Perfil de la informalidad

Modelo de Probabilidad Lineal & Modelo Probit

Tomás Abeledo Julián Macaggi Tomás Bustos

Econometría II

Segundo Cuatrimestre de 2018

Resumen

El ejercicio realizado busca encontrar patrones sistemáticos en el perfil de informalidad laboral a partir de la EAHU que realiza el INDEC.

Objetivo

Encontrar los factores que inciden sobre la probabilidad de tener un trabajo informal y detectar posibles prácticas discriminatorias en el mercado laboral.

- Pool de corte transversal
- Modelo de Probabilidad Lineal
- Modelo Probit (de variable limitada)

Resumen

Variables	Tipo
Informal	Binaria
Migrante Sudamericano	Binaria
= 1 Mujer	Binaria
De 25 a 49 Ańos	Binaria
De 50 Ańos y Más	Binaria
Primario Completo	Binaria
Secundario Incompleto	Binaria
Secundario Completo	Binaria
Universitario Incompleto	Binaria
Universitario Completo	Binaria
NOA	Binaria
NEA	Binaria
Cuyo	Binaria
Pampeana	Binaria
Patagónica	Binaria
Actividades Primarias	Binaria
Explotación Minas y Canteras	Binaria
Industria Manufacturera	Binaria
Suministro de Electricidad y Gas	Binaria
Suministro de Agua	Binaria
Construcción	Binaria
Comercio	Binaria
Transporte y Almacenamiento	Binaria
[3;6] Meses	Binaria
[6;12] Meses	Binaria
[1;5] Ańos	Binaria
>5 Ańos	Binaria

Table of Contents

Modelo de Probabilidad Lineal (LPM)

$$Y_i = \begin{cases} 1 & tiene\ descuento\ jubilatorio \\ 0 & no\ tiene\ descuento\ jubilatorio \end{cases} \tag{1}$$

Se modela la variable dependiente como una variable aleatoria de Bernoulli en donde su función de probabilidad condicional está dada por (3). Esto implica que se modelará la probabilidad de éxito para Y_i condicional en los regresores, según (4).

$$Y_i \sim \mathcal{B}ernoulli(P(\boldsymbol{X})) \tag{2}$$

$$P[Y_i = y_i|\boldsymbol{X}] = P^{y_i}(\boldsymbol{X}) * [1 - P(\boldsymbol{X})]^{1-y_i} \tag{3}$$

$$E[Y_i|\boldsymbol{X}] = P(\boldsymbol{X}) = \boldsymbol{X}\boldsymbol{\beta} \tag{4}$$

Ventajas del LPM

1. Estimación consistente mediante OLS

$$\hat{\boldsymbol{\beta}}^{OLS} = \arg\min_{\boldsymbol{\beta}} \sum_{i=1}^{N} (Y_i - \boldsymbol{\beta} \boldsymbol{X_i})^2 \tag{5}$$

Para una muestra aleatoria obtenida de un modelo poblacional lineal en sus parámetros que satisface

$$E[\boldsymbol{X}'u] = 0 \tag{6}$$

$$Rg(E[\boldsymbol{X}'\boldsymbol{X}]) = K \tag{7}$$

la solución para $\hat{\boldsymbol{\beta}}^{OLS}$ es única y, además, es un estimador consistente para $\boldsymbol{\beta}$.

$$Plim\ \hat{\beta}_j^{OLS} = \beta_j, \quad \forall j = 1, \ldots, K \tag{8}$$

Ventajas del LPM

2. Estadísticos e intervalos de confianza válidos

El modelo presentado en 3 es heterocedástico por construcción.

$$Var[Y_i|\boldsymbol{X}_i] = P(\boldsymbol{X}_i) * [1 - P(\boldsymbol{X}_i)] \tag{9}$$

Sin embargo, la corrección propuesta por White (1980) a la matriz de covarianzas, nos permite estimar consistentemente $\sigma^2_{\hat{\beta}}$ y construir estadísticos de prueba robustos a cualquier patrón de heterocedasticidad.

$$t = \frac{\hat{\beta}_j^{OLS} - \beta_j}{S^{white}_{\hat{\beta}_j^{OLS}}} \sim \mathcal{T}Student_{N-8}$$

$$P\left(\hat{\beta}_j - t_{1-\alpha/2} * S^{white}_{\hat{\beta}_j^{OLS}} < \beta_j < \hat{\beta}_j + t_{1-\alpha/2} * S^{white}_{\hat{\beta}_j^{OLS}}\right) = 1 - \alpha$$

Ventajas del LPM

3. Interpretación de los efectos parciales

El cambio en la probabilidad de éxito ante un incremento unitario de X_j permaneciendo constantes los otros regresores está dado por (10) así se trate de una variable contínua o binaria.

$$\frac{\Delta P[Y_i = 1|\boldsymbol{X}]}{\Delta X_j} = \beta_j \tag{10}$$

Serán de interés la estimación de los efectos parciales para *migrante sudamericano*, *mujer*, entre otros; controlando por otros factores que afectan la probabilidad de ser informal y que podrían estar correlacionadas con estas: el nivel de estudios alcanzado, la región del hogar, la actividad económica en la que se emplea, la edad, etc.

