Análisis Factorial Confirmatorio

Análisis Avanzado de Datos II

Material preparado por

Anais Herrera Leighton

Medición en Ciencias Sociales

- En ocasiones un indicador puede ser suficiente para capturar el concepto que se quiere medir
  - Variables observadas (ej. edad, sexo)

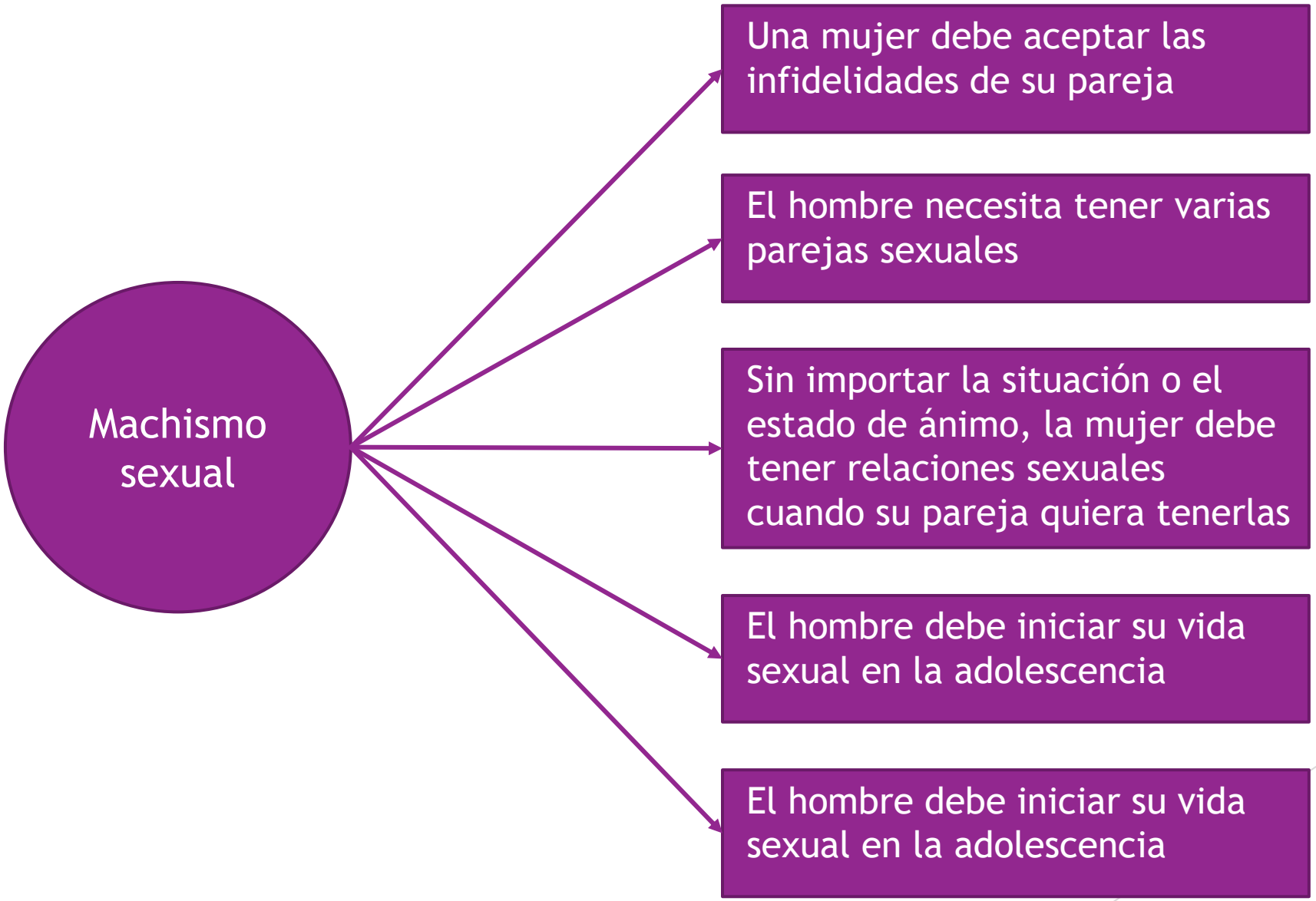
- Pero algunos conceptos no pueden ser medidos directamente, para lo cual requerimos distintos indicadores
  - Variables latentes (ej. Actitudes machistas, clase social)

Medición en Ciencias Sociales

Ejemplo de medición de una variable latente (unidimensional): *Escala de Machismo Sexual* (Díaz, Rosas & González, 2010).

