According to the so-called self-esteem movement in education (which I suspect you've all been victimized by), a student's self-esteem is critical to his/her school achievement and, therefore, every opportunity should be taken to build the student's sense of self-esteem. Do the data support the role of self-esteem in achievement?

PART ONE: DATA SUMMARIZATION (and dealing with missing values)

We will attempt to duplicate an analysis in Keith's textbook Multiple Regression and Beyond (1st ed., 2006). Get and summarize the data as follows.

> nels = read.csv("http://ww2.coastal.edu/kingw/psyc480/data/nels.csv")
> dim(nels)
[1] 1000  5
> summary(nels)    # note: NA indicates the presence of missing values

std.achieve.score  SES  prev.grades  self.esteem
  Min.   :28.94 Min. : -2.41400 Min. : 0.50 Min. : -2.30000
  1st Qu.:43.85 1st Qu.: -0.56300 1st Qu.:2.50 1st Qu.: -0.45000
  Median :51.03 Median : -0.08650 Median :3.00 Median : -0.06000
  Mean   :50.92 Mean   : -0.03121 Mean   :2.97 Mean   : 0.03973
  3rd Qu.:59.04 3rd Qu.:  0.53300 3rd Qu.:3.50 3rd Qu.:  0.52000
  Max.   :69.16 Max.   :  1.87400 Max.   :4.00 Max.   :  1.35000
  NA's   :77          NA's   :17     NA's   :59

locus.of.control
  Min.   :-2.16000
  1st Qu.:-0.36250
  Median : 0.07000
  Mean   : 0.04703
  3rd Qu.: 0.48000
  Max.   : 1.43000
  NA's   :60

NELS stands for National Educational Longitudinal Study. NELS was a study begun in 1988, when a large, nationally representative sample of 8th graders was tested, and then retested again in 10th and 12th grades. These aren't the complete NELS data but 1000 cases chosen, I assume, at random. (Read all about NELS here if you're curious: https://nces.ed.gov/surveys/nels88/.)

The variables (also a very much incomplete set of the originals) are:
std.achieve.score - a standardized achievement test score (I believe this one was from a history or social studies test.)
SES - standardized assessment of socioeconomic status of the student's family
prev.grades - student's previous GPA on the usual 0-4 scale
self.esteem - standardized self-esteem score (extracted from multiple questions asked on a survey constructed by the NELS researchers)
locus.of.control - standardized locus of control score (extracted from multiple questions asked on a survey constructed by the NELS researchers - caution: unlike most such scales, on this one low scores indicate external LOC)

Let's begin by shortening up those column names a bit.

> colnames(nels) = c("achieve","SES","grades","esteem","LOC")

It appears that Keith used SPSS in his analysis. Let's see how close we can come to getting everything (of value) that Keith got from SPSS.

We already have the means, but let's get them again. Since this data frame is all numeric, we can use the apply() function. (Note: For the means, we could also use colMeans(nels).)

> apply(nels, 2, mean)    # 2 indicates column means (1 would be row means)
   achieve  SES  grades  esteem  LOC
NA  -0.031205  NA  NA  NA
Uh oh! What went wrong? Notice in our previous summary that the variables have missing values. When there are missing values, the mean() function in R will fail, because R is smart enough to know, when there are missing values, that there really isn't a mean in the conventional sense. We need to tell R to ignore the missing values by using the appropriate option.

```r
> apply(nels, 2, mean, na.rm=T)   # remove NAs
  achieve       SES      grades      esteem         LOC
         50.91808234 -0.03120500  2.97039674  0.03973433  0.04703191

Standard deviations can be obtained the same way. (Do they seem right for standardized scores?)

```r
> apply(nels, 2, sd, na.rm=T)
  achieve       SES    grades    esteem       LOC
         9.9415281 0.7787984 0.7522191 0.6728771 0.6235697

Sample sizes are another problem. The na.rm= option does not work with the R length() function. We know that we have a thousand cases, and we know how many missing values there are in each variable (above), so we could rely on simple subtraction, but it seems there should be a more elegant way. So I'm going to write a function (sum(is.na(x)) counts NAs--try it) to do it. The function will be placed in your workspace and will remain there until you delete it.

```r
> sampsize = function(x) length(x) - sum(is.na(x))
> apply(nels, 2, sampsize)
  achieve     SES  grades  esteem     LOC
         923    1000     983     941     940

The next thing Keith got from SPSS was a correlation matrix. The cor() function in R will also fail if there are missing values, because R wants to know how we prefer to deal with them. Do we want to use pairwise complete cases? Or do we want to include only complete observations, i.e., cases in which there are no missing values on ANY variable? That was not clear in the SPSS output, so I tried it both ways. It turns out SPSS was using only complete observations.

