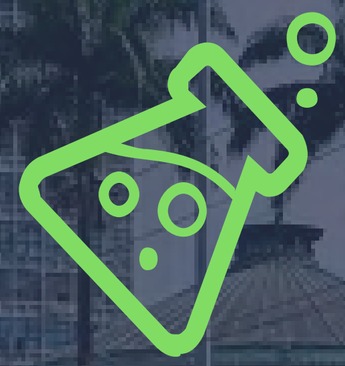


# Modelos mixtos para estudiar conjuntamente selección residencial, uso de tiempo y estilo de vida

**Sebastian Astroza,**

Departamento de Ingeniería Industrial,
Universidad de Concepción

# Contenidos

# Introducción

# Diferentes tipos de variables

**CONTINUA**

- Distancia
- Área

**ORDINAL**

- Escala de likert
- Nivel de uso/frecuencia
- Alternativas ordenadas

**NOMINAL**

- Modo de transporte
- Elección de un producto en una estantería
- Elección de una alternativa dentro de un set

**CONTEO**

- Número de automoviles
- Número de habitantes de un hogar

# Variables múltiple discreta-continuas (MDC)

- **Modelos clásicos de elección discreta:**
  - Alternativas son mutuamente exclusivas
  - Sólo una alternativa puede ser elegida

- **Modelos múltiple discreto-continuos:**
  - Consumidores eligen múltiples alternativas al mismo tiempo,
  - junto con la dimension continua de la cantidad consumida

# Ejemplo: asignación de tiempo

| Individuo | Tiempo asignado a actividades (horas por día) | | | | | |
|---|---|---|---|---|---|---|
| | En el hogar | Viaje | Recreación fuera del hogar | Trabajo fuera del hogar | Compras o trámites fuera del hogar | Total |
| 1 | 12.5 | 1.8 | 0.5 | 8.0 | 1.2 | 24.0 |
| 2 | 16.0 | 2.2 | 5.8 | 0.0 | 0.0 | 24.0 |
| 3 | 10.1 | 1.9 | 0.0 | 12.0 | 0.0 | 24.0 |
| 4 | 24.0 | 0.0 | 0.0 | 0.0 | 0.0 | 24.0 |

# Modelación conjunta de diversas variables

- De importante interés en muchas áreas de investigación
- Considera la naturaleza simultánea de decisiones que están inter-relacionadas.
- Deben ser estudiadas conjuntamente debido al impacto (en las múltiples decisiones) de:
  - variables exógenas observables en común
  - variables exógenas inobservables en común
  - combinación de las dos anteriores

# Ejemplo: variable observable en común

- Es posible que individuos de hogares con bajo ingreso
  - Elijan (o esten obligados a) ubicarse en barrios con alta densidad poblacional,
  - tengan bajos niveles de posesión de automóvil, y
  - pasen menos tiempo en actividades de recreación

- Si los efectos en común son solo debido a factores observables → se puede modelar cada variable independientemente

# Ejemplo: variable inobservable en común

- Individuos que tienen un estilo de vida activo
  - podrían buscar domicilio en áreas que ofrezcan mayor accessibilidad a centros de actividades,
  - podrían poseer menos autos, e
  - invertir tiempo importante para propósitos recreacionales

- Cuando variables inobservables impactan las diferentes variables

  Modelar independientemente → estimación ineficiente de los efectos de las variables independientes en las variables dependientes

- Cuando variables endógenas son usadas para explicar otras variables endógenas

  Modelar independientemente en presencia de efectos inobservables→ estimación inconsistente

# Modelación conjunta de diversas variables

- Literatura dominada por la modelación conjunta de multiples variables continuas

- ¿Qué pasa cuando las variables son no-conmesurables? (mix de variables continuas y discretas)

- Falta de una distribución multivariada conveniente para conjuntamente representar las relaciones entre variables dependendientes discretas y continuas

- Caso particular: variable dependiente de naturaleza multiple discreta-continua (MDC)

# Nuestro modelo

- Nuevo enfoque econométrico para la estimación conjunta de modelos mixtos que incluyen:
  - variable dependiente MDC
  - variables dependientes nominales,
  - variables dependientes de conteo, ordinales y continuas

- Estimación conjunta: factores latentes de estilo de vida (o personalidad) impactan las diferentes variables dependientes