Variable	O2011	O2012	O2013	O2014
Migrante Sudamericano	0.08**	0.10**	0.06	0.02
= 1 Mujer	0.08***	0.07***	0.09***	0.07***
De 25 a 49 Ańos	-0.14***	-0.17***	-0.16***	-0.16***
De 50 Ańos y Más	-0.11***	-0.15***	-0.13***	-0.16***
Primario Completo	-0.08***	-0.08***	-0.01	-0.12***
Secundario Incompleto	-0.09***	-0.10***	-0.04	-0.11***
Secundario Completo	-0.20***	-0.23***	-0.16***	-0.22***
Universitario Incompleto	-0.23***	-0.25***	-0.20***	-0.22***
Universitario Completo	-0.28***	-0.30***	-0.24***	-0.31***
NOA	0.07***	0.06***	0.06***	0.08***
NEA	0.06***	0.08***	0.09***	0.10***
Cuyo	0.05***	0.03**	0.01	0.02
Pampeana	0.02	0.02	0.01	0.03*
Patagónica	-0.09***	-0.09***	-0.12***	-0.10***
Actividades Primarias	0.08***	0.09***	0.10	0.14***
Explotación Minas y Canteras	-0.13***	-0.10***	-0.11***	-0.11***
Industria Manufacturera	-0.00	0.01	0.03	0.04**
Suministro de Electricidad y Gas	-0.14***	-0.15***	-0.12***	-0.09***
Suministro de Agua	-0.02	-0.08*	0.01	0.07
Construcción	0.23***	0.25***	0.31***	0.31***
Comercio	0.06***	0.12***	0.11***	0.12***
Transporte y Almacenamiento	0.11***	0.12***	0.15***	0.08*
[3;6] Meses	-0.16***	-0.18***	-0.10***	-0.17***
[6;12] Meses	-0.26***	-0.24***	-0.25***	-0.28***
[1;5] Ańos	-0.36***	-0.37***	-0.35***	-0.37***
>5 Ańos	-0.52***	-0.52***	-0.51***	-0.50***
Constant	0.90***	0.95***	0.86***	0.93***
N	11582873	11956623	12115128	11759149
R^2	0.32	0.36	0.35	0.34

Argumentos en contra del LPM

1. Efectos parciales constantes

Según (10), el efecto parcial de X_j sobre la probabilidad de ser informal es constante y, por tanto, independiente del nivel de X_j, del signo de la variación, del nivel de otras variables, etc.

Soluciones convencionales

Introducir no linealidades en el modelo mediante:

- Descomposición del recorrido de variables continuas en binarias categóricas
- Interacciones entre variables
- Niveles en logaritmos

Argumentos en contra del LPM

2. Las probabilidades ajustadas pueden no pertenecer al intervalo unitario

Dado que el modelo no restringe el recorrido de los pronósticos y tenemos efectos marginales constantes, habrá individuos para los cuales se estima $\hat{P}^{MCO}[Y_i = 1|\boldsymbol{X}] < 0$ o $\hat{P}^{MCO}[Y_i = 1|\boldsymbol{X}] > 1$.

Soluciones convencionales

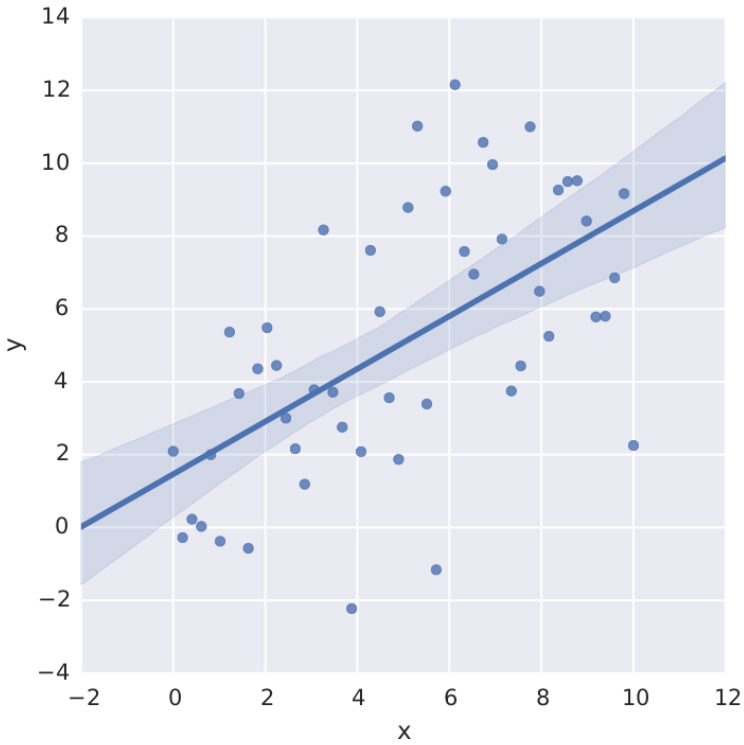
Corregir de forma *ad-hoc* las probabilidades estimadas para cada individuo según:

- Si $\hat{P}^{MCO} \leq 0 \Rightarrow \hat{P} = 0{,}01$
- Si $\hat{P}^{MCO} \geq 1 \Rightarrow \hat{P} = 0{,}99$

Argumentos en contra del LPM

3. Malas estimaciones conforme X se aleja del promedio muestral

Así como cualquier modelo de RLM, su capacidad de predicción decrece a medida que los regresores se alejan de sus promedios muestrales. Por eso, depende de cual sea la pregunta de investigación.



