Exercise 1: Use the following code to read in the Harvard Forest Dataset HF143 Data. It is in the file datasoil.txt. It consists of soil properties along three transects in Harvard Forest, collected by Richard Bowden, Charles McClaugherty and, Timothy Sipe.

a. Look at the output. Is a decent amont of the variability explained by the first 4 factors? Use the test of hypothesis of sufficient numbers of factors to find a suitable number of factors to use. What proportion of the variability in the data is explained by the factor analysis?

Solution:
Cleaning the data, we exclude the first four columns since they are not numerical. The first column of data ff.thickness seems to have a lot of missing values. Experimenting with how we can clean up the data, I found that trying to impute the missing values ff.thickness with the mean, MLR estimation or even just removing the whole column we were able to retain a lot of the observations. Unfortunately It seems as though with 300 plus observations the factor analysis algorithm becomes unstable (I'd imagine it is problematic for the cholesky-esque factorization that is needed to compute the loadings). So I ended up just removing the first four columns and removing any observations with NAs, reducing the number of observations to 155.

Running the factor analysis with for 4 factors, with varimax rotation, we find that the data is highly variable (at least when we consider only orthogonal rotations) since the first and largest factor only explains 16.4% of the variance, with only about 44% of the variance being explained by the first four factors. The factor analysis with only 4 factors rejects the hypothesis that 4 factors are sufficient with a p-value of 4.43e-12. Interestingly a minimum of 8 factors were necessary in order to accept the null hypothesis at an $\alpha = .05$, In that case it's clear that adding more factors is explaining the variance little by little, until we've explained a majority of the variance.
Code:

```
f <- file.choose()
dat <- read.csv(f, header=TRUE)
dat <- dat[,5:23]
dat <- na.omit(dat)
out <- factanal(dat, factors=4, rotation="varimax", scores="regression")

out <- factanal(dat, factors=4, rotation="varimax", scores="regression")
Call:
factanal(x = dat, factors = 4, scores = "regression", rotation = "varimax")

Uniquenesses:
ff.thickness bulk.density soil.mass
       0.311        0.922     0.560
           c            n        om
       0.045        0.219     0.149
         no3        n.min      nitr
       0.961        0.879     0.936
```

```
                     mg                       k            soil.moisture
                  0.007                   0.703                   0.369

                 ph.h2o                  ph.cacl2
                  0.622                   0.292
                      p                     nh4
                  0.727                   0.884
                    cec                      ca
                  0.815                   0.476
soil.moisture.capacity
                  0.720
```

```
Loadings:
                       Factor1 Factor2 Factor3 Factor4
ff.thickness            0.239   0.752  -0.221  -0.133
bulk.density           -0.226          0.106
soil.mass                       0.383  -0.146  -0.519
ph.h2o                 -0.182                  0.575
ph.cacl2               -0.437                  0.713
c                       0.932   0.215          -0.197
n                       0.864   0.157
om                      0.875   0.246          -0.156
p                       0.367  -0.152   0.288   0.179
nh4                     0.176  -0.248           0.152
no3                                             0.165
n.min                  -0.102  -0.149           0.290
nitr                                            -0.229
cec                             0.354   0.185  -0.157
ca                     -0.201  -0.121   0.620   0.291
mg                              0.274   0.946  -0.138
k                              -0.113   0.533
soil.moisture           0.361   0.684  -0.151  -0.101
soil.moisture.capacity  0.214   0.479

                Factor1 Factor2 Factor3 Factor4
SS loadings       3.125   1.904   1.806   1.567
Proportion Var    0.164   0.100   0.095   0.082
Cumulative Var    0.164   0.265   0.360   0.442

Test of the hypothesis that 4 factors are sufficient.
The chi square statistic is 230.17 on 101 degrees of freedom.
The p-value is 4.43e-12
```

```
-----------------------------------------------------------------------
out <- factanal(dat, factors=8, rotation="varimax", scores="regression")
out
...........
...........
                Factor1 Factor2 Factor3 Factor4 Factor5 Factor6 Factor7
SS loadings       2.666   2.050   1.839   1.401   1.366   1.221   1.007
```

```
Proportion Var    0.140   0.108   0.097   0.074   0.072   0.064   0.053
Cumulative Var    0.140   0.248   0.345   0.419   0.491   0.555   0.608

Test of the hypothesis that 7 factors are sufficient.
The chi square statistic is 80.73 on 59 degrees of freedom.
The p-value is 0.0317
```

b. Look at the factor loading. Can you roughly interpret the loadings on the first factor? Is it always possible to do so? Why or why not?

