3.1 | Path Diagrams for Structural Equations

Recall that for a given data set, exploratory factor analysis can be used to identify clusters of highly correlated dimensions, both among those considered to be *predictors* and among those considered to be *responses*, and represent them as factors or linear combinations of said dimensions. In many data analyses, the next stage is to quantify and investigate the effect that the resulting predictor factors appear to have on the resulting response factors. This can be achieved through the use of a structural equation model, which begins by specifying and visualizing these and other relationships among factors and dimensions using a path diagram.

3.1.1 | Revisiting the Loading Matrix

The exact relationships in a path diagram can be defined using a variety of methods, including expert opinion and previous research. A more quantitative approach, however, is to use the results of exploratory factor analysis to determine how factors and dimensions relate to each other.

Consider an example data set from a conference presentation on the epidemiology of toxoplasmosis infections and corresponding socioeconomic risk factors among people living in urban, rural, and farm communities throughout southern Chile. Toxoplasmosis is a globally prevalent infection caused by the parasite *Toxoplasma gondii* with a wide range of clinical symptoms in humans. *T. gondii* has a complex lifecycle with many opportunities for human infection, including direct contact with contaminated soil through gardening and consumption of contaminated food.

```
SIG <- matrix(c(1.54, .10, .10,-.07, .21, .05, .22, .02, .07,-.13,
                 .10, .89, .02, .23, .12, .03, .08, .08, .10, .73,
                 .10, .02,8.96, .45, .06,-.22, .15, .22, .12, .54,
                -.07, .23, .45,1.33, .52, .02, .24, .29, .28, .66,
                 .21, .12, .06, .52,1.47, .09, .22, .16, .27, .43,
                 .05, .03,-.22, .02, .09, .26, .07,-.01, .08,-.13,
                 .22, .08, .15, .24, .22, .07, .69, .32, .40, .16,
                 .02, .08, .22, .29, .16,-.01, .32, .56, .36, .17,
                 .07, .10, .12, .28, .27, .08, .40, .36, .84, .15,
                -.13, .73, .54, .66, .43,-.13, .16, .17, .15,6.69),
              nrow = 10, ncol = 10, byrow = T,
              dimnames = list(c("Sex","Age","Income","WorksGardens",
                                "CleansBarns","DrainsLands",
                                "ButchersCows","MilksCows",
                                "BirthsCows","OpticalDensity"),
                              c("Sex","Age","Income","WorksGardens",
                                "CleansBarns","DrainsLands",
                                "ButchersCows","MilksCows",
                                "BirthsCows","OpticalDensity")))
```

These data include the covariance structure between a set of nine predictor dimensions and a single response dimension measured on 303 study participants. The predictor dimensions include the Sex, Age, and Income of the study participants, as well as a variety of risk factors such as whether the individual WorksGardens, CleansBarns, or DrainsLands that have been flooded. The single response dimension, on the other hand, measures the OpticalDensity (an indicator of antibody concentration, clinical diagnosis, and infection) from the results of a Colorimetric Enzyme-Linked Immunosorbent Assay (ELISA) Detection Platform.

Given the need to work separately with the two different sets of dimensions (i.e., predictors and responses), it is helpful to use matrix notation to separate out the covariance matrix for the set of predictor dimensions from that for the set of response dimensions.

```
SIG_p <- SIG[1:9,1:9]

SIG_r <- SIG[10, 10]
SIG_r <- as.matrix(SIG_r)
```

In our example data, we use $\boldsymbol{\Sigma_p}$ to refer to the covariance matrix for the set of predictor dimensions and $\boldsymbol{\Sigma_r}$ to refer to set of response dimensions.

Because the covariance matrix for the set of response dimensions is simply a single variance (that of the OpticalDensity dimension), we use the as.matrix() function to ensure that it is maintained as a matrix rather than a vector in R. This also means that there is no need to perform factor analysis on this set of dimensions. Recall that the purpose of factor analysis is to reduce the dimensionality of a data set, and a dimensionality of one cannot be reduced any further! The approach that would be taken had this set included more than one dimension, however, would be identical to that demonstrated below for the set of predictor dimensions.

