Heterocedasticidad en modelos de regresión ordinal con R

José Luis Cañadas Reche
Instituto de Estudios Sociales Avanzados IESA-CSIC

Noviembre 2011

Modelo logit ordinal de ventajas proporcionales

Modelo logit acumulativo con ventajas proporcionales

$$\text{logit}\left[P(Y \leq j|X = x)\right] = \ln \frac{P(Y \leq j|x)}{P(Y > j|x)} = \alpha_j - \beta_1 X_{i1} - \ldots - \beta_k X_{ik}$$

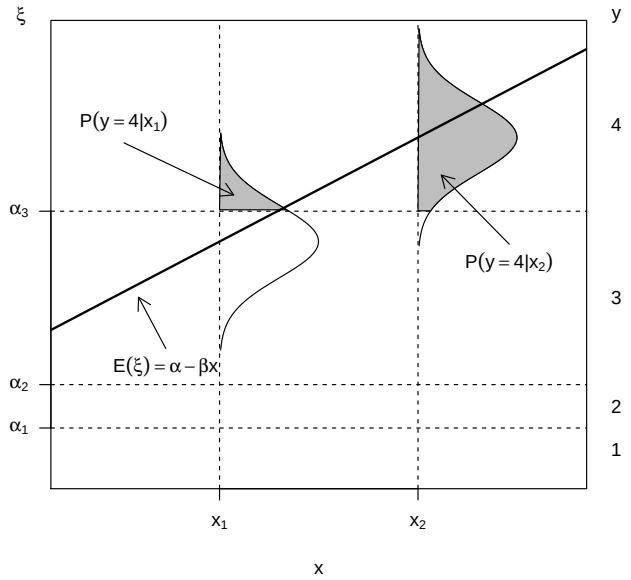
Motivación del modelo

- ξ de forma que

$$\xi_i = \alpha + \beta_1 X_{i1} + \ldots + \beta_k X_{ik} + \varepsilon_i$$

Donde Y toma los valores

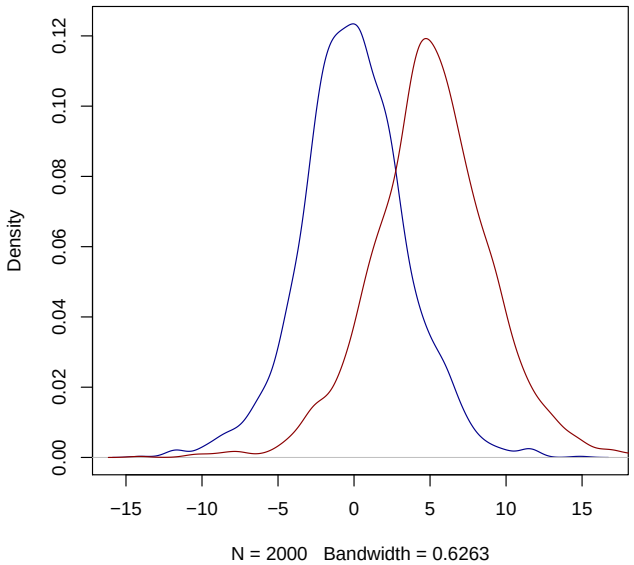
$$Y_i = \begin{cases} \text{Cat}_1 & \text{Si } \xi_i \leq \alpha_1 \\ \text{Cat}_2 & \text{Si } \alpha_1 \leq \xi_i \leq \alpha_2 \\ . & \\ \text{Cat}_{m-1} & \text{Si } \alpha_{m-2} \leq \xi_i \leq \alpha_{m-1} \\ \text{Cat}_m & \text{Si } \alpha_{m-1} \leq \xi_i \end{cases}$$

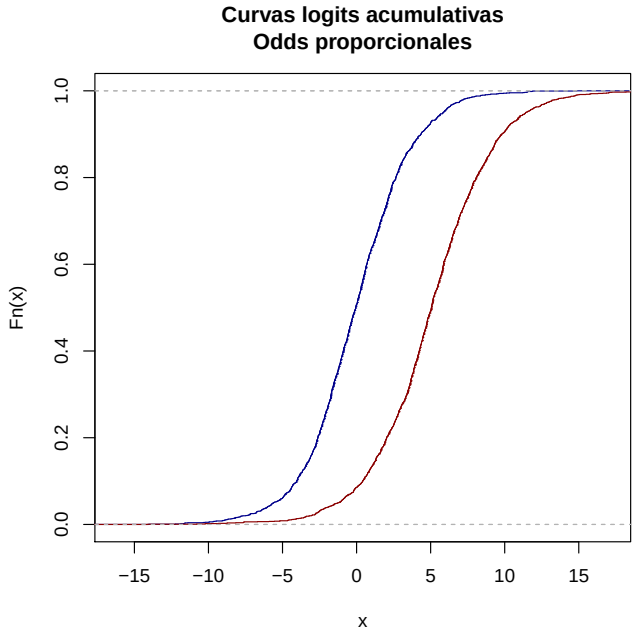


Con homocedasticidad

En estos modelos se asume un "orden estocástico".

