Notas de Econometría

Pavel Solís

2025

7 Análisis de Regresión Múltiple con Información Cualitativa: Variables Binarias

Hasta ahora: Variables cuantitativas (ej. $salario$, $educ$, $calif$, nox, $ventas$)

- Su magnitud de información útil

Aquí: Variables cualitativas (ej. género, raza, industria, regiones)

- Pueden aparecer como variables independientes o como variable dependiente

7.1 Descripción de información cualitativa

Variables cualitativas frecuentemente vienen en forma de información binaria

- Ej. hombre/mujer, doméstico/extranjero, cliente/no cliente

Definimos una variable binaria para capturar la información relevante

- Variable cero-uno, indicadora, dummy, ficticia
- Expanden el tipo de análisis que podemos hacer

Generalmente, asignamos un nombre para el evento que toma el valor 1

- Útil para ecuaciones y para una explicación clara
  - Ej. género vs hombre/mujer, origen vs doméstico/extranjero
- Presencia o ausencia de una característica
  - Ej. Personas (mujer) con estado civil (casada)

Los valores 0/1 son arbitrarios

- En principio, podemos usar 2 valores cualquiera
- Pero con 0/1 sus parámetros en RLM son interpretables

7.2 Una variable independiente dummy

Para incorporar información binaria en modelo de regresión, solo agregamos variable dummy como variable independiente

$$salario = \beta_0 + \delta_0 mujer + \beta_1 educ + u$$

- δ_0 para resaltar interpretación de parámetros de dummies por ahora
- Solo 2 variables afectan al salario: género y educación

- $mujer = 1$ si la persona es mujer, $mujer = 0$ si es hombre

Interpretación:

- δ_0 es diferencia promedio en salario por hora entre hombres y mujeres
  - Dada la misma cantidad de educación
- δ_0 determina si hay discriminación contra las mujeres
  - Si $\delta_0 < 0$, mujeres ganan menos por el mismo nivel de otros factores

Usando esperanza y suponiendo RLM.4 ($\mathbb{E}(u|mujer, educ) = 0$)

$$\mathbb{E}(salario|mujer = 1, educ) = \beta_0 + \delta_0 + \beta_1 educ$$
$$\mathbb{E}(salario|mujer = 0, educ) = \beta_0 + \beta_1 educ$$
$$\implies \delta_0 = \mathbb{E}(salario|mujer, educ) - \mathbb{E}(salario|hombre, educ)$$

- Mismo nivel de educ en ambos
- Única diferencia es por género

δ_0 es constante por lo que gráficamente se muestra como un cambio en el intercepto

- Líneas paralelas porque la diferencia no depende del nivel de educación
- Si $\delta_0 < 0$, hombres ganan una cantidad fija más que las mujeres

[Gráfica]

- ¿Qué supuesto se viola si incluimos una dummy *hombre*?

- β_0: Intercepto para hombres
- $\beta_0 + \delta_0$: Intercepto para mujeres
- Son 2 grupos por lo que solo se necesitan 2 interceptos diferentes
- Entonces, en adición a β_0, solo necesitamos 1 variable dummy
- Usar 2 dummies introduce colinealidad perfecta
  - *hombre* es función lineal perfecta de *mujer*: $hombre + mujer = 1$

Trampa de variable dummy:

- Cuando muchas dummies describen un número dado de grupos

Grupo base o de referencia es contra el que se hace la comparación

- En ejemplo anterior, *hombre* es el grupo base
  - β_0 es el intercepto para hombres
  - δ_0 es la diferencia en interceptos entre mujeres y hombres

- Para escoger mujeres como grupo base:

$$salario = \alpha_0 + \gamma_0 hombre + \beta_1 educ + u$$

  - α_0: Intercepto para mujeres
  - $\alpha_0 + \gamma_0$: Intercepto para hombres

- No importa qué grupo base escojamos, pero tenemos que saber cuál es

A pesar de que el siguiente modelo es correcto:

$$salario = \gamma_0 hombre + \delta_0 mujer + \beta_1 educ + u$$

- Difícil probar diferencia en interceptos
- Diferentes formas de calcular R^2
- Entonces, incluir intercepto para grupo base

Podemos incluir más variables independientes

$$salario = \beta_0 + \delta_0 mujer + \beta_1 educ + \beta_2 exper + \beta_3 antig + u$$

