Diego Muñoz. 2024. El Path Analysis como herramienta de identificación de dependencias entre variables. Revista Sigma, 20 (2). Páginas 1–8.

REVISTA SIGMA

Departamento de Matemáticas y Estadística

Universidad de Nariño

Volumen XX Nº 2 (2024), páginas 1–8

El Path Analysis como herramienta de identificación de dependencias entre variables

Diego Fernando Muñoz Muñoz[1]

Abstract: Path analysis (PA) is a relevant statistical method for evaluating hypothetical relationships between variables. Emerging in 1920, it has remained current due to its ability to estimate the magnitude and significance of causal connections. In this article, to address the characteristics of the AP, a model represented by a path diagram is formulated, in which the arrows indicate the relationships between variables and the Path coefficients represent the relationship between the variables connected by the arrows; likewise, the estimators of the analysis are identified. Finally, relevant aspects of the hypothesized model are discussed, as well as some considerations regarding the use of PA. *Keywords.* Path analysis, trail analysis, multiple regression, endogenous variables, exogenous variables.

Resumen: El Path analysis o análisis de senderos (PA) es un método estadístico relevante para evaluar relaciones hipotéticas entre variables. Surgido en 1920, se ha mantenido vigente debido a su capacidad para estimar la magnitud y significancia de las conexiones causales. En el presente artículo, para abordar las características del PA, se formula un modelo representado por un diagrama de senderos, en el cual las flechas indican las relaciones entre variables y los coeficientes Path representan la relación entre las variables conectadas por las flechas; así mismo, se identifican los estimadores del análisis. Finalmente, se discuten aspectos relevantes tanto del modelo hipotetizado, como algunas consideraciones respecto al uso de PA.

Palabras Clave. Path analysis, análisis de sendero, regresión múltiple, variables endógenas, variables exógenas.

[1]Docente, Departamento de Psicología, Universidad de Nariño. Investigador, Grupo Psicología y Salud, Departamento de Psicología, Universidad de Nariño, Colombia. Correo: dife@udenar.edu.co Orcid: https://orcid.org/0000-0003-2375-4019 Google Schoolar: https://scholar.google.com.mx/citations?hl=es&user=jRF6JMgAAAAJ

1. Introducción

El mundo, en sus múltiples aristas, está permeado por una serie de intrincadas relaciones entre eventos, los cuales se influyen recíprocamente, y generan una realidad compleja. Por tal razón, ha sido de especial interés el desarrollo de estrategias que permitan analizar ciertas relaciones de dependencia mutua entre variables, entendiendo tales relaciones como susceptibles de explicación a través de modelos teóricos. Más aún, en términos de las ciencias sociales, dada la múltiple serie de interrelaciones entre distintas características o *constructos* que determinan distintos aspectos de los contextos de interacción individual y grupal, se han planteado y desarrollado métodos orientados hacia la medición y compresión de las mismas (Coolican, 2014).

Tal es el caso del Path Analysis, o análisis de senderos (en adelante PA). Según Pérez, Medrano y Sánchez (2013),"es un método que permite evaluar el ajuste de modelos teóricos en los que se proponen un conjunto de relaciones de dependencia entre variables" (pág. 52). Este método es el miembro "más viejo" en la familia de los modelos de ecuaciones estructurales (SEM, por sus siglas en inglés), aunque tal característica no opaca su amplio uso en la investigación (Kline, 2016). Su utilidad y practicidad han aportado en su vigencia, y, dada su importancia, en el presente artículo se abordan aspectos relacionados con su objetivo, sus estimadores y coeficientes, sus usos, y un ejemplo práctico en el software estadístico R.

2. Path analysis: características y consideraciones.

El PA es un método desarrollado por el genetista Sewall Wright en el año 1920 (Loehlin y Beaujean, 2017). Desde su creación ha sido una herramienta ampliamente usada en ciencias sociales y del comportamiento, dado que su objetivo es proveer la estimación de la magnitud y la significancia de conexiones causales entre grupos de variables (Stage, Carter y Nora, 2004), ayudando a seleccionar o inferir entre hipótesis sobre la influencia entre ellas (Pérez et al, 2013).

Como menciona Stage *et al* (2004), el PA se puede calificar como un método muy relacionado con la regresión múltiple. Esto dado que, como señalan Pérez *et al* (2013), además de verificar la contribución directa que existe entre un grupo de variables independientes sobre una dependiente, también se verifica y analiza la interacción entre variables predictoras y la influencia indirecta de las mismas sobre las variables dependientes, las cuales, a su vez, pueden operar como variables independientes de otras variables que se especifiquen en el modelo.