- Expresa en tu opinión tu grado de acuerdo o desacuerdo con las siguientes frases. Por favor responde honestamente utilizando estas opciones: (1) Totalmente en desacuerdo; (2) En desacuerdo; (3) Sin opinión; (4) De acuerdo; (5) Totalmente de acuerdo
  - Una mujer debe aceptar las infidelidades de su pareja
  - El hombre necesita tener varias parejas sexuales
  - Sin importar la situación o el estado de ánimo, la mujer debe tener relaciones sexuales cuando su pareja quiera tenerlas
  - El hombre debe iniciar su vida sexual en la adolescencia
  - El hombre debe hacer que su hijo hombre inicie su vida sexual

Medición en Ciencias Sociales



Medición en Ciencias Sociales

Ejemplos de variables latentes (con más de una dimensión)

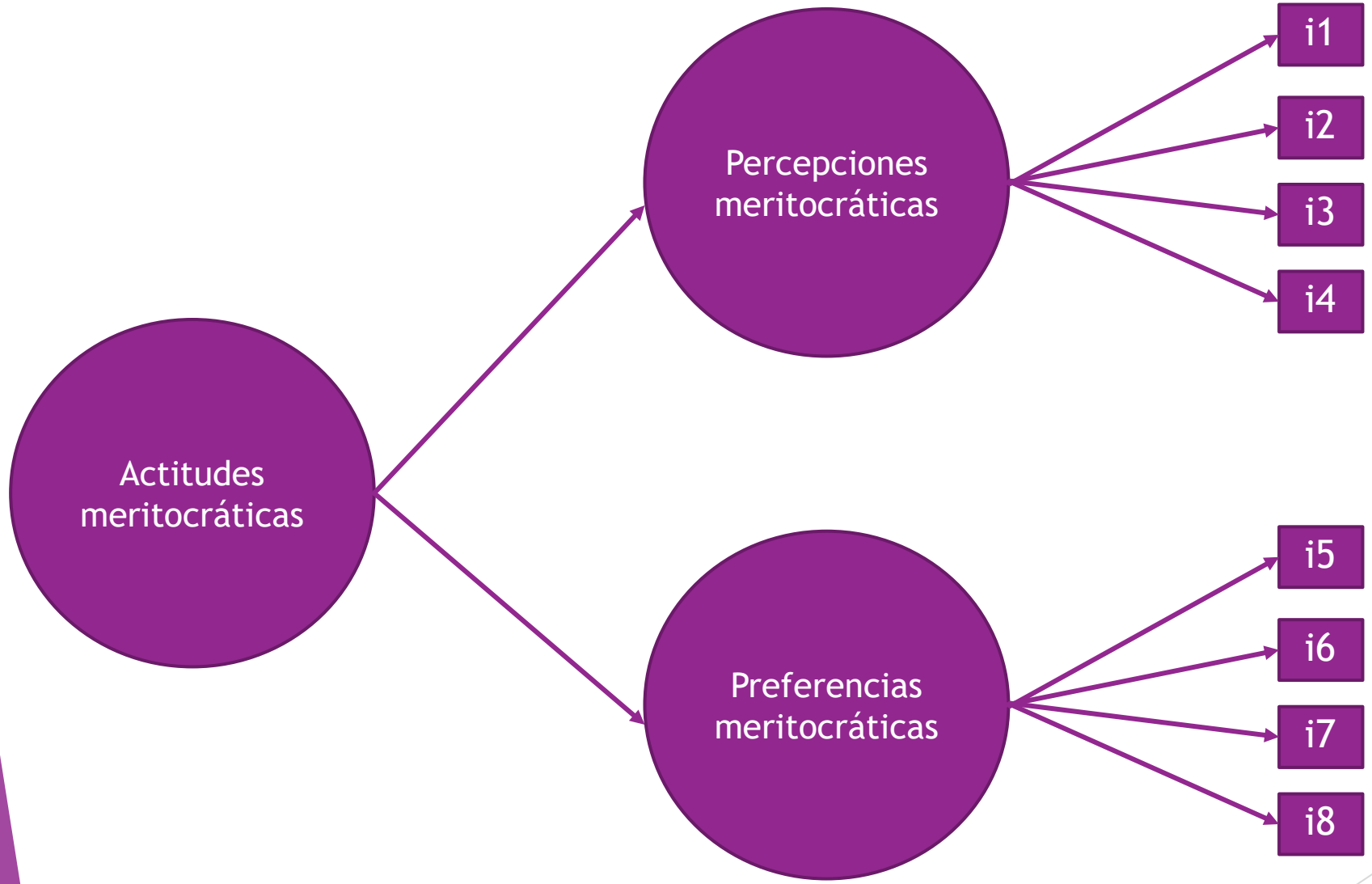
- Actitudes meritocráticas
  - Percepciones meritocráticas
  - Preferencias meritocráticas
- Actitudes hacia la violencia de carabineros
  - Justificación uso de violencia para disolver marchas
  - Justificación uso de violencia al allanar comunidades mapuche
- Inteligencia
  - Habilidades verbales
  - Habilidades matemáticas

