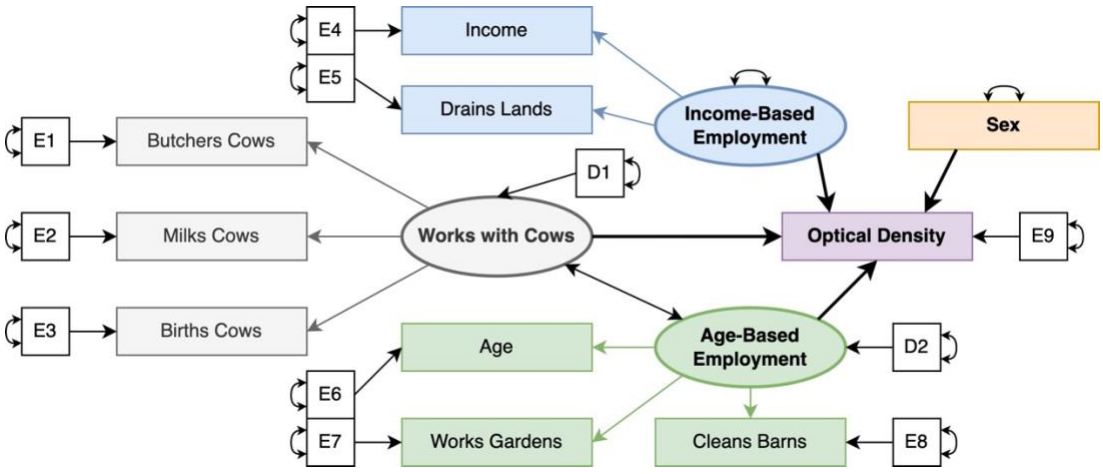
3.2 | Identification and Matrix Representation

After visualizing both the measurement and structural parts of a structural equation model using a path diagram, the next step is to estimate the value of each path in this diagram. These paths include both those that connect one construct to another, referred to as regression coefficients, and those that connect a construct to itself, referred to as variances. In order to estimate each of these parameters, however, we must first ensure that we have sufficient data for proper parameterization.

3.2.1 | Establishing Model Identifiability

The easiest way to determine if there are sufficient data to properly parameterize each regression coefficient and variance required by a structural equation model is to first visualize each of these parameters using a path diagram.

Consider again the path diagram of the relationships between the set of predictor dimensions and factors and the single response dimensions in our example data.



Recall that each line in this path diagram represents a parameter that needs to be estimated through structural equation modeling. Counting them up, we see that our current structural equation model would need to estimate a total of 37 parameters (24 regression coefficients and 13 variances).

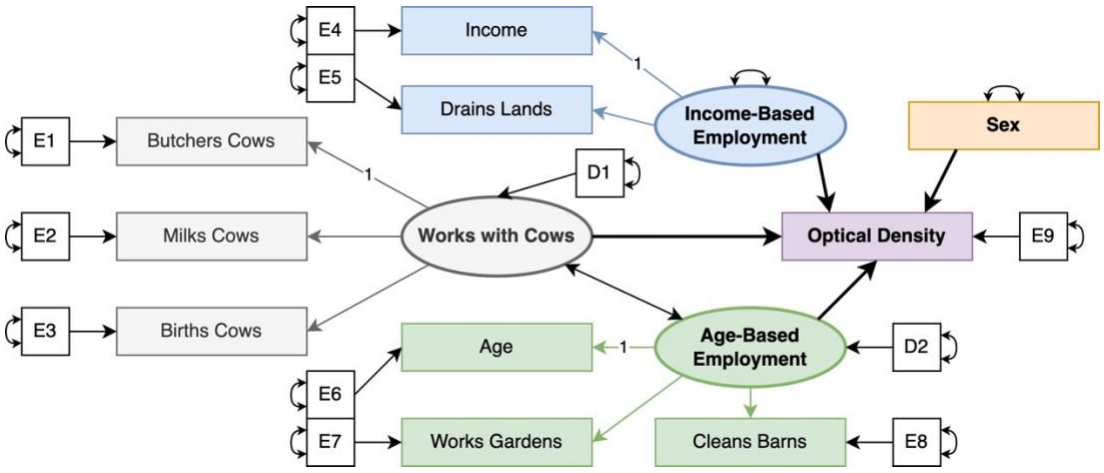
Unfortunately, not all of these parameters can be uniquely estimated. Of particular concern is the need to establish the *scale* of each factor in the measurement model. Because factors are simply linear combinations of dimensions, there are an infinite number of combinations of estimates for paths from a factor to the corresponding dimensions that would result in the same model, so long as they are proportionally identical.