- $H_0 : \delta_0 = 0$: No hay diferencia entre hombres y mujeres
- $H_a : \delta_0 < 0$: Hay discriminación contra las mujeres
- Todo sigue igual (estimación, estadístico t, etc.)
- Solo cambia la interpretación de δ_0

7.2.1 Aplicación: Análisis de políticas públicas

Podemos usar variables dummy para evaluar programas sociales

- 2 grupos:
  - Grupo de control: No participa en programa social
  - Grupo de tratamiento: Participa en programa social
- Nombres vienen de ciencias experimentales
  - Pero en ciencias sociales la asignación rara vez es aleatoria
- Usamos RLM para controlar y determinar efecto causal

Ejemplo. Empresas 1988

$$\widehat{hrscapac} = \underset{(43.41)}{46.67} \quad \underset{(5.59)}{+26.25\, apoyo} \quad \underset{(3.54)}{-0.98 \log(venta)} \quad \underset{(3.88)}{-6.07 \log(emplead)}$$

$$n = 105,\ R^2 = 0.237$$

- ¿t_{apoyo} es significativa?

- Controlando por ventas y empleados, empresas que recibieron apoyo capacitaron a cada trabajador en promedio 26.25 horas más
  - Apoyo tiene efecto grande en la capacitación ($\overline{x}_{apy} = 17$, $\max_{apy} = 164$)
- ¿El efecto es causal? ¿El apoyo aumenta las horas de capacitación?
  - El apoyo puede estar indicando otra cosa
    - Empresas que recibieron el apoyo de todos modos hubieran capacitado más a sus trabajadores aún sin el apoyo
  - Nada nos dice que estimamos un efecto causal
    - Necesitamos saber cómo se asignaron los apoyos
  - Esperamos haber controlado por suficientes factores relacionados con
    - Horas de capacitación
    - Si recibió apoyo

7.2.2 Interpretación de coeficientes de dummies cuado $\log(y)$

Los coeficientes tienen una interpretación de porcentaje al multiplicarlos por 100

Ejemplo.

$$\begin{aligned}\widehat{\log(precio)} = \underset{(0.65)}{-1.35} &+ \underset{(0.038)}{0.168\,\log(lote)} + \underset{(0.093)}{0.707\,\log(pies2)} \\ &+ \underset{(0.029)}{0.027\,habit} + \underset{(0.045)}{0.054\,colonial} \\ n = 88, &\qquad R^2 = 0.649\end{aligned}$$

- Una casa de estilo colonial se predice que se venderá por 5.4% más
  - Fijando los otros factores

Cuando el coeficiente de la dummy sugiere un cambio proporcional grande en y

- Aproximación es mala
- Necesitamos cambio exacto

Ejemplo.

$$\widehat{\log(salario)} = \underset{(0.099)}{0.417} - \underset{(0.036)}{0.297\,mujer} + \underset{(0.007)}{0.8\,educ} + \underset{(0.005)}{0.029\,exper}$$

$$\underset{(0.0001)}{-0.00058\,exper^2} + \underset{(0.007)}{0.032\,antg} - \underset{(0.00023)}{0.00059\,antg^2}$$

$$n = 526, \qquad \mathrm{R}^2 = 0.441$$

- Para mismos niveles de *educ*, *exper*, *antg*, las mujeres ganan
  - $100(0.297) = 29.7\%$ menos que los hombres, pero aproximación es mala
  - Queremos la diferencia porcentual exacta: $\frac{salario_M - salario_H}{salario_H}$
- Sabemos que :

$$-0.297 = \log(salario_M) - \log(salario_H) = \log\left(\frac{salario_M}{salario_H}\right)$$

$$\implies \frac{salario_M}{salario_H} = \exp(-0.297) \implies \frac{salario_M}{salario_H} - 1 = \exp(-0.297) - 1 \approx 0.257$$

  - Salario de mujeres es en promedio 25.7% más bajo que para hombres

Diferencia porcentual exacta en $\widehat{y}$ cuando la variable dependiente es $\log(y)$, x_1 es dummy, y $x_1 = 1$ vs $x_1 = 0$

$$\%\Delta\widehat{y} = 100\left[\exp(\widehat{\beta}_1) - 1\right]$$

- Si cambiamos grupo base:

$$\%\Delta\widehat{salario} = 100[\exp(0.297) - 1] \approx 34.6\%$$

- Aproximación $100\widehat{\beta}_1 = 29.7\%$ está entre 25.7 y 34.6
  - Diferencia en salarios predichos entre hombres y mujeres es alrededor de 29.7%

