HW4

Ran Zhang

2/13/2019

Store HW4 working environment

```
set.seed(02112019)
```

Problem 1a

Read the data

```
data1 <- read.table("hemangioma.txt",header = TRUE)
```

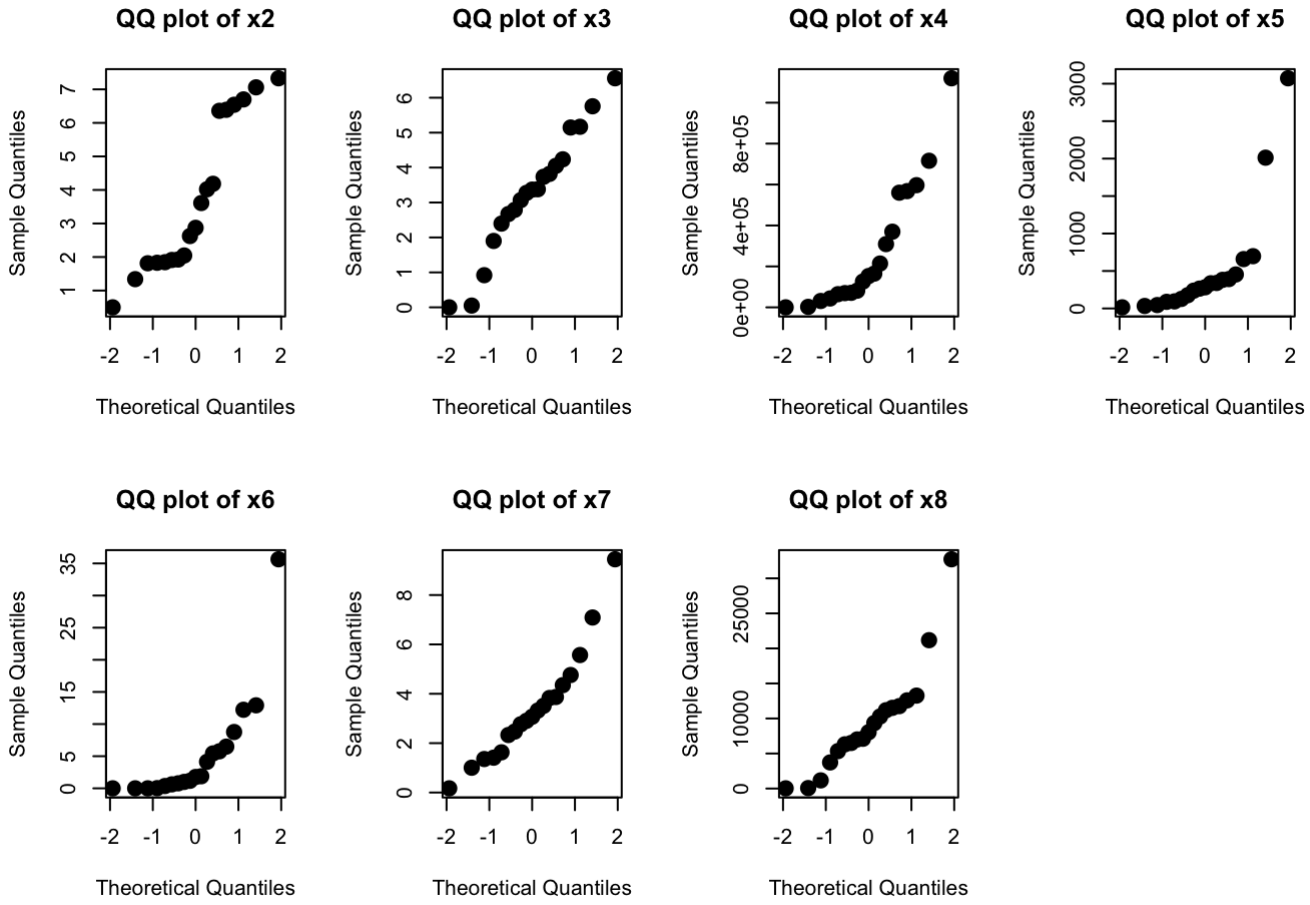
Build the dataset

```
data1 <- matrix(unlist(data1), ncol=8, byrow=FALSE)
colnames(data1) <- c("Age","RB","p16","DLK","Nanog","C.Myc","EZH2","IGF.2")
```

