R: A Primer, Part I

David R. Cross, Ph.D.[*]
Texas Christian University

September 23, 2008

I have three objectives for this Primer:

1. The first is to provide a brief introduction to R—*the open-source programming language and software system for statistics.*
2. The second is to introduce some basic R commands, using an extensive example based on the recent *U. S. News & World Report* rankings of "America's Best Colleges."
3. The third is to suggest some advanced capabilities that should make R attractive to statisticians and research scientists.

A Brief Introduction

R is to statistics what linux is to operating systems: R is cutting edge, open source, user built and maintained, and dynamically evolving. R is an attractive alternative to commercial statistical packages such as SPSS, SYSTAT and SAS, in the same way that linux is an attractive alternative to Windows XP and Vista. R has many of the same virtues (and vices) as does linux, and in the eyes of this beholder, R is beautiful. Here are some good R resources:

- *The R Project for Statistical Computing* (`http://www.r-project.org/`).
- *The Comprehensive R Archive Network* (`http://cran.r-project.org/`).
- *Wikipedia: R* (`http://en.wikipedia.org/wiki/R_programming_language`).
- *Quick-R* (`http://www.statmethods.net/index.html`).
- *Introduction to R* (`http://www.biostat.wisc.edu/~kbroman/Rintro/`).
- W. Revelle, *Using R for psychological research: A simple guide to an elegant package* (`http://personality-project.org/r/`).

[*] www.davidcross.us

- W. N. Venables, D. M. Smith, and the R Development Core Team, *An Introduction to R* (Revised and updated), Network Theory Ltd, 2005. (Also available as a pdf file: `cran.r-project.org/doc/manuals/R-intro.pdf`.)

You should also know that R is "Gnu S," which means that R is the open source version of the S-Plus data analysis system. Resources for S-Plus include the following:

- *Wikipedia: S-Plus* (`http://en.wikipedia.org/wiki/S-PLUS`).
- *StatLib—Software and extensions for the S (Splus) language* (website) (`http://lib.stat.cmu.edu/S/`).
- A. Krause & M. Olson, *The Basics of S-Plus* (4th ed.), Springer Statistics, 2005.

R is free software (open source), which means that it falls under the purview of the Free Software Foundation (`http://www.fsf.org/`) and the GNU General Public License (`http://www.gnu.org/copyleft/gpl.html`).

An Extensive Example

I illustrate some basic R capabilities using an example taken from a recent issue of *U. S. News & World Report*, namely, the rankings of "America's Best Colleges." Relevant sources include the following:

- *U. S. News & World Report*, September 1–8, 2008.[1]
- *America's Best Colleges* (2009 Edition), published by the editors and staff of *U. S. News & World Report*, 2008.
- *Best Colleges 2009* (website) (`http://colleges.usnews.rankingsandreviews.com/college`)
- *Wikipedia: College and university rankings* (`http://en.wikipedia.org/wiki/College_rankings`)
- E. F. Farrell and M. Van Der Werf, Playing the rankings game, *Chronicle of Higher Education* (Special Report), May 27, 2007. (see `http://chronicle.com/free/v53/i38/38a01101.htm`)
- R. Grewal, J. A. Dearden & G. L. Lilien, The university rankings game: Modeling the competition among universities for ranking, *The American Statistician*, 2008, Vol. 62, No. 3, pp. 232–237.

[1] I extracted the data set used in the example from this publication (see **Appendix**).

My analysis of the university rankings is designed not only to introduce R, but also to shed some light on the rankings themselves. The analysis will include data input, descriptive statistics, graphical displays, and multiple regression analyses. All of the statistical procedures illustrated here can be found in the following textbook, which is highly recommended:

- P. Dalgaard, *Introductory Statistics with R*, Springer Statistics and Computing, 2002.

This analysis does not exhaust all possibilities with these data, but is a start. More can be done later.

Reading the Data

As an R novice, I keep it simple when entering data. For this example, I created a plain text file (see **Appendix**) and read it into R using the following commands:[2]

```
> ranks <- read.table("/Users/davidcross/Documents/TCU/USNews/USNews.txt",
           header=T)
> attach(ranks)
```

The `read.table` command will read data from the file listed in parentheses, and assign the data to the "object" on the LHS of the assignment symbol (`<-`). In the first of the above commands, the data in `USNews.txt` are assigned to `ranks`. The `attach` command makes all eleven variables in `ranks` individually available for future commands.

R possesses an impressive array of data reading and writing commands, as is indicated in the following post to EDSTAT-L:[3]

```
From: Brett Magill <magillb@SBCGLOBAL.NET>
Subject: Re: R Data Import/Export
To: EDSTAT-L@LISTS.PSU.EDU
Date: Tue, 26 Aug 2008 14:05:03 -0700
Reply-To: Teaching and Learning Statistics <EDSTAT-L@lists.psu.edu>

Yes, R can read pretty much anything.  Basic text (CSV
or other Delimited Files) are the easiest, using a
command like

read.csv("A:/Path/to/file.csv")

With binary files from other packages you might need
to use the "foreign" library by
```

[2] I will denote commands and output in typewriter font, e.g., `attach(ranks)`.

[3] See also Robert Kabacoff's *Quick-R* website—`http://www.statmethods.net/index.html`—which is impressive.

```
1. Invoking it...

library(foreign)

2. Executing one of its data handling functions.

read.spss("A:/Path/to/file.sav")
read.xprt(...) SAS transport
read.ssd(...) SAS data
read.systat(...) systat
read.mtp(...) minitab
read.dbf(...) dbf database file
...

Of course there is also a write.XXX() method for each
of these too.
```

Thank you, Brett!

Univariate Statistics and Boxplots

Univariate statistics are available using the `summary` command, which is applied here to variables 2–11, omitting the first (School):[4]

```
> summary(ranks[2:11])

     Score            Peer           Fresh          Grad            U20
Min.   : 37.00   Min.   :2.50   Min.   :79.0   Min.   :56.00   Min.   :22.00
1st Qu.: 44.00   1st Qu.:3.10   1st Qu.:88.0   1st Qu.:73.00   1st Qu.:40.50
Median : 54.00   Median :3.50   Median :92.0   Median :80.00   Median :49.00
Mean   : 58.88   Mean   :3.57   Mean   :91.4   Mean   :79.75   Mean   :51.14
3rd Qu.: 71.50   3rd Qu.:4.00   3rd Qu.:96.0   3rd Qu.:88.50   3rd Qu.:62.50
Max.   :100.00   Max.   :4.90   Max.   :99.0   Max.   :97.00   Max.   :76.00

     X50M            FTF              Top10           Accept          Alumni
Min.   : 0.40   Min.   : 69.00   Min.   :22.00   Min.   : 9.00   Min.   : 7.00
1st Qu.: 7.50   1st Qu.: 85.50   1st Qu.:40.00   1st Qu.:32.50   1st Qu.:14.00
Median :10.00   Median : 91.00   Median :60.00   Median :50.00   Median :19.00
Mean   :11.72   Mean   : 89.84   Mean   :61.44   Mean   :47.79   Mean   :21.94
3rd Qu.:16.00   3rd Qu.: 95.00   3rd Qu.:86.00   3rd Qu.:60.50   3rd Qu.:28.00
Max.   :30.00   Max.   :100.00   Max.   :99.00   Max.   :89.00   Max.   :60.00
```

This makes a servicable readout of the data, but box-and-whisker plots reveal more. In what follows I present these ten variables in Figures 1 and 2. The

[4] The output has been reformatted to better fit the page.

commands for the first pair of boxplots are given just below, beginning with the `par` command, which tells R to output the following graphics commands in a single row and two columns, using `(mfcol=c(1,2)`. The boxplots are shown in Figure 1.

```
> par(mfcol=c(1,2))
> boxplot(Score,xlab="Overall Score")
> boxplot(Peer,xlab="Peer Assessment")
```