$$P\left(Y \leq j\right) \mid x_1) \leq P\left(Y \leq j\right) \mid x_2) \; \forall j$$

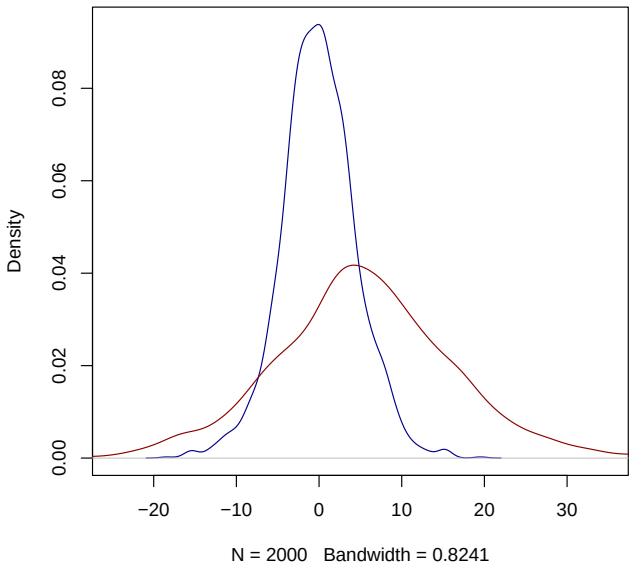


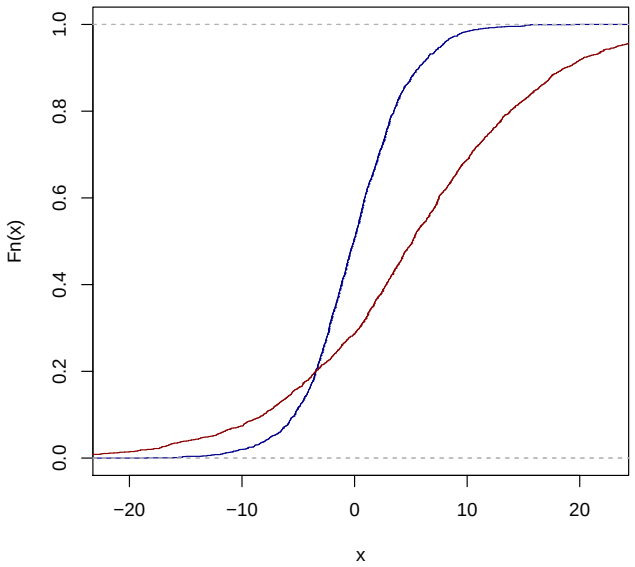


IESA
Instituto de Estudios Sociales Avanzados

Con heterocedasticidad

Si no existe orden estocástico puede pasar que
$P(Y \leq j \mid x_1) \leq P(Y \leq j \mid x_2)$ para valores pequeños de j
$P(Y \leq j \mid x_1) \geq P(Y \leq j \mid x_2)$ para valores grandes de j.
Es decir, en x_1 las respuestas están más concentradas que en x_2





IESA
Instituto de Estudios Sociales Avanzados

Modelo extendido

- Incorporación de los efectos de la dispersión

Modificación del modelo

$$\text{logit}\left[P(Y \leq j|X = x, Z = z)\right] = \ln \frac{P(Y \leq j|x)}{1 - P(Y \leq j|x)} = \frac{\alpha_j - \beta^t X}{\exp(\gamma^t Z)}$$

- Donde en el conjunto de covariables Z pueden incluir o no a las covariables X
- Diferenciamos dos partes en el modelo
  - Modelo de localización: $\alpha_j - \beta^t X$
  - Modelo de escala: $\exp(\gamma^t Z)$

IESA
Instituto de Estudios Sociales Avanzados

Proceso

1. Selección de variables, para un modelo logit acumulativo sin heterocedasticidad.
   - `polr` y `stepAIC` del paquete `MASS`
2. Comprobación hipótesis
   - Ventajas proporcionales. Uso de `vglm` del paquete VGAM
   - Bondad del ajuste global. Funciones ad-hoc para adaptar el estadístico de Hosmer-Lemeshow
3. Modelo con heterocedasticidad
   - Ajuste con la función `clm` del paquete `ordinal`[1]
   - LR test
   - Bondad del ajuste.