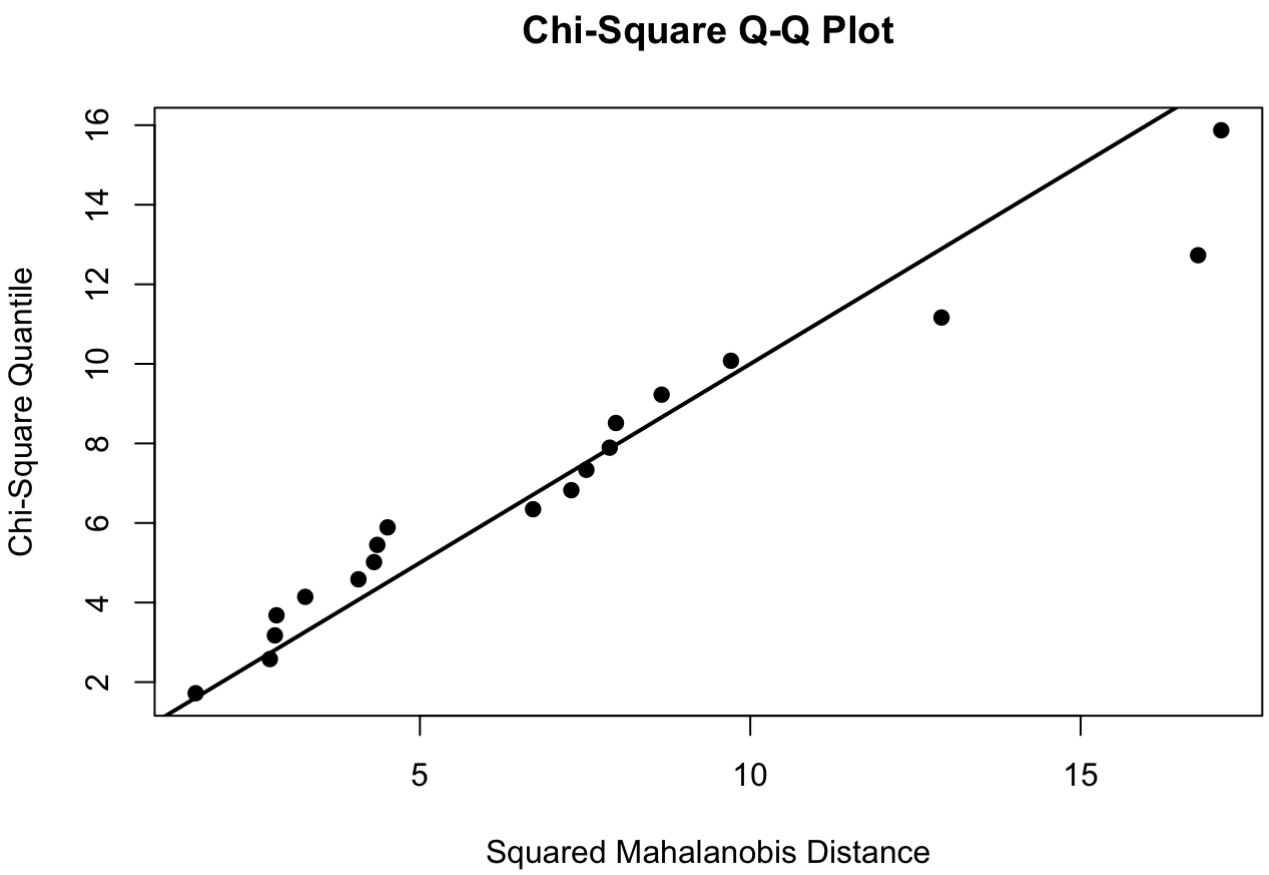
Check for the normality

```
par(mfrow=c(2,4))
for (ii in 2:8){
  qqnorm(data1[,ii],
         main=paste0("QQ plot of x", ii), pch=19, cex=1.5)
}
par(mfrow=c(1,1))
library(MVN)
```

```
## sROC 0.1-2 loaded
```




```
mvn(data1[,2:8], mvnTest = "royston", multivariatePlot = "qq")
```