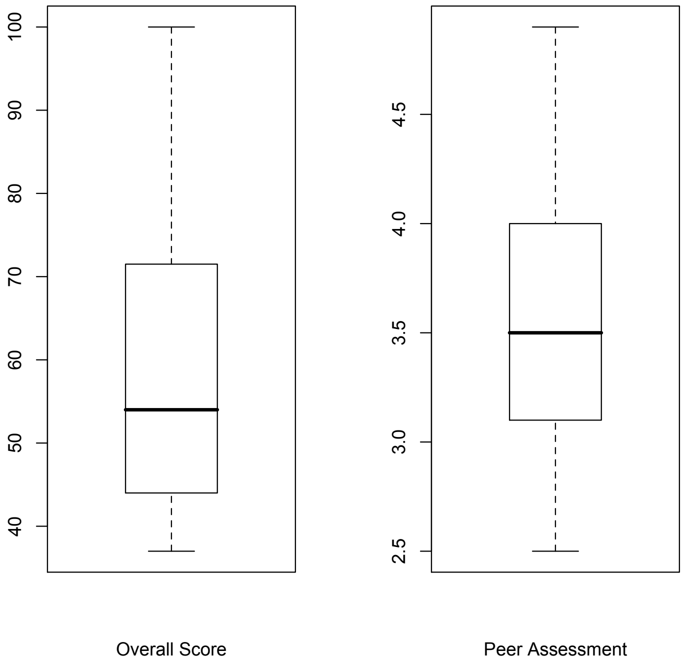

Figure 1: Box-and-whisker plots for Overall Score and Peer Assessment.

The boxplots for `Score` and `Peer` reveal that the distribution is positively skewed for "Overall Score" but fairly symmetric for "Peer Assessment." In terms of the overall score, it appears that there are relatively few universities with high scores.[5] This clumping towards the bottom of the distribution can be seen in the rankings (not included in my data set), where there are more ties towards the bottom than towards the top. One implication of this phenomenon

[5] These are scaled scores that can range from 0 to 100; the peer assessment ratings can range from 1 to 5.

is that the the rankings will be *less stable* towards the bottom of the distribution of ranked universities.

The remaining boxplots were constructed in a similar fashion, while making the appropriate substitutions for variable names. The remaining eight boxplots, shown in Figure 2, were generated using the following commands:

```
> par(mfcol=c(2,4))
> boxplot(Grad, xlab="Graduation Rate")
> boxplot(Fresh, xlab="Freshman Retention")
> boxplot(U20, xlab="Classes Under 20")
> boxplot(X50M, xlab="Classes Over 50")
> boxplot(FTF, xlab="Full Time Faculty")
> boxplot(Alumni, xlab="Alumni Giving")
> boxplot(Top10, xlab="Top 10% HS")
> boxplot(Accept, xlab="Acceptance Rate")
```

Note that the boxplots show the median, which the center line, the 1st and 3rd quartiles, which are the lower and upper "hinges" of the boxes, and the minimum and maximum values, which are the endpoints of the "whiskers." From the box-and-whisker plots shown in Figure 2 you can see the following:

- The median freshman retention rate is about 92% for the top 113 national universities.
- The median graduation rate is about 80%.
- The median percentage of classes under 20 is about 50%.
- The median percentage of classes more than 50 is about 10%; this distribution is positively skewed, with an outlier at 30%.
- The median percentage of faculty that are full time is about 90%, with an outlier down around 70%.
- The median rate of alumni giving is about 19 or 20%, with three outliers up in the 50–60% range.
- The median percentage of incoming students who come from the top 10% of their high school classes is about 60%; this percentage ranges from a low of 20% to a high of almost 100%.
- The median acceptance rate for the top 113 national universities is about 50%, and ranges from about 10% to almost 90%.

Bivariate Scatterplots

In the next stage of the analysis I construct two series of bivariate scatterplots. The first series plots the overall score versus each of the other variables, the second series plots the peer rankings versus each of the remaining variables. For reasons that will become clear below, the focal association for this problem is the association between the overall score and peer assessment. A scatterplot for these two variables can be created in R using the following command:

```
> plot(Peer,Score, xlab="Peer Assessment", ylab="Overall Score")
```

As can be seen in Figure 3, the association between peer ratings and the overall scores is strong and positive: As peer ratings increase, so do the overall scores, and hence also the rankings.[6] An important feature of these data is that the data set represents only the top 113 universities in the "National Universities" category. The remaining 149 (I think) are not included, and hence are not shown in this scatterplot. The lowest overall score included among the top 113 national universities is 37, which effectively truncates the distribution at that value.

For the multiple regressions presented in the following subsection, the peer ratings are by far the strongest predictor of the overall scores. However, four other variables also make significant contributions to this regression, and these are shown in the following set of bivariate scatterplots (see Figure 4). The commands for this set are:

```
> par(rmcol(2,2))
> plot(Grad, Score, xlab="Graduation Rate", ylab="Overall Score")
> plot(U20, Score, xlab="Classes Under 20 (%)", ylab="Overall Score")
> plot(Top10, Score, xlab="Freshmen Top 10%", ylab="Overall Score")
> plot(Alumni, Score, xlab="Alumni Giving (%)", ylab="Overall Score")
```

The predictors plotted in Figure 4 are graduation rate, entering freshmen coming from the top ten percent of their high school classes, percentage of classes with class size under 20, and the rate of alumni giving. All four predictors have a positive association with the overall scores, although there is some nonlinearity (graduation rate and alumni giving) and heteroscedacity (freshmen in the top ten percent and classes under 20). A similar set of four `plot` commands can be used to generate the remaining four scatterplots in this series (see Figure 5):

```
> plot(Fresh, Score, xlab="Freshman Retention (%)", ylab="Overall Score")
> plot(X50M, Score, xlab="Classes Over 50 (%)", ylab="Overall Score")
> plot(FTF, Score, xlab="Full-Time Faculty (%)", ylab="Overall Score")
> plot(Accept, Score, xlab="Acceptance Rate", ylab="Overall Score")
```

The predictors plotted in Figure 5 are freshmen retention rate, percentage of full-time facult, percentage of classes that have 50 or more students, and the acceptance rate for new students. Three of the four predictors have positive associations with the overall score, whereas the fourth (acceptance rate) has a negative association. As with the previous set in this series (Figure 4), there are instances of nonlinearity (freshman retention rate) and instances of

[6] The Pearson correlation for the overall score and peer ratings is .91; all of the bivariate correlations are presented below.

heteroscedasticity (full-time faculty). Somewhat surprising is the scatterplot between overall score and percentage of classes over 50, which indicates a complex association between these two variables.

The second series of scatterplots is shown in Figures 6 and 7. Associations between the peer ratings and those predictors that are signficant in the multiple regressions of the next section are plotted in Figure 6. Associations between the peer ratings and those predictors that are *not* signficant in the multiple regressions of the next section, are plotted in Figure 7. The R commands used to generate the plots shown in Figure 6 are as follows:

```
> plot(Grad, Peer, xlab="Graduation Rate", ylab="Peer Assessment")
> plot(FTF, Peer, xlab="Full-Time Faculty (%)", ylab="Peer Assessment")
> plot(X50M, Peer, xlab="Classes 50 or More (%)", ylab="Peer Assessment")
> plot(Accept, Peer, xlab="Acceptance Rate", ylab="Peer Assessment")
```

As can be seen in Figure 6, three of these predictors are positively associated with peer ratings, and one (acceptance rate) is negatively associated with peer ratings. Further, there is some evidence for heteroscedasticity, especially in the bottom row of the figure. The R commands used to generate the plots shown in Figure 7 are as follows:

```
> plot(Fresh, Peer, xlab="Freshman Retention (%)", ylab="Peer Assessment")
> plot(Top10, Peer, xlab="Freshman Top 10%", ylab="Peer Assessment")
> plot(U20, Peer, xlab="Classes Under 20 (%)", ylab="Peer Assessment")
> plot(Alumni, Peer, xlab="Alumni Giving (%)", ylab="Peer Assessment")
```

As can be seen in Figure 7, all four predictors are positively associated with Peer Ratings, although the strength of the association is greater for those in the top row (freshman retention and freshman in the top 10%). There is also some evidence of heterscedasticity (classes under 20) and nonlinearity (alumni giving).