Si se quiere aproximar el efecto medio del tratamiento (APE), entonces el modelo permite estimarlo consistentemente.

Si se quieren aproximar probabilidades individuales, el modelo puede no tener un buen desempeño para valores extremos de X.

Table of Contents

Modelo de variable limitada

$$Y_i = \begin{cases} 1 & tiene\ descuento\ jubilatorio \\ 0 & no\ tiene\ descuento\ jubilatorio \end{cases} \tag{11}$$

Se utiliza un modelo que restringe la forma en que la variable de respuesta binaria depende de los regresores según (13). Esto implica que se modelará la probabilidad de éxito para Y_i condicional en los regresores, según (14).

$$Y_i \sim \mathcal{B}ernoulli(P(\boldsymbol{X})) \tag{12}$$

$$f(Y|\boldsymbol{X};\boldsymbol{\beta}) = [G(\boldsymbol{X}\boldsymbol{\beta})]^{y_i}\,[1 - G(\boldsymbol{X}\boldsymbol{\beta})]^{1-y_i} \tag{13}$$

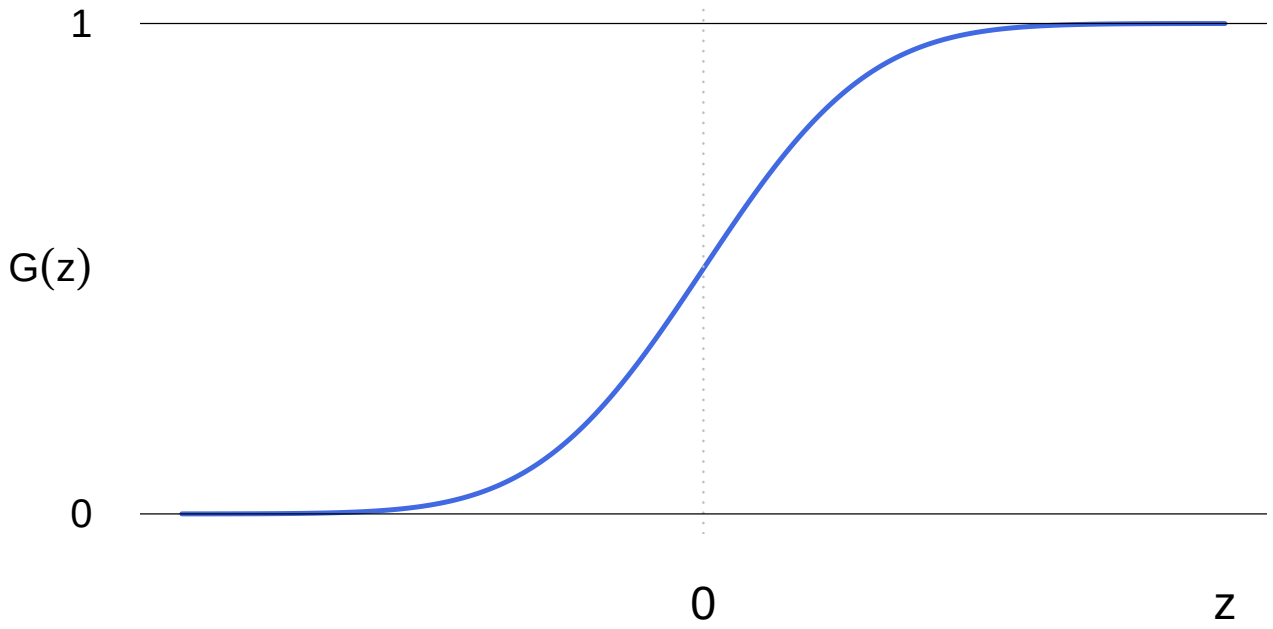
$$P[Y = 1|\boldsymbol{X}] = G(\boldsymbol{X}\boldsymbol{\beta}) \qquad con\ \ 0 < G(z) < 1\ \ \forall z \epsilon \mathbb{R} \tag{14}$$

Modelo Probit

Función de distribución acumulada

Siendo el modelo **Probit** el caso particular en que

$$G(z) = \int_{-\infty}^{z} \frac{e^{-v^2/2}}{\sqrt{2\pi}} dv \tag{15}$$



Modelo Probit

Derivación del modelo a partir del modelo de variable latente

Sea y^* una variable "latente" dada por

$$y^* = \boldsymbol{X}\boldsymbol{\beta} + \varepsilon \qquad con \quad \varepsilon \sim \mathcal{N}(0,1) \tag{16}$$

y siendo $Y = 1[y^* > 0]$, se tiene que la probabilidad de éxito para y está dada por

$$\begin{aligned} P[Y = 1|\boldsymbol{X}] &= P[y^* > 0\,|\boldsymbol{X}] \\ &= P[\varepsilon > -\boldsymbol{X}\boldsymbol{\beta}\,|\boldsymbol{X}] \\ &= 1 - G(-\boldsymbol{X}\boldsymbol{\beta}) \\ &= G(\boldsymbol{X}\boldsymbol{\beta}) \end{aligned}$$

Modelo Probit

Estimación del modelo

La estimación de los parámetros del modelo no lineal se realiza por el método de máxima verosimilitud, donde $\hat{\beta}^{Probit}$ verifica

$$\hat{\beta}^{Probit} = \arg\max_{\beta} \mathcal{L}[\boldsymbol{X};\beta] \tag{17}$$

siendo $\mathcal{L}[\boldsymbol{X};\beta]$ la función de verosimilitud de la muestra.

$$\ln \mathcal{L}[\boldsymbol{X};\beta] = \sum_{i=1}^{N} y_i * \ln[G(\boldsymbol{X}\beta)] + (1-y_i) * \ln[1-G(\boldsymbol{X}\beta)] \tag{18}$$

Bajo los supuestos del modelo, $\hat{\beta}^{Probit}$ es consistente, asintóticamente normal y pueden construirse estadísticos de prueba e intervalos de confianza asintóticamente válidos.

