MODELOS DE RESPUESTA NO ORDENADA (NOMINAL)

ORLANDO JOAQUI BARANDICA

Universidad del Valle

Introducción

Las respuestas nominales se encuentran en múltiples áreas de las ciencias sociales: Economía laboral, de la educación, del transporte, redes sociales.

Es usual la utilización de los modelos de respuesta nominal aun cuando la variable dependiente es ordinal, algunas razones:

- Para eludir el supuesto de regresiones paralelas.
- Por ambigüedad al decidir si la variable dependiente es nominal u ordinal
- Por facilidad, el investigador está más familiarizado con el modelo nominal.

Introducción

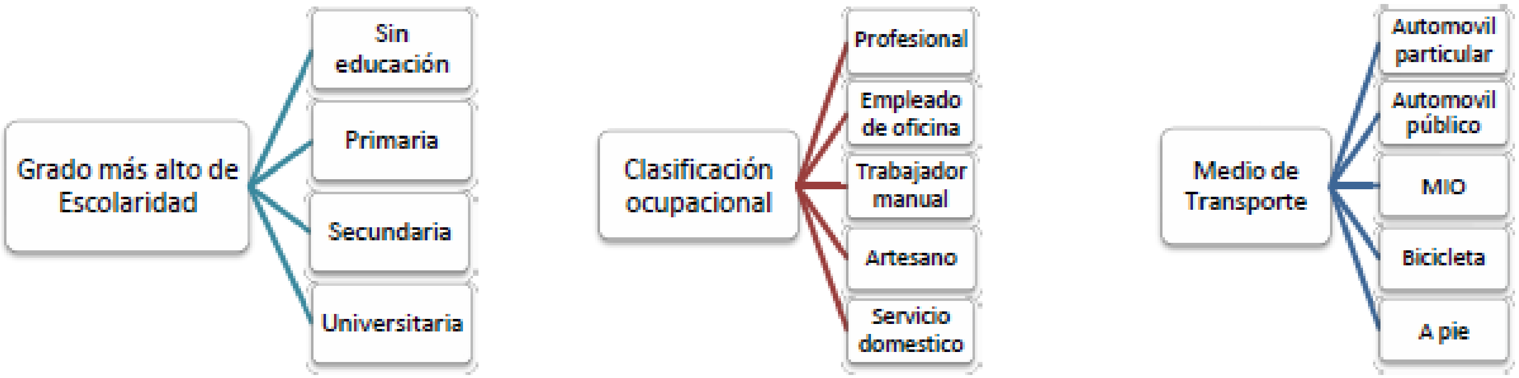
Si la variable dependiente es ordinal y se utiliza un modelo para variables nominales se incurre en pérdida de eficiencia dado que se tiene información que está siendo ignorada.

Por el contrario, cuando un método de variables ordinales es aplicado a una variable nominal, los estimadores resultantes serán sesgados.

En resumen, si la variable dependiente es ordinal, la potencial perdida de eficiencia causada por el uso de una técnica nominal es mucho más fuerte que evitar el potencial sesgo.

Introducción

Se presenta cuando el proceso de elección – o las categorías de la variable dependiente – no implica ningún ordenamiento.



- Ingreso del hogar,
- Edad,
- Escolaridad de los padres,
- Nro. de hermanos,
- Raza.
- Género

- Nivel de educación,
- Edad,
- Experiencia,
- Género,
- Posición en el hogar
- Estado civil

- Nivel económico,
- Años de escolaridad,
- Edad,
- Distancia a recorrer
- Tiempo de desplazamiento
- Costo del viaje

Modelo Logit Multinomial

El modelo logit multinomial (MLM) puede ser pensado como la estimación simultánea de modelos logit binarios para todas las comparaciones posibles de las alternativas.

Por tanto, el MLM es una simple extensión del modelo logit binario, sin embargo, la dificultad viene del hecho, que son muchas las alternativas a comparar.

Modelo Logit Multinomial

Ejemplo: si la variable categórica tiene tres alternativas, el MLM es equivalente a estimar tres logit binarios que comparan las categorías 1 y 2, 1 y 3, y 2 y 3.

Sean las alternativas A, B y C de Yi con nA, nB y nC observaciones cada una.

Para analizar la relación entre la variable dependiente categórica (Yi) y una única variable explicativa (Xi) se debe estimar una serie de regresiones binarias.

Modelo Logit Multinomial

Utilizando Odds Ratios, la estimación del efecto de Xi sobre la razón A vs B, con nA + nB observaciones, se hace a partir del modelo logit binario:

$$ln\left[\frac{\Pr\left(Y_i = A|X_i\right)}{\Pr\left(Y_i = B|X_i\right)}\right] = \beta_{1,A|B} + \beta_{2,A|B}X_i$$

La variable dependiente de este modelo es el logaritmo del Odds. Un incremento de una unidad en Xi aumenta el odds en un factor

$$exp\left(\beta_{2,A|B}\right) = e^{\beta_{2,A|B}}$$

Las otras comparaciones pueden ser obtenidas de la misma forma:

$$ln\left[\frac{\Pr\left(Y_i = B|X_i\right)}{\Pr\left(Y_i = C|X_i\right)}\right] = \beta_{1,B|C} + \beta_{2,B|C}X_i; \qquad ln\left[\frac{\Pr\left(Y_i = A|X_i\right)}{\Pr\left(Y_i = C|X_i\right)}\right] = \beta_{1,A|C} + \beta_{2,A|C}X_i$$

Modelo Logit Multinomial

¿Son necesarios los tres logit binarios?