Medición en Ciencias Sociales

Ejemplos de variables latentes (con más de una dimensión): *Actitudes meritocráticas* (Castillo, Iturra, Meneses & Maldonado, 2021)

Dimensión	Subdimensión	Indicador
Percepción	Meritocrática	Quienes más se esfuerzan logran obtener mayores recompensas que quienes se esfuerzan menos.(i1)
		Quienes poseen más talento logran obtener mayores recompensas que quienes poseen menos talento. (i2)
	No meritocrática	Quienes tienen padres ricos logran salir adelante. (i3)
		Quienes tienen buenos contactos logran salir adelante. (i4)
Preferencia	Meritocrática	Quienes más se esfuerzan deberían obtener mayores recompensas que quienes se esfuerzan menos. (i5)
		Quienes poseen más talento deberían obtener mayores recompensas que quienes poseen menos talento. (i6)
	No meritocrática	Está bien que quienes tienen padres ricos salgan adelante. (i7)
		Está bien que quienes tienen buenos contactos salgan adelante (i8)

Medición en Ciencias Sociales



Medición en Ciencias Sociales

- Los conceptos complejos no se pueden medir directamente, por lo que se realiza una operacionalización
  - Variable latente o factor
- Generamos mediciones que se aproximan a medir lo que representa el concepto
  - Variables observadas, indicadores o ítems
- No observamos directamente la variable latente, si no que esta es **deducida a partir de las correlaciones entre las variables observadas**

Medición de variables latentes

- Podemos utilizar modelos estadísticos para entender constructos sociales y responder preguntas cómo
  - ¿Cómo se relacionan entre sí distintos indicadores de un mismo concepto?
  - ¿Son las variables adecuadas para capturar un determinado concepto?
  - ¿Cuántas dimensiones tiene un concepto?
- Análisis factorial (variables continuas)
  - Permite estudiar la interrelación (o interdependencia) de variables observadas
  - Se agrupan las variables en un factor o en un número reducido de factores
  - Cada indicador tiene una varianza común que es explicada por el factor latente y una varianza única

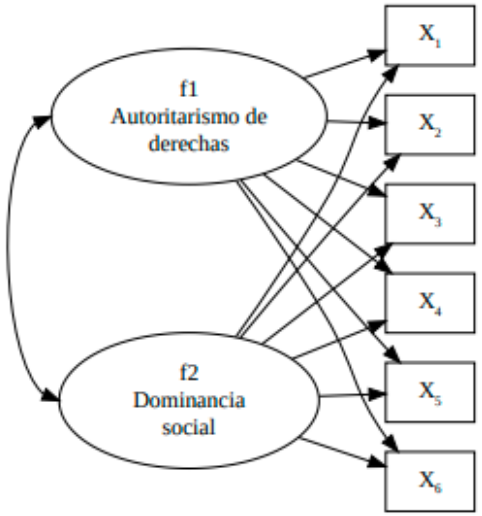
Análisis factorial

Análisis Factorial Exploratorio (AFE o EFA)	Análisis Factorial Confirmatorio (AFC o CFA)
Cuando no hay un modelo teórico que sustenta la manera en que las variables observadas se relacionan entre sí	Un modelo teórico especifica qué variables observadas se relacionan con qué variable(s) latente(s)
Busca identificar las variables latentes que subyacen a un conjunto de indicadores correlacionados entre sí	Generalmente, cada variable observada se relaciona solamente con una variable latente
El modelo plantea que cada variable observada se relaciona con todas las variables latentes	Podemos evaluar si el modelo planteado se ajusta a los datos observados

