- Cuando una var. ordinal toma muchos valores, no incluimos una dummy p/c/valor
  ▷ En ese caso: partimos la var. en categorías

- Vgr. ranking de la escuela: top10, r11_25, r26_40, r41_60, r61_100
  definir vars. dummy igual a 1 cuando ranking caiga en rango
  ▷ gpo. base: escuelas con un ranking debajo de 100
  ▷ podemos calcular el cambio porcentual exacto con $\exp(\hat{\beta}_1) - 1$
  ▷ p/ evaluar si partir var. en categs ayuda: comparar $\bar{R}^2$'s

Interacciones con Vars. Dummy

- Interacciones entre Vars.
- En ejemplo de género y edo. civil podemos agregar una término de interacción ÷ mujer – casada p/ q' prima de matrimonio dependa del género

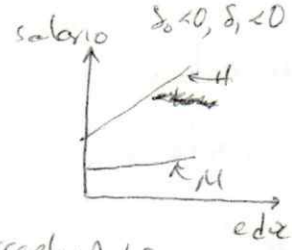
$$\widehat{\log(salario)} = \underset{(0.100)}{0.321} - \underset{(0.056)}{0.11}\, mujer + \underset{(0.055)}{0.213}\, casada - \underset{(0.072)}{0.301}\, mujer * casada + \ldots$$ ← idéntico

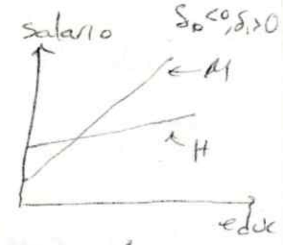
* interacción estadísticamente significativa

- Podemos estimar la dif. de salario ÷ los 4 gpos. ← cuidado con selección de 0's y 1's
  ▷ h soltero (gpo. base): mujer = 0, casada = 0 → intercepto: 0.321
  ▷ h casado: mujer = 0, casada = 1 → intercepto: 0.321 + 0.213 = 0.534
  ▷ m soltera: mujer = 1, casada = 0 → intercepto: 0.321 – 0.11 = 0.211
  ▷ m casada: mujer = 1, casada = 1 → intercepto: 0.321 – 0.11 + 0.213 – 0.301 = 0.123
- Dif. forma de obtener diferencias de salario ÷ combinaciones de género y edo. civil
  ▷ forma con interacc. ÷ dummies permite probar Ho: diferencial de género no depende de edo. civil; diferencial de edo. civil no depende de género
  ▷ forma g-1 categ. mejor p/ probar dif. salariales ÷ cualquier gpo. y gpo. base (h soltero)

- Modelar dif. Pendientes (herramienta poderosa)
- Incluir dif. vars. dummy permite tener dif. interceptos p/ cualquier # de gpos.
- P/ dif. pendientes: interactuar vars. dummy con vars. indep. q' no son dummies
- Vgr. dif. salarial cte. ÷ H–M pero ahora queremos probar si rendim. de educ es igual p/ H–M

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





▷ p/ H: mujer = 0 < intercepto: β_0 / pendiente educ: β_1

p/ M: mujer = 1 < intercepto: $\beta_0 + \delta_0$ / pendiente educ: $\beta_1 + \delta_1$

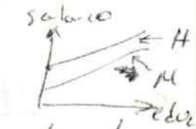
M ganan menos q' H con baja educ y brecha se abre educ↑

M ganan menos q' H con poca educ, pero brecha se cierra educ↑

▷ interpretación < δ_0: mide dif. en interceptos ÷ H/M ; δ_1: mide dif. en rendim. educ. ÷ H/M

- p/ estimar define interacción mujer * educ y estima

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



▷ $H_0: \delta_1 = 0$ → mismo rendim. a educ ÷ H/M, misma pendiente de log(salario) wrt educ

▷ si $\delta_0 \neq 0$ → dif. interceptos, i.e. hay dif. salarial bajo H_0 pero es la misma p/ dif. educ.

▷ $H_0: \delta_0 = 0,\ \delta_1 = 0$ (usamos prueba F) → salarios promedio son iguales p/ H/M con misma educ.