In a more thorough analysis of these data, the researcher may want to investigate the impact, if any, of the heteroscedasticity and/or nonlinearity in some of the scatterplots shown in Figures 4–7. I chose not to dig that deep, at least for now, and instead turned to the regression analyses that inspired the grouping of scatterplots.

Multiple Regression

As a prelude to the multiple regression analyses that follow, I present here correlations among the ten variables in the dataset:[7]

```
> cor(ranks[2:11])
            Score       Peer       Fresh       Grad        U20
Score   1.0000000  0.9109817  0.87500684  0.8898054  0.6999411
Peer    0.9109817  1.0000000  0.78433615  0.7361488  0.4640000
Fresh   0.8750068  0.7843362  1.00000000  0.9141843  0.5237235
Grad    0.8898054  0.7361488  0.91418432  1.0000000  0.6251940
U20     0.6999411  0.4640000  0.52372349  0.6251940  1.0000000
X50M   -0.1486512  0.1467712 -0.01198980 -0.2073705 -0.5879330
FTF     0.2857388  0.4302485  0.20157556  0.1367771 -0.0739485
Top10   0.8432630  0.7575381  0.83216386  0.7991461  0.5204247
Accept -0.8328961 -0.6710517 -0.78254498 -0.7921744 -0.6739636
Alumni  0.7422412  0.5557688  0.60805924  0.6522845  0.5688418

              X50M          FTF       Top10       Accept     Alumni
Score  -0.14865117  0.285738846  0.84326304 -0.832896144  0.7422412
Peer    0.14677120  0.430248545  0.75753814 -0.671051660  0.5557688
Fresh  -0.01198980  0.201575559  0.83216386 -0.782544980  0.6080592
Grad   -0.20737045  0.136777083  0.79914610 -0.792174408  0.6522845
U20    -0.58793296 -0.073948496  0.52042469 -0.673963576  0.5688418
X50M    1.00000000  0.401561690  0.02775489  0.279859889 -0.3166975
FTF     0.40156169  1.000000000  0.19672745  0.005055447  0.2328474
Top10   0.02775489  0.196727445  1.00000000 -0.741397584  0.5325166
Accept  0.27985989  0.005055447 -0.74139758  1.000000000 -0.6719766
Alumni -0.31669747  0.232847381  0.53251664 -0.671976608  1.0000000
```

This is fairly typical output for modern statistical packages, which displays far too many "significant" digits for each correlation. In Table 1 the correlations are displayed in a more suitable fashion.[8] The cor command used to obtain the correlations can be used to obtain various kinds of correlations (Pearson, Kendall, Spearman), and its cousins var and cov can be used to obtain variances and covariances, respectively. The default for cor is Pearson, which is what we have here. cor can also be used to obtain blocks of correlations, between two different sets of variables. This is a good time to point out that R has a nice help facility, so that all you have to do is type, for example, help(cor), which opens a new window with the documentation.

Table 1 shows that the first eight variables tend to be strongly correlated, whereas the last two (Full Time Faculty and Classes 50 or More) tend not to correlate with the first eight. The following multiple regressions elaborate on this general impression.

[7] The output has been reformatted to better fit the page.

[8] See, for example, G. A. F. Seber, *Multivariate Observations* (Ch. 4), Wiley, 2004.

	OS	PR	GR	FR	T10	AR	AG	U20	FTF
PR	.9								
GR	.9	.7							
FR	.9	.8	.9						
T10	.8	.8	.8	.8					
AR	-.8	-.7	-.8	-.8	-.7				
AG	.7	.6	.7	.6	.5	-.7			
U20	.7	.5	.6	.5	.5	-.7	.6		
FTF	.3	.4	.1	.2	.2	.0	.2	-.1	
50M	-.1	.1	-.2	.0	.0	.3	-.3	-.6	.4

Table 1: Pearson correlations among the ten variables, displayed using a single digit, and with rows and columns ordered according to the correlation with overall score (OS-Overall Score; PR-Peer Ratings; GR-Graduation Rate; FR-Freshman Retention; T10-Incoming Class in Top 10%; AR-Acceptance Rate; AG-Alumni Giving; U20-Classes Under 20; FTF-Full Time Faculty; 50M-Classes 50 or More).

Predicting the Overall Score

The goal of this first analysis is to determine which variables predict the Overall Score, which in turn determines the rankings. At first glance the answer to this question might seem preordained, since the Overall Score is a weighted linear combination of the other variables. However, a simple weighting does not take into account the associations among the predictors themselves, and because of this it could turn out that the weights in a multiple regression analysis would be different than the weighting scheme used by *U. S. News and World Report*. In any case, it makes an interesting example for getting our feet wet with R.[9]

The regression analyses are carried out using the `lm` (linear model) command, which takes a formula as its primary input. The general format is something like `lm(Y~X1+X2+X3+X4)`, where there is one criterion variable and four predictor variables. Typical usage requires that the result of `lm` be stored in an object. For example, `lm.out <- lm(Y~X1+X2+X3+X4)` would store the result in the object `lm.out`. Various things can be done with the output object, including `summary`, which produces a nice summary of the regression analysis. These R commands are used below in a sequence of multiple regression analyses, which as a whole constitute a manual backwards elimination algorithm. I start with `Score` as the criterion variable, and the remaining nine variables as predictors. At each step of the algorithm, the predictor with the largest p-value is eliminated, until only predictors that are statistically significant remain. In the analysis presented here, `Fresh` was the first variable to be dropped, followed by `FTF`, `Accept`, and `X50M`. The final regression contains five predictors, with an R^2 of nearly .99. What follows is a listing of the R commands and output for this analysis.

