessed by Levine, Ramsey, and Smidt, also contribute to the amount of education debt that a student is willing to incur.*

Introduction

On March 29, 2013, the Utah Board of Regents approved a five percent increase in student tuition at the 11 colleges and universities of in the Utah Higher System of Education (Whitehurst 2013), and from the 2006-2007 to the 2010-2011 academic years, student paid tuition at those colleges and universities increased by an average of 12.9% per year – a rate of increase that was several times the rate of increase in the consumer price index. In the 1980s, nationally student tuition financed about 25% of the expenses of colleges and universities, but currently, students finance about 47% of the national expense (*id*). In Utah during the 2006-2007 academic year, student tuition financed 36.1% of expenses of the eleven Utah Higher Education system colleges and universities, but by 2010-2011 academic year, that proportion rose to 48.5% (*id*).

The rapid rate of increase in student tuition raises concerns about the affordability of higher education, and using 2001 data reported in Levine, Ramsey, and Smidt (2001) for 30 public and 50 private colleges and universities (Table A 1, p. 15), the relationship between the total cost of attending a college to student debt was explored. Our null hypothesis was that there was no combination of six variables (semester type, public or private college, SAT score, score on English as a second language exam, cost of room and board, total cost of attendance) in the Levine, Ramsey, and Smidt data that was statistically significantly associated to a level α = 0.05 with student indebtedness, the seventh variable in that data set. Our alternative hypothesis was that some combination of six variables in the data could be statistically significantly associated to a level α = 0.05 with student indebtedness.

Initially, we found that in 2001, where *TC* is total annual cost of one year of attendance including room and board in thousands of U.S. dollars (USD), *I* is the total predicted student indebtedness in USD, the best-fit linear regression model predicting student debt ($M_{I,TC}$) [$R^{2=}$ 0.36, F(2,72)=20.55, sig. level $\alpha = 0.05$, $p<0.001$] is:

$$M_{y,x} = \beta_0 - (\beta_1 \times x) + (\beta_2 \times x^2);\ \text{(Eq. 1).}$$

$$M_{I,TC} = -12.196 + (2.328 \times TC) + (-0.0499 \times TC^2);\ \text{(Eq. 2).}$$

Equation 1 is the generalized formula; Equation 2 is the instant best-fit-model. The best-fit-model is plotted in Figure 1, below p. 3. Equation 2 is valid for the domain of total cost modeled, (10.2, 28.9 K USD).

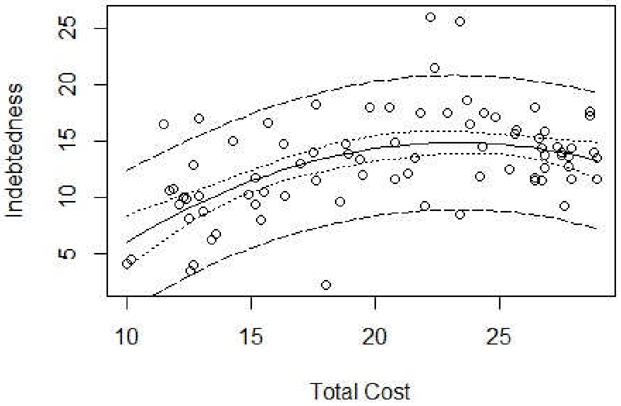

Figure 1. Initial best-fit-model of total cost to student indebtedness (unstandardized). Model No. 6 in Table A 9, below.

For the standardized form of the best-fit model, the regression equation [R^2= 0.31, F(2,27)=5.96, sig. level $\alpha = 0.05$, p<0.001] is:

$$M_{y,\acute{x}} = \beta_0 - \left(\beta_1 \times \left(\frac{(\acute{x}-\bar{x})}{Sxx}\right)\right) + \left(\beta_2 \times \left(\frac{(\acute{x}-\bar{x})}{Sxx}\right)^2\right);\ \text{(Eq. 3).}$$

$$S_{xx} = \sqrt{\frac{\sum x_i^2 - \frac{(\sum x_i)}{n}}{n-1}};\ \text{(Eq. 4).}$$

$$M_{I,\acute{T}C} = -12.196 + \left(2.328 \times \left(\frac{(\acute{T}C-14.16)}{2.741}\right)\right) + \left(-0.0499 \times \left(\frac{(\acute{T}C-14.16)}{2.741}\right)^2\right);\ \text{(Eq. 5).}$$

Equations 3 and 4 are the generalized formula for a standardized model; Equation 5 is the instant standardized best-fit-model. The standardized best-fit-model is plotted in Figure B 1, p. 23. Equation 5 is valid for the domain of total cost modeled above.

However, this initial model was based on the observation of a quadratic dispersion in the residuals when both private and public total cost were associated with both private and public student indebtedness, Figure B 4, p. 24. Closer examination of the cohorts of colleges subsetted by 30 public total cost with public student indebtedness and 50 colleges subsetted by private total cost with private student indebtedness revealed that this quadratic dispersion was an artifact because Levine, Ramsey, and Smidt's total cost data consists of two distinct populations, and one of those populations has non-random residuals. A statistically significant linear regression model for the subset of 30 public total cost to public student indebtedness colleges was found (Figure B 2, p. 23), and the residuals of that model are reasonably randomly distributed for regression modeling [Shapiro-Wilk p-value=0.49], Figure 2and Figure B 5, p. 25. But the residuals from a linear regression on the subset of private total cost to private student indebtedness colleges are grossly non-random [Shapiro-Wilk p-value=0.003], Figure 2 and Figure B 6, p. 26, because there is a very weak relationship between the cost of attending a private college and the amount of private student debt, Figure B 3, p. 23. Figure 2 shows the residuals of the public only subset model (left) and the private only subset model (right):

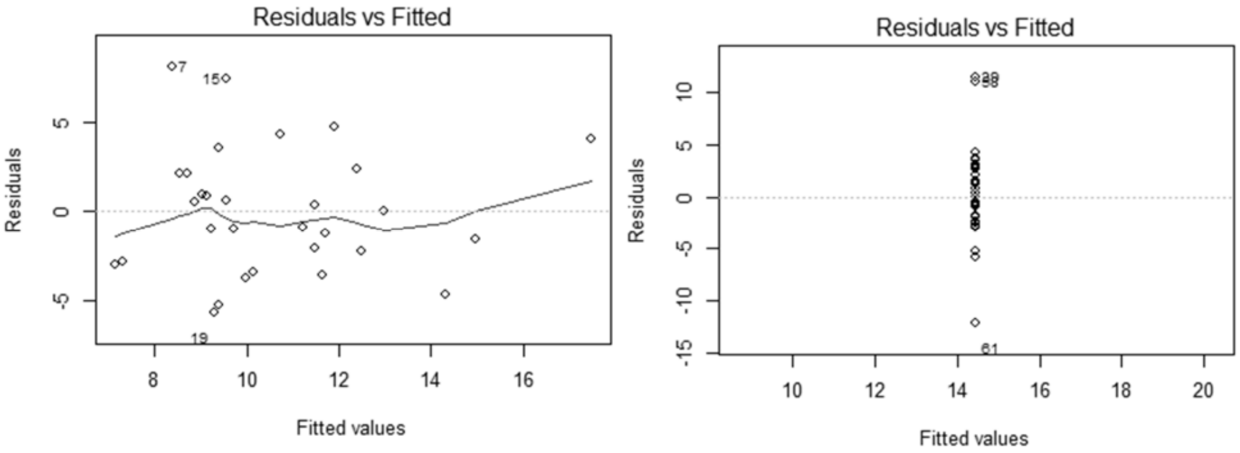

Figure 2 - Residuals vs. fitted values of linear model of public only data subset (left) and private only data subset (right). Both regression models are I~TC.

The quadratic residual curve in the combined cohort model data, Figure 1 and Figure B 4, is an artifact of the private college only residuals distorting the randomly distributed residuals of the public college model into a false polynomial-like curve, Figure 2, above. The distribution of residuals of the private only subset indicated that independent variables had been omitted during the experiment's design, and therefore, no statistically valid regression model can be formed from the data because the private college total cost would have to be included in any multivariable model based on the entire Levine, Ramsey, and Smidt table.

An important lesson-learned from this modeling exercise is that sometimes a quick initial assessment of a data's distribution can indicate whether a model is viable without engaging in detailed modeling. Table 1 summarizes the distribution of public and private subsets of the combined data:

Table 1- Summary of College Cost Data by Public and Private Institution.

Type	N	Mean(Total Cost)	SD(Total Cost)	Mean(Student Indebtedness)	SD(Student Indebtedness)	Ratio – Debt/TC
Public	30	14.2	2.7	10.6	4.2	0.75
Private	50	24.3	3.5	14.5	3.8	0.60

Source: Table A 1, below.

That important independent variables were not captured in the original survey is evident in this distribution of the subsetted data. In 2001, the means of total student debt for public and private colleges students overlap (10.6±4.2 K USD vs. 14.5 ±3.8 K), but the mean annual total cost of attending a private colleges was 170% higher than for a student attending a public university (24.3K USD verses 14.2K USD), and the means of total cost do not overlap. These estimates quickly lead to the inference that private school students pay much more to attend private institutions, but they debt finance the same gross amount as less well-off students. Low and moderate students attend less expensive public institutions which they cannot afford without taking on the same debt are private students. The most probable cause of this disparity in the ratio of student-debt-to-student-wealth between private and public students is family income and assets. Levine, Ramsey, and Smidt's data did not report these important free variables, and, as a consequence, no predictive model would be expected to capture different variations in public vs. private student debt.

Methods

With Levine, Ramsey, and Smidt's data, a best-fit linear regression model was developed using generally-accepted statistical practices. These practices included examination of the distribution of the data; identification of outliers; identification of the best independent variables using a correlation matrix, exploratory factor analysis, principal component analysis, and a scatterplot matrix; examination of the distribution of residuals; determination of the coefficients of the best-fit-model using Akaike's information criterion (AIC) and the elimination method; and, consideration of interaction effects. Analysis utilized computer-aided-software, that is the open source program *R*™ enhanced with specialty linear regression packages contributed by individual statisticians. The use of *R* for statistical analysis has been subjected to the criticism on the grounds that results may not be replicable or accurate due to the lack of a responsible corporate party that oversees the software's quality control. *R* was a sufficient and appropriate analysis tool, but for replication purposes, the open source software packages used to develop this model are listed in Table A 2, p. 17, below. Particularly useful functions for model analysis in either the base *R* program or in its specialty packages are listed in Table A 3, below.

Development of the initial facially valid regression model

Our initial regression modeling indicated that in 2001, total cost of attending a college or university is statistically significantly associated with the level of debt incurred by a student; however, total cost only is associated with about 36% of the variation in debt incurred. The null hypothesis was rejected; the alternative hypothesis was accepted based on the following analysis and results.

First, the data's distribution was analyzed (Table 1, above), outliers were identified, and outliers were removed from further modeling. Figure B 7, below p. 27, shows the distribution of the sample's room and board, total annual cost, and student indebtedness data, and two outliers can be seen in the figure. Further analysis identified the outliers as the private universities Case Western Reserve and Northeastern (Figure B 8), and those outliers were excluded in subsequent analysis steps (*see* Outlier column in Table A 1 and Table A 4, p. 18).

Second, potential candidates for independent regression variables of total annual cost and type were selected, and those selections were identified based on results from a correlation matrix, exploratory factor analysis, principal component analysis, and a scatterplot matrix. Combinations of candidates for independent regression variables were excluded through correlation matrix analysis, Table A 5, p. 19, were only one of Room-Board and Total-Cost, SAT and Type, or Type and Total-Cost were needed to model indebtedness. The correlation matrix was interpreted as follows. One or more of the candidate independent variables that are highly correlated between themselves should be excluded. For example, Type is correlated with SAT, Room-Board, and Total-Cost. Table A 5, p. 19. These correlation scores indicate that probably only one of Room-Board and Total-Cost are needed to model Indebtedness. Adding both Room-Board and Total-Cost adds only the noise of the covariance between Room-Board and Total-Cost to a model. Scoring the results of exploratory factor analysis (Table A 6 , p. 19) suggested type of institution, average SAT scores, and/or total annual cost as candidate independent regression variables that may predict an unknown third dependent variable. Exploratory factor analysis yields candidate independent variables that are contradictory to variables suggested by the correlation matrix. Kabacoff (2001) demonstrates how in exploratory factor analysis, the correlations of the covariances of combinations of candidate independent variables are computed (*id* at p. 343-345). High values in the right-hand column of the exploratory factor analysis matrix indicate that factor might be used in a regression model, either alone or in combination. The result of principal component analysis (PCA) are shown graphically in a correlogram ordered by each factors' PCA score, Figure B 9, p. 28, and by values in Table A 7. PCA can also be used to estimate the number of factors sufficient to capture all the variation in all candidate factors, Teetor (2011), p. 349-350. PCA suggested only one of any four factors - total annual cost, room and board, SAT score and type of institution – would be needed as candidate independent variables in order to capture 93% or more of the variation in Indebtedness. Factor analysis indicated that two factors would have been sufficient, but principal component analysis suggested that only one factor would have been sufficient.

A scatterplot matrix was prepared to further explore candidate independent variable selection, Figure B 10, p. 29. Kabacoff (2011) describes a scatterplot matrix as a *n* by *n* grid with the dimensions of numeric columns in the source data table(*id* at p. 264-271). The *n* entries of the diagonal contain the names of each variable, and the remaining cells of the grid contain a x,y plot of a simple linear regression of the intersected row and column variables. For the instant case, the college data table has 7 numeric column variables, and Figure B 10 contains 42 (7^2-7) linear regression graphs. The proposed independent variable – Indebtedness- is in the lower right-hand corner of Figure B 10. Looking at the bottom row of Figure B 10, the plot at the intersections of Indebtedness-Room Board, Indebtedness-Total Cost, and Indebtedness-Type of institution suggest a cubic, quadratic and linear relationship respectively between indebtedness and each individual independent variable.

Third, the foregoing various tests for regression candidates are signs that informed our judgment call concerning how many and which of college data variables are the best potential factors for modeling. Table A 8, p. 20, summarizes the signs provided by each test, and based on those tests and signs, and the finding that one or two factors would be sufficient for modeling, the two primary candidate factors selected for regression modeling were total cost and type-of-institution.

Fourth, based on preliminary regression modeling, influence points were identified, and bias in the model was reduced by removing such points from further modeling, Figure B 11, p. 30 and Table A 4, p. 18. The influence plot, Figure B 11, is a graphical variation of the Cook's distance plot shown in Figure B 12 (lower right hand plot), p. 31. Cook's distance D_i indicates whether a given point has a high influence on the accuracy of the regression slope, and a rule-of-thumb is where D_i.> 1, the point may represent a questionable observation. In Figure B 11, Cook's distance D_i is plotted as the radius of circle. Three influence points, Id No.s 7, 6 and 61, are apparent, and those points correspond to Florida Int'l. University, Florida State University, and Rice University. For the purpose of this paper, receipt of extrinsic evidence concerning the questionable nature of these influence points was assumed, and those three observations were excluded from further modeling.