- Vgr

$$\underset{(0.119)}{\log(salario) = 0.389} - \underset{(0.168)}{0.227}\, mujer + \underset{(0.008)}{0.082}\, educ - \underset{(0.0131)}{0.0056}\, mujer \cdot educ + \underset{(0.005)}{0.029}\, exper - \underset{(0.00011)}{0.00058}\, exper^2 + \underset{(0.007)}{0.032}\, antig - \underset{(0.00024)}{0.00059}\, antig^2$$

$n = 526,\ R^2 = 0.441$

$t = -1.35$; $t = -0.43$

▷ rendim. educ < H: 8.2% ; M: $0.082 - 0.0056 = 0.0764 = 7.6\%$; dif. −0.56% no signif. → no evidencia en contra de H_0 (mismo rendim. educ.)

† mujer no indica si hay menor pago p/ mujeres ceteris paribus

coef. mujer menos preciso x multicol ÷ mujer / mujer.educ

δ_0 mide dif. salarial ÷ H/M cuando educ = 0 (pocas obs. en muestra)

p/ estimar dif. salarial a $\overline{educ} = 12.5$: cambiar mujer.educ por mujer(educ − 12.5) → solo cambia $\hat{\beta}_{mujer}$ en est($\hat{\beta}_0$)

▷ p/ $H_0: \delta_0 = 0,\ \delta_1 = 0$, estad. $F = 34.33$ con g.l. num = 2 y g.l. den. = 518, valor-p ≈ 0

- Probar si funciones de regr. varían por grupos.

- Queremos probar si mismo mod. de regr. describe las calif. de atletas univ. H/M

$GPA = \beta_0 + \beta_1\, examen\ admisión + \beta_2\, prepa\ percentil + \beta_3\, tothrs\ clase + u$

▷ p/ dif. en interceptos: dummy p/ < H o M

▷ p/ dif en cualquier pendiente con base en género: interactuar var. indep. con < H o M

▷ aquí: queremos probar si hay cualquier dif. ÷ H/M ⇒ intercepto y todas pendientes } pueden ser dif. p/ H/M

- $GPA = \beta_0 + \delta_0 mujer + \beta_1 examen + \delta_1 mujer \cdot examen + \beta_2 prepapct + \delta_2 mujerprepapct + \beta_3 totlrs + \delta_3 mujer \cdot totlrs$ to

admisión

dif. en intercepto — dif. en pend. examen

- $H_0: \delta_0 = 0, \delta_1 = 0, \delta_2 = 0, \delta_3 = 0$ ← GPA sigue mismo mod. p/ $\frac{H}{M}$ (con cualquier $\delta_j \neq 0$ mod. es dif.)

- Vgr. (mod. no restringido)

$GPA = 1.48 - 0.353\, mujer + 0.0011\, exam + 0.00075\, \frac{mujer}{exam} - 0.0085\, prepact - 0.00055\, \frac{mujer}{prepapct} + 0.0023\, totlrs - 0.00012\, \frac{mujer}{totlrs}$, $n = 366$, $R^2 = 0.406$, $\bar{R}^2 = 0.394$

(0.21) (0.411) (0.002) (0.00039) (0.0014) (0.00316) (0.0009) (0.00163)

$t \approx 2$

▷ p/ estad. F _ mod. restringido (sin mujer): $R_r^2 = 0.352$, $F \approx 8.14$, valor-$p \approx 0 \Rightarrow$ rechazamos H_0 (H y M siguen dif. mod.) aunq. indiv. no signif.