[9] The weights used by *U. S. News and World Report* can be found at: `http://chronicle.com/free/v53/i38/38a01301.htm`

```
> lm.ranks <- lm(Score~Peer+Fresh+Grad+U20+X50M+FTF+Top10+Accept+Alumni)
> summary(lm.ranks)

Call:
lm(formula = Score ~ Peer + Fresh + Grad + U20 + X50M + FTF +
    Top10 + Accept + Alumni)

Residuals:
      Min        1Q    Median        3Q       Max
-5.649724 -1.548107 -0.007142  1.691242  4.804091

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -48.67296    9.85337  -4.940 3.44e-06 ***
Peer         14.55833    0.72692  20.027  < 2e-16 ***
Fresh         0.10829    0.13779   0.786  0.43392
Grad          0.30309    0.06315   4.800 6.05e-06 ***
U20           0.18621    0.02799   6.654 1.95e-09 ***
X50M         -0.17065    0.06361  -2.682  0.00865 **
FTF           0.05994    0.04359   1.375  0.17239
Top10         0.08958    0.01728   5.185 1.26e-06 ***
Accept       -0.04294    0.02142  -2.005  0.04790 *
Alumni        0.23507    0.03187   7.375 6.70e-11 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 2.159 on 93 degrees of freedom
Multiple R-squared: 0.987,Adjusted R-squared: 0.9857
F-statistic: 782.5 on 9 and 93 DF,  p-value: < 2.2e-16

>
> lm.ranks <- lm(Score~Peer+Grad+U20+X50M+FTF+Top10+Accept+Alumni)
> summary(lm.ranks)

Call:
lm(formula = Score ~ Peer + Grad + U20 + X50M + FTF + Top10 +
    Accept + Alumni)

Residuals:
    Min      1Q  Median      3Q     Max
-5.5929 -1.5074  0.1136  1.6074  4.9963

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -41.77818    4.47670  -9.332 4.83e-15 ***
Peer         14.62023    0.72117  20.273  < 2e-16 ***
```

```
Grad          0.33777    0.04507   7.494 3.63e-11 ***
U20           0.18559    0.02792   6.648 1.94e-09 ***
X50M         -0.15691    0.06104  -2.571   0.0117 *
FTF           0.05848    0.04346   1.346   0.1817
Top10         0.09208    0.01695   5.433 4.34e-07 ***
Accept       -0.04645    0.02091  -2.222   0.0287 *
Alumni        0.23653    0.03176   7.448 4.51e-11 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 2.155 on 94 degrees of freedom
Multiple R-squared: 0.9869,Adjusted R-squared: 0.9858
F-statistic: 883.8 on 8 and 94 DF,  p-value: < 2.2e-16

> lm.ranks <- lm(Score~Peer+Grad+U20+X50M+Top10+Accept+Alumni)
> summary(lm.ranks)

Call:
lm(formula = Score ~ Peer + Grad + U20 + X50M + Top10 + Accept +
    Alumni)

Residuals:
    Min      1Q  Median      3Q     Max
-5.6443 -1.5618  0.2362  1.5709  5.3962

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -38.59289    3.81574 -10.114  < 2e-16 ***
Peer         14.98132    0.67223  22.286  < 2e-16 ***
Grad          0.33576    0.04524   7.422 4.88e-11 ***
U20           0.18553    0.02804   6.618 2.15e-09 ***
X50M         -0.13813    0.05968  -2.315   0.0228 *
Top10         0.09137    0.01701   5.371 5.55e-07 ***
Accept       -0.03723    0.01984  -1.877   0.0636 .
Alumni        0.25140    0.02990   8.409 4.12e-13 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 2.164 on 95 degrees of freedom
Multiple R-squared: 0.9866,Adjusted R-squared: 0.9856
F-statistic:  1001 on 7 and 95 DF,  p-value: < 2.2e-16

> lm.ranks <- lm(Score~Peer+Grad+U20+X50M+Top10+Alumni)
> summary(lm.ranks)

Call:
```

```
lm(formula = Score ~ Peer + Grad + U20 + X50M + Top10 + Alumni)

Residuals:
   Min     1Q Median     3Q    Max
-5.723 -1.606  0.202  1.462  5.561

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -43.55055    2.78982 -15.611  < 2e-16 ***
Peer         15.15156    0.67478  22.454  < 2e-16 ***
Grad          0.35437    0.04472   7.925 4.09e-12 ***
U20           0.19409    0.02802   6.926 4.93e-10 ***
X50M         -0.15318    0.05991  -2.557   0.0121 *
Top10         0.10032    0.01654   6.064 2.62e-08 ***
Alumni        0.26396    0.02952   8.941 2.80e-14 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 2.192 on 96 degrees of freedom
Multiple R-squared: 0.9861,Adjusted R-squared: 0.9853
F-statistic:  1138 on 6 and 96 DF,  p-value: < 2.2e-16

> lm.ranks <- lm(Score~Peer+Grad+U20+Top10+Alumni)
> summary(lm.ranks)

Call:
lm(formula = Score ~ Peer + Grad + U20 + Top10 + Alumni)

Residuals:
    Min      1Q  Median      3Q     Max
-5.6769 -1.4169  0.1455  1.4215  6.2248

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -46.60744    2.59166 -17.984  < 2e-16 ***
Peer         14.28221    0.59928  23.832  < 2e-16 ***
Grad          0.38264    0.04455   8.590 1.47e-13 ***
U20           0.24177    0.02151  11.240  < 2e-16 ***
Top10         0.08877    0.01636   5.425 4.26e-07 ***
Alumni        0.28135    0.02954   9.526 1.42e-15 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 2.254 on 97 degrees of freedom
Multiple R-squared: 0.9852,Adjusted R-squared: 0.9844
F-statistic:  1290 on 5 and 97 DF,  p-value: < 2.2e-16
```

As can be seen in the summary statement just above, five variables are significant predictors of the overall score: peer ratings, graduation rates, percentage of classes under 20, percentage of students coming from the top 10% of their high school classes, and rate of alumni giving. However, not all of these predictors contribute equally to the equation, for peer ratings appear to have a greater impact than the other four. This raises the question of whether the other four are necessary. In order to test this, I used the R `anova` command to compare two models: One with peer ratings only, and one with peer ratings plus the other four predictors. Here are the commands and output:

```
> lm1.ranks <- lm(Score~Peer+Grad+U20+Top10+Alumni)
> lm2.ranks <- lm(Score~Peer)
> summary(lm1.ranks)

Call:
lm(formula = Score ~ Peer + Grad + U20 + Top10 + Alumni)

Residuals:
    Min      1Q  Median      3Q     Max
-5.6769 -1.4169  0.1455  1.4215  6.2248

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -46.60744    2.59166 -17.984  < 2e-16 ***
Peer         14.28221    0.59928  23.832  < 2e-16 ***
Grad          0.38264    0.04455   8.590 1.47e-13 ***
U20           0.24177    0.02151  11.240  < 2e-16 ***
Top10         0.08877    0.01636   5.425 4.26e-07 ***
Alumni        0.28135    0.02954   9.526 1.42e-15 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 2.254 on 97 degrees of freedom
Multiple R-squared: 0.9852,Adjusted R-squared: 0.9844
F-statistic:  1290 on 5 and 97 DF,  p-value: < 2.2e-16

> summary(lm2.ranks)

Call:
lm(formula = Score ~ Peer)

Residuals:
     Min       1Q   Median       3Q      Max
-17.0075  -4.9820   0.7783   5.4719  16.7783

Coefficients:
```

```
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -36.922      4.379  -8.432 2.49e-13 ***
Peer          26.837      1.209  22.197  < 2e-16 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 7.486 on 101 degrees of freedom
Multiple R-squared: 0.8299,Adjusted R-squared: 0.8282
F-statistic: 492.7 on 1 and 101 DF,  p-value: < 2.2e-16

> anova(lm2.ranks,lm1.ranks)
Analysis of Variance Table

Model 1: Score ~ Peer
Model 2: Score ~ Peer + Grad + U20 + Top10 + Alumni
  Res.Df    RSS  Df Sum of Sq      F    Pr(>F)
1    101 5659.7
2     97  492.8   4    5166.9 254.23 < 2.2e-16 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

The output from the anova command indicates that all five variables make a significant contribution to prediction of the overall score. However, based on the t-tests, it appears that peer ratings are more important than the others. One way to look at this is in terms of variance accounted for: Peer ratings alone account for 83% of variance in overall scores, whereas the other four variables collectively account for an additional 15% or 16% beyond that accounted for by peer ratings. Another way to look at this is in terms of the amount of change in the dependent variable (overall score) that would occur with changes in each of the independent variables. A summary of this approach is presented in Table 2, which also indicates that the peer ratings are the strongest determinant of the overall score. In fact, just a .1 increase in the peer assessment would yield an increase of 1.43 in the overall score, which could cause a school like TCU to leapfrog several schools in the *U. S. News & World Report* rankings.

Variable	Change	Weight
Peer Assessment	1.43	.25
Graduation Rate	.38	.16
Classes Under 20	.24	.06
Freshmen Top 10	.09	.06
Alumni Giving	.28	.05

Table 2: Relative impact of five predictors on the overall score, based on the regression analysis and the original weighting. Change driven by peer ratings is based on an increment of .1, change driven by the remaining variables is based on an increment of 1.

Predicting Peer Ratings

Given that the peer ratings are a significant predictor of the overall score, the question naturally arises, "What predicts the peer ratings?" The goal of the analysis presented in this subsection is to determine which, if any, of the variables in this data set (excluding the overall score) are significant predictors of peer assessments. The same algorithm—manual backwards elimination—used in the previous analysis of the overall score is used here in the analysis of the peer ratings. Four variables are dropped in sequence: alumni giving, freshman retention, top 10 percent, and classes under 20. The remaining four predictors each have p-values less than .001, and collectively account for 74% of the variance in peer ratings.

