

Profesor: Héctor Bahamonde, PhD.

e: hector.bahamonde@uoh.cl

w: www.hectorbahamonde.com

Curso: MLE.

TA: Gonzalo Barria.

I. OUTCOMES ORDENADOS: ORDERED LOGIT/PROBIT

Los *outcomes ordenados* son aquellos donde la variable dependiente es categórica, pero representa cierto orden. Uno de los ejemplos más típicos, es el de la *escala de Likert*. La escala de Likert se utiliza para caracterizar, por ejemplo, niveles de aprobación/desaprobación de candidatos, políticas públicas, etc. La escala de Likert tiene en general cinco niveles de respuesta: *muy de acuerdo*, *acuerdo*, *neutral*, *desacuerdo*, *muy en desacuerdo*.

Lamentablemente, analistas insisten en analizar estas variables dependientes intervalares usando métodos lineales OLS (Long 1997, p. 115). Esto genera sesgos en los análisis porque asume El otro punto que sugiera la figura (en el panel inferior) es que los errores son heteroesquedásticos. que los intervalos numéricos entre cada categoría son constantes. Es decir, la distancia (numérica) que existe entre *muy de acuerdo* y *acuerdo* es la misma que *desacuerdo* y *muy en desacuerdo*. Y esto no es cierto. Es por esto que debemos considerar este *data generating process* distinto, y utilizar métodos de MLE. En otras palabras, no podemos asumir que el proceso ordinal sea necesariamente intervalar.

Modelo Latente Una manera de motivar este modelo es vía modelos latentes. En esta motivación, tú verías una variable dependiente y_i sólo con 1's, 2's, 3's, 4's y 5's (continuando con el ejemplo de la escala de Likert). Sin embargo, el *data generating process* de y_i sigue un proceso *latente* y_i^* (que no ves), que de manera análoga a la motivación logit, es gatillado por umbrales (o “thresholds”) τ . Formalmente,

$$y_i = \left\{ \begin{array}{ll} 1_{\text{muy de acuerdo}} & \text{si } \tau_0 = -\infty \leq y_i^* < \tau_1 \\ 2_{\text{acuerdo}} & \text{si } \tau_1 \leq y_i^* < \tau_2 \\ 3_{\text{neutral}} & \text{si } \tau_2 \leq y_i^* < \tau_3 \\ 4_{\text{desacuerdo}} & \text{si } \tau_3 \leq y_i^* < \tau_4 \\ 5_{\text{muy en desacuerdo}} & \text{si } \tau_4 \leq y_i^* < \tau_5 = \infty \end{array} \right\} \quad (1)$$

Como lo podrás notar, esta motivación es muy similar al modelo latente del modelo logit,

$$y_i = \left\{ \begin{array}{ll} 1 & \text{si } y_i^* > \tau \\ 0 & \text{si } y_i^* \leq \tau \end{array} \right\} \quad (2)$$

Y de hecho, el modelo estructural ordered probit/logit te tendría que ser muy familiar,

$$y_i^* = x_i\beta + \epsilon_i \quad (3)$$

Una manera gráfica de ver esta motivación vía modelos latentes, es a través de la siguiente figura,