Solution:
Looking at the magnitudes of the loadings we find that the first factor seems to heavily emphasize the c,n, and om variables. Looking at the documentation provided by the data we can see that theses variables correspond to percent carbon, percent nitrogen, and percent organic matter. Given that these are primary factors in evaluating soil health/fertility, it seems like this first factor maybe telling us this about our data. It is not always possible to interpret the loadings on any of the factors, most of the time the first factor will be the easiest to interpret, as it captures the most variance (is able to most of the signal in the data). It really depends on the signal to noise ratio that is found in the data. Data that are entirely noise will result in practically useless, uninterpretable loadings

c. Try a promax rotation, What is this and how does it differ from varimax rotation? Did the proportion of variation explained or the hypothesis test change by much? Why is this result reasonable? Why did it show factor correlations for promax and not for varimax?.

Solution:
Using the promax parameter allows for non-orthogonal factors. Generally when we do factor analysis, we want to generate the following factorization for our data X,

$$X = L'F + E$$

With varimax the columns of F are orthogonal, with promax the columns of F are allowed to be non-orthogonal. Performing the factor analysis with promax rotation we find that, allowing for non-orthogonal factors does not significantly effect the analysis with respect to the amount of variance explained, and similarly we need 8 factors in order to accept the null hypothesis at an $\alpha = .05$. Factor correlations are displayed by promax because, in this case the factors are allowed to be correlated because of the non-orthogonality.
Code:

```
out <- factanal(dat, factors=4, rotation="promax", scores="regression")

Call:
factanal(x = dat, factors = 4, scores = "regression", rotation = "promax")

Uniquenesses:
          ff.thickness           bulk.density              soil.mass
                 0.311                  0.922                  0.560
                     c                      n                     om
                 0.045                  0.219                  0.149
                   no3                 n.min                   nitr
                 0.961                  0.879                  0.936
                    mg                      k         soil.moisture
                 0.007                  0.703                  0.369

                ph.h2o                ph.cacl2
                 0.622                  0.292
                     p                    nh4
                 0.727                  0.884
                   cec                     ca
                 0.815                  0.476
 soil.moisture.capacity
                 0.720


Loadings:
                        Factor1 Factor2 Factor3 Factor4
ff.thickness             0.115   0.865   0.120
bulk.density            -0.202
soil.mass               -0.187   0.319  -0.439
ph.h2o                           0.282   0.706
ph.cacl2                -0.311   0.161   0.827
c                        0.877   0.119  -0.229
n                        0.862   0.133
om                       0.822   0.167  -0.168
p                        0.419  -0.199          0.180
nh4                      0.238  -0.260         -0.130
no3                      0.131           0.164
n.min                                    0.283
nitr                                    -0.225
cec                              0.344          0.302
ca                      -0.142  -0.127   0.267   0.566
mg                               0.140  -0.124   1.010
k                               -0.217          0.480
soil.moisture            0.252   0.774   0.116
soil.moisture.capacity   0.136   0.530          0.131

                Factor1 Factor2 Factor3 Factor4
SS loadings       2.747   2.220   1.771   1.755
Proportion Var    0.145   0.117   0.093   0.092
Cumulative Var    0.145   0.261   0.355   0.447

Factor Correlations:
        Factor1 Factor2 Factor3 Factor4
```

```
Factor1    1.000    0.119    0.191   -0.108
Factor2    0.119    1.000   -0.184    0.133
Factor3    0.191   -0.184    1.000   -0.564
Factor4   -0.108    0.133   -0.564    1.000

Test of the hypothesis that 4 factors are sufficient.
The chi square statistic is 230.17 on 101 degrees of freedom.
The p-value is 4.43e-12
>
```

d. What does the plot of scores tell you? What are scores?