7.3 Uso de variables dummy para varias categorías

Podemos usar varias variables dummies independientes en la misma ecuación

- Ej. Mujer casada
- Pero meterlas por separado supone misma prima de matrimonio para H y M
- Podemos permitir diferencias salariales entre 4 grupos:
  - H casado, M casada, H soltero, M soltera
- Necesitamos escoger el grupo base: H soltero
- Definimos variables dummy para los otros 3 grupos y las incluimos en el modelo
  - Quitamos *mujer* porque es redundante

Ejemplo.

$$\widehat{\log(salario)} = \underset{(0.1)}{0.321} \underset{(0.055)}{+0.213\,hcas} \underset{(0.058)}{-0.198\,mcas} \underset{(0.056)}{-0.11\,msolt}$$

$$\underset{(0.007)}{+0.079\,educ} \underset{(0.005)}{+0.027\,xper} \underset{(0.00011)}{-0.00054\,xper^2} \underset{(0.007)}{+0.029\,ant} \underset{(0.00023)}{-0.00053 ant^2}$$

- Intercepto representa el grupo base
- Coeficientes son estadísticamente significativos
- Interpretación de los coeficientes de dummies depende del grupo base
  - Miden diferencia proporcional en el salario relativo a H soltero
  - H casado gana 21.3% más que H soltero ceteris paribus
  - M casada gana 19.8% menos que H soltero ceteris paribus
- Solo incluimos 3 de 4 grupos para evitar trampa de variable dummy
- Aunque grupo base es H soltero, podemos estimar diferencias entre grupos
  - Ignoramos el intercepto porque es el mismo para todos
  - Ej. Diferencia proporcional entre M soltera y M casada:
    - $-0.11 - (-0.198) = 0.088$, M soltera gana 8.8% más que M casada
  - Para probar si esa diferencia es estadísticamente significativa, escogemos uno de los 2 grupos como base y reestimamos

$$\widehat{\log(salario)} = \underset{(0.106)}{0.123} \underset{(0.056)}{+0.411\,hcas} \underset{(0.058)}{+0.198\,hsolt} \underset{(0.052)}{+0.088\,msolt + \ldots}$$

  - ¿Cuál es el grupo base?
  - ¿t_{msolt}?

Si el modelo tendrá intercepto para g categorías, incluimos $g - 1$ variables dummy

- De lo contrario, trampa de variables dummy
- El intercepto para el grupo base es el intercepto del modelo
- Coeficiente de dummy para una categoría representa diferencia estimada en interceptos entre esa categoría y el grupo base

7.4 Uso de variables dummy para incorporar información ordinal

Variable ordinal tiene categorías ordenadas pero distancias entre ellas no tienen significado

- Ej. Calificación crediticia: CC de 0 (peor) a 4 (mejor)

- ¿Cómo incorporarla al modelo para explicar tasa interés soberana (TIS)?

$$TIS = \beta_0 + \beta_1 CC + \text{ otros factores } + u$$

- ¿Diferencia entre 4 y 3 es la misma que entre 1 y 0?
  - Modelo restringido
  - No permite capturar efecto no constante

Dado que CC toma pocos valores, es mejor definir variables dummy para cada valor

- ¿Cuántas categorías tiene CC?
- ¿Cuántas variables dummy incluimos en el modelo?

$$TIS = \beta_0 + \delta_1 CC_1 + \delta_2 CC_2 + \delta_3 CC_3 + \delta_4 CC_4 + \text{ otros factores } + u$$

- Modelo no restringido
- ¿Cuál es el grupo base?
- $CC_1 = 1$ si $CC = 1$ y $CC_1 = 0$ de otra forma, $CC_2 = 1$ si $CC = 2$ ...
- Interpretación:
  - δ_1: diferencia en TIS (ceteris paribus) entre países con $CC = 1$ y $CC = 0$
  - ...
- Efecto de cada categoría puede ser diferente
  - Usar varias dummies es más flexible
- Primer modelo es caso especial del segundo:

$$TIS = \beta_0 + \delta_1(CC_1 + 2CC_2 + 3CC_3 + 4CC_4) + \text{ otros factores } + u$$

- Podemos probar restricción de efecto parcial constante
  - Estadístico F con $q = 3$, R^2_{sr}, R^2_r

Cuando variable ordinal toma muchos valores, no incluimos una dummy para cada valor