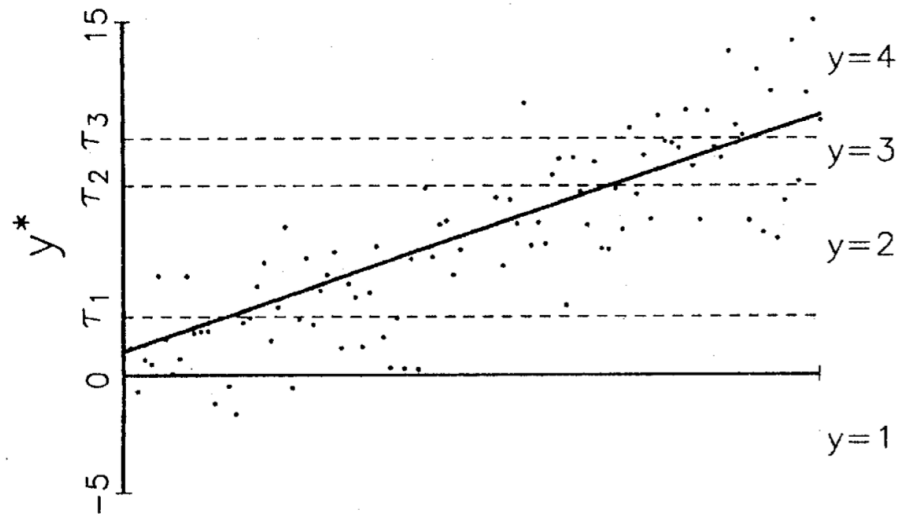
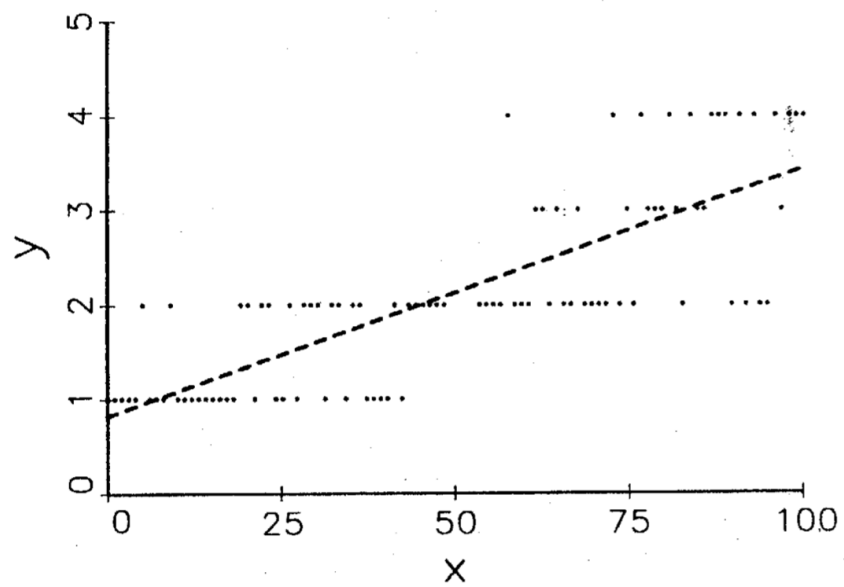
Panel A: Regression of Latent y^* Panel B: Regression of Observed y 

Figure 5.1. Regression of a Latent Variable y^* Compared to the Regression of the Corresponding Observed Variable y

Como sabemos, la estimación vía modelos latentes no es posible, no podemos estimar una regresión entre y_i^* y x (Long 1997, p. 117). El otro punto que sugiera la figura (en el panel inferior) es que los errores son heteroeskedásticos.

Supuestos Distribucionales Debido a que esta es una extensión directa del modelo logit/probit, tenemos dos opciones de distribuciones, logit y probit. Como ya sabemos, **estas son distribuciones de los errores**. El PDF del modelo ordered probit es formalmente,

$$\phi(e) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{e^2}{2}\right) \quad (4)$$

donde $E[\phi(e)] \sim (0, 1)$.

El PDF del modelo ordered logit es formalmente definido como sigue,

$$\lambda(\epsilon) = \frac{\exp(\epsilon)}{[1 + \exp(\epsilon)]^2} \quad (5)$$

donde $E[\lambda(\epsilon)] \sim (0, \frac{\pi^2}{3})$.