Solution:
Scores of an observations we can see how much of each factor influences it. Considering the score plots, we can see any trends among the factors as well as to see if any outliers have affected our factor analysis. With both factor analysis' there do not appear to be extreme outliers which define each factor. There seems to be a grouping of observations with high 1 factor scores, however the factor does not seem to be defined by outliers, there is a good spread of observations across both factors.

Figure 1: Varimax Score Plot

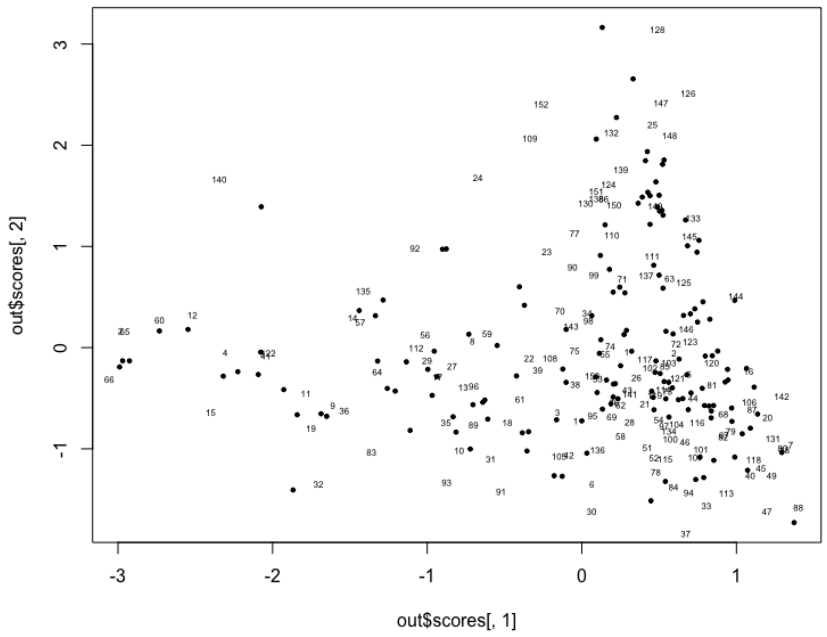
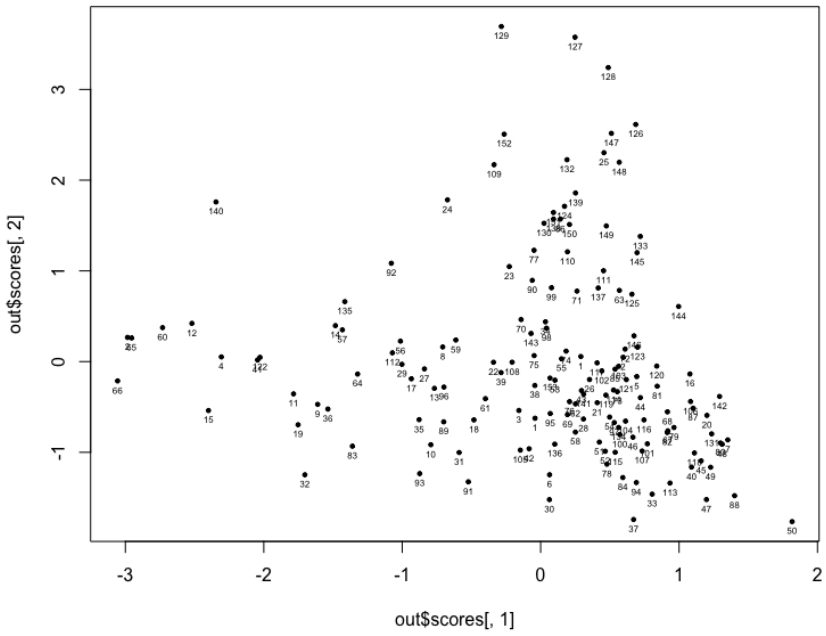