- En ese caso, partimos la variable en categorías
- Ej. Ranking de escuelas: $top10$, $r11\_25$, $r26\_40$, $r41\_60$, $r61\_100$
  - Definimos variables dummy igual a 1 cuando el ranking caiga en cada categoría
  - Grupo base: Escuelas con un ranking debajo de 100
- Podemos calcular el cambio porcentual exacto con $\exp(\widehat{\beta}_j) - 1$
- Para evaluar si partir variable en categorías ayuda comparar $\overline{\mathrm{R}}^2$'s

7.5 Interacciones con variables dummy

7.5.1 Interacciones entre variables

En ejemplo de género y estado civil podemos agregar un término de interacción

- Interacción $mujer \cdot casad$ permite que prima de matrimonio dependa del género

Ejemplo.

$$\widehat{\log(salario)} = \underset{(0.100)}{0.321} \quad \underset{(0.056)}{-0.11\, mujer} \quad \underset{(0.055)}{+0.213\, casad} \quad \underset{(0.072)}{-0.301\, mujer \cdot casad} + \ldots$$

- ¿Cuál es el grupo base?
- ¿Interacción es estadísticamente significativa?
- Podemos estimar la diferencia del salario entre los 4 grupos
  - Cuidado con selección de 0's y 1's
  - H soltero: $mujer = 0$, $casad = 0$, intercepto: 0.321
  - H casado: $mujer = 0$, $casad = 1$, intercepto: 0.321 + 0.213 = 0.534
  - M soltera: $mujer = 1$, $casad = 0$, intercepto: 0.321 - 0.11 = 0.211
  - M casada: $mujer = 1$, $casad = 1$, intercepto: 0.321 - 0.11 + 0.213 - 0.301 = 0.123

Diferente forma de obtener diferencias de salario entre combinaciones de género y estado civil

- Forma con interacciones entre dummies permite probar
  - H_0: Diferencial de género no depende del estado civil
  - H_0: Diferencial de estado civil no depende del género
- Forma $g - 1$ categorías es mejor para probar diferencias salariales entre cualquier grupo y el grupo base

7.5.2 Modelar diferentes pendientes

Para tener diferentes interceptos para cualquier número de grupos

- Incluir diferentes variables dummy
- Ej. Diferencia salarial entre H y M

Para diferentes pendientes:

- Interactuar variables dummy con variables independientes que no son dummies
- Ej. Diferente rendimiento de educación entre H y M

El siguiente modelo permite capturar diferentes interceptos y rendimiento de educación

$$\log(salario) = (\beta_0 + \delta_0\, mujer) + (\beta_1 + \delta_1\, mujer)\, educ + u$$

- Para H: $mujer = 0$
  - Intercepto: β_0
  - Pendiente de $educ$: β_1
- Para M: $mujer = 1$
  - Intercepto: $\beta_0 + \delta_0$
  - Pendiente de $educ$: $\beta_1 + \delta_1$
- Interpretación:
  - δ_0 mide diferencia en interceptos entre H y M
  - δ_1 mide diferencia en rendimiento de educación entre H y M
- Caso: $\delta_0 < 0$ y $\delta_1 < 0$
  - M gana menos que H con poca educación y brecha se abre con más educación
- Caso: $\delta_0 < 0$ y $\delta_1 > 0$
  - M gana menos que H con poca educación y brecha se cierra con más educación

[Gráficas]

Para estimar, primero generamos variable interacción $mujer \cdot educ$ y luego estimamos

$$\log(salario) = \beta_0 + \delta_0\, mujer + \beta_1\, educ + \delta_1\, mujer \cdot educ + u$$

- $H_0 : \delta_1 = 0$
  - Mismo rendimiento de educación entre H y M
  - Misma pendiente de $\log(salario)$ con respecto a $educ$
- Si $\delta_0 \neq 0$, diferentes interceptos
  - Hay diferencia salarial bajo H_0 pero es la misma para diferente educación
- $H_0 : \delta_0 = 0,\ \delta_1 = 0$ (usamos prueba F)
  - Salarios promedio son iguales para H y M con misma educación

Ejemplo.