Análisis factorial

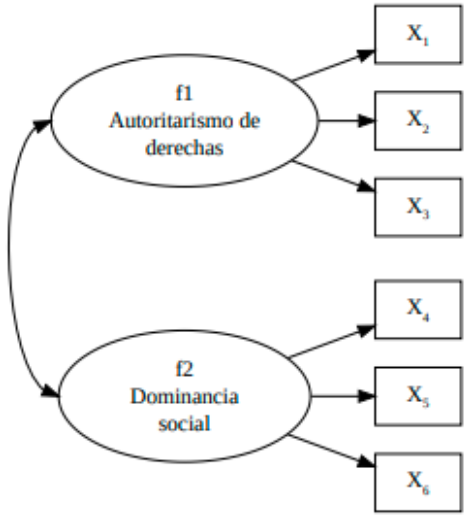
¿Corresponden autoritarismo y dominancia a dos dimensiones distintas del conservadurismo?

Exploratorio



Relaciona todos los indicadores con todos los factores

Confirmatorio



Restringe relaciones indicadores y factores

Términos relevantes

- **Factores**: variables latentes a la base de las correlaciones entre los indicadores
- **Cargas factoriales:** medida estandarizada de asociación entre el indicador y la variable latente
- **Comunalidad**: proporción del indicador que se asocia a factor(es) comun(es)

Supuestos Análisis Factorial Confirmatorio

- Un número importante de variables observadas de nivel de medición intervalo/razón (mínimo 5 categorías de respuesta en escala).
- Base de datos suficientemente grande. Esto depende de la calidad de los datos, de las correlaciones y del número de ítems. Existen distintos criterios. En general, utilizar bases de al menos 200 casos.
- El modelo está correctamente especificado
- Relación lineal entre variables
- Ausencia de multicolinealidad
- Normalidad multivariante en las variables (al usar el método de estimación de máxima verosimilitud)

Pasos AFC

1. Identificación modelo
2. Evaluación ajuste del modelo
3. Cargas factoriales y comunalidades
4. Comparación modelos
5. Reespecificar modelo (si es necesario)

Identificación del modelo

- Un requisito para poder realizar el AFC es que el modelo se encuentre “identificado”, es decir, que exista una solución única para todos los parámetros del modelo, es decir, contamos con información suficiente para estimarlo.
- Esto depende de la relación entre la cantidad de variables que utilizamos y la cantidad de parámetros a estimar.
- Los grados de libertad del modelo deben ser mayores a 0.
- En caso de que no se cumpla con esto, el software nos los hará saber, debemos cambiar la especificación del modelo.

Identificación del modelo

- En caso de que el modelo no se encuentre identificado debemos chequear ciertas condiciones necesarias, pero no suficientes (es decir, que pueden cumplirse y aun así el modelo no se encontrará identificado).
- Regla t: condición necesaria pero no suficiente $t \leq \frac{1}{2} \cdot p(p+1)$
- Donde: t = número total de parámetros a ser estimados (coeficientes de regresión, varianzas de los errores y correlaciones) y p = número de variables observadas
- La escala de las variables latentes debe estar fijada (ya sea fijando una de las cargas factoriales para cada variable latente en 1 o asignando una varianza de 1 a la variable latente).
- Deben haber suficientes indicadores para cada variable latente:
  - Al menos 2 indicadores por variable latente cuando el modelo tiene al menos 2 variables latentes
  - Al menos 3 indicadores por variable latente cuando el modelo tiene solamente 1 variable latente

Especificación y estimación del modelo

- Utilizaremos el paquete {lavaan} (Rosseel, 2012)
- Comenzamos especificando el modelo en un objeto:

```
mod_conf <- 'autoritarismo =~ aut1 + aut2 + aut3
             dominancia =~ dom1 + dom2 + dom3'
```

- El símbolo =~ significa que la variable latente (izquierda) es medida por los indicadores (derecha)
- La función cfa() se usa para estimar el modelo

```
mod_conf_cfa <- cfa(mod_conf, # modelo especificado
                    data = datos) # base de datos con ítems
```

- mod_conf_cfa es un objeto con los resultados del modelo ajustado

Evaluación ajuste del modelo

Prueba de Chi-Cuadrado

- Evalúa si existen discrepancias significativas entre la matriz de covarianza observada y la que es estimada por el modelo
- La hipótesis nula plantea que no son significativamente distintas
  - En este caso, $p > 0,05$ indica un buen ajuste del modelo con 95% de confianza
  - Buscamos mantener la hipótesis nula
- Sin embargo, es sensible al tamaño de la muestra
  - Con muestras grande ($n > 400$) es muy difícil encontrar un buen ajuste
- Criterio menos estricto: Chi2/ grados de libertad < 2
  - Aunque hay autores que proponen que si el valor es igual o menor a 4 el ajuste es adecuado