- Indicadores subjetivos de opiniones/actitudes son reportados para las variables latentes

# The GHDM

Generalized Heterogeneous Data Model

# Structural Model

- $L$ latent constructs $(z_1^*, z_2^*, \ldots, z_L^*)$

$$z_l^* = \boldsymbol{\alpha}_l' \boldsymbol{w} + \eta_l$$

$z_l^*$ Latent construct
$\boldsymbol{w}$ : Individual-specific covariates
$\boldsymbol{\alpha}_l$ : Vector of coefficients
$\eta_l$ : Standard multivariate normally distributed random error term

- Define

$$\boldsymbol{\alpha} = (\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2, \ldots \boldsymbol{\alpha}_L)', \; \boldsymbol{z}^* = (z_1^*, z_2^*, \ldots, z_L^*)', \; \boldsymbol{\eta} = (\eta_1, \eta_2, \eta_3, \ldots, \eta_L)'.$$

- Then

$$\boldsymbol{z}^* = \boldsymbol{\alpha}\boldsymbol{w} + \boldsymbol{\eta}, \quad \boldsymbol{\eta} \sim \mathrm{MVN}_L(\mathbf{0}_L, \boldsymbol{\Gamma})$$

where $\boldsymbol{\Gamma}$ is a $(L \times L)$ correlation matrix

# Measurement Model

$H$ continuous variables $(y_1, y_2, ..., y_H)$ $\boldsymbol{y} = \boldsymbol{\gamma x} + \boldsymbol{dz}^* + \boldsymbol{\varepsilon}$

$N$ ordinal variables $(\tilde{y}_1^*, \tilde{y}_2^*, ..., \tilde{y}_N^*)$ $\tilde{\boldsymbol{y}}^* = \tilde{\boldsymbol{\gamma}}\boldsymbol{x} + \tilde{\boldsymbol{d}}\boldsymbol{z}^* + \tilde{\boldsymbol{\varepsilon}}, \quad \tilde{\boldsymbol{\psi}}_{low} < \tilde{\boldsymbol{y}}^* < \tilde{\boldsymbol{\psi}}_{up}$

$C$ count variables $(\breve{y}_1^*, \breve{y}_2^*, ..., \breve{y}_C^*)$ $\breve{\boldsymbol{y}}^* = \breve{\boldsymbol{d}}\boldsymbol{z}^* + \breve{\boldsymbol{\varepsilon}}, \quad \breve{\boldsymbol{\psi}}_{low} < \breve{\boldsymbol{y}}^* < \breve{\boldsymbol{\psi}}_{up}$

$\boldsymbol{\gamma}, \tilde{\boldsymbol{\gamma}}$ and $\breve{\boldsymbol{\gamma}}$ : Vector of coefficients capturing effect of exogenous and possible endogenous variables

$\boldsymbol{d}, \tilde{\boldsymbol{d}},$ and $\breve{\boldsymbol{d}}$ : Vector of latent variable loading on dependent variables

$\boldsymbol{\varepsilon}, \tilde{\boldsymbol{\varepsilon}},$ and $\breve{\boldsymbol{\varepsilon}}$ : Vector of error terms of dependent variables

$\tilde{\boldsymbol{\psi}},$ and $\breve{\boldsymbol{\psi}}$ : Vector of thresholds for ordinal, and count outcomes

# Nominal Choice Model

$$U_i = \boldsymbol{b}_i' \boldsymbol{x} + \boldsymbol{\vartheta}_i'(\boldsymbol{\beta}_i \boldsymbol{z}^*) + \varsigma_i,$$

$$\boldsymbol{U} = \tilde{\boldsymbol{b}}\boldsymbol{x} + \tilde{\boldsymbol{\varpi}}\,\boldsymbol{z}^* + \tilde{\boldsymbol{\varsigma}},$$

$$\tilde{\boldsymbol{\varpi}} = (\boldsymbol{\vartheta}\boldsymbol{\beta}), \qquad \tilde{\varsigma} \sim MVN_I(\mathbf{0}\ , \tilde{\Lambda})$$