Consider the two paths from IncomeBasedEmployment to Income and DrainsLands. Recall that we can use the matrix of regression coefficients to determine how the values in these two dimensions should be combined to produce the corresponding factor.

```
B <- solve(R_p) %*% A
B
                      TC1         TC3          TC2         TC4
Sex           -0.056204811 -0.04985096  0.097513016  0.87215057
Age           -0.303591104  0.55310759 -0.001547613  0.11745089
Income        -0.015574168  0.06900230  0.660421753  0.22097160
WorksGardens  -0.089056562  0.54421216  0.090432269 -0.24360455
CleansBarns   -0.166728910  0.50275990 -0.111264669  0.07994143
DrainsLands    0.006469611  0.07698805 -0.602763664  0.06935292
ButchersCows   0.437988663 -0.19292313 -0.028058164  0.15935836
MilksCows      0.456495942 -0.13504612  0.123824229 -0.14988388
BirthsCows     0.445746324 -0.13802702 -0.097458688 -0.07799929
```

Using the highlighted in bold values in this matrix, we see that values of this factor can be found by taking the value of Income multiplied by 0.66 and adding it to the value of DrainsLands multiplied by -0.60. It is important to remember, however, that factors are simply mathematical constructs. As such, any two values that are proportionally identical to the values above, such as 0.33 and -0.30 or 0.22 and -0.20, would produce the same relative values of the factor (just divided by a factor of two and three, respectively).

This makes finding a unique set of regression coefficients for the paths between this or any other factor and the corresponding dimensions impossible. To address this issue, we need to establish the scale of this and other factors ahead of time by *fixing* one of the paths from each factor to its corresponding dimensions to a value of 1 (or some other fixed number).



Fixing a parameter prevents it from being estimated by the structural equation model and, as such, should only be used when necessary. Although establishing the scale of a factor in this manner prevents us from finding a unique estimate of it, it does allow us to find a unique solution for the structural equation model as a whole.

After establishing the scale of the three factors in our example model, there are only 34 parameters that remain to be estimated.

Now that we know the total number of parameters that need to be estimated, we need to ensure that we have enough data to properly estimate each of them. Although many statistical models, including regression, use the sample size as a measure of how much data are available to estimate model parameters, structural equation models instead use the number of unique variances and covariances to achieve this goal. Because structural equation modeling uses the covariance structure of a data matrix as the basis for analysis, we can use the number of unique values in this covariance structure to serve as this measure.

```
SIG
                  Sex   Age  Income  WorksGardens CleansBarns
Sex              1.54  0.10    0.10         -0.07        0.21
Age              0.10  0.89    0.02          0.23        0.12
Income           0.10  0.02    8.96          0.45        0.06
WorksGardens    -0.07  0.23    0.45          1.33        0.52
CleansBarns      0.21  0.12    0.06          0.52        1.47
DrainsLands      0.05  0.03   -0.22          0.02        0.09
ButchersCows     0.22  0.08    0.15          0.24        0.22
MilksCows        0.02  0.08    0.22          0.29        0.16
BirthsCows       0.07  0.10    0.12          0.28        0.27
OpticalDensity  -0.13  0.73    0.54          0.66        0.43
                DrainsLands ButchersCows MilksCows BirthsCows
Sex                    0.05         0.22      0.02       0.07
Age                    0.03         0.08      0.08       0.10
Income                -0.22         0.15      0.22       0.12
WorksGardens           0.02         0.24      0.29       0.28
CleansBarns            0.09         0.22      0.16       0.27
DrainsLands            0.26         0.07     -0.01       0.08
ButchersCows           0.07         0.69      0.32       0.40
MilksCows             -0.01         0.32      0.56       0.36
BirthsCows             0.08         0.40      0.36       0.84
OpticalDensity        -0.13         0.16      0.17       0.15
                OpticalDensity
Sex                      -0.13
Age                       0.73
Income                    0.54
WorksGardens              0.66
CleansBarns               0.43
DrainsLands              -0.13
ButchersCows              0.16
MilksCows                 0.17
BirthsCows                0.15
OpticalDensity            6.69
```