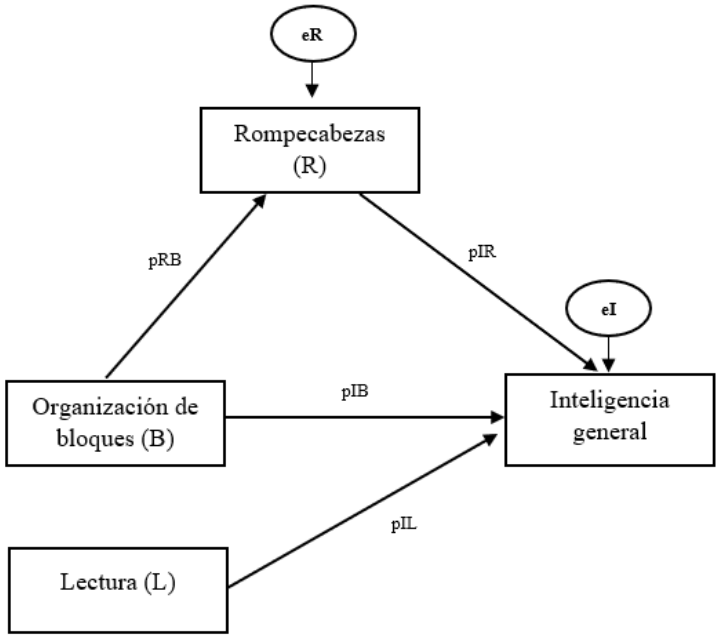
Por otro lado, a diferencia de los modelos de ecuaciones estructurales, en los cuales se estiman las relaciones causales entre variables latentes, a través de medidas múltiples (a modo de indicadores-variables manifiestas), el PA se utiliza cuando existe una única medida de cada constructo (Kline, 2016; Pérez *et al*, 2013).

Al realizar un PA, se formula un modelo que pretende explicar, de manera teórica, la relación existente entre las variables, modelo que puede ser representado a través de ecuaciones o diagramas (Pérez, 2013). Tal es el caso del Path Diagram, diagrama de senderos o diagrama de trayectorias (Lohelin y Beaujean, 2017). En este diagrama, se representan las relaciones entre variables a través de flechas, también denominadas senderos (de ahí la denominación de análisis de senderos), para las cuales se estiman coeficientes *Path*, que son semejantes a los coeficientes beta de una regresión, y que, como mencionan Pérez *et al* (2013), indican "en qué medida un cambio en a variable al comienzo de una flecha se relaciona con un cambio

en la variable al final de la flecha" (p. 53).

Para el desarrollo y explicación del PA, se revisará un modelo sobre el nivel de inteligencia general. Para tal objetivo, se ha retomado la base de datos denominada "*Ability and intelligence test*". Este dataset es una matriz de covarianzas correspondiente a 112 observaciones acerca de cinco test relacionados con inteligencia (lectura, vocabulario, rompecabezas, bloques e imágenes), así como una medida de inteligencia general. Esta base de datos, obtenida en el estudio de Bartholomew y Knott (1990), se encuentra a través de la librería *datasets* del software R. En el presente ejercicio, para aspectos prácticos, se ha decidido tomar tres variables, siguiendo el ejemplo presentado por Nayebi (2020), se ha hipotetizado que la medida de inteligencia general está explicada por la contribución de la organización de bloques, la lectura y los rompecabezas. Del mismo modo, se plantea la existencia de una contribución de la organización de bloques en la realización de rompecabezas. Así, en el diagrama de trayectoria de la figura 1 se presenta el modelo propuesto:

Figura 1: diagrama del modelo.



En el gráfico se pueden identificar distintos aspectos. Las variables se han enmarcado en cuadros, dado que son variables observables. Para el caso de variables latentes suelen usarse círculos. Retomando lo propuesto por Pérez *et al* (2013), Inteligencia general y Rompecabezas reciben cierto grado de influencia de las demás variables, por lo cual se pueden denominar variables endógenas, mientras que organización de bloques y la lectura, que no reciben influencia de otras, se denominan variables exógenas. Los círculos en los cuales se identifican las siglas eD y eR hacen referencia a los residuales, los cuales muestran la varianza de la variable endógena que no alcanza a ser explicada por el modelo. Adicionalmente, los coeficientes Path son aquellos que se encuentran sobre las flechas o senderos, e indican el signo y la magnitud del efecto de una variable sobre otra (Beaujean, 2014). Finalmente, es importante considerar que se pueden calcular *efectos indirectos*, que hacen referencia a efectos acumulativos de influencia de una variable sobre otra, a través de una variable mediadora; su cálculo se obtiene a través de la multiplicación de los Path coeficientes. Para el caso, $pRB * pIR =$ es el efecto indirecto de bloques sobre inteligencia artificial a través de rompecabezas.