```r
> cor(nels, use="complete.obs")   # use="pairwise.complete" is the alternative
  achieve       SES    grades    esteem       LOC
achie 1.0000000 0.4299589 0.4977062 0.1727001 0.2477357
SES  0.4299589 1.0000000 0.3254019 0.1322685 0.1942161
grades 0.4977062 0.3254019 1.0000000 0.1670206 0.2283377
esteem 0.1727001 0.1322685 0.1670206 1.0000000 0.5854103
LOC  0.2477357 0.1942161 0.2283377 0.5854103 1.0000000

Note the correlations with achievement scores. The correlation with esteem is the lowest, but it is nevertheless positive and statistically significant. How many cases contributed to this correlation? Interpret the other correlations.

```r
> sum(complete.cases(nels))   # counts the number of complete cases
[1] 887
> source("http://ww2.coastal.edu/kingw/psyc480/functions/rcrit.R")
> r.crit(df=885)   # critical r for 887 cases and alpha=.05
   df    alpha   1-tail   2-tail
885.0000  0.0500   0.0553   0.0658

PART TWO: STANDARD MULTIPLE REGRESSION

The next thing Keith got was a standard multiple regression. Easy peasy! Note: by doing the regression in the following fashion, we are tacitly assuming that there are no interactions between these variables. In fact, there are significant interactions. For example, SES:LOC is significant. What does that mean?

Nevertheless, we are following Keith's analysis, so on we go.

```r
> lm.out = lm(achieve ~ SES + grades + esteem + LOC, data=nels)
> summary(lm.out)
```
Call:
```
lm(formula = achieve ~ SES + grades + esteem + LOC, data = nels)
```

Residuals:
```
       Min        1Q  Median        3Q       Max
-24.8402 -5.3648   0.2191   5.6748  20.3933
```

Coefficients:
```
                Estimate  Std. Error t value Pr(>|t|)
(Intercept)     35.5174     1.2256  28.981   <2e-16 ***
SES             3.6903     0.3776   9.772   <2e-16 ***
grades          5.1502     0.3989  12.910   <2e-16 ***
estem           0.2183     0.5007   0.436    0.663
LOC             1.5538     0.5522   2.814    0.005 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Residual standard error: 8.041 on 882 degrees of freedom
(113 observations deleted due to missingness)
Multiple R-squared: 0.3385, Adjusted R-squared: 0.3355
F-statistic: 112.8 on 4 and 882 DF,  p-value: < 2.2e-16

Notice that the effect of esteem is not significant. How do we reconcile this with the finding above that there is a significant positive correlation between achievement and esteem? Two ways! Interpret the rest of the table.

The next thing Keith got were confidence intervals for his regression coefficients.
```
> confint(lm.out)
          2.5 %   97.5 %
(Intercept) 33.1120344 37.922735
SES          2.9491568  4.431499
grades       4.3672598  5.933147
esteem      -0.7643607  1.201002
LOC          0.4700688  2.637597
```

Interpret. What do these intervals mean?

The next thing Keith got was an ANOVA table summarizing the regression. This isn't really necessary, but just for kicks...
```
> print(anova(lm.out), digits=7)  # anova(lm.out) is the basic function
Analysis of Variance Table
Response: achieve
            Df Sum Sq  Mean Sq   F value   Pr(>F)
SES          1 15938.64 15938.645 246.49535 <.001 ***
grades      1 12344.62 12344.616 190.91274 <.001 ***
estem       1   391.62   391.623   6.05655  0.0140453 *
LOC          1  512.00   512.001   7.91823  0.0050028 **
Residuals   882 57031.03    64.661
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

This is exactly the result we would have gotten by using aov() to begin with. It is a little more elaborate than SPSS's table (i.e., R is giving us more information than SPSS gave Keith). To get that table, we have to collapse the first four lines of this table and add a total line.