Si conocemos como afecta Xi el Odds de A vs B y el de B vs C, es razonable pensar que se tiene la información del Odds A vs C , dado que:

$$ln\left[\frac{\Pr\left(Y_i = A|X_i\right)}{\Pr\left(Y_i = B|X_i\right)}\right] - ln\left[\frac{\Pr\left(Y_i = B|X_i\right)}{\Pr\left(Y_i = C|X_i\right)}\right] = ln\left[\frac{\Pr\left(Y_i = A|X_i\right)}{\Pr\left(Y_i = C|X_i\right)}\right]$$

La igualdad del lado izquierdo implica que:

$$\beta_{1,A|B} - \beta_{1,B|C} = \beta_{1,A|C}$$

$$\beta_{2,A|B} - \beta_{2,B|C} = \beta_{2,A|C}$$

Modelo Logit Multinomial

Es decir, algunas de las comparaciones son redundantes. No obstante hay algunas complicaciones.

Las anteriores igualdades describen una relación entre los parámetros en una población, relación que podría no mantenerse necesariamente en una muestra.

La razón es que los tres logit binarios son estimados sobre muestras diferentes, en el modelo multinomial todos los logit son estimados de manera simultánea, lo que sugiere un uso más eficiente de los datos.

Modelo Logit Multinomial

Este enfoque se fundamenta en la teoría de la utilidad aleatoria. Se supone un agente económico racional que elige la alternativa que le proporciona la mayor utilidad. El modelo se puede concebir como un problema de decisión en el sentido de que se debe elegir una opción entre un conjunto de alternativas.

Supongamos que $U_{i0}, U_{i1}, \ldots, U_{i(M-1}$ representan las utilidades asociadas a las alternativas, para el individuo i, con vector de variables

$$\boldsymbol{X}_i = \{X_{i1}, X_{i2}, \ldots, X_{ik}\}$$

que representa el conjunto de características personales asociados al individuo i.

Modelo Logit Multinomial

Asumiendo linealidad en las funciones de utilidad, de tal forma que:

$$U_{im} = V_{im} + \varepsilon_{im}, \quad \text{con } m = 0, 1, 2, \ldots, (M-1).$$

El individuo i selecciona una determinada alternativa si la utilidad que le proporciona es mayor que la utilidad que le proporciona el resto:

$$Y_i = \begin{cases} 0 & si \;\; U_{i0} > U_{is} \text{ para toda } s \neq 0 \\ 1 & si \;\; U_{i1} > U_{is} \text{ para toda } s \neq 1 \\ & \qquad \qquad \ldots \\ M-1 & si \;\; U_{i(M-1)} > U_{is} \text{ para toda } s \neq M-1 \end{cases}$$

Planteamiento del Modelo

Supongamos que M, el número de alternativas, ejemplo: medios de transporte; servicio público, vehículo particular, bicicleta o a pie, son etiquetadas por un índice m = 1, 2, …, M de forma que la respuesta sea una variable aleatoria nominal.

Un posible modelo definido en términos de la utilidad aleatoria, está dado por:

$$U_i^j = V_{ij} + \varepsilon_{ij} = \boldsymbol{X}_i'\boldsymbol{\beta}_j + \varepsilon_{ij}$$

con Xi un vector de variables explicativas para el individuo i. Beta_j un vector k*1 de parámetros para la alternativa j que mide el peso relativo de las características en la utilidad derivada. Vij = Xi*Beta_j representa la utilidad sistemática de la alternativa j para un individuo con vector de características Xi, y un término de error Eij de preferencias individuales.

Estimación

Suponiendo una variable dependiente con m alternativas, m = 1, 2, 3, …, M definimos la probabilidad de que la alternativa j sea elegida.

Asumiendo que el i-ésimo individuo escoge la alternativa para la cual la utilidad U_i^j es máxima de manera independiente, se tiene:

$$p_{ij} = Pr[Y_i = j|\boldsymbol{X}] = Pr\left[V_{ij} + \varepsilon_{ij} > V_{is} + \varepsilon_{is}\right] \text{ para toda } \quad j \neq s$$

con $V_{ij} = \boldsymbol{X}_i'\boldsymbol{\beta}_j$. Luego,

$$p_{ij} = \frac{e^{X_i'\beta_j}}{\sum_{m=1}^{M} e^{X_i'\beta_m}} \qquad m = 1,2, \dots, M$$