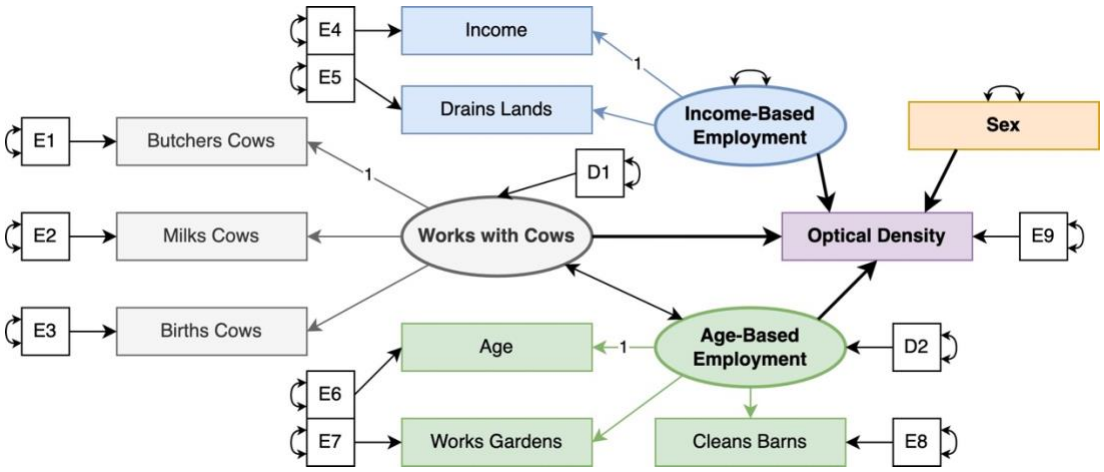
3.4 | Parameter Estimation and Significance

After using the system of matrix equations to reconstruct the covariability structure between the original measured dimensions, the next step is to find the optimal set of model parameters that results in the absolute smallest difference between the observed and reconstructed covariance matrix. This optimized set of parameter estimates can then be investigated to determine which effects are significant, which are strongest, and which should be used in a predictive model of the response dimensions or factors. These estimates can then also be used to construct an updated path diagram that illustrates the magnitude and significance of these different effects and provides a concise visualization of the entire structural equation model.

3.4.1 | Constructing a Structural Equation Model in R

Recall that structural equation modeling is essentially an optimization problem, wherein the goal is to find the set of parameter estimates for each line in the diagram that minimizes the objective function Q that quantifies the differences between the reconstructed and observed covariance matrices.

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.



After fixing three paths to properly establish the scale of each factor, this path diagram has a total of 34 parameters that need to be estimated.

Given the wide use of structural equation modeling in various disciplines, numerous packages have been developed in R to quickly calculate these optimized estimates, as well as a variety of additional metrics and information. The `lavaan` package, for example, includes an `sem()`

function that we can use to define each of the regression coefficients visualized in our path diagram.

Using the `sem()` function begins by specifying a *formula* that describes each of the non-hidden relationships visible in the path diagram. This formula can be separated into three components, the first of which corresponds to the factors constructed in the measurement model.

```
'
 IncomeBasedEmployment =~ Income + DrainsLands
 WorksWithCows =~ ButchersCows + MilksCows + BirthsCows
 AgeBasedEmployment =~ Age + WorksGardens + CleansBarns
'
```

In our example data, the measurement model only included three factors, each of which is recreated within this component of the formula. Note that we have used the `=~` operator to denote which dimensions load on to each individual factor.

The next component of this formula corresponds to the covariances specified in the measurement model.

```
'
 WorksWithCows ~~ AgeBasedEmployment
'
```

In our example data, the only covariability was between two of the factors, and is denoted using the `~~` operator. A similar approach can be taken for any complex dimensions or other instances of covariability depicted in the path diagram.

The third and final component of this formula corresponds to the effects of the set of predictor dimensions and factors on the set of response dimensions and factors, which are identified in the structural model.

```
'
 OpticalDensity ~ Sex + IncomeBasedEmployment + WorksWithCows +
                  AgeBasedEmployment
'
```

In our example data, there was only one dimension that comprised the set of responses, allowing us to define each of the effects in a single line using the `~` operator. If there had been more responses, each of them would have required a similar but separate line of code.