Chi-Square Q-Q Plot
Chi-Square Quantile
2
4
6
8
10
12
14
16
5
10
15
Squared Mahalanobis Distance

```
## $multivariateNormality
##      Test       H      p value MVN
## 1 Royston 68.8651 2.755476e-12  NO
##
## $univariateNormality
##           Test  Variable Statistic   p value Normality
## 1 Shapiro-Wilk    RB        0.8723  0.0158      NO
## 2 Shapiro-Wilk    p16       0.9692  0.7601      YES
## 3 Shapiro-Wilk    DLK       0.8235  0.0026      NO
## 4 Shapiro-Wilk   Nanog      0.6049  <0.001      NO
## 5 Shapiro-Wilk   C.Myc      0.6365  <0.001      NO
## 6 Shapiro-Wilk   EZH2       0.9223  0.1246      YES
## 7 Shapiro-Wilk   IGF.2      0.8556  0.0083      NO
##
## $Descriptives
##         n          Mean      Std.Dev       Median   Min          Max
## RB     19 3.731693e+00 2.277890e+00      2.87076  0.50 7.330000e+00
## p16    19 3.278360e+00 1.764494e+00      3.37000  0.00 6.556375e+00
## DLK    19 2.768669e+05 3.045860e+05 153060.60000 94.47 1.119258e+06
## Nanog  19 5.109990e+02 7.621973e+02    282.29727 14.96 3.072500e+03
## C.Myc  19 5.209137e+00 8.433340e+00      1.77000  0.00 3.564539e+01
## EZH2   19 3.413648e+00 2.214785e+00      3.07000  0.17 9.448814e+00
## IGF.2  19 9.427143e+03 7.615040e+03   8038.53000 29.93 3.272181e+04
##                25th          75th       Skew   Kurtosis
## RB         1.878818 6.375000e+00  0.3653069 -1.5520357
## p16        2.533415 4.143347e+00 -0.1950667 -0.6171533
## DLK    66929.905000 4.649044e+05  1.2153334  0.6156620
## Nanog    110.739865 4.229038e+02  2.3256168  4.5486159
## C.Myc      0.490000 6.124258e+00  2.4540173  6.0464772
## EZH2       1.979461 4.110000e+00  1.0167412  0.7754154
## IGF.2   5825.205000 1.164033e+04  1.4093342  2.2524263
```

Comment: As the results from Shapiro-Wilk test, the p16, EZH2 marginal distributions are normally distributed.