Fifth, a polynomial best-fit-model (BFM) was selected using Akaike's information criterion (AIC) and the elimination method, and quadratic BFM Model No. 6 ($I \sim TC+TC^2$) was chosen from ten regression simulations described in Table A 9, p. 20. Akaike's information criteria incorporates a measure of the maximum likelihood estimator for a given regression, and the lower the AIC measure, the more likely that a regression model is the best fit model. AIC supplements other model elimination criterion shown in Table A 9, such as normality, the model's residual, the model's F statistic, and the p-value of the intercept β_0 and the p-value of the dominant independent variable β_1. The model elimination reasoning applied to Model No.s 1 through 10 in Table A 9 leading to the selection of Model No. 6 was as follows. Models No.s 1-2: With outliers and influence points, simple linear models involving total cost violated the normality of residuals condition. The Shaprio Wilk statistic indicates that the residuals were not normally distributed. Models No. 3 to 5: Removing the outlier and influence points and examining permutations of first degree variables improved normality, but the change did not substantially improve the coefficient of the determination R^2. Model No. 6: A quadratic model on total cost substantially improved the coefficient of the determination and Akakie's information criteria, and Model No. 6 also had strong p-values for β_0 and β_1. Model No.s 7, 8 and 9: Model permutations based on combining the quadratic TC^2 with first degree terms for SAT and type of institution did not lower AIC or substantially change the coefficient of determination, and that lack of change indicated that adding SAT and type of institution to the TC^2 model did not affect regression results. Therefore, SAT and type of institution should be eliminated as factors since they provide no additional information that contributed to the regression. Model No. 10: A model combining a cubic for room and board with a quadratic term for total cost achieved the lowest AIC and highest coefficient of determination, but the p-values for β_0 and β_1 are in the region of acceptance of the null hypothesis. The foregoing elimination process left Model No. 6 (I ~ $TC+TC^{2)}$ as the best-fit-model of choice.

Sixth, a regression equation for the polynomial BFM was prepared, as recited in Equation No 2 at p. 3, above, and as shown in Figure 1, p. 3 below. The normality of the residuals and the sufficiency of other distributions of the residuals and y-hats were examined by a standard four-box plot, Figure B 12, p. 31. The upper-right hand residual to fitted-plot of Figure B 12 shows a clump of residuals at the right-hand side of the figure, and at first, it was felt that this pattern was inherent to a polynomial regression. The coefficients and detailed characteristics of the model are described in Table A 10, p 21. Seventh, interaction between variables in the final model was considered, but because there is only one independent variable, there are no interactions to consider between TC and TC^2 (Figure B 13, p. 32). Eighth and finally, a regression equation for a standardized polynomial BFM was prepared, as recited in Equation No 5 at page 3, above, and as shown in Figure B 1, below.

In conclusion, based on the foregoing analysis and results, the initial alternative hypothesis was accepted because a statistically significant linear model could be found.

Subsequent rejection of the initial hypothesis based on hidden non-random distribution of residuals

Because of the clump at the right-hand side of the fitted to y-hat graph in the upper right corner of Figure B 12, p. 31, it was decided to investigate whether a linear regression model could be created for sub-populations of public and private universities. Examination of the public college only data showed that a statistically significant linear model of public total cost to public student indebtedness could be constructed, Figure B 5, p. 25 and Table A 11, p. 22 [R^2=0.295, F(1,28)=11.8, p-value= 0.0019]. The residuals of that public only college model were normally distributed [Shapiro-Wilk p-value=0.49], Figure B 5, p. 25. On generation of a similar model for the private college only model, a grossly non-random distribution of model residuals was found [Shapiro-Wilk p-value=0.003], Figure B 6, p. 26. It was concluded that the polynomial residual pattern seen in the initial best-fit model, Figure B 4, p. 24, was an artifact of combining residuals from these two heterogeneous samples.

Results

The alternative simulations explored in Table A 9, p. 20, and the correlation matrix, Table A 5, p. 19, show that no statistically valid model could be generated without including the combined total cost information from both public and private colleges because no other variable has a sufficiently strong coefficient of correlation to support a linear model. Thus, no statistically valid model can be generated of student indebtedness from the Levine, Ramsey, and Smidt data that does not include the non-random dispersion of residuals hidden in the private college total cost data.

Discussion

A statistically valid model might be developed by gathering and including additional variables concerning family income and wealth. The ability of the public only college model based on total cost data (I~TC) to predict student debt is low [R^2=0.295], but this was expected given that Levine, Ramsey, and Smidt's data did not report factors, such as the income and assets of students' families, that reasonably would be expected to influence a student and family decision-making concerning debt finance students college education.

The decision on how much to debt financing to incur in order to obtain a degree is a complex decision by a student and his or her family, and all of the factors that might go into that decision are not captured here. A likely explanation for the weak relationship between the cost of attending a private college and the amount of private student debt is that the price of attended prestigious private colleges is inelastic as high income, high-wealth parents have a high willingness to pay any amount to send their adult children to colleges were they can make social and economic connections with ultra-wealthy students. In contrast, the total cost of attending public college is more elastic, since low and middle income parents and students are forced by wealth constraints to choose between different qualities of instruction and of the college experience. Gender may also have a role in the level of debt financing a family or student is willing to incur. Levine, Ramsey, and Smidt's data did not report these important free variables, and, as a consequence, the predictive strength of any model derived from that data would be low. Future work will involve collecting more recent data that includes the family income and asset indicators, and we expect an improved model would follow. Nevertheless, the exercise of preparing the inconclusive models here provided useful insights for designing an improved future improved model.

Conclusion

A statistically significant polynomial regression equation that associated the total cost of attendance at a college or university with the amount of debt incurred by students could not be formulated from Levine, Ramsey, and Smidt's data, due to hidden non-randomness in the total cost model residuals of the private college only sub-population.

Kurt Allen Fisher

Elijah Neilson

References

Comprehensive R Archive Network CRANa. (2013). Homepage. Retrieved Apr. 1, 2013, from http://cran.r-project.org/.

Comprehensive R Archive Network. CRANb. (2012). R (Version 2.14.1) [Computer Software].

Fox, J., Weisberg, S., Bates, D. et al. (2013, Feb. 18). Car Package for R (Version 2.0-16) [Computer software]. Retrieved Apr. 1, 2013, from http://cran.r-project.org/web/packages/car/index.html.

Fox, J., Weisberg, S. & Hong, J. et al. (2013, Apr. 18). Effects Package for R.. [Computer software]. (Version 2.2-4). Retrieved Apr. 1, 2013, from http://cran.r-project.org/web/packages/effects/index.html

Ihaka, R, Murrell, P., Hornik, K. et al. (2013, Jan. 24). Colorspace Package for R (Version 1.2-1) [Computer software]. Retrieved Apr. 1, 2013, from http://cran.r-project.org/web/packages/colorspace/index.html.

Kabacoff, R. I. (2011). *R in Action: Data analysis and graphics in R*. Shelter Island, N.Y.: Manning. Publ.

Levine, D. M., Ramsey, P. P., & Smidt, R. K. (2001). *Applied statistics for engineers and scientists: Using Microsoft Excel and MINITAB*. Upper Saddle River, N.J: Prentice Hall.

Teetor, P. 2011. *R Cookbook*. Sebastopol, Calif.: O'Reilly.

Wei, T. (2013, Jan. 24). Corrplot Package for R (Version 0.71) [Computer software]. Retrieved April 1, 2013, from http://cran.r-project.org/web/packages/corrplot/index.html.

Whitehurst, L. (Salt Lake City Tribune). (2013, March 29). As state money is slashed, Utah students are paying more of college cost. *Salt Lake City Tribune*, p. A1, A4. Salt Lake City, Utah. Retrieved from http://www.sltrib.com/sltrib/news/56070355-78/utah-tuition-percent-increase.html.csp.

Addendum A - Tables

Table A 1. College costs data for 2001

Id	School	Term	Type	SAT	TOEFL	Room Board (USD 000s)	Total Cost (USD 000s)	Indebt-edness (USD 000s)	Outlier Influence Point
1	Arizona St. U.	1	0	1080	0	4.3	12.7	12.9	0
2	Ball St. U.	1	0	985	1	4	12.5	8.21	0
3	Cal St. U. Fresno	1	0	955	0	5.4	13.1	8.76	0
4	Clemson U.	1	0	1130	1	3.9	12.4	9.98	0
5	Col. of William Mary	1	0	1295	1	4.5	19.4	13.42	0
6	Florida Int'l. U.	1	0	1135	0	2.7	10	4.14	2
7	Florida St. U.	1	0	1180	1	4.5	11.5	16.5	2
8	George Mason U.	1	0	1055	1	5	17	13	0
9	Georgia St. U.	0	0	1115	0	7.4	15.4	8.08	0
10	Montclair St. U.	1	0	1025	0	5.3	10.2	4.5	0
11	North Carolina St. U.	1	0	1145	1	4	14.3	14.99	0
12	Oregon St. U.	0	0	1072	1	4.4	15.5	10.5	0
13	Purdue U.	1	0	1095	1	4.5	15.2	11.84	0
14	San Diego St. U.	1	0	945	1	6.2	13.6	6.75	0
15	Slippery Rock U. Penn.	1	0	955	0	3.6	12.9	17	0
16	SUNY Binghamton	1	0	1039	1	4.6	13.4	6.25	0
17	Texas AM U.	1	0	1150	1	3.9	12.7	4.1	0
18	U. Georgia	0	0	1180	1	4	11.9	10.8	0
19	U. Hawaii Manoa	1	0	1075	0	4.7	12.6	3.62	0
20	U. Houston	1	0	1065	1	4.1	12.1	9.4	0
21	U. Maryland	1	0	1170	1	5.5	15.7	16.64	0
22	U. Mass. Amherst	1	0	1100	1	4.2	16.4	10.2	0
23	U. Nevada Las Vegas	1	0	980	0	5.5	12.3	10	0
24	U. New Hampshire	1	0	1110	1	4.4	18.6	9.66	0
25	U. North Carolina CH	1	0	1225	1	4.5	15.2	9.41	0
26	U. Texas Austin	1	0	1215	1	3.9	12.9	10.2	0
27	U. Vermont	1	0	1115	1	5.1	22.4	21.5	0
28	Virginia C. U.	1	0	1005	1	4.3	16.3	14.73	0
29	Virginia Tech	1	0	1265	1	3.5	14.9	10.33	0
30	West Virginia U.	1	0	1025	1	4.6	11.7	10.7	0
31	Babson Col.	1	1	1165	1	7.6	26.4	18	0
32	Boston Col.	1	1	1285	1	7.5	26.8	15.86	0
33	Boston U.	1	1	1235	1	7	27.9	14.46	0
34	Bowdoin Col.	1	1	1345	1	6	27.8	13.64	0
35	Bryant Col.	0	1	1080	1	6.7	20.6	18	0

Id	School	Term	Type	SAT	TOEFL	Room Board (USD 000s)	Total Cost (USD 000s)	Indebt-edness (USD 000s)	Outlier Influence Point
36	Bucknell U.	1	1	1255	1	5	25.4	12.5	0
37	Canisius Col.	1	1	1143	0	5.9	18.8	14.82	0
38	Carnegie Mellon U.	1	1	1335	1	6.1	25.6	15.68	0
39	Case Western Reserve U.	1	1	1330	1	5	22.2	26.03	1
40	Clark U.	1	1	1121	1	4.4	24.4	17.5	0
43	Col. of HolyCross	1	1	1275	1	6.7	26.8	12.63	0
41	Colby Col.	0	1	1275	1	5.7	27.9	11.63	0
42	Collgate U.	1	1	1300	1	5.9	27.6	9.24	0
44	Emory U.	1	1	1310	1	6.5	26.6	15.31	0
45	Fordham U.	1	1	1150	1	7.4	23.4	8.59	0
46	Franklin Marshall Col.	1	1	1260	1	4.5	26.4	11.5	0
47	George Washington U.	1	1	1235	1	6.9	26.7	14.37	0
48	Georgetown U.	1	1	1330	1	7.5	27.5	14.01	0
49	Gettysburg Col.	1	1	1200	1	4.8	26.4	11.75	0
50	Harvard U.	1	1	1465	1	7	28.9	11.65	0
51	Iona Col.	1	1	955	1	7.3	19.8	18	0
52	Lafayette Col.	1	1	1185	1	6.3	26.7	11.5	0
53	LaSalle U.	1	1	1105	0	6.7	20.8	11.7	0
54	Lehigh U.	1	1	1225	1	6	26.8	13.84	0
55	Manhattan Col.	1	1	952	1	7.1	22	9.27	0
56	NewYork U.	1	1	1260	1	7.8	28.6	17.32	0
57	Niagara U.	1	1	1065	0	5.4	17.6	11.58	0
58	Northeastern U.	0	1	1055	1	8.2	23.4	25.6	1
59	Northwestern U.	0	1	1350	1	6.1	24.2	11.98	0
60	Providence Col.	1	1	1185	1	6.7	22.9	17.5	0
61	Rice U.	1	1	1395	1	6	18	2.32	2
62	Rochester Inst. Tech	0	1	1185	0	6.1	21.8	17.5	0
63	Seattle U.	0	1	1100	1	5.3	19.5	12	0
64	Seton Hall U.	1	1	1030	1	7.1	20.8	14.9	0
65	Siena Col.	1	1	1095	0	5.4	17.6	18.25	0
66	Southern Methodist U.	1	1	1150	1	5.3	21.3	12.11	0
67	St. Bonaventure U.	1	1	1098	1	5.1	17.5	14	0
68	Stanford U.	0	1	1430	1	7.3	27.8	12.77	0
69	Syracuse U.	1	1	1180	1	7.2	24.3	14.5	0
70	Tulane U.	1	1	1270	1	6.3	27.5	13.85	0
71	U. Chicago	0	1	1370	1	7.3	28.8	14.07	0
72	U. Miami	1	1	1145	1	7.1	25.7	16.07	0
73	U. Notre Dame	1	1	1320	1	4.8	23.8	16.57	0
75	U. Portland	1	1	1135	0	4.5	18.9	13.9	0

Id	School	Term	Type	SAT	TOEFL	Room Board (USD 000s)	Total Cost (USD 000s)	Indebt-edness (USD 000s)	Outlier Influence Point
76	U. Scranton	0	1	1115	0	6.6	21.6	13.5	0
74	U. Pennsylvania	1	1	1355	1	7.5	28.6	17.62	0
77	Vanderbilt U.	1	1	1295	1	7.1	27.3	14.5	0
78	Villanova U.	1	1	1242	1	7	24.8	17.13	0
79	Wake Forest U.	1	1	1280	1	5.2	23.7	18.7	0
80	Yale U.	1	1	1450	1	6.7	28.9	13.57	0

Note: Source: Levine, D. M., Ramsey, P. P., & Smidt, R. K. (2001) Term: Semester-0, Other-1; Type: Public-0, Private-1, TOEFL – Test score English as a foreign language <550 – 0, >550 – 1; Outlier Influence Point: Neither – Outlier – 1, Influence Point – 2. Column "Outlier Influence Point" was added here, and that column is not part of the original Levine, Ramsey, and Smidt data.

Table A 2. Open source software used for analysis

Software	Version	Function	Citation
R	2.14.	Statistical analysis.	CRAN (2012)
Car	2.0-16	Enhanced linear regression modeling utilities, including scatter plot matrix.	Fox, Weisberg & Batres (2013)
Effects	2.2-	Enhanced interaction analysis and graphs for linear regression modeling.	Fox, Weisberg & Hong (2013)
Colorspace	1.2-1	Required graphics package for corrplot.	Ihaka et al (2013)
Corrplot	0.71	Enhanced correlation graphics and principal component analysis	Wei (2013)

Table A 3. Useful *R* functions for model analysis.