$$\begin{array}{lcccc} \widehat{\log(salar)} = & 0.389 & -0.227\, mujer & +0.082\, educ & -0.0056\, mujer \cdot educ \\ & (0.119) & (0.168) & (0.008) & (0.0131) \\ & +0.029\, xper & -0.00058\, xper^2 & +0.032\, ant & -0.00059\, ant^2 \\ & (0.005) & (0.00011) & (0.007) & (0.00024) \end{array}$$

$$n = 526 \qquad \mathrm{R}^2 = 0.441$$

- ¿t_{mujer}? ¿$\mathrm{t}_{mujer \cdot educ}$?
- Rendimiento de educación
  - H: 8.2%
  - M: 0.082 - 0.0056 = 0.0764 = 7.6%
  - Diferencia (-0.56%) no es significativa
  - No hay evidencia en contra de H_0 (mismo rendimiento de educación)
- t_{mujer} no indica si hay menor pago para mujeres ceteris paribus
  - Coeficiente de $mujer$ es menos preciso
    - MC entre $mujer$ y $mujer \cdot educ$
  - δ_0 mide diferencia salarial entre H y M cuando $educ = 0$
    - Pocas observaciones en la muestra con $educ = 0$
  - Para estimar diferencia salarial a $\overline{educ} = 12.5$
    - Cambiar $mujer \cdot educ$ por $mujer(educ - 12.5)$
- Para $\mathrm{H}_0 : \delta_0 = 0,\ \delta_1 = 0$, $\mathrm{F} = 34.33$ con $q = 2$ y $n-k-1 = 518$, valor $-$ p ≈ 0

7.5.3 Probar si función de regresión varía por grupos

Queremos probar si mismo modelo de regresión describe calificaciones de atletas H y M

$$GPA = \beta_0 + \beta_1 exam + \beta_2 prepaprct + \beta_3 hrsclase + u$$

- Para diferencia en interceptos:
  - Dummy de género (H o M)
- Para diferencia en cualquier pendiente con base en género:
  - Interactuar variable independiente con H o M
- Aquí: Queremos probar si hay <u>cualquier</u> diferencia entre H y M
  - Intercepto y todas las pendientes pueden ser diferentes para H y M

El siguiente modelo permite probar diferencias en intercepto y todas las pendientes

$$\begin{aligned} GPA = \beta_0 + \delta_0 mujer + \beta_1 exam + \delta_1 mujer \cdot exam + \beta_2 prepa \\ + \delta_2 mujer \cdot prepa + \beta_3 hrsclase + \delta_3 mujer \cdot hrsclase + u \end{aligned}$$

- $\mathrm{H}_0 : \delta_0 = 0,\ \delta_1 = 0,\ \delta_2 = 0,\ \delta_3 = 0$
- GPA sigue mismo modelo para H y M
- Si cualquier $\delta_j \neq 0$, el modelo es diferente

Ejemplo.
Modelo no restringido:

$$\widehat{GPA} = \underset{(0.21)}{1.48} \quad \underset{(0.411)}{-0.353\,mujer} \quad \underset{(0.002)}{+0.011\,exam} \quad \underset{(0.00039)}{+0.00075\,mujer \cdot exam}$$

$$\underset{(0.0014)}{-0.0085\,prepa} \quad \underset{(0.00316)}{-0.00055\,mujer \cdot prepa} \quad \underset{(0.0009)}{+0.0023\,hrscls} \quad \underset{(0.00163)}{-0.00012\,mujer \cdot hrscls}$$

$$n = 366 \qquad \mathrm{R}^2 = 0.4406 \qquad \overline{\mathrm{R}}^2 = 0.394$$

- Para estadístico F, modelo restringido sin $mujer$
  - $\mathrm{R}^2 = 0.352$, $\mathrm{F} \approx 8.44$, valor $-$ p ≈ 0, rechazamos H_0
  - H y M siguen diferente modelo aunque individualmente no significativos
- Para interpretar diferencias, tomar en cuenta interacciones
  - M menor que H en 0.353 solo cuando $exam = 0$, $prepa = 0$, $hrsclase = 0$
  - $-353 + 0.00075(1,100) - 0.00055(10) - 0.00012(50) \approx 0.461$
  - M más que H a esos niveles de las variables independientes

Para cualquier número de variables independientes, usamos forma SCR del estadístico F
En el modelo general, tenemos k variables independientes, 1 intercepto y 2 grupos

$$y = \beta_{g,0} + \beta_{g,1}k_1 + \beta_{g,2}k_2 + \ldots + \beta_{g,k}k_k + u$$

