CSSS 589 Homework 3

Cameron Marsden

November 10, 2016

We are interested in examining whether the correlation structure of school attendance and economic activity for age ranges 5-11, 12-14, and 15-17 among 16 countries can be adequately described by a latent factor model. The data for 16 countries come from a series of Child Labor Surveys administered between 1999 to 2009 that were commissioned by the International Labour Organization:

```
#Read in data as subset of six variables of interest:
#(school[age] and eco[age] for age=1 (5-11),2 (12-14),3 (15-17))
labor <- read.csv("childhood_labor.csv")[,c(1,11:13,17:19)]

colnames(labor) <- c("country","school.5to11","school.12to14",
                     "school.15to17","eco.5to11","eco.12to14","eco.15to17")
attach(labor)

kable(labor,digits=4)
```

country	school.5to11	school.12to14	school.15to17	eco.5to11	eco.12to14	eco.15to17
Albania	97.5	94.2	78.2	2.4	9.4	16.8
Benin	78.1	76.0	64.5	29.5	39.1	45.2
Bolivia	98.6	93.5	84.1	17.9	33.5	39.3
Cameroon	98.0	91.9	73.5	30.5	53.6	57.4
Egypt	97.0	88.9	77.6	4.0	13.3	22.6
Indonesia	97.2	93.2	75.4	3.3	10.8	21.9
Jordan	98.5	96.1	86.1	0.3	1.9	5.6
Kyrgyzstan	98.7	99.1	89.2	25.1	44.5	54.2
Madagascar	73.3	76.6	48.9	16.3	36.7	54.4
Mali	56.9	56.3	43.8	61.8	77.2	82.6
Niger	60.5	53.2	26.3	43.3	62.9	67.4
Peru	98.2	92.0	70.3	33.3	49.6	54.1
Moldova	99.4	99.4	88.4	13.8	43.3	42.4
Rwanda	91.9	91.9	66.6	3.4	12.6	33.8
Senegal	60.7	50.2	41.8	12.3	24.0	34.4
Uruguay	99.5	94.1	71.4	3.5	11.7	29.3

Problem 1

The table below gives the correlation values between our six variables of interest:

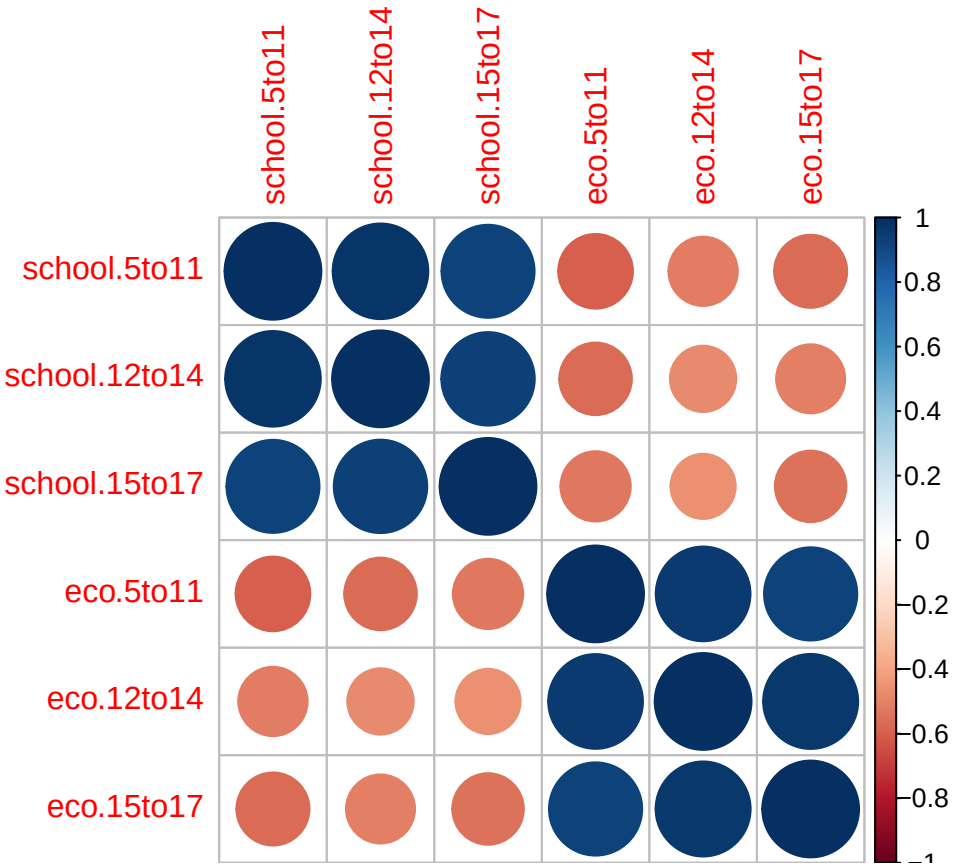
```
X <- labor[2:7]
corr.X <- cor(X)

#Denote p as number of variables and n as no. of countries
p <- length(X)
n <- length(country)

kable(corr.X,digits=4)
```

	school.5to11	school.12to14	school.15to17	eco.5to11	eco.12to14	eco.15to17
school.5to11	1.0000	0.9756	0.9238	-0.5955	-0.5132	-0.5700
school.12to14	0.9756	1.0000	0.9318	-0.5623	-0.4701	-0.5089
school.15to17	0.9238	0.9318	1.0000	-0.5276	-0.4540	-0.5464
eco.5to11	-0.5955	-0.5623	-0.5276	1.0000	0.9576	0.9271
eco.12to14	-0.5132	-0.4701	-0.4540	0.9576	1.0000	0.9637
eco.15to17	-0.5700	-0.5089	-0.5464	0.9271	0.9637	1.0000