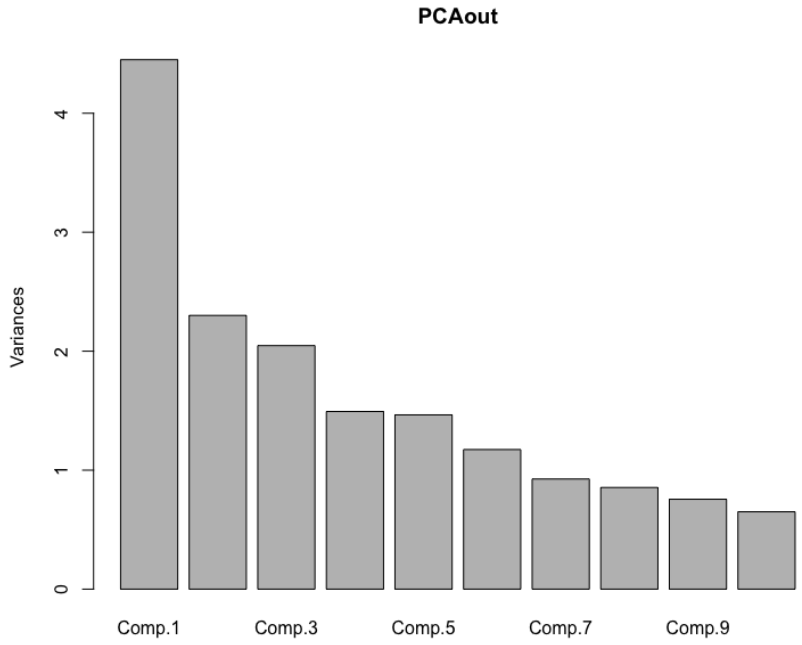
Figure 2: Promax Score Plot



Exercise 2: Using the same dataset as in problem one, run a principal components analysis. Use a screeplot to select the number of important PCs. Do the number of PCs seem to match what you got with the factor analysis. Do the loadings look similar ot those from the factors you calculated in b and c? Why do you think this happened?

Solution:
Performing the principal component analysis we get the following screeplot. Considering the 'keep PCs' which contribute to 10 percent or more of the variance explained we see that PCA gives us around 8 or 9 components. This agrees with our factor analysis. Considering the loadings, from the PCA we do see some similarity among the first factor, with respect to the loadings from the factor analysis.

Figure 3: PCA Scree Plot



Code:

```
> summary(PCAout)
Importance of components:
                          Comp.1     Comp.2     Comp.3     Comp.4     Comp.5
Standard deviation     2.109532  1.5167763  1.4306648  1.22197267  1.21023194
Proportion of Variance 0.234217  0.1210848  0.1077264  0.07859038  0.07708744
Cumulative Proportion  0.234217  0.3553018  0.4630282  0.54161858  0.61870601

                           Comp.6     Comp.7     Comp.8     Comp.9    Comp.10
Standard deviation     1.08320125 0.96211091 0.92438319 0.86923383 0.80643962
Proportion of Variance 0.06175394 0.04871881 0.04497286 0.03976671 0.03422868
Cumulative Proportion  0.68045996 0.72917877 0.77415163 0.81391834 0.84814701

                          Comp.11    Comp.12    Comp.13    Comp.14    Comp.15
Standard deviation     0.75520074 0.68643519 0.6613945 0.62964684 0.57675458
Proportion of Variance 0.03001727 0.02479965 0.0230233 0.02086606 0.01750768
Cumulative Proportion  0.87816428 0.90296393 0.9259872 0.94685329 0.96436097

                          Comp.16    Comp.17     Comp.18     Comp.19
Standard deviation     0.49299528 0.44332640 0.403017189 0.274109644
Proportion of Variance 0.01279181 0.01034412 0.008548571 0.003954531
Cumulative Proportion  0.97715278 0.98749690 0.996045469 1.000000000
```

```
PCAout$loadings
Loadings:
                        Comp.1 Comp.2 Comp.3 Comp.4 ... Comp.19
ff.thickness             0.318  0.212         0.137
bulk.density            -0.145         0.195  0.345
soil.mass                0.186  0.297  0.313 -0.167
ph.h2o                  -0.168        -0.188  0.240
ph.cacl2                -0.320                0.224
c                        0.415 -0.174 -0.135
n                        0.333 -0.275 -0.190  0.146
om                       0.405 -0.207 -0.127
p                             -0.430         0.268
nh4                           -0.168 -0.314 -0.306
no3                                  -0.406 -0.177
n.min                   -0.151 -0.127         0.512
nitr                          0.160  0.355  0.168
cec                      0.140         0.200 -0.169
ca                      -0.195 -0.364  0.219
mg                            -0.389  0.375 -0.107
k                             -0.385  0.337 -0.163
soil.moisture            0.339  0.138         0.132
soil.moisture.capacity   0.231         0.116  0.356
```