```
> lm.ranks <- lm(Peer~Fresh+Grad+U20+X50M+FTF+Top10+Accept+Alumni)
> summary(lm.ranks)

Call:
lm(formula = Peer ~ Fresh + Grad + U20 + X50M + FTF + Top10 +
    Accept + Alumni)

Residuals:
      Min        1Q    Median        3Q       Max
-0.828374 -0.211791  0.001987  0.199470  0.715443

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.989580   1.382943  -1.439 0.153569
Fresh        0.020539   0.019436   1.057 0.293324
Grad         0.010981   0.008888   1.235 0.219750
U20          0.009743   0.003842   2.536 0.012857 *
X50M         0.029039   0.008515   3.410 0.000958 ***
FTF          0.022441   0.005736   3.913 0.000173 ***
Top10        0.003295   0.002428   1.357 0.177958
Accept      -0.006217   0.002971  -2.093 0.039082 *
Alumni       0.002116   0.004517   0.468 0.640633
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 0.3064 on 94 degrees of freedom
Multiple R-squared: 0.7698,Adjusted R-squared: 0.7502
F-statistic:  39.3 on 8 and 94 DF,  p-value: < 2.2e-16

> lm.ranks <- lm(Peer~Fresh+Grad+U20+X50M+FTF+Top10+Accept)
> summary(lm.ranks)

Call:
lm(formula = Peer ~ Fresh + Grad + U20 + X50M + FTF + Top10 +
    Accept)

Residuals:
      Min        1Q    Median        3Q       Max
-0.834624 -0.205374  0.001467  0.192531  0.719938

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -2.083232   1.362775  -1.529 0.129669
Fresh        0.021121   0.019316   1.093 0.276956
Grad         0.011289   0.008827   1.279 0.204053
U20          0.009786   0.003825   2.559 0.012089 *
X50M         0.028002   0.008188   3.420 0.000925 ***
FTF          0.023507   0.005243   4.484 2.05e-05 ***
Top10        0.003248   0.002416   1.344 0.182058
Accept      -0.006648   0.002813  -2.364 0.020133 *
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 0.3051 on 95 degrees of freedom
Multiple R-squared: 0.7693,Adjusted R-squared: 0.7523
F-statistic: 45.25 on 7 and 95 DF,  p-value: < 2.2e-16

> lm.ranks <- lm(Peer~Grad+U20+X50M+FTF+Top10+Accept)
> summary(lm.ranks)

Call:
lm(formula = Peer ~ Grad + U20 + X50M + FTF + Top10 + Accept)

Residuals:
     Min       1Q   Median       3Q      Max
-0.76555 -0.19262  0.00799  0.18924  0.76677

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.755236   0.618832  -1.220 0.225294
Grad         0.018343   0.006031   3.041 0.003035 **
U20          0.009789   0.003829   2.557 0.012132 *
X50M         0.030926   0.007747   3.992 0.000128 ***
FTF          0.023651   0.005247   4.508 1.85e-05 ***
Top10        0.003776   0.002369   1.594 0.114280
Accept      -0.007484   0.002710  -2.762 0.006880 **
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 0.3054 on 96 degrees of freedom
```

```
Multiple R-squared: 0.7664,Adjusted R-squared: 0.7518
F-statistic: 52.49 on 6 and 96 DF,  p-value: < 2.2e-16

> lm.ranks <- lm(Peer~Grad+U20+X50M+FTF+Accept)
> summary(lm.ranks)

Call:
lm(formula = Peer ~ Grad + U20 + X50M + FTF + Accept)

Residuals:
      Min        1Q    Median        3Q       Max
-0.785155 -0.197299 -0.008085  0.206521  0.794412

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.981335   0.607114  -1.616 0.109257
Grad         0.023097   0.005283   4.372 3.09e-05 ***
U20          0.011110   0.003767   2.949 0.003993 **
X50M         0.036025   0.007111   5.066 1.94e-06 ***
FTF          0.023835   0.005287   4.508 1.83e-05 ***
Accept      -0.008844   0.002592  -3.412 0.000943 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 0.3078 on 97 degrees of freedom
Multiple R-squared: 0.7602,Adjusted R-squared: 0.7478
F-statistic:  61.5 on 5 and 97 DF,  p-value: < 2.2e-16

> lm.ranks <- lm(Peer~Grad+X50M+FTF+Accept)
> summary(lm.ranks)

Call:
lm(formula = Peer ~ Grad + X50M + FTF + Accept)

Residuals:
     Min       1Q   Median       3Q      Max
-0.60580 -0.18729 -0.02698  0.17077  0.86866

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.610706   0.616850  -0.990 0.324593
Grad         0.026641   0.005343   4.986 2.66e-06 ***
X50M         0.023924   0.006032   3.966 0.000139 ***
FTF          0.025863   0.005444   4.751 6.94e-06 ***
Accept      -0.011467   0.002528  -4.535 1.63e-05 ***
---
```