We see strong, positive correlations between age groups for school attendance and economics (i.e., top left and bottom right quadrants of matrix). We also see a moderately strong, negative relationship between school attendance and economic activity for all age groups.

We can further visualize the correlation patterns with the following plot:

```
corrplot(corr.X,method="circle")
```



Problem 2

The tables below give summary values when we perform factor analysis with no rotation and two factors (as principal component analysis suggests in 2A).

```
labor.fa <- factanal(X,
                     factors=2,
                     rotation="none",
                     scores="regression")
```

Uniquenesses

```
kable(t(labor.fa$uniquenesses),digits=4)
```

school.5to11	school.12to14	school.15to17	eco.5to11	eco.12to14	eco.15to17
0.0282	0.0179	0.1167	0.0645	0.005	0.0619

Loadings

```
loadings <- as.matrix(labor.fa$loadings)[c(1:6),c(1:2)]

SS.loadings <- c(sum(loadings[,1]^2),sum(loadings[,2]^2))
prop.var <- SS.loadings/p
cum.var <- cumsum(prop.var)

summary.loadings <- rbind(SS.loadings,prop.var,cum.var)
colnames(summary.loadings) <- c("Factor 1","Factor 2")
rownames(summary.loadings) <- c("SS loadings","Proportion Var","Cumulative Var")

kable(loadings,digits=4)
```

	Factor1	Factor2
school.5to11	-0.6943	0.6999
school.12to14	-0.6583	0.7407
school.15to17	-0.6338	0.6940
eco.5to11	0.9620	0.1003
eco.12to14	0.9709	0.2291
eco.15to17	0.9571	0.1482

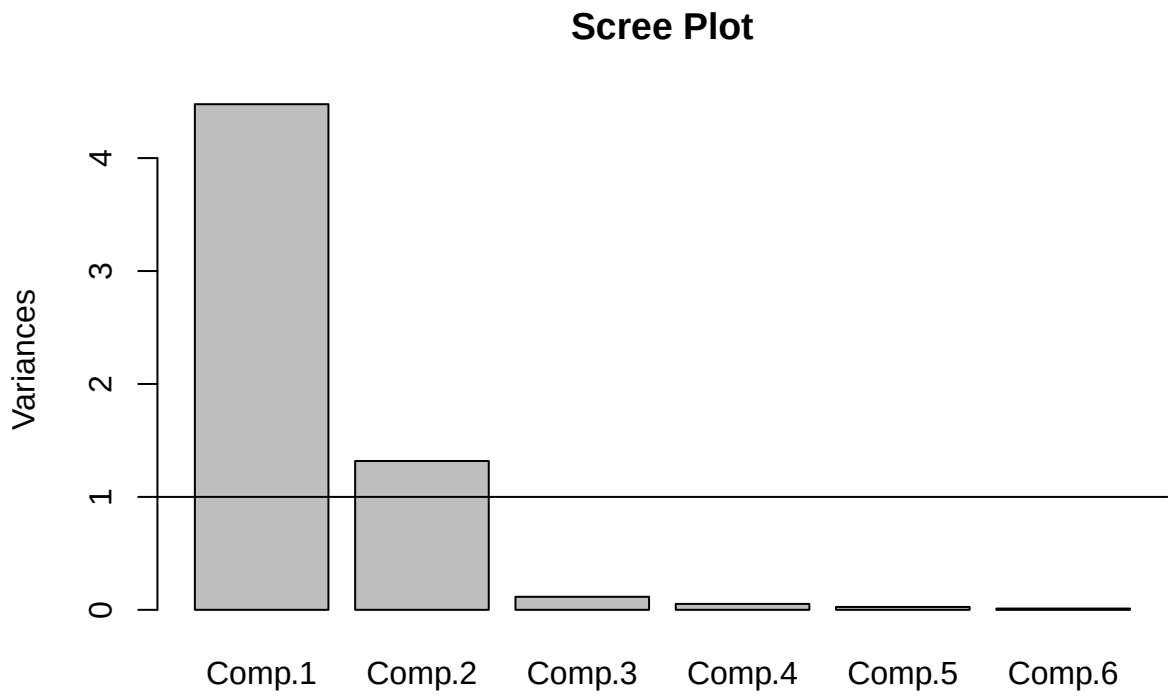
```
kable(summary.loadings,digits=4)
```

	Factor 1	Factor 2
SS loadings	4.1012	1.6046
Proportion Var	0.6835	0.2674
Cumulative Var	0.6835	0.9510

Part A

From our principal component analysis, we found that the first two components explained the vast majority of variability among our six variables. As shown in the Scree Plot, only the first two components exceed our eigenvalue threshold $\lambda > 1$.

```
#Perform PCA
pc.lab <- princomp(X,cor=TRUE)
#Plot results
plot(pc.lab,main="Scree Plot")
abline(h=1)
```



When performing factor analysis, we can look at the corresponding principal component analysis to assess the goodness-of-fit for factor analysis. So for these data, we expect a two-factor model will be most appropriate in explaining the variability between the six variables without overfitting/overcomplicating the analysis.

