Path Analysis Introduction and Example

Joel S Steele, PhD

Winter 2017

Path Analysis

Model specification

There two main ways of communicating the system of equations that represents a theoretical model. Either with a set of simultaneous equations, or with a path diagram. Below we explore both and provide an example.

Path Model Assumptions

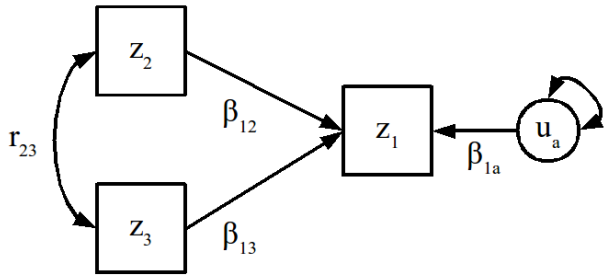
For this example we will be accepting a number of assumptions.

1. All causal relations are *linear* and *additive*
2. All models are recursive
   - results in uncorrelated error terms
   - no *two-way* causal relations
   - no feedback loops
3. Error terms are uncorrelated with other independent variables
4. There is a *weak* causal ordering
5. Causal closure, meaning all of the relevant causal variables are included in the model

If these assumptions are met, then we can use least squares regression for our estimation. In what follows we will be fitting our model to the standardized data.

A simple example

Figure 1: Regression path diagram



Based on Figure 1 we have a simple multiple regression, this is not any more difficult than what we have seen previously. Everything that we know from multiple regression should replicate in this situation. However, there is another aspect to this illustration that is important, namely that the goal of path modeling, and the multivariate extensions such as SEM and latent variable modeling, is to reproduce the variance-covariance matrix of the variables included. In this example we will be using z-scores, so we will be interested in reproducing the correlation matrix among the variables z_1, z_2, and z_3.

Simultaneous equation modeling approach

The equation that represents the path model above in Figure 1 can be expressed as,

$$z_{1i} = \beta_{12} z_{2i} + \beta_{13} z_{3i} + \beta_{1a} u_{ai}. \tag{1}$$

In our following steps we will work to compute the correlations among each of the variables, based on the model. That is, we will compute the correlations using Equation 1 above to see how each relations is decomposed based on our theoretical arrangement.

Correlation r_{12}

In order to compute the model-based expected correlation between z_1 and z_2 we will multiply both sides of the equation by z_2 and simplify.

$$\begin{aligned} \tfrac{1}{n}\Sigma z_{1i} z_{2i} &= \tfrac{1}{n}\Sigma\beta_{12} z_{2i} z_{2i} + \tfrac{1}{n}\Sigma\beta_{13} z_{3i} z_{2i} + \tfrac{1}{n}\Sigma\beta_{1a} u_{ai} z_{2i} \\ \tfrac{1}{n}\Sigma z_{1i} z_{2i} &= \beta_{12}\tfrac{1}{n}\Sigma z_{2i} z_{2i} + \beta_{13}\tfrac{1}{n}\Sigma z_{3i} z_{2i} + \beta_{1a}\tfrac{1}{n}\Sigma u_{ai} z_{2i} \\ r_{12} &= \beta_{12}(1) + \beta_{13} r_{23} + \beta_{1a} r_{a2} \end{aligned}$$

It is important to note that, by assumption errors are uncorrelated with all other predictors, thus $r_{a2} = 0$. Making this substitution we obtain,

$$r_{12} = \beta_{12} + \beta_{13} r_{23} \tag{2}$$

represents our model based estimation of the correlation between z_1 and z_2.

Correlation r_{13}

In order to compute the model-based expected correlation between z_1 and z_3 we will multiply both sides of the equation by z_3 and simplify.

$$\begin{aligned} \tfrac{1}{n}\Sigma z_{1i} z_{3i} &= \tfrac{1}{n}\Sigma\beta_{12} z_{2i} z_{3i} + \tfrac{1}{n}\Sigma\beta_{13} z_{3i} z_{3i} + \tfrac{1}{n}\Sigma\beta_{1a} u_{ai} z_{3i} \\ r_{13} &= \beta_{12} r_{23} + \beta_{13}(1) + 0 \\ r_{13} &= \beta_{12} r_{23} + \beta_{13} \end{aligned} \tag{3}$$

Parameter estimation of β_{12} and β_{13}

Now that we have the model implied correlations for both r_{12} and r_{13}, we can focus on the estimation of the parameters β_{12} and β_{13}. Starting from the model implied relations among the variables, the estimation of these parameters can be expressed using our earlier solutions in equations 2 and 3.

