# Funciones de influencia

Gabriel V. Montes-Rojas

# Derivadas direccionales de estadísticos de distribución

- Sea $y$ una variable aleatoria con función de distribución $F(y)$ y con densidad $F'(y) = f(y)$.
- Sea $T(F)$ un estadístico de distribución que es infinitesimalmente robusto. Ejemplos: media, cuantil $\tau$, Gini.
- La función de influencia es la derivada direccional de $T(F)$ evaluada en $F$, que mide el efecto de una perturbación en $F$. La definición de **derivada direccional** es la siguiente: Si $H$ es otra distribución, si los datos siguen una distribución *un poco* diferente de $F$, pero que se "aproxima a" $F$, entonces medir la influencia de usar $H$ en vez de $F$ es la derivada direccional de $T$ en $F$ en la dirección de $H$:

$$IF(T) = \nabla T_{F \to H} = \frac{d}{dt} T(tH + (1-t)F)|_{t=0} = \lim_{t \to 0} \frac{T(tH + (1-t)F) - T(F)}{t}$$

- Otra forma de verlo es asumir que $x$ es un valor adicional en una muestra grande que "perturba" la distribución con masa de probabilidad $\delta_x$. $H$ es entonces $H(y) = 1[x \geq y]$ y $h(y)$ es una función de densidad que es cero excepto en $x$. Entonces,

$$IF(x; T; F) = \lim_{t \to 0} \frac{T(t\delta_x + (1-t)F) - T(F)}{t}$$

# Derivadas direccionales de estadísticos de distribución

- Una propiedad de IF es $E[IF(X;T;F)] = \int_{-\infty}^{\infty} IF(x;T;F)f(x)dx = 0$.
- Entonces definamos la función de influencia recentrada (recentered influence function, RIF)

$$RIF(x;T;F) = T(F) + IF(x;T;F)$$

Tenemos que $E[RIF(x;T;F)] = E[T(F)] + \int_{-\infty}^{\infty} IF(x;T;F)f(x)dx = T(F)$.

# Jacknife

- Una forma empírica de aproximar y estimar la IF es usar el método llamado *jacknife*.
- Definamos $T_n$ como el estimador de una muestra empírica $F_n$, $\{y_1, y_2, \dots, y_n\}$. Entonces el pseudo-valor jacknife de la observación $i$ es $T^*_{ni} = nT_n - (n-1)T_{n-1}(y_1, y_2, \dots, y_{i-1}, y_{i+1}, \dots, y_n) = nT_n - (n-1)T_{n-1}$. Entonces una aproximación a IF es usar $F_n$ en vez de $F$ y $-1/(n-1)$ en vez de $t$, tal que

$$IF(y_i; T; F_n) = \frac{T\left(\frac{n}{n-1}F_n - \frac{1}{n-1}\delta_{y_i}\right) - T(F_n)}{-1/(n-1)} = (n-1)(T_n - T_{n-1}) = T^*_{ni} - T_n$$

# IF y RIF de la media

- Sea $y$ una variable aleatoria con media $\mu_F = E[y] = T_\mu(F)$.
- $$IF(T_\mu) = \nabla T_{\mu,F\rightarrow H} = \lim_{t\rightarrow 0} \frac{T_\mu(tH + (1-t)F) - T_\mu(F)}{t} = \mu_H - \mu_F$$
- $$IF(x; T_\mu; F) = x - \mu_F$$
- $$RIF(x; T_\mu; F) = x$$

  Esto muestra que la influencia de una observación puede ser tan grande como sea esa observación.

# IF y RIF del cuantil $\tau$

- Sea $y$ una variable aleatoria con cuantil $\tau$, $q_\tau = T_{q_\tau}(F) = F^{-1}(\tau)$.
- $$IF(T_{q_\tau}) = [\tau - H(q_\tau)]/f(q_\tau)$$

  Prueba: Sea $w(t)$ el cuantil $\tau$ de $tH + (1-t)F$, tal que $tH(w(t)) + (1-t)F(w(t)) = \tau$ y además $w(0) = q_\tau$.

  $$\frac{d\tau}{dt} = 0 = H(w(t)) - F(w(t)) + w'(t)[th(w(t)) + (1-t)f(w(t))]$$

  tal que

  $$w'(t) = \frac{F(w(t)) - H(w(t))}{th(w(t)) + (1-t)f(w(t))}.$$

  Entonces,

  $$IF(T_{q_\tau}) = \nabla T_{q_\tau, F\to H} = w'(0) = [\tau - H(q_\tau)]/f(q_\tau),$$

  $$RIF(x; T_{q_\tau}; F) = q_\tau + \frac{\tau - 1[y \leq_\tau]}{f(q_\tau)}.$$

# IF y RIF de la varianza

- Sea $y$ una variable aleatoria con varianza $\sigma_F^2 = \int (y - \mu_F)^2 f(y) dy = \int (y - \mu_F)^2 dF$.
- $$IF(T_{\sigma_F^2}) = \sigma_H^2 - \sigma_F^2 + (\mu_H - \mu_F)^2$$

