Ciencia de Datos

Modelos de Elección Discreta

Roberto Ponce López

Tecnológico de Monterrey

rpl@tec.mx

10 de febrero del 2020

Tabla de Contenidos

Marco Teórico de la Elección Discreta

- Tomador de Decisión
  - Individuos, personas o hogares
  - Características Socioeconómicas: edad, educación, ingreso
- Cuando no podemos aleatorizar, tenemos que conducir estudios observacionales con datos del mundo real
- Alternativas (ϵ)
  - El tomador de decisión n elige solamente una alternativa de un set the decisión $C_n = \{1, 2, 3..., i, ...J_n\}$ con J_n alternativas.
- Atributos de las alternativas (precio y calidad)
- Regla de decisión (maximizar utilidad)

Ejemplo: Elección de Automóviles

- Tomador de Decisión: individuos u hogares

Ejemplo: Elección de Automóviles

- Tomador de Decisión: individuos u hogares
- Choice set: alternativas de automóviles

Ejemplo: Elección de Automóviles

- Tomador de Decisión: individuos u hogares
- Choice set: alternativas de automóviles
- Regla de decisión: maximización de utilidad

Ejemplo: Elección de Automóviles

- Tomador de Decisión: individuos u hogares
- Choice set: alternativas de automóviles
- Regla de decisión: maximización de utilidad
- Función de utilidad: función de los atributos del automóvil, tales como precio y calidad

Ejemplo: Elección de Automóviles

- Tomador de Decisión: individuos u hogares
- Choice set: alternativas de automóviles
- Regla de decisión: maximización de utilidad
- Función de utilidad: función de los atributos del automóvil, tales como precio y calidad
- Características individuales

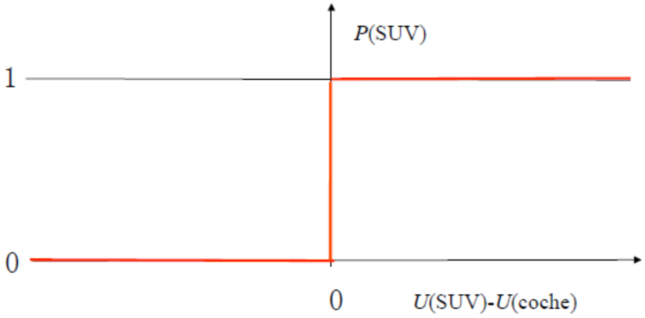
Elección del Tomador de Decisión

- Maximización de Utilidad
- El tomador de decisión elige la alternativa que le reporta la mayor utilidad (y que está dentro del set de alternativas)
- Si $U(SUV) > U(coche)$, elige la SUV
- Si $U(SUV) < U(coche)$, elige el coche
- ¿Cómo calculamos estas utilidades?

Construyendo la Función de Utilidad

- $U_{SUV} = U(precio_{SUV}, comfort_{SUV})$
- $U_{coche} = U(precio_{coche}, comfort_{coche})$
- Asumimos linealidad en los parámetros:
  $U(SUV) = \beta_1 \cdot precio_{SUV} + \beta_2 \cdot comfort_{SUV} + \ldots$
- Los parámetros representan gustos, los cuales varían entre personas
  - Incluye características socioeconómicas como tamaño del hogar, edad e ingreso
  - $U_{SUV} = U(precio_{SUV}/ingreso + \ldots)$

- Si $(U_{SUV} - U_{coche}) > 0$, Probabilidad(SUV)=1
- Si $(U_{SUV} - U_{coche}) < 0$, Probabilidad(SUV)=0



Elección Binaria Probabilística

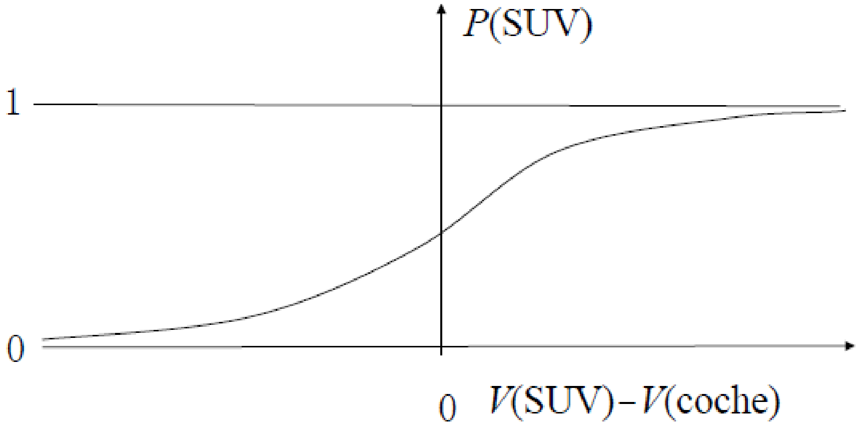
- Modelo de Utilidad Aleatoria (Random Utility Model - RUM)

$$U_i = V(\text{Atributos de } i; parametros) + \epsilon_i$$

- ¿Qué hay en la ϵ
  1. Atributos no observados
  2. Cambios en preferencias no observados
  3. Errores de medición

$$U(SUV) = \beta_1 \cdot (precio_{SUV}/ingreso) + \beta_2(comfort_{SUV}) + \epsilon_{SUV}$$