Package	Function	Does
Outliers	outlier()	finds outliers
Base	pairs()	plots scatterplot matrix.
Car	scatterplotMatrix()	plots scatterplot matrix.
Corrplot	corrplot	prepares correlograms
Base	factanal()	does factor analysis in order to estimate minimum number of independent variables from data.
Base	prcomp()	does principal component analysis in order to estimate minimum number of independent variables from data.
Car	influencePlot()	makes influential point plot
Effects	allEffects()	makes interaction chart for multivariable linear regression

Note: Except outliers, these functions are used in the *R* code listed in Addendum C.

Table A 4. Outlier and influence points in college cost data.

Id	School	Term	Type	SAT	TOEFL	Room Board (USD 000s)	Total Cost (USD 000s)	Indebt-edness (USD 000s)	Outlier Influence Point
6	Florida Int'l. U.	1	0	1135	0	2.7	10	4.14	2
7	Florida St. U.	1	0	1180	1	4.5	11.5	16.5	2
39	Case Western Reserve U.	1	1	1330	1	5	22.2	26.03	1
58	Northeastern U.	0	1	1055	1	8.2	23.4	25.6	1
61	Rice U.	1	1	1395	1	6	18	2.32	2

Note: Source: Levine, D. M., Ramsey, P. P., & Smidt, R. K. (2001) Term: Semester-0, Other-1; Type: Public-0, Private-1, TOEFL – Test score English as a foreign language <550 – 0, >550 – 1; Outlier Influence Point: Neither – Outlier – 1, Influence Point – 2. Column "Outlier Influence Point" was added here, and that column is not part of the original Levine, Ramsey, and Smidt data.

Table A 5. Correlation matrix for college cost data.

					Room	Total
	Term	**Type**	**SAT**	**TOEFL**	**Board**	**Cost**
Term	1.00	-0.11	-0.07	0.07	-0.20	-0.07
Type	-0.11	1.00	0.48 #	0.16	0.68 #	0.84 #
SAT	-0.07	0.48	1.00	0.37	0.29	0.67 #
TOEFL	0.07	0.16	0.37	1.00	0.14	0.39
Room Board	-0.20	0.68	0.29	0.14	1.00	0.70 #
Total Cost	-0.07	0.84	0.67	0.39	0.70	1.00

Source: Table A 1, above. # - Correlation values discussed in main text, above.

Table A 6. Exploratory factor analysis for college cost data.

	Indebtedness
Term	-0.08
Type	0.43 #
SAT	0.17
TOEFL	0.18
Room Board	0.35
Total Cost	0.45 #
Indebtedness	1.00

Source: Table A 1, above. # - Common factor values discussed in main text, above.

Table A 7 Principal component analysis of college cost data variables.

Rotation	**PC1**	**PC2**	**PC3**	**PC4**	**PC5**	**PC6**
Term	-0.0002	0.0040	-0.1109	0.4610	0.8804	-0.0097
Type	0.0019	-0.0758	0.0517	-0.3136	0.1608	-0.9313
SAT	0.9995	0.0314	0.0035	-0.0002	0.0006	-0.0002
TOEFL	0.0012	-0.0160	-0.0868	0.8182	-0.4432	-0.3555
RoomBoard	0.0030	-0.2005	0.9679	0.1396	0.0501	0.0317
TotalCost	0.0314	-0.9761	-0.2017	-0.0158	-0.0119	0.0716
Importance	**PC1**	**PC2**	**PC3**	**PC4**	**PC5**	**PC6**
Standard deviation	125.886	4.4788	0.8602	0.3717	0.3403	0.2342
Proportion of Variance	0.9987	0.0013	0.0001	0.0000	0.0000	0.0000
Cumulative Proportion	0.9987	0.9999	1.0000	1.0000	1.0000	1.0000

Source: Table A 1, above.

Table A 8. Signs from tests that evaluated candidate independent variables for use in regression modeling.

Test/ Variable	Correlation Matrix	EFA	PCA	Scatterplot	Decision on use in modeling
Term	N	N	N	N	N
Type	Y	Y	Y	Y	Y
SAT	N	Y	Y	N	Possible
TOEFL	N	N	N	N	N
Room Board	N	Y	Y	Y	Effect is subsumed in Total Cost
Total Cost	Y	?	Y	Y	Y

Note: *See* main text for discussed of the tests. Total Cost and Type were selected as the key candidates for factors in modeling. EFA – exploratory factor analysis; PCA – principal component analysis.

Table A 9. Linear regression models explored to initially select a best-fit-model.

No.	Model	Shapiro Wilk of e*	R^2	adj. R^2	Last Term AIC	Last Term Residual	F stat	df	B_0 p-value	B_{TC} p-value
1	Total Cost, with outliers-influence points	0.09	0.20	0.19	220.76	1201.76	19.92	78	2.34E-04	2.69E-05
2	Total Cost - Room Board, with outliers-influence points	0.06	0.19	0.18	222.14	1222.58	18.25	78	5.89E-07	5.42E-05
3	Total Cost, no outlier-influence points	0.84	0.24	0.23	175.54	738.59	23.31	73	7.56E-06	7.38E-06
4	Type, no outlier-influence points	0.14	0.24	0.23	176.17	744.79	22.50	73	5.60E-28	1.01E-05
5	Type+TC, no outlier-influence points	0.69	0.24	0.23	175.54	738.59	23.31	73	7.56E-06	7.38E-06
6	**TC+TC^2, no outlier-influence points**	**0.93**	**0.36**	**0.35**	**164.46**	**620.34**	**20.55**	**72**	**2.22E-02**	**6.55E-05**
7	SAT+Type+TC+TC^2, no outlier-influence points	0.97	0.36	0.35	164.46	620.34	20.55	72	2.22E-02	6.55E-05
8	SAT+TC+TC^2, no outlier-influence points	0.96	0.36	0.35	164.46	620.34	20.55	72	2.22E-02	6.55E-05
9	Type+TC+TC^2, no outlier-influence points	0.93	0.36	0.35	164.46	620.34	20.55	72	2.22E-02	6.55E-05
10	RB+RB^2+RB^3+TC+TC^2, no outlier-influence points	0.97	0.38	0.35	166.16	601.63	10.84	70	7.86E-02	1.51E-01

Note: The initial best-fit-model was Model No. 6.

Table A 10. Coefficients of the initial best-fit-model.

```
                              Formula
M = Indebtedness ~ TotalCost + I(TotalCost^2), data = collegeData, subset = (isOutlier ==
0), direction = "forward")
                              Residuals:
Min           1Q           Median 3Q        Max
-6.3129       -2.1838      0.0496  1.9723   7.4865
                              Coefficients:
                    Estimate     Std.Error    t value      Pr(>|t|)
(Intercept)         -12.19595    5.21699      -2.338       0.022184 *
TotalCost           2.32769      0.54902      4.240        6.55e-05 ***
I(TotalCost^2)      -0.04998     0.01349      -3.705       0.000412 ***
Signif. codes:   0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1


Residual standard error: 2.935 on 72 degrees of freedom
Multiple R-squared: 0.3634,     Adjusted R-squared: 0.3457
F-statistic: 20.55 on 2 and 72 DF,  p-value: 8.713e-08
```

Note: Best-fit-model characteristics returned from *R* software for Model No. 6 in Table A 9, above.

Table A 11. Coefficients of the public college only total cost model.

```
lm(formula = Indebtedness ~ TotalCost, data = collegeData, direction = "forward")
Residuals:
Min            1Q             Median         3Q             Max
-5.6814        -2.6668        -0.4329        2.1332         8.1169


Coefficients:
               Estimate       Std. Errorr    t value  Pr(>|t|)
(Intercept)    -1.2170        3.5027         -0.347         0.73086
TotalCost      0.8348         0.2430         3.435          0.00186 **
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1


Residual standard error: 3.587 on 28 degrees of freedom
Multiple R-squared: 0.2965,     Adjusted R-squared: 0.2714
F-statistic:  11.8 on 1 and 28 DF,  p-value: 0.001865
```

Note: Model characteristics returned from *R* software for only public colleges in Table A 9, above.

Addendum B - Figures

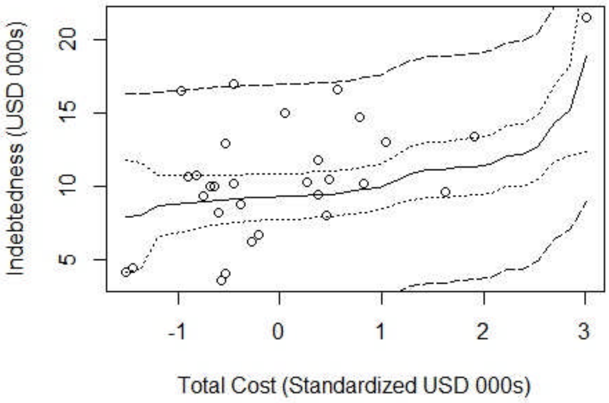

Figure B 1. Initial best-fit-model of total cost to student indebtedness (standardized). Model No. 6 in Table A 9, above.

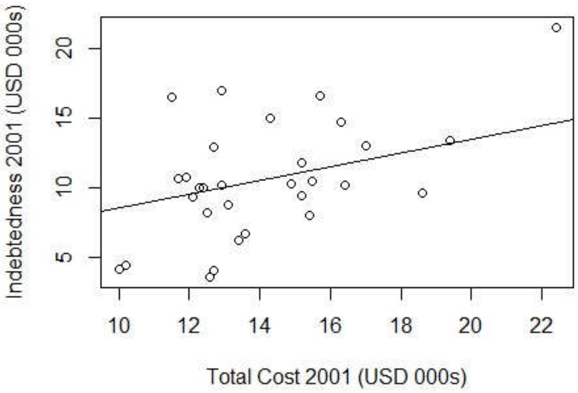

Figure B 2. The indebtedness ~ TC linear model for only the public colleges in Table A 9, above.

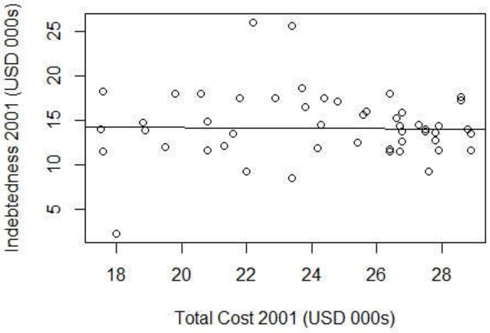

Figure B 3. The indebtedness ~ TC linear model for only the private colleges in Table A 9, above.

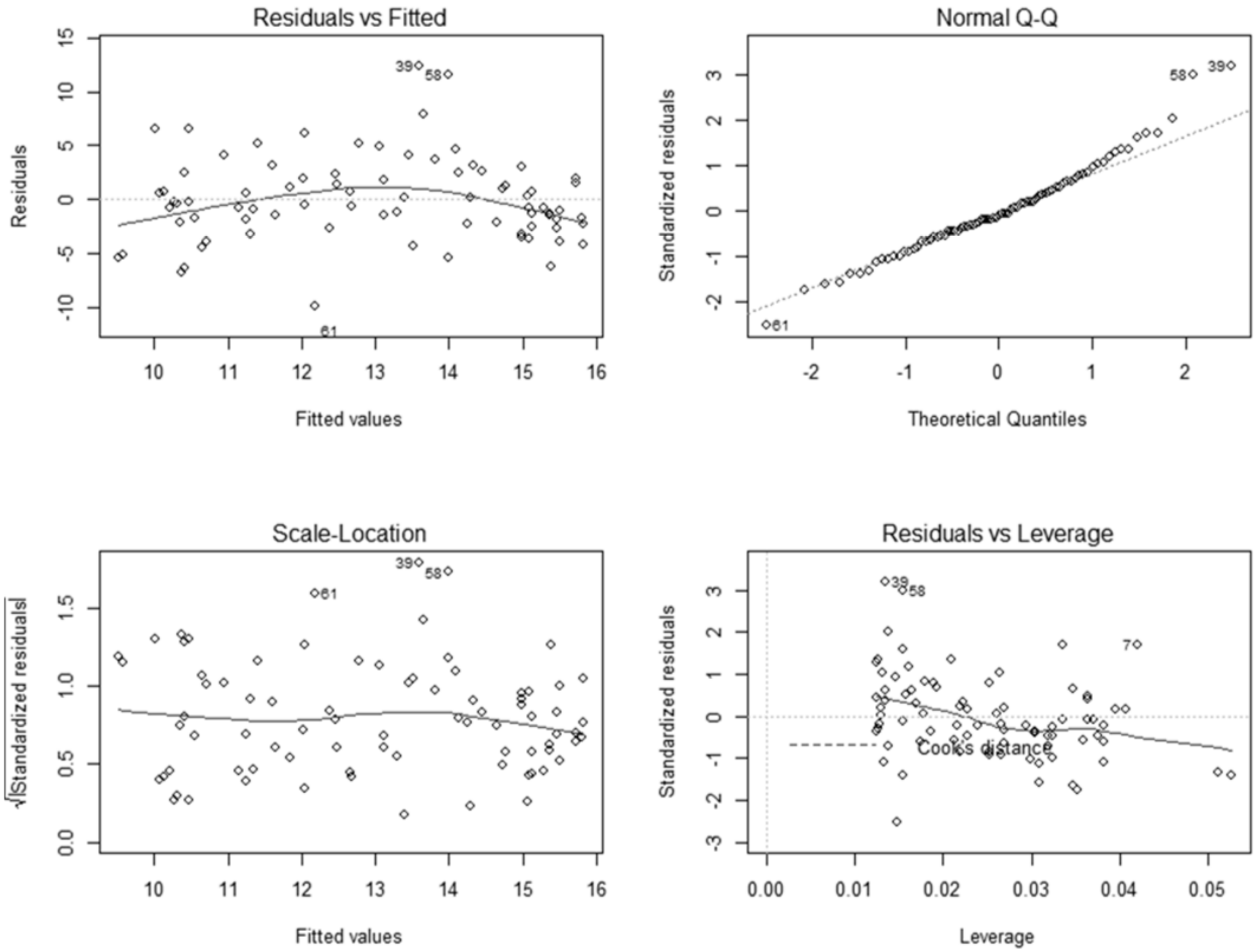

Figure B 4. Four box residuals analysis plot of Model No. 1 (I~TC for all data) in Table A 9 , above.