```
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 0.3197 on 98 degrees of freedom
Multiple R-squared: 0.7387,Adjusted R-squared: 0.728
F-statistic: 69.26 on 4 and 98 DF,  p-value: < 2.2e-16
```

The four remaining variables are graduation rates, classes 50 or more, percent full time faculty, and the acceptance rate (which has a negative association with peer ratings). This analysis suggests that these four characteristics of national universities are at least correlated with perceptions that guide peer assessments in the *U. S. News & World Report* rankings. These findings may present clues about the basis of those perceptions, and how they might be changed.

The analyses presented in this Primer are just a beginning. Other analyses are possible that might shed further light on these rankings. Here I list a few possibilities:

- To begin with, the data set is incomplete. I didn't include test score (SAT/ACT) because I didn't have a way to convert the two tests to a common scale (i.e., percentiles). Further, the published data set only includes the top 113 universities, but there are another 149 or so that are ranked. I believe a complete data set is available on the *Chronicle of Higher Education* website, but I haven't verified this.

- A natural next step is to fit a path model to these data, with peer ratings as a mediator, overall score as the dependent variable, and the remaining variables as predictors. The analyses presented here suggest that the peer ratings would mediate some, but not all, of the effects on the overall score. This will be Part II of the Primer, if and when time permits.

- Also instructive would be a set of analyses that determine how my own institution—TCU—stands relative to certain comparison groups (e.g., universities in the former Southwestern Conference, or the current Mountain West Conference). Analyses that focus on deviations from what is predicted would be especially useful here.

- Longitudinal data are also available, which would make possible analyses of trend. What characteristics define those schools that are improving their scores? What characteristics define those schools that are falling behind? How susceptible to change are the peer ratings?

- Finally, is it possible to shed light on the validity of these rankings themselves, using these data? Are they an exercise in sophistry, or do they provide value? Are they merely advertising, or do they have substantive meaning?

Some Advanced Capabilities

I conclude this Primer with a few comments about R's advanced capabilities.

Graphical Display

The graphics (boxplots and scatterplots) presented in this Primer barely scratch the surface of R's capabilities. Here are some key resources:

- *The R Graphics Package* (http://stat.ethz.ch/R-manual/R-patched/library/graphics/html/00Index.html)
- *The R Graphics Gallery* (http://addictedtor.free.fr/graphiques/)
- P. Murrell, *R Graphics*, Chapman & Hall, 2006. (See also http://www.stat.auckland.ac.nz/ paul/RGraphics/rgraphics.html)
- D. Sarkar, *Lattice: Multivariate Data Visualization with R*, Springer, 2008. (See also http://lmdvr.r-forge.r-project.org/figures/figures.html)

Linear Models

As with R's graphics capabilities, the regression analyses presented here barely scratch the surface of R's modeling potential. Here are some key resources:

- M. J. Crawley, *The R Book*, Wiley, 2007. (http://www.wiley.com/WileyCDA/WileyTitle/productCd-0470510242.html)
- J. Fox, *An R and S-Plus Companion to Applied Regression*, Sage, 2002. (See also http://socserv.mcmaster.ca/jfox/)
- J. J. Faraway, *Linear Models with R*, Chapman & Hall, 2005. (Available as a pdf file: http://cran.r-project.org/doc/contrib/Faraway-PRA.pdf)

Hierarchical Linear Models

The nlme package in R can be used to estimate and test a wide range of linear and nonlinear mixed effects models, including hierarchical linear models (HLM):