- Para $g = 1$, $g = 2$
- H_0 : Intercepto y todas las pendientes son iguales
- En modelo restringido, misma β para los 2 grupos
  - Restricciones: $k + 1$ (dummy de grupo más k interacciones)
    - Ej. En GPA, $k + 1 = 4$
  - Grados de libertad: $(n - k - 1) - (k + 1) = n - 2(k + 1)$
    - Ej. En GPA, $n - 2(4) = n - 8 = 366 - 8 = 358$
- Idea clave: SCR del modelo no restringido se puede obtener de 2 regresiones
  - $\mathrm{SCR}_{sr} = \mathrm{SCR}_1 + \mathrm{SCR}_2$
    - SCR_1 de estimar modelo para grupo $g = 1$ con n_1 observaciones
    - SCR_2 de estimar modelo para grupo $g = 2$ con n_2 observaciones
    - Ej. $n_1 = 90$ mujeres, $n_2 = 276$ hombres, $n = n_1 + n_2$
  - SCR_r de estimar modelo como 1 solo grupo

Estadístico F conocido como **estadístico de Chow**

$$\mathrm{F} = \frac{(\mathrm{SCR}_r - (\mathrm{SCR}_1 + \mathrm{SCR}_2))}{\mathrm{SCR}_1 + \mathrm{SCR}_2} \cdot \frac{[n - 2(k + 1)]}{k + 1}$$

- Solo valido bajo homocedasticidad ($\sigma_1^2 = \sigma_2^2$)
- Forma R^2 solo se puede usar si se incluyen interacciones para modelo sin restricciones

> Ejemplo. GPA
> $SCR_r = 85.515$, $SCR_1 = 19.603$, $n_1 = 90$, $SCR_2 = 58.752$, $n_2 = 276$, entonces
>
> $$SCR_{sr} = 19.603 + 58.752 = 78.355$$
>
> $$F = \frac{(85.515 - 78.355)}{78.355} \cdot \frac{358}{4} \approx 8.18$$

Limitación de la prueba de Chow: H_0 no permite cambios entre grupos

- Común permitir diferencia en interceptos y probar para diferencias en pendientes
- 2 formas de probar una diferencia en interceptos
  - Incluir dummy para el grupo y sus interacciones, y probar significancia conjunta de las interacciones
  - Obtener una forma SCR del estadístico F en la que SCR_r se obtiene de una regresión que solo tiene un cambio de intercepto

    $$F = \frac{(SCR_r - (SCR_1 + SCR_2))}{SCR_1 + SCR_2} \cdot \frac{[n - 2(k+1)]}{k}$$

    * Probamos k restricciones en lugar de $k + 1$
- Si no se rechaza H_0, el mejor modelo permite una diferencia en interceptos

Otros usos de dummies: En series de tiempo

- Cambio estructural
- Estacionalidad
- Estimación por tramos

7.6 Variable dependiente binaria: Modelo de probabilidad lineal

Hasta ahora: Variable dependiente y ha tenido significado cuantitativo
También podemos usar RLM para explicar un evento cualitativo

- Ej. Si adulto tiene prepa, si universitario consume drogas, si empresa es adquirida
- y puede cambiar de 0 a 1, de 1 a 0 o no cambiar

Interpretación de β_j's cambia

- Ya no es Δy ante $\Delta x = 1$ ceteris paribus

Si suponemos RLM.4 ($\mathbb{E}(u \mid x_1, x_2, \ldots, x_k) = 0$), se cumple que

$$\mathbb{E}(y \mid \vec{x}) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k$$

Como y es binaria, siempre se cumple que

- $\mathbb{P}(y = 1 \mid \vec{x}) = \mathbb{E}(y \mid \vec{x})$
  - Probabilidad de éxito es igual al valor esperado de y
- $\mathbb{P}(y = 1 \mid \vec{x}) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k$
  - $\mathbb{P}(y = 1 \mid \vec{x})$ se llama la probabilidad de respuesta
  - $p(\vec{x}) = \mathbb{P}(y = 1 \mid \vec{x})$: Probabilidad de éxito es una función lineal de las x_j's
  - Ejemplo de un modelo de respuesta binaria (otros logit, probit)
- $\mathbb{P}(y = 0 \mid \vec{x}) = 1 - \mathbb{P}(y = 1 \mid \vec{x})$ porque probabilidades suman 1
  - También es función lineal de las x_j's

Modelo de probabilidad lineal (MPL):