IESA
Instituto de Estudios Sociales Avanzados

[1] La última versión del paquete ordinal (Septiembre de 2011) permite uso de stepAIC

Ejemplo: ¿Los gobernantes tienen en cuenta la opinión de los ciudadanos?

```
> table(datos1$p19)

          Nunca     Pocas veces   Algunas veces Bastantes veces         Siempre
            331             728             817             177              19

> library(MASS)
> fit0 <- polr(p19 ~ 1 , data=datos1, method="logistic")
> fit.sup <- polr(p19 ~ sexo + edad + estudios + ideologia + participacion+p101,
+ data=datos1, method="logistic")
> fit.1 <- stepAIC(fit0,scope = list(upper = fit.sup, lower = fit0))

> fit.1$call$formula

p19 ~ participacion + ideologia + p101

> fit.sup.inter <- polr(p19 ~ ideologia * participacion * p101,
+     data = datos1, method = "logistic")
> fit.final <- stepAIC(fit0, scope = list(upper = fit.sup.inter,
+     lower = fit0))

> fit.final$call$formula

p19 ~ participacion + ideologia + p101 + participacion:ideologia
```

IESA
Instituto de Estudios Sociales Avanzados

Ventajas proporcionales

```
> library(VGAM)
> # Modelo con ventajas proporcionales
> modelo.vglm <- vglm(fit.final$call$formula, cumulative(parallel = TRUE), data=datos1)
> # Comprobamos si se cumple la asunción de ventajas proporcionales
> #en general
> modelo.vglm2 <- vglm(fit.final$call$formula, cumulative(parallel = FALSE), data=datos1)
> pchisq(deviance(modelo.vglm) - deviance(modelo.vglm2), df = df.residual(modelo.vglm) -
+ df.residual(modelo.vglm2), lower.tail = FALSE)

[1] 0.01091455

> modelo.vglm4 <- vglm(fit.final$call$formula, cumulative(parallel = FALSE~1+ideologia),
+ data=datos1)
> head(coef(modelo.vglm4, matrix = TRUE))[1:4,]

                                 logit(P[Y<=1]) logit(P[Y<=2]) logit(P[Y<=3])
(Intercept)                         -0.6601493     1.06301640     2.58031711
participacionVotó a algún partido   -1.4023409    -1.40234094    -1.40234094
participacionVotó en blanco         -0.2389180    -0.23891797    -0.23891797
ideologia                           -0.1034373    -0.09983339     0.07224897
                                 logit(P[Y<=4])
(Intercept)                           4.3205019
participacionVotó a algún partido    -1.4023409
participacionVotó en blanco          -0.2389180
ideologia                             0.2574202

> pchisq(deviance(modelo.vglm) - deviance(modelo.vglm4),
+ df = df.residual(modelo.vglm) - df.residual(modelo.vglm4), lower.tail = FALSE)

[1] 0.002377577
```

IESA
Instituto de Estudios Sociales Avanzados

Bondad de ajuste global

- Probabilidades observadas $\Longleftrightarrow$ probabilidades predichas
- Problemas con los estadísticos G^2como el X^2
- Posible solución: Test de Hosmer-Lemeshow

Test de Hosmer-Lemeshow

```
> hosmerlem = function(y, yhat, g = 10) {
+     cutyhat = cut(yhat, breaks = quantile(yhat, probs = seq(0,
+         1, 1/g)), include.lowest = TRUE)
+     obs = xtabs(cbind(1 - y, y) ~ cutyhat)
+     expect = xtabs(cbind(1 - yhat, yhat) ~ cutyhat)
+     chisq = sum((obs - expect)^2/expect)
+     P = 1 - pchisq(chisq, g - 2)
+     return(list(chisq = chisq, p.value = P))
+ }
```

Hosmer-Lemeshow II

```
> # tests de Hosmer-Lemeshow
> hosmerlem(y1,acumuladas[,1])

$chisq
[1] 4.359372

$p.value
[1] 0.823332

> hosmerlem(y2,acumuladas[,2])

$chisq
[1] 11.68138

$p.value
[1] 0.165996

> hosmerlem(y3,acumuladas[,3])

$chisq
[1] 9.194202

$p.value
[1] 0.3261793

> hosmerlem(y4,acumuladas[,4])

$chisq
[1] 7.085214

$p.value
[1] 0.5274673
```

IESA
Instituto de Estudios Sociales Avanzados

Ajuste del modelo con heterocedasticidad.