It is important to note that R, like many programming languages, is very sensitive to spelling and cases of variables; as such, it is critical that each of the dimension names used in this formula appear in the exact same manner in the original data matrix or the corresponding covariance matrix, and that the names of each factor appears in the exact same way each time

it is referred to in the formula. Failure to do so is one of the most common causes of errors and warnings when the sem() function is used.

Now that we have identified each of the separate components of this formula, the next step is to combine them into a single object.

```
EQN <- '
        # Measurement Model (Factor Definition)
        IncomeBasedEmployment =~ Income + DrainsLands
        WorksWithCows =~ ButchersCows + MilksCows + BirthsCows
        AgeBasedEmployment =~ Age + WorksGardens + CleansBarns

        # Measurement Model (Covariance Specification)
        WorksWithCows ~~ AgeBasedEmployment

        # Structural Model
        OpticalDensity ~ Sex + IncomeBasedEmployment +
                         WorksWithCows + AgeBasedEmployment
'
```

Although the use of the # operator to provide additional comments and structure to the formula is not strictly necessary, it does make it easier later on to make adjustments or to identify errors within the corresponding structural equation model.

Now that we have a complete formula to describe each of the non-hidden parameters in our path diagram, the final step is to use the sem() function to define the corresponding structural equation model by setting the model argument equal to the equation object defined above.

```
MOD <- sem(model = EQN, sample.cov = SIG, sample.nobs = 303)
Warning message:
In lav_object_post_check(object) :
  lavaan WARNING: some estimated lv variances are negative

MOD
lavaan 0.6-11 ended normally after 113 iterations

  Estimator                                         ML
  Optimization method                           NLMINB
  Number of model parameters                        24

  Number of observations                           303

Model Test User Model:

  Test statistic                                78.786
  Degrees of freedom                                30
  P-value (Chi-square)                           0.000
```

Before moving on to the output itself, it is important to note that the sem() function can be used either with raw data (by setting the data argument equal to the original data matrix) or, as

illustrated above, with the covariance matrix derived from the raw data. This allows for the construction of structural equation models even when the original data matrix is not available (as is often the case with published research and reports), so long as the covariance structure and the corresponding sample size are known (as is also often the case with published research and reports) and specified by the `sample.cov` and `sample.nobs` arguments, respectively.

Note also that unlike many of the functions and examples used previously, this function call produced a *warning* when the model was constructed. In this instance, although the algorithm was able to properly converge to an optimal solution, the corresponding set of parameters included at least one variance that was negative. Recall that variance is used to measure the dispersion of a set of numerical values and, as such, cannot be negative.

In our example, this warning is being caused by largely differing variances in our original dimensions, wherein the variance of `Income` (8.96) is over thirty times that of `DrainsLands` (0.26). In most instances, this issue, which is known as a *Heywood case*, is relatively minor and will not substantially alter model results or estimates; however, simply standardizing or scaling the original data matrix prior to analysis (assuming it is available) can help to alleviate this.

Moving on, the output itself includes a number of interesting pieces of information. First is the `$estimator` value, which identifies which approach to finding the best set of parameter estimates was used. By default, the `sem()` function uses maximum likelihood or `ML` estimation although, as we briefly mentioned earlier in Section 3.3, other approaches may also be used. These can be specified using the `estimator` argument in the function call.

The next line of output specifies which algorithm was used to find the optimal solution under the chosen estimation approach. In our example data, the *non-linear minimization with box constraints* or *NLMINB* algorithm was used and was able to find an optimal solution after 113 iterations. The next two lines of code simply list the number of parameters that were estimated by the model and the total sample size on which this estimation was based.

The final segment of output involves several metrics that are used to evaluate how well this optimized model actually fits the data, a topic we will explore in more detail in Section 3.5.

3.4.2 | Extracting Parameter Estimates

Now that we have constructed the structural equation model as an object in R, we can use this object to extract the optimized parameter estimates that together produce the smallest difference between the observed and reconstructed covariance matrices.