Part B

In addition to our comparison with principal component analysis, we can use a reconstructed correlation matrix to see how it compares with our observed correlation matrix.

```
reconstr.cor <- loadings(labor.fa)%*%t(loadings(labor.fa))+diag(labor.fa$uniquenesses)

comp.cor <- reconstr.cor-corr.X
kable(comp.cor,digits=4)
```

	school.5to11	school.12to14	school.15to17	eco.5to11	eco.12to14	eco.15to17
school.5to11	0.0000	-0.0001	0.0018	-0.0022	-0.0005	0.0092
school.12to14	-0.0001	0.0000	-0.0006	0.0033	0.0006	-0.0114
school.15to17	0.0018	-0.0006	0.0000	-0.0124	-0.0023	0.0426
eco.5to11	-0.0022	0.0033	-0.0124	0.0000	-0.0007	0.0085
eco.12to14	-0.0005	0.0006	-0.0023	-0.0007	0.0001	-0.0005
eco.15to17	0.0092	-0.0114	0.0426	0.0085	-0.0005	0.0000

From our comparison, we can see that the reconstructed correlations between variables—particularly the school-school and eco-eco comparisons between groups—are nearly identical to the correlations we observed. We will still conduct a statistical test to assess the goodness of fit, but it appears that our two-factor loadings explain the correlations between the six variables very well.

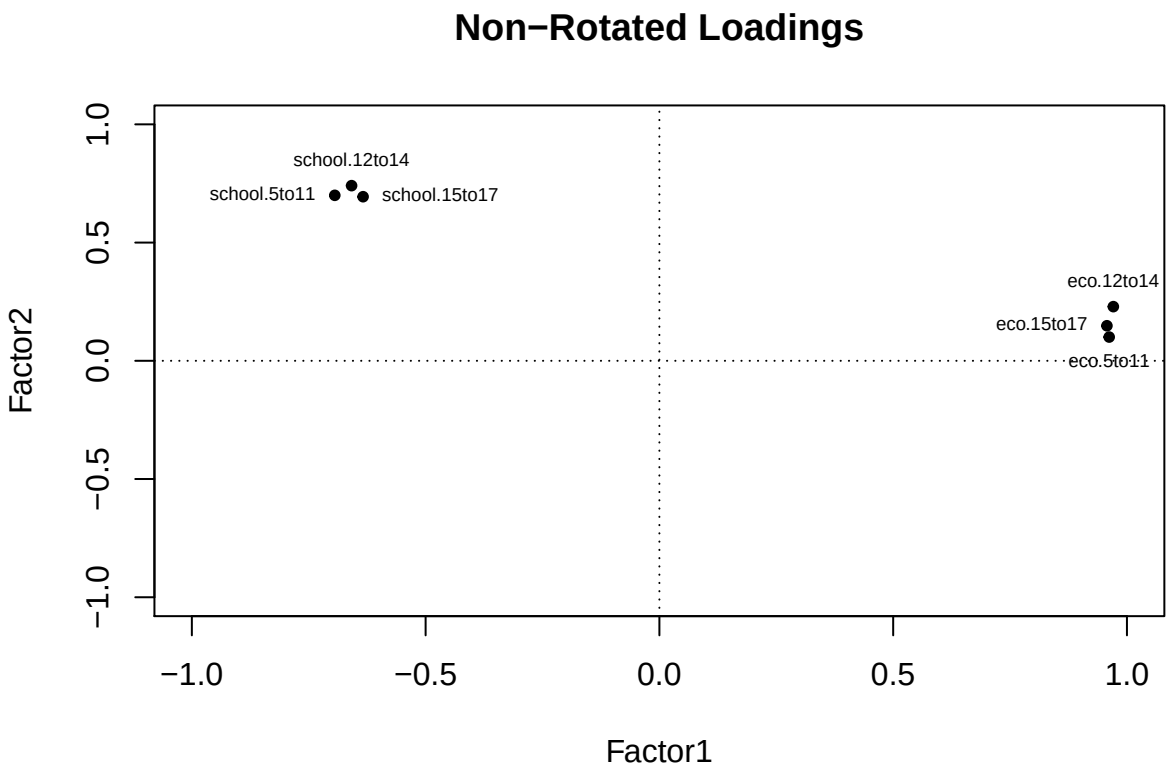
Part C

```
labor.STAT <- labor.fa$STATISTIC
labor.dof <- labor.fa$dof
labor.pval <- labor.fa$PVAL
```

The two-factor analysis in part A returns a χ^2 statistic of 6.8814 with 4 degrees of freedom—resulting in a p-value of 0.1423. With a high p-value, we fail to reject our null hypothesis that the correlation matrix of the manifest variables has the form specified by the factor model. In other words, the two-factor model we specified is a good fit for the observed correlations.

