

Anexo Suplementario
Anexo 1. Variables del Modelo de Regresión Logit

Variable dependiente Y

Código de la Variable	Variable	Valor categoría	Descripción categoría	Nombre variable dummy (en corrida del modelo)
QS411	Prueba de Papanicolau	0	No demanda PAP	
		1	Si demanda PAP	QS411

Variables independientes X

Variables categóricas

Variable	Descripción variable	Valor categoría	Descripción categoría	Nombre variable dummy	En la presentación de resultados
V025	Tipo de Área de residencia	0	Rural		Área de Residencia Rural
		1	Urbano	V025	Área de residencia Urbana
QS102	Hipertensión Arterial o Presión Alta	0	Diagnóstico negativo o no sabe		Sin hipertensión
		1	Diagnóstico positivo	QS102	Con hipertensión
QS109	Diabetes o azúcar alta	0	Diagnóstico negativo o no sabe	QS109	Sin Diabetes
		1	Diagnóstico positivo		Con Diabetes
CONOCIMIENTO	Conocimiento de prevención de cáncer	0	No tiene conocimiento		Sin Conocimiento CCU
		1	Si tiene conocimiento	CONOCIMIENTO	Con Conocimiento CCU
V190	Índice de riqueza	1	Más pobre	V190_1	Índice de Riqueza: Muy Pobre
		2	Pobre	V190_2	Índice de Riqueza: Pobre
		3	Medio	V190_3	Índice de Riqueza: media
		4	Rico	V190_4	Índice de Riqueza: Rico
		5	Más rico	V190_5	Índice de Riqueza: Muy Rico
ETNIA	Origen / Lengua étnico	1	Español	ETNIA_1	Lengua Materna: español
		2	Quechua / aimara	ETNIA_2	Lengua Materna: quechua/aymara
		3	Otras lenguas originarias	ETNIA_3	Lengua Materna: originaria distinta a quechua/aymara
		4	Otras lenguas extranjeras	ETNIA_4	Lengua Materna: extranjera
EDUCACIÓN	Nivel educativo	1	Inicial/Primaria/Secundaria	EDUCACION_1	Educación básica
		2	Superior	EDUCACION_2	Educación Superior

		3	Postgrado	EDUCACION_3	Educación de Posgrado
EST_CIVIL	Estado Civil	1	Soltero	EST_CIVIL_1	Soltero
		2	Conviviente/Casado	EST_CIVIL_2	Conviviente/Casado
		3	Viudo/Divorciado/Separado	EST_CIVIL_3	Viuda/Divorciada/Separada
PESO	Nivel de peso (relación peso/edad)	1	Bajo peso	PESO_1	Bajo peso
		2	Normal	PESO_2	Peso normal
		3	Sobre Peso	PESO_3	Sobrepeso
		4	Obesidad	PESO_4	Obesidad
REGION	Región Natural	1	Costa	COSTA	Región natural: Costa
		2	Región Sierra_ Zona Bajoandina	SIERRA - Zona Bajoandina	Región natural: Sierra - Zona Bajoandina
		3	Región Sierra_ Zona Mesoandina	SIERRA - Zona Mesoandina	Región natural: Sierra - Zona Mesoandina
		4	Región Sierra - Zona Altoandina	SIERRA - Zona Altoandina	Región natural: Sierra - Zona Altoandina
		5	Región Selva - Selva Baja	SELVA - Piedemonte Amazonico y Selva Baja	Región natural: Selva - Piedemonte Amazonico y Selva Baja
		6	Región Selva - Ceja de Selva	SELVA - Ceja de Selva	Región natural: Selva - Ceja de Selva
		7	Región Selva - Selva Alta	SELVA - Selva Alta	Región natural: Selva - Selva Alta
HV026	Urbanización de lugar de Residencia	0	Capital, Gran ciudad	HV0260	Ciudad Capital
		1	Pequeña ciudad	HV0261	Vive en una Ciudad pequeña
		2	Pueblo	HV0262	Vive en un Pueblo
		3	Campo	HV0263	Vive en el Campo

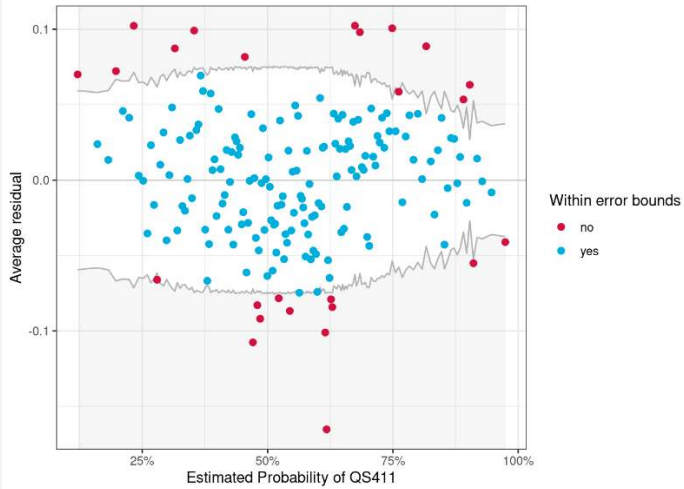
Variables Numéricas

Variable	Descripción variable	Descripción categoría	En la aplicación del Modelo)	En la presentación de resultados
QS23	Edad (día entrevista)	Edad en años cumplidos	QS23	Edad
V201	Número de niños nacidos	Número de hijos	V201	Número de hijos
V525	Edad primera relación sexual	Edad en años	V525	Edad primera RS
V136	Nº de miembros del hogar	Número de miembros	V136	Número de miembros hogar