Much like using `summary()` for linear regression models built using the `lm()` function, we can use the `parameterEstimates()` function in R to extract the estimates, standard errors, and associated metrics for each parameter in a structural equation model built using `sem()`.

```
parameterEstimates(MOD)
```

	lhs	op	rhs	est	se	z	pvalue	ci.lower	ci.upper
1	IncomeBasedEmployment	=~	Income	1.000	0.000	NA	NA	1.000	1.000
2	IncomeBasedEmployment	=~	DrainsLands	0.425	0.380	1.118	0.264	-0.320	1.170
3	WorksWithCows	=~	ButchersCows	1.000	0.000	NA	NA	1.000	1.000
4	WorksWithCows	=~	MilksCows	0.898	0.093	9.664	0.000	0.716	1.080
5	WorksWithCows	=~	BirthsCows	1.135	0.116	9.756	0.000	0.907	1.363
6	AgeBasedEmployment	=~	Age	1.000	0.000	NA	NA	1.000	1.000
7	AgeBasedEmployment	=~	WorksGardens	2.385	0.552	4.323	0.000	1.304	3.466
8	AgeBasedEmployment	=~	CleansBarns	1.915	0.460	4.165	0.000	1.014	2.816
9	WorksWithCows	~~	AgeBasedEmployment	0.109	0.029	3.801	0.000	0.053	0.165
10	OpticalDensity	~	Sex	-0.092	0.114	-0.806	0.420	-0.316	0.132
11	OpticalDensity	~	IncomeBasedEmployment	0.543	0.356	1.525	0.127	-0.155	1.241
12	OpticalDensity	~	WorksWithCows	-0.776	0.467	-1.662	0.097	-1.691	0.139
13	OpticalDensity	~	AgeBasedEmployment	3.475	1.173	2.962	0.003	1.176	5.775
14	Income	~~	Income	9.446	0.939	10.056	0.000	7.605	11.288
15	DrainsLands	~~	DrainsLands	0.352	0.102	3.457	0.001	0.153	0.552
16	ButchersCows	~~	ButchersCows	0.335	0.040	8.348	0.000	0.256	0.414
17	MilksCows	~~	MilksCows	0.274	0.033	8.399	0.000	0.210	0.337
18	BirthsCows	~~	BirthsCows	0.383	0.049	7.843	0.000	0.287	0.479
19	Age	~~	Age	0.783	0.069	11.422	0.000	0.649	0.918
20	WorksGardens	~~	WorksGardens	0.735	0.112	6.552	0.000	0.515	0.955
21	CleansBarns	~~	CleansBarns	1.085	0.110	9.856	0.000	0.869	1.300
22	OpticalDensity	~~	OpticalDensity	5.791	0.611	9.473	0.000	4.593	6.989
23	IncomeBasedEmployment	~~	IncomeBasedEmployment	-0.516	0.480	-1.074	0.283	-1.458	0.426
24	WorksWithCows	~~	WorksWithCows	0.353	0.057	6.186	0.000	0.241	0.464
25	AgeBasedEmployment	~~	AgeBasedEmployment	0.104	0.042	2.448	0.014	0.021	0.187
26	IncomeBasedEmployment	~~	WorksWithCows	0.114	0.095	1.208	0.227	-0.071	0.300
27	IncomeBasedEmployment	~~	AgeBasedEmployment	0.063	0.055	1.146	0.252	-0.045	0.171
28	Sex	~~	Sex	1.535	0.000	NA	NA	1.535	1.535