Evaluación ajuste del modelo

RMSEA (Root Mean Square Error of Approximation)

- Evalúa el ajuste del modelo considerando el tamaño de la muestra y la complejidad del modelo.
- Valores menores a 0,05 indican un buen ajuste
- Valores entre 0,05 y 0,08 indican un ajuste razonable
- Valores sobre 0,10 indican un mal ajuste

Evaluación ajuste del modelo

CFI (Comparative Fit Index)

- Índice que compara el ajuste del modelo con un modelo base o nulo (modelo sin covarianzas entre las variables)
- Da cuenta del incremento en el ajuste al pasar de un modelo base al modelo propuesto
- Es menos sensible al tamaño de la muestra
- Obtiene valores entre 0 y 1. Mientras más cercano a 1, mejor es el ajuste
- En general, se considera que un modelo tiene un ajuste razonable si CFI > 0,90 y que tiene un ajuste bueno si CFI > 0,95.

Cargas factoriales y comunalidades

- Estimamos el modelo y obtenemos cargas para cada variable en su factor correspondiente
- Obtenemos soluciones estandarizadas y no estandarizadas
- Un buen modelo tiene cargas factoriales estandarizadas sobre 0,7 (o al menos sobre 0,5)
- *Recordatorio*: las cargas factoriales son una medida estandarizada de asociación entre el indicador y la variable latente
- Las comunalidades corresponden al R-cuadrado, esto es, el porcentaje de varianza de una variable que es explicado por la variable latente (carga estandarizada al cuadrado).
- Las cargas (y comunalidades) más altas implican que las variables son más relevantes a la hora de definir un factor (mediciones más “puras” de este)
- La información de las cargas sirve para determinar eventualmente que ítem sería mejor eliminar (si es que es necesario eliminar algún ítem), pudiéndose descartar la(s) variable(s) con la(s) carga(s) más baja(s).

Cargas factoriales y comunalidades

```
Latent Variables:
                   Estimate  Std.Err  z-value  P(>|z|)   Std.lv  Std.all
  autoritarismo =~
    aut1              1.000                               1.646    0.907
    aut2              0.921    0.060   15.457    0.000    1.517    0.887
    aut3              0.929    0.069   13.502    0.000    1.530    0.803
  dominancia =~
    dom1              1.000                               1.201    0.919
    dom2              1.096    0.080   13.729    0.000    1.316    0.882
    dom3              0.861    0.084   10.298    0.000    1.034    0.685

Covariances:
                   Estimate  Std.Err  z-value  P(>|z|)   Std.lv  Std.all
  autoritarismo ~~
    dominancia        0.678    0.174    3.891    0.000    0.343    0.343
```

- Estimate: cargas no estandarizadas
  - aut1 y dom1 igual a 1
  - Fijado así para dar escala de los factores
  - lavaan asume que el primer indicador de un factor se fija en 1
- Std.lv: solución con factores estandarizados
- Std.all: solución con factores e indicadores estandarizados

Cargas factoriales y comunalidades

```
Latent Variables:
                   Estimate  Std.Err  z-value  P(>|z|)   Std.lv  Std.all
  autoritarismo =~
    aut1              1.000                               1.646    0.907
    aut2              0.921    0.060   15.457    0.000    1.517    0.887
    aut3              0.929    0.069   13.502    0.000    1.530    0.803
  dominancia =~
    dom1              1.000                               1.201    0.919
    dom2              1.096    0.080   13.729    0.000    1.316    0.882
    dom3              0.861    0.084   10.298    0.000    1.034    0.685

Covariances:
                   Estimate  Std.Err  z-value  P(>|z|)   Std.lv  Std.all
  autoritarismo ~~
    dominancia        0.678    0.174    3.891    0.000    0.343    0.343
```

- Todos los indicadores p<0,05
  - Su relación con los factores es significativa al 95% de confianza
- aut1 y dom1 no tienen valor p porque fueron fijados
- Los coeficientes estandarizados (Std.all) > 0,5
- aut1 y dom1 son los indicadores más puros de sus factores

Cargas factoriales y comunalidades

Latent Variables:

	Estimate	Std.Err	z-value	P(>|z|)	Std.lv	Std.all
autoritarismo =~						
aut1	1.000				1.646	0.907
aut2	0.921	0.060	15.457	0.000	1.517	0.887
aut3	0.929	0.069	13.502	0.000	1.530	0.803
dominancia =~						
dom1	1.000				1.201	0.919
dom2	1.096	0.080	13.729	0.000	1.316	0.882
dom3	0.861	0.084	10.298	0.000	1.034	0.685