Estimacion: Probabilidades y Likelihood Continuando con el Equation 3, $x_i\beta + \epsilon_i$ es posible de ser calculado en términos de probabilidades de la siguiente manera,

$$\Pr(y_i = 1|x_i) = \Pr(\tau_0 \leq x_i\beta + \epsilon_i < \tau_1|x_i) \quad (6)$$

Y asumiendo que las observaciones son independientes entre sí, el likelihood está dado por,

$$L(\beta, \tau|y, X) = \prod_{i=1}^N \Pr(y_i) \quad (7)$$

II. PROGRAMACIÓN

Carguemos los datos

```
p_load(foreign)
dat = read.dta("https://github.com/hbahamonde/MLE/raw/master/Datasets/nas92_ordered.dta")
```

Hagamos un resumen,

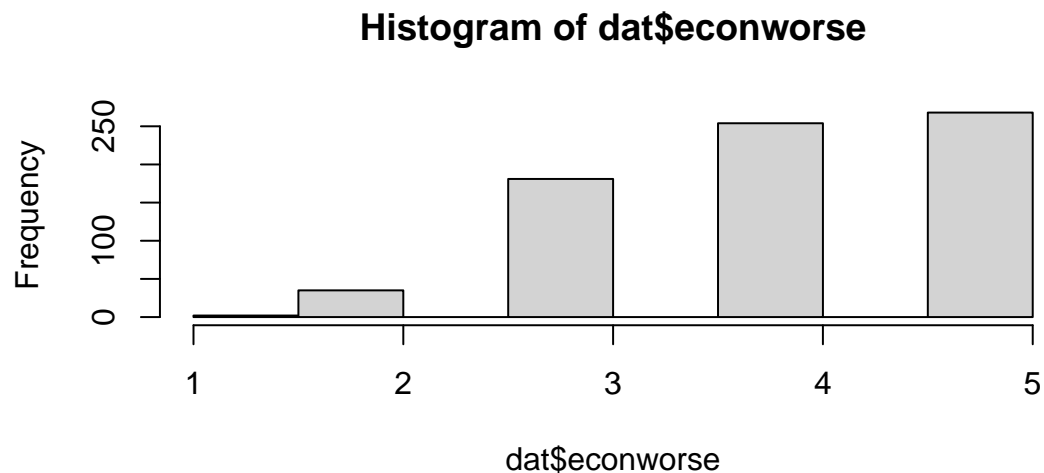
```
summary(dat)
```

```
##      bushapp      ideology      bushideo      clintideo      distbushideo
## Min.      :0.00   Min.      :1.000   Min.      :1.00   Min.      :1.000   Min.      :0.000
## 1st Qu.:0.00   1st Qu.:2.000   1st Qu.:4.00   1st Qu.:2.000   1st Qu.:1.000
## Median :1.00   Median :5.000   Median :6.00   Median :3.000   Median :2.000
## Mean   :1.25   Mean   :4.245   Mean   :5.15   Mean   :3.096   Mean   :2.106
## 3rd Qu.:2.00   3rd Qu.:6.000   3rd Qu.:6.00   3rd Qu.:4.000   3rd Qu.:3.000
## Max.    :3.00   Max.    :7.000   Max.    :7.00   Max.    :7.000   Max.    :6.000
## NA's    :23    NA's    :151   NA's    :72    NA's    :84    NA's    :175
## distclintideo      econworse      oppforce      gulfwarworthit
## Min.      :0.000   Min.      :1.000   Min.      :1.000   Min.      :0.0000
## 1st Qu.:1.000   1st Qu.:3.000   1st Qu.:3.000   1st Qu.:0.0000
## Median :2.000   Median :4.000   Median :3.000   Median :1.0000
## Mean   :2.068   Mean   :4.015   Mean   :2.964   Mean   :0.5835
## 3rd Qu.:3.000   3rd Qu.:5.000   3rd Qu.:3.000   3rd Qu.:1.0000
## Max.    :6.000   Max.    :5.000   Max.    :5.000   Max.    :1.0000
## NA's    :190   NA's    :10    NA's    :8     NA's    :37
##      pid      educyears      govtemp      union
## Min.    :-3.0000   Min.     : 2.00   Min.     :0.000   Min.     :0.0000
## 1st Qu.: -2.0000   1st Qu.:12.00   1st Qu.:0.000   1st Qu.:0.0000
## Median : 0.0000   Median :13.00   Median :0.000   Median :0.0000
## Mean    :-0.1092   Mean     :13.57   Mean     :0.136   Mean     :0.1653
## 3rd Qu.: 2.0000   3rd Qu.:16.00   3rd Qu.:0.000   3rd Qu.:0.0000
## Max.     : 3.0000   Max.     :17.00   Max.     :1.000   Max.     :1.0000
## NA's     :8       NA's     :5
##      faminc      minority      _est_m2      _est_m1
## Min.     : 1.50   Min.     :0.0000   Min.     :0.0000   Min.     :0.0000
## 1st Qu.: 21.00   1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.0000
## Median : 37.50   Median :0.0000   Median :1.0000   Median :1.0000
## Mean     : 41.93   Mean     :0.1347   Mean     :0.6787   Mean     :0.6787
```

```
## 3rd Qu.: 55.00 3rd Qu.:0.0000 3rd Qu.:1.0000 3rd Qu.:1.0000
## Max. :140.00 Max. :1.0000 Max. :1.0000 Max. :1.0000
## NA's :55
```