Choice model in utility difference form:

$$\boldsymbol{u} = \mathbf{M}\boldsymbol{U} = \mathbf{M}\tilde{\boldsymbol{b}}\boldsymbol{x} + \mathbf{M}\tilde{\boldsymbol{\varpi}}\,\boldsymbol{z}^* + \mathbf{M}\tilde{\boldsymbol{\varsigma}} = \boldsymbol{b}\boldsymbol{x} + \boldsymbol{\varpi}\,\boldsymbol{z}^* + \boldsymbol{\varsigma},$$

$$\boldsymbol{b} = \mathbf{M}\tilde{\boldsymbol{b}},\ \boldsymbol{\varpi} = \mathbf{M}\tilde{\boldsymbol{\varpi}}\text{, and } \boldsymbol{\varsigma} = \mathbf{M}\tilde{\boldsymbol{\varsigma}}.\ \ \boldsymbol{\varsigma} \sim MVN_{I-1}(\mathbf{0}\ , \boldsymbol{\Lambda})$$

$\boldsymbol{b}_i$ : Vector of coefficients capturing effect of exogenous and possible endogenous variables

$\boldsymbol{\vartheta}_i$ : Matrix of 1 and 0 to determine the loading of latent constructs on the alternatives

$\tilde{\boldsymbol{\varpi}}$ : Vector of coefficients on latent constructs

$\tilde{\boldsymbol{\varsigma}}$ : Random error vector distributed multivariate normal

$I$ : Number of alternatives

# MDC Model

## Utility function

$$\max \tilde{U}(\boldsymbol{t}) = \sum_{k=1}^{K-1} \frac{\tau_k}{\alpha_k} \psi_k \left( \left( \frac{t_k}{\tau_k} + 1 \right)^{\alpha_k} - 1 \right) + \frac{1}{\alpha_K} \psi_K (t_K)^{\alpha_K}$$

$$s.t. \sum_{k=1}^{K} t_k = T$$

$\tilde{U}(\boldsymbol{t})$ : A quasi-concave, increasing, and continuously differentiable function with respect to the consumption quantity (time investment) vector $\boldsymbol{t}$

$\tau_k, \alpha_k$ and $\psi_k$ : parameters associated with an activity purpose $k$

Further, specify the baseline utility as a function of exogenous variables and latent variables as follows:

$$\psi_k = \exp(\boldsymbol{x}, \boldsymbol{z}^*, \tilde{\xi}_k) = \exp(\tilde{\boldsymbol{\delta}}_k' \boldsymbol{x}_k + \tilde{\boldsymbol{\mu}}_k' \boldsymbol{z}^* + \tilde{\xi}_k) \text{ or } \overline{\psi}_k^* = \ln(\psi_k) = \tilde{\boldsymbol{\delta}}_k' \boldsymbol{x}_k + \tilde{\boldsymbol{\mu}}_k' \boldsymbol{z}^* + \tilde{\xi}_k$$

$\tilde{\xi}_k$: normally distributed random error term, which captures the idiosyncratic characteristics that impact baseline utility of activity purpose $k$

# KKT First Order Conditions

## Lagrangian Function

$$L = \sum_{k=1}^{K-1} \frac{\tau_k}{\alpha_k} \exp(\boldsymbol{\delta}_k' \boldsymbol{x}_k + \boldsymbol{\mu}_k' \boldsymbol{z} + \xi_k) \left( \left( \frac{t_k}{\tau_k} + 1 \right)^{\alpha_k} - 1 \right) + \frac{1}{\alpha_K} \psi_K (t_K)^{\alpha_K} - \tilde{\lambda} \left[ \sum_{k=1}^{K} t_k - T \right]$$

## KKT first-order conditions for optimal time investments

$$\exp(\boldsymbol{\delta}_k' \boldsymbol{x}_k + \boldsymbol{\mu}_k' \boldsymbol{z} + \xi_k) \left( \frac{t_k^*}{\tau_k} + 1 \right)^{\alpha_k - 1} - \tilde{\lambda} = 0, \text{if } t_k^* > 0, \quad k = 1,2,\ldots,K-1$$

$$\exp(\boldsymbol{\delta}_k' \boldsymbol{x}_k + \boldsymbol{\mu}_k' \boldsymbol{z} + \xi_k) \left( \frac{t_k^*}{\tau_k} + 1 \right)^{\alpha_k - 1} - \tilde{\lambda} < 0, \text{if } t_k^* = 0 \quad k = 1,2,\ldots,K-1$$