To begin, we will focus on the estimation of β_{12}. Our first step is to solve for the parameter β_{13} from equation 3. We do this in order to get an equation that expresses β_{13} in terms of β_{12}, we will need this to solve for β_{12}.

$$\begin{aligned} r_{13} &= \beta_{12} r_{23} + \beta_{13} \\ \beta_{13} &= r_{13} - \beta_{12} r_{23} \end{aligned}.$$

Substituting this expression into equation 2 we obtain,

$$\begin{aligned} r_{12} &= \beta_{12} + (r_{13} - \beta_{12} r_{23}) r_{23} \\ r_{12} &= \beta_{12} + r_{13} r_{23} - \beta_{12} r_{23}^2 \\ r_{12} - r_{13} r_{23} &= \beta_{12} - \beta_{12} r_{23}^2 \\ r_{12} - r_{13} r_{23} &= \beta_{12}(1 - r_{23}^2) \end{aligned} ,$$

$$\beta_{12} = \frac{r_{12} - r_{13} r_{23}}{1 - r_{23}^2} \tag{4}$$

A similar process can be performed for the estimation of β_{13}.

Standard Error of Estimation

Finally, we will solve for the model based correlation of z_{1i} with itself. We multiply through our structural equation by z_{1i},

$$\begin{aligned} \tfrac{1}{n}\Sigma z_{1i} z_{1i} &= \tfrac{1}{n}\Sigma \beta_{12} z_{2i} z_{1i} + \tfrac{1}{n}\Sigma \beta_{13} z_{3i} z_{1i} + \tfrac{1}{n}\Sigma \beta_{1a} u_{ai} z_{1i} \\ 1 &= \beta_{12} r_{12} + \beta_{13} r_{13} + \beta_{1a} r_{1a} \\ \beta_{1a} r_{1a} &= 1 - (\beta_{12} r_{12} + \beta_{13} r_{13}) \end{aligned}$$

Recall that the multiple R^2 for a model is equal to $\Sigma_{p=1}^{k} \beta_{yp} r_{yp}$, where k is the number of predictors for the variable y. In our above equation this translates to $R^2 = \beta_{12} r_{12} + \beta_{13} r_{13}$, thus we can express the above equation as,

$$\beta_{1a} r_{1a} = 1 - R^2. \tag{5}$$

You may also notice that since u_{ai} is uncorrelated with any other predictor, the correlation $r_{1a} = \beta_{1a}$. This results in our final expression of the equation 5,

$$\begin{aligned} \beta_{1a}^2 &= 1 - R^2 \\ \beta_{1a} &= \sqrt{1 - R^2}. \end{aligned} \tag{6}$$

This last expression is our standard error of the estimate from the model.

Data Example

Motivation

The difference from what we have seen before is that now we are considering multiple equations with multiple outcomes possible. Note that each equation is still for a single outcome, but we can consider the entire system of equations. This allows us to not only see the influence of other inputs on relations among predictors and outcomes, as with **Moderation**, in this framework we are interested in the possible mechanisms of causation. These causal relations can be either *direct* or *indirect* meaning that they can operate through other variables.

These data represent a subset of 62 academic professionals who were measured on a number of variables including:

- *sex* : Biological sex of respondent (*male*=1)
- *time* : Time, in years, since earning their PhD
- *pub* : Number of publications
- *cit* : Number of citations
- *salary* : Annual salary in dollars

Table 1: Descriptive statistics

	mean	sd	min	max	range	se
time	6.790	4.278	1	21	20	0.543
pub	18.177	14.004	1	69	68	1.779
sex	0.565	0.500	0	1	1	0.063
cit	40.226	17.172	1	90	89	2.181
salary	54815.758	9706.023	37939	83503	45564	1232.666

Below we present a path diagram in Figure 2, as well as the mathematical specification of the system of equations in Equation 7.

Zero-order correlations

It is always informative to look at the raw associations among the variables before any modeling is proposed. Below is the correlation table for these data.

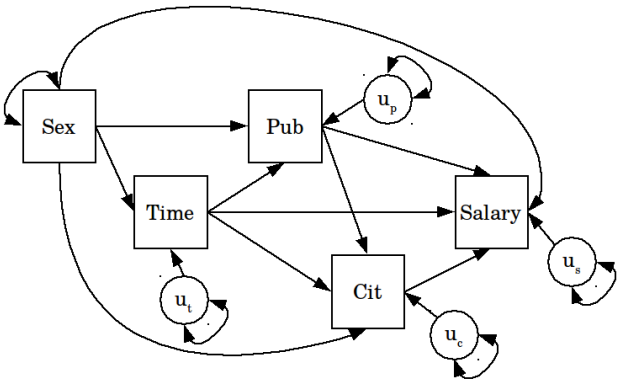
Table 2: correlation raw data

	time	pub	sex	cit	salary
time	1.000	0.651	0.210	0.373	0.608
pub	0.651	1.000	0.159	0.333	0.506
sex	0.210	0.159	1.000	0.149	0.201
cit	0.373	0.333	0.149	1.000	0.550
salary	0.608	0.506	0.201	0.550	1.000

The entire system can be expressed as,

$$\begin{aligned} time &\sim sex \\ pub &\sim sex + time \\ cit &\sim sex + time + pub \\ salary &\sim sex + time + pub + cit \end{aligned} \tag{7}$$

Figure 2: Path diagram



Model fit using linear multiple regression

Next we explore what the estimates will be for each of our linear equations using the multiple regression estimation framework.