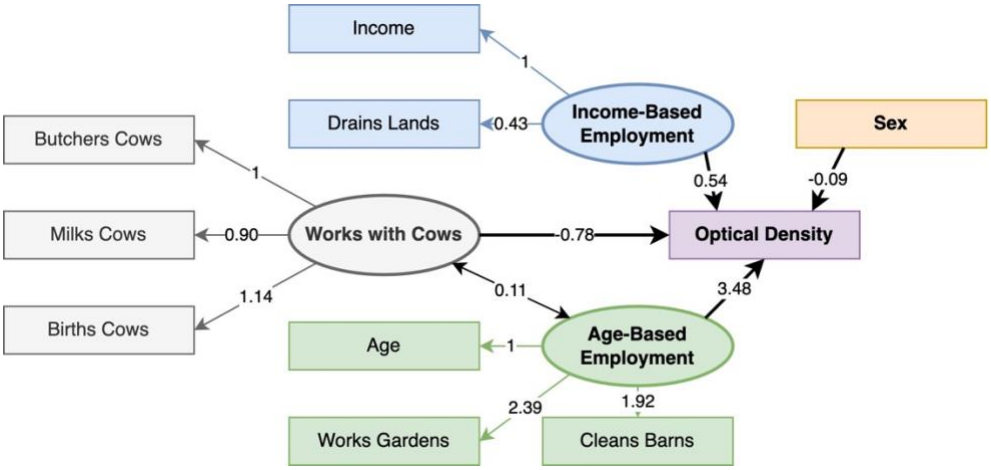
The first thing to note is that, in addition to providing estimates for each of the non-hidden parameters identified in our formula, this function also provides estimates for several hidden variances in the model. Because these variances are rarely investigated or interpreted, we will use matrix algebra to isolate the output only to the regression coefficients and covariances specified in our original formula.

```
parameterEstimates(MOD)[1:13,]

                     lhs op                   rhs    est    se
1  IncomeBasedEmployment =~                Income  1.000 0.000
2  IncomeBasedEmployment =~           DrainsLands  0.425 0.380
3           WorksWithCows =~          ButchersCows  1.000 0.000
4           WorksWithCows =~             MilksCows  0.898 0.093
5           WorksWithCows =~            BirthsCows  1.135 0.116
6      AgeBasedEmployment =~                   Age  1.000 0.000
7      AgeBasedEmployment =~           WorksGardens  2.385 0.552
8      AgeBasedEmployment =~            CleansBarns  1.915 0.460
9           WorksWithCows ~~    AgeBasedEmployment  0.109 0.029
10         OpticalDensity  ~                   Sex -0.092 0.114
11         OpticalDensity  ~ IncomeBasedEmployment  0.543 0.356
12         OpticalDensity  ~         WorksWithCows -0.776 0.467
13         OpticalDensity  ~    AgeBasedEmployment  3.475 1.173

         z pvalue ci.lower ci.upper
1       NA     NA    1.000    1.000
2    1.118  0.264   -0.320    1.170
3       NA     NA    1.000    1.000
4    9.664  0.000    0.716    1.080
5    9.756  0.000    0.907    1.363
6       NA     NA    1.000    1.000
7    4.323  0.000    1.304    3.466
8    4.165  0.000    1.014    2.816
9    3.801  0.000    0.053    0.165
10 -0.806  0.420   -0.316    0.132
11  1.525  0.127   -0.155    1.241
12 -1.662  0.097   -1.691    0.139
13  2.962  0.003    1.176    5.775
```

Using the same operators as before to differentiate factor loadings, oblique or complex covariances, and the effects of predictors on responses, we see that Sex and WorksWithCows have negative effects on OpticalDensity (an indicator of antibody concentration and infection), whereas IncomeBasedEmployment and AgeBasedEmployments have positive effects.

Note that, similar to the output of a regression model, each of these estimates is accompanied by a standard error that, among other things, provides a measure of how far the estimated effect size is likely to be from the actual effect size throughout the entire population of interest. As we will see in Section 3.4.3 below, these standard errors can be used to perform hypothesis tests regarding and construct confidence intervals surrounding each parameter estimate.

Now that we have an estimate for each of our parameters, it is common practice to update our original path diagram to include these estimates for each non-hidden path, line, and arrow.



Before moving on, note that even without specifying in the formula that at least one path from each of the three factors to one of their corresponding dimensions had to be fixed in order to properly establish the scale of each factor, the resulting parameter estimates still accounted for this need (by fixing whichever constituent dimension was listed first in the equation).

Now that we have these estimates, we can interpret each of the regression coefficients (but not the covariances) in much the same way as we would any linear regression coefficient.

For instance, as AgeBasedEmployment (a linear combination of Age, WorksGardens, and CleansBarns) increases by one unit, OpticalDensity increases by 3.48. This would indicate that individuals who are older, who work in gardens, or who clean barns are at a substantially higher risk of contracting toxoplasmosis, all else held constant. Taking a closer look at how strongly these dimensions are related to their corresponding factor, we can see that AgeBasedEmployment is driven much stronger by WorksGardens than either of the other dimensions. As such, it stands to reason that working in gardens is comparatively a much greater risk factor than either of the other two dimensions.

Similarly, as Sex increases by one unit, OpticalDensity decreases by 0.09. Given that this binary dimension was coded such that females were represented as zeroes and males as ones, this would indicate that being male slightly decreases the risk of contracting toxoplasmosis, all else held constant.