Variable	P2011	P2012	P2013	P2014
Migrante Sudamericano	0.26*	0.34**	0.20	0.09
= 1 Mujer	0.27***	0.24***	0.32***	0.23***
De 25 a 49 Años	-0.43***	-0.53***	-0.51***	-0.50***
De 50 Años y Más	-0.33***	-0.47***	-0.40***	-0.49***
Primario Completo	-0.28***	-0.28***	-0.02	-0.42***
Secundario Incompleto	-0.31***	-0.34***	-0.10	-0.40**
Secundario Completo	-0.67***	-0.78***	-0.52***	-0.75***
Universitario Incompleto	-0.79***	-0.84***	-0.67***	-0.75***
Universitario Completo	-0.97***	-1.11***	-0.89***	-1.17***
NOA	0.23***	0.24***	0.23***	0.30***
NEA	0.21***	0.32***	0.32***	0.37***
Cuyo	0.18***	0.15**	0.03	0.08
Pampeana	0.06	0.08	0.01	0.11*
Patagónica	-0.36***	-0.40***	-0.54***	-0.42***
Actividades Primarias	0.25***	0.30***	0.38*	0.46***
Explotación Minas y Canteras	-1.00***	-0.72***	-1.49***	-0.77***
Industria Manufacturera	0.01	0.08	0.15*	0.16**
Suministro de Electricidad y Gas	-0.77***	-1.21***	-1.07***	-0.67**
Suministro de Agua	-0.08	-0.31	0.02	0.24
Construcción	0.74***	0.85***	1.04***	1.03***
Comercio	0.23***	0.43***	0.40***	0.42***
Transporte y Almacenamiento	0.40***	0.46***	0.55***	0.33**
[3;6] Meses	-0.47***	-0.57***	-0.30***	-0.51***
[6;12] Meses	-0.74***	-0.73***	-0.74***	-0.84***
[1;5] Años	-1.02***	-1.10***	-1.00***	-1.08***
>5 Años	-1.60***	-1.67***	-1.61***	-1.57***
Constant	1.21***	1.39***	1.08***	1.34***
N	11582873	11956623	12115128	11759149
R^2_{pseudo}	0.27	0.30	0.30	0.29

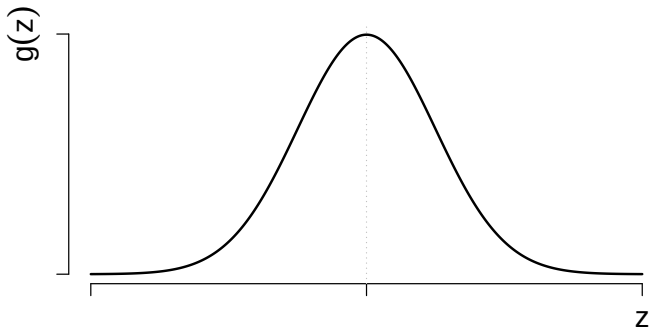
Modelo Probit

Efectos parciales

Probit permite la estimación de efectos parciales no lineales sobre la probabilidad de informalidad.

Variable continua

$$\frac{\partial P[y=1|\boldsymbol{X}]}{\partial x_j} = g(\boldsymbol{X}\boldsymbol{\beta}) * \beta_j \tag{19}$$



Variable discreta

$$G(\beta_0 + \beta_1 x_1 + \cdots + \beta_{k-1} x_{k-1} + \beta_k) - G(\beta_0 + \beta_1 x_1 + \cdots + \beta_{k-1} x_{k-1}) \tag{20}$$

Deben elegirse los niveles de interés de X para evaluar el efecto específico sobre la probabilidad de éxito.

Soluciones convencionales

- Efecto parcial en la media (PEA)

$$g(\bar{X}\hat{\beta}) * \beta_j \qquad \textit{Caso continuo}$$

$$G(\bar{x}\beta + \beta_j(c+1)) - G(\bar{x}\beta + \beta_j c) \qquad \textit{Caso discreto}$$

- Efecto parcial medio (APE)

$$\left[\sum_{i=1}^{N} \frac{g(X_i\hat{\beta})}{N}\right] * \beta_j \qquad \textit{Caso continuo}$$

$$\sum_{i=1}^{N} \frac{G(\bar{x}_i\beta + \beta_j(c+1)) - G(\bar{x}_i\beta + \beta_j c)}{N} \qquad \textit{Caso discreto}$$

Modelo Probit

Balance

Ventajas

- Rango limitado para $P[Y = 1|\boldsymbol{X}]$
- Efectos parciales no lineales
- Efectos parciales dependientes del nivel de $\boldsymbol{X}$