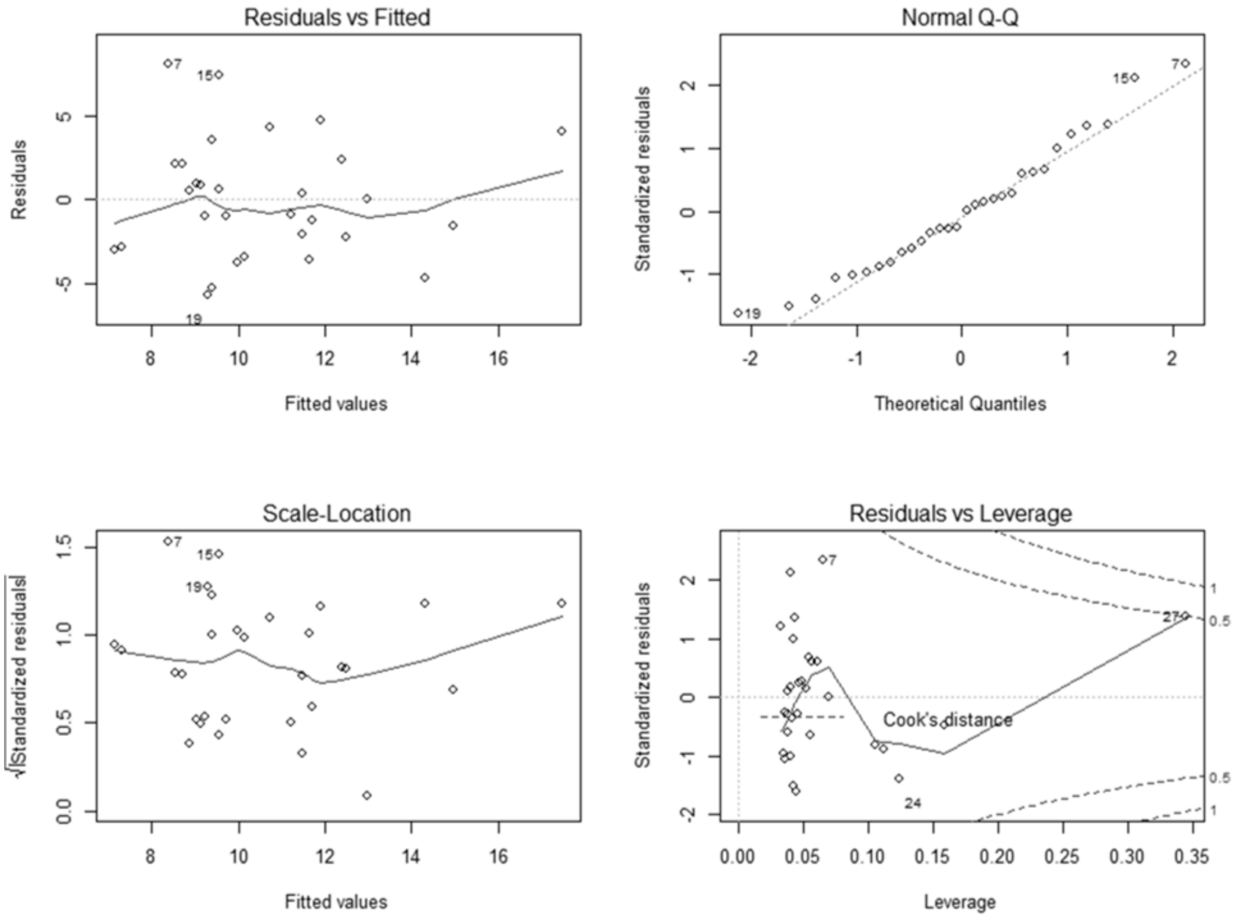

Figure B 5. Four box residuals analysis plot of the public college only data model (I~TC).

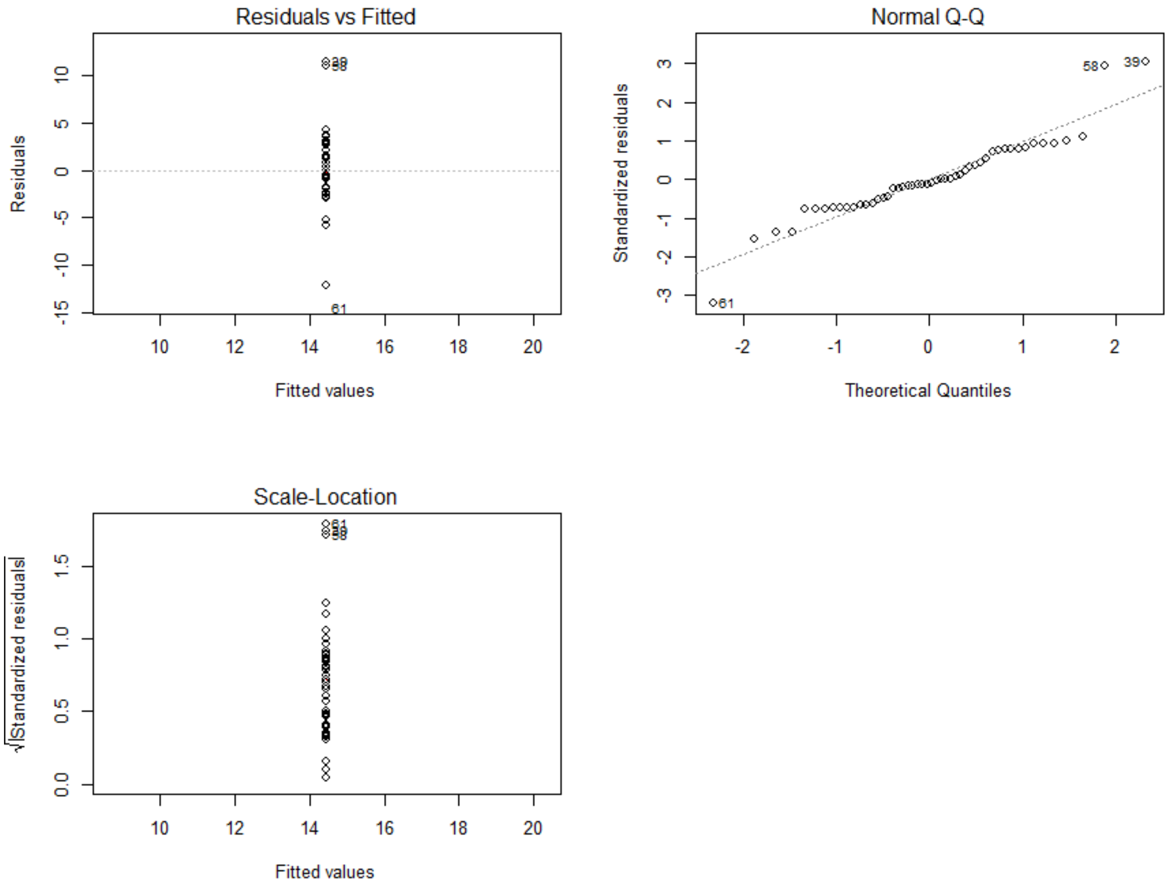

Figure B 6. Four box residuals analysis plot of the private college only data model (I~TC).

Note: The residuals are not randomly dispersed. *Compare* the dispersal of residuals in this figure (private only total cost linear model) and in Figure B 5 (public only total cost linear model) to Figure B 4 (private and public total cost linear model).

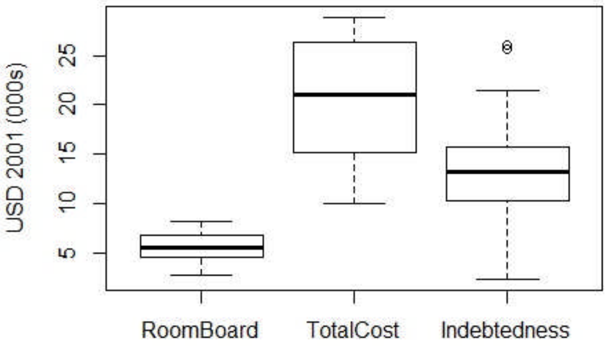

Figure B 7. Distribution of college cost data. Source: Table A 1, above. Outlier points in Indebtedness columns are Case Western Reserve Univ. and Northeastern Univ.

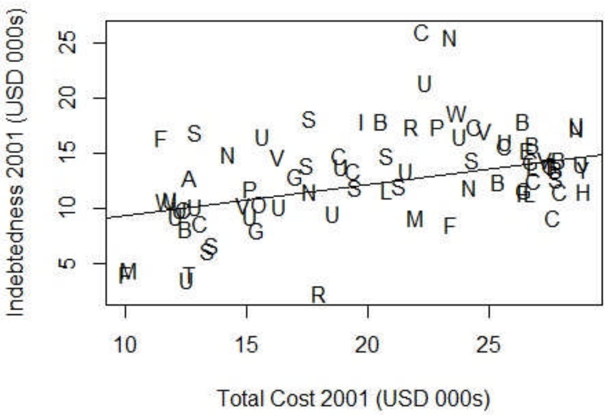

Figure B 8. Outlier identification from preliminary Indebtedness to Total Cost linear model. Source: Table A 1, above. Outlier points near (22,25) are Case Western Reserve Univ. (C) and Northeastern Univ. (N).

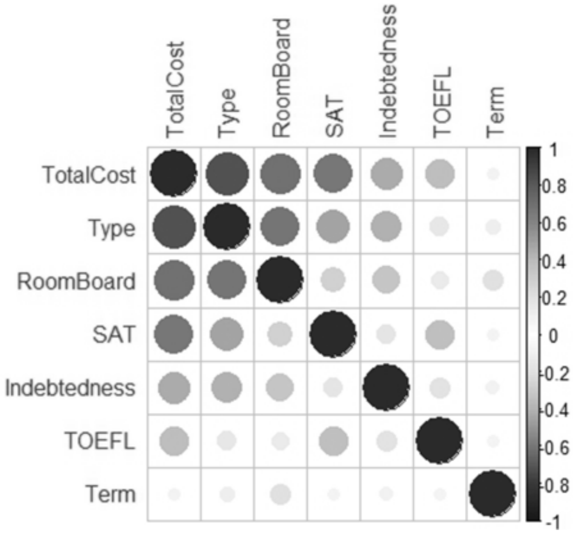

Figure B 9. Principal component analysis correlogram for college cost data in Table A 1, above. The grid is organized by PCA strength, and the size and color of the dots indicate the value of Pearson's correlation coefficient R. The scale on the right-hand side of the figure indicates the numerical value of R.

[Intentional Blank]

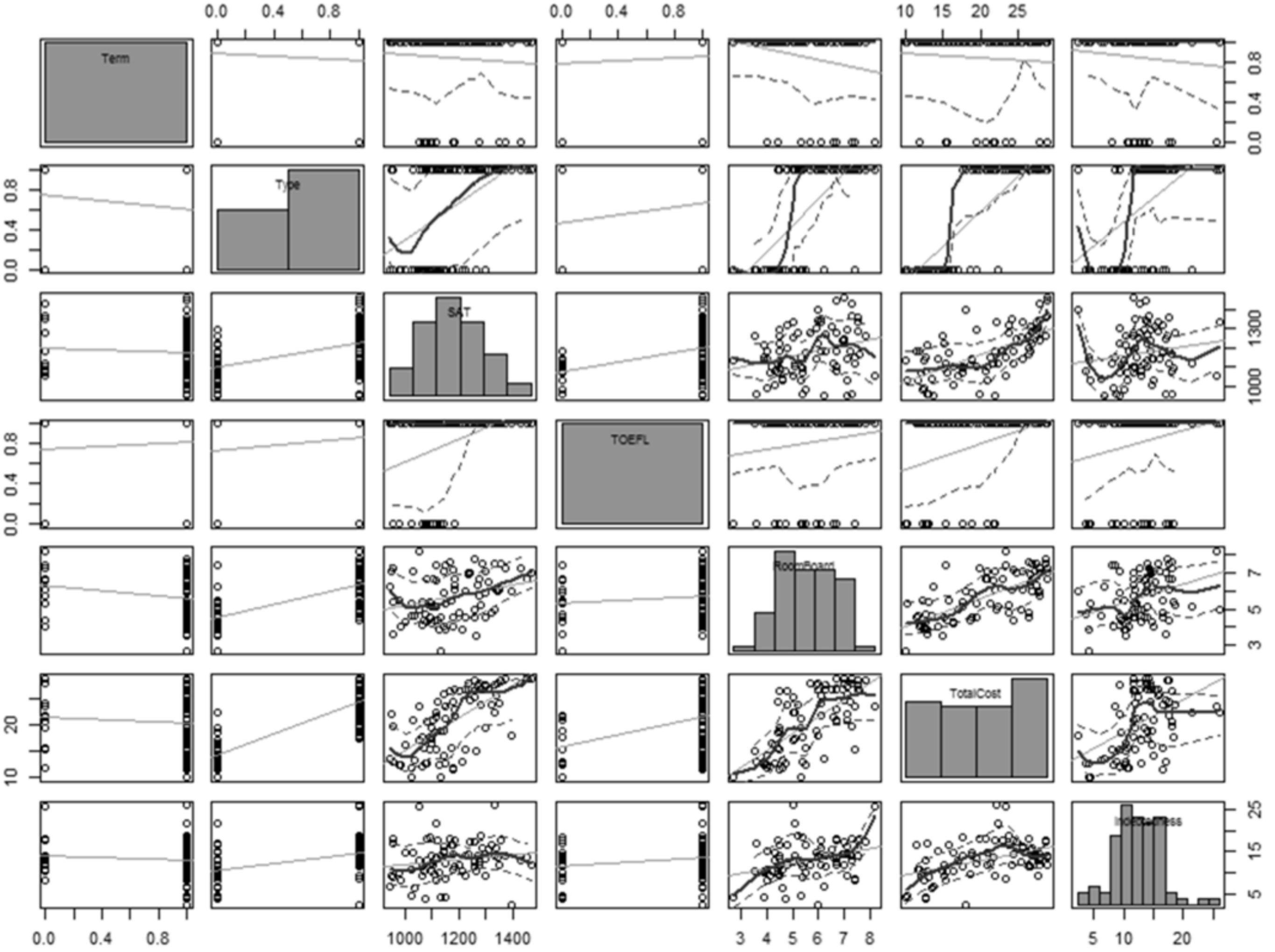

Figure B 10. Scatterplot matrix of college cost data in Table A 1, above.

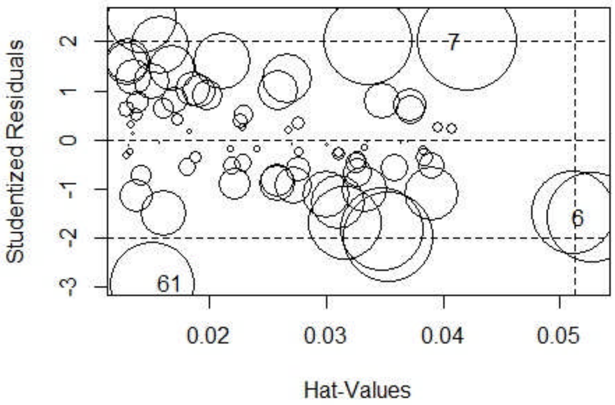

Figure B 11. Influence point plot based on preliminary model Indebtedness to Total Cost, for college cost data in Table A 1, above.