Covariances:

	Estimate	Std.Err	z-value	P(>|z|)	Std.lv	Std.all
autoritarismo ~~						
dominancia	0.678	0.174	3.891	0.000	0.343	0.343

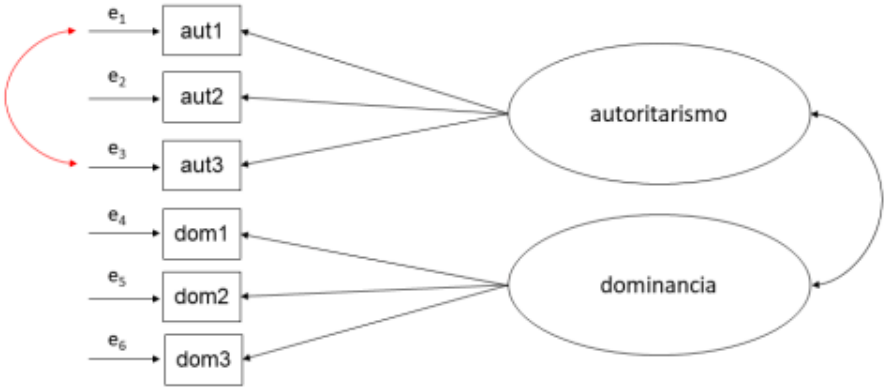
- La covarianza entre ambos factores es significativa al 95% de confianza, $p<0,05$
- La correlación entre ambas variables es positiva (Std.all = 0,343)

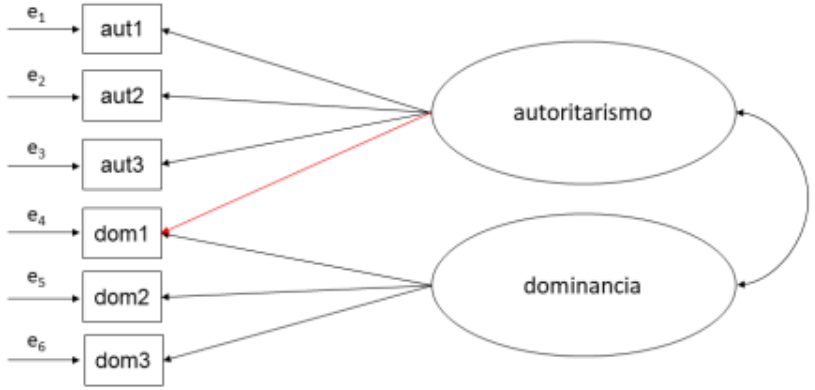
Cargas factoriales y comunalidades

- Las comunalidades están expresadas como R cuadrado (R-Square)
- Estas se calculan como la carga factorial estandarizada al cuadrado
- En general, observamos comunalidades altas.
- El ítem dom3 tiene la comunalidad más baja: el factor “dominancia” explica el 46,9% de la varianza de dom3.
  - Podríamos evaluar si sacar este ítem del modelo, en caso de querer reducir la escala

Reespecificar modelo

- Eventualmente, si el modelo no ajusta adecuadamente, se recomienda considerar índices de modificación
- En general, se proponen cambios relativos a la especificación de:
  - **Correlaciones entre los errores** de indicadores
    - Implica una asociación indicadores que no es explicada por la asociación con el factor, ej.: (a) Indicadores medidos en un mismo momento en el tiempo o (b) Indicadores medidos utilizando pruebas parecidas
  - **Cargas cruzadas**: un indicador carga en más de un factor
    - En caso de realizarse, debiera ser justificado teóricamente





Reespecificar modelo

- Observamos una lista de sugerencias de modificación bajo Modification Indices
- Mayores *mi* indican que este cambio significará un mayor cambio en la bondad de ajuste del modelo
- =~: indica agregar un efecto de una variable observada en una variable latente
- ~~: indica agregar una correlación entre los errores de dos variables observadas
- En este caso, los índices de modificación más altos se relacionan con el dom3, se sugiere:
  - Agregar dom3 como un indicador de autoritarismo (mi = 6,437)
  - Agregar una correlación entre los errores de dom1 y dom2 (mi = 6,437)
  - Agregar una correlación entre los errores de aut2 y dom3 (mi = 4,180)