Desventajas

- Exposición a mismas fuentes de endogeneidad que **LPM**
- Potenciales errores de especificación
  - Heterocedasticidad en ε
  - No normalidad en ε

Table of Contents

	OLS				Probit			
	2011	2012	2013	2014	2011	2012	2013	2014
Migrante Sudamericano	0,083	0,101	0,065	0,024	0,099	0,131	0,076	0,034
= 1 Mujer	0,078	0,066	0,087	0,067	0,098	0,089	0,118	0,083
De 25 a 49 Años	-0,137	-0,168	-0,157	-0,162	-0,158	-0,197	-0,187	-0,184
De 50 Años y Más	-0,111	-0,150	-0,129	-0,158	-0,114	-0,158	-0,136	-0,163
Primario Completo	-0,082	-0,078	-0,008	-0,120	-0,097	-0,097	-0,009	-0,140
Secundario Incompleto	-0,092	-0,097	-0,036	-0,112	-0,107	-0,116	-0,037	-0,133
Secundario Completo	-0,199	-0,230	-0,159	-0,220	-0,217	-0,254	-0,175	-0,238
Universitario Incompleto	-0,235	-0,246	-0,200	-0,221	-0,236	-0,251	-0,208	-0,223
Universitario Completo	-0,276	-0,303	-0,243	-0,315	-0,290	-0,323	-0,273	-0,331
NOA	0,065	0,063	0,062	0,083	0,086	0,090	0,088	0,114
NEA	0,060	0,083	0,086	0,103	0,080	0,121	0,121	0,139
Cuyo	0,048	0,034	0,007	0,020	0,066	0,054	0,010	0,031
Pampeana	0,019	0,019	0,005	0,029	0,023	0,028	0,005	0,039
Patagónica	-0,089	-0,095	-0,123	-0,097	-0,119	-0,132	-0,169	-0,136
Actividades Primarias	0,077	0,085	0,100	0,135	0,096	0,114	0,145	0,178
Explotación Minas y Canteras	-0,127	-0,096	-0,112	-0,110	-0,255	-0,210	-0,307	-0,212
Industria Manufacturera	-0,003	0,013	0,029	0,038	0,005	0,029	0,054	0,060
Suministro de Electricidad y Gas	-0,135	-0,154	-0,118	-0,091	-0,216	-0,285	-0,265	-0,193
Suministro de Agua	-0,024	-0,082	0,005	0,066	-0,027	-0,104	0,009	0,091
Construcción	0,230	0,248	0,306	0,305	0,288	0,328	0,396	0,392
Comercio	0,065	0,121	0,114	0,118	0,086	0,163	0,154	0,159
Transporte y Almacenamiento	0,111	0,122	0,150	0,080	0,152	0,178	0,211	0,123
[3;6] Meses	-0,156	-0,185	-0,104	-0,168	-0,150	-0,179	-0,102	-0,159
[6;12] Meses	-0,257	-0,239	-0,251	-0,280	-0,216	-0,216	-0,217	-0,233
[1;5] Años	-0,362	-0,374	-0,346	-0,372	-0,321	-0,347	-0,319	-0,331
>5 Años	-0,518	-0,525	-0,508	-0,499	-0,499	-0,519	-0,508	-0,492
_cons	0,899	0,950	0,859	0,928				

Evaluación de los modelos

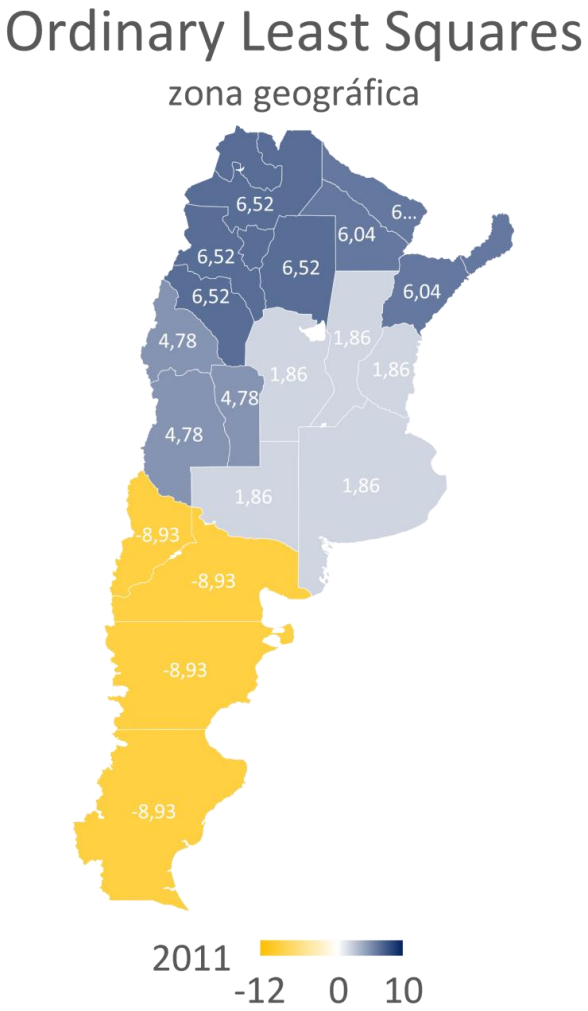
	LPM	PROBIT
AIC (2011)	1,14E+07	1,11E+07
BIC (2011)	1,14E+07	1,11E+07
R2 (2011)	0,3192	0,2667
AIC (2012)	1,13E+07	1,10E+07
BIC (2012)	1,13E+07	1,10E+07
R2 (2012)	0,3554	0,3024
AIC (2013)	1,16E+07	1,12E+07
BIC (2013)	1,16E+08	1,12E+08
R2 (2013)	0,348	0,2974
AIC (2014)	1,11E+07	1,09E+07
BIC (2014)	1,11E+07	1,09E+07
R2 (2014)	0,3434	0,2916