[Intentional Blank]

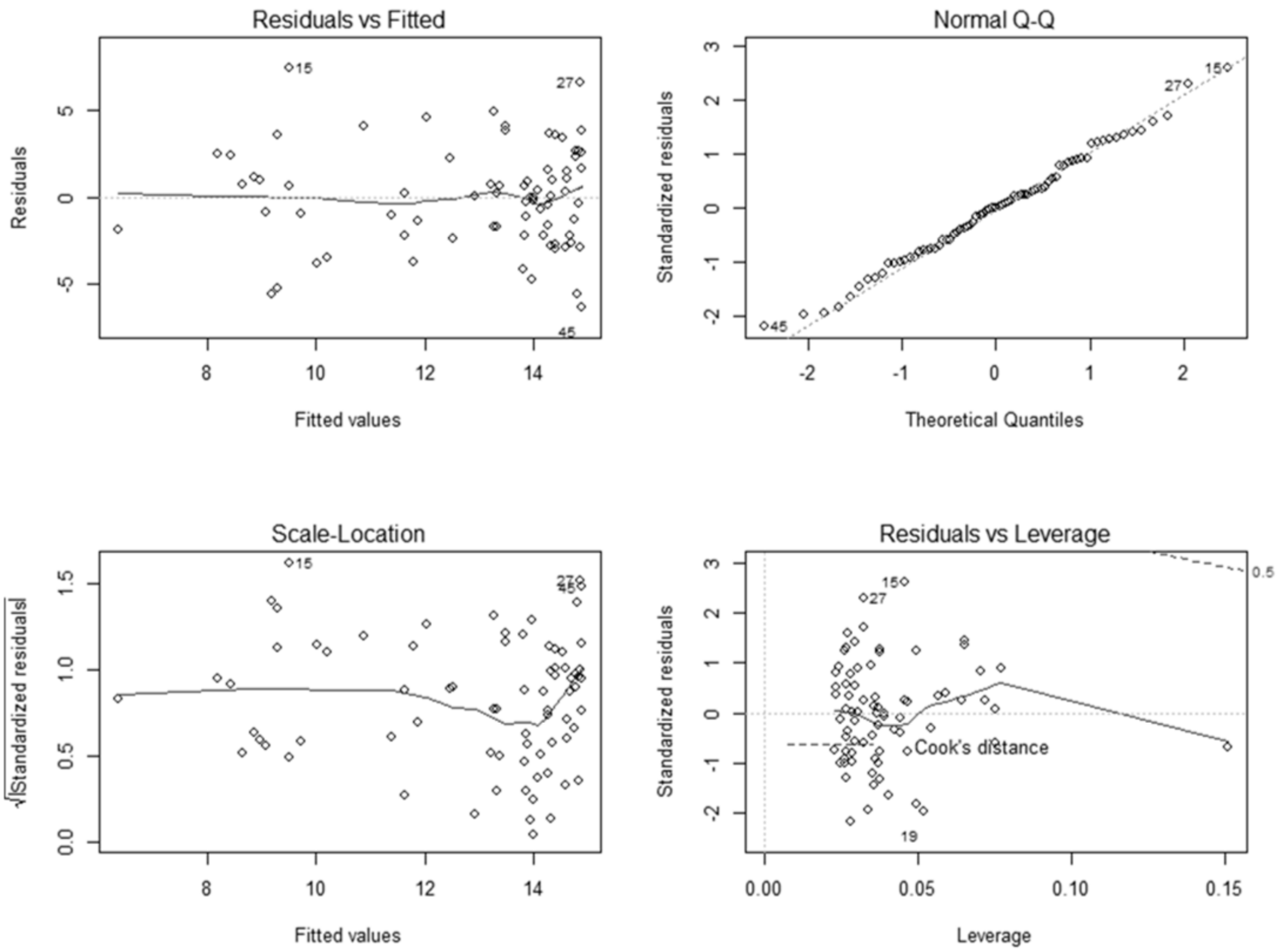

Figure B 12. Four box residuals analysis plot of Model No. 6, the initial best-fit-model, in Table A 9 , above.

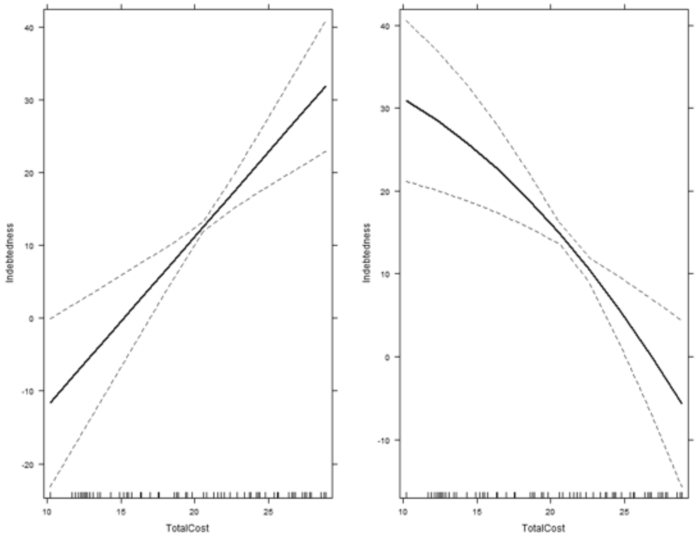

Figure B 13. Interaction plot of Model No. 6, the best-fit-model, in Table A 9 , above. (a) Left. Interaction of Total Cost. (b) Right. Interaction of Total Cost^2.

Note: Because there is only one independent variable, the graphs show no interaction between two independent variables. The graphs do show how the slope of TC and TC^2 change across the domain of TC.

Addendum C – R Code

Listing C 1. A List of R Code Listings

Addendum C No.	Purpose of R Code Listed
2	Distribution analysis
3	Correlation, factor and principal component analysis
4	Select best-fit-model by step AIC and elimination
5	Compute standardized linear regression model coefficients
6	Supplemental factor analysis and principal component analysis in order to estimate the sufficient number of factors
Omitted	Derivative code of the above used in evaluating public only and private only college distributions and models.

Listing C 2. R Code for distribution analysis.

```
# Stat 3080, R Group Chun Project 4
# Kurt Fisher and Elijah Neilson
# Task 1 - Get and Clean Data, Analyze Distribution, and Identify Outliers
# Version Date: 3-31-2013
# Version: 1.1
# File: 20130331RProT1Distribution1_3.R

# Uses: schooldatamodASCII.txt, schoolacronyms.txt

# Makes the following figures and tables that will be used in writing the report.
# Makes automatically:
#   Listing1DataCleaned.csv - Cleaned data with outlier's marked.
#   Listing2DistributionSummary. csv - Compares means of Total Cost and Indebtedness at
#      public and private colleges.
#   Listing3FiveNumberSummary.csv - Five number summary of independent variables
#   Listing4CorrelationMatrix.csv - Correlation matrix for all variables
#   Listing5CorrelationIndebtedness.csv - Just the key correlation column for Indebtedness
Make manually:
#   Fig01CostDebtDistribution.jpeg - Boxplot of Roomboard TotalCost Indebtedness
#      This figure identifies two outliers in Indebtedness - Northwestern and Case Western Res.
#   Fig02TotalCosttoIndebtPlot.jpeg - Scatter plot that shows outliers for Total Cost to Indebtedness
#   Fig03TotalCosttoIndebtIds.jpeg  - Scatter plot that shows outliers with points as letters.

# Set directory
#   Usage Note:
#   Change this path to your personal working directory.
#   Put data files schooldatamodASCII.txt and schoolacronyms.txt into that directory for reading.
setwd("C:\\Users\\fisherka\\Documents\\00000000UofUSpr2013\\3080Stats\\ProjectSec")
#
source("C:\\Users\\fisherka\\Documents\\00000000UofUSpr2013\\3080Stats\\ProjectSec\\20130331RProT1Dist
ribution1_3", echo=TRUE) #    Note: source is with echo ON option.
#
source("C:\\Users\\fisherka\\Documents\\00000000UofUSpr2013\\3080Stats\\ProjectSec\\20130331RProT1Dist
ribution1_3", echo=FALSE) #    Note: source is with echo OFF option.

# 1) Get data
collegeData = read.table("schooldatamodASCII.txt", header=TRUE)
collegeData <- data.frame(collegeData)
# Coerce the data into a numeric type in order to allow use with later histogram plotting.
#    Coercion is not needed for visual data analysis of correlation packages.
collegeData$Term <- as.numeric(collegeData$Term)
collegeData$Type <- as.numeric(collegeData$Type)
collegeData$SAT <- as.numeric(collegeData$SAT)
collegeData$TOEFL <- as.numeric(collegeData$TOEFL)
collegeData$RoomBoard <- as.numeric(collegeData$RoomBoard)
collegeData$TotalCost <- as.numeric(collegeData$TotalCost)
collegeData$Indebtedness <- as.numeric(collegeData$Indebtedness)
# append table of acronyms for scatterplotting outliers
collegeAcronyms <- read.table("schoolacronyms.txt", header=TRUE) # read the acronym list
collegeAcronyms <- collegeAcronyms[3] # get the acronym column
```

```
collegeData[9] <- collegeAcronyms # append the acronym column to the data table
col10 <- rep(0,80) # make a variable to exclude outliers 0=False 1 = True
col10 <- list(isOutlier=col10) # initialize to all False
collegeData <- cbind(collegeData,col10) # append the column to the data table
collegeData # Display the data

# 2) Build the distribution summary of means and sd's.
c1 <- c("public") # get mean and sd of TC and Indebtedness for public colleges.
c2 <- length(which(collegeData$Type == 0))
c3 <- sapply(subset(collegeData, Type==0, select=c(TotalCost,Indebtedness)),mean)
c4 <- sapply(subset(collegeData, Type==0, select=c(TotalCost,Indebtedness)),sd)
row1 <-
cbind(typeCollege=c1,countColleges=c2,meanCollegeTC=c3[1],sdCollegeTC=c4[1],meanCollegeIndebtedness
=c3[2],sdCollegeIndebtedness=c4[2])
row1 <- data.frame(row1)
d1 <- c("private") # get mean and sd of TC and Indebtedness for public colleges.
d2 <- length(which(collegeData$Type == 1))
d3 <- sapply(subset(collegeData, Type==1, select=c(TotalCost,Indebtedness)),mean)
d4 <- sapply(subset(collegeData, Type==1, select=c(TotalCost,Indebtedness)),sd)
row2 <-
cbind(typeCollege=d1,countColleges=d2,meanCollegeTC=d3[1],sdCollegeTC=d4[1],meanCollegeIndebtednes
s=d3[2],sdCollegeIndebtedness=d4[2])
row2 <- data.frame(row2)
distributionSummary <- rbind(row1,row2) # combine the data for public and private into one table.
distributionSummary # display the table.

# 3) Display a distribution summary of the variables and distribution plots
#    Discover and mark outliers
#    a) Display the summary
collegeData5NumSummary = summary(collegeData[2:8]) # columns 2 to 8 contain numeric data that can be
summarized.
#  Display the five number summary of Total Cost and Indebtedness
summary(collegeData$TotalCost)
summary(collegeData$Indebtedness)

#    b) Display a boxplot and TotalCost to Indebtedness x,y plots. Save Plots
#  Distribution plot
boxplot(collegeData[6:8], ylab="USD 2001 (000s)")
# Manually save Fig01CostDebtDistribution.jpeg # Boxplot of Roomboard TotalCost Indebtedness. Show
outliers.
#  Total Cost to Indebtedness Plot
plot(collegeData$TotalCost,collegeData$Indebtedness, ylab="Indebtedness 2001 (USD 000s)", xlab="Total
Cost 2001 (USD 000s)")
z <- line(collegeData$TotalCost,collegeData$Indebtedness)
abline(coef(z))
# Manually save Fig02TotalCosttoIndebtPlot.jpeg # Scatter plot that shows outliers for Total Cost to
Indebtedness
#  Total Cost to Indebtedness Plot with Letter Identifiers
plot(collegeData$TotalCost,collegeData$Indebtedness, ylab="Indebtedness 2001 (USD 000s)", xlab="Total
Cost 2001 (USD 000s)", pch=as.character(collegeData$SchoolAcron))
abline(coef(z))
# Manually save Fig03TotalCosttoIndebtIds.jpeg # Scatter plot that shows outliers with points as letters.

#    c) Identify and mark the Indebtedness outliers
collegeData[order(collegeData$Indebtedness, decreasing=TRUE),][1:2,]
```

```
collegeData[39,10]= 1 # Set Case Western Reserve as an indebtedness outlier.
collegeData[58,10]= 1 # Set Northeastern Univ. as an indebtedness outlier.
#       Note there is one low Indebtedness value that is not an outlier - Rice Univ.
collegeData[order(collegeData$Indebtedness, decreasing=FALSE),][1,]
#       Redisplay the data table.
collegeData # Redisplay the data table.

# 4) Display a correlation matrix
cData <- cor(collegeData[2:8])
cData # Display a correlation matrix
cDataIndebtedness <- cor(collegeData[2:7],collegeData[8]) # Display just the Indebtedness correlations
#    Note that the highest correlated variables to Indebtedness are Type of Total Cost.
cDataIndebtedness # Indebtedness correlation matrix
#    Note that Roomboard and TotalCost are highly correlated. Room Board does not provide much new information beyond TotalCost.
cor(collegeData$RoomBoard,collegeData$TotalCost)

# 5) Export data tables
write.csv(collegeData, file="Listing1DataCleaned.csv")
write.csv(distributionSummary, file="Listing2DistributionSummary.csv")
write.csv(collegeData5NumSummary, file="Listing3FiveNumberSummary.csv")
write.csv(cData, file="Listing4CorrelationMatrix.csv")
write.csv(cDataIndebtedness, file="Listing5CorrelationIndebtedness.csv")
```