En esta aplicación pensaremos en la variable `econworse`: *Cree usted que la economía ha empeorado?* [Muy de acuerdo, de acuerdo, neutral, desacuerdo, muy en desacuerdo]. Veámos cómo se ve esta variable.

```
hist(dat$econworse)
```



El paquete de R que usaremos se llama `polr`—éste especifica que la variable dependiente de ser `factor`.

```
dat$econworse.f = as.factor(dat$econworse) # transforma a factor
head(dat$econworse.f) # ve como queda

## [1] 4 5 5 5 4 5
## Levels: 1 2 3 4 5
```

Ahora estimemos un ologit y un oprobit.

	Model 1	Model 2
ideology	-0.27*** (0.05)	-0.17*** (0.03)
educyears	-0.09* (0.04)	-0.06* (0.02)
faminc	-0.00 (0.00)	-0.00 (0.00)
1 2	-8.97*** (1.16)	-4.63*** (0.48)
2 3	-5.57*** (0.62)	-3.27*** (0.35)
3 4	-3.44*** (0.58)	-2.11*** (0.33)
4 5	-1.91*** (0.57)	-1.17*** (0.33)
AIC	1353.41	1350.17
BIC	1383.68	1380.44
Log Likelihood	-669.70	-668.09
Deviance	1339.41	1336.17
Num. obs.	558	558

*** $p < 0.001$; ** $p < 0.01$; * $p < 0.05$

Table 1: *Statistical models*

```
p_load(MASS)
o.logit = polr(econworse.f ~ ideology + educyears + faminc, data = dat, method = "logistic") # o-l
o.probit = polr(econworse.f ~ ideology + educyears + faminc, data = dat, method = "probit") # o-pr
```

Desde ahora en adelante, prestaremos más atención a la presentación de resultados. Hagamos una tabla.

```
p_load(texreg)
texreg(list(o.logit, o.probit)) # usa "screenreg" no "texreg".
```

Ya que los resultados son (casi) siempre similares, durante el resto de la clase solo veremos el `o.logit`.

Fíjate que vemos mas interceptos, uno por cada τ . Debido a que y_i tiene cinco valores, hay cuatro τ . Esto se puede interpretar así,