- *R Documentation: Linear and nonlinear mixed effects models* (http://web.njit.edu/all_topics/Prog_Lang_Docs/html/library/nlme/html/00Index.html)
- J. Fox, *Linear Mixed Models: Appendix to An R and S-Plus Companion to Applied Regression* (cran.r-project.org/doc/contrib/Fox-Companion/appendix-mixed-models.pdf)
- J. Fox, *Statistical Computing Using R/S* (http://socserv.mcmaster.ca/jfox/Courses/R-course/index.html)
- J. J. Faraway, *Extending the Linear Model with R: Generalized Linear, Mixed Effects and Nonparametric Regression*, Chapman & Hall, 2005. (See also http://www.maths.bath.ac.uk/~jjf23/)

Structural Equation Models

The sem package in R is a capable alternative to AMOS, LISREL, and EQS:

- J. Fox, *Structural Equation Models (sem)* (`http://socserv.mcmaster.ca/jfox/Misc/sem/index.html`)
- W. Revelle, *Chapter 4: sem in R and in LISREL* (`www.personality-project.org/r/sem.chap4.pdf`)

Rasch Models

eRm is an R package that allows researchers to investigate Rasch psychometric models:

- *R-Forge: eRm* (`http://r-forge.r-project.org/projects/erm/`)
- P. Mair & R. Hatzinger, Extended Rasch Modeling: The eRm Package for the Application of IRT Models in R, *Journal of Statistical Software*, Vol. 20, Issue 9, Feb 2007. (`http://www.jstatsoft.org/v20/i09`)
- *Journal of Statistical Software: Special Issue on Psychometrics in R* (`http://www.jstatsoft.org/v20`)[10]

The list of topics and sources presented here is highly selective. Indeed, as I prepared this document, I was surprised again and again about how many sources are available. I have tried to present the best and most relevant sources, but have little faith in the result. There is much more available, and much of it is at least as good as what is listed here. Part of the problem is that as far as I can tell, in the world of statistical software, R is "where the action is." The growth of R is impressive, from several standpoints, including the number of users, available texts, and the software itself. The only solution is to jump in and begin using R, and at the same time begin exploring the many resources that are available. Best wishes with your R adventure!

[10] Specials issues on R are available for Chemistry, Ecology, and Econometrics.

Appendix: "Best National Universities"

Here are the data used in the "America's Best Colleges" example:[11]

School	Score	Peer	Fresh	Grad	U20	50M	FTF	Top10	Accept	Alumni
Harvard	100	4.9	97	97	75	9	93	95	9	41
Prince	99	4.8	98	95	73	10	93	96	10	60
Yale	98	4.8	99	96	75	8	87	97	10	43
MIT	94	4.9	98	93	64	12	90	97	12	37
Stanford	94	4.9	98	95	74	11	99	91	10	36
CalTech	93	4.6	98	89	69	8	98	99	17	29
Penn	93	4.5	98	95	74	7	86	96	16	38
Columbia	90	4.5	98	94	76	8	91	94	11	36
Duke	90	4.4	97	94	70	5	97	90	23	40
Chicago	90	4.6	97	90	72	4	87	83	35	32
Dartmouth	89	4.3	98	93	64	9	93	91	15	53
Northwest	87	4.3	97	93	75	7	96	85	27	31
WashU-SL	87	4.1	97	92	73	9	94	95	17	37
Cornell	86	4.5	96	92	60	17	98	87	21	34
JohnsHop	85	4.5	97	91	65	11	97	82	24	33
Brown	84	4.3	98	95	70	9	94	92	14	40
Rice	80	4.0	97	91	68	7	93	83	25	34
Emory	79	3.9	94	88	68	7	95	88	27	36
NotreDame	79	3.9	98	95	56	10	97	86	24	51
Vandy	79	4.0	96	91	67	6	95	80	33	25
Calif	77	4.7	97	88	62	14	90	99	23	14
CMU	75	4.1	94	87	65	9	93	73	28	22
GeoTown	74	4.0	97	93	58	7	80	90	21	28
Virgina	74	4.3	97	93	49	14	98	87	35	24
UCLA	73	4.2	97	90	53	20	90	97	24	14
Michigan	72	4.4	96	88	45	18	92	92	50	18
USC	71	3.9	96	85	64	12	82	86	25	38
Tufts	70	3.6	96	89	75	4	84	80	27	23
WkForest	70	3.5	94	89	57	2	92	64	42	32
UNC	69	4.1	96	83	44	12	98	76	35	23
Brandeis	67	3.5	95	88	66	6	89	79	34	33
WillMary	65	3.7	95	92	49	7	91	79	34	23
NYU	64	3.8	92	84	58	12	75	66	37	11
BostonC	63	3.5	96	91	48	7	76	80	27	21
GaTech	62	4.0	92	78	40	22	100	66	63	31
Lehigh	62	3.2	94	83	47	10	88	93	32	33
UCSD	62	3.8	94	84	44	30	93	99	43	8
Rochester	62	3.4	94	81	62	12	86	72	41	18