Using this variance covariance matrix, we can see that there are a total of 10 variances and 90 covariances among our original 10 dimensions. It is important to note, however, that because this matrix is symmetric across the primary diagonal, only half of the 90 covariances are unique. As such, these data actually provide only 55 unique data points for estimating the parameters in a structural equation model.

Although manually counting the number of unique variances and covariances is always an option, we can simply use the number of dimensions in the original data matrix to find this number of data points:

$$Data\ Points = \frac{d \times (d+1)}{2} \tag{3.1}$$

Now that we have both the number of parameters to be estimated (34) and the number of unique data points provided by the data (55), the next step is to compare them.

If there are more data points than there are parameters that need to be estimated, the model is said to be *over-identified*. If, on the other hand, the number of parameters exceeds the number of data points, the model is said to be *under-identified*, which indicates that there are not enough data to properly estimate each regression coefficient and variance in the model. Finally, if the number of parameters is exactly the same as the number of data points, the model is said to be *just identified*.

In order to properly estimate each parameter in a structural equation model, as well as measures of uncertainty such as standard around these estimates, we require more data points than parameters, which occurs only when the model is over-identified.

Although it is technically possible to estimate each parameter if a model is just identified, it would be impossible to then calculate any measures of uncertainty around these estimates. This is similar to fitting a linear regression line, which includes two parameters of slope and intercept, to just two data points. The resulting equation would be a perfect fit that would go directly through the two data points and, as such, would have no uncertainty around it.

Because these uncertainties are necessary for calculating confidence intervals, conducting hypothesis tests, and determining which predictor variables have a significant effect on the response variables, both in linear regression and in structural equation modeling, our goal is to develop a model that is over-identified.

In our example data, we have more data points (55) than parameters to be estimated (34), indicating that our model is already over-identified. In such situations, we can proceed to the next step of the analysis, which we discuss in Section 3.2.2 below.

On the other hand, if a model is either under-identified or just identified, we can establish identifiability by fixing paths in a sequential manner until the number of data points exceeds the number of parameter paths that remain unfixed and free to be estimated.

Although we could technically fix any paths until the model is over-identified, there are certain paths we should avoid fixing at all costs. Key among these is any regression coefficient connecting predictor variables to response variables in the structural portion of the model. Remember, these regression coefficients will allow us to determine how strong of an effect

predictors have on the responses, which is the original goal of our analysis, so fixing these paths would limit the interpretability and usefulness of our model.

The paths from errors to measured dimensions, on the other hand, while necessary for proper parameter estimation, are not the main focus of the analysis and are rarely discussed or interpreted afterward. As such, these paths are usually the first to be fixed to establish identifiability.

If a model is still not over-identified even after fixing all the paths from each error to its corresponding dimension, the next set of paths to fix are those from disturbances to their corresponding factors.

As the dimensionality of the data increases, the number of data points increases much faster than the number of parameters to be estimated in a structural equation model. As such, it is very rare to need to fix any paths beyond those described here unless the dimensionality is very low. In such situations, structural equation modeling may simply not be the best option.

3.2.2 | Constructing the Two Matrices of Regression Coefficients

Now that we have an over-identified structural equation model, the next step is to represent each regression coefficient and variance using a set of matrices. These matrices can then be used in a series of matrix equations to find estimates for each model parameter, a topic we will discuss in more detail in Sections 3.3 and 3.4. This approach to structural equation modeling is known as the *Bentler-Weeks method* and is by far the most common used strategy for parameter estimation.