Elección Binaria Probabilística



Modelo de Utilidad Aleatoria

- Regla de decisión: maximización de utilidad
- El individuo n selecciona la alternativa con el pago más alto de utilidad U_{in}, entre aquéllas en el set de decisión C_n
- Utilidad: $U_{in} = V_{in} + \epsilon_{in}$
- Utilidad: V_{in} es el componente sistemático de utilidad, expresado como una función de k variables observables, e.g. $\sum_k \beta_k x_{ink}$
- ϵ_{in} es el componente aleatorio de la utilidad

Modelo de Utilidad Aleatoria

- Probabilidad de Elección

$$\begin{aligned} P(i|C_n) &= P(U_{in} \geq U_{jn}), \forall j \in C_n \\ &= P(U_{in} - U_{jn} \geq 0), \forall j \in C_n \\ &= P(U_{in}) = max_j U_{jn}, \forall j \in C_n \end{aligned}$$

- En el caso de elección binaria:

$$\begin{aligned} P_n(1) &= P(U_{1n} \geq U_{2n}) \\ &= P(U_{1n} - U_{2n} \geq 0) \end{aligned} \tag{1}$$

Ejemplo con Atributos

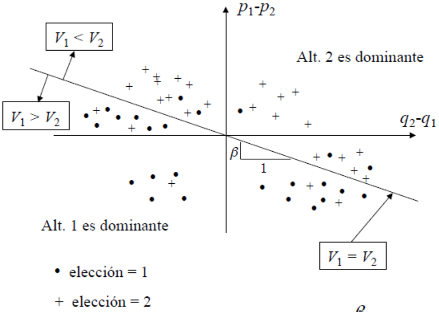
- Utilidades ordinales
- Las decisiones se construyen a partir de diferencias en las utilidades de las alternativas

$$U_1 = (\beta_1 q_1 - \beta_2 p_1 + \epsilon_1 + \alpha)\mu$$

$$U_2 = (\beta_1 q_2 - \beta_2 p_2 + \epsilon_2 + \alpha)\mu$$

$$\beta_1, \beta_2, \mu > 0$$

Ejemplo con Atributos



$$U_1 = \frac{\beta_1}{\beta_2} q_1 - p_1 + \varepsilon_1$$

$$U_2 = \frac{\beta_1}{\beta_2} q_2 - p_2 + \varepsilon_2$$

$$\beta = \frac{\beta_1}{\beta_2} = \text{"valor de comfort"}$$

$$V_1 = \frac{\beta_1}{\beta_2} q_1 - p_1$$

$$V_2 = \frac{\beta_1}{\beta_2} q_2 - p_2$$

$$U_1 - U_2 = -\frac{\beta_1}{\beta_2}(q_2 - q_1) - (p_1 - p_2) + (\varepsilon_1 - \varepsilon_2)$$

Utilidad Sistemática ($V_i n$)

- Atributos: describen a una alternativa
- Los atributos pueden ser genéricos o específicos por alternativa (ejemplos: kilómetros por litro de gasolina, interiores de piel o tela)

Utilidad Sistemática ($V_i n$)

- Atributos: describen a una alternativa
- Los atributos pueden ser genéricos o específicos por alternativa (ejemplos: kilómetros por litro de gasolina, interiores de piel o tela)
- Características: describen al tomador de la decisión o de la elección
- Las características típicamente son variables socioeconómicas

Fuentes de datos sobre comportamiento de usuarios

- Datos de Preferencias Reveladas
  - Base de datos de clientes o usuarios
  - Encuestas
  - Decisiones observadas de casos concretos

Fuentes de datos sobre comportamiento de usuarios

- Datos de Preferencias Reveladas
  - Base de datos de clientes o usuarios
  - Encuestas
  - Decisiones observadas de casos concretos
- Datos de Preferencias Establecidas
  - Basado en encuestas
  - El entrevistado establece valuaciones a alternativas que se le brindan

Recolección de datos

La recolección de datos para cada individuo en la muestra:

- Set de Decisión: alternativas disponibles
- Características Socio-demográficas
- Atributos de las alternativas disponibles
- Decisión del individuo n sobre la alternativa i

n	Edad	precio_coche	precio_suv	Elección
1	35	15.4	58.2	coche
2	45	14.2	31.0	suv
3	37	19.6	43.6	coche
4	42	50.8	59.9	suv
5	32	55.5	33.8	coche
6	15	N/A	48.4	suv

Estimación por Máxima Verosimilitud

La recolección de datos para cada individuo en la muestra:

- Encuentra los valores de β que es más probable que resulten en las elecciones o decisiones observadas en la muestra:

$$y_n = \begin{cases} 1 & \text{si el individuo } n \text{ elige la alternativa 1;} \\ 0 & \text{si el individuo } n \text{ elige la alternativa 2.} \end{cases}$$

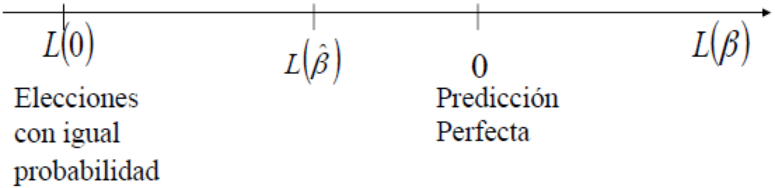
- Poteriormente, maximizamos sobre $\{\beta_1, \beta_2, ..., \beta_K\}$, la siguiente expresión (función de verosimilitud):

$$L(\beta_1, \beta_2, ..., \beta_k) = \prod_{n=1}^{N} P_n(1)^{y_n} \cdot P_n(2)^{(1-y_n)}$$

- Log-Máxima Verosimilitud: $L(\beta_1, \beta_2, ..., \beta_K)$
- Estimador de máxima verosimilitud: $\hat{\beta} = argmax_\beta L(\beta_1, \beta_2, ..., \beta_K)$

Máxima Verosimilitud

- Encontramos los estimadores de β que maximizan el logaritmo de la función de verosimilitud
- La función tiene un valor negativo



Goodness of Fit

- ρ y $\bar{\rho}^2$ son análogos a la R y R^2
- El índice de máxima verosimilitud ρ^2 mide la fracción de la verosimilitud inicial es explicada por el modelo

$$\rho^2 = 1 - \frac{L(\hat{\beta})}{L(0)}$$

- El valor de ρ^2 es entre 0 y 1

Modelo de Utilidad Aleatoria

- Regla de decisión: maximización de utilidad
- El individuo n selecciona la alternativa que le reporta la mayor utilidad U_{in} entre las alternativas en el set C_n
- Utilidad: $U_{in} = V_{in} + \epsilon_{in}$
- El componente sistemático V_{in} de la utilidad se expresa como una función de las variables observables:

$$V_{in} = \beta' X_{in} = \sum_{k=1}^{K} B_k X_{ink}$$

- ϵ_{in}: componente aleatorio de utilidad

Modelo de Utilidad Aleatoria

- Probabilidad de Elección

$$
\begin{aligned}
P(i|C_n) &= P(U_{in} \geq U_{jn}), \forall j \in C_n \\
&= P(U_{in} - U_{jn} \geq 0), \forall j \in C_n \\
&= P(U_{in}) = max_j U_{jn}, \forall j \in C_n
\end{aligned}
$$

- Para elección binaria:

$$
\begin{aligned}
P_n(1) &= P(U_{1n} \geq U_{2n}) \\
&= P(U_{1n} - U_{2n} \geq 0) \\
&= P(\epsilon_{2n} - \epsilon_{1n} \leq V_{1n} - V_{2n}) \\
&= F_{\epsilon_2 - \epsilon_1}(V_{1n} - V_{2n})
\end{aligned} \tag{2}
$$

Distribución de los Términos de Utilidad Aleatoria: Valor Extremo

- Distribución de los errores de valor extremo
- Supuesto:los términos del error son el máximo de muchas variables aleatorias que capturan atributos no observados (e.g. habilidad, estatus de la marca) y también errores de medición
- Teorema de Gumbel: el máximo de muchas variables independientes e idénticamente distribuidas siguen una distribución de valor extremo
- $\epsilon_{in} \sim EV(0, \mu)$