Unfortunately, simply knowing the magnitude of an effect or relationship is not enough, because each of these estimates is, by definition, a function of the units in which the original dimensions were measured. For instance, if the frequency with which a study participant MilksCows was measured on a yearly rather than a monthly scale, the corresponding effect size would increase by a factor of twelves (mirroring the fact that there are twelve months in a year).

As such, in addition to interpreting the magnitude of each variable's effect on the response variable or variables, it is important to also determine if the effect is significant.

3.4.3 | Determining Effect Significance

Much like in linear regression, determining whether or not a model coefficient is significant is done using a simple *hypothesis test*.

Recall that the *null hypothesis* in such a test states that, on average, there is no effect of the predictor variable on the response variable from the perspective of the population. The *alternative hypothesis*, on the other hand, is that there is some effect. Using the available sample data, we can calculate a *test statistic* to compare the validity of these competing hypothesis and then use it to find the corresponding *p-value*. By comparing this p-value to a *critical alpha-level*, we can determine whether to *reject* or *fail to reject* the null hypothesis.

Fortunately, all of the numerical aspects of such a hypothesis test are already presented in the output of the `parameterEstimates()` function.

```
parameterEstimates(MOD)[10,]
              lhs op rhs    est    se      z pvalue
10 OpticalDensity  ~ Sex -0.092 0.114 -0.806   0.42
                                  ci.lower ci.upper
10                                  -0.316    0.132
```

For instance, the test statistic (also known as a *z-statistic* because of its derivation from the equation for a standardized z-score) for a hypothesis test of the effect of `Sex` on `OpticalDensity` is -0.806, which corresponds to a p-value of 0.42. Using a common alpha-level of 0.05, we see that this p-value is not less than alpha and, as such, we would fail to reject the null hypothesis.

This would lead us to conclude that there is no evidence that `Sex` has an effect on `OpticalDensity` throughout the population. Essentially, the negative effect observed in our sample data can be ascribed to *sampling variability*, which refers to the fact that the any statistical information derived from sample data (in this instance, the effect of `Sex` on `OpticalDensity`) will vary from one random sample to another.

If, on the other hand, the p-value is below the alpha-level, we would reject the null hypothesis and conclude that the corresponding variable does have a significant effect on `OpticalDensity`. Using again an alpha-level of 0.05, the only variable that appears to have such a significant effect on our response variable is `AgeBasedEmployment`.

Although not strictly necessary, many researchers will then use the results of these hypothesis tests to bold, highlight, or otherwise accentuate significant effects in the original path diagram.

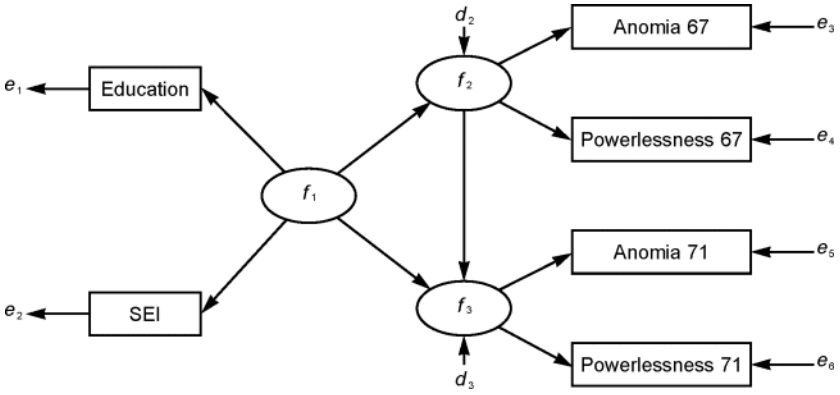
One last thing to note is that the `parameterEstimates()` function also provides the lower and upper bound for a *95% confidence interval* of the effect of each variable on the response. However, we will not discuss the interpretation of such intervals here, instead encouraging interested readers to review how similar intervals are interpreted in linear regression.

Exercises for Section 3.4

3.13 The Stability of Alienation. Feelings and experiences of alienation are of growing concern in many countries around the world. But are these feelings of alienation highly volatile and, therefore, amenable to change, or are they relatively stable over time? An article published in 1977 in *Sociological Methodology* investigated how alienation changed over time and how this change was influenced by education and socioeconomic status.