This conclusion confirms that the differences between the reconstructed and observed correlation matrices in Part B are small enough to consider that the two-factor model fits the data well. (In other words, the slightly larger deviance we saw between school and eco variable correlations is likely due to chance variation).

```
plot(labor.fa$loadings, pch=20,
     xlim = c(-1, 1), ylim = c(-1, 1),
main = "Non-Rotated Loadings")
abline(v = 0, h = 0, lty = 3)
text(labor.fa$loadings, labels = colnames(corr.X),
     pos=c(2,3,4,1,3,2),
     cex = .6)
```



Problem 3

To investigate the meaning of the two factors, we can use Varimax and Oblimin rotations for more pronounced factor loadings.

```
labor.fa.varimax <- update(labor.fa,
                           rotation="varimax")
labor.fa.oblimin <- update(labor.fa,
                           rotation="oblimin")
```

Part A

Varimax Rotation

Varimax rotation imposes orthogonality of factors, so the correlation matrix will simply be identity:

```
labor.scores.varimax <- labor.fa.varimax$scores
factor.corr.varimax <- cor(labor.scores.varimax)
rownames(factor.corr.varimax) <- colnames(factor.corr.varimax) <- c("Factor 1","Factor 2")

kable(factor.corr.varimax,digits=1)
```

	Factor 1	Factor 2
Factor 1	1	0
Factor 2	0	1

Oblimin Rotation

The Oblimin rotation allows for correlation of factors to produce stronger distinction among variables:

```
labor.scores.oblimin<- labor.fa.oblimin$scores
factor.corr.oblimin <- cor(labor.scores.oblimin)
rownames(factor.corr.oblimin) <- colnames(factor.corr.oblimin) <- c("Factor 1","Factor 2")

kable(factor.corr.oblimin,digits=4)
```

	Factor 1	Factor 2
Factor 1	1.0000	0.5386
Factor 2	0.5386	1.0000

We see there is a 0.5386 correlation between factors 1 and 2 when we use the Oblimin rotation. (Note: The factors 1 and 2 are inverted from Varimax as explained in part B.)

Part B

Varimax Rotation

The following two tables show the loading values and summary when performing two-factor analysis with the Varimax rotation:

```
loadings <- as.matrix(labor.fa.varimax$loadings)[c(1:6),c(1:2)]

SS.loadings <- c(sum(loadings[,1]^2),sum(loadings[,2]^2))
prop.var <- SS.loadings/p
cum.var <- cumsum(prop.var)

summary.loadings <- rbind(SS.loadings,prop.var,cum.var)
colnames(summary.loadings) <- c("Factor 1","Factor 2")
rownames(summary.loadings) <- c("SS loadings","Proportion Var","Cumulative Var")

kable(loadings,digits=4)
```

	Factor1	Factor2
school.5to11	-0.3253	0.9306
school.12to14	-0.2753	0.9520
school.15to17	-0.2732	0.8992
eco.5to11	0.9114	-0.3236
eco.12to14	0.9749	-0.2112
eco.15to17	0.9277	-0.2784

```
kable(summary.loadings,digits=4)
```

	Factor 1	Factor 2
SS loadings	2.898	2.8078
Proportion Var	0.483	0.4680
Cumulative Var	0.483	0.9510

For Factor 1 using Varimax-rotated loadings, we see moderate negative correlation with all three school attendance variables—while retaining independence from Factor 2. On the other hand, we see very high correlation values (all above 0.9) for each economic variable.