▷ p/ interpretar difs. tomar en cuenta interacciones

M menor q' H en 0.353 solo cuando examen = 0, prepapct = 0, totlrs = 0

$-.353 + 0.00075(1,100) - 0.00055(10) - 0.00012(50) \approx 0.461$ M más q' H a esos niveles de vars. indep.

sat — prepa — totlrs

* P/ cualquier # de vars. indep. usamos la forma SCR del estad. F

- En mod. gral.: k vars. indep, 1 intercepto, 2 gpos. $g \leq \frac{1}{2}$, H_0: intercepto, todas pendientes son iguales

$y = \beta_{g,0} + \beta_{g,1} x_1 + \beta_{g,2} x_2 + \ldots + \beta_{g,k} x_k + u$ p/ $g = 1, g = 2$

- H_0: misma β p/ los 2 gpos involucra $k+1$ restricciones (vgr GPA $k+1 = 4$)

- mod. restringido (dummy de gpo. + k interaccs.): $(n-k-1)-(k+1) = n - 2(k+1)$ g.l. (vgr. GPA $366 - 2(4) = 358$)

- Idea clave: SCR del mod. no restringido se puede obtener de 2 regs. (p/c/gpo.)

▷ SCR_1 de estimar mod. p/ gpo. $g=1$ con n_1 obs. ($n_1 = 90$ mujeres)

▷ SCR_2 " " " $g=2$ con n_2 obs ($n_2 = 276$ hombres) $\Big\}\, n = n_1 + n_2$

▷ $SCR_{sr} = SCR_1 + SCR_2$

▷ SCR_r de estimar mod. como 1 solo gpo.

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

estadístico F conocido como " de Chow

- Solo válido bajo homos. ($\sigma_1^2 = \sigma_2^2$)

- Vgr. GPA

$SCR_r = 85.515$, $SCR_1 = 19.603$ $n_1 = 90$, $SCR_2 = 58.752$ $n_2 = 276 \Rightarrow SCR_{sr} = 19.603 + 58.752 = 78.355$

$$F = \frac{85.515 - 78.355}{78.355} \cdot \frac{358}{4} \approx 8.18$$

- forma R^2 solo se puede usar si se incluyen interacciones p/ mod. s/restr.

- Limitación de la prueba de Chow: Ho no permite cambios ÷ gpos.
  ▷ también permitir dif. en interceptos y probar por difs. en pendientes
  ▷ 2 formas
    - incluir dummy p/el gpo. y sus interacciones, y probar signif. conj. de las interaccs.
    - obtener una forma SCR del estadístico F en la q SCR$_r$ se obtiene de una regr. q 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 restrices. no k+1

- Si no se rechaza Ho, el mejor modelo permite una dif. en interceptos
- Otros usos de dummies: en series de tiempo cambio estructural, estacionalidad, estimación por tramos

Mod. Lineal de Probabilidad (Var. Dep. Binaria)

- Hasta ahora var. dep. y ha tenido significado cuantitativo
- Podemos usar RLM p/ explicar un evento cualitativo
  - si, ¿ébito tiene prepa universitaria / consumió drogas / empresa adquirida por otra
  ▷ y puede cambiar de 0 a 1, de 1 a 0 o no cambiar
- Interpretación de β_j's cambia (ya no Δy ante $\Delta x = 1$ ceteris paribus)
- Si suponemos RLM.4 se cumple $(E(u|x_1, \ldots, x_k) = 0)$: $E(y|x) = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k$
- Como y es binaria, siempre se cumple $P(y=1|x) = E(y|x)$ ← prob. de éxito = valor esperado de y
  ∴ $P(y=1|x) = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k$ ← prob. de éxito $p(x) = P(y=1|x)$ es una func. lineal de las x_j
  - ejemplo de mod. de respuesta binaria (otros logit, probit)
  - $P(y=1|x)$ se llama la probabilidad de respuesta
- Como probabilidades suman 1, $P(y=0|x) = 1 - P(y=1|x)$ también es func. lineal de las x_j
- Mod. lineal de probab. = RLM con y binaria xq prob. de resp. es lineal en paráms. β_j
- En MPL, β_j mide el cambio en la prob. de éxito cuando cambia x_j ceteris paribus
  $\Delta P(y=1|x) = \beta_j \Delta x_j$
- La mecánica de MCO es la misma q antes
- Ec. estimada: $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \ldots + \hat{\beta}_k x_k$
  - $\hat{y}$: prob. de éxito predicha
  - $\hat{\beta}_0$: " " " cuando c/ $x_j = 0$
  - $\hat{\beta}_1$: cambio en prob. de éxito predicha cuando x_1 ↑ en 1