# Overall Model System

$\Gamma$ : Correlation matrix of structural equation error terms $(\boldsymbol{\eta})$

$\boldsymbol{\Sigma}$ : Covariance matrix of continuous variables error terms $(\boldsymbol{\varepsilon})$

$\mathbf{IDEN_C}$ : Identity matrix of size c

$$\overline{\boldsymbol{\Sigma}} = \begin{bmatrix} \boldsymbol{\Sigma} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{IDEN_N} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{IDEN_C} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \boldsymbol{\Lambda} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \boldsymbol{\Omega} \end{bmatrix}$$

$A =$ # of exogenous and endogenous variables

$E = (H + N + C)$

$\bar{y} = \left( y', [\tilde{y}^*]', [\breve{y}^*]' \right)' \; [E \times 1 \; vector]$

$\bar{\gamma} = (\gamma', \tilde{\gamma}', \boldsymbol{0}_{AC})' \; [E \times A \text{ matrix}]$

$\bar{d} = (d', \tilde{d}', \breve{d}')' \; [E \times L \text{ matrix}]$

$\bar{\varepsilon} = (\varepsilon', \tilde{\varepsilon}', \breve{\varepsilon}')' \; (E \times 1 \text{ vector})$

## GHDM-MDCP Model

$$yu = \tilde{V} + \pi z^* + \kappa$$

$$yu = \tilde{V} + \pi z^* + \kappa = \tilde{V} + \pi(\alpha w + \eta) + \kappa = \tilde{V} + \pi\alpha w + \pi\eta + \kappa$$

Then $yu \sim MVN_{E+I+K-2}(B, \Theta)$,

where $B = \tilde{V} + \pi\alpha w$, and $\Theta = \pi\Gamma\pi' + \overline{\Sigma}$

$\tilde{V} = [(\bar{\gamma}x)', (bx)', V']'$, $\pi = (\bar{d}', \varpi', \mu')'$, $\kappa = (\bar{\varepsilon}', \varsigma', \xi')'$

$yu = (\bar{y}', u', \tilde{u}')' \left[ (E + I + K - 2) \times 1 \text{ vector} \right]$

$\tilde{E} = E + I + K - 2$

# Likelihood function

$$L(\vec{\boldsymbol{\theta}}) = \det(\mathbf{J}) \times f_{H+\tilde{F}_C}\left((\boldsymbol{y}, \mathbf{0}_{\tilde{F}_C}) \mid \tilde{\boldsymbol{B}}_1, \tilde{\boldsymbol{\Theta}}_1\right) \times \Pr\left[\vec{\boldsymbol{\psi}}_{low} \le \boldsymbol{y}\tilde{\boldsymbol{u}}_2 \le \vec{\boldsymbol{\psi}}_{up}\right],$$

$$= \det(\mathbf{J}) \times f_{H+\tilde{F}_C}\left((\boldsymbol{y}, \mathbf{0}_{\tilde{F}_C}) \mid \tilde{\boldsymbol{B}}_1, \tilde{\boldsymbol{\Theta}}_1\right) \times \int_{D_r} f_{N+I+\tilde{F}_{NC}}\left(\boldsymbol{r} \mid \vec{\boldsymbol{B}}_2, \vec{\boldsymbol{\Omega}}_2\right) dr$$

- The maximum likelihood estimation of the model involves the evaluation of an $(N + I + \tilde{F}_{NC})$ -dimensional rectangular integral for each decision-maker
- So, we use the Maximum Approximate Composite Marginal Likelihood (MACML) approach of Bhat (2011).

$$L_{CML}(\vec{\boldsymbol{\theta}}) = f_H(\boldsymbol{y} \mid \tilde{\boldsymbol{B}}_y, \tilde{\boldsymbol{\Omega}}_y) \times \left( \prod_{n=1}^{N-1} \prod_{n'=n+1}^{N} \Pr(j_n = a_n, j_{n'} = a'_n) \right) \times$$

$$\times \left( \prod_{n=1}^{N} \Pr(j_n = a_n, g = r) \right) \times \left( \prod_{n=1}^{N} \Pr(j_n = a_n, i = m) \right) \times \left( \Pr(g = r, i = m) \right)$$