We begin with the dependent variables or DVs. Recall that every construct that has a path leading to it constitutes a dependent variable in a structural equation model. We refer to each of these dependent variables as η_i, and the total number of dependent variables as q. In our example path diagram, there are eleven dependent variables (nine of the ten original dimensions and two of the three factors).

We can construct the *matrix of regression coefficients among the DVs* $\boldsymbol{\beta}$ by using each of these variables to define the row and column names of a matrix and the corresponding existence of a path or lack thereof to define the corresponding values in the matrix:

$$\boldsymbol{\beta} = \begin{matrix} \\ \eta_1 \\ \vdots \\ \eta_q \end{matrix} \begin{matrix} \begin{matrix} \eta_1 & \cdots & \eta_q \end{matrix} \\ \begin{bmatrix} 0 & \ldots & * \\ \vdots & \ddots & \vdots \\ 0 & \ldots & 0 \end{bmatrix} \end{matrix} \tag{3.2}$$

It is important to note that the columns represent the starting point of each path, whereas the rows represent the ending point of each path.

If there is no path from one dependent variable to a different dependent variable, the corresponding element in the matrix is recorded as zero, indicating that the first variable has zero effect on the other. If, on the other hand, there is a fixed path from a dependent variable to a different variable, the corresponding element in the matrix is recorded as one (or whichever value the path was fixed to in the path diagram).

Finally, if there is a path from a dependent variable to a different variable but that path is not fixed but instead needs to be estimated using structural equation modeling, the corresponding element in the matrix is recorded using an asterisk. Although other symbols or characters can certainly be used to denote parameter paths that need to be estimated, an asterisk is the most common choice by far.

In R, we can construct this matrix for our example path diagram using the `as.matrix()` function.

```
BT <- matrix(c(  0,  0,  0,  0,  0,  0,  0,  0,  0,  1,  0,
                 0,  0,  0,  0,  0,  0,  0,  0,  0, NA,  0,
                 0,  0,  0,  0,  0,  0,  0,  0,  0, NA,  0,
                 0,  0,  0,  0,  0,  0,  0,  0,  0,  0,  0,
                 0,  0,  0,  0,  0,  0,  0,  0,  0,  0,  0,
                 0,  0,  0,  0,  0,  0,  0,  0,  0,  0,  1,
                 0,  0,  0,  0,  0,  0,  0,  0,  0,  0, NA,
                 0,  0,  0,  0,  0,  0,  0,  0,  0,  0, NA,
                 0,  0,  0,  0,  0,  0,  0,  0,  0, NA, NA,
                 0,  0,  0,  0,  0,  0,  0,  0,  0,  0, NA,
                 0,  0,  0,  0,  0,  0,  0,  0,  0, NA,  0),
             nrow = 11, ncol = 11, byrow = T,
             dimnames = list(c("ButchersCows","MilksCows",
                               "BirthsCows","Income","DrainsLands",
                               "Age","WorksGardens","CleansBarns",
                               "OpticalDensity","WorksWithCows",
                               "AgeBasedEmployment"),
                             c("ButchersCows","MilksCows",
                               "BirthsCows","Income","DrainsLands",
                               "Age","WorksGardens","CleansBarns",
                               "OpticalDensity","WorksWithCows",
                               "AgeBasedEmployment")))
```

Because using an asterisk in R would constitute a non-numerical entry into this matrix, we used `NA` to denote each parameter that needs to be estimated. For instance, the path from `AgeBasedEmployment` to `OpticalDensity` is represented by the `NA` in the eight row tenth column (highlighted in bold above).

Although not strictly necessary, it is common practice to place these variables in the same order as their errors to simplify some of the other matrix construction needed for structural equation modeling. Further, the variables themselves are usually placed in such an order that the dimensions are first, the factors are second, the errors are next, and the disturbances are last.

Next, we turn our attention to the independent variables or IVs. Recall that every construct that has no path leading to it, including errors and disturbances, constitutes an independent variable

in a structural equation model. We refer to each of these independent variables as ζ_i, and the total number of independent variables as r. In our example path diagram, there are twelve independent variables (one of the original dimensions, one of the factors, eight errors, and two disturbances).