```
df  SS    MS     F    p
Regression  4 29186.88 7296.72 112.846 <.001
Residuals  882 57031.03  64.661
Total      886 86217.91
```

Where have we seen that F-value before? Some more demonstrations...
> 29186.88 / 86217.91   # what is this?
[1] 0.3385246
> sqrt(15938.64 / 86217.91)   # what is this? what if we didn't take sqrt?
[1] 0.4299589
Here's something new. What is this? (Where have we already seen this?)

> 1 - (64.661 / (86217.91 / 886))
[1] 0.335525

Before we leave this behind, notice that the effect of esteem is now significant. What the heck is going on? Two possible interpretations! Interpret the entire R table while you're at it.

Now we want the beta coefficients, or standardized regression coefficients. These tell us, very approximately, which of the predictors are "more important" in predicting the response. Why can't we just compare the regression coefficients themselves?

We have to be careful and recall that only complete cases are being used in this analysis, so before we use scale() to convert our variables to z-scores, we need to get rid of the cases with missing values. We'll create a second data frame with only complete cases. (We could have done this right from the start if only we'd known this was the analysis we were trying to mimic. However, Keith's means and SDs are from the full data.)

> nelscomplete = na.omit(nels)   # omit cases with missing values
And NOW we'll use scale() to convert this to standardized z-scores.

> nelz = scale(nelscomplete)   # nelz is a matrix, which we don't want
> nelz = as.data.frame(nelz)   # convert from matrix to data frame

If we feel that we have to confirm that we now have z-scores...

> apply(nelz, 2, mean)   # means should be zero (why aren't they?)
achieve           SES        grades        esteem           LOC
-2.659451e-16  4.284218e-18  2.214149e-16  1.838963e-17 -5.529263e-18
> apply(nelz, 2, sd)   # SDs should be one
achieve     SES  grades  esteem     LOC
1       1       1       1       1

Looks like z-scores to me! Now we'll do the standard multiple regression with nelz. If we leave out the summary() part, R will just report the coefficients, which in this case are the betas.

> lm(achieve ~ ., data=nelz)   # ~ . means all other variables

Call:
  lm(formula = achieve ~ ., data = nelz)
Coefficients:
(Intercept)          SES       grades       esteem          LOC
  -3.265e-16    2.855e-01    3.802e-01    1.474e-02    9.684e-02

Beta coefficients are directly comparable because the variables are now all on the same scale. ROUGHLY comparable, that is. Which predictors appears to be most important (have the largest betas)?

PART THREE: REGRESSION DIAGNOSTICS

Keith doesn't show these for this analysis, but let's do it anyway.

> par(mfrow=c(2,2))   # set the graphics window to show 2 rows of 2 graphs
> plot(lm.out, 1)
> plot(lm.out, 2)
> plot(lm.out, 4)
PART FOUR: REVIEW OF SOME MORE STUFF KEITH DIDN'T DO

What's going on with self-esteem? Based on my profound knowledge and absolutely thorough review of the literature (!), I believe there is an effect of self-esteem on achievement, but I believe it is mediated by locus of control. High self-esteem makes a person feel in control of his/her destiny, and in turn this causes him to excel scholastically. What effects would this mediation effect have in the correlations and standard regression analysis above?

I propose we conduct the following mediation analysis. (Diagram it in the more conventional "triangle" diagram.)

esteem -> LOC -> achieve, esteem -> achieve

We could test this using just the lm() function, but we have the mediate() function to automate this. Why not use it?

> source("http://ww2.coastal.edu/kingw/psyc480/functions/mediate.R")

Now, I haven't previously commented about doing a mediation analysis with missing values in the data frame. In a word, DON'T! So we want to be sure to use the nelscomplete data frame as the data source.

> with(nelscomplete, mediate(x=esteem, m=LOC, y=achieve, diagram=T))

Test of Simple Mediation Effect

<table>
<thead>
<tr>
<th>X on M</th>
<th>M on Y</th>
<th>total</th>
<th>direct</th>
<th>indirect</th>
</tr>
</thead>
<tbody>
<tr>
<td>statistic</td>
<td>0.540276</td>
<td>3.579580</td>
<td>2.557419</td>
<td>0.623457</td>
</tr>
<tr>
<td>std.err</td>
<td>0.025152</td>
<td>0.644502</td>
<td>0.490301</td>
<td>0.594812</td>
</tr>
<tr>
<td>p.value</td>
<td>0.000000</td>
<td>0.000000</td>
<td>0.000000</td>
<td>0.000000</td>
</tr>
<tr>
<td>Sobel.z</td>
<td>5.38</td>
<td>n</td>
<td>887.00</td>
<td></td>
</tr>
</tbody>
</table>
Is it possible that we are mistaking an interaction between esteem and LOC for a mediation effect? (Important note: It is possible for a variable to be both a mediator and a moderator variable, but discussion of that is beyond the scope of this course.)