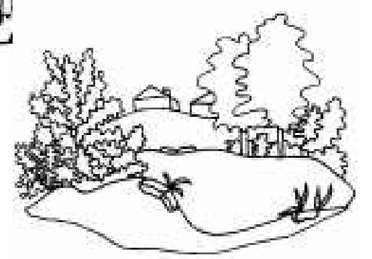
$$\times \left( \prod_{n=1}^{N} \Pr\left(\boldsymbol{t}^*; j_n = a_n\right) \right) \times \Pr\left(\boldsymbol{t}^*; g = r\right) \times \Pr\left(\boldsymbol{t}^*; i = m\right)$$

# Aplicación:

Efectos de **auto-selección residencial** en un modelo de uso de tiempo y actividades

# Problema de auto-selección residencial

- Efecto del entorno construido (BE) en demanda de transporte → casual? asociativo? ambos?
- Usualmente datos de sección transversal → Localización residencial y patrones de viajes son observados conjuntamente
- Buscamos el "verdadero" efecto del BE
- Algunas características de personalidad y estilo de vida (no observables) pueden impactor ambos: localización residencial y comportamiento de actividades-viajes.
- Ignorar estos efectos de auto-selección puede llevar a sobre-estimar el efecto causal de atributos del entorno en el comportamiento de acitividades-viajes
- Potenciales políticas de diseño de BE desinformadas



# Una solución al problema de auto-selección residencial

- Capturar lo “inobservable” tradicionalmente en encuestas e incluirlo como variables explicatorias “observadas”

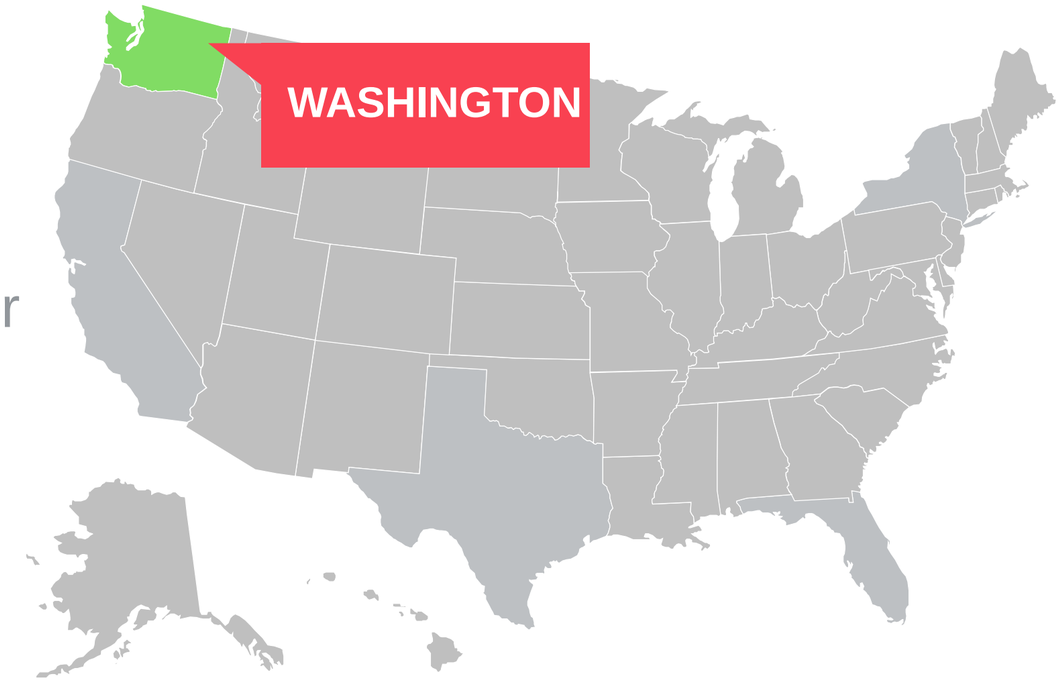
- Ejemplos de estas variables “blandas”:
  - Consciencia medioambiental
  - Inclinación a una vida activa
  - Nivel de stress/ansiedad
  - Nivel de flexibilidad en el proceso de toma de decisiones
  - Conocimiento del transporte público
  - Inclinación por una vida de lujos
  - Familiaridad con la tecnología

# Datos y variables dependientes

- Puget Sound household travel survey 2014

- Unidad de análisis: hogar (con al menos un trabajador que trabajaba fuera del hogar)