Now, although structural equation modeling uses the covariance structure as the basis for analysis, exploratory factor analysis instead uses the correlation structure, a standardized form of the covariance structure. As such, we often need to use matrix algebra or the cov() function in R to calculate the correlation matrices for one or, if necessary, both sets of dimensions.

```
R_p <-  (solve(sqrt(diag(diag(SIG_p))))) %*% SIG_p %*%
       t(solve(sqrt(diag(diag(SIG_p)))))
dimnames(R_p) <- list(c("Sex","Age","Income","WorksGardens",
                        "CleansBarns","DrainsLands",
                        "ButchersCows","MilksCows","BirthsCows"),
                      c("Sex","Age","Income","WorksGardens",
                        "CleansBarns","DrainsLands",
                        "ButchersCows","MilksCows","BirthsCows"))
```

Because using matrix algebra to find the correlation matrix often results in a loss of dimension names, we used the dimnames() function to reestablish the proper row and column names before proceeding with the actual factor analysis.

Recall that the first step in exploratory factor analysis is to identify the intrinsic dimensionality.

```
eigen(R_p)$values
[1] 2.5308671 1.2333196 1.0816741 1.0335368 0.9081244 0.8047493
[7] 0.5494241 0.4411884 0.4171161
```

Based on Kaiser's criterion, this set of dimensions can be represented using only four factors.

We can now use the `pca()` function to investigate which dimensions are highly correlated and if oblique rotation is necessary, specifying the most common oblique rotation algorithm `"oblimin"` for the `rotate` argument.

```
pca(r = R_p, nfactors = 4, rotate = "oblimin")

Principal Components Analysis
Call: principal(r = r, nfactors = nfactors, residuals = residuals,
    rotate = rotate, n.obs = n.obs, covar = covar, scores = scores,
    missing = missing, impute = impute,
    oblique.scores = oblique.scores,
    method = method, use = use, cor = cor, correct = 0.5,
    weight = NULL)
Standardized loadings (pattern matrix) based upon correlation matrix
                TC1   TC3   TC2   TC4   h2   u2 com
Sex            0.03  0.00  0.04  0.94 0.89 0.11 1.0
Age           -0.20  0.67 -0.02  0.14 0.43 0.57 1.3
Income         0.06  0.11  0.77  0.20 0.63 0.37 1.2
WorksGardens   0.17  0.73  0.12 -0.23 0.70 0.30 1.4
CleansBarns    0.05  0.67 -0.15  0.13 0.52 0.48 1.2
DrainsLands    0.09  0.14 -0.73  0.13 0.60 0.40 1.2
ButchersCows   0.81 -0.03 -0.06  0.22 0.72 0.28 1.2
MilksCows      0.82  0.03  0.15 -0.12 0.72 0.28 1.1
BirthsCows     0.81  0.03 -0.12 -0.03 0.69 0.31 1.0

                       TC1  TC3  TC2  TC4
SS loadings           2.09 1.49 1.19 1.11
Proportion Var        0.23 0.17 0.13 0.12
Cumulative Var        0.23 0.40 0.53 0.65
Proportion Explained  0.36 0.25 0.20 0.19
Cumulative Proportion 0.36 0.61 0.81 1.00

 With component correlations of
     TC1  TC3   TC2   TC4
TC1 1.00 0.33  0.01  0.08
TC3 0.33 1.00  0.02  0.05
TC2 0.01 0.02  1.00 -0.08
TC4 0.08 0.05 -0.08  1.00

Mean item complexity =  1.2
Test of the hypothesis that 4 components are sufficient.

The root mean square of the residuals (RMSR) is  0.11

Fit based upon off diagonal values = 0.72
```

Using the factor correlation matrix in this output (highlighted in bold above), we can see that oblique rotation was necessary because at least one of the correlations between factors is beyond the common threshold of 0.30. If, on the other hand, this matrix indicated that oblique rotation was not necessary, we would run the factor analysis again under an orthogonal rotation.