Anexo 2. Pruebas aplicadas al Modelo de entrenamiento

1. Test de autocorrelación de los errores

Aproximadamente el 86% de los residuos están dentro de los límites de error.



2. Test de colinealidad entre las X

Check for Multicollinearity

Low Correlation

Parameter	VIF	Increased SE
QS102	1.16	1.08
QS109	1.22	1.10
CONOCIMIENTO	1.00	1.00
V190	1.07	1.03
ETNIA	1.22	1.11
EDUCACION	1.02	1.01
EST_CIVIL	1.00	1.00
PESO	1.06	1.03
QS23	1.00	1.00
V201	1.00	1.00
V525	1.00	1.00
V136	1.00	1.00
REGION	1.22	1.11

High Correlation

Parameter	VIF	Increased SE
V025	16665594.48	4082.35
HV026 1	6778822.42	4096.20

3. Test de Distribución del error y de la variable dependiente (y)

Distribution of Model Family

Predicted Distribution of Residuals

Distribution Probability

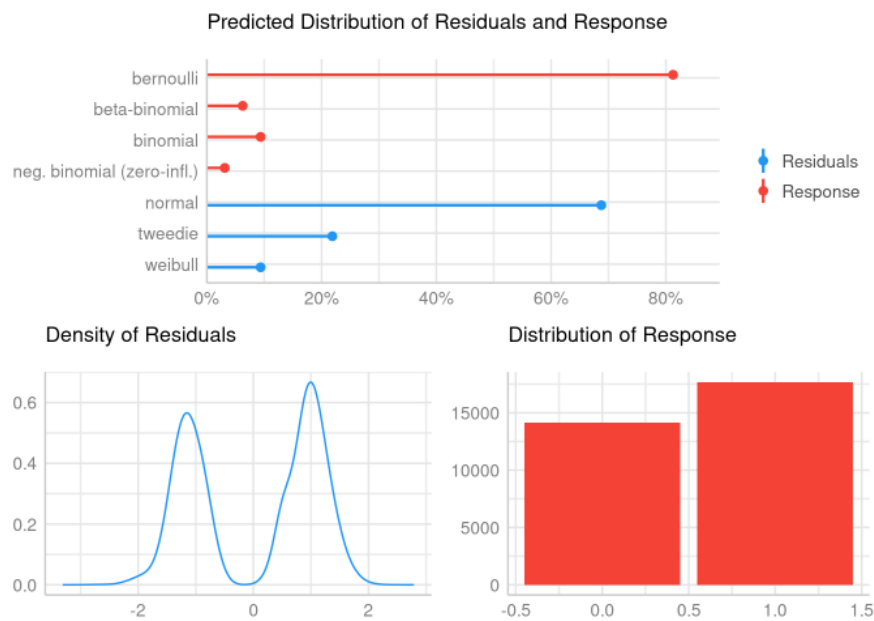
normal	69%
tweedie	22%
weibull	9%

Predicted Distribution of Response

Distribution Probability

bernoulli	81%
binomial	9%
beta-binomial	6%

```
plot(check_distribution(mod_logit))
```



Tests del Modelo (TESTS DE BONDAD DE AJUSTE)

4. Tests de significancia del modelo

Test de R2

	test	chiSq	df	pVal
1:	PrI	32033.19	31775	1.529021e-01
2:	drl	38970.66	31775	1.226577e-156
3:	PrG	31707.46	29036	2.499623e-27
4:	drG	38556.65	29036	4.944156e-282
5:	PrCT	31707.46	29036	2.499623e-27
6:	drCT	38556.65	29036	4.944156e-282

5. Test de R2 (pseudos R2)

	method	R2
1:	ssl	0.1378440
2:	ssG	0.1594316
3:	III	0.1082024
4:	IIG	0.1092373

6. Tabla de contingencia del test de Hosmer-Lemeshow

P	y1	y1hat	y0	y0hat	n	Pbar
1: 0.321	958	885.9149	2576	2648.0851	3534	0.2506833
2: 0.406	1333	1293.2470	2199	2238.7530	3532	0.3661515
3: 0.473	1550	1559.3960	1986	1976.6040	3536	0.4410057
4: 0.526	1665	1769.2329	1869	1764.7671	3534	0.5006318
5: 0.579	1894	1956.3630	1640	1577.6370	3534	0.5535832
6: 0.632	2003	2139.7146	1531	1394.2854	3534	0.6054654
7: 0.695	2422	2338.7560	1109	1192.2440	3531	0.6623495
8: 0.806	2711	2623.0935	826	913.9065	3537	0.7416153
9: 0.996	3134	3104.2823	400	429.7177	3534	0.8784047