1. Ajuste de modelos con el paquete `ordinal`.
2. Modelo con y sin heterocedasticidad.
3. Modelos logit y probit.
4. LR test

Ajuste con clm (cumulative link models)

```
> library(ordinal)
> modelo.clm <- clm(fit.final$call$formula, data=datos1, Hess=T)
> modelo.clm.heter <- clm(fit.final$call$formula ,scale= ~ p101, data=datos1, Hess=TRUE)
> anova(modelo.clm.heter,modelo.clm)

Likelihood ratio tests of cumulative link models:

                 formula:                scale: link: threshold:
modelo.clm       fit.final$call$formula ~1     logit flexible
modelo.clm.heter fit.final$call$formula ~p101  logit flexible

                 no.par    AIC  logLik LR.stat df Pr(>Chisq)
modelo.clm           13 5263.3 -2618.6
modelo.clm.heter     17 5257.2 -2611.6  14.111  4   0.006949 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Resumen del modelo

```
> summary(modelo.clm.heter)
formula: fit.final$call$formula
scale:   ~p101
data:    datos1

 link  threshold nobs logLik   AIC     niter max.grad cond.H
 logit flexible  2072 -2611.59 5257.18 9(0)  4.43e-07 3.5e+04

Coefficients:
                                          Estimate Std. Error z value Pr(>|z|)
participacionVotó a algún partido          1.40535    0.29632   4.743 2.11e-06 ***
participacionVotó en blanco                0.37235    1.01866   0.366 0.714717
ideologia                                  0.09417    0.05556   1.695 0.090082 .
p101Con menos frecuencia                   0.17847    0.10279   1.736 0.082535 .
p1011 2 días por semana                    0.33318    0.11293   2.950 0.003174 **
p1013 4 días por semana                    0.41822    0.12626   3.312 0.000925 ***
p101Todos los días                         0.35124    0.12079   2.908 0.003638 **
participacionVotó a algún partido:ideologia -0.21828    0.06088  -3.585 0.000337 ***
participacionVotó en blanco:ideologia      -0.10734    0.21977  -0.488 0.625257
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

log-scale coefficients:
                         Estimate Std. Error z value Pr(>|z|)
p101Con menos frecuencia -0.23846    0.06465  -3.689 0.000226 ***
p1011 2 días por semana  -0.05921    0.06328  -0.936 0.349463
p1013 4 días por semana  -0.07666    0.07119  -1.077 0.281534
p101Todos los días       -0.03182    0.06599  -0.482 0.629698
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Threshold coefficients:
                            Estimate Std. Error z value
Nunca|Pocas veces            -0.6359     0.2761  -2.303
Pocas veces|Algunas veces     0.9917     0.2758   3.595
Algunas veces|Bastantes veces 3.0986     0.2970  10.433
Bastantes veces|Siempre       5.4205     0.3878  13.977
```

IESA
Instituto de Estudios Sociales Avanzados

Bondad del ajuste del modelo con heterocedasticidad

```
> hosmerlem(y1,acumuladas.heter[,1])

$chisq
[1] 5.42188

$p.value
[1] 0.7116788

> hosmerlem(y2,acumuladas.heter[,2])

$chisq
[1] 10.52692

$p.value
[1] 0.2299714

> hosmerlem(y3,acumuladas.heter[,3])

$chisq
[1] 7.868574

$p.value
[1] 0.4464125

> hosmerlem(y4,acumuladas.heter[,4])