$$time \sim sex$$

	Estimate	Std. Error	t value	Pr(>|t|)
sex	0.21	0.125	1.674	0.099

$$pub \sim sex + time$$

	Estimate	Std. Error	t value	Pr(>|t|)
sex	0.023	0.1	0.234	0.816
time	0.646	0.1	6.442	0.000

$$cit \sim sex + time + pub$$

	Estimate	Std. Error	t value	Pr(>|t|)
sex	0.071	0.122	0.578	0.566
time	0.257	0.159	1.620	0.110
pub	0.155	0.157	0.983	0.330

$$salary \sim sex + time + pub + cit$$

	Estimate	Std. Error	t value	Pr(>|t|)
sex	0.047	0.095	0.498	0.621
time	0.378	0.126	3.002	0.004
pub	0.134	0.123	1.089	0.281
cit	0.357	0.101	3.542	0.001

Structural Equation Modeling of the System

Next we will use the R package *lavaan* to fit the above model to the our data.

```
suppressMessages(library(lavaan))
fig12.2.1_mod = '
time    ~ sex
pub     ~ sex + time
cit     ~ sex + time + pub
salary  ~ sex + time + pub + cit'
fit = sem(fig12.2.1_mod, data=dat)
summary(fit,fit.measures=T)
```

```
lavaan (0.5-22) converged normally after 135 iterations

  Number of observations                            62

  Estimator                                         ML
  Minimum Function Test Statistic                0.000
  Degrees of freedom                                 0

Model test baseline model:

  Minimum Function Test Statistic               91.009
  Degrees of freedom                                10
  P-value                                        0.000

User model versus baseline model:

  Comparative Fit Index (CFI)                    1.000
  Tucker-Lewis Index (TLI)                       1.000

Loglikelihood and Information Criteria:

  Loglikelihood user model (H0)              -1348.081
  Loglikelihood unrestricted model (H1)      -1348.081

  Number of free parameters                         14
  Akaike (AIC)                                2724.162
  Bayesian (BIC)                              2753.942
  Sample-size adjusted Bayesian (BIC)         2709.893

Root Mean Square Error of Approximation:

  RMSEA                                          0.000
  90 Percent Confidence Interval          0.000  0.000
  P-value RMSEA <= 0.05                            NA

Standardized Root Mean Square Residual:

  SRMR                                           0.000

Parameter Estimates:

  Information                                 Expected
```

```
  Standard Errors                             Standard

Regressions:
                   Estimate      Std.Err     z-value  P(>|z|)
  time ~
    sex                1.794       1.063    1.688    0.091
  pub ~
    sex                0.657       2.762    0.238    0.812
    time               2.114       0.323    6.548    0.000
  cit ~
    sex                2.426       4.096    0.592    0.554
    time               1.034       0.622    1.661    0.097
    pub                0.190       0.188    1.008    0.314
  salary ~
    sex              917.767    1783.362    0.515    0.607
    time             857.006     276.091    3.104    0.002
    pub               92.746      82.391    1.126    0.260
    cit              201.931      55.141    3.662    0.000

Variances:
                   Estimate      Std.Err     z-value  P(>|z|)
   .time              17.214       3.092    5.568    0.000
   .pub              111.191      19.971    5.568    0.000
   .cit              244.239      43.867    5.568    0.000
   .salary         46042901.212 8269549.178    5.568    0.000
```

Estimation comparisons

Below we present tables of estimates from both the SEM as well as the multiple equations using linear regression.

Table 7: Standardized estimates from SEM

lhs	rhs	std.all	z	pvalue
time	sex	0.210	1.688	0.091
pub	sex	0.023	0.238	0.812
pub	time	0.646	6.548	0.000
cit	sex	0.071	0.592	0.554
cit	time	0.257	1.661	0.097
cit	pub	0.155	1.008	0.314
salary	sex	0.047	0.515	0.607
salary	time	0.378	3.104	0.002
salary	pub	0.134	1.126	0.260
salary	cit	0.357	3.662	0.000

Table 8: Estimates from linear regression models

	Estimate	Std. Error	t value	Pr(>|t|)
time ~ sex	0.210	0.125	1.674	0.099
pub ~ sex	0.023	0.100	0.234	0.816
pub ~ time	0.646	0.100	6.442	0.000
cit ~ sex	0.071	0.122	0.578	0.566
cit ~ time	0.257	0.159	1.620	0.110
cit ~ pub	0.155	0.157	0.983	0.330
salary ~ sex	0.047	0.095	0.498	0.621
salary ~ time	0.378	0.126	3.002	0.004
salary ~ pub	0.134	0.123	1.089	0.281
salary ~ cit	0.357	0.101	3.542	0.001

Table 9: correlation raw data

	time	pub	sex	cit	salary
time	1.000	0.651	0.210	0.373	0.608
pub	0.651	1.000	0.159	0.333	0.506
sex	0.210	0.159	1.000	0.149	0.201
cit	0.373	0.333	0.149	1.000	0.550
salary	0.608	0.506	0.201	0.550	1.000