We see that the first component of the PCA again favors variables c, n, and om similarly to the factor analysis loadings. The rest of the loadings do not look particularly similar. This makes sense that they are not exactly the same. PCA generates loadings by solving the eigenvalue eigenvector problem on the correlation matrix, factor analysis generates loadings via an algorithm which minimizes the error in the following factorization,

$$\Sigma_x = LL' + D.$$

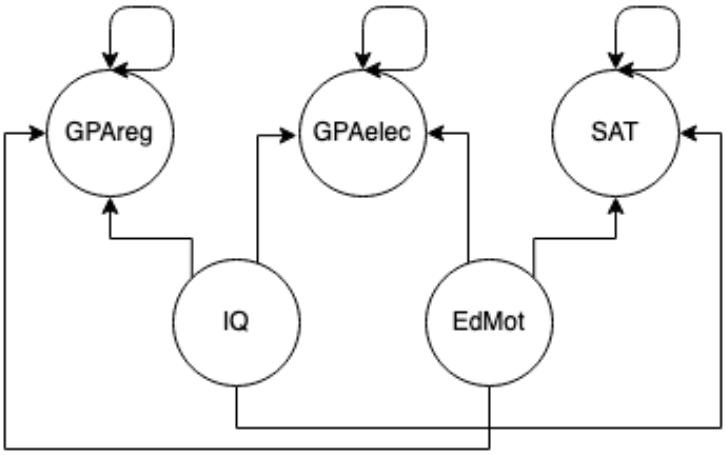
Exercise 3: Below is the covariance matrix based on $N = 150$ first year college students.

```
M <-
structure(c(0.594, 0.483, 3.993, 0.426, 0.5, 0.483, 0.754, 3.626,
1.757, 0.722, 3.993, 3.626, 47.457, 4.1, 6.394, 0.426, 1.757,
4.1, 10.267, 0.525, 0.5, 0.722, 6.394, 0.525, 2.675), .Dim = c(5L,
5L), .Dimnames = list(c("GPAreq", "GPAelec", "SAT", "IQ", "EdMot"
), c("GPAreq", "GPAelec", "SAT", "IQ", "EdMot")))
```

a. Try to make up two or three reasonable path analysis models for this data. Draw a structural diagram for each. Also, list all of the parameters of each model.