$chisq
[1] 8.255067

$p.value
[1] 0.4089594
```

Que nos da un mejor ajuste que el modelo sin heterocedasticidad

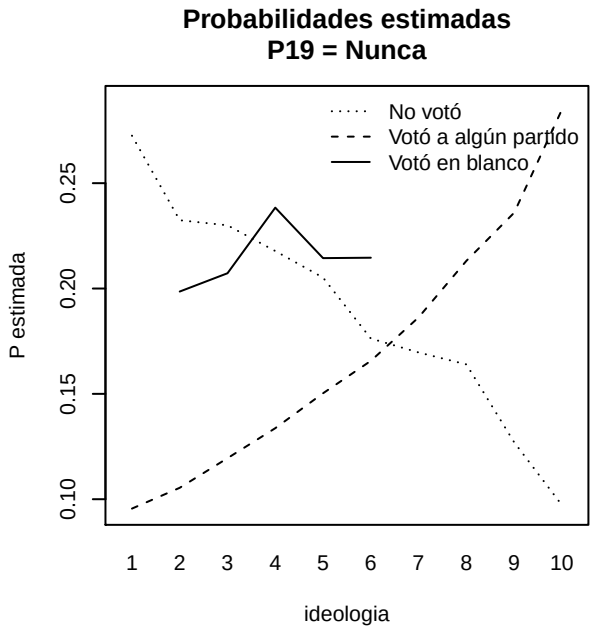
IESA
Instituto de Estudios Sociales Avanzados

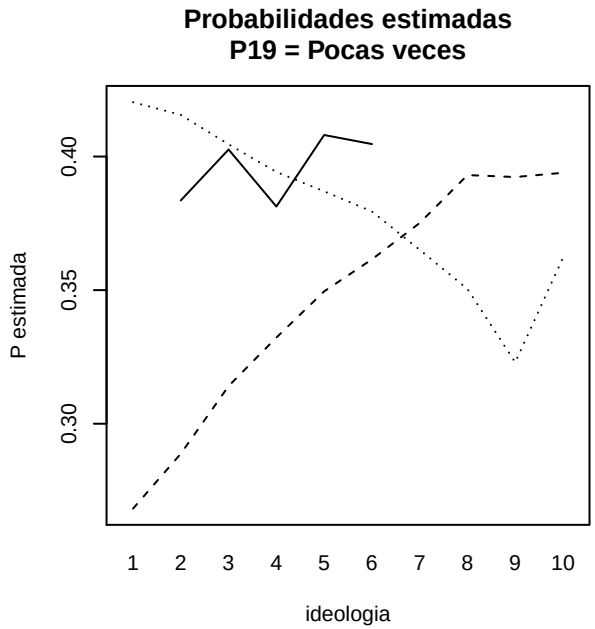
- La formulación general del modelo

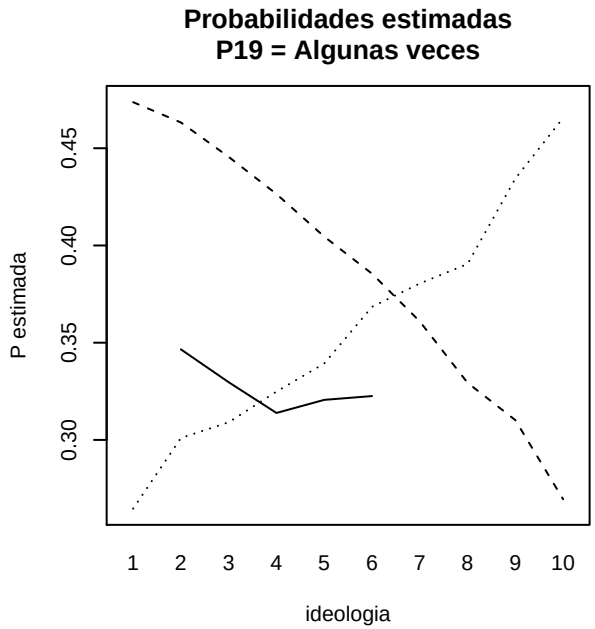
$$\text{logit}\left[P(Y \leq j|X = x, Z = z)\right] = \ln \frac{P(Y \leq j|x)}{1 - P(Y \leq j|x)} = \frac{\alpha_j - \beta^t X}{\exp(\gamma^t Z)}$$

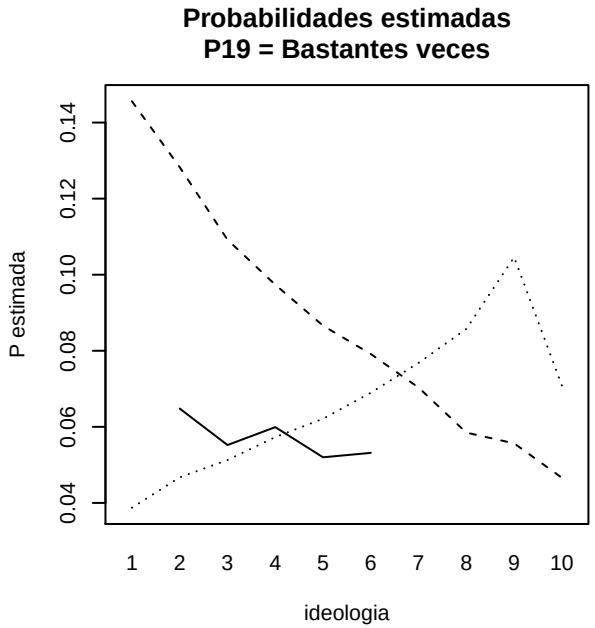
- Las estimaciones para el modelo de localización se modifican mediante $\exp(\gamma^t Z)$ para las diferentes categorías de la variable P101
  - $1^{\underline{a}}$ categoría : $\exp(\gamma^t Z) = e^0 = 1$
  - $2^{\underline{a}}$ categoría: $\exp(\gamma^t Z) = e^{-0,238} = 0,788$
  - $3^{\underline{a}}$ categoría : $\exp(\gamma^t Z) = e^{-0,059} = 0,943$
  - $4^{\underline{a}}$ categoría : $\exp(\gamma^t Z) = e^{-0,077} = 0,926$
  - $5^{\underline{a}}$ categoría : $\exp(\gamma^t Z) = e^{-0,032} = 0,969.$

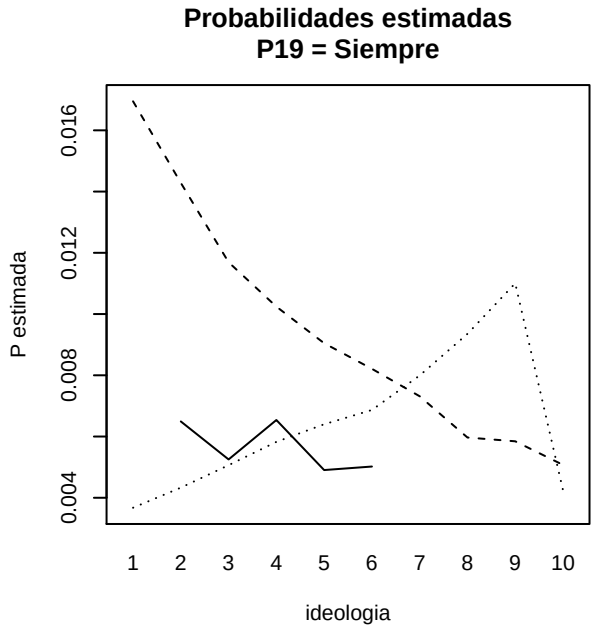
IESA
Instituto de Estudios Sociales Avanzados


Probabilidades estimadas
P19 = Nunca
No votó
Votó a algún partido
Votó en blanco
P estimada
0.10
0.15
0.20
0.25
1 2 3 4 5 6 7 8 9 10
ideologia


Probabilidades estimadas
P19 = Pocas veces
P estimada
0.30
0.35
0.40