  Prueba:

$$\frac{d}{dt} \int (y - t\mu_H - (1-t)\mu_F)^2 [th(y) + (1-t)f(y)] dy \|_{t=0} = \int (y - \mu_F)^2 (dH - dF) + (\mu_F - \mu_H) \int (y - \mu_F)^2 dF$$
$$= -\sigma_F^2 + \int (y - \mu_F)^2 dH = -\sigma_F^2 + \int (y - \mu_H + \mu_H + \mu_F)^2 dH = \sigma_H^2 - \sigma_F^2 + (\mu_H - \mu_F)^2$$

$$RIF(x; T_{q_\tau}; F) = (x - \mu_F)^2$$

# IF y RIF del Gini

- Sea $y$ una variable aleatoria con dominio en $\mathbb{R}^+$. El coeficiente de Gini es una medida estándar de la desigualdad. La forma más fácil de entender es a partir de la curva de Lorenz (el porcentage acumulado hasta distintos niveles de $y$), ver gráfico.

$$G_F = \frac{1}{2\mu_F}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x)f(y)|x-y|dxdy = \frac{1}{\mu_F}\int F(y)(1-F(y)dy$$
$$= \frac{1}{2\mu_F}\int_0^1\int_0^1 |Q_y(\tau_1)-Q_y(\tau_1)|d\tau_1 d\tau_2$$

- $$IF(T_{G_F}) = \frac{\mu_F-\mu_H}{\mu_F}G_F + \frac{1}{\mu_F}\int[H(y)-F(y)][1-2F(y)]dy$$

Prueba: Notemos que $\mu_F G_F = \int F(y)(1-F(y))dy$, entonces definamos $\mu_{tH+(1-t)F}G_{tH+(1-t)F} = \int (tH(y)+(1-t)F(y))(1-tH(y)-(1-t)F(y))dy$. Tomemos derivadas con respecto a $t$, y evaluemos en $t=0$:

$$\frac{d}{dt}[\mu_{tH+(1-t)F}G_{tH+(1-t)F}]|_{t=0} = (\mu_H-\mu_F)G_F + \mu_F IF(T_{G_F})$$

$$\frac{d}{dt}[\int tH(y)+(1-t)F(y)(1-tH(y)+(1-t)F(y))dy]|_{t=0} = \int[H(y)-F(y)][1-2F(y)]dy$$

# IF y RIF del Gini

- $IF(x; T_G; F) = -\frac{\mu_F + x}{\mu_F} G_F + 1 - \frac{x}{\mu_F} + \frac{2}{\mu_F} \int_0^x F(y)dy$
- $RIF(x; T_G; F) = -\frac{x}{\mu_F} G_F + 1 - \frac{x}{\mu_F} + \frac{2}{\mu_F} \int_0^x F(y)dy$

# IF y RIF de la tasa de pobreza

- Sea $y$ una variable aleatoria con dominio en $\mathbb{R}^+$ y sea $z$ el valor de la canasta que define la línea de pobreza (el individuo $i$ es pobre si $y_i \leq z$). La tasa de pobreza es $P_F = \int_0^z F(y)dy$.
- $IF(T_P; F) = \frac{d}{dt} \int_0^z [th(y) + (1-t)f(y)]dy|_{t=0} = \int_0^z (dH - dF)$
- $RIF(x; T_P; F) = 1[x \leq z]$
- El índice FGT($\alpha$) ( Foster, Greer, Thorbecke, 1984) de pobreza es $T_{FGT_\alpha}(F) \int_0^z (1-y/z)^\alpha f(y)dy$, con $\alpha \geq 0$.
- $IF(T_{FGT_\alpha}; F) = \frac{d}{dt} \int_0^z (1-y/z)^\alpha [th(y) + (1-t)f(y)]dy|_{t=0} = \int_0^z (1-y/z)^\alpha h(y)dy - T_{FGT_\alpha}$
- $RIF(x; T_{FGT_\alpha}; F) = (1-x/z)^\alpha 1[x \leq z]$