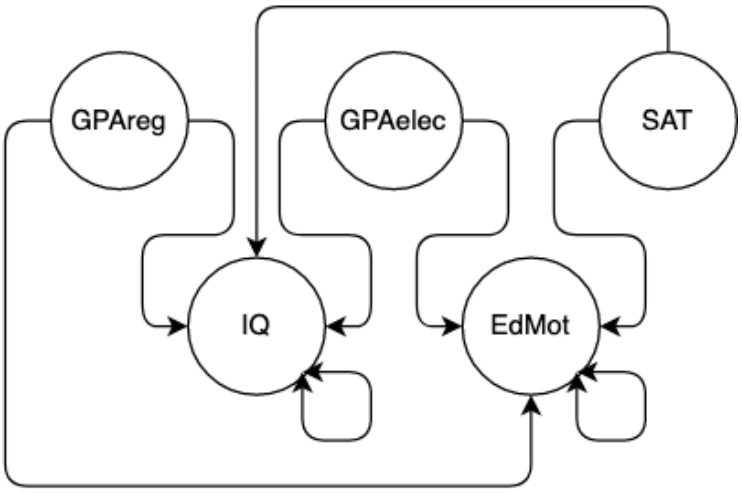
Solution:
Here is my first attempt at a structural diagram. It seems as though IQ and EdMot are exogenous variables, they seem to be outside variables used for predicting student success in metrics like GPA, and SAT scores. In this case we have six regression parameters, three error variances, and three covariances. This is the following structural diagram,

Figure 4: First Structural Diagram



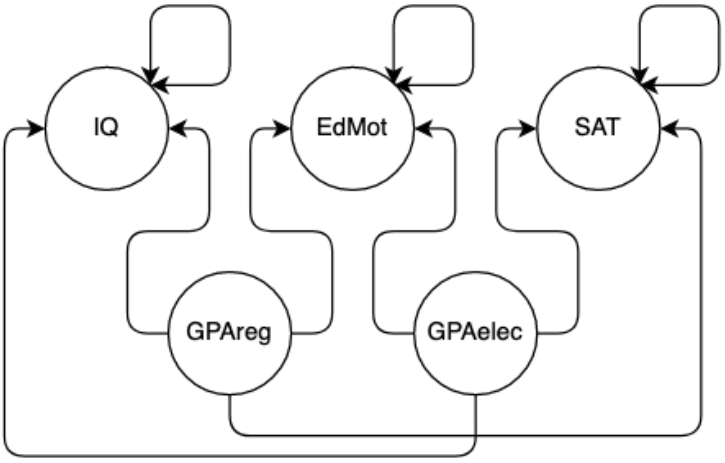
My second structural diagram uses school performance metrics like GPA and SAT scores to predict IQ and EdMot, sort of the inverse of the previous diagram. Here we assume that GPA, and SAT Scores are exogenous and IQ and EdMot are endogenous(seems unlikely but we'll test it anyway). This structural diagram has six regression parameters, two error variances, and one covariance.

Figure 5: Second Structural Diagram



My third structural diagram uses GPA to predict IQ, EdMot, and SAT Scores. In this model we are treating GPA as an exogenous variable and the rest as endogenous. This model continues six regression parameters and three error variances, and three covariances.

Figure 6: Third Structural Diagram



b, Using lavaan, run each model. Which model seems to fit the data best? how good is the fit of the model?

Solution:
Calling AIC on the fitted models we find that the second structure diagram has the

best fit. Interestingly this is the model with the least degrees of freedom/parameters. All models had Standardized Root Mean Square Residual of 0 and CFI and TLI goodness of fit values of 1.