We can construct the *matrix of regression coefficients between the IVs and DVs* $\boldsymbol{\gamma}$ by using the independent variables to define the columns of the matrix and the dependent variables to define the rows:

$$\boldsymbol{\gamma} = \begin{matrix} \\ \eta_1 \\ \vdots \\ \eta_q \end{matrix} \begin{matrix} \begin{matrix} \zeta_1 & \cdots & \zeta_r \end{matrix} \\ \begin{bmatrix} 0 & \dots & 0 \\ \vdots & \ddots & \vdots \\ * & \dots & 0 \end{bmatrix} \end{matrix} \tag{3.3}$$

We can again construct this matrix manually in R.

```
GM <- matrix(c(  0,  0, NA,  0,  0,  0,  0,  0,  0,  0,  0,  0,  0,
                 0,  0,  0, NA,  0,  0,  0,  0,  0,  0,  0,  0,  0,
                 0,  0,  0,  0, NA,  0,  0,  0,  0,  0,  0,  0,  0,
                 0,  1,  0,  0,  0, NA,  0,  0,  0,  0,  0,  0,  0,
                 0, NA,  0,  0,  0,  0, NA,  0,  0,  0,  0,  0,  0,
                 0,  0,  0,  0,  0,  0,  0, NA,  0,  0,  0,  0,  0,
                 0,  0,  0,  0,  0,  0,  0,  0, NA,  0,  0,  0,  0,
                 0,  0,  0,  0,  0,  0,  0,  0,  0, NA,  0,  0,  0,
                NA, NA,  0,  0,  0,  0,  0,  0,  0,  0, NA,  0,  0,
                 0,  0,  0,  0,  0,  0,  0,  0,  0,  0,  0, NA,  0,
                 0,  0,  0,  0,  0,  0,  0,  0,  0,  0,  0,  0, NA),
            nrow = 11, ncol = 13, byrow = T,
            dimnames = list(c("ButchersCows","MilksCows",
                              "BirthsCows","Income","DrainsLands",
                              "Age","WorksGardens","CleansBarns",
                              "OpticalDensity","WorksWithCows",
                              "AgeBasedEmployment"),
                            c("Sex","IncomeBasedEmployment",
                              "E1","E2","E3","E4","E5","E6",
                              "E7","E8","E9","D1","D2")))
```

The need to maintain the same order between errors and disturbances and their corresponding dimensions and factors now becomes clear – it helps to ensure that most of the values in this matrix (and the one defined below in Section 3.2.3) line up perfectly along one of the diagonals!

3.2.3 | Constructing the One Matrix of Variances

Now that we have represented all the paths corresponding to regression coefficients in our structural equation model, the next step is to represent the variances in the model, which are depicted as lines that curve back on themselves.

We can construct the *matrix of variances among the IVs* $\boldsymbol{\Phi}$ by using the independent variables to define both the rows and columns:

$$\boldsymbol{\Phi} = \begin{matrix} \\ \zeta_1 \\ \vdots \\ \zeta_r \end{matrix} \begin{matrix} \begin{matrix} \zeta_1 & \cdots & \zeta_r \end{matrix} \\ \begin{bmatrix} * & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & * \end{bmatrix} \end{matrix} \tag{3.4}$$

Because variances are only assigned to independent variables, these are the only variables that we need to consider in this matrix. Further, because a variance can only exist between a variable and itself, only the primary diagonal of this matrix contains non-zero values.