[11] This data file is taken from *U. S. News & World Report*, September 1–8, 2008, and is available upon request; a more complete data set is available on the *Chronicle of Higher Ed* website (`http://chronicle.com/free/v53/i38/38a01101.htm`).

Wisconsin	62	4.1	93	80	44	18	94	60	56	13
Illinois	61	4.0	92	82	38	19	99	55	71	14
CaseWest	59	3.4	91	81	62	10	92	66	75	14
RPI	59	3.5	92	82	53	10	95	64	49	18
UWash	59	3.9	93	75	35	17	93	86	65	17
UCD	58	3.8	90	79	35	28	94	95	59	12
UCI	58	3.5	94	80	49	17	90	96	56	13
UCSB	58	3.5	91	85	50	17	95	96	54	19
PennSt	57	3.7	94	84	33	17	96	45	51	22
UTexas	57	4.0	93	78	35	23	97	69	51	16
Florida	56	3.6	94	81	41	20	99	76	42	17
Yeshiva	55	2.8	88	85	69	1	87	61	69	23
Tulamr	54	3.3	88	76	62	8	82	50	44	23
MiamiFlor	54	3.2	89	76	47	6	87	65	38	19
GeoWash	53	3.4	92	78	57	11	69	66	37	10
Syracuse	53	3.3	92	82	62	8	83	42	51	19
Maryland	53	3.6	93	80	34	14	89	71	47	14
OSU	52	3.6	91	71	35	19	89	52	59	16
Pepper	52	3.1	89	80	68	3	80	46	35	13
Georgia	51	3.4	93	78	38	11	92	53	54	14
Pitt	51	3.4	90	75	43	15	92	48	56	15
BostonU	50	3.4	91	82	52	10	82	51	59	7
Clemson	49	3.1	89	78	48	11	95	52	50	28
Fordham	49	3.1	89	80	47	1	80	43	42	21
Minnesota	49	3.6	87	63	43	16	94	44	57	15
Rutgers	48	3.3	89	73	41	20	85	40	56	15
TexasAM	48	3.5	92	78	22	23	94	45	76	17
MiamiOhio	47	3.2	90	80	35	9	85	35	75	18
Purdue	47	3.7	85	69	34	19	96	31	79	17
SMU	47	3.0	88	71	58	9	85	40	50	14
UConn	47	3.1	92	75	44	16	90	40	49	19
Iowa	47	3.5	84	66	51	10	98	23	83	14
Indiana	46	3.7	88	72	39	17	94	31	70	13
MichState	46	3.4	91	74	25	22	95	29	74	15
Delaware	46	3.0	90	78	43	13	93	41	56	16
VaTech	46	3.3	90	78	23	22	95	40	67	21
WPI	46	2.7	92	76	70	9	92	48	66	18
Baylor	45	3.1	84	72	41	9	91	45	44	26
Marquette	44	3.0	90	75	38	11	80	34	67	19
SUNY-Bing	44	3.0	90	79	41	14	86	49	39	10
Colorado	44	3.4	84	67	50	14	85	25	82	8
Clark	43	2.8	87	76	58	5	94	32	56	24
ColoSM	43	3.0	83	62	43	13	86	53	61	28
StLouis	43	2.9	85	75	52	6	85	37	80	17
American	42	2.9	88	73	46	3	79	50	53	14
NCState	42	3.1	90	69	32	16	96	34	60	24

SIT	42	2.6	89	76	38	8	80	47	51	22
SUNY-SB	40	3.2	88	59	36	22	84	36	43	13
Arizona	40	3.5	79	56	38	12	99	34	80	7
UCSC	40	3.1	89	68	32	23	88	96	82	12
Missouri	40	3.2	85	68	47	14	99	26	86	13
FloridaSt	39	3.0	88	69	34	15	92	33	55	22
Howard	39	2.8	89	68	63	5	85	23	54	19
IIT	39	2.7	84	67	58	3	78	43	57	13
UMass	39	3.2	83	67	40	17	95	22	66	12
SanDiego	39	2.7	85	74	41	.4	73	38	48	11
Pacific	39	2.5	84	67	51	5	80	41	59	12
Dayton	38	2.5	87	76	36	5	80	23	82	24
Oklahoma	38	2.9	85	62	49	10	91	33	89	21
Oregon	38	3.3	85	67	39	16	87	23	87	17
SCarolina	38	2.9	85	63	46	10	88	29	59	22
Tennessee	38	3.0	81	58	32	8	98	39	71	12
BYU	37	2.9	90	73	47	10	89	49	74	15
TCU	37	2.6	84	69	46	7	82	30	49	23
UNH	37	2.8	86	73	44	15	87	24	59	11

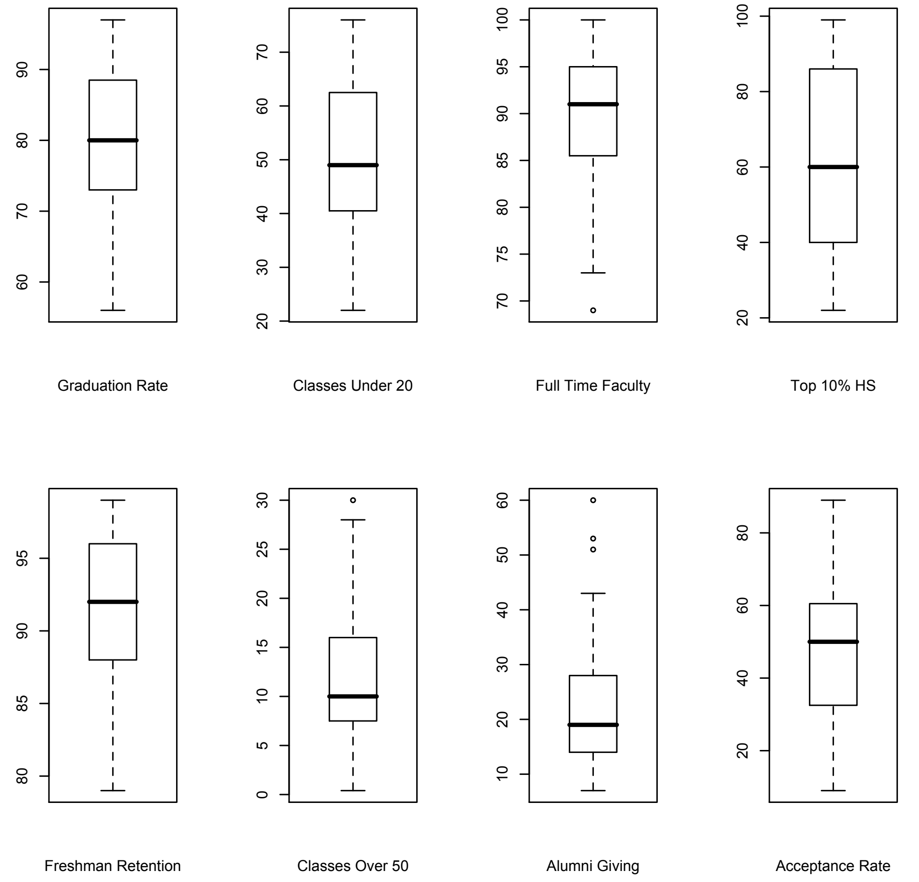

Figure 2: Box-and-whisker plots for the eight predictor variables.

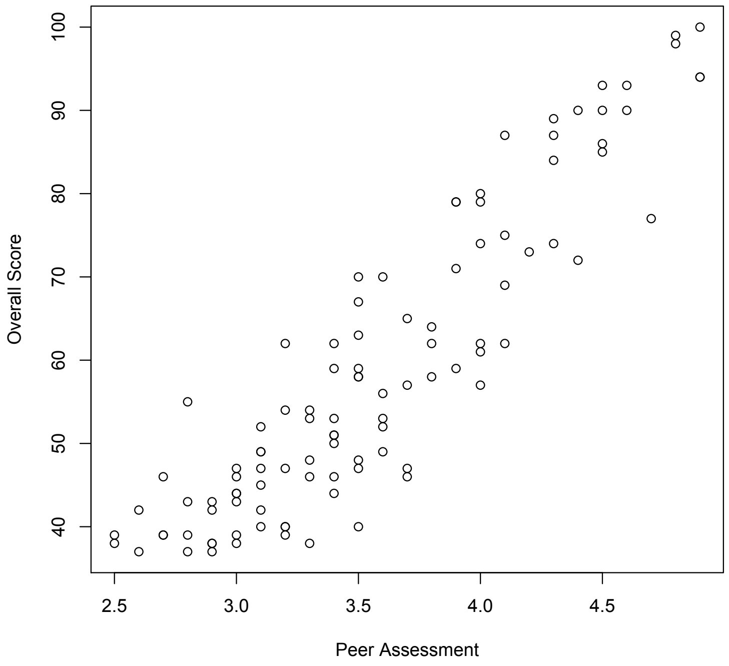

Figure 3: Scatterplot showing the association between the overall score and peer ratings.

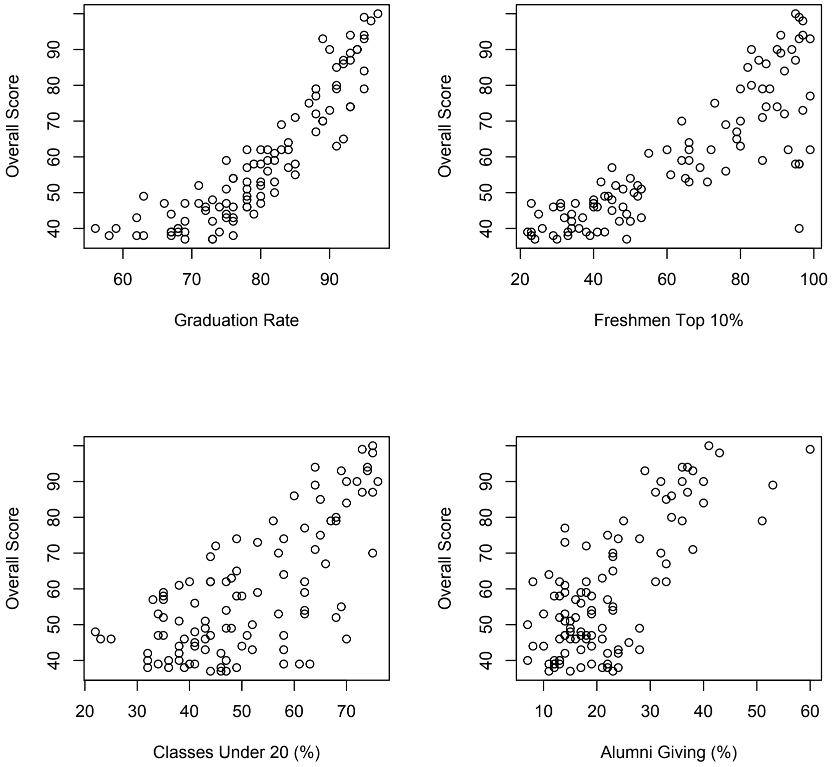

Figure 4: Scatterplot showing the association between the overall score and those predictor variables (e.g., graduation rate) that are significant predictors in the multiple regression analysis analysis reported in the text.

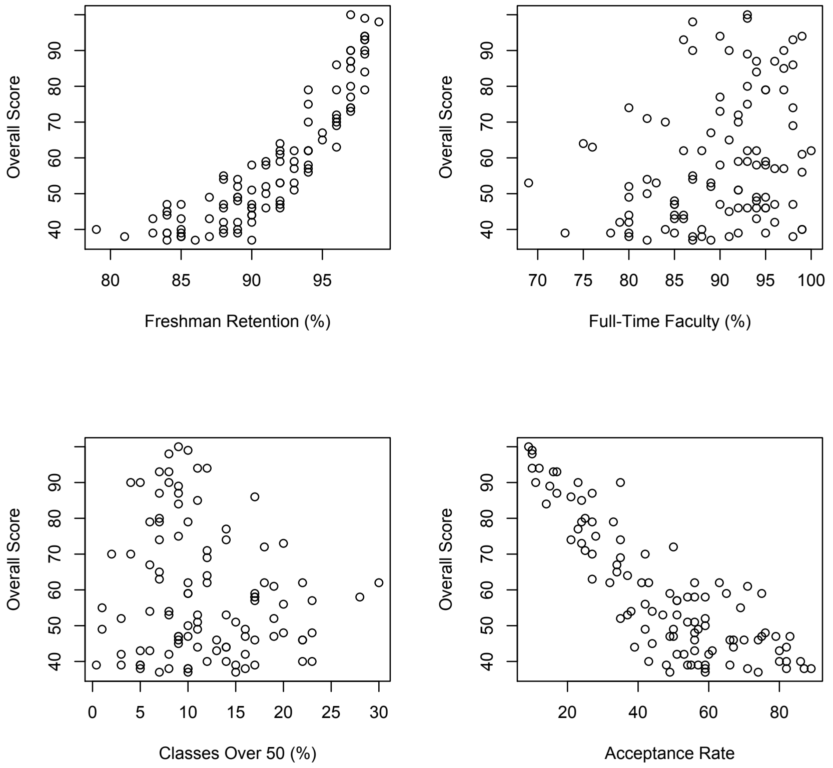

Figure 5: Scatterplot showing the association between the overall score and those predictor variables (e.g., freshman retention) that are *not* significant predictors in the multiple regression analysis reported in the text.

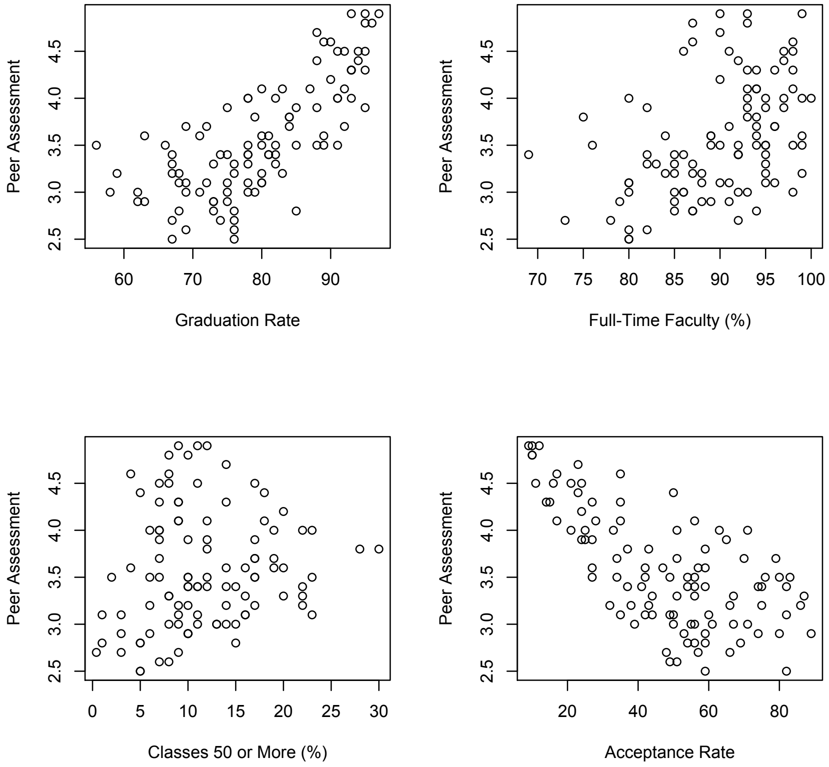

Figure 6: Scatterplot showing the association between the peer ratings and those predictor variables (e.g., graduation rate) that are significant predictors in the multiple regression analysis reported in the text.

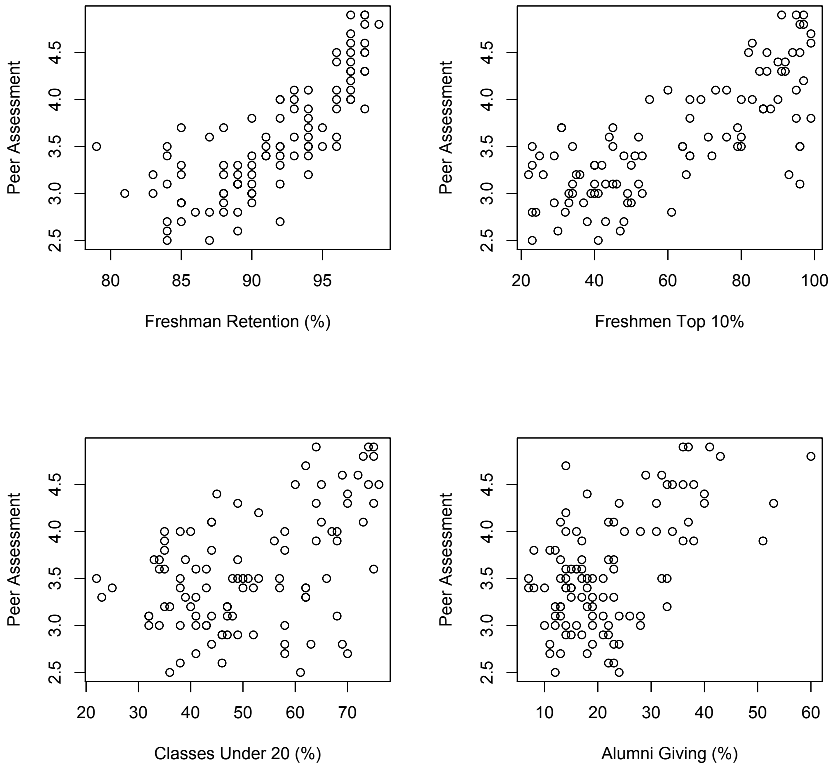

Figure 7: Scatterplot showing the association between the peer ratings and those predictor variables (e.g., freshman retention) that are *not* significant predictors in the multiple regression analysis reported in the text.