Code:

```
> myModel1 <- '
+   #regressions
+   GPAreq ~ IQ + EdMot
+   GPAelec ~ IQ + EdMot
+   SAT ~ IQ + EdMot
+ '
> fit1 <- sem(myModel1, sample.cov = M, sample.nobs = 150)
> summary(fit1, fit.measures=TRUE)
lavaan 0.6-10 ended normally after 40 iterations

  Estimator                                         ML
  Optimization method                           NLMINB
  Number of model parameters                        12

  Number of observations                           150

Model Test User Model:

  Test statistic                                 0.000
  Degrees of freedom                                 0

Model Test Baseline Model:

  Test statistic                               461.731
  Degrees of freedom                                 9
  P-value                                        0.000

User Model versus Baseline Model:

  Comparative Fit Index (CFI)                    1.000
  Tucker-Lewis Index (TLI)                       1.000

Loglikelihood and Information Criteria:

  Loglikelihood user model (H0)               -635.396
  Loglikelihood unrestricted model (H1)       -635.396

  Akaike (AIC)                                1294.791
  Bayesian (BIC)                              1330.919
  Sample-size adjusted Bayesian (BIC)         1292.941

Root Mean Square Error of Approximation:

  RMSEA                                          0.000
  90 Percent confidence interval - lower         0.000
  90 Percent confidence interval - upper         0.000
  P-value RMSEA <= 0.05                             NA
```

```
Standardized Root Mean Square Residual:

  SRMR                                            0.000

Parameter Estimates:

  Standard errors                             Standard
  Information                                 Expected
  Information saturated (h1) model          Structured

Regressions:
                   Estimate  Std.Err  z-value  P(>|z|)
  GPAreq ~
    IQ                0.032    0.018    1.799    0.072
    EdMot             0.181    0.035    5.142    0.000
  GPAelec ~
    IQ                0.159    0.014   11.284    0.000
    EdMot             0.239    0.028    8.652    0.000
  SAT ~
    IQ                0.280    0.143    1.951    0.051
    EdMot             2.335    0.281    8.309    0.000

Covariances:
                   Estimate  Std.Err  z-value  P(>|z|)
 .GPAreq ~~
   .GPAelec           0.294    0.039    7.465    0.000
   .SAT               2.688    0.386    6.957    0.000
 .GPAelec ~~
   .SAT               1.438    0.276    5.210    0.000

Variances:
                   Estimate  Std.Err  z-value  P(>|z|)
   .GPAreq            0.487    0.056    8.660    0.000
   .GPAelec           0.300    0.035    8.660    0.000
   .SAT              31.168    3.599    8.660    0.000

-----------------------------------------------------------------

> myModel2 <- '
+   IQ ~ GPAreq + GPAelec + SAT
+   EdMot ~ GPAreq + GPAelec + SAT
+ '
> fit2 <- sem(myModel2, sample.cov = M, sample.nobs = 150)
> summary(fit2, fit.measures=TRUE)
lavaan 0.6-10 ended normally after 17 iterations

  Estimator                                         ML
  Optimization method                           NLMINB
  Number of model parameters                         9

  Number of observations                           150

Model Test User Model:
```

```
  Test statistic                                0.000
  Degrees of freedom                                0

Model Test Baseline Model:

  Test statistic                              224.880
  Degrees of freedom                                7
  P-value                                       0.000

User Model versus Baseline Model:

  Comparative Fit Index (CFI)                   1.000
  Tucker-Lewis Index (TLI)                      1.000

Loglikelihood and Information Criteria:

  Loglikelihood user model (H0)              -560.704
  Loglikelihood unrestricted model (H1)      -560.704

  Akaike (AIC)                               1139.409
  Bayesian (BIC)                             1166.504
  Sample-size adjusted Bayesian (BIC)        1138.021

Root Mean Square Error of Approximation:

  RMSEA                                         0.000
  90 Percent confidence interval - lower        0.000
  90 Percent confidence interval - upper        0.000
  P-value RMSEA <= 0.05                            NA

Standardized Root Mean Square Residual:

  SRMR                                          0.000

Parameter Estimates:

  Standard errors                             Standard
  Information                                 Expected
  Information saturated (h1) model          Structured

Regressions:
                   Estimate  Std.Err  z-value  P(>|z|)
  IQ ~
    GPAreq           -2.387    0.393   -6.067    0.000