Regardless of whether orthogonal or oblique rotation is used, the next step is to extract the corresponding rotated loading matrix.

```
A <- pca(r = R_p, nfactors = 4, rotate = "oblimin")$loadings[]
round(A, 3)
                 TC1    TC3    TC2    TC4
Sex            0.030 -0.004  0.038  0.943
Age           -0.196  0.671 -0.019  0.145
Income         0.056  0.109  0.766  0.196
WorksGardens   0.167  0.732  0.118 -0.232
CleansBarns    0.046  0.669 -0.149  0.127
DrainsLands    0.095  0.137 -0.728  0.134
ButchersCows   0.806 -0.026 -0.056  0.221
MilksCows      0.819  0.031  0.152 -0.125
BirthsCows     0.813  0.033 -0.119 -0.029
```

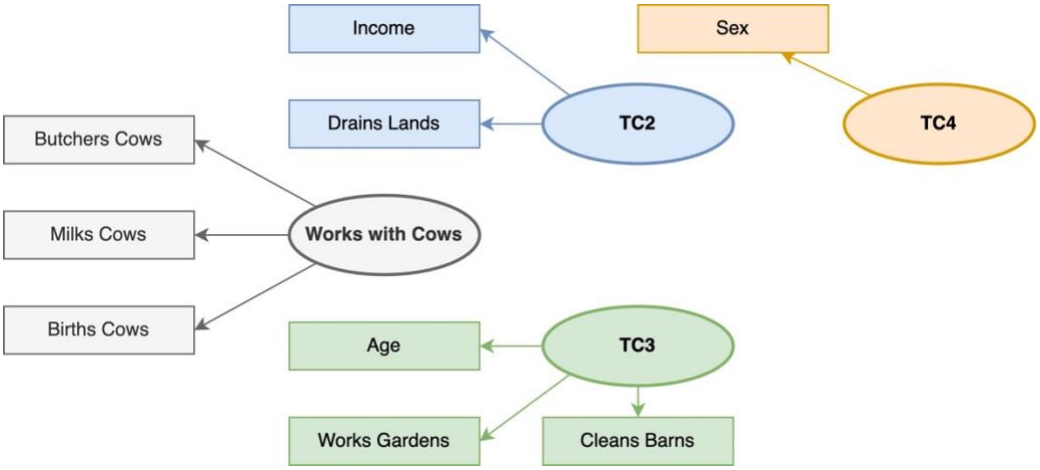
Using the common threshold of an absolute loading above 0.3, we can see that ButchersCows, MilksCows, and BirthsCows all load substantially on to first same factor, which we can refer to as Working with Cows. Age, WorksGardens, and CleansBarns, on the other hand, load substantially on to the second factor, whereas as Income and DrainsLands load on to the third factor. The remaining dimension of Sex loads on to the fourth factor and by itself comprises the entirety of that factor, something that we will address below in Section 3.1.2.

3.1.2 | Defining the Measurement Model

Structural equation models can be thought of as the combination of two underlying models that relate different factors and dimensions to each other. The first is known as the *measurement model*, which is used to define the relationships between the original dimensions and the corresponding factors separately for the sets of predictor and response measurements. The second model, which is known as the structural model, is described below in Section 3.1.3.

We can visualize the hypothesized set of relationships in these models using a *path diagram*. We create a path diagram by displaying the original measured dimensions as boxes and the corresponding factors as circles and use lines with arrows to specify how one affects the other. It is important to note that in different disciplines, dimensions may be referred to as *measured* or *observed variables*, *indicators*, or *manifest variables*, whereas factors may be referred to as *latent variables*, *constructs*, or *unobserved variables*. More generally, any construct in a path diagram, whether a dimension or factor, may be referred to simply as a variable.

Although we can, with some effort, create path diagrams for structural equation models in R, it is often easier to use an external flow chart editor to achieve this. One popular choice is available at https://app.diagrams.net/.



Using our example data, we can see that each factor has a line connecting it to its constituent dimensions that were identified using exploratory factor analysis. Lines such as these that have only one arrow represent a *directed* relationship between a factor and dimension.