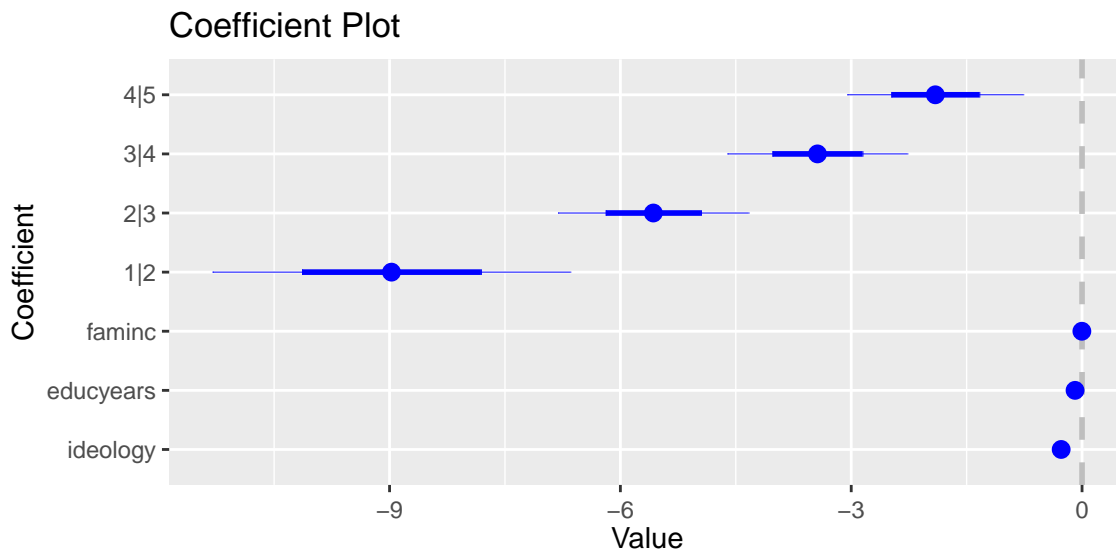
$$\begin{aligned}
\text{logit}(Pr(y_i \leq 1)) &= -8.97 - 0.27 \times \text{ideology}_i - 0.09 \times \text{educyears}_i - 0 \times \text{faminc}_i \\
\text{logit}(Pr(y_i \leq 2)) &= -5.57 - 0.27 \times \text{ideology}_i - 0.09 \times \text{educyears}_i - 0 \times \text{faminc}_i \\
\text{logit}(Pr(y_i \leq 3)) &= -3.44 - 0.27 \times \text{ideology}_i - 0.09 \times \text{educyears}_i - 0 \times \text{faminc}_i \\
\text{logit}(Pr(y_i \leq 4)) &= -1.91 - 0.27 \times \text{ideology}_i - 0.09 \times \text{educyears}_i - 0 \times \text{faminc}_i
\end{aligned} \tag{8}$$

III. INTERPRETACIÓN

Ahora interpretaremos el modelo.

Intervalos de Confianza Inspeccionemos los intervalos de confianza,

```
p_load(coefplot)
coefplot(o.logit)
```



El eje x del gráfico está en escala de logit, o *log-odds*. Es decir, si subo una unidad en **ideology**, esperamos que **econworse.f** suba -0.27 en la escala logit, o **log-odds** manteniendo las otras variables constantes en sus medias.

Odds Ratios Calculemos ahora los *odds ratios*.


```
exp(coef(o.logit))

## ideology educyears faminc
## 0.7622486 0.9131977 0.9977306
```

Esto quiere decir que cuando subo una unidad en **ideology** (i.e. me vuelvo mas derechista) es 0.76 más *posible* que encuentre la economía peor (**econworse**), manteniendo el resto de las variables constantes en sus medias. El supuesto que permite esta comparacion, i.e. de que los odds ratios se aplican a cualquier nivel de la y_i , se llama **parallel regression assumption** (Long 1997, p. 140). Por esto es que estos *odds ratios* son *proporcionales* (aplican en cualquier intervalo de **ideology**).