# Regresión RIF

- Las funciones RIF se pueden usar en modelos de regresión en modelos condicionales, $Y|X$.
- Estamos interesados en el efecto marginal de un cambio en $X$. Sea $F_X$ la distribución de $X$ y $G_X$ una perturbación. Definamos $G_Y^*$ como la distribución inducida en $Y$ luego de un cambio en $X$.
- $$F_Y(y) = \int F_{Y|X}^{-1}(y|X=x) \cdot dF_X(x)$$
  $$G_Y^*(y) = \int F_{Y|X}^{-1}(y|X=x) \cdot dG_X(x)$$
- $T(F_Y) = \int RIF(y;T;F_Y)dF_Y(y) = \int\int RIF(y;T;F_Y)dF_{Y|X)}(y|X = x)dF_X(x) = \int E[RIF(y;T;F_Y)|X=x]dF_X(x)$. Entonces,
  $$\frac{dIF(F_Y \to (1-t)F_y + tG_Y^*)}{dt}|_{t=0} = \int E[RIF(y;T;F_Y)|X=x]d(F_X - G_X)(x).$$
- Si $X_j$ es contínua, el cambio marginal es
  $$\int dE[RIF(y;T;F_Y)|X=x]/dx \; d(F_X)(x).$$

# Regresión RIF

- Firpo, Fortin and Lemieux (2009) propone el siguiente modelo lineal:

$$E[RIF(Y;T)] = X\beta + \epsilon$$

para cualquier estadístico distribucional $T(F_Y)$.

- La IF de un criterio de evaluación social se puede ver como una medida local del impacto de una política. Rothe (2010) lo llama **efectos de política distribucional**.

# Regresión RIF

- Firpo, Fortin and Lemieux (2009) aplican este criterio a los cuantiles no condicionales, **unconditional quantile regression**.
- Para cuantiles no podemos usar la ley de esperanzas iteradas, tal que $E[Y] = E\left[E(Y|X)\right]$ pero

$$F_Y^{-1}(\tau) \neq E_X[F_{Y|X}^{-1}(\tau|X)],$$

  lo que significa que no podemos recuperar el efecto sobre los cuantiles no condicionales a través de cuantiles condicionales.
- Entonces proponen una regresión RIF de $\widehat{RIF(Y, \hat{q}_\tau)}$ on X, donde $\hat{q}_\tau$ es el cuantil $\tau$ no condicional de $Y$ y $RIF()$ es la RIF de cuantiles.

# Regresión RIF

- Mirando a los resultados de cada individuo como una función de política y de los tipos de individuos, podemos usar la descomposición contrafactual à la Oaxaca-Blinder.
- Supongamos $A$ y $B$ dos grupos (tiempo $t_0$ y $t_1$; hombres y mujeres; sindicalizados y no sindicalizados). Entonces,

$$T(F_Y^B) - T(F_Y^A) = \bar{X}_B(\beta_B - \beta_A) + (\bar{X}_B - \bar{X}_A)\beta_A,$$

donde el primer término se refiere el **efecto explicado, dotación o composición** y el segundo como **efecto precio o no explicado**.

# RIFREG en STATA

- Usar el contenido de `rifreg.zip` en `https://faculty.arts.ubc.ca/nfortin/datahead.html`
- Copiar los comandos al mismo directorio en el que se va a trabajar.
- `rifreg` $Y$ $X$, `quantile(0.1)`
- `rifreg` $Y$ $X$, `quantile(0.5)`
- `rifreg` $Y$ $X$, `quantile(0.9)`
- `rifreg` $Y$ $X$, `gini`
- `rifreg` $Y$ $X$, `variance`

# Bibliografía


- Ariza, J.F. and Montes-Rojas, G.V. (2018) "Decomposition Methods for Analyzing Inequality Changes in Latin America 2002-2014," forthcoming Empirical Economics.
- Bitler, M. P., Gelbach, J. B. and Hoynes, H. W. (2006) "What Mean Impacts Miss: Distributional Effects of Welfare Reform Experiments, American Economic Review, 96(4), 9881012. https://gspp.berkeley.edu/assets/uploads/research/pdf/BGH-AER-2006.pdf
- Essama-Nssah, B. and Lambert, P.J (2015) "Chapter 6. Influence Functions for Policy Impact Analysis, in Inequality, Mobility and Segregation: Essays in Honor of Jacques Silber Research on Economic Inequality, Volume 20, 135159. http://www.ecineq.org/milano/WP/ECINEQ2011-236.pdf
- Huber, P.J and Ronchetti, E.M. (2009) Robust Statistics (2nd ed.). Wiley.
- Firpo, S., Fortin, N.M. and Lemieux, T. (2009) "Unconditional Quantile Regressions, Econometrica, 77(3), 953-973. https://economics.ubc.ca/files/2013/05/pdf_paper_thomas-lemieux-unconditional-quantile-regressions.pdf
- Firpo, S., Fortin, N.M. and Lemieux, T. (2011) "Chapter 1. Decomposition Methods in Economics, in Orley Ashenfelter and David Card (eds.) Handbook of Labor Economics, Vol 4A, 2-102. https://www.nber.org/papers/w16045.pdf
- Rothe, C. (2010) "Nonparametric Estimation of Distributional Policy Effects," Journal of Econometrics, 155, 5670.
- https://www.stata.com/meeting/germany16/slides/de16_vankerm.pdf