> summary(lm(achieve ~ esteem * LOC, data=nelscomplete))  # why nelscomplete?

Call:
  lm(formula = achieve ~ esteem * LOC, data = nelcom)

Residuals:
  Min     1Q Median     3Q    Max
-23.5112 -6.7235  -0.1815   7.5945  25.5315

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)   51.22933    0.35850 142.913  < 2e-16 ***
estime        0.60772    0.59383   1.023   0.3064
LOC           3.77150    0.65019   5.801 9.18e-09 ***
estime:LOC  -1.37750    0.67343  -2.046   0.0411 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 9.545 on 883 degrees of freedom
Multiple R-squared:  0.06696, Adjusted R-squared:  0.06379
F-statistic: 21.12 on 3 and 883 DF,  p-value: 3.206e-13

Interpret. (Keith did not consider possible interactions, so on we go.)

PART FIVE: HIERARCHICAL REGRESSION

Why? Let's talk a little about what the hierarchical regression does that the standard regression does not. I think Keith's exposition on this is so clear that I'm going to essentially paraphrase it, and I'll refer you to his textbook for more. (He covers hierarchical regression only briefly, but it is an excellent summary.)
Standard regression enters all of the predictor variables at once into the regression, and each predictor is treated as if it were entered last. That is, each predictor variable in the regression table above is controlled for all of the other predictors in the table. Thus, we can say that the effect of SES is 3.6903 when grades, esteem, and LOC are held constant or are held at their mean values. Another way to say it: with grades, esteem, and LOC removed.

The following diagram, which I have redrawn from Keith (Figure 5.7, page 83), shows what this means most clearly.

In the little regression universe that we've created for ourselves, the predictors can influence the response potentially over several pathways. For example, SES can have a direct effect on achievement scores, indicated by the arrow that goes straight to the response, but it can also have effects that follow indirect pathways. It can have an effect on locus of control, which then affects achievement scores. It can have an effect on self esteem, which can then act directly on achievement scores, but can also act through locus of control. Following the other arrows around (other than the arrows showing direct effects) will show us how WE THINK each of these predictors might have an indirect effect on achievement scores.

Standard regression reveals only the direct effects of each predictor. Because SES is treated as if it is entered last in a standard regression, all of the indirect effects of SES have already been pruned away, having been handed over to one of the other predictors, before SES gets to attempt to claim variability in the response. Thus, SES is left only with its direct pathway.

However, if SES were entered alone into a regression analysis (as the only predictor, that is), we would get to see its total effect on achievement score, which is its effect through both direct and indirect pathways, because it wouldn't be competing with the other predictors for response variability.

That's what a hierarchical regression does. If SES were entered first into a hierarchical regression, it would get to claim all the response variability it could without having to compete with the other predictors. In other words, we wouldn't be seeing just the direct effect of SES on achievement scores, we'd be seeing the total effect. Then when we enter Previous Grades second, we'd get to see its total effect, provided none of that effect goes through SES and, therefore, has already been claimed. Then we enter Self Esteem third, and we see it's total effect, provided none of that goes through either Previous Grades or SES. And so on.
In other words, standard regression reveals direct effects only. Hierarchical regression reveals total effects, PROVIDED we have entered the predictors in the correct order. And that is a BIG provided! Order of entering the predictors is the biggest catch in hierarchical regression. How do we know what order to use? Keith recommends drawing a diagram like the one above. Such a diagram should make it clear in what order the variables should be entered. If you can't draw such a diagram, then you probably shouldn't be using hierarchical regression.

Here are the hierarchical (or sequential as Keith calls it) multiple regression results (Figure 5.3, page 79) as SPSS (I assume) calculated them. The variables were entered one at a time in the order shown in the table.

![Change Statistics Table]

We already have some of these results. We need to recall the total sum of squares, 86217.92, and we need to recall the ANOVA table we produced earlier.