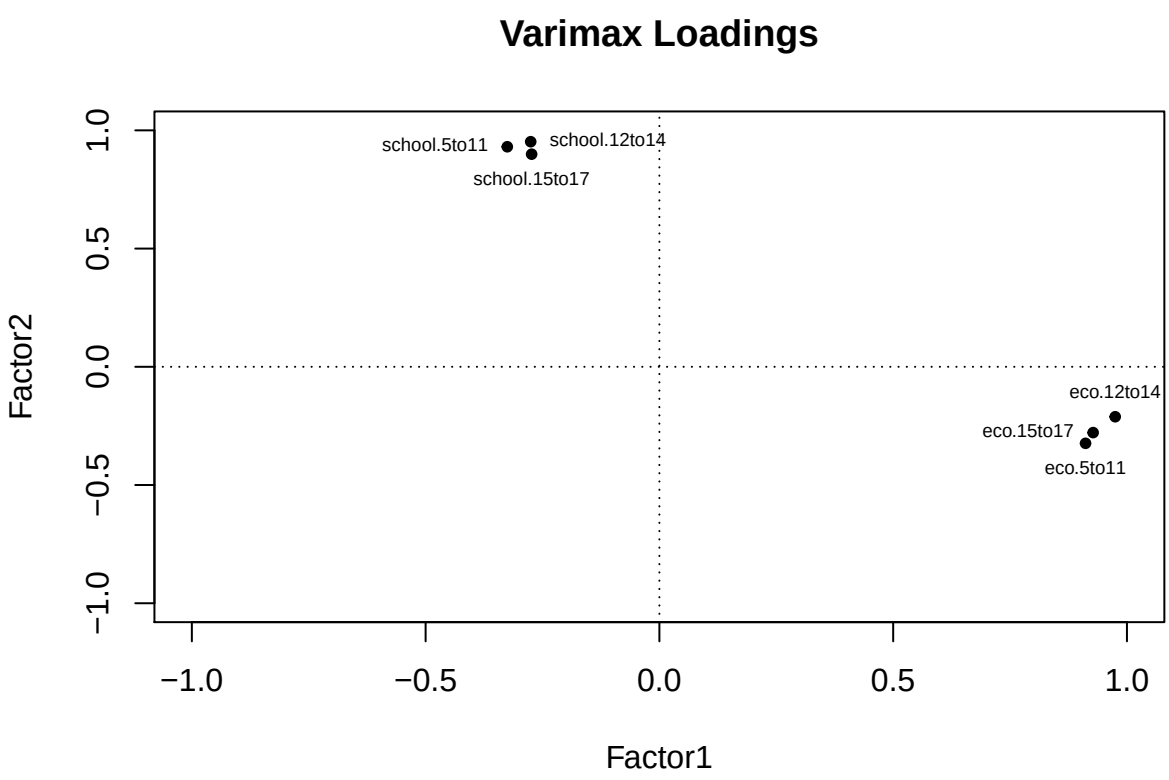
**Factor 1 describes the latent variable of economic activity among each country's children between ages 5-17.*

For Factor 2, we have a reversal of variable effects. The three economic variables are moderately negative while the three school variables all have strong, positive correlations (all above 0.89).

Factor 2 describes the latent variable of school attendance among each country's children between ages 5-17.

The plot below shows the Varimax-rotated loadings:

```
plot(labor.fa.varimax$loadings, pch=20,
     xlim = c(-1, 1), ylim = c(-1, 1),
main = "Varimax Loadings")
abline(v = 0, h = 0, lty = 3)
text(labor.fa.varimax$loadings, labels = colnames(corr.X),
     pos=c(2,4,1,1,3,2),
     cex = .6)
```



Oblimin Rotation

The following two tables show the loading values and summary when performing two-factor analysis with the Oblimin rotation:

```
loadings <- as.matrix(labor.fa.oblimin$loadings)[c(1:6),c(1:2)]

SS.loadings <- c(sum(loadings[,1]^2),sum(loadings[,2]^2))
prop.var <- SS.loadings/p
cum.var <- cumsum(prop.var)

summary.loadings <- rbind(SS.loadings,prop.var,cum.var)
colnames(summary.loadings) <- c("Factor 1","Factor 2")
rownames(summary.loadings) <- c("SS loadings","Proportion Var","Cumulative Var")

kable(loadings,digits=4)
```

	Factor1	Factor2
school.5to11	0.9630	-0.0410
school.12to14	1.0035	0.0234
school.15to17	0.9436	0.0071
eco.5to11	-0.0803	0.9213
eco.12to14	0.0688	1.0332
eco.15to17	-0.0232	0.9558

```
kable(summary.loadings,digits=4)
```

	Factor 1	Factor 2
SS loadings	2.8366	2.8321
Proportion Var	0.4728	0.4720
Cumulative Var	0.4728	0.9448

With Oblimin Factor Analysis, we allow factors to be correlated with one another. As we might expect, the school attendance for children within a country will be correlated with the amount of economic activity of children. In the rotation we are using, we see a moderate correlation (which I will interpret as the correlation between change of school attendance and change of economic acitivity between age groups). Once we account for the between factor correlation, we see a stronger distinction of the same pattern we saw with the Varimax rotation:

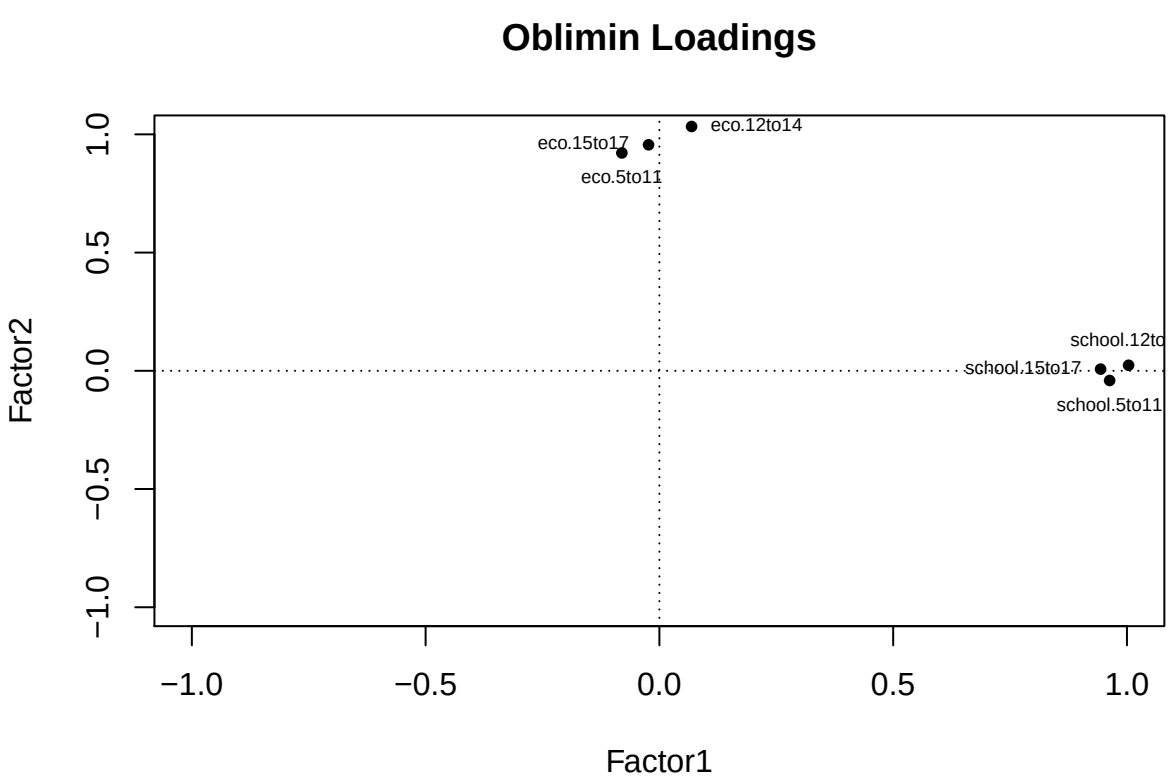
Note: Since we are treating the covariance between the two factors as positive (as given by the default R output when updating the unrotated matrix to the oblimin rotated matrix), the factors switch latent dimensions from Varimax.

Factor 1 describes the latent dimension of school attendance among each country's children between ages 5-17 with very little correlation to the economic activity variables.

Factor 2 describes the latent dimension of economic activity among each country's children between ages 5-17 with very little correlation to school attendance.

The plot below shows the Oblimin-rotated loadings. (Note that the patterns are very similar to the ones seen with non-rotated and Varimax-rotated loadings, however, the variables match with opposite factors—which does not change our overall conclusions as we are interested in the presence of two distinct factors regardless of order—as shown by the nearly equal variance explained by each factor.)

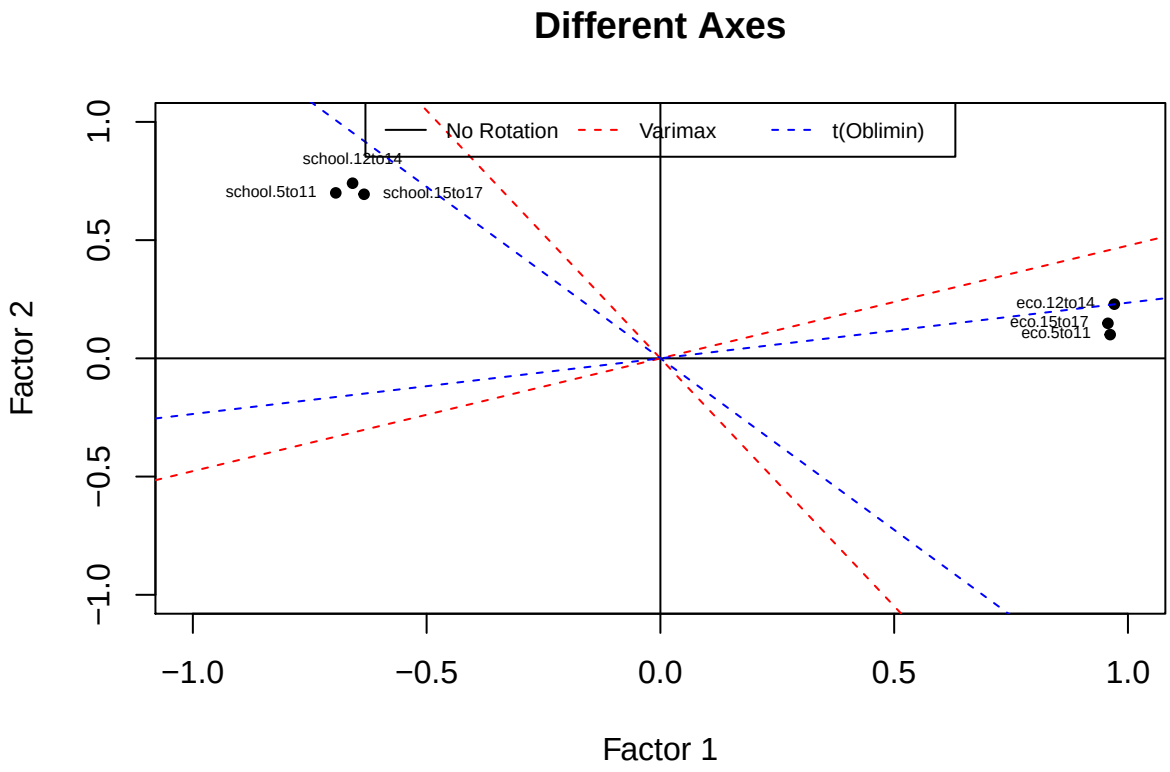
```
plot(labor.fa.oblimin$loadings, pch=20,
     xlim = c(-1, 1), ylim = c(-1, 1),
main = "Oblimin Loadings")
abline(v = 0, h = 0, lty = 3)
text(labor.fa.oblimin$loadings, labels = colnames(corr.X),
     pos=c(1,3,2,1,4,2),
     cex = .6)
```