Although it is common to think of dimensions as making up or comprising the factors, structural equation modeling takes the opposite approach. In our example data, the implication is that a study participant who `WorkswithCows` results in or produces an individual who also, to some degree, `ButchersCows`, `MilksCows`, and `BirthsCows`.

There are a few issues with the current path diagram, however, that need to be addressed.

First, we have a factor that is leading to only a single dimension of `Sex`. Although all dimensions must align with one or more factors in exploratory factor analysis, a common practice in structural equation modeling is to instead retain any dimension that by itself comprises a factor in its original unfactored state. This avoids any unnecessary one-to-one relationships.

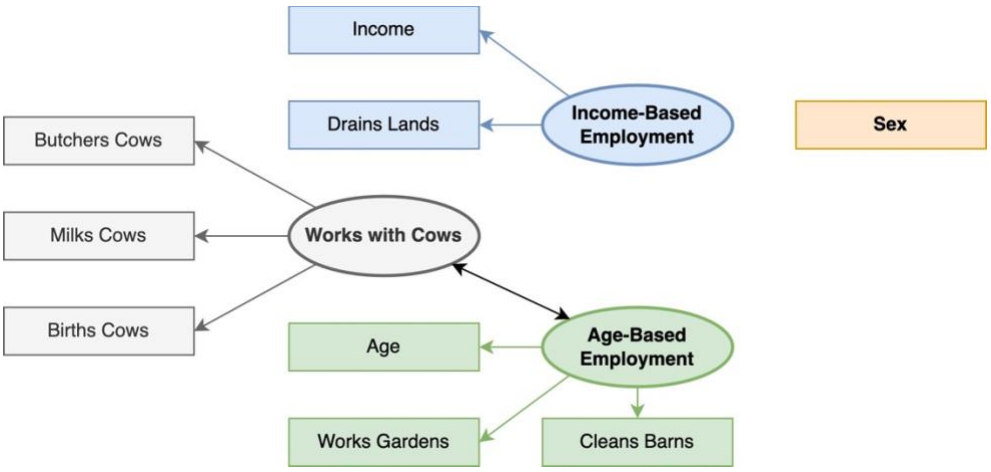
Second, exploratory factor analysis indicated that oblique rotation was necessary, implying that at least some of the factors are correlated. Using the factor correlation matrix in the `pca()` function output, we can see that the first factor `T1` (here renamed to `WorkswithCows`) and second factor `T3` have substantial correlation beyond the standard absolute threshold of 0.3. We can use a line with two arrows to visualize such an *undirected* relationship, which indicates some level of covariability between the factors without any specific direction as to which factor is driving or causing the other.

Third, although we have named one of our now three remaining factors, the other two factors are simply listed as `TC2` and `TC3`. An informative path diagram should name all factors in a way that eases interpretation of results. One thing to note is that this may not always be easy; in

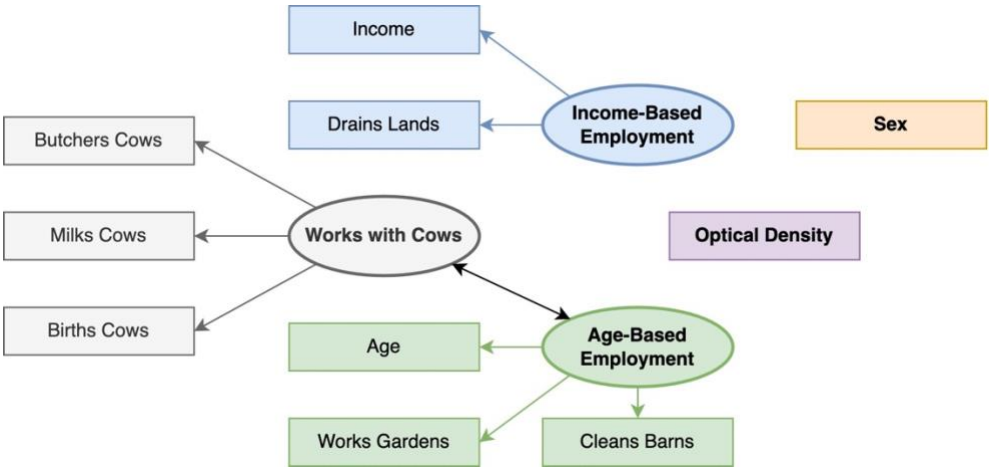
such situations, it may help to revisit the intrinsic dimensionality and investigate if a different number would have resulted in factors that were easier to name or interpret. Alternatively, we can use one of the dimensions comprising such hard-to-name factors as the basis for the name.