    GPAelec           3.917    0.289   13.537    0.000
    SAT              -0.012    0.038   -0.315    0.753
  EdMot ~
    GPAreq           -0.653    0.237   -2.751    0.006
    GPAelec           0.733    0.174    4.201    0.000
    SAT               0.134    0.023    5.786    0.000

Covariances:
                   Estimate  Std.Err  z-value  P(>|z|)
 .IQ ~~
```

```
    .EdMot            -1.026    0.233   -4.399    0.000

Variances:
                   Estimate  Std.Err  z-value  P(>|z|)
    .IQ               4.421    0.510    8.660    0.000
    .EdMot            1.607    0.186    8.660    0.000
```

```
> myModel3 <- '
+   IQ ~ GPAreq + GPAelec
+   EdMot ~ GPAreq + GPAelec
+   SAT ~ GPAreq + GPAelec
+ '
> fit3 <- sem(myModel3, sample.cov = M, sample.nobs = 150)
> summary(fit3, fit.measures=TRUE)
lavaan 0.6-10 ended normally after 26 iterations

  Estimator                                         ML
  Optimization method                           NLMINB
  Number of model parameters                        12

  Number of observations                           150

Model Test User Model:

  Test statistic                                 0.000
  Degrees of freedom                                 0

Model Test Baseline Model:

  Test statistic                               352.874
  Degrees of freedom                                 9
  P-value                                        0.000

User Model versus Baseline Model:

  Comparative Fit Index (CFI)                    1.000
  Tucker-Lewis Index (TLI)                       1.000

Loglikelihood and Information Criteria:

  Loglikelihood user model (H0)               -998.534
  Loglikelihood unrestricted model (H1)       -998.534

  Akaike (AIC)                               2021.067
  Bayesian (BIC)                             2057.195
  Sample-size adjusted Bayesian (BIC)        2019.217

Root Mean Square Error of Approximation:

  RMSEA                                          0.000
  90 Percent confidence interval - lower         0.000
```

```
  90 Percent confidence interval - upper         0.000
  P-value RMSEA <= 0.05                             NA

Standardized Root Mean Square Residual:

  SRMR                                           0.000

Parameter Estimates:

  Standard errors                             Standard
  Information                                 Expected
  Information saturated (h1) model          Structured

Regressions:
                   Estimate  Std.Err  z-value  P(>|z|)
  IQ ~
    GPAreq           -2.458    0.323   -7.610    0.000
    GPAelec           3.905    0.287   13.621    0.000
  EdMot ~
    GPAreq            0.132    0.215    0.612    0.541
    GPAelec           0.873    0.191    4.569    0.000
  SAT ~
    GPAreq            5.869    0.688    8.528    0.000
    GPAelec           1.050    0.611    1.718    0.086

Covariances:
                   Estimate  Std.Err  z-value  P(>|z|)
 .IQ ~~
   .EdMot            -1.058    0.256   -4.137    0.000
   .SAT              -0.243    0.770   -0.315    0.753
 .EdMot ~~
   .SAT               2.684    0.558    4.811    0.000

Variances:
                   Estimate  Std.Err  z-value  P(>|z|)
   .IQ                4.424    0.511    8.660    0.000
   .EdMot             1.966    0.227    8.660    0.000
   .SAT              20.082    2.319    8.660    0.000
-----------------------------------------------------------------
> AIC(fit1, fit2, fit3)
     df       AIC
fit1 12 1294.791
fit2  9 1139.409
fit3 12 2021.067
```

c. In your best fitting model, are any of the links candidates for removal? How could you determine this?

Solution:
From the summary report we can see that the link between SAT and IQ can likely be

a candidate for removal. The MLR which produces IQ finds that the SAT variable has a p-value of 0.753, considerably higher than the rest and insignificant at the $\alpha = .05$ level.

d. Why can't we use bootsrapping to get standard errors in this situation?

Solution:
In this situation we don't have the data, only the covariance/correlation matrix.

e. Is the model you selected at best recursive, or non-recursive? How can you tell?

Solution:
All of the models I selected are non-recursive. We setup up multiple regressions, and none of the results of those regressions feed into other regressions causing a loop or indirect effects.

f. Are there any indirect effects in your best model? What are they?

Solution:
My model has no indirect effects, IQ is directly modeled with GPA, and SAT and similarly with EdMot.

g, What variables in your best model are exogenous, and which are endogenous?

Solution:
Like previously stated the second model has GPA and SAT scores as exogenous variables and IQ and EdMot are endogenous.