- RLM con y binaria
- Probabilidad de respuesta es lineal en parámetros

En MPL, β_j mide el cambio en la probabilidad de éxito cuando cambia x_j ceteris paribus

$$\Delta\mathbb{P}(y = 1 \mid \vec{x}) = \beta_j \Delta x_j$$

La mecánica de MCO es la misma que antes

- Ecuación estimada: $\widehat{y} = \widehat{\beta}_0 + \widehat{\beta}_1 x_1 + \widehat{\beta}_2 x_2 + \ldots + \widehat{\beta}_k x_k$
- $\widehat{y}$: Probabilidad de éxito predicha
- $\widehat{\beta}_0$: Probabilidad de éxito predicha cuando cada $x_j = 0$
- $\widehat{\beta}_j$: Cambio en probabilidad de éxito predicha cuando $\Delta x_j = 1$

MPL's son fáciles de estimar e interpretar

Ejemplo. Participación en fuerza laboral (fl) de mujeres en el año 1975

$$\begin{array}{llll} \widehat{fl} = & 0.586 & -0.0034\, ingresp & +0.038\, educ \quad +0.039\, exper \\ & (0.154) & (0.0014) & (0.007) \quad (0.006) \end{array}$$

$$\begin{array}{llll} -0.0006\, exper^2 & -0.016\, edad & -0.262\, menor6 & +0.013\, mayor6 \\ (0.00018) & (0.002) & (0.034) & (0.013) \end{array}$$

$$n = 753 (428\ 1\text{'s}) \qquad \overline{\mathrm{R}}^2 = 0.264$$

- Coeficientes con signos esperados y t's estadísticamente significativas
- Interpretación: Efecto en probabilidad de estar en la fuerza laboral

  - Si $\Delta educ = 10$ (años), probabilidad sube 0.38 (efecto grande)
  - Si $\Delta ingresp = 10$ (miles), probabilidad baja 0.034 (efecto pequeño)
  - Si $\Delta menor6 = 1$, probabilidad baja 0.262 (efecto grande)
- Si fijamos $ingresp = 50$, $exper = 5$, $edad = 30$, $menor6 = 1$, $mayor6 = 0$
  [Gráfica]
  - Otros valores cambian rango pero no la pendiente
- Probabilidad negativa si $educ < 3.84$, pero en datos $\min(educ) = 5$
- Probabilidad predicha es 0.5 cuando $\max(educ) = 17$
- Efecto decreciente de $exper$ en probabilidad por $exper^2$ (ceteris paribus)

$$\Delta\widehat{fl} = 0.039 - 2(0.0006)exper = 0.039 - 0.0012exper$$

  - No efecto en probabilidad cuando $exper = 0.039/0.0012 = 32.5$
    * Umbral alto, solo 13 de 753 tienen $exper > 32$

MPL es útil sobre todo cerca de promedios de variables independientes

- Restringir atención a esos valores

Deficiencias del MPL

- Predicciones pueden ser menores a 0 o mayores a 1
  - Para ciertos valores de variables independientes
- Efecto marginal constante
  - Probabilidad no puede estar relacionada linealmente a todos los valores
  - Ej. $\Delta\widehat{fl} = 0.262(\Delta menor6) = 0.262 * 4 = 1.048$
- Tiene heteroscedasticidad (ver abajo)

Forma de usar probabilidades predichas para predecir variable binaria

- $\widehat{y}$ denota valores ajustados (puede que sean menores a 0 o mayores a 1)
- Define valor predicho como:

$$\widetilde{y}_i = 1 \text{ si } \widehat{y}_i \geq 0.5 \text{ y } \widetilde{y}_i = 0 \text{ si } \widehat{y}_i < 0.5$$

  - De esta forma, $\widetilde{y}_i \in \{0, 1\}\, \forall i$ igual que y_i
  - Porcentaje correctamente predicho: Medida de bondad de ajuste para variables dependientes binarias (comparar $\widetilde{y}_i$ vs y_i)

MPL viola un supuesto G-M cuando y es binaria: Hay heterocedasticidad en MPL

$$\text{Var}\left(y \mid \vec{x}\right) = \mathbb{E}\left(y^2 \mid \vec{x}\right) - \mathbb{E}\left(y \mid \vec{x}\right)^2 = p\left(\vec{x}\right) - p\left(\vec{x}\right)^2 = p\left(\vec{x}\right)\left[1 - p\left(\vec{x}\right)\right]$$

- $p\left(\vec{x}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k$, depende de $\vec{x}$