Finally, although this issue does not appear in our example data, many factor analyses yield dimensions that despite either orthogonal or oblique rotation, remain complex. In such situations, these dimensions are often retained outside of any factor, much like a dimension that is the only constituent of a factor. Additionally, a line with two arrows is then drawn from such retained dimensions to any factor that they substantially loaded on to, indicating some degree of covariability between them (similar to that exhibited by oblique factors).

We can now update our path diagram to encompass these adjustments.



Now that we have visualized the paths in the measurement model for the set of predictor variables, we can do the same for the set of response variables. However, because the set of response variables only includes a single dimension, this visualization would only include a single square to denote the `OpticalDensity` dimension. As such, we can proceed directly to combining the predictor and response measurement models into a single path diagram.

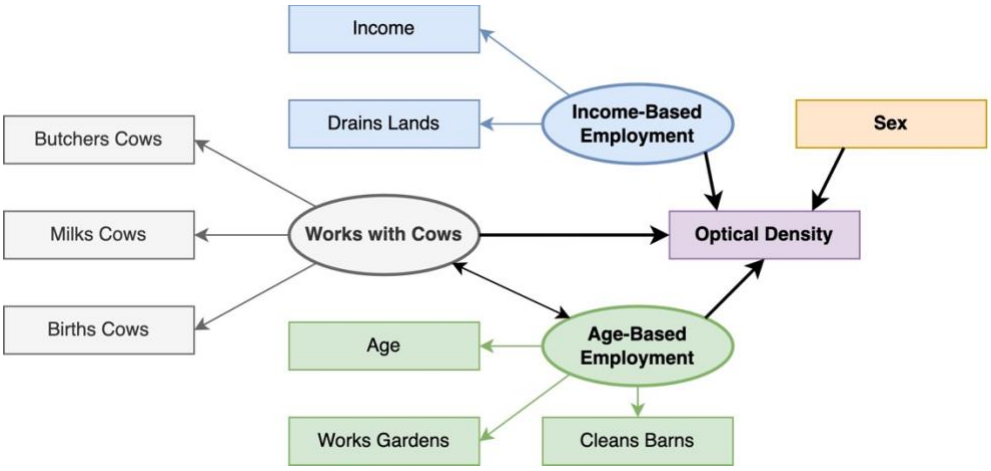


At this stage of the visualization, the response dimension (or set of factors and dimensions) can be placed anywhere, whether to the side or in the middle of the set of predictor variables.

3.1.3 | Defining the Structural Model

Recall that structural equation models can be thought of as the combination of two underlying models that relate different factors and dimensions to each other.

The second of these, known as the *structural model*, is used to define the hypothesized relationships between predictor and response variables. The most common approach is to simply hypothesize a relationship between every pairwise combination of constructs on either side of the predictor and response divide, and then use the subsequent results of the structural equation model to investigate which are significant or substantial and which can be ignored.



Using our example data, we hypothesize that each of our four retained variables have some effect on the single response dimension of Optical Density, as is depicted by lines with a single arrow. Although not overly common, one approach to distinguish these relationships from those between factors and their corresponding dimensions is to bold these new lines.

We also note that relationships are only drawn between factors or dimensions that were retained outside of any factors, which can be distinguished by using bold lettering. Dimensions that comprise existing factors are not considered to have a direct effect on other constructs because any such effect would first pass through the corresponding factor.