Calculate the Mahalanobis distribution

```
s1 <- cov(data1[,2:8])
x1.cen <- scale(data1[,2:8], center=T, scale=F)
d2 <- diag(x1.cen%*%solve(s1)%*%t(x1.cen))
d2.matrix <- matrix(d2,ncol=1)
colnames(d2.matrix) <- c("d2")
data2 <- cbind(data1,d2.matrix)
data2
```

```
##        Age       RB      p16        DLK      Nanog     C.Myc     EZH2
##  [1,]   81 2.046149 3.067127  308974.72   94.17336  6.489601 2.764101
##  [2,]   95 6.540000 1.900000   70988.30  381.83000  1.000000 7.090000
##  [3,]   95 3.610000 3.820000  153060.60  237.28000  0.000000 5.570000
##  [4,]  165 1.912267 3.735868  596991.60   88.23737  0.000000 2.469633
##  [5,]  286 2.625436 5.168293  369600.56  282.29727 12.225828 1.628923
##  [6,]  299 2.870760 5.755246 1119257.50  176.75143  8.764235 3.511469
##  [7,]  380 1.925978 2.396830  214070.92   45.26692  5.758915 1.411180
##  [8,]  418 7.060000 3.380000   69511.45  264.62000  1.170000 3.070000
##  [9,]  420 6.390000 3.370000   81457.12  658.75000  1.880000 3.870000
## [10,]  547 6.360000 4.050000   64348.36  336.11000  0.780000 4.760000
## [11,]  590 1.813758 5.147162  164881.06 2012.47920 35.645386 9.448814
## [12,]  635 6.700000 2.670000  126015.95 3072.50000  0.000000 4.350000
## [13,]  752 1.845369 3.275246  567857.56  127.30637  4.129052 1.004505
## [14,]  760 7.330000 0.920000   43438.04  697.57000  1.770000 3.320000
## [15,] 1171 1.828868 6.556375  716259.50  392.08760 12.917779 2.904105
## [16,] 1277 1.340000 0.050000      94.47   14.96000  0.360000 3.830000
## [17,] 1520 4.020000 2.790000   31124.69  453.72000  0.620000 2.330000
## [18,] 2138 0.500000 0.000000    2330.59   33.11000  0.030000 0.170000
## [19,] 3626 4.183583 4.236695  560208.30  339.93090  5.432815 1.356589
##           IGF.2        d2
##  [1,] 11175.689  1.521731
##  [2,]  5340.170  7.547304
##  [3,]  6310.240  7.457835
##  [4,]  7008.523  3.855230
##  [5,]  7104.238  7.123146
##  [6,]  9342.126 12.215844
##  [7,]  3725.515  2.657346
##  [8,]  8038.530  4.125148
##  [9,] 12583.250  2.680522
## [10,]  6505.150  4.079116
## [11,] 32721.809 15.894008
## [12,] 11762.760 16.225957
## [13,] 10283.141  2.585439
## [14,] 11517.890  8.202557
## [15,] 13263.792  4.273072
## [16,]    29.930  6.358907
## [17,]  1162.690  3.092665
## [18,]    66.490  6.907369
## [19,] 21173.781  9.196805
```

Explanation: As we the combination of Mahalanobis distance and the chi-square plot, observation 11 and 12 which has the biiggest distances are outliers.

Problem 1b

Perform EFA on orginal data

```
par(mfrow=c(2,3))
fa.orginal1 <- factanal(x=data1[,2:8], factors = 1)
fa.orginal1
```

```
##
## Call:
## factanal(x = data1[, 2:8], factors = 1)
##
## Uniquenesses:
##    RB   p16   DLK Nanog C.Myc  EZH2 IGF.2
## 0.999 0.661 0.931 0.748 0.303 0.738 0.139
##
## Loadings:
##       Factor1
## RB
## p16    0.582
## DLK    0.263
## Nanog  0.502
## C.Myc  0.835
## EZH2   0.512
## IGF.2  0.928
##
##                Factor1
## SS loadings      2.482
## Proportion Var   0.355
##
## Test of the hypothesis that 1 factor is sufficient.
## The chi square statistic is 32.05 on 14 degrees of freedom.
## The p-value is 0.00395
```

```
fa.orginal2 <- factanal(x=data1[,2:8], factors = 2)
fa.orginal2
```

```
##
## Call:
## factanal(x = data1[, 2:8], factors = 2)
##
## Uniquenesses:
##    RB   p16   DLK Nanog C.Myc  EZH2 IGF.2
## 0.846 0.335 0.060 0.578 0.316 0.527 0.166
##
## Loadings:
##       Factor1 Factor2
## RB     0.107  -0.377
## p16    0.427   0.695
## DLK            0.970
## Nanog  0.615  -0.209
## C.Myc  0.780   0.273
## EZH2   0.647  -0.234
## IGF.2  0.878   0.252
##
##                Factor1 Factor2
## SS loadings      2.370   1.802
## Proportion Var   0.339   0.257
## Cumulative Var   0.339   0.596
##
## Test of the hypothesis that 2 factors are sufficient.
## The chi square statistic is 14.92 on 8 degrees of freedom.
## The p-value is 0.0607
```

```
fa.orginal3 <- factanal(x=data1[,2:8], factors = 3)
fa.orginal3
```

```
##
## Call:
## factanal(x = data1[, 2:8], factors = 3)
##
## Uniquenesses:
##    RB   p16   DLK Nanog C.Myc  EZH2 IGF.2
## 0.050 0.293 0.005 0.609 0.005 0.490 0.249
##
## Loadings:
##       Factor1 Factor2 Factor3
## RB     0.141  -0.144   0.954
## p16    0.366   0.757
## DLK   -0.163   0.961  -0.211
## Nanog  0.559           0.275
## C.Myc  0.841   0.295  -0.448
## EZH2   0.682           0.193
## IGF.2  0.780   0.377
##
##                Factor1 Factor2 Factor3
## SS loadings      2.274   1.757   1.269
## Proportion Var   0.325   0.251   0.181
## Cumulative Var   0.325   0.576   0.757
##
## Test of the hypothesis that 3 factors are sufficient.
## The chi square statistic is 1.86 on 3 degrees of freedom.
## The p-value is 0.603
```