- No hay sesgo pero pruebas t y F no son válidas
  - Se deben corregir los errores estándar, pero en la práctica son similares

Podemos incluir dummies como variables independientes en MPL

- Coeficientes miden diferencia predicha en probabilidad en relación con grupo base
- Ej. $\widehat{arrest86} = 0.350 +$ controles $+ 0.17 rnegra + 0.096 hispano$
  - Probabilidad de arresto es 17% más alta para raza negra que raza blanca

7.7 Consideraciones adicionales para análisis de políticas y evaluación de programas

Tener cuidado al evaluar programas

- Generalmente, grupos de control y tratamiento no asignados aleatoriamente

Incluir factores que puedan estar sistemáticamente relacionados con variable independiente binaria de interés

- Ej. Discriminación en autorizaciones de créditos por raza
- Aprobación depende de otros factores: ingreso, riqueza, calificación crediticia
- Necesitamos controlar por ellos si hay diferencias sistemáticas en ellas entre razas

  MPL: $aprob = \beta_0 + \beta_1 noblanco + \beta_2 ingreso + \beta_3 riqueza + \beta_4 calcred +$ otros factores $+ u$
- Hay discriminación contra minorías si se rechaza $H_0 : \beta_1 = 0$ en favor de $H_a : \beta_1 < 0$

Problemas de auto-selección pueden hacer que RLM esté sesgado por falta de controles suficientes

- Individuos se auto seleccionan en comportamiento o programas
  - Participación no aleatoria
- Ej. Individuos seleccionados para participar en programa, pueden participar o no
- Se usa cuando indicador binario de participación puede relacionarse sistemáticamente con factores no observados
  - Ej. $y = \beta_0 + \beta_1 particp + u$
  - Nos preocupa que $\mathbb{E}(u \mid particp = 1) \neq \mathbb{E}(u \mid particp = 0)$, u depende de participación
  - Variable independiente es endógena, β_1 estará sesgada
    - No encontraremos verdadero efecto de participación
    - Podemos encontrar efectos espurios
- Si se observan factores correlacionados con *particp*, RLM mitiga el problema
- Si no se observan factores correlacionados con *particp*, RLM genera sesgo por falta de controles
  - Datos panel, variables instrumentales, logit, probit

7.8 Interpretación de resultados de regresión con variables dependientes discretas

Cuando y es una variable discreta categórica no binaria (pocos valores enteros), ¿cómo interpretamos los coeficientes?

Ejemplo. Mujeres de Botswana

$$\begin{array}{rccc} \widehat{hijos} = & -1.997 & +0.175\,edad & -0.09\,educ \\ & (0.094) & (0.003) & (0.006) \end{array}$$

$$n = 4,361,\ \mathrm{R}^2 = 0.56$$

- MCO estima efectos de x_j sobre valor esperado/promedio de y
  - Bajo RLM.1 a RLM.4, $\mathbb{E}(y \mid \vec{x}) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k$
    - β_j efecto ceteris paribus de $\Delta x_j = 1$ sobre el valor esperado de y
  - Para una muestra dada, $\mathbb{E}(y \mid \vec{x}) = \widehat{\beta}_0 + \widehat{\beta}_1 x_1 + \widehat{\beta}_2 x_2 + \ldots + \widehat{\beta}_k x_k$
    - $\widehat{\beta}_j$ estimado de cómo cambia el promedio de y cuando $\Delta x_j = 1$
- $\widehat{\beta}_{educ} = -0.09$
  - Estimamos que la fertilidad promedio cae 0.09 hijos dado 1 año más de educación
  - Si cada mujer de un grupo de 100 tiene 1 año más de educación, habrá 9 niños menos entre ellas

Agregar variables independientes dummy cuando y es discreta, no afecta interpretación

Ejemplo. Mujeres de Botswana

$$\begin{array}{rcccc} \widehat{hijos} = & -2.071 & +0.177\,edad & -0.079\,educ & -0.362\,electric \\ & (0.095) & (0.003) & (0.006) & (0.068) \end{array}$$

$$n = 4,358,\ \mathrm{R}^2 = 0.562$$

- Comparando 100 mujeres con electricidad vs 100 mujeres sin ella (y con misma educación y edad), estimamos que el primer grupo tenga 36 niños menos

RLM da buena aproximación a efectos parciales sobre $\mathbb{E}(y \mid \vec{x})$

- Pero hay mejores modelos para variables dependientes discretas
  - Logit, probit, tobit, Poisson