3.1.4 | Identifying the Hidden Constructs

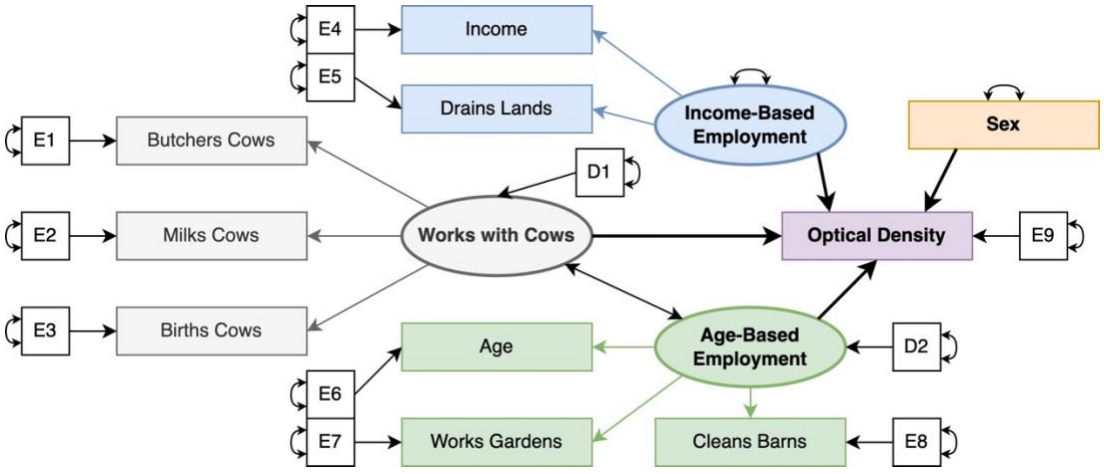
Now that the relationships both within and between the sets of predictor and response variables have been established, the final step is to consider what are sometimes referred to as

the *hidden constructs*. These constructs include both variables and relationships that, although are necessary for the proper parameterization of a structural equation model, are rarely investigated or interpreted in great detail.

Before identifying these hidden constructs, it is important to note that any construct in these path diagrams with an arrow pointing to it is referred to as a *dependent variable* because its value depends on the construct and constructs from which the arrow or arrows originated. Similarly, any construct that only has arrows pointing away from it is referred to as an *independent variable*, because its value is independent of and does not depend on that of any other construct.

The first step in identifying hidden constructs is to assign each dependent variable a construct that points directly to it. It is common practice to refer to those constructs that point at dimensions as *errors* and those that point at factors as *disturbances*, although that is the only distinction between them. These errors and disturbances are meant to indicate that no variable can be predicted perfectly and without error through the use of the variables that point to it, much like the error term used in linear regression.

Next, each independent variable is assigned a *variance* to account for the fact there is likely to be variability in its values, regardless of whether or not they are a dimension or a factor. It is important to note that because error and disturbance terms are technically independent variables, each of them must be assigned a variance, as well.



Each of the lines in this completed path diagram represents a relationship between a dimension, factor, error, or disturbance that can be quantified using a regression coefficient, similar to that used in linear regression or the calculation of factor scores. The lines that double back on themselves similarly represent the variances that also need to be quantified.

In this way, path diagrams provide a detailed visualization of which regression coefficients and variances need to be estimated using structural equation modeling. Precisely how these values are estimated is what we will explore in depth throughout the remainder of this chapter.

Exercises for Section 3.1

3.01 Online Banking Performance (continued). Exercise 2.07 introduced data on the cost efficiency focused operational orientations of 32 banks in Taiwan that were published in 2009 in *Computer and Operations Research*, which are recreated below:

	PC1	PC2
A	0.744	-0.563
B	0.588	-0.221
C	0.756	-0.129
D	0.708	-0.374
AB	0.739	-0.574
BC	0.755	-0.193
AC	0.802	-0.409
AD	0.825	-0.548
BD	0.842	-0.422
CD	0.856	-0.400
ABC	0.783	-0.456
BCD	0.855	-0.405
ABD	0.830	-0.525
ACD	0.850	-0.460
ABCD	0.839	-0.487

The loading matrix above includes the correlations between two extracted factors or principal components (PC) and dimensions related to the estimated efficiency of deposits (A), operational costs (B), number of employees (C), and equipment (D) at each bank in 2005.