Antes de continuar con el desarrollo del ejercicio, es importante tener en cuenta que, para la realización de un análisis de sendero, se deben comprobar los siguientes supuestos: normalidad multivariada, multicolinealidad, independencia de los residuales, así como linealidad de las variables (Pérez *et al*, 2013). Si bien para efectos prácticos, se considerará que las variables del modelo no incumplen ninguno de los supuestos, para un análisis más detallado de la comprobación de los supuestos, se recomienda la lectura del capítulo 1.8 de Nayebi (2020).

2.1. Especificación del análisis en R

De primera mano, se cargan las librerías necesarias para el ejercicio, y se organiza la base sobre la cual se realizarán los análisis, con el siguiente código:

```
library(lavaan)
library(semPlot)
library(datasets)
datasets:: ability.cov
data<- ability.cov [,1:6]
names(data) <- gsub ("cov\\.", "", names(data))
names(data)
```

Una vez ajustado el dataset, se procede a la definición del modelo, como se muestra a continuación:

```
model2 <- '
    general ~ a*maze + b*reading + c*blocks
    maze ~ d*blocks
    ind:= c*d
    # var resid
    general ~~ e*general
    blocks ~~ f*blocks
    maze ~~ g*maze
    reading ~~ h*reading'
```

**__general:__ inteligencia general; __maze:__ rompecabezas; __Reading:__ lectura; __blocks:__ organización de bloques*

Sobre el código es importante comentar que el símbolo $\sim$ implica la relación de influencia: las variables hacia la derecha son las variables endógenas, y las variables a la izquierda son las exógenas. Las letras a, b, c, etc. Que se multiplican por cada variable, representan los coeficientes Path que se calcularán. $Ind := c * d$ calcula el efecto indirecto de blocks sobre general a través de maze. Para una revisión más exhaustiva sobre la especificación de la sintaxis de *lavaan*, se sugiere revisar a Beaujean (2014).

Así, una vez definido el modelo, se procede a calcular su ajuste. Para tal caso, se recurre a la función SEM de la librería *lavaan*. Es importante considerar que, como menciona Beaujean, también se puede hacer uso de la función CFA (análisis factorial confirmatorio):

```
fit2 <- sem(model2, data = data)
summary(fit2)
```

Tal código generará la siguiente salida:

```
lavaan 0.6.15 ended normally after 15 iterations

  Estimator                                         ML
  Optimization method                           NLMINB
  Number of model parameters                         8

  Number of observations                             6

Model Test User Model:

  Test statistic                                 0.716
  Degrees of freedom                                 2
  P-value (Chi-square)                           0.699

Parameter Estimates:

  Standard errors                             Standard
  Information                                 Expected
  Information saturated (h1) model          Structured

Regressions:
                   Estimate  Std.Err  z-value  P(>|z|)
  general ~
    maze       (a)   -0.235    0.513   -0.458    0.647
    reading    (c)    0.255    0.075    3.378    0.001
    blocks     (b)    0.165    0.065    2.536    0.011
  maze ~
    blocks     (d)    0.103    0.030    3.371    0.001

Variances:
                   Estimate  Std.Err  z-value  P(>|z|)
   .general    (e)   18.310   10.571    1.732    0.083
    blocks     (f) 2088.633 1205.873    1.732    0.083
    reading    (g)  535.843  309.369    1.732    0.083
   .maze       (h)   11.599    6.697    1.732    0.083

Defined Parameters:
                   Estimate  Std.Err  z-value  P(>|z|)
    ind               0.017    0.008    2.027    0.043
```

En el modelo, respecto a la significancia de las relaciones hipotetizadas, los coeficientes de regresión estandarizados indicaron que las variables exógenas reading, y blocks tuvieron efectos significativos sobre la variable endógena general. Específicamente, los coeficientes para las relaciones reading $\longrightarrow$ general y blocks $\longrightarrow$ general fueron 0.255 y 0.165 respectivamente, con valores $p < 0.05$. Maze no tuvo efectos significativos sobre general. Así mismo, se identificó que blocks tuvo un coeficiente estimado de 0.103 en términos de su influencia respecto a la variable maze ($p = 0.001$). Finalmente, existe un efecto indirecto significativo de la variable blocks sobre general, a través de maze.