Perfom EFA on removed outliers's data

```
fa.remove1 <- factanal(x=data1[-c(11,12),2:8 ], factors = 1)
fa.remove1
```

```
##
## Call:
## factanal(x = data1[-c(11, 12), 2:8], factors = 1)
##
## Uniquenesses:
##    RB   p16   DLK Nanog C.Myc  EZH2 IGF.2
## 0.939 0.291 0.265 0.999 0.369 0.961 0.681
##
## Loadings:
##       Factor1
## RB    -0.247
## p16    0.842
## DLK    0.857
## Nanog
## C.Myc  0.794
## EZH2  -0.197
## IGF.2  0.565
##
##                Factor1
## SS loadings      2.494
## Proportion Var   0.356
##
## Test of the hypothesis that 1 factor is sufficient.
## The chi square statistic is 29.55 on 14 degrees of freedom.
## The p-value is 0.0088
```

```
fa.remove2 <- factanal(x=data1[-c(11,12),2:8 ], factors = 2)
fa.remove2
```

```
##
## Call:
## factanal(x = data1[-c(11, 12), 2:8], factors = 2)
##
## Uniquenesses:
##    RB   p16   DLK Nanog C.Myc  EZH2 IGF.2
## 0.005 0.205 0.237 0.373 0.349 0.683 0.472
##
## Loadings:
##       Factor1 Factor2
## RB    -0.103   0.992
## p16    0.888
## DLK    0.823  -0.291
## Nanog  0.110   0.784
## C.Myc  0.769  -0.243
## EZH2  -0.111   0.552
## IGF.2  0.642   0.339
##
##                Factor1 Factor2
## SS loadings      2.506   2.170
## Proportion Var   0.358   0.310
## Cumulative Var   0.358   0.668
##
## Test of the hypothesis that 2 factors are sufficient.
## The chi square statistic is 5.4 on 8 degrees of freedom.
## The p-value is 0.714
```

```
par(mfrow=c(1,1))
```

Conclusion: We got the different conclusions about how many factors we should use after removing outliers, as the original data we need 3 factors but we only need 2 factors after removing outliers. Because as the fa.remove2 shows we fail to reject the hypothesis test as the p-value is far bigger than 0.05.

Problem 2a

Build the correlation matrix

```
library(sem)
lt <- readMoments("EverittEx5.5.txt", diag = T)
R <- (lt + t(lt)) - diag(1, 6)
colnames(R) <- c("French", "English", "History", "Arithmetic", "Algebra", "Geometry")
rownames(R) <- c("French", "English", "History", "Arithmetic", "Algebra", "Geometry")
R
```

```
##            French English History Arithmetic Algebra Geometry
## French       1.00    0.44    0.41       0.29    0.33     0.25
## English      0.44    1.00    0.35       0.35    0.32     0.33
## History      0.41    0.35    1.00       0.16    0.19     0.18
## Arithmetic   0.29    0.35    0.16       1.00    0.59     0.47
## Algebra      0.33    0.32    0.19       0.59    1.00     0.46
## Geometry     0.25    0.33    0.18       0.47    0.46     1.00
```

Perform the test with k=2 from MLE w/t rotation

```
library(psych)
n_2a <- nrow(R)
fa.mle_2a <- fa(r=R,n.obs=n_2a,nfactors= 2, rotate = "none",fm="ml")
print(fa.mle_2a$loadings,digits = 2,cutoff = 0.3)
```

```
##
## Loadings:
##            ML1   ML2
## French      0.56  0.42
## English     0.57
## History     0.39  0.45
## Arithmetic  0.74
## Algebra     0.72
## Geometry    0.59
##
##                  ML1  ML2
## SS loadings    2.20 0.60
## Proportion Var 0.37 0.10
## Cumulative Var 0.37 0.47
```

Interpreatation：The loadings here are difficult to interpret as all variables have impacts to the first factor.