a. Construct this loading matrix in R with the proper row and column names
b. Explain which dimensions are complex and, as such, should be retained outside of any factor
c. Explain which factor or factors the remaining non-complex dimensions substantially load on to
d. Use this orthogonally rotated loading matrix to construct a path diagram of the corresponding measurement model, including only non-hidden constructs and relationships – the `include_graphics()` function can be used to insert JPEG and other type of images directly into R

3.02 Cavity Trees, Den Order, and Fishers. Exercise 2.13 introduced data on the characteristics of eight cavity trees used as dens by fishers in Minnesota that were published in 2020 in the *Canadian Journal of Forest Research*. Because female fishers can use up to five different cavities during a single season, sequentially relocating kits among them as they grow bigger, these characteristics can be used as predictors of the `Order` in which a tree was used.

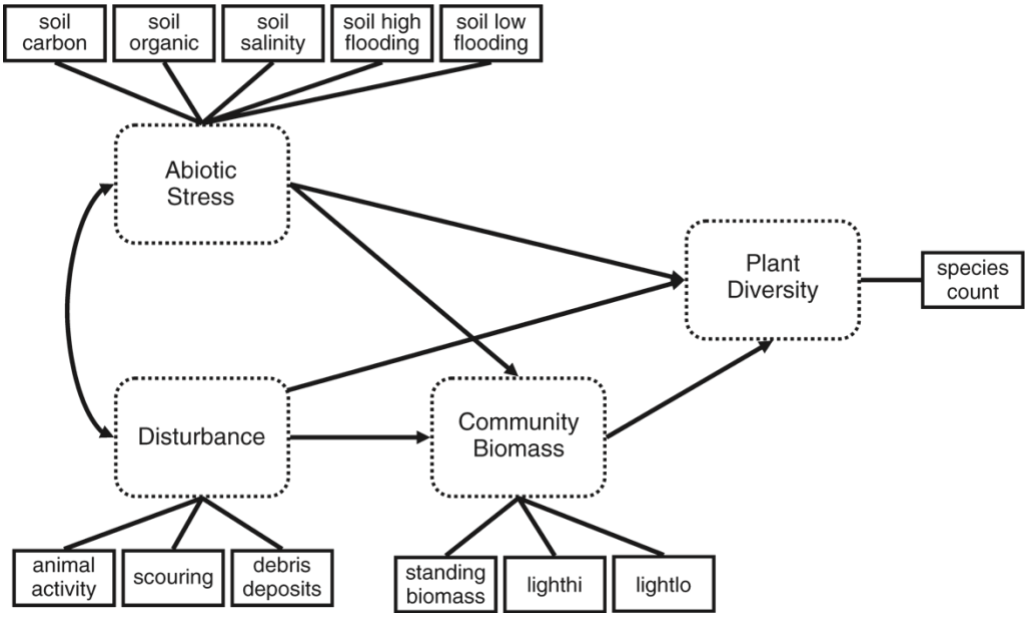
	DBH	Type	Slope	Aspect	Order
DBH	78.45	-2.96	-4.73	-183.75	-1.96
Type	-2.96	0.21	0.62	12.30	0.07
Slope	-4.73	0.62	34.23	-147.01	2.07

Aspect	-183.75	12.30	-147.01	2942.23	-12.38
Order	-1.96	0.07	2.07	-12.38	0.21

The covariance matrix above was calculated using the DBH (diameter at breast height in cm) and Type (0 for coniferous and 1 for deciduous) of the tree itself, as well as the Slope (in degrees from horizontal) and Aspect (in degrees from due South) of the ground the tree is on and the Order in which it was used by fishers at the Camp Ripley military facility in Minnesota.

a. Construct this covariance matrix in R with the proper row and column names
b. Use matrix algebra to find the correlation matrix for the set of predictor dimensions (this should be identical to the correlation matrix used in Exercise 2.12)
c. Find the factor correlation matrix for an oblique rotation of this set of predictor dimensions using the "oblimin" algorithm and an intrinsic dimensionality of two
d. Explain why oblique rotation is not necessary for these data
e. Find the orthogonally rotated loading matrix for this set of predictor dimensions using the "varimax" algorithm and an intrinsic dimensionality of two
f. Use this loading matrix to construct a path diagram of the corresponding structural equation model, including all predictors, responses, and both hidden and non-hidden constructs and relationships

