For this post, I’m going to show the basics of path analysis. I’ll show how to reproduce the correlation matrix and covariance matrix.

The Model

I will use the HSB dataset, which is at the time of this post publicly available on the IDRE site. The file is in sas7bdat format so we’ll first need to get the data into R and then we’ll ready to go.

Import the data into R from SAS format

Fortunately, there is a package that reads sas7bdat files into *R* so I’ll use that.

```
library(sas7bdat)
hsb<-read.sas7bdat("http://www.ats.ucla.edu/stat/data/hsb2.sas7bdat")
path.data<-cbind(hsb$read, hsb$math, hsb$science)
colnames(path.data)<-c("read","math","science")
path.data<-as.data.frame(path.data)
head(path.data)
```

```
## read math science
## 1  57  41  47
## 2  68  53  63
## 3  44  54  58
## 4  63  47  53
## 5  47  57  53
## 6  44  51  63
```

The variables I want to use are Math, Reading, and Science. Let’s estimate the model below. This is a pretty straightforward multiple regression model.
library(lavaan)
path.model<-'
  science ~ read + math
  read~~math
  science~~science
',
path.fitted<-sem(path.model, data=path.data, fixed.x=FALSE)
summary(path.fitted, standardized=TRUE)

## lavaan (0.5-16) converged normally after 31 iterations
##
## Number of observations 200
##
## Estimator ML
## Minimum Function Test Statistic 0.000
## Degrees of freedom 0
## P-value (Chi-square) 1.000
##
## Parameter estimates:
##
## Information Expected
## Standard Errors Standard
##
## Regressions:
## science ~
## read  0.365  0.066  5.551  0.000  0.365  0.378
## math  0.402  0.072  5.576  0.000  0.402  0.380
Reproducing the correlation matrix

To reproduce the correlation matrix, we need to use the standardized estimates. Let me add the path diagram that includes the unstandardized estimates.

![Path Diagram](image)

The actual correlation is:

```r
round(cor(path.data),2)
## read math science
## read 1.00 0.66 0.63
## math 0.66 1.00 0.63
## science 0.63 0.63 1.00
```

To reproduce the correlation matrix, we need to apply the tracing rules we discussed in class.

**Correlation between reading and math**

This correlation can be estimated directly from the data.

```r
round(cor(path.data$read, path.data$math),2)
## [1] 0.66
```
**Correlation between reading and science**

What are all of the possible paths we could take from reading (R) to science (S)?

1. \( RS = p_{rs} = .378 \)
2. \( RMS = r_{rm} \times p_{ms} = .662 \times .380 = 0.2516 \)

Now add these together to get the correlation between reading and science: 0.63.

**Correlation between math and science**

What are all of the possible paths we could take from math (M) to science (S)?

1. \( MS = p_{ms} = .380 \)
2. \( MRS = r_{rm} \times p_{rs} = .662 \times .378 = 0.2502 \)

Now add these together to get the correlation between reading and science: 0.63.

**Reproducing the covariance matrix**

To reproduce the covariance matrix, we need to use the unstandardized estimates. Let me add the path diagram that includes the unstandardized estimates.

The actual covariance is:

```r
cov(path.data), 2
```

<table>
<thead>
<tr>
<th></th>
<th>read</th>
<th>math</th>
<th>science</th>
</tr>
</thead>
<tbody>
<tr>
<td>read</td>
<td>105.12</td>
<td>63.61</td>
<td>63.97</td>
</tr>
<tr>
<td>math</td>
<td>63.61</td>
<td>87.77</td>
<td>58.50</td>
</tr>
<tr>
<td>science</td>
<td>63.97</td>
<td>58.50</td>
<td>98.03</td>
</tr>
</tbody>
</table>

To reproduce the covariance matrix, we need to apply the tracing rules we discussed in class.
Covariance between reading and math
This covariance can be estimated directly from the data.

\[
\text{round}(\text{cov(path.data$read, path.data$math),2)}
\]
\#
[1]
63.61

Covariance between reading and science
What are all of the possible paths we could take from reading (R) to science (S)?

1. RS = \(\sigma_R^2 \times p_{rs} = 105.1 \times .365 = 38.36\)
2. RMS = \(\sigma_m \times p_{ms} = 63.61 \times .402 = 25.57\)

Now add these together to get the correlation between reading and science: 63.93.

Covariance between math and science
What are all of the possible paths we could take from math (M) to science (S)?

1. MS = \(\sigma_M^2 \times p_{ms} = 87.8 \times .402 = 35.3\)
2. MRS = \(\sigma_r \times p_{rs} = 63.61 \times .365 = 23.22\)

Now add these together to get the correlation between reading and science: 58.51.

Variance of science scores
Just for fun, why don’t we see what estimate we get the for variance of the outcome variable, science scores? Let’s find all possible paths from and back to science scores (S).

1. SMS = \(p_{ms} \times \sigma_M^2 \times p_{ms} = .402 \times 87.8 \times .402 = 14.19\)
2. SRS = \(p_{rs} \times \sigma_R^2 \times p_{rs} = .365 \times 105.1 \times .365 = 14\)
3. SMRS = \(p_{ms} \times \sigma_r \times p_{rs} = .402 \times 63.61 \times .365 = 9.33\)
4. SRMS = \(p_{rs} \times \sigma_r \times p_{ms} = .365 \times 63.61 \times .402 = 9.33\)

Now add these together to get the correlation between reading and science: 46.86.

Wait a minute! The actual variance of the science scores is 98.0, but we just computed 46.86. Using the math and reading scores, we were only able to explained a proportion of the variance in science scores. The proportion is:

\[
\frac{46.86}{98} = 0.4782
\]

If that sounds familiar, there’s a good reason for why. It’s \(R^2\) from regression! Let’s estimate a regression model and see what the estimated \(R^2\) value is.
\texttt{summary(lm(science~read+math, path.data))}$r\.squared$

## [1] 0.4782