Below is the plot of the three analysis methods (i.e., no rotation, Varimax rotation, and (transposed) Oblimin rotation):

```
plot(labor.fa$loadings, pch=20,
     xlim=c(-1,1), ylim=c(-1,1),
     main="Different Axes", xlab="Factor 1",ylab="Factor 2")

abline(v=0,h=0,lty=1)
text(labor.fa$loadings, labels=colnames(corr.X),cex=0.5,
     pos=c(2,3,4,2,2,2))

varimax.basis <- solve(labor.fa.varimax$rotmat)
oblimin.basis <- t(labor.fa.oblimin$rotmat)

abline(a=0, b=varimax.basis[2]/varimax.basis[1],lty=2,col="red")
abline(a=0, b=varimax.basis[4]/varimax.basis[3],lty=2,col="red")
text(t(varimax.basis)*2,c("V1","V2"))

abline(a=0, b=oblimin.basis[2]/oblimin.basis[1],lty=2,col="blue")
abline(a=0, b=oblimin.basis[4]/oblimin.basis[3],lty=2,col="blue")
text(t(oblimin.basis)*2,c("O1","O2"))

legend("top",
       col = c("black", "red", "blue"),
       lty = c(1,2,2),
legend = c("No Rotation", "Varimax", "t(Oblimin)"),
```

```
ncol = 3, cex = .7)
```