Problem 2b

Perform the test with k=2 from MLE with rotation

```
n_2b <- nrow(R)
fa.mle_2b <- fa(r=R,n.obs=n_2b,nfactors= 2, rotate = "varimax",fm="ml")
print(fa.mle_2b$loadings,digits = 2,cutoff = 0.3)
```

```
##
## Loadings:
##            ML1  ML2
## French          0.66
## English    0.32 0.55
## History         0.59
## Arithmetic 0.77
## Algebra    0.72
## Geometry   0.57
##
##                  ML1  ML2
## SS loadings    1.59 1.21
## Proportion Var 0.27 0.20
## Cumulative Var 0.27 0.47
```

Interpretation: After roration, we can easily found that last three variables contribute to the first factor, and first three variables mainly contribut to the second factor.

Problem 2c

Probability of the Emprical Chi Square given the hypothesis

```
fa.mle_2a$EPVAL
```

```
## [1] 0.9998481
```

```
fa.mle_2b$EPVAL
```

```
## [1] 0.9998481
```

As the p-value of chi-square test is much bigger than 0.05, so we fail to reject the hypothesis test, so this two factor model is sufficient enough.

Problem 3a Read the data

```
library(sem)
lt1 <- readMoments("EverittEx7.1.txt", diag = T)
R1 <- (lt1 + t(lt1)) - diag(1, 9)
R2 <- R1[-9, -9]
R2
```

```
##       X1    X2    X3    X4    X5    X6    X7    X8
## X1  1.00 -0.04  0.61  0.45  0.03 -0.29 -0.30  0.45
## X2 -0.04  1.00 -0.07 -0.12  0.49  0.43  0.30 -0.31
## X3  0.61 -0.07  1.00  0.59  0.03 -0.13 -0.24  0.59
## X4  0.45 -0.12  0.59  1.00 -0.08 -0.21 -0.19  0.63
## X5  0.03  0.49  0.03 -0.08  1.00  0.47  0.41 -0.14
## X6 -0.29  0.43 -0.13 -0.21  0.47  1.00  0.63 -0.13
## X7 -0.30  0.30 -0.24 -0.19  0.41  0.63  1.00 -0.26
## X8  0.45 -0.31  0.59  0.63 -0.14 -0.13 -0.26  1.00
```

Input the model in R

```
ability_model <- specifyModel(file="hw4.txt")
```

```
## NOTE: it is generally simpler to use specifyEquations() or cfa()
##       see ?specifyEquations
```

```
ability_sem <- sem::sem(model=ability_model,S=R2,N=123)
par(mfrow=c(2,1))
summary(ability_sem)
```

```
##
##  Model Chisquare =  63.2304   Df =  19 Pr(>Chisq) = 1.180358e-06
##  AIC =  97.2304
##  BIC =  -28.20111
##
##  Normalized Residuals
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## -2.0597 -0.3085 -0.0200  0.0122  0.6249  1.9168
##
##  R-square for Endogenous Variables
##     X1     X3     X4     X8     X2     X5     X6     X7
## 0.4449 0.6502 0.5695 0.5720 0.2804 0.3494 0.6927 0.5350
##
##  Parameter Estimates
##          Estimate   Std Error  z value   Pr(>|z|)
## lambda11  0.6670173 0.08657199  7.704771 1.310781e-14 X1 <--- DR
## lambda31  0.8063408 0.08164213  9.876528 5.262474e-23 X3 <--- DR
## lambda41  0.7546241 0.08341040  9.047122 1.467856e-19 X4 <--- DR
## lambda81  0.7562965 0.08335235  9.073487 1.152677e-19 X8 <--- DR
## lambda22  0.5295387 0.09314848  5.684888 1.308985e-08 X2 <--- PR
## lambda52  0.5911277 0.09149425  6.460818 1.041384e-10 X5 <--- PR
## lambda62  0.8323020 0.08700206  9.566463 1.106251e-21 X6 <--- PR
## lambda72  0.7314687 0.08859275  8.256531 1.499678e-16 X7 <--- PR
## psi1      0.5550877 0.08394571  6.612461 3.779834e-11 X1 <--> X1
## psi2      0.7195887 0.10156760  7.084826 1.392193e-12 X2 <--> X2
## psi3      0.3498144 0.07038309  4.970148 6.690186e-07 X3 <--> X3
## psi4      0.4305423 0.07437535  5.788777 7.090077e-09 X4 <--> X4
## psi5      0.6505678 0.09583263  6.788584 1.132394e-11 X5 <--> X5
## psi6      0.3072733 0.08761380  3.507133 4.529624e-04 X6 <--> X6
## psi7      0.4649534 0.08655697  5.371646 7.802124e-08 X7 <--> X7
## psi8      0.4280155 0.07422086  5.766782 8.079941e-09 X8 <--> X8
## rho      -0.3049759 0.10136386 -3.008725 2.623469e-03 PR <--> DR
##
##  Iterations =  17
```