1 2 3 4 5 6 7 8 9 10
ideologia


Probabilidades estimadas
P19 = Algunas veces
P estimada
0.30
0.35
0.40
0.45
1 2 3 4 5 6 7 8 9 10
ideologia


Probabilidades estimadas
P19 = Bastantes veces
P estimada
0.04
0.06
0.08
0.10
0.12
0.14
1 2 3 4 5 6 7 8 9 10
ideologia


Probabilidades estimadas
P19 = Siempre
P estimada
0.004
0.008
0.012
0.016
1 2 3 4 5 6 7 8 9 10
ideologia

GRACIAS

IESA
Instituto de Estudios Sociales Avanzados

Referencias

Alan Agresti.

Categorical Data Analysis.
WILEY, second edition, 2002.
ISBN 0-471-36093-7.

R. Michael Alvarez and John Brehm.

Hard Choices, Easy Answers: Values, Information, and American Public Opinion.
Princeton University Press, 2002.
ISBN 069109635X.

R. H. B. Christensen.

ordinal—regression models for ordinal data, 2010.
R package version 2011.09-14 http://www.cran.r-project.org/package=ordinal/.

Simon Sheather.

A Modern Approach to Regression with R (Springer Texts in Statistics).
Springer, 2009.
ISBN 9780387096078.

Laura A. Thompson.

S-PLUS (and R) Manual to Acccompany Agresti's Categorical Data Analysis, 2007.

W.N Venables and B.D. Ripley.

Modern Applied Statistics with S.
Springer, fourth edition edition, 2002.

Thomas W. Yee.

The vgam package for categorical data analysis.
Journal of Statistical Software, 32(10):1–34, 2010.

IESA
Instituto de Estudios Sociales Avanzados