La función *summary ()* de R tiene la funcionalidad de generar indicadores de bondad de ajuste, así como el estadístico R2, como menciona Beajuean (2014), a través del siguiente cómputo:

```
summary(fit2, fit.measures=T,rsquare=T)
```

El argumento $fit.measures = T$ retorna los índices de ajuste, mientras que *rsquare=T* provee el estadístico R^2 para evaluar el porcentaje de varianza explicado por el modelo. De tal modo, se presentan los resultados a continuación:

```
R-Square:
                   Estimate
     general          0.808
     maze             0.654
```

Se encontró que las variables endógenas (picture, maze) explicaban una cantidad significativa de varianza en sus respectivas variables exógenas, con valores de R-cuadrado de 0.808 y 0.654 respectivamente.

Por otro lado, el modelo propuesto demostró un buen ajuste a los datos observados, como lo sugieren los índices de ajuste comparativo (CFI) y Tucker-Lewis (TLI), que fueron de 1.000 y 1.318 respectivamente. Sin embargo, el Root Mean Square Error of Approximation (RMSEA) fue de 0.000, con un intervalo de confianza del 90 % que incluye valores cercanos a cero en su valor inferior, pero un intervalo superior de 0.597, lo cual puede implicar un alto grado de incertidumbre en la estimación del RMSEA. El SMRM, adicionalmente, no tiene un adecuado ajuste a los datos, puesto que es mayor a 0.1, por lo cual la viabilidad del modelo debe ser analizada con precaución. Si bien estos resultados sugieren la necesidad de re - especificar el modelo, para efectos del desarrollo del ejercicio, se continúa con el cómputo del modelo.

```
User Model versus Baseline Model:

  Comparative Fit Index (CFI)                    1.000
  Tucker-Lewis Index (TLI)                       1.318

Loglikelihood and Information Criteria:

  Loglikelihood user model (H0)                -91.914
  Loglikelihood unrestricted model (H1)        -91.556

  Akaike (AIC)                                 199.828
  Bayesian (BIC)                               198.162
  Sample-size adjusted Bayesian (SABIC)        175.039

Root Mean Square Error of Approximation:

  RMSEA                                          0.000
  90 Percent confidence interval - lower         0.000
  90 Percent confidence interval - upper         0.597
  P-value H_0: RMSEA <= 0.050                    0.703
  P-value H_0: RMSEA >= 0.080                    0.291

Standardized Root Mean Square Residual:

  SRMR                                           0.113
```

Posteriormente, se procede a generar el Path diagram, mediante el siguiente código:

```
semPaths(fit2, "eq", layout = "tree", edge.label.cex = 1, style =
           "OpenMx", intercepts = TRUE, residuals = TRUE, rotation
           = 3, rescale = FALSE, sizeMan = 11, sizeMan2 = 4,
           whatLabels = "std")
```

Obteniendo la siguiente gráfica:

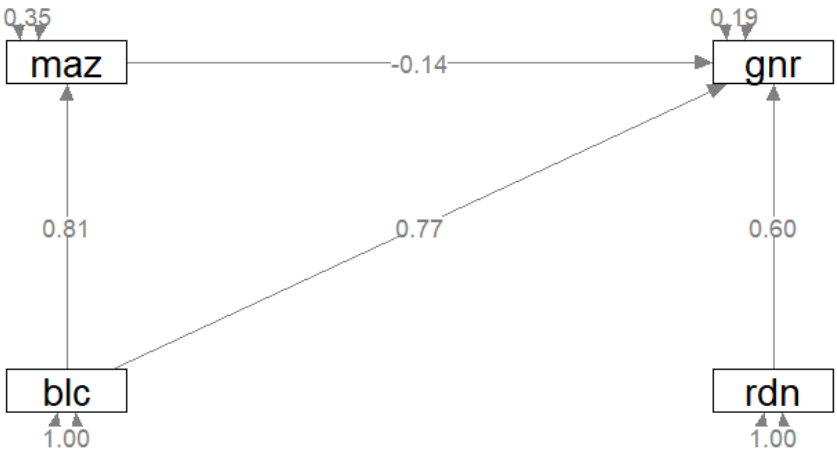


Figura 2: **Maz: rompecabezas;* ***gnr****: inteligencia general;* ***blc****: Organización de bloques;* ***rdn****: lectura.*