Listing C 3. R Code for correlation, factor and principal component analysis.

```
# Stat 3080, R Group Chun Project 4
# Kurt Fisher and Elijah Neilson
# Task 2 - Visual Analysis of Contributing Correlation Variables
# Version Date: 4-19-2013
# Version: 1.3
# File: 20130331RProT2FactorAnalysis1_3.R

# Usage note: This task requires specialty packages car, corrplot, and nnet. However,
#    incompatiablities in those packages case runtime errors in the boxplot package.
#    After running this task, remember to check that packages car, corrplot, and nnet
#    are disabled. Those packages should be automatically unloaded.

# Uses: schooldatamodASCII.txt, schoolacronyms.txt

# Makes the following figures and tables that will be used in writing the report.
# Make automatically:
#    Listing6EFATable.csv - Exploratory factor analysis
#    Listingz9PCATable1.csv - Principal component analysis rotation matrix
#    Listingz9PCATable2.csv - Principal component analysis summary.

# Make manually
#    Fig04CorrelationScatter.jpeg - A scatterplot correlation matrix. (Use the Pairs Function Plot).
#    Fig05PrincipalComponentAnalysis.jpeg - Identifies candidate principal independent variables
#       as TotalCost and Type-of-Institution. (Use the cData.FPC plot - ordered by PCA result.)
#    Usage note: Zoom the images before saving.

# Shows: TotalCost and Type-of-Institution are the best candidate independent variables.
#        SAT is a potential third variable.

# Set directory
#   Usage Note:
#   Change this path to your personal working directory.
#   Put data files schooldatamodASCII.txt and schoolacronyms.txt into that directory for reading.
setwd("C:\\Users\\fisherka\\Documents\\00000000UofUSpr2013\\3080Stats\\ProjectSec")
#
source("C:\\Users\\fisherka\\Documents\\00000000UofUSpr2013\\3080Stats\\ProjectSec\\20130331RProT2Fac
torAnalysis1_2.R", echo=TRUE) #    Note: source is with echo ON option.
#
source("C:\\Users\\fisherka\\Documents\\00000000UofUSpr2013\\3080Stats\\ProjectSec\\20130331RProT2Fac
torAnalysis1_2.R", echo=FALSE) #    Note: source is with echo OFF option.

# 1) Get data
collegeData = read.table("schooldatamodASCII.txt", header=TRUE)
collegeData <- data.frame(collegeData)
# Coerce the data into a numeric type in order to allow use with later histogram plotting.
#    Coercion is not needed for visual data analysis of correlation packages.
collegeData$Term <- as.numeric(collegeData$Term)
collegeData$Type <- as.numeric(collegeData$Type)
collegeData$SAT <- as.numeric(collegeData$SAT)
collegeData$TOEFL <- as.numeric(collegeData$TOEFL)
```

```
collegeData$RoomBoard <- as.numeric(collegeData$RoomBoard)
collegeData$TotalCost <- as.numeric(collegeData$TotalCost)
collegeData$Indebtedness <- as.numeric(collegeData$Indebtedness)
# append table of acronyms for scatterplotting outliers
collegeAcronyms <- read.table("schoolacronyms.txt", header=TRUE) # read the acronym list
collegeAcronyms <- collegeAcronyms[3] # get the acronym column
collegeData[9] <- collegeAcronyms # append the acronym column to the data table
col10 <- rep(0,80) # make a variable to exclude outliers 0=False 1 = True
col10 <- list(isOutlier=col10) # initialize to all False
collegeData <- cbind(collegeData,col10) # append the column to the data table

#2) Mark known outliers previously identified
collegeData[39,10]= 1 # Set Case Western Reserve as an indebtedness outlier.
collegeData[58,10]= 1 # Set Northeastern Univ. as an indebtedness outlier.
#       Note there is one low Indebtedness value that is not an outlier - Rice Univ.
collegeData[order(collegeData$Indebtedness, decreasing=FALSE),][1,]
collegeData # Display the data

# 3) Rebuild the correlation matrix
cData <- cor(collegeData[2:8])
cData # Display a correlation matrix

# 4) Do visual correlation analysis

#     a) Do visual analysis using R's built-in pairs function.
pairs(~Term+Type+SAT+TOEFL+RoomBoard+TotalCost+Indebtedness, data=collegeData,
panel=panel.smooth, main="Scatter Plot Matrix of College Data - Pairs Function")

#     b) Do visual analysis using car - Companion for Applied Analysis package
library(car) # install the Companion for Applied Analysis package
#  Plot and save scattergram matrix analysis
scatterplotMatrix(~Term+Type+SAT+TOEFL+RoomBoard+TotalCost+Indebtedness, data=collegeData,
diagonal="histogram", main="Scatter Plot Matrix of College Data - CAR Package")
# Manually save Fig04CorrelationScatter.jpeg # A scatterplot correlation matrix. (Use the Pairs Function Plot).
#   Usage Note: Recommend manually opening and saving this image at a higher resolution.
#   At a higher resolution, the file is a png file.

#     c) Do visual analysis use corrplot
library(corrplot) # install the correlation plot analysis package
# Make ordered sets of the data using corrplot package
order.AOE <- corrMatOrder(cData, order="AOE") # Arrange correlations by angular order of eigenvectors
order.FPC <- corrMatOrder(cData, order="FPC") # Arrange correlations by first principal component analysis
cData.AOE <- cData[order.AOE,order.AOE ]
cData.FPC <- cData[order.FPC,order.FPC ]
corrplot(cData) # Make the correlation plots - unordered
corrplot(cData.AOE)  # Make the plot using Angular order of Eigenvectors
#  Plot and save the PCA color correlogram.
corrplot(cData.FPC)  # Make the correlation plot using Principal Component Analysis order
# Manually save Fig05PrincipalComponentAnalysis.jpeg # Identifies candidate principal independent variables

5) Prepare principal component analysis data matrix
collegeDataDF <- data.frame(collegeData[2:7])
pc.cr <- prcomp(collegeDataDF)
pc.cr$rotation
write.csv(pc.cr$rotation, file="Listingz9PCATable1.csv")
```

```
pc.crsum <- summary(pc.cr)
pc.crsum$importance
write.csv(pc.crsum$importance, file="Listingz9PCATable2.csv")

# 6) Do Exploratory factor analysis (EFA)
#     R. Kabacoff. 2001. Exploratory factor analysis. Sec. 14.3. In: R in Action. Manning Publ. p. 342-343.
#     Use: EFA looks for common factors in the all the dependent variables without
#          reference to the dependent variable. EFA looks for how the independent
#          variables might combine to predict and unobserved and unknown dependent variable,
#          called a common factor.
options(digits=2)
covariances <- cov(collegeData[2:7]) # covariances without Indebtedness
correlations <- cov2cor(covariances)
correlations # Display the EFA table.
# Note: The highest factor is 0.838 from the combination TotalCost,Type.
#       The fourth highest factor is SAT at 0.669. The third highest, RoomBoard,
#       is autocorrelated with TotalCost.
write.csv(correlations, file="Listing6EFATable.csv")

detach("package:car", unload=TRUE)
detach("package:corrplot", unload=TRUE)
detach("package:nnet", unload=TRUE)
# Usage Note:  Remember, after running this task, check that packages car, corrplot, and nnet
# are disabled.
```