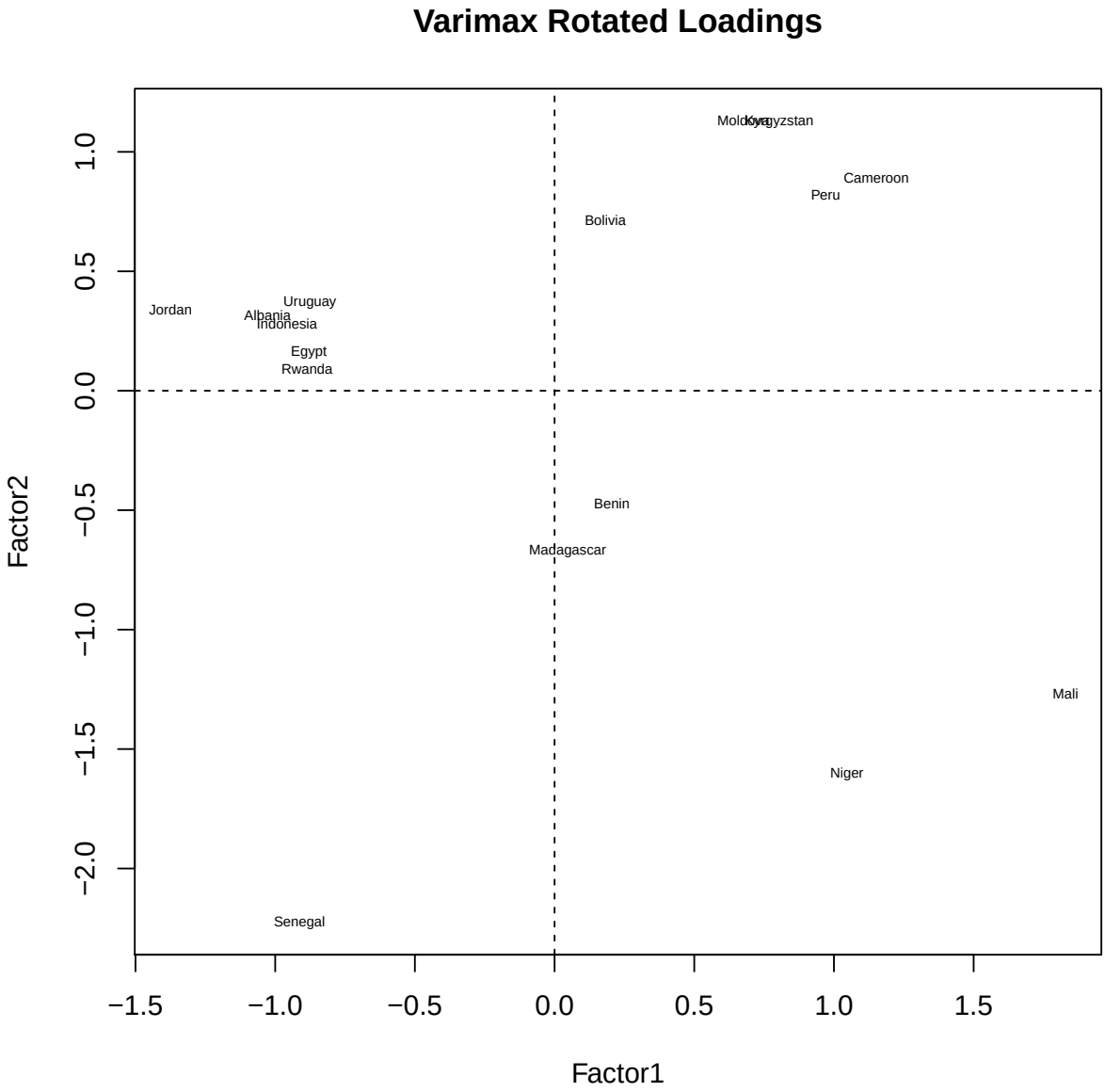


Problem 4

Due to the strong relationships viewed with both Oblimin and Varimax-rotated loadings, it is easier to interpret the scores when using Varimax-rotation as we can treat the two factors as independent. en factors 1 and 2 as it accounts for this correlation.

The plot below demonstrates the Varimax-rotated scores for all sixteen countries from our dataset:

```
plot(labor.fa.varimax$scores,
     pch=NA,
     main="Varimax Rotated Loadings")
abline(v=0,h=0,lty=2)
text(labor.fa.varimax$scores,
     labels=country,
     cex=0.5)
```



In the plot, we can see which countries have high rates of economic activity for children ages 5-17 (factor 1) and high school attendance records for the same age range (factor 2):

We expect the two latent variables to be negatively associated with one another: if a country has high school attendance rates, we would expect there to be low economic activity among children, and vice versa. With Varimax-rotated loadings, we force independence between the two latent variables, so we should expect to see fewer countries that have either high rates of school attendance combined with high rates of children economic activity or low rates for the same latent variables. The few countries that break this expectation are listed below:

(Note that high vs. low rates of either variable for country i is measured in contrast with the other n=15 countries.)

- In the top right quadrant, we see countries with high rates of each latent variable: Cameroon, Peru, Kyrgyzstan, and Bolivia

- In the bottom left quadrant, we see one country with low rates of each latent variable: Senegal.

As expected, we see many more countries that demonstrate an inverse relationship between the two latent dimensions:

- In the top left quadrant, we see countries with high school attendance and low children's economic activity: Jordan, Uruguay, Albania, Indonesia, Egypt, Rwanda
- In the bottom right quadrant, we see countries with low school attendance and high children's ecomonic activity: Madagascar, Benin, Niger, Mali

Looking at geographical patterns, we see trends between countries of similar combinations of the two latent variables:

- We see a clear trend of African countries having a low school attendance. In fact, all five countries with negative scores for factor 2 (school attendance) are from Sub-Saharan Africa.
- We also see high children's economic activity in Latin America with the exception of Uruguay.
- Our sample of Middle Eastern countries,Jordan and Egypt, both appear in the top left quadrant, indicating low amounts of children's economic activity and high school attendance—a combination of factor dimension values which generally indicates a country with higher socioeconomic status.

Problem 5

Principal Component Analysis vs. Factor Analysis for our data

Compared with Principal Component Analysis of the same dataset, we saw many similarities:

- Most importantly, we determined that there were two components to explaining the vast majority of variance between the 16 sampled countries.
- Both of the components dealt with the same structure of economic activity and school attendance.
- Specifically, we analyzed Kyrgyzstan and Senegal using Principal Component Analysis, and we saw similar patterns as we see now with factor analysis: Kyrgystan has a very high level of school attendance while Senegal stands out from the entirety of the other 15 sampled countries with its low rates of both school attendance and economic activity.

There are certain advantages to each type of analysis as well. With principal component analysis, we saw that the about 75% of the variance between countries was explained by the first component and 22% for the second. For unrotated factor analysis, the differences between the two proportions of variance weren't as pronounced. With principal component analysis

- Building on the last point, with Principal Component Analysis, we could see that the first component explained the vast majority of the variance. Instead of explaining one dimension, we concluded that the first component indicated the stark contrast between countries with high school attendance and countries with high economic activity.
- For the rotated factor analysis models, we have the benfit of segmenting the two factors into more easily interpretable latent dimensions—particularly with Varimax rotation. For this type of rotation, the two factors are viewed as independent, so we were able to analyze each quadrant of country scores easily. We could also interpret the Oblimin rotation which allows for correlation between the two factors while allowing each factor to have the same amount of importance (as shown by the proportion of explained variance).

Overall comparisons

- Overall, Principal Component Analysis allows for more in-depth, complex components (e.g., contrasts) at the cost of more easily interpretable factor analysis models. In addition, Prinicpal Component Analysis creates a clearer hierarchy of importance for its components while factor analysis with different rotations allows for interpretation of each component with equal weight. Depending on the data itself, both methods are useful for discovering underlying patterns in multidimensional data.