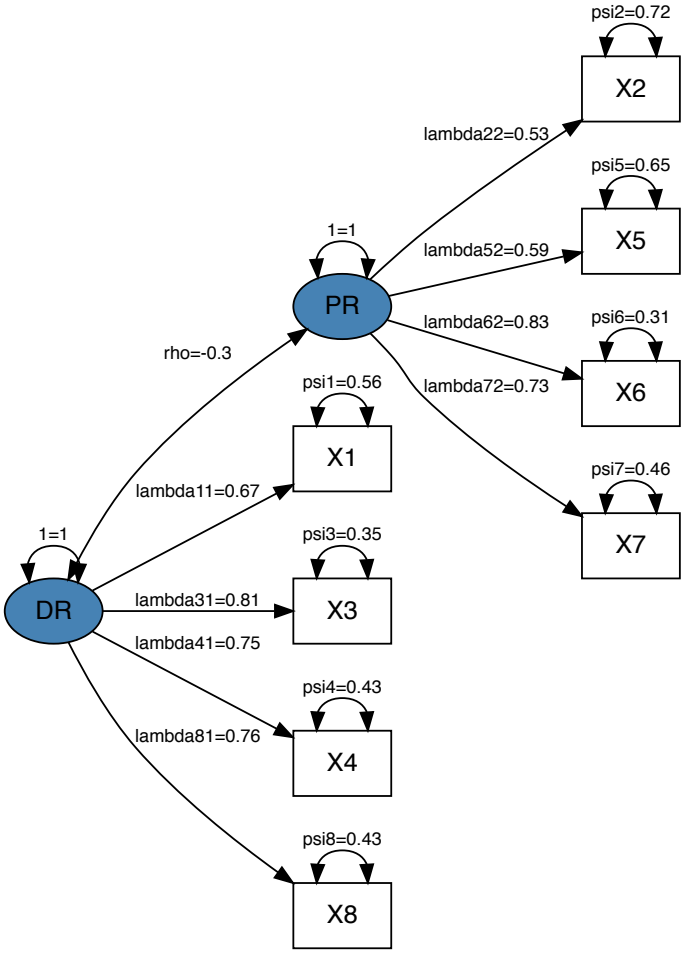
```
par(mfrow=c(1,1))
```

Graph the results

```
library(pathdiagram)
```

```
## Loading required package: shape
```

```
library(shape)
pathDiagram(ability_sem, ignore.double = FALSE, edge.labels = "both",file="ability_seb_f
itted",output.type = "dot", node.colors = c("steelblue", "transparent"))
library(DiagrammeR)
grViz("ability_seb_fitted.dot")
```



Problem 3b 95% Confidence interval of two latent variables

```
CI <- c(-0.305-1.96*0.101, -0.305+1.96*0.101)
CI
```

```
## [1] -0.50296 -0.10704
```