Cambios Marginales Calculemos ahora los cambios marginales. Pensemos en dos perfiles.

```
p_load(margins)

# 1
margins(o.logit, at = list(
  ideology = max(dat$ideology, na.rm = T), # derechista
  educyears = min(dat$educyears, na.rm = T)) # sin educ
)

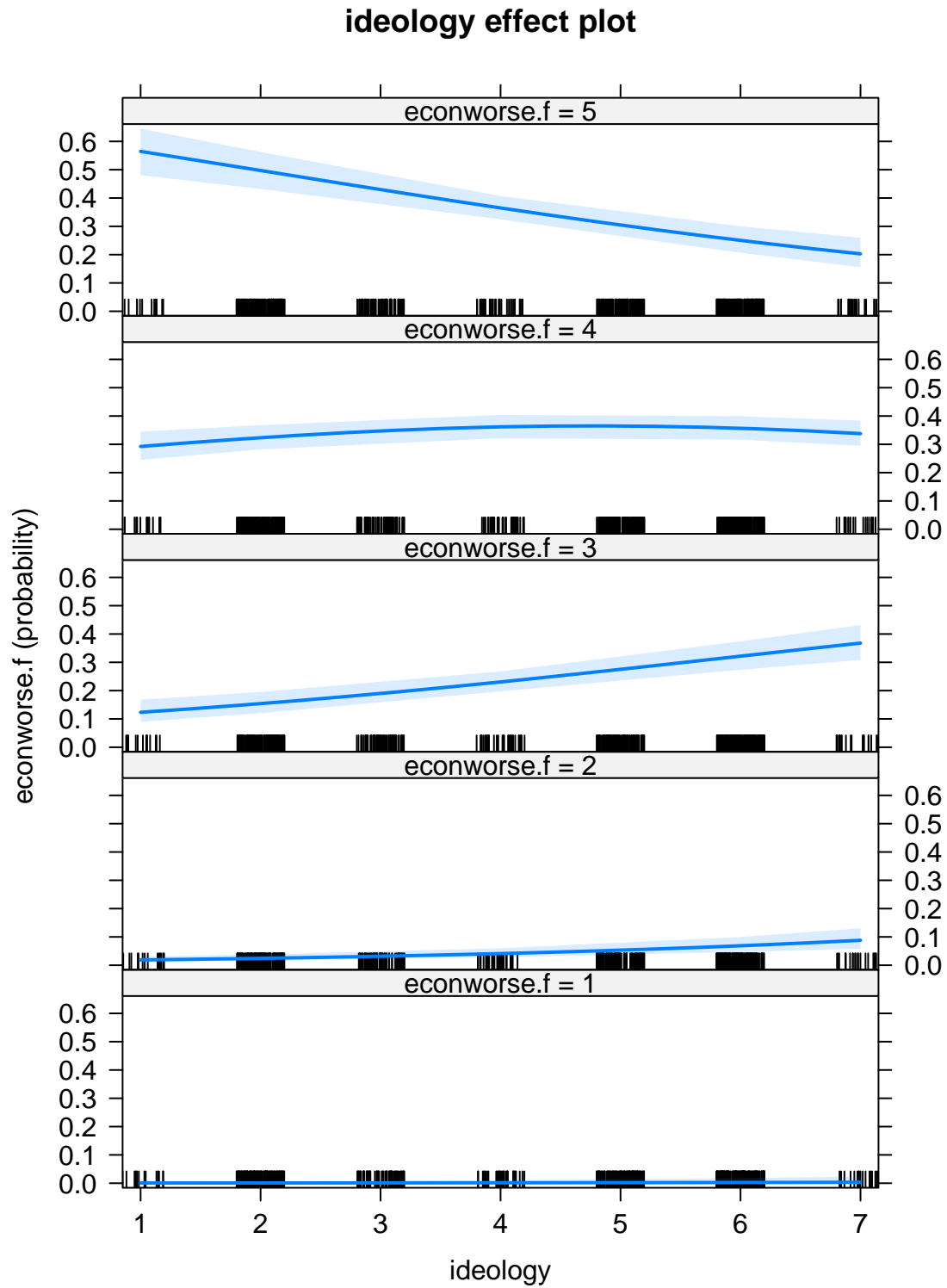
## at(ideology) at(educyears) ideology educyears faminc
##          7          2 0.0003049 0.000102 0.000002552

# 2
margins(o.logit, at = list(