Distribución de los Términos de Utilidad Aleatoria: Valor Extremo

- Asumimos que:

$$\epsilon_1 \sim EV(0, \mu)$$

$$\epsilon_2 \sim EV(0, \mu)$$

- ϵ_1 y ϵ_2 son independientes

Distribución de los Términos de Utilidad Aleatoria: Valor Extremo

- Asumimos que:

$$\epsilon_1 \sim EV(0, \mu)$$

$$\epsilon_2 \sim EV(0, \mu)$$

- ϵ_1 y ϵ_2 son independientes
- La porción de utilidad no observada de una alternativa no está relacionada con la porción no observada de la utilidad de otra alternativa
- De esta forma, $\epsilon = \epsilon_1 + \epsilon_2 \sim Logistic(0, \mu)$
- Resultando que:

$$P(1|\{1,2\}) = \frac{1}{1 + e^{\mu(V_1 - V_2)}}$$

- Tenemos una probabilidad con forma cerrada

Extensión a más de Dos Alternativas

Set de alternativas C_n: J_n alternativas, $J_n \geq 2$

$$\begin{aligned} P(i|C_n) &= P[V_{in} + \epsilon_{in} \geq V_{jn} + \epsilon_{jn}, \forall j \in C_n] \\ &= P[\epsilon_{jn} - \epsilon_{in} \leq V_{in} - V_{jn}, \forall j \in C_n] \end{aligned} \quad (3)$$

Ejemplo: caso con tres alternativas

$$\begin{aligned}
P_n(1) = P(1|C_n) &= P(U_{1n}+ \geq U_{2n} \text{ y } U_{1n} \geq U_{3n}) \\
&= P(V_{1n} + \epsilon_{1n} \geq V_{2n} + \epsilon_{2n} \text{ y } V_{1n} + \epsilon_{1n} \geq V_{3n} + \epsilon_{3n}) \\
&= P(\epsilon_{2n} - \epsilon_{1n} \leq V_{1n} - V_{2n} \text{ y } \epsilon_{3n} - \epsilon_{1n} \leq V_{1n} - V_{3n}) \\
&= F_{\epsilon_2 - \epsilon_1, \epsilon_3 - \epsilon_1}(V_{1n} - V_{2n}, V_{1n} - V_{3n}) \qquad (4)
\end{aligned}$$

Propiedades del Logit

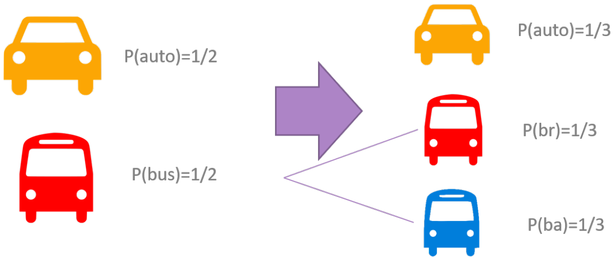
- Independencia de Alternativas Irrelevantes
- El modelo

$$P_n(i|C_n) = \frac{e^{\mu V_{in}}}{\sum_{j \in C_n} e^{\mu V_{jn}}}$$

- El parámetro μ de escala de las utilidades

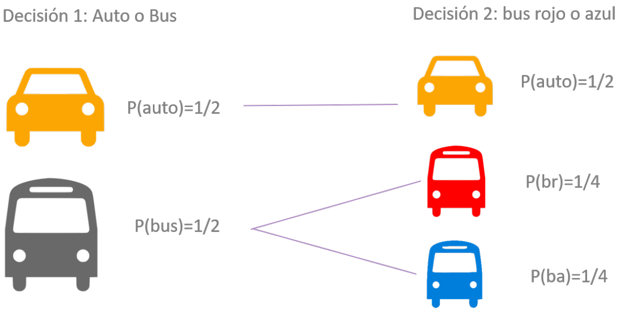
Independencia de Alternativas Irrelevantes

- Paradoja del autobús rojo y azul
- Elección de modo de transporte



Independencia de Alternativas Irrelevantes

- Solución: estructura de decisión anidada
- Nested Logit



Nested Logit

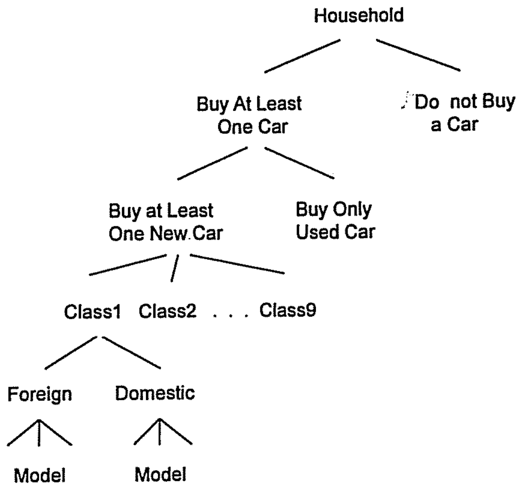


Figura: Modelo de decisión de un autómovil. Fuente: Pinelopi, Koujianou, Goldberg (1995)