- Tamaño muestral: 3,637




- **Variables dependientes**
  - Densidad residencial (hh/sq. mile) : menos que 750; 750-1,999; 2,000-2,999; 3,000 o más (variable nominal)
  - Promedio de distancia al trabajo/estudio en millas (variable continua)
  - Posesión de automóvil (variable de conteo)
  - Fracción del tiempo invertido participando en las siguientes actividades: Tramites personales, compras, recreación, comer afuera, social, pasar a buscar/dejar, en casa (variable MDC)

| Variable dependiente continua | | | | |
|---|---|---|---|---|
| **Variable** | Mean | Std. Dev. | Min. | Max. |
| **Distancia al trabajo promedio en el hogar** | 14.47 | 13.78 | 0.05 | 99.95 |

| Variable dependiente de conteo | | | | | | | |
|---|---|---|---|---|---|---|---|
| **Número de vehículos** | Frequencia | | | | | | |
| | 0 | 1 | 2 | 3 | 4 | 5 | >6 |
| **Number** | 304 | 1,378 | 1,354 | 413 | 135 | 36 | 17 |
| **%** | 8.4 | 37.8 | 37.2 | 11.4 | 3.7 | 1.0 | 0.5 |

| Variable dependiente nominal | |
|---|---|
| **Densidad residencial (hogares por sq. mile)** | Número de observaciones (%) |
| **<750** | 478 (13.2) |
| **750-2,000** | 866 (23.8) |
| **2,000-3,000** | 525 (14.4) |
| **>3,000** | 1,768 (48.6) |

| Variable dependiente MDC | | | | |
|---|---|---|---|---|
| **Actividad** | Participación (%) | Fracción promedio | Número de hogares (% del total) que asignan tiempo… | |
| | | | Sólo a este tipo de actividad | En otras actividades también |
| **En el hogar** | 3,637 (100.0) | 0.780 | 533 (14.7) | 3,104 (85.3) |
| **Tramites personales** | 1,607 ( 44.2) | 0.202 | 216 (13.4) | 1,391 (86.6) |
| **Compras** | 1,664 ( 45.8) | 0.060 | 355 (21.3) | 1,309 (78.7) |
| **Recreación** | 1,011 ( 27.8) | 0.131 | 148 (14.6) | 863 (85.4) |
| **Comer afuera** | 1,092 ( 30.0) | 0.081 | 203 (18.6) | 889 (81.4) |
| **Social** | 659 ( 18.1) | 0.180 | 82 (12.4) | 557 (87.6) |
| **Pasar a buscar/dejar** | 751 ( 20.6) | 0.047 | 26 ( 3.5) | 725 (96.5) |