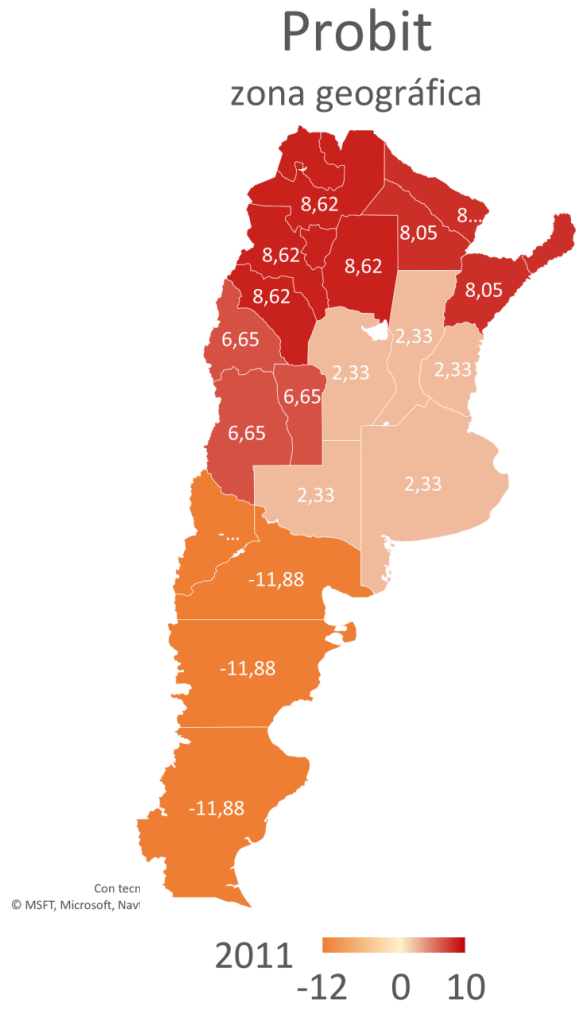
Si querés poner texto acá se puede. Sino va centrado.

	O2011	O2012	O2013	O2014
Caso más cercano a 1	1,36	1,45	1,38	1,41
Caso más cercano a 0	-0,01	0	-0,03	-0,03