Listing C 4. R Code to select a best-fit-model by step AIC and elimination analysis.

```
# Stat 3080, R Group Chun Project 4
# Kurt Fisher and Elijah Neilson
# Task 3 - Use step AIC and elimination analysis to find the best fit model
# Version Date: 4-1-2013
# Version: 1.2
# File: 20130401RProT3BestFitModel1_2.R

# Usage note: This task requires specialty packages
#    car, colorspace, effects, nnet. It uses the standard lattice package. However,
#    incompatiablities in those packages case runtime errors in the boxplot package.
#    After running this task, remember to check that packages
#    car, colorspace, effects, and nnet
#    are disabled. Disabling of those packages should occur automatically.

# Uses: schooldatamodASCII.txt, schoolacronyms.txt

# Makes the following figures and tables that will be used in writing the report.
# Make automatically:
#     Listing7BestFitAnalysis.csv - Best Fit Analysis of several models
#     Listing8BestFitModelDescrip.txt - Best Fit Model Summary
# Manually save
#  Note: automatic save is not working due to package conflicts.
#              Fig06InfluencePointPlot.jpeg - Influence point plot of college data. See R-in-Action, p. 204.
#     Fig07FourBoxPlotAllPoints.jpeg - R's standard four box analysis plot with outliers-influence points
regressed on model TotalCost only.
#        Shows quadratic residuals on TotalCost. Shows effect of outliers and influence points on regression.
#     Fig08PlotBestFitModelFourBox.jpeg - R's standard four box analysis plot without outlier-influence points
regressed on model TC+TC^2.
#        Shows best fit model analysis.
#     Fig09PlotBestFitModelInteraction.jpeg = Shows specialty plot of effects. See Help(effect) and R-in-Action,
p. 187.
#        Useage Note: Second interaction plot for TC^2 is mislabeled. Label manually.
#     Fig10BestFitModelPredConfInt.jpeg - Plot of Best-Fit-Model without x transforms with confidence and
prediction lines.

# Shows - Optimized best-fit mixed model using automatic step AIC computation
#        and elimination testing.  TC+TC^2 is the best model.

# Set directory
#   Usage Note:
#   Change this path to your personal working directory.
#   Put data files schooldatamodASCII.txt and schoolacronyms.txt into that directory for reading.
setwd("C:\\Users\\fisherka\\Documents\\00000000UofUSpr2013\\3080Stats\\ProjectSec")
#
source("C:\\Users\\fisherka\\Documents\\00000000UofUSpr2013\\3080Stats\\ProjectSec\\20130401RProT3Bes
tFitModel1_2.R", echo=TRUE) #    Note: source is with echo ON option.
#
source("C:\\Users\\fisherka\\Documents\\00000000UofUSpr2013\\3080Stats\\ProjectSec\\20130401RProT3Bes
tFitModel1_2.R", echo=FALSE) #    Note: source is with echo OFF option.
```

```
# 1) Get data
collegeData = read.table("schooldatamodASCII.txt", header=TRUE)
collegeData <- data.frame(collegeData)
# Coerce the data into a numeric type in order to allow use with later histogram plotting.
#    Coercion is not needed for visual data analysis of correlation packages.
collegeData$Term <- as.numeric(collegeData$Term)
collegeData$Type <- as.numeric(collegeData$Type)
collegeData$SAT <- as.numeric(collegeData$SAT)
collegeData$TOEFL <- as.numeric(collegeData$TOEFL)
collegeData$RoomBoard <- as.numeric(collegeData$RoomBoard)
collegeData$TotalCost <- as.numeric(collegeData$TotalCost)
collegeData$Indebtedness <- as.numeric(collegeData$Indebtedness)
# append table of acronyms for scatterplotting outliers
collegeAcronyms <- read.table("schoolacronyms.txt", header=TRUE) # read the acronym list
collegeAcronyms <- collegeAcronyms[3] # get the acronym column
collegeData[9] <- collegeAcronyms # append the acronym column to the data table
col10 <- rep(0,80) # make a variable to exclude outliers 0=False 1 = True
col10 <- list(isOutlier=col10) # initialize to all False
collegeData <- cbind(collegeData,col10) # append the column to the data table

# 2) Mark known outliers previously identified
collegeData[39,10]= 1 # Set Case Western Reserve as an indebtedness outlier.
collegeData[58,10]= 1 # Set Northeastern Univ. as an indebtedness outlier.
#       Note there is one low Indebtedness value that is not an outlier - Rice Univ.
# collegeData[order(collegeData$Indebtedness, decreasing=FALSE),][1,] # Disabled, use in prior tasks

# 3) Find and remove influence points
# Identify influence points
lm1 <- lm(Indebtedness ~ TotalCost, subset=(isOutlier==0), data = collegeData, direction="forward")
library(car)
#   Make influence plot and save.
influencePlot(lm1)
# Useage Note: Manually save Fig06InfluencePointPlot.png - Influence point plot of college data. See R-in-
Action, p. 204.

#   Mark points for exclusion from best model fit.
collegeData[6,10]= 2 # Mark influence point - Florida Int'l Univ.
collegeData[7,10]= 2 # Mark influence point - Florida State Univ.
collegeData[61,10]= 2 # Mark influence point - Rice Univ.

collegeData # Display the data table with revisions

# 4) Optimize a linear model using AIC (Akaike's Information Criterion)
#    The better-fit model has a lower AIC.

# 4a) Do a linear model with degree 1 and only the most significant term
# Total Cost, with outliers-influence points
summary(lm1 <- lm(Indebtedness ~ TotalCost, data = collegeData, direction="forward")) # Make the model
slm1 <- step(lm1) # summarize the model
slm1Summary <- summary(slm1) # make the summary
summary(slm1) # display the summary
slm1$anova # check the df and residuals

#  Plot and save the four box analysis plot including outlier-influence points.
par(mfrow=c(2,2))
```

```
plot(slm1) # Plot all boxes on first run in order to see the big regression analysis picture
#                  The first box plots fitted (x-axis) to residuals (y-axis).
par(mfrow=c(1,1))
# Usage Note: Manually save Fig07FourBoxPlotAllPoints.jpeg # R's standard four box analysis plot with outlier-influence points regressed on model TotalCost only.
#         Shows quadratic residuals on TotalCost. Shows effect of outliers and influence points on regression.
# Usage Note: Recommend zooming image and manually saving at a higher resolution.

# Report normality test
lm1ShaprioWTest <- shapiro.test(residuals(lm1)) # Check normality of the residuals
lm1ShaprioWTest
# Build an analysis results row
c1 <- c("Total Cost, with outliers-influence points")   #  Make the model name
c2 <- lm1ShaprioWTest$p.value # store Shapiro-Wilk test on residuals
c3 <- slm1Summary$r.squared # get model r^2
c4 <- slm1Summary$adj.r.squared # get model r^2
c5 <- slm1$anova$AIC[length(slm1$anova$AIC)]  # get last AIC
c6 <- slm1$anova[length(slm1$anova$AIC),5] # get last residual
c7 <- slm1Summary$fstatistic[1]  # get fstat value
c8 <- slm1Summary$fstatistic[3]  # get fstat df
c9 <- slm1Summary$coefficients[[1,4]] # p-value of Intercept
c10 <- slm1Summary$coefficients[[2,4]] # p-value of TotalCost
rowTemp <- list(TestName=c1,ShapiroWilkPValue=c2,RSqd=c3,AdjRSqd=c4,LastAIC=c5,LastResidual=c6,Fstat=c7,FstatDF=c8,InterceptPValue=c9,TotalCostPValue=c10,c10) # store a list with column names
rowTemp <- rowTemp[-11] # Delete error column
rowTemp <- data.frame(rowTemp)
resultsBestFitAnalysis <- rbind(rowTemp) # Append the analysis results to the model analysis summary table
resultsBestFitAnalysis

# 4b) Do a linear model with degree 1 and a net cost model
# Net Cost = Total Cost - Room Board, with outliers-influence points
netCost <- collegeData$TotalCost - collegeData$RoomBoard
summary(lm1 <- lm(collegeData$Indebtedness ~ netCost, data = collegeData, direction="forward")) # Make the model
slm1 <- step(lm1) # summarize the model
slm1Summary <- summary(slm1) # make the summary
summary(slm1) # display the summary
slm1$anova # check the df and residuals
par(mfrow=c(1,2)) # Plot the residuals and normality analysis graphs
plot(slm1, which=1)
plot(slm1, which=2)
par(mfrow=c(1,1))
lm1ShaprioWTest <- shapiro.test(residuals(lm1)) # Check normality of the residuals
lm1ShaprioWTest
# Build an analysis results row
c1 <- c("Total Cost - Room Board, with outliers-influence points")   #  Make the model name
c2 <- lm1ShaprioWTest$p.value # store Shapiro-Wilk test on residuals
c3 <- slm1Summary$r.squared # get model r^2
c4 <- slm1Summary$adj.r.squared # get model r^2
c5 <- slm1$anova$AIC[length(slm1$anova$AIC)]  # get last AIC
c6 <- slm1$anova[length(slm1$anova$AIC),5] # get last residual
c7 <- slm1Summary$fstatistic[1]  # get fstat value
c8 <- slm1Summary$fstatistic[3]  # get fstat df
c9 <- slm1Summary$coefficients[[1,4]] # p-value of Intercept
```

```
c10 <- slm1Summary$coefficients[[2,4]] # p-value of TotalCost
rowTemp <-
list(TestName=c1,ShapiroWilkPValue=c2,RSqd=c3,AdjRSqd=c4,LastAIC=c5,LastResidual=c6,Fstat=c7,FstatD
F=c8,InterceptPValue=c9,TotalCostPValue=c10,c10) # store a list with column names
rowTemp <- rowTemp[-11] # Delete error column
rowTemp <- data.frame(rowTemp)
resultsBestFitAnalysis <- rbind(resultsBestFitAnalysis,rowTemp) # Append the analysis results to the model
analysis summary table

# 4c) Do a linear model with degree 1 and only the most significant term
#     Now exclude the outliers and influence points
summary(lm1 <- lm(Indebtedness ~ TotalCost, subset=(isOutlier==0), data = collegeData, direction="forward"))
slm1 <- step(lm1) # summarize the model
slm1Summary <- summary(slm1) # make the summary
summary(slm1) # display the summary
slm1$anova # check the df and residuals
par(mfrow=c(1,2)) # Plot the residuals and normality analysis graphs
plot(slm1, which=1)
plot(slm1, which=2)
par(mfrow=c(1,1))
lm1ShaprioWTest <- shapiro.test(residuals(lm1)) # Check normality of the residuals
lm1ShaprioWTest
# Build an analysis results row
c1 <- c("Total Cost, no outlier-influence points")   #  Make the model name
c2 <- lm1ShaprioWTest$p.value # store Shapiro-Wilk test on residuals
c3 <- slm1Summary$r.squared # get model r^2
c4 <- slm1Summary$adj.r.squared # get model r^2
c5 <- slm1$anova$AIC[length(slm1$anova$AIC)]  # get last AIC
c6 <- slm1$anova[length(slm1$anova$AIC),5] # get last residual
c7 <- slm1Summary$fstatistic[1]  # get fstat value
c8 <- slm1Summary$fstatistic[3]  # get fstat df
c9 <- slm1Summary$coefficients[[1,4]] # p-value of Intercept
c10 <- slm1Summary$coefficients[[2,4]] # p-value of TotalCost
rowTemp <-
list(TestName=c1,ShapiroWilkPValue=c2,RSqd=c3,AdjRSqd=c4,LastAIC=c5,LastResidual=c6,Fstat=c7,FstatD
F=c8,InterceptPValue=c9,TotalCostPValue=c10,c10) # store a list with column names
rowTemp <- rowTemp[-11] # Delete error column
rowTemp <- data.frame(rowTemp)
resultsBestFitAnalysis <- rbind(resultsBestFitAnalysis,rowTemp) # Append the analysis results to the model
analysis summary table

# Continue to exclude outlier and influence points for all remaining runs.

# 4d) Do a linear model with degree 1 and only the most significant term
#     Now exclude the outliers and influence points
summary(lm1 <- lm(Indebtedness ~ Type, subset=(isOutlier==0), data = collegeData, direction="forward"))
slm1 <- step(lm1) # summarize the model
slm1Summary <- summary(slm1) # make the summary
summary(slm1) # display the summary
slm1$anova # check the df and residuals
par(mfrow=c(1,2)) # Plot the residuals and normality analysis graphs
plot(slm1, which=1)
plot(slm1, which=2)
par(mfrow=c(1,1))
lm1ShaprioWTest <- shapiro.test(residuals(lm1)) # Check normality of the residuals
```

```
lm1ShaprioWTest
# Build an analysis results row
c1 <- c("Type, no outlier-influence points")   #  Make the model name
c2 <- lm1ShaprioWTest$p.value # store Shapiro-Wilk test on residuals
c3 <- slm1Summary$r.squared # get model r^2
c4 <- slm1Summary$adj.r.squared # get model r^2
c5 <- slm1$anova$AIC[length(slm1$anova$AIC)]  # get last AIC
c6 <- slm1$anova[length(slm1$anova$AIC),5] # get last residual
c7 <- slm1Summary$fstatistic[1]  # get fstat value
c8 <- slm1Summary$fstatistic[3]  # get fstat df
c9 <- slm1Summary$coefficients[[1,4]] # p-value of Intercept
c10 <- slm1Summary$coefficients[[2,4]] # p-value of TotalCost
rowTemp <-
list(TestName=c1,ShapiroWilkPValue=c2,RSqd=c3,AdjRSqd=c4,LastAIC=c5,LastResidual=c6,Fstat=c7,FstatD
F=c8,InterceptPValue=c9,TotalCostPValue=c10,c10) # store a list with column names
rowTemp <- rowTemp[-11] # Delete error column
rowTemp <- data.frame(rowTemp)
resultsBestFitAnalysis <- rbind(resultsBestFitAnalysis,rowTemp) # Append the analysis results to the model
analysis summary table

# 4e) Do a linear model with degree 2 and replace with a different significant term
# Type+TC, no outlier-influence points
summary(lm1 <- lm(Indebtedness ~ Type + TotalCost, subset=(isOutlier==0), data = collegeData,
direction="forward"))
slm1 <- step(lm1) # summarize the model
slm1Summary <- summary(slm1) # make the summary
summary(slm1) # display the summary
slm1$anova # check the df and residuals
par(mfrow=c(1,2)) # Plot the residuals and normality analysis graphs
plot(slm1, which=1)
plot(slm1, which=2)
par(mfrow=c(1,1))
lm1ShaprioWTest <- shapiro.test(residuals(lm1)) # Check normality of the residuals
lm1ShaprioWTest
c1 <- c("Type+TC, no outlier-influence points")   #  Make the model name
c2 <- lm1ShaprioWTest$p.value # store Shapiro-Wilk test on residuals
c3 <- slm1Summary$r.squared # get model r^2
c4 <- slm1Summary$adj.r.squared # get model r^2
c5 <- slm1$anova$AIC[length(slm1$anova$AIC)]  # get last AIC
c6 <- slm1$anova[length(slm1$anova$AIC),5] # get last residual
c7 <- slm1Summary$fstatistic[1]  # get fstat value
c8 <- slm1Summary$fstatistic[3]  # get fstat df
c9 <- slm1Summary$coefficients[[1,4]] # p-value of Intercept
c10 <- slm1Summary$coefficients[[2,4]] # p-value of TotalCost
rowTemp <-
list(TestName=c1,ShapiroWilkPValue=c2,RSqd=c3,AdjRSqd=c4,LastAIC=c5,LastResidual=c6,Fstat=c7,FstatD
F=c8,InterceptPValue=c9,TotalCostPValue=c10,c10) # store a list with column names
rowTemp <- rowTemp[-11] # Delete error column
rowTemp <- data.frame(rowTemp)
resultsBestFitAnalysis <- rbind(resultsBestFitAnalysis,rowTemp) # Append the analysis results to the model
analysis summary table

# 4f) Do a linear model with degree 2 and only the most significant term
# TC+TC^2, no outliers, no influence points
#     Note: This is the selected Best-Fit-Model. Additional plots of this Best-Fit-Model are run and saved.
```

```
summary(lm1 <- lm(Indebtedness ~ TotalCost + I(TotalCost^2), subset=(isOutlier==0), data = collegeData,
direction="forward"))
slm1 <- step(lm1) # summarize the model
slm1Summary <- summary(slm1) # make the summary
summary(slm1) # display the summary
slm1$anova # check the df and residuals

#  Plot and save the four box analysis plot exclusing outlier-influence points for Best-Fit-Model.
par(mfrow=c(2,2))
plot(slm1)
# Usage Note: Manually save Fig08PlotBestFitModelFourBox.jpeg" # R's standard four box analysis plot
without outlier-influence points regressed on model TC+TC^2.
#         Shows best fit model residuals-normality analysis.
# Usage Note: Recommend zooming image and manually saving at a higher resolution.
par(mfrow=c(1,1))

# Make interaction effect plot. See Help(effect) and R-in-Action, p. 187.
library(colorspace)
library(effects)
mod.BFM <- lm1
mod.eff <- allEffects(mod.BFM)
plot(mod.eff, multiline=TRUE)
# Usage Note: Manually save Fig09PlotBestFitModelInteraction.jpeg" # Shows specialty plot of effects. See
Help(effect) and R-in-Action, p. 187.
#         Useage Note: Second plot for TC^2 is mislabeled. Label manually.
# Programming note:  R's built-in interaction-plot command is for two-way anova analysis.
#         The specialty interaction package effects does interaction plots for
#         linear regression models.

# Report normality test
lm1ShaprioWTest <- shapiro.test(residuals(lm1)) # Check normality of the residuals
lm1ShaprioWTest
# Build an analysis results row
c1 <- c("TC+TC^2, no outlier-influence points")   #  Make the model name
c2 <- lm1ShaprioWTest$p.value # store Shapiro-Wilk test on residuals
c3 <- slm1Summary$r.squared # get model r^2
c4 <- slm1Summary$adj.r.squared # get model r^2
c5 <- slm1$anova$AIC[length(slm1$anova$AIC)]  # get last AIC
c6 <- slm1$anova[length(slm1$anova$AIC),5] # get last residual
c7 <- slm1Summary$fstatistic[1]  # get fstat value
c8 <- slm1Summary$fstatistic[3]  # get fstat df
c9 <- slm1Summary$coefficients[[1,4]] # p-value of Intercept
c10 <- slm1Summary$coefficients[[2,4]] # p-value of TotalCost
rowTemp <-
list(TestName=c1,ShapiroWilkPValue=c2,RSqd=c3,AdjRSqd=c4,LastAIC=c5,LastResidual=c6,Fstat=c7,FstatD
F=c8,InterceptPValue=c9,TotalCostPValue=c10,c10) # store a list with column names
rowTemp <- rowTemp[-11] # Delete error column
rowTemp <- data.frame(rowTemp)
resultsBestFitAnalysis <- rbind(resultsBestFitAnalysis,rowTemp) # Append the analysis results to the model
analysis summary table

# 4g) Do a linear model with degree 2 and add the second significant term
# SAT+Type+TC+TC^2, no outlier-influence points
summary(lm1 <- lm(Indebtedness ~ SAT + Type + TotalCost + I(TotalCost^2), subset=(isOutlier==0), data =
collegeData, direction="forward"))
```

```
slm1 <- step(lm1) # summarize the model
slm1Summary <- summary(slm1) # make the summary
summary(slm1) # display the summary
slm1$anova # check the df and residuals
par(mfrow=c(1,2)) # Plot the residuals and normality analysis graphs
plot(slm1, which=1)
plot(slm1, which=2)
par(mfrow=c(1,1))
lm1ShaprioWTest <- shapiro.test(residuals(lm1)) # Check normality of the residuals
lm1ShaprioWTest
c1 <- c("SAT+Type+TC+TC^2, no outlier-influence points")   #  Make the model name
c2 <- lm1ShaprioWTest$p.value # store Shapiro-Wilk test on residuals
c3 <- slm1Summary$r.squared # get model r^2
c4 <- slm1Summary$adj.r.squared # get model r^2
c5 <- slm1$anova$AIC[length(slm1$anova$AIC)]  # get last AIC
c6 <- slm1$anova[length(slm1$anova$AIC),5] # get last residual
c7 <- slm1Summary$fstatistic[1]  # get fstat value
c8 <- slm1Summary$fstatistic[3]  # get fstat df
c9 <- slm1Summary$coefficients[[1,4]] # p-value of Intercept
c10 <- slm1Summary$coefficients[[2,4]] # p-value of TotalCost
rowTemp <-
list(TestName=c1,ShapiroWilkPValue=c2,RSqd=c3,AdjRSqd=c4,LastAIC=c5,LastResidual=c6,Fstat=c7,FstatD
F=c8,InterceptPValue=c9,TotalCostPValue=c10,c10) # store a list with column names
rowTemp <- rowTemp[-11] # Delete error column
rowTemp <- data.frame(rowTemp)
resultsBestFitAnalysis <- rbind(resultsBestFitAnalysis,rowTemp) # Append the analysis results to the model
analysis summary table

# 4h) Do a linear model with degree 2 and replace with a different significant term
# SAT+TC+TC^2, no outlier-influence points
summary(lm1 <- lm(Indebtedness ~ SAT + TotalCost + I(TotalCost^2), subset=(isOutlier==0), data =
collegeData, direction="forward"))
slm1 <- step(lm1) # summarize the model
slm1Summary <- summary(slm1) # make the summary
summary(slm1) # display the summary
slm1$anova # check the df and residuals
par(mfrow=c(1,2)) # Plot the residuals and normality analysis graphs
plot(slm1, which=1)
plot(slm1, which=2)
par(mfrow=c(1,1))
lm1ShaprioWTest <- shapiro.test(residuals(lm1)) # Check normality of the residuals
lm1ShaprioWTest
c1 <- c("SAT+TC+TC^2, no outlier-influence points")   #  Make the model name
c2 <- lm1ShaprioWTest$p.value # store Shapiro-Wilk test on residuals
c3 <- slm1Summary$r.squared # get model r^2
c4 <- slm1Summary$adj.r.squared # get model r^2
c5 <- slm1$anova$AIC[length(slm1$anova$AIC)]  # get last AIC
c6 <- slm1$anova[length(slm1$anova$AIC),5] # get last residual
c7 <- slm1Summary$fstatistic[1]  # get fstat value
c8 <- slm1Summary$fstatistic[3]  # get fstat df
c9 <- slm1Summary$coefficients[[1,4]] # p-value of Intercept
c10 <- slm1Summary$coefficients[[2,4]] # p-value of TotalCost
rowTemp <-
list(TestName=c1,ShapiroWilkPValue=c2,RSqd=c3,AdjRSqd=c4,LastAIC=c5,LastResidual=c6,Fstat=c7,FstatD
F=c8,InterceptPValue=c9,TotalCostPValue=c10,c10) # store a list with column names
```

```
rowTemp <- rowTemp[-11] # Delete error column
rowTemp <- data.frame(rowTemp)
resultsBestFitAnalysis <- rbind(resultsBestFitAnalysis,rowTemp) # Append the analysis results to the model
analysis summary table

# 4i) Do a linear model with degree 2 and replace with a different significant term
# Type+TC+TC^2, no outlier-influence points
summary(lm1 <- lm(Indebtedness ~ Type + TotalCost + I(TotalCost^2), subset=(isOutlier==0), data =
collegeData, direction="forward"))
slm1 <- step(lm1) # summarize the model
slm1Summary <- summary(slm1) # make the summary
summary(slm1) # display the summary
slm1$anova # check the df and residuals
par(mfrow=c(1,2)) # Plot the residuals and normality analysis graphs
plot(slm1, which=1)
plot(slm1, which=2)
par(mfrow=c(1,1))
lm1ShaprioWTest <- shapiro.test(residuals(lm1)) # Check normality of the residuals
lm1ShaprioWTest
c1 <- c("Type+TC+TC^2, no outlier-influence points")   #  Make the model name
c2 <- lm1ShaprioWTest$p.value # store Shapiro-Wilk test on residuals
c3 <- slm1Summary$r.squared # get model r^2
c4 <- slm1Summary$adj.r.squared # get model r^2
c5 <- slm1$anova$AIC[length(slm1$anova$AIC)]  # get last AIC
c6 <- slm1$anova[length(slm1$anova$AIC),5] # get last residual
c7 <- slm1Summary$fstatistic[1]  # get fstat value
c8 <- slm1Summary$fstatistic[3]  # get fstat df
c9 <- slm1Summary$coefficients[[1,4]] # p-value of Intercept
c10 <- slm1Summary$coefficients[[2,4]] # p-value of TotalCost
rowTemp <-
list(TestName=c1,ShapiroWilkPValue=c2,RSqd=c3,AdjRSqd=c4,LastAIC=c5,LastResidual=c6,Fstat=c7,FstatD
F=c8,InterceptPValue=c9,TotalCostPValue=c10,c10) # store a list with column names
rowTemp <- rowTemp[-11] # Delete error column
rowTemp <- data.frame(rowTemp)
resultsBestFitAnalysis <- rbind(resultsBestFitAnalysis,rowTemp) # Append the analysis results to the model
analysis summary table

# 4f) Do a linear model with degree 3 for Roomboard and only the most significant term
# RB+RB^2+RB^3+TC+TC^2, no outliers, no influence points
summary(lm1 <- lm(Indebtedness ~ RoomBoard + I(RoomBoard^2) + I(RoomBoard^3) + TotalCost +
I(TotalCost^2), subset=(isOutlier==0), data = collegeData, direction="forward"))
slm1 <- step(lm1) # summarize the model
slm1Summary <- summary(slm1) # make the summary
summary(slm1) # display the summary
slm1$anova # check the df and residuals

#  Plot and save the four box analysis plot exclusing outlier-influence points for Best-Fit-Model.
par(mfrow=c(2,2))
plot(slm1)
# Usage Note: Manually save Fig08PlotBestFitModelFourBox.jpeg" # R's standard four box analysis plot
without outlier-influence points regressed on model TC+TC^2.
#        Shows best fit model residuals-normality analysis.
# Usage Note: Recommend zooming image and manually saving at a higher resolution.
par(mfrow=c(1,1))
```

```
# Make interaction effect plot. See Help(effect) and R-in-Action, p. 187.
library(colorspace)
library(effects)
mod.BFM <- lm1
mod.eff <- allEffects(mod.BFM)
plot(mod.eff, multiline=TRUE)
# Usage Note: Manually save Fig09PlotBestFitModelInteraction.jpeg" # Shows specialty plot of effects. See Help(effect) and R-in-Action, p. 187.
#	Useage Note: Second plot for TC^2 is mislabeled. Label manually.
# Programming note:  R's built-in interaction-plot command is for two-way anova analysis.
#	The specialty interaction package effects does interaction plots for
#	linear regression models.

# Report normality test
lm1ShaprioWTest <- shapiro.test(residuals(lm1)) # Check normality of the residuals
lm1ShaprioWTest
# Build an analysis results row
c1 <- c("RB+RB^2+RB^3+TC+TC^2, no outlier-influence points")   #  Make the model name
c2 <- lm1ShaprioWTest$p.value # store Shapiro-Wilk test on residuals
c3 <- slm1Summary$r.squared # get model r^2
c4 <- slm1Summary$adj.r.squared # get model r^2
c5 <- slm1$anova$AIC[length(slm1$anova$AIC)]  # get last AIC
c6 <- slm1$anova[length(slm1$anova$AIC),5] # get last residual
c7 <- slm1Summary$fstatistic[1]  # get fstat value
c8 <- slm1Summary$fstatistic[3]  # get fstat df
c9 <- slm1Summary$coefficients[[1,4]] # p-value of Intercept
c10 <- slm1Summary$coefficients[[2,4]] # p-value of TotalCost
rowTemp <- list(TestName=c1,ShapiroWilkPValue=c2,RSqd=c3,AdjRSqd=c4,LastAIC=c5,LastResidual=c6,Fstat=c7,FstatDF=c8,InterceptPValue=c9,TotalCostPValue=c10,c10) # store a list with column names
rowTemp <- rowTemp[-11] # Delete error column
rowTemp <- data.frame(rowTemp)
resultsBestFitAnalysis <- rbind(resultsBestFitAnalysis,rowTemp) # Append the analysis results to the model analysis summary table


# 5) Display and export the best fit analysis table
write.csv(resultsBestFitAnalysis, file="Listing7BestFitAnalysis.csv")
resultsBestFitAnalysis

# The results are counter-intuitive because, as shown above
#   the means and sd's of Indebtedness overlap for private and public colleges,
#   but the means and sd's of Total Cost does not.

# Conclusion: The Best-Fit-Model is TC+TC^2

# 6) Replot the the best-fit model with prediction and confidence lines.
#   4f) Do a linear model with degree 2 and only the most significant term
# TC+TC^2, no outliers, no influence points
summary(lm1 <- lm(Indebtedness ~ TotalCost + I(TotalCost^2), subset=(isOutlier==0), data = collegeData, direction="forward"))
slm1 <- step(lm1) # summarize the model
slm1Summary <- summary(slm1) # make the summary
summary(slm1) # display the summary
slm1$anova # check the df and residuals
```

```
# plot 95% confidence interval
# plotting prediction and confidence bands #
fit = slm1
x.val=seq(min(collegeData$TotalCost),max(collegeData$TotalCost), length=50)  # get 50 points
pb=predict(fit,data.frame(TotalCost=x.val),int="p")
cb=predict(fit,data.frame(TotalCost=x.val),int="c")

#  Plot and save Best-Fit-Model with confidence and prediction intervals
plot(collegeData$TotalCost,collegeData$Indebtedness,xlab=c("Total Cost"), ylab=c("Indebtedness"))
# abline(lm1) # regression line (y~x)
lines(predict(fit) ~ x.val)
matlines(x.val,cb,lty=c(1,3,3),col="black")
matlines(x.val,pb,lty=c(1,5,5),col="black")
# Manually save Fig10BestFitModelPredConfInt.jpeg - Plot of Best-Fit-Model without x transforms with
confidence and prediction lines.

# Export the Best-Fit-Model-Description
zz <- file("Listing8BestFitModelDescrip.txt", open="wt")
sink(zz)
slm1Summary
sink()
unlink(zz)

# Next Step in Next Application: Transform quadratic x's on Best-Fit-Model.

detach("package:car", unload=TRUE)
detach("package:effects", unload=TRUE)
detach("package:nnet", unload=TRUE)
detach("package:colorspace", unload=TRUE)
detach("package:lattice", unload=TRUE)
# Usage Note:  Remember, after running this task, check that packages
# car, colorspace, effects, and nnet
# are disabled.

# 7)  Check the best-fit-model y returns

funcBFM <- function(xhat)
{
  yhat <- -12.19595 + (2.32769 * xhat) + (-0.04998 * xhat^2)
  return(yhat)
}

xhat <- seq(10,30,1)

funcBFM(xhat)
```