7. Test de Bondad de ajuste

	test	stat	val	df	pVal
1:	HL	chiSq	71.263935	7	8.201991e-13
2:	mHL	F	33.032694	8	2.876306e-52
3:	OsRo	Z	25.313213	NA	2.285191e-141
4:	SstPgeq0.5	Z	3.467473	NA	5.253773e-04
5:	SstPI0.5	Z	7.109394	NA	1.165532e-12
6:	SstBoth	chiSq	62.566856	2	2.592864e-14
7:	SIIPgeq0.5	chiSq	12.664028	1	3.727583e-04
8:	SIIPi0.5	chiSq	47.393805	1	5.806589e-12

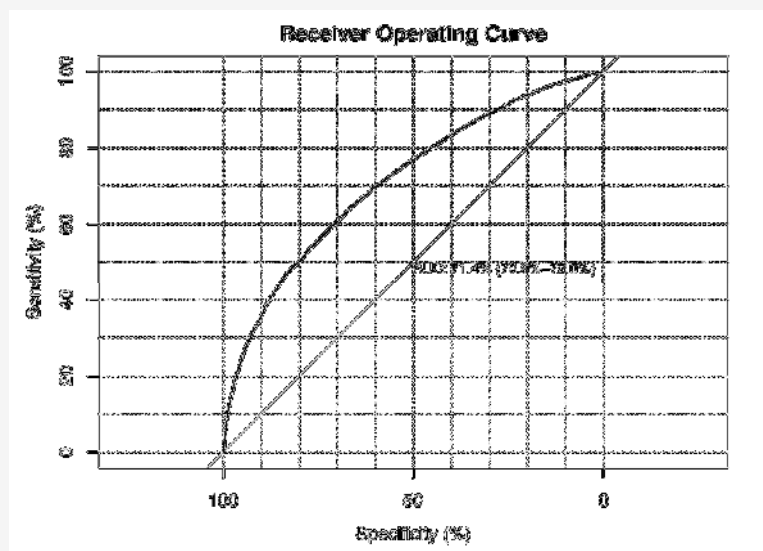
9: SIlBoth chiSq 47.585644 2 4.644177e-11

8. Test de performance AUC (Área bajo la curva). Test en el mismo set de datos del modelo

auc	lower 95% CI	upper 95% CI
71.42018	70.86312	71.97725

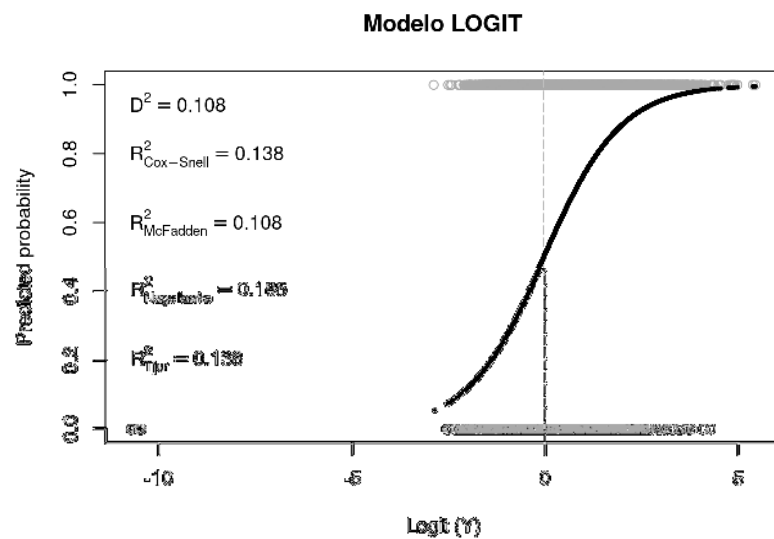
(,"interpret")

[1] "auc = 0.5 --> useless" "0.7 < auc < 0.8 --> good" "0.8 < auc < 0.9 --> excellent"



9. Plot del Modelo

plotGLM(model = mod_logit, xlab = "Logit (Y)", ylab = "Predicted probability", main = "Modelo LOGIT")



10. Test de Porcentaje de Predicciones Correctas. Test en el mismo set de datos del modelo

Confusion Matrix and Statistics

predic_logit		0	1
0		7687	4596
1		6449	13074

Accuracy: 0.6527

95% CI: (0.6475, 0.658)

No Information Rate: 0.5556

P-Value [Acc > NIR]: < 2.2e-16

Kappa : 0.2875

Mcnemar's Test P-Value : < 2.2e-16

Sensitivity: 0.7399

Specificity: 0.5438

Pos Pred Value: 0.6697

Neg Pred Value: 0.6258

Prevalence: 0.5556

Detection Rate: 0.4111

Detection Prevalence: 0.6138

Balanced Accuracy: 0.6418

'Positive' Class: 1