Contribución a la informalidad

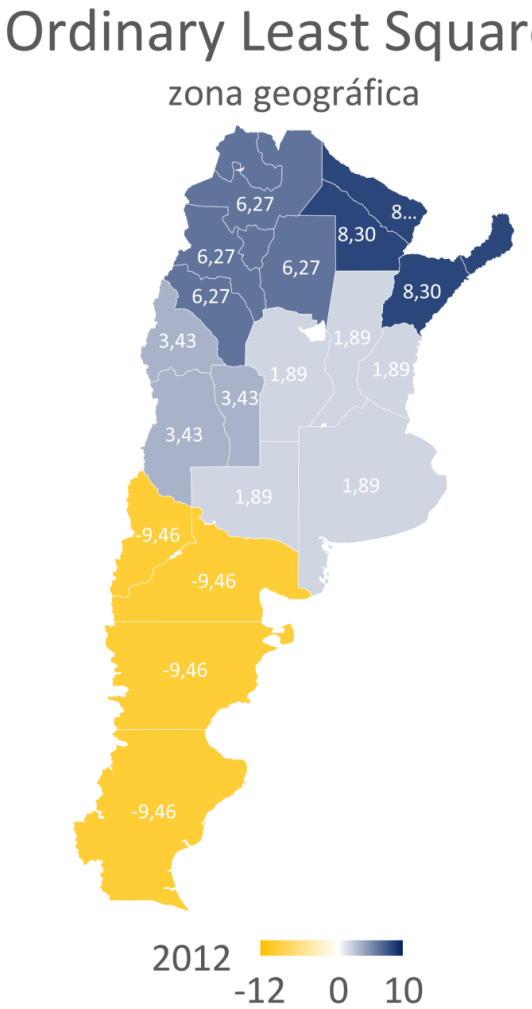
Región geográfica - 2011

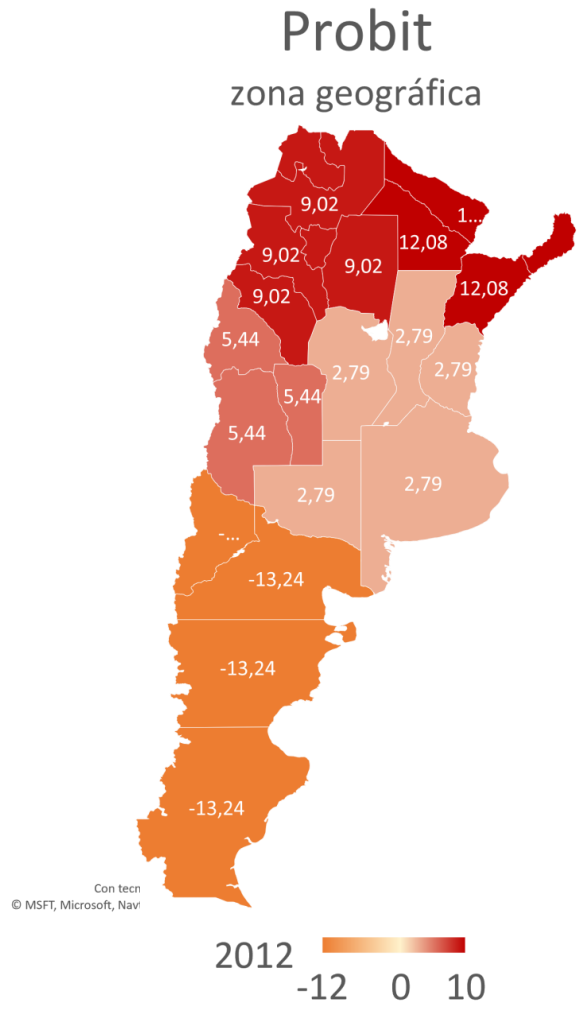




Contribución a la informalidad

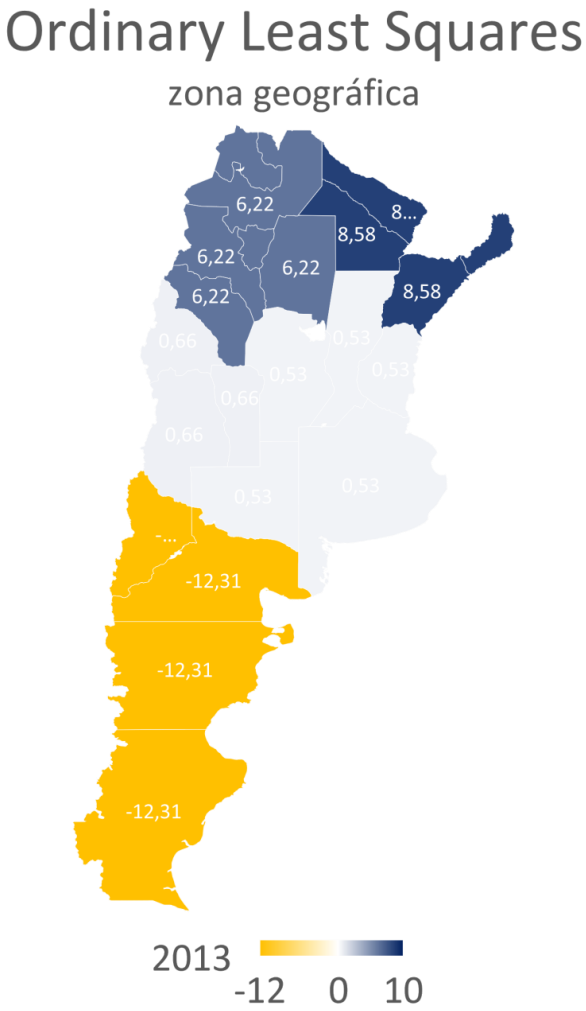
Región geográfica - 2012

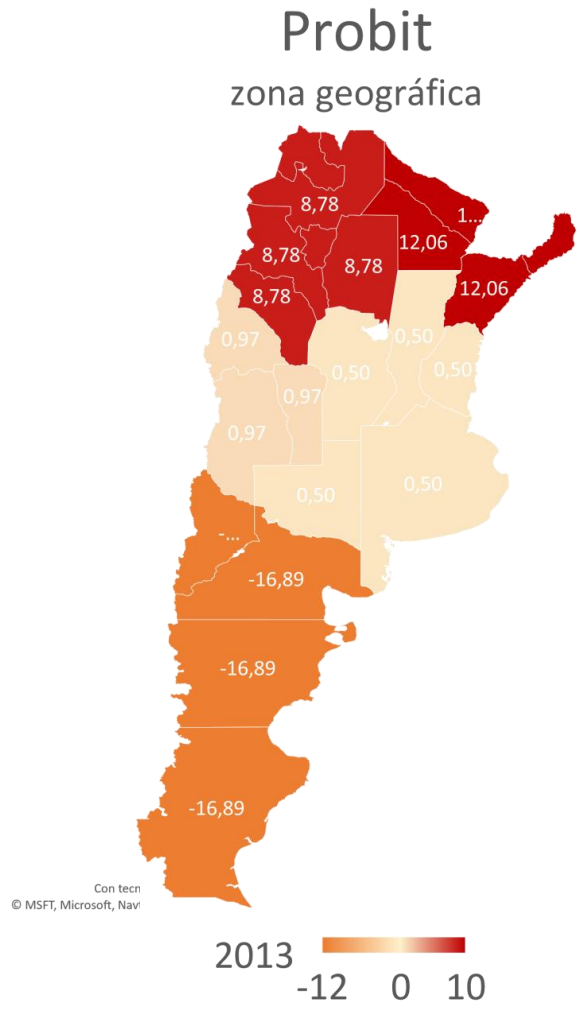




Contribución a la informalidad

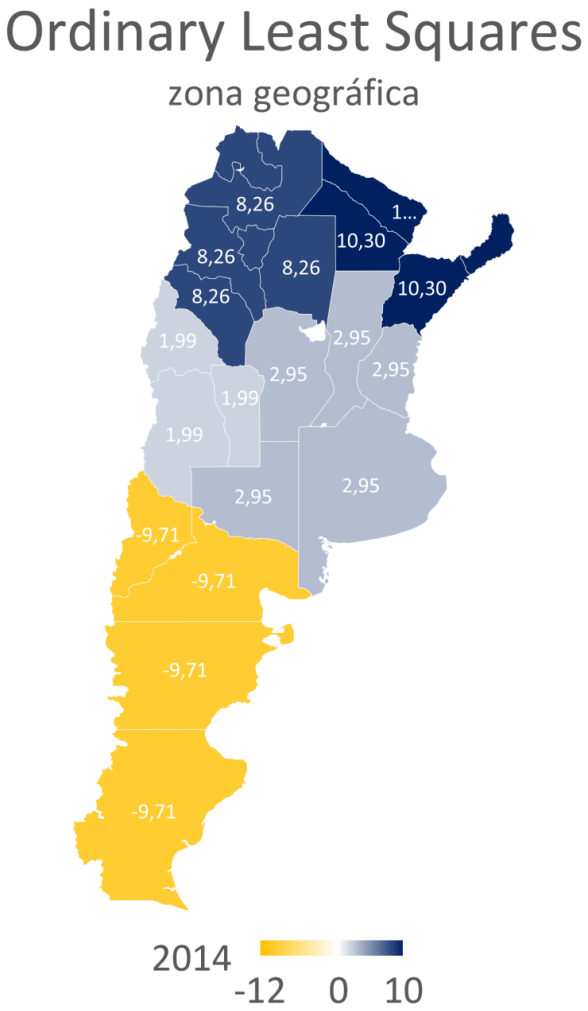
Región geográfica - 2013

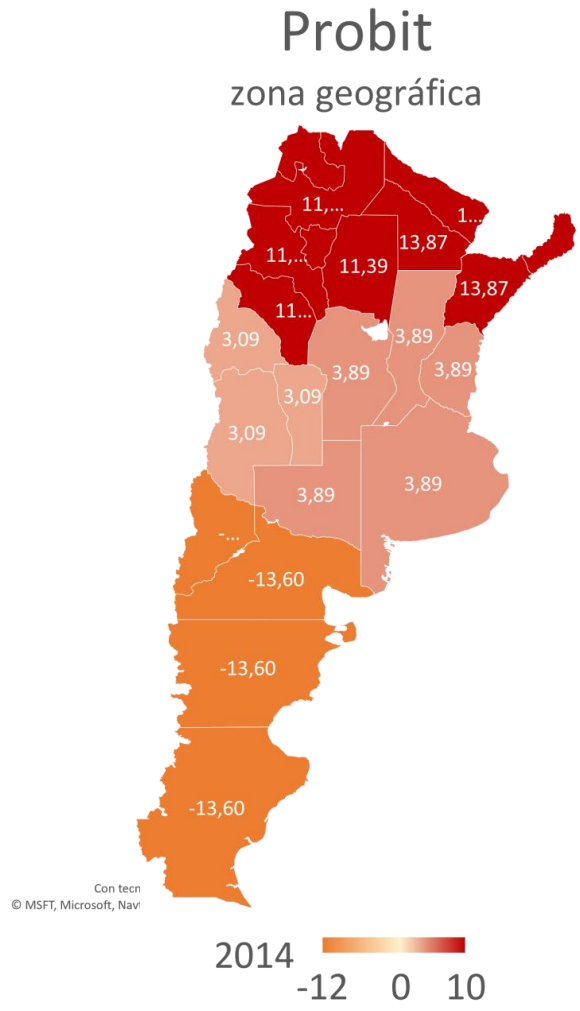




Contribución a la informalidad

Región geográfica - 2014





Estimaciones individuales

Modelo LPM vs. Probit

	OLS				Probit			
	2011	2012	2013	2014	2011	2012	2013	2014
Penalidad por sexo								
Primario incompleto	0,078	0,066	0,087	0,067	0,107	0,095	0,127	0,090
Universitario completo	0,078	0,066	0,087	0,067	0,072	0,060	0,079	0,056
Penalidad por condición de migrante								
Primario incompleto	0,083	0,101	0,065	0,024	0,103	0,134	0,079	0,035
Universitario completo	0,083	0,101	0,065	0,024	0,069	0,085	0,050	0,022
Actividades excluidas	0,083	0,101	0,065	0,024	0,087	0,112	0,065	0,029
Construcción	0,083	0,101	0,065	0,024	0,102	0,132	0,074	0,033
Casos extremos								
Penalidad por sexo[1]	0,078	0,066	0,087	0,067	0,003	0,000	0,001	0,002
Penalidad por condición de migrante[2]	0,083	0,101	0,065	0,024	0,061	0,057	0,044	0,020

[1] Mujer, de 25 a 49 años, Universitario completo, Patagonia, Electricidad y Gas, con más de 5 años de antigüedad.

[2] Hombre, de 25 a 49 años, Universitario completo, GBA, Agua, entre 1 y 5 años de antigüedad.

Table of Contents