```r
> print(anova(lm.out), digits=7)   # print() is used to get more digits
Analysis of Variance Table
Response: achieve
   Df Sum Sq Mean Sq     F value   Pr(>F) 
SES      1 15938.64 15938.645  246.49535 < 2.22e-16 ***
grades   1 12344.62 12344.616  190.91274 < 2.22e-16 ***
estee    1   391.62   391.623    6.05655  0.0140453 *
LOC      1  512.00  512.001    7.91823  0.0050028 **
Residuals 882 57031.03    64.661
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

We can get the change in R-squared values by dividing each SS by the total.

```r
> c(15938.64, 12344.62, 391.62, 512.00) / 86217.92
[1] 0.1848646 0.1431793 0.0045422 0.0059384
```

Of course, the R Square column is just the cumulative sum of those.

```r
> cumsum(c(15938.64, 12344.62, 391.62, 512.00) / 86217.92)
[1] 0.1848646 0.3280439 0.3325861 0.3385245
```

And if you want to be really anal about it (like SPSS), the R column is the square root of that.

```r
> sqrt(cumsum(c(15938.64, 12344.62, 391.62, 512.00) / 86217.92))
[1] 0.4299588 0.5727511 0.5767028 0.5818286
```

Notice that our ANOVA table does not give the same tests, however (except for the last one). Why? The answer is that the ANOVA table tests are done with a common error term, namely, the one shown in the table. That's standard operating procedure in ANOVA, but it is not in hierarchical regression. The tests in the SPSS table can be obtained from the ANOVA for any term by pooling all the terms
below it in the table with the error term. For example, the test for grades
would be:

```r
> ((12344.62)/1) / ((391.62+512.00+57031.03)/(1+1+882))  # F value
[1] 188.3613
> pf(188.3613, 1, 884, lower=F)  # p-value
[1] 5.310257e-39
```

Here's another way to get the tests in the table above that, perhaps, makes it
a little clearer how these tests are "sequential."

```r
> model1 = lm(achieve ~ 1, data=nelscomplete)  # nothing but an intercept
> model2 = lm(achieve ~ SES, data=nelscomplete)  # adding SES to the model
> anova(model1, model2)  # compare "reduced" model to "full" model
Analysis of Variance Table

Model 1: achieve ~ 1
Model 2: achieve ~ SES

Res.Df   RSS Df Sum of Sq      F    Pr(>F)
1   886  86218
2   885  70279  1     15939 200.71 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

> model3 = lm(achieve ~ SES + grades, data=nelscomplete)  # add in grades
> anova(model2, model3)
Analysis of Variance Table

Model 1: achieve ~ SES
Model 2: achieve ~ SES + grades

Res.Df   RSS Df Sum of Sq      F    Pr(>F)
1   885  70279
2   884  57935  1     12345 188.36 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

> model4 = lm(achieve ~ SES + grades + esteem, data=nelscomplete)  # add esteem
> anova(model3, model4)
Analysis of Variance Table

Model 1: achieve ~ SES + grades
Model 2: achieve ~ SES + grades + esteem

Res.Df   RSS Df Sum of Sq      F   Pr(>F)
1   884  57935
2   883  57543  1    391.62 6.0095 0.01442 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

> model5 = lm(achieve ~ ., data=nelscomplete)  # adding LOC; or use lm.out
> anova(model4, model5)
Analysis of Variance Table

Model 1: achieve ~ SES + grades + esteem
Model 2: achieve ~ SES + grades + esteem + LOC

Res.Df   RSS Df Sum of Sq      F   Pr(>F)
1   883  57543
2   882  57031  1     512 7.9182 0.005003 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Now for a thought question. What is this? And what does it imply? (Okay, two
thought questions.)

```r
> sqrt(12344.62 / 86217.92)
[1] 0.3783904
```
Variables don't have to be entered one at a time into a hierarchical regression. They can also be entered in blocks. This might happen, for example, if the researcher wanted to enter all psychological variables at one time, or if the researcher was uncertain about the order in which they should be entered. Let's say in the current example we wanted to enter esteem and LOC at the same time. The increase in R-squared can be obtained by summing the SSes in the ANOVA table and dividing by the total SS.

\[
\frac{(391.62 + 512.00)}{86217.92} = 0.01048065
\]

Which is just the sum of the individual increases in R-squared. The test on this would be the following.

\[
\text{anova(model3, lm.out)} \quad \# \text{ or anova(model3, model5)}
\]

Analysis of Variance Table

<table>
<thead>
<tr>
<th>Res.Df</th>
<th>RSS</th>
<th>Df</th>
<th>Sum of Sq</th>
<th>F</th>
<th>Pr(&gt;F)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>884</td>
<td></td>
<td>57935</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>882</td>
<td>2</td>
<td>903.62</td>
<td>6.9874</td>
<td>0.0009754 ***</td>
</tr>
</tbody>
</table>

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

Model 3 was the model before this block of variables was entered, i.e., the model that included SES and grades, and lm.out was the model after this block of variables was entered. This is in agreement with the SPSS output that is in Figure 5.9 in Keith. So good, SPSS got it right again!
PART SEVEN: CONCLUSIONS

What is the effect of self-esteem on scholastic achievement? There are (at least) four possible competing conclusions that can be drawn from the above analyses. Make sure you use the CORRECT WORDING.

1) From the correlation coefficient:

2) From the standard multiple regression:

3) From the mediation analysis:

4) From the hierarchical regression:

How do we know which of these is the correct conclusion?