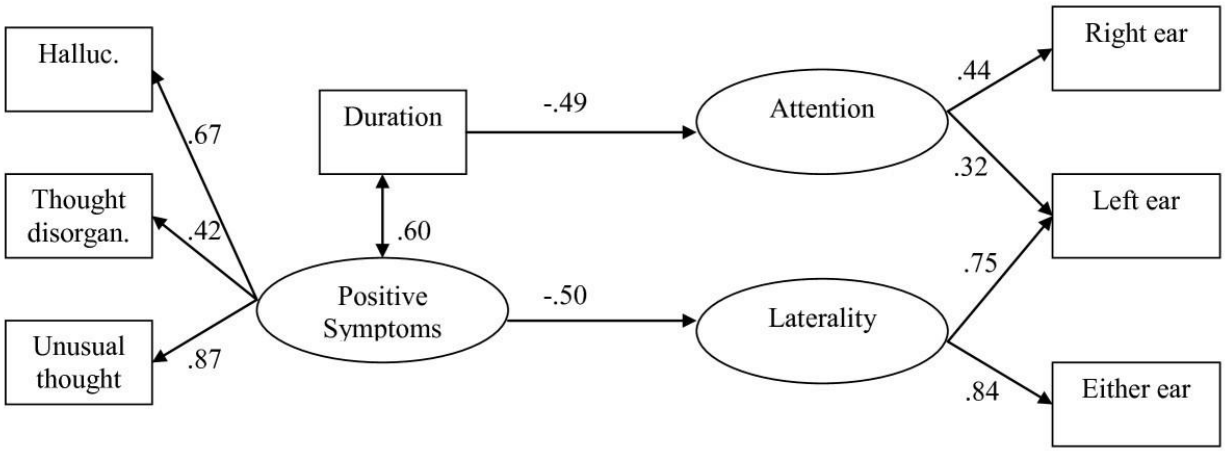
3.03 Plant Diversity and Environmental Factors. Ecological research, especially the study of communities and ecosystems, has been accused of lacking sufficient cohesion to support robust generalizations. Recently, ecologists have become attracted to the possibility that structural equation modeling can be used to address this challenge by providing a way to link specific system attributes to general, theoretical concepts using latent variables. An article published in 2010 in *Ecological Monographs* discuss characteristics of ecological theory and some of the challenges for proper specification of theoretical ideas in structural equation models.



The path diagram above represents major categories of influences on spatial variations in plant diversity based on a theoretical framework developed in 1997.

a. Explain how many dimensions and how many factors are contained in this path diagram
b. Explain how many errors are hidden from this path diagram
c. Explain how many disturbances are hidden from this path diagram
d. Explain how many dependent variables are indicated by this path diagram
e. Explain how many independent variables, whether hidden or not, are indicated by this path diagram
f. Explain how many variances are hidden from this path diagram

3.04 Schizophrenia and Language Processing. Dichotic listening tasks are used as a means of assessing functioning within the left temporal lobe language areas. Previous research suggested increased impairment in left temporal lobe language processing among patients with a high number of positive symptoms (e.g., hallucinations and delusions) of schizophrenia. An article published in 2010 in the *BMC Research Notes* explored how structural equation modeling has been applied in psychiatry to understanding patients' experiences of schizophrenia.



The path diagram above includes the hypothesized relationships among positive symptoms, duration of schizophrenia, and dichotic listening, using a sample of 129 patients from clinics in Norway and California who were diagnosed with schizophrenia.

a. Explain how many errors are hidden from this path diagram
b. Explain how many disturbances are hidden from this path diagram
c. Explain how many variances are hidden from this path diagram
d. Explain how many additional regression coefficients and variances will need to be estimated using structural equation modeling because of these hidden constructs