As before, we can manually construct this matrix in R using the `as.matrix()` function.

```
PH <- matrix(c( NA,  0,  0,  0,  0,  0,  0,  0,  0,  0,  0,  0,  0,
                 0, NA,  0,  0,  0,  0,  0,  0,  0,  0,  0,  0,  0,
                 0,  0, NA,  0,  0,  0,  0,  0,  0,  0,  0,  0,  0,
                 0,  0,  0, NA,  0,  0,  0,  0,  0,  0,  0,  0,  0,
                 0,  0,  0,  0, NA,  0,  0,  0,  0,  0,  0,  0,  0,
                 0,  0,  0,  0,  0, NA,  0,  0,  0,  0,  0,  0,  0,
                 0,  0,  0,  0,  0,  0, NA,  0,  0,  0,  0,  0,  0,
                 0,  0,  0,  0,  0,  0,  0, NA,  0,  0,  0,  0,  0,
                 0,  0,  0,  0,  0,  0,  0,  0, NA,  0,  0,  0,  0,
                 0,  0,  0,  0,  0,  0,  0,  0,  0, NA,  0,  0,  0,
                 0,  0,  0,  0,  0,  0,  0,  0,  0,  0, NA,  0,  0,
                 0,  0,  0,  0,  0,  0,  0,  0,  0,  0,  0, NA,  0,
                 0,  0,  0,  0,  0,  0,  0,  0,  0,  0,  0,  0, NA),
            nrow = 13, ncol = 13, byrow = T,
            dimnames = list(c("Sex","IncomeBasedEmployment",
                              "E1","E2","E3","E4","E5","E6",
                              "E7","E8","E9","D1","D2"),
                            c("Sex","IncomeBasedEmployment",
                              "E1","E2","E3","E4","E5","E6",
                              "E7","E8","E9","D1","D2")))
```

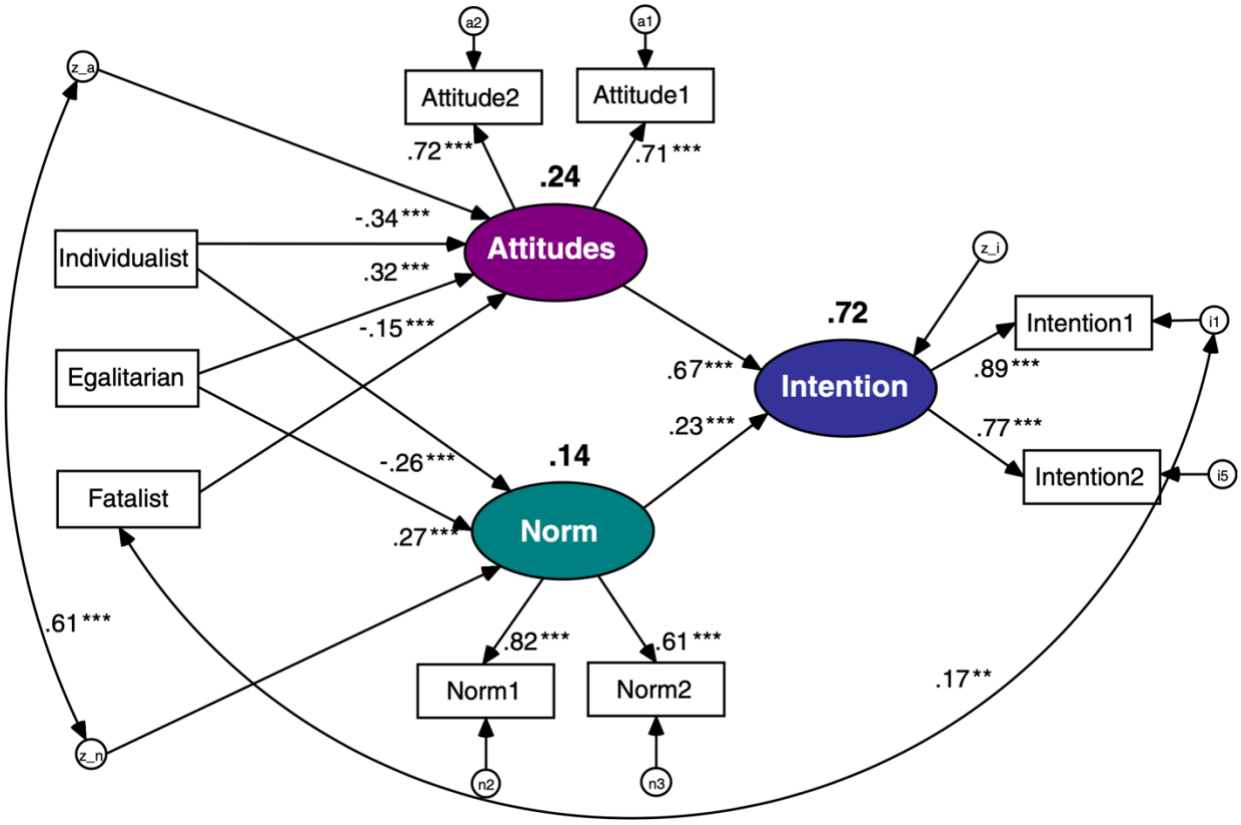
Together, the matrices of regression coefficients and the matrix of variances should represent every single path that was visualized in the path diagram. As such, it is good practice to check that the number of parameters matches the number of non-zero elements in these matrices.