En el gráfico, se observa la identificación de los coeficientes Path. El coeficiente de 0.77 entre blocks y general indica que, por cada desviación estándar de cambio en blocks, se espera un cambio de 0.77 desviaciones estándar en general, manteniendo constantes las otras variables en el modelo. El coeficiente de 0.60 entre reading y general tiene una interpretación similar: por cada desviación estándar de cambio en reading, se espera un cambio de 0.60 desviaciones estándar en general. El coeficiente path de 0.81 entre blocks y maze indica una relación positiva entre estas dos variables, mientras que el coeficiente de -0.14 entre maze y general indica una relación negativa entre estas dos variables endógenas. Finalmente, se observa valores residuales de 0.35 para maze y 0.19 para general, los cuales son el porcentaje de varianza de las variables que no se explican por el modelado de las variables exógenas.

En términos aplicados al problema estudiado, los resultados sugieren relaciones significativas entre las medidas de tareas cognitivas específicas (organización de bloques, lectura y realización de rompecabezas y la medida de inteligencia general en la muestra de referencia. En primer lugar, las puntuaciones en la tarea de organización de bloques (blocks) y lectura (reading) están positivamente asociadas con la medida de inteligencia general, lo que sugiere que un alto desempeño en estas tareas se relaciona con un mayor nivel de inteligencia general en los participantes. Estos resultados son consistentes con la literatura previa, en la cual se han encontrado vínculos entre habilidades específicas, como la habilidad visoespacial y la comprensión lectora, y la inteligencia general (Deary *et al.*, 2007; Sternberg, 2000). Además, la asociación positiva entre las tareas de organización de bloques (blocks) y realización de rompecabezas (maze) sugiere una posible relación entre la habilidad visoespacial medida por la tarea de "blocksƨ las habilidades cognitivas implicadas en la resolución de laberintos. Esto es consistente con la idea de que ciertas habilidades cognitivas subyacentes pueden influir en múltiples tareas, lo que resalta la complejidad de la inteligencia y la necesidad de un enfoque multidimensional en su estudio (Deary, 2012).

Así, a través el PA adelantado, se ha generado una exploración respecto al modelo hipotetizado, que genera información valiosa de cara tanto a análisis más sofisticados, como a una compresión teórica sobre la inteligencia general, que subyace de variables observables.

3. Conclusiones

Hasta el momento, se ha adelantado una breve revisión sobre el Path analysis, sus usos, su cómputo y aplicación a un ejercicio sintético. Para finalizar, se retoman algunos aspectos que deber ser considerados en el momento de llevar a cabo análisis de senderos. Estos aspectos, retomados de Stage *et al* (2004), resaltan tanto fortalezas y debilidades, a saber:

Su principal fortaleza es permitir estudiar efectos directos e indirectos con una gran variedad de variables dependientes e independientes. Además, permite "diagramar" tal conjunto de variables, las cuales se traducen en ecuaciones necesarias para el análisis. Por otro lado, aunque en algunos casos se puede usar el PA para evaluar entre dos o más hipótesis causales, no es posible establecer de manera absoluta la dirección de la causalidad. Así mismo, no debe ser usado en el caso de que hayan "bucles de retroalimentación", es decir, cuando no hay una progresión causal establecida a lo largo del diagrama de senderos (Stage *et al*, 2004). Aún con lo anterior, el uso de este tipo de análisis permite a los a los investigadores, y en mayor medida a aquellos pertenecientes a las ciencias sociales, tener mayor comprensión sobre las relaciones existentes entre grupos de variables, a través del rastreo de implicaciones de una serie de suposiciones causales que se hace sobre un sistema de relaciones entre variables (Stage *et al*, 2004).

Referencias

[1] Coolican, H. (2014). *Research methods and statistics in psychology* (6th ed.). Routledge.

[2] Pérez, E., Medrano, L y Sánchez, J. (2013). El Path Analysis: conceptos básicos y ejemplos de aplicación. Revista Argentina de Ciencias del Comportamiento, 5 (1). 52-66.

[3] Beaujean, A. (2014). *Latent Variable Modeling Using* R. Routledge.

[4] Kline, R. (2016). *Principles and Practice of Structural Equation Modeling*. The Guilford Press.

[5] Nayebi, H. (2020). A*dvanced Statistics for Testing Assumed Casual Relationships*. Springer.

[6] Stage, F., Carter, H y Nora, A. (2004) Path Analysis: An Introduction and Analysis of a Decade of Research, *The Journal of Educational Research*, 98 (1). 5-13.

[7] Loehlin, J y Beaujean, A. (2017). *Latent Variable Models An Introduction to Factor, Path, and Structural Equation Analysis*. Routledge.