	Anomia 67	Powerlessness 67	Anomia 71	Powerlessness 71	Education	SEI
Anomia 67	11.834	6.947	6.819	4.783	-3.839	-2.190
Powerlessness 67	6.947	9.364	5.090	5.028	-3.889	-1.883
Anomia 71	6.819	5.090	12.532	7.495	-3.841	-2.175
Powerlessness 71	4.783	5.028	7.495	9.986	-3.625	-1.878
Education	-3.839	-3.889	-3.841	-3.625	9.610	3.552
SEI	-2.190	-1.883	-2.175	-1.878	3.552	4.503

The covariance matrix above was calculated using the anomia and powerlessness subscales of alienation that were measured in 1967 and again in 1971 on $N = 932$ people from Illinois, as well as their education (years of schooling completed) and Duncan's Socioeconomic Index (SEI).



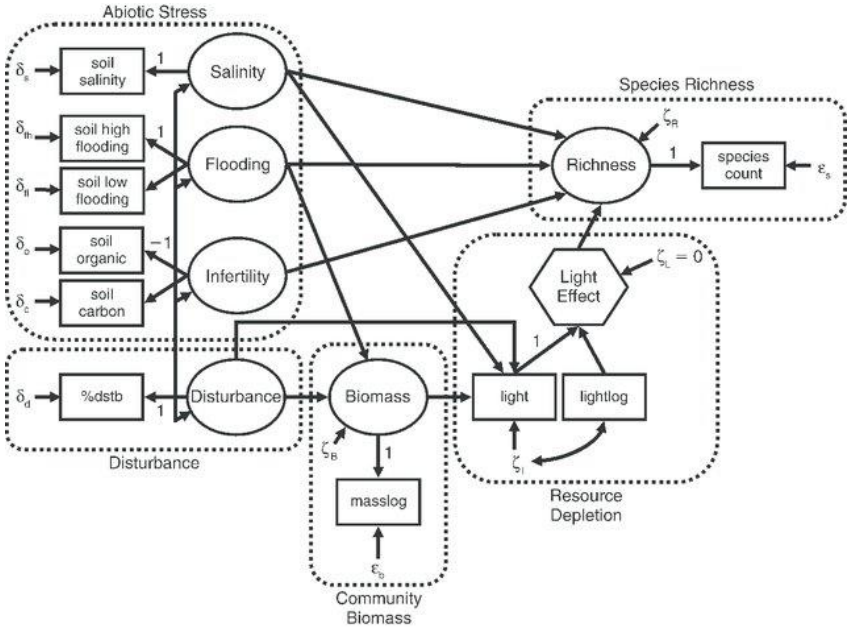
The path diagram above represents the theoretical framework for this analysis.

a. Construct this covariance matrix in R with the proper row and column names
b. Use the `sem()` function to define a structural equation model based on this covariance matrix and the non-hidden elements in the corresponding path diagram
c. Find the resulting optimized estimates for each parameter in the path diagram
d. Explain if `f1` has a positive or negative effect on the alienation factor from 1967
e. Explain if `f1` has a stronger effect on alienation in 1967 or on alienation in 1971
f. Explain if alienation in 1967 has a significant effect on alienation in 1971 and, by extension, that alienation is a relatively stable phenomenon over time

3.14 Stress, Disturbance, and Species Richness. Exercise 3.03 introduced data on the relative importance of abiotic conditions, disturbance, and community biomass on plant species richness in a coastal wetland, which were first published in 1997 in *The American Naturalist*.

	1	2	3	4	5	6	7	8	9	10
1. light log	1.00	.858	.667	-.251	-.699	.060	.012	.552	.547	.327
2. light	.858	1.00	.776	-.404	-.794	.157	.120	.439	.462	.321
3. %dstb	.667	.776	1.00	-.228	-.686	.218	.186	.249	.290	.216
4. species count	-.251	-.404	-.228	1.00	.291	.119	.132	-.374	-.406	-.292
5. mass log	-.699	-.794	-.686	.291	1.00	-.096	-.071	-.426	-.466	-.138
6. soil carbon	.060	.157	.218	.119	-.096	1.00	.973	-.170	-.150	.249
7. soil organic	.012	.120	.186	.132	-.071	.973	1.00	-.211	-.188	.244
8. soil low flooding	.552	.439	.249	-.374	-.426	-.170	-.211	1.00	.959	.073
9. soil high flooding	.547	.462	.290	-.406	-.466	-.150	-.188	.959	1.00	.052
10. soil salinity	.327	.321	.216	-.292	-.138	.249	.244	.073	.052	1.00