For the time being, we have left the matrix elements corresponding to unfixed paths that will be estimated using structural equation modeling equal to `NA`. It is important to note that this is simply a placeholder and that the goal of structural equation modeling is to actually find the best combination of numerical values to replace these `NA` values.

Although the exact mechanisms of how to find these best estimates will be discussed later in Sections 3.3 and 3.4, it is important to remember that the goal of structural equation modeling is to capture and reconstruct the covariance structure between dimensions. As such, this process will rely heavily on the covariance matrix derived from the original data matrix.

Exercises for Section 3.2

3.05 Climate Change, Theory of Planned Behavior, and Values. A recent development in how societies have delt with climate change is the shift to individuals, indicating that every single person has to bear the responsibility to get involved with the climate change issue. An article published in 2011 in *Climate Change* investigated climate-friendly behavioral intentions and the underlying psychological processes as hypothesized in the Theory of Planned Behavior.



The path diagram above includes the hypothesized relationships between Attitudes and subjective Norms (the predictor variables) and the Intention to use public transport (the response variable) among a sample of 3541 staff and students at a Swiss technical university.

a. Find the number of unique data points provided by the dimensions in this diagram
b. Explain how many additional variances are hidden from this path diagram
c. Explain which paths (if any) need to be fixed to properly establish the scale of each factor
d. Explain which additional paths (if any) need to be fixed to properly establish model identifiability

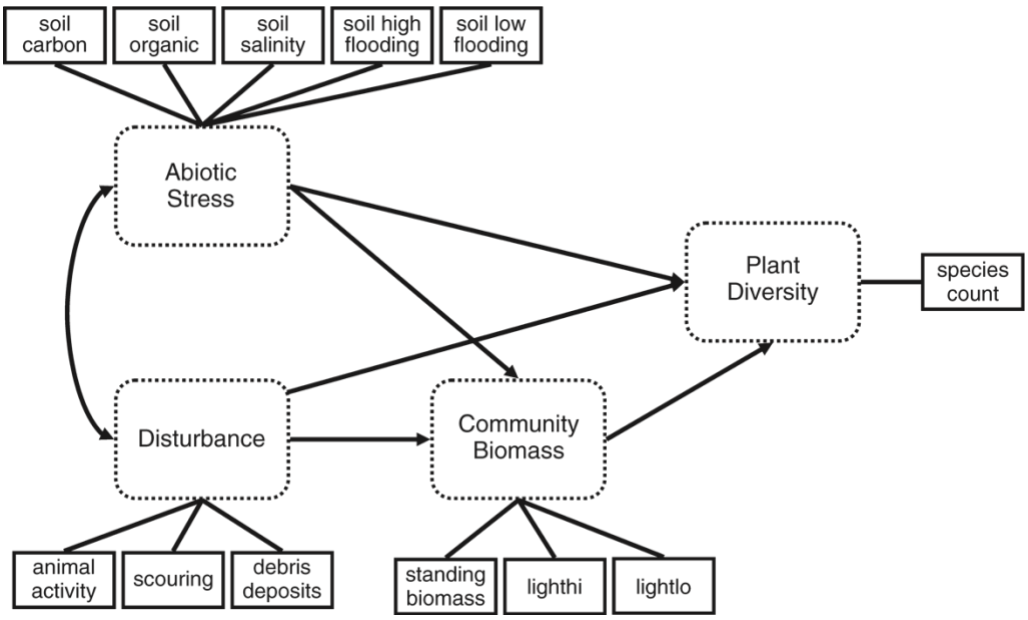
3.06 Cavity Trees, Den Order, and Fishers (continued). Exercise 3.02 introduced additional data on the characteristics of eight cavity trees used as dens by fishers in Minnesota that were published in 2020 in the *Canadian Journal of Forest Research*, which are recreated below:

	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 orthogonally rotated loading matrix for a two-factor representation of the set of predictor dimensions using the `"varimax"` algorithm
c. Use this loading matrix to construct a path diagram of the corresponding structural equation model, including both hidden and non-hidden constructs, as well as fixing any paths needed to properly establish the scale of each factor and model identifiability
d. Use this path diagram to construct the matrix of regression coefficients among the dependent variables
e. Construct the matrix of regression coefficients between the independent and the dependent variables
f. Construct the matrix of variances among the independent variables

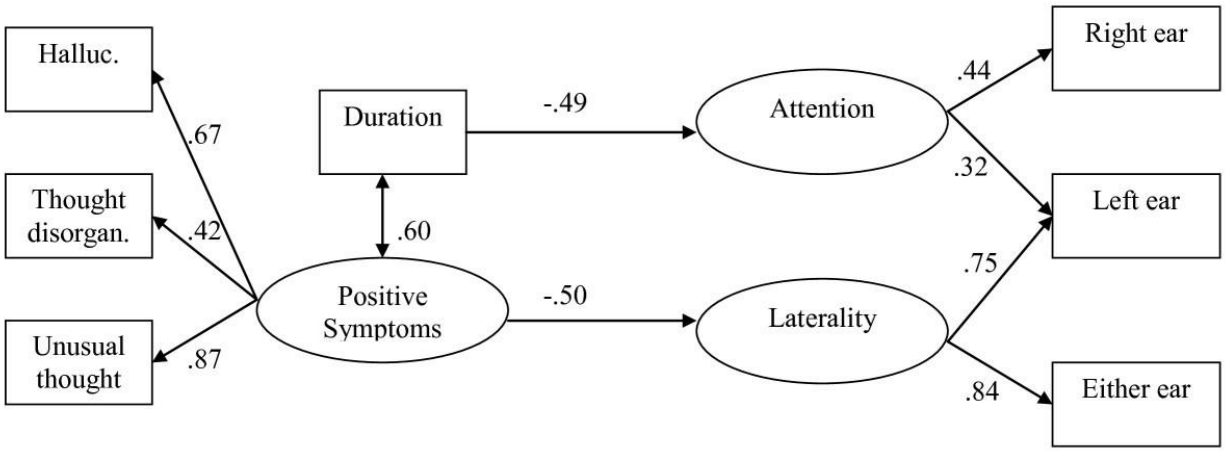
3.07 Plant Diversity and Environmental Factors (continued). Exercise 3.03 introduced data on how structural equation modeling can be used to support robust generalizations that were published in 2010 in *Ecological Monographs*, which are recreated below:



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 which paths need to be fixed to properly establish the scale of each factor
c. Explain how many additional paths that represent regression coefficients are hidden from this path diagram
d. Explain how many additional paths that represent variances are hidden from this path diagram
e. Find the number of unique data points provided by this covariance matrix in terms of variances and covariances
f. Explain if the corresponding structural equation model, without any additional fixing of paths, is over-identified, just identified, or under-identified

3.08 Schizophrenia and Language Processing (continued). Exercise 3.04 introduced data on impairment in left temporal lobe language processing and positive symptoms of schizophrenia that were first published in 2010 in the *BMC Research Notes*, which are recreated below:



The path diagram above includes the hypothesized relationships among PositiveSymptoms, Duration of schizophrenia, and dichotic listening (Attention and Laterality), using a sample of 129 patients from clinics in Norway and California who were diagnosed with schizophrenia.

a. Explain how many hidden and non-hidden independent and dependent variables are in this path diagram
b. Use this path diagram to explain what the row and column names would be in the matrix of regression coefficients among the dependent variables
c. Use this path diagram to explain what the row and column names would be in the matrix of regression coefficients between the independent variables and the dependent variables
d. Use this path diagram to explain what the row and column names would be in the matrix of variances among the independent variables