  ideology = min(dat$ideology, na.rm = T), # izquierdista
  educyears = min(dat$educyears, na.rm = T)) # sin educ
)

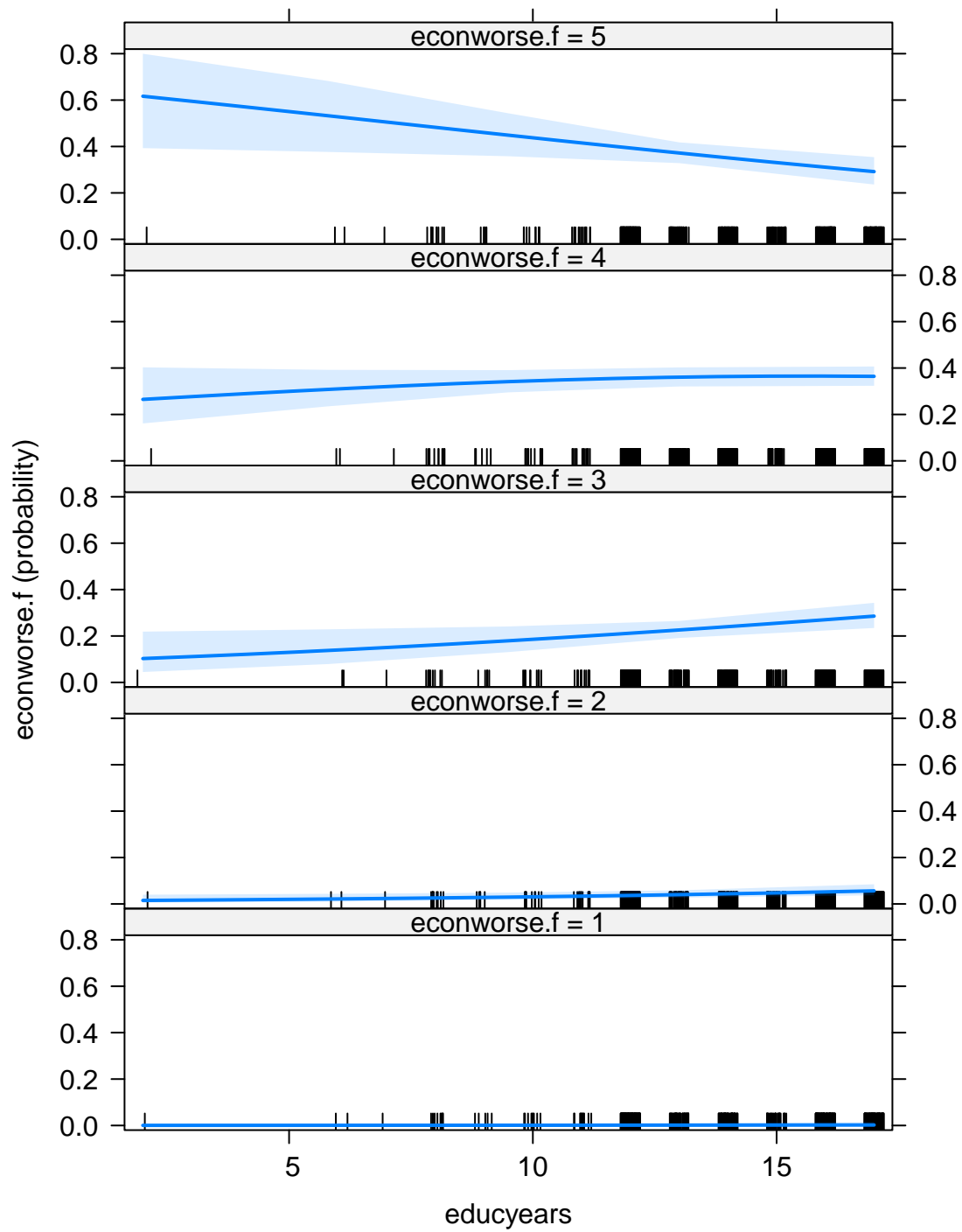
## at(ideology) at(educyears) ideology educyears faminc
##          1          2 0.00005992 0.00002004 0.0000005014
```

Predicted probabilities Calculemos ahora los predicted probabilities.