- Vgr. participación en fuerza laboral de mujeres en 1975 (FL)

$n = 753$ (428 1's)
$R^2 = 0.264$

$$\underset{(0.154)}{\widehat{enflab}} = 0.586 - \underset{(0.0014)}{0.0034}\ \underset{\text{miles USD}}{ingresposo} + \underset{(0.007)}{0.038}\ educ + \underset{(0.006)}{0.039}\ exper - \underset{(0.00018)}{0.0006}\ exper^2 - \underset{(0.002)}{0.016}\ edad - \underset{(0.034)}{0.262}\ \underset{\text{niños}}{menor6} + \underset{(0.013)}{0.013}\ mayor6$$

▷ t's estadíst. signif. y signos esperados

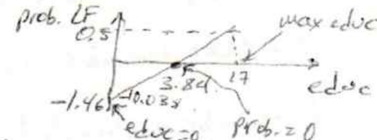
▷ interpretación: educ ↑ 10 y → probab. de estar en FL ↑ $0.038(10) = 0.38$ ← incremento gde.

▷ si fijamos ingresposo = 50, exper = 5, edad = 30, menor6 = 1, mayor6 = 0

▷ prob. < 0 si educ < 3.84 pero en datos min(educ) = 5

▷ max(educ) = 17 → probabilidad predicha de 0.5

otros valores cambian rango pero no pendiente



▷ interpretación: Δ ingresposo = 10 (10k) → probab. en FL ↓ 0.034 ← efecto pequeño

▷ exper en cuadrado p/permitir q' tenga efecto decreciente en prob.

$\Delta \widehat{enflab} = 0.039 - 2(0.0006)exper = 0.039 - 0.0012\ exper$ ceteris paribus

⇒ no efecto en prob. cuando $exper = \frac{0.039}{0.0012} = 32.5$ ← alto, solo 13 de 753 tienen exper > 32

▷ niño menor de 6 ↑ reduce prob. de particip. en 0.262

- MPL's son fáciles de estimar e interpretar
- Deficiencias del MPL
  - ▷ p/ciertos valores de vars. indep., predicciones < 0 o > 1 cuando prob.'s $\in [0, 1]$
  - ▷ efecto marginal constante: una prob. no puede estar relacionada linealmente a todos valores de vars. indep. $\Delta \widehat{enflab} = 0.262(\Delta \overset{=4}{menor6}) = 1.048$ no en muestra
  - ▷ tiene heteroscedasticidad
- MPL útil ~~sobre~~ todo cerca de promedios de vars. indep., restringir atención a esos valores
- Forma de usar prob.'s predichas aún si < 0 o > 1 p/ predecir var. binaria
  - ▷ $\hat{y}$: ~~mdo~~ denota valores ajustados (puede q' haya < 0 y > 1)
  - ▷ define valor predicho como: $\tilde{y}_i = 1$ si $\hat{y}_i \geq 0.5$ y $\tilde{y}_i = 0$ si $\hat{y} < 0.5$ → $\tilde{y}_i \in \{0, 1\}\ \forall i$ como y_i
  - ▷ porcentaje correctamente predicho: medida de bondad de ajuste p/ vars. dep. binarias (comparar $\tilde{y}_i$ vs y_i)
- MPL violan supuesto G-M xq' y es binaria y su varianza

$Var(y|x) = E(y^2|x) - E(y|x)^2 = E(y|x) - E(y|x)^2 = p(x) - p(x)^2 = p(x)[1 - p(x)]$

▷ donde $p(x) = \beta_0 + \beta_1 x_1 + \dots + \beta_k x_k$ ⇒ var. depende de x ⇒ hay heterosc en MPL

▷ no hay sesgo pero pruebas t y F no son válidas → corregir errest's pero en práctica similar