# Variables latentes

## ❑Inclinación a un estilo de vida verde

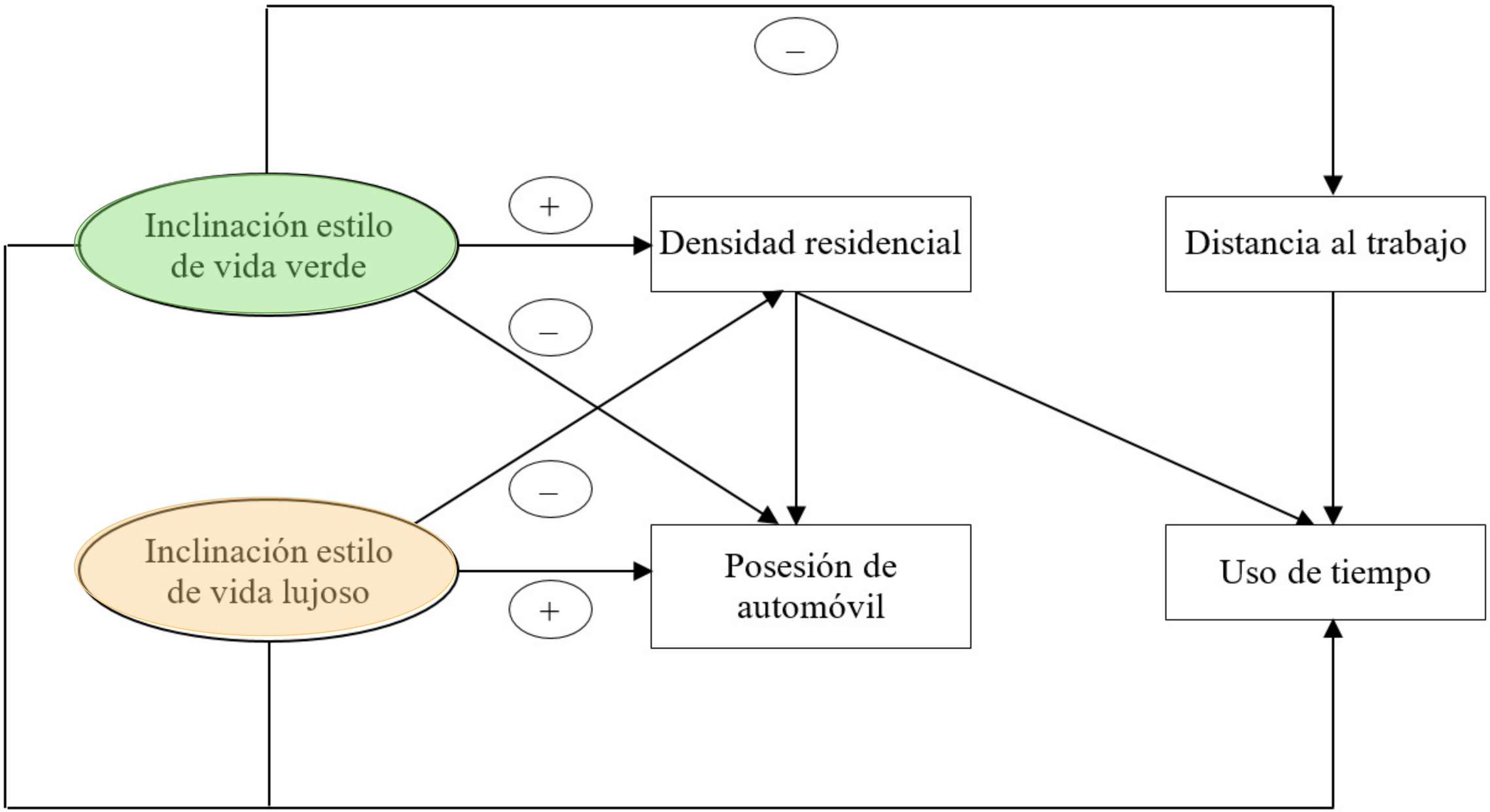




## ❑Inclinación a un estilo de vida lujoso




# Efectos de variables latentes en el conjunto de decisiones



# Examinando el “verdadero” efecto de políticas neo-urbanistas de densificación

Efecto de transferir un hogar al azar desde un vecindario de muy baja densidad (<750 hh/sq. mile) a otro de muy alta densidad (>3000 hh/sq. mile) (standard error en paréntesis)

| Variable | ATE del GHDM | ATE del IHDM | % de la diferencia atribuible a | |
|---|---|---|---|---|
| | | | “Verdadero” efecto | Efecto de auto-selección |
| **Posesión de automovil** | 0.143 (0.011) | 0.340 (0.021) | 42 | 58 |
| **Participación en** | | | | |
| **Tramites personales** | -0.037 (0.013) | -0.041 (0.013) | 90 | 10 |
| **Compras** | 0.011 (0.004) | 0.019 (0.007) | 65 | 35 |
| **Recreación** | 0.134 (0.021) | 0.190 (0.014) | 71 | 29 |
| **Comer afuera** | 0.094 (0.020) | 0.119 (0.021) | 79 | 21 |
| **Social** | -0.056 (0.014) | -0.078 (0.017) | 72 | 28 |
| **Pasar a buscar/dejar** | -0.156 (0.033) | -0.162 (0.025) | 96 | 4 |

# Conclusiones

# Conclusiones

- Proponemos y aplicamos un marco integrado para modelar multiples tipos de variables (incluyendo variables continuas, ordinals, conteo, nominales, and MDC)
- Esperamos que este enfoque
  - Abrirá nuevas puertas en la exploración del nexo entre distintas variables en diversas áreas de investigación
  - Tendrá vital importancia en este nuevo mundo de datos masivos

# MUCHAS GRACIAS

¿Preguntas?

**EMAIL**
sastroza@udec.cl