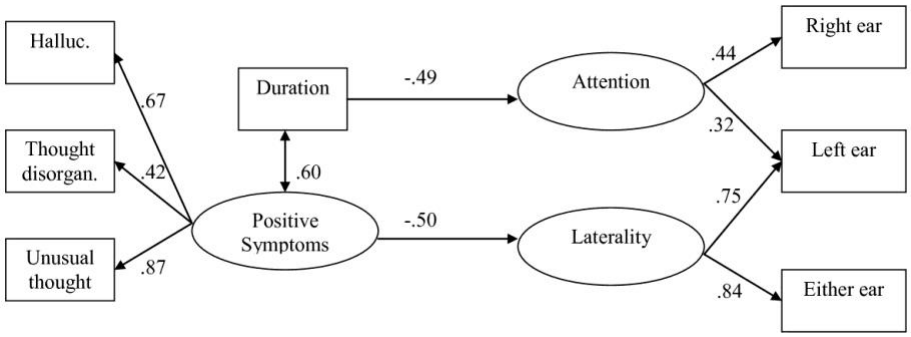
The standardized covariance matrix above was calculated using data on $N = 190$ field plots throughout coastal marsh communities of the northern Gulf of Mexico.



The path diagram above represents the direct effects of four predictive factors on a single response factor, as illustrated in an article published in 2010 in *Ecological Monographs*.

a. Construct this covariance matrix in R with the proper row and column names
b. Use the `sem()` function to define a structural equation model based on this covariance matrix and the non-hidden elements in the corresponding path diagram (note that the arrows from `LightEffect` to `light` and `lightlog` are incorrectly reversed)
c. Find the resulting optimized estimates for each parameter in the path diagram
d. Explain which of predictive factor had the most significant direct effect on `Richness`

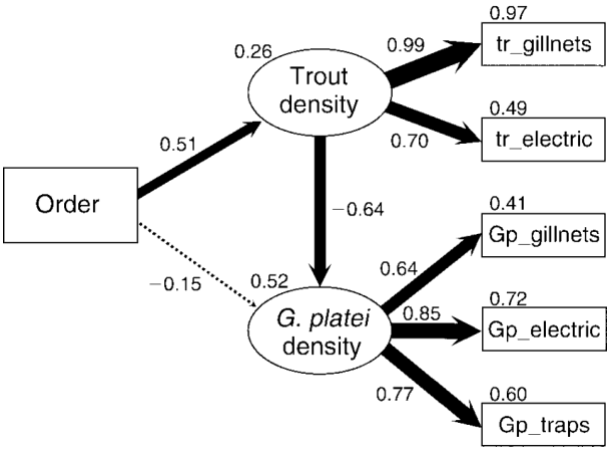
3.15 Schizophrenia and Language Processing (continued). Exercise 3.04 introduced data on impairment in left temporal lobe language processing and positive symptoms of schizophrenia that were first published in 2010 in the *BMC Research Notes*, which are recreated below:



The path diagram above includes the hypothesized relationships among 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. Find the number of rows and columns in the resulting reconstructed covariance matrix
b. Explain if the path diagram includes any instances of structural covariability
c. Explain which dimension was most indicative of PositiveSymptoms of schizophrenia
d. Explain if LeftEar was more indicative of Attention or Laterality
e. Explain if PositiveSymptoms had a positive or negative effect on Laterality
f. Explain if this supports the original framework of the article that there is "increased impairment in left temporal lobe language processing among schizophrenia patients"

3.16 Invasive Salmonids, Lake Order, and Galaxiid Decline (continued). Exercise 3.09 introduced data on the effects of hydrological position and trout density on the density of galaxiids that were published in 2012 in *Ecological Applications*, which are recreated below:



The path diagram above represents an optimized set of path parameters based on a sample of 25 lakes in the Aysén region of Chile that were surveyed in 2007 and 2009. Note that the number next to each construct represents the path of the corresponding error or disturbance.

a. Explain which metric of galaxiid density was the strongest indicator of G.plateiDensity
b. Explain if TroutDensity had a positive or negative effect on G.plateiDensity
c. Explain if TroutDensity had a significant effect on G.plateiDensity
d. Explain if Order or TroutDensity had a stronger direct effect on G.plateiDensity