- Podemos incluir ~~dum~~ dummies como vars. indep. en modelos con vars. dep. dummies
  - ▷ Coefs. miden la dif. predicha en la prob. en relación con el gpo. base
  - ▷ $\widehat{arrest86} = 0.380 + \text{controles} + 0.17\ negra + 0.096\ hispano$

prob. de arresto es 17% más alta p/raza negra q' raza blanca

Consideraciones Adicionales p/ Análisis de Políticas y Eval. de Programas

- Tener cuidado al evaluar programas x q' gralte. en ciencias sociales gpos.: control / tratam. — no asignados aleatoriam.
- Incluir factores q' puedan estar sistemáticamente relacionados con la var. indep. binaria de interés
  - ▷ vgr. discriminación en autorizaciones de créditos por raza
  - ▷ la aprobación depende de otros factores: ingreso, riqueza, calif. crediticia
  - ▷ necesitamos controlar por ellos si hay diferencias sistemáticas en ellos entre razas
  - ▷ $aprobación = \beta_0 + \beta_1 noblanco + \beta_2 ingreso + \beta_3 riqueza + \beta_4 califcred + otros\ factores$ ← MPL
  - ▷ discriminación contra minorías si se rechaza $H_0: \beta_1 = 0$ en favor de $H_1: \beta_1 < 0$
- Problemas de auto-selección pueden hacer q' RLM esté sesgado por falta de controles suficientes
  - ▷ individuos se autoseleccionan en comportamiento o programas ⇒ participación no aleatoria
  - ▷ vgr. individuos seleccionados p/ particip. en programa, pueden participar o no ← no aleatorio
  - ▷ se usa cuando indicador binario de participación puede relacionarse sistemáticamente con factores no observados
  - ▷ vgr. $y = \beta_0 + \beta_1 particip + u$ nos preocupa q' $E(u|partic=1) \neq E(u|partic=0)$ ← u depende de particip.
    indep. var. es endógena, β_1 estará sesgada ← no encontraremos verdadero efecto de partic. podemos encontrar efectos espurios
  - ▷ si factores correl. con particip. son observados, RLM mitiga el problema
  - ▷ " " " " no son observados, RLM genera sesgo x falta de controles
    en cuyo caso — datos panel
    — variables instrumentales
    — logit, probit

Interpretación de Resultados de Regr. con Vars. Dep. Discretas

- Cuando y es var. discreta categórica no binaria (pocos valores enteros), cómo interpretar coefs.?
- Vgr: $\underset{(0.094)}{hijos = -1.997} + \underset{(0.003)}{0.175}\,edad - \underset{(0.006)}{0.09}\,educ$, $n = 4,361$, $R^2 = 0.56$ (mujeres de Botswana)
- MCO estima efectos de x_j sobre el valor esperado/promedio de y
  - ▷ Bajo RLM.1 a RLM.4, $E(y|x) = \beta_0 + \beta_1 x_1 + \dots + \beta_k x_k \Rightarrow \beta_j$ efecto cet. par. de $x_j \uparrow$ sobre valor esp. de y
  - ▷ p/ una muestra dada, $\hat{E}(y|x) = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \dots + \hat{\beta}_k x_k \Rightarrow \hat{\beta}_j$ estimado de cómo promedio de y cambia cuando $\Delta x_j = 1$
- ∴ $\hat{\beta}_{educ} = -0.09$: estimamos q' fertilidad promedio cae 0,09 hijos dado 1 año más educ.
  si c/mujer de un gpo. de 100 tiene +1 A de educ, habrá 9 niños menos ÷ ellas
- Agregar vars. dummy con y discreta no afecta interpretación
  - ▷ Vgr: $\underset{(0.095)}{hijos = -2.071} + \underset{(0.003)}{0.177}\,edad - \underset{(0.006)}{0.079}\,educ - \underset{(0.068)}{0.362}\,electric.$, $n = 4,358$, $R^2 = 0.562$
  - ▷ comparando 100 mujeres con electric. a 100 mujeres sin ella — y misma educ y edad — estimamos q' el 1er gpo. tenga 36 niños menos
- RLM buena aprox. (inicial por MCO) a efectos parciales sobre $E(y|x)$ pero mejores modelos: logit, probit, tobit, Poisson