Listing C 5. R Code to compute standardized linear regression model coefficients.

```
# Stat 3080, R Group Chun Project 4
# Kurt Fisher and Elijah Neilson
# Task 4 - Prepare standardized polynomial Best-Fit-Model
# Version Date: 3-31-2013
# Version: 1.0
# File: 20130331RProT4TransformsPoly1_0.R

# Usage note: This task requires specialty packages car, corrplot, and nnet. However,
#    incompatiablities in those packages case runtime errors in the boxplot package.
#    After running this task, remember to check that packages car, corrplot, and nnet
#    are disabled. Disabling of those packages should occur automatically.

# Uses: schooldatamodASCII.txt, schoolacronyms.txt

# Makes the following figures and tables that will be used in writing the report.
# Make automatically:
#     Listing9BestFitModelStandardized.csv - Best Fit Analysis of several models
# Make manually:
#     Fig11BFMStandardizedPredConfInt.jpeg

# Shows - Prepares standarized plot of previsously determined Best-Fit-Model TC+TC^2.
#         Computes variables for writing the unstandardized version of the BFM linear
#         regression equation.

# Set directory
#   Usage Note:
#   Change this path to your personal working directory.
#   Put data files schooldatamodASCII.txt and schoolacronyms.txt into that directory for reading.
setwd("C:\\Users\\fisherka\\Documents\\00000000UofUSpr2013\\3080Stats\\ProjectSec")
#
source("C:\\Users\\fisherka\\Documents\\00000000UofUSpr2013\\3080Stats\\ProjectSec\\20130331RProT4Tra
nsformsPoly1_0.R", echo=TRUE) #    Note: source is with echo ON option.
#
source("C:\\Users\\fisherka\\Documents\\00000000UofUSpr2013\\3080Stats\\ProjectSec\\20130331RProT4Tra
nsformsPoly1_0.R", echo=FALSE) #    Note: source is with echo OFF option.

# 1) Get data
collegeData = read.table("schooldatamodASCII.txt", header=TRUE)
collegeData <- data.frame(collegeData)
# Coerce the data into a numeric type in order to allow use with later histogram plotting.
#    Coercion is not needed for visual data analysis of correlation packages.
collegeData$Term <- as.numeric(collegeData$Term)
collegeData$Type <- as.numeric(collegeData$Type)
collegeData$SAT <- as.numeric(collegeData$SAT)
collegeData$TOEFL <- as.numeric(collegeData$TOEFL)
collegeData$RoomBoard <- as.numeric(collegeData$RoomBoard)
collegeData$TotalCost <- as.numeric(collegeData$TotalCost)
collegeData$Indebtedness <- as.numeric(collegeData$Indebtedness)
# append table of acronyms for scatterplotting outliers
collegeAcronyms <- read.table("schoolacronyms.txt", header=TRUE) # read the acronym list
collegeAcronyms <- collegeAcronyms[3] # get the acronym column
collegeData[9] <- collegeAcronyms # append the acronym column to the data table
col10 <- rep(0,80) # make a variable to exclude outliers 0=False 1 = True
col10 <- list(isOutlier=col10) # initialize to all False
collegeData <- cbind(collegeData,col10) # append the column to the data table

# 2) Mark known outliers previously identified
collegeData[39,10]= 1 # Set Case Western Reserve as an indebtedness outlier.
collegeData[58,10]= 1 # Set Northeastern Univ. as an indebtedness outlier.
#       Note there is one low Indebtedness value that is not an outlier - Rice Univ.
# collegeData[order(collegeData$Indebtedness, decreasing=FALSE),][1,] # Disabled, use in prior tasks

# 3) Remove previously identified influence points
#   Mark points for exclusion from best model fit.
collegeData[6,10]= 2 # Mark influence point - Florida Int'l Univ.
collegeData[7,10]= 2 # Mark influence point - Florida State Univ.
collegeData[61,10]= 2 # Mark influence point - Rice Univ.

collegeData # Display the data table with revisions

# 4) Recreate an instance of the final selected Best-Fit-Model.
#   4f) Do a linear model with degree 2 and only the most significant term
# TC+TC^2, no outliers, no influence points
summary(lm1 <- lm(Indebtedness ~ TotalCost + I(TotalCost^2), subset=(isOutlier==0), data = collegeData,
direction="forward"))
slm1 <- step(lm1) # summarize the model
slm1Summary <- summary(slm1) # make the summary
summary(slm1) # display the summary
slm1$anova # check the df and residuals

# 5) Build a standardized Best-Fit-Model

# abstract the x-y pair from the original model
xymodel <- subset(collegeData, Type==0, select=c(TotalCost,Indebtedness))
colnames(xymodel)[1] <- "x"
colnames(xymodel)[2] <- "y"

# Find Sxx of undstandardized model
n <- length(xymodel$x)
xymodelSxx <- sd(xymodel$x)

# standardize the x values
col1xyStd <- (xymodel$x-mean(xymodel$x))/xymodelSxx
col2xystd <- xymodel$y
xymodelStd <- cbind(x=col1xyStd,y=col2xystd) # new x,y's with x's standardized.
xymodelStd <- data.frame(xymodelStd)
xymodelStd <- xymodelStd[order(xymodelStd$x),]  # Reorder the table by ascending x's.
# Programming note: failure to sort the x's will cause the regression line to print in the
# wrong order).
print(xymodelStd)

# make polynomial model
#  Note key use of I qualifier in lm.
xymodellm <- lm( xymodelStd$y~xymodelStd$x+I(xymodelStd$x^2)) # make polynomial model
fitlm <- xymodellm # port model to generic code fragment.
```

```
fitlmSummary <- summary(fitlm)
print(fitlmSummary) # Print the standardized model summary.

# Extract the coefficient of determination
fitlmSummary$r.squared

# 6) Display the standardized Best-Fit-Model

# Plot the model with 95% confidence and prediction confidence bands
#  Compute the x values
fit = fitlm
x.val=seq(min(xymodelStd$x),max(xymodelStd$x), length=30)  # get 15 points
# Programming note: The number of predicted points must be less than the
#     number of x points in the data.
pb=predict(fit,data.frame(x=x.val),int="p")
cb=predict(fit,data.frame(x=x.val),int="c")

plot(xymodelStd$x,xymodelStd$y, xlab="Total Cost (Standardized USD 000s)", ylab="Indebtedness (USD
000s)")
# lines(xymodelStd$x,fitted(fitlm)) # regression line (y~x)
lines(predict(fit) ~ x.val)
matlines(x.val,cb,lty=c(1,3,3),col="black")
matlines(x.val,pb,lty=c(1,5,5),col="black")

# Manually save Fig11BFMStandardizedPredConfInt.jpeg - Plot of Best-Fit-Model, standardized with x
transforms and with confidence and prediction bands.

# 7) Compute other values to write standardized regression equation
#    using unstandardized x's.  See. Devore, Ex. 13-4, and Math 3080, RLab No. 11, Part 2.

# Note the values need to write a standardized equation form are:
n  # for the t-table lookup
xymodelSxx   # standard deviation of the original model's x's.
meanX <- mean(xymodel$x)  # mean of the x's in the original model
meanX
```

Listing C 6. R Code to run PCA and factor analysis tests in order to determine the likely number of factors.

```
# Stat 3080, R Group Chun Project 4
# Kurt Fisher and Elijah Neilson
# Task 5 - Prepare standardized polynomial Best-Fit-Model
# Version Date: 4-02-2013
# Version: 1.0
# File: 20130402RProT5SuppEFAPCA1_0.R

# Uses: schooldatamodASCII.txt, schoolacronyms.txt

# Shows – Estimates the number of factors using PCA and factor analysis after Teetor pp. 349-350.

# Set directory
#   Usage Note:
#   Change this path to your personal working directory.
#   Put data files schooldatamodASCII.txt and schoolacronyms.txt into that directory for reading.
setwd("C:\\Users\\fisherka\\Documents\\00000000UofUSpr2013\\3080Stats\\ProjectSec")
#
source("C:\\Users\\fisherka\\Documents\\00000000UofUSpr2013\\3080Stats\\ProjectSec\\20130402RProT5Sup
pEFAPCA1_0.R", echo=TRUE) #    Note: source is with echo ON option.
#
source("C:\\Users\\fisherka\\Documents\\00000000UofUSpr2013\\3080Stats\\ProjectSec\\20130402RProT5Sup
pEFAPCA1_0.R", echo=FALSE) #    Note: source is with echo OFF option.

# 1) Get data
collegeData = read.table("schooldatamodASCII.txt", header=TRUE)
collegeData <- data.frame(collegeData)
# Coerce the data into a numeric type in order to allow use with later histogram plotting.
#    Coercion is not needed for visual data analysis of correlation packages.
collegeData$Term <- as.numeric(collegeData$Term)
collegeData$Type <- as.numeric(collegeData$Type)
collegeData$SAT <- as.numeric(collegeData$SAT)
collegeData$TOEFL <- as.numeric(collegeData$TOEFL)
collegeData$RoomBoard <- as.numeric(collegeData$RoomBoard)
collegeData$TotalCost <- as.numeric(collegeData$TotalCost)
collegeData$Indebtedness <- as.numeric(collegeData$Indebtedness)
# append table of acronyms for scatterplotting outliers
collegeAcronyms <- read.table("schoolacronyms.txt", header=TRUE) # read the acronym list
collegeAcronyms <- collegeAcronyms[3] # get the acronym column
collegeData[9] <- collegeAcronyms # append the acronym column to the data table
col10 <- rep(0,80) # make a variable to exclude outliers 0=False 1 = True
col10 <- list(isOutlier=col10) # initialize to all False
collegeData <- cbind(collegeData,col10) # append the column to the data table

# 2) Mark known outliers previously identified
collegeData[39,10]= 1 # Set Case Western Reserve as an indebtedness outlier.
collegeData[58,10]= 1 # Set Northeastern Univ. as an indebtedness outlier.
#       Note there is one low Indebtedness value that is not an outlier - Rice Univ.
# collegeData[order(collegeData$Indebtedness, decreasing=FALSE),][1,] # Disabled, use in prior tasks
# 3) Remove previously identified influence points
#   Mark points for exclusion from best model fit.
collegeData[6,10]= 2 # Mark influence point - Florida Int'l Univ.
collegeData[7,10]= 2 # Mark influence point - Florida State Univ.
collegeData[61,10]= 2 # Mark influence point - Rice Univ.

# 4) Subset the data for FA and PCA.
collegeDataSubset <- subset(collegeData,isOutlier==0, select=c(Term, Type, SAT, TOEFL, RoomBoard,
TotalCost))

collegeDataSubset # Display the data table with revisions

#  Display the five number summary of Total Cost and Indebtedness without outliers and influence points.
summary(collegeDataSubset$TotalCost)
summary(collegeDataSubset$Indebtedness)

# 5) Do PCA analysis
prcomp(collegeDataSubset)
princomp(collegeDataSubset)
plot(prcomp(collegeDataSubset))

# 6) Do factor analysis to estimate the number of factors

# 6a) n=1 factors
factanal(collegeDataSubset, factors=1)

# 6b) n=2 factors
factanal(collegeDataSubset, factors=2)

# 6c) n=3 factors
factanal(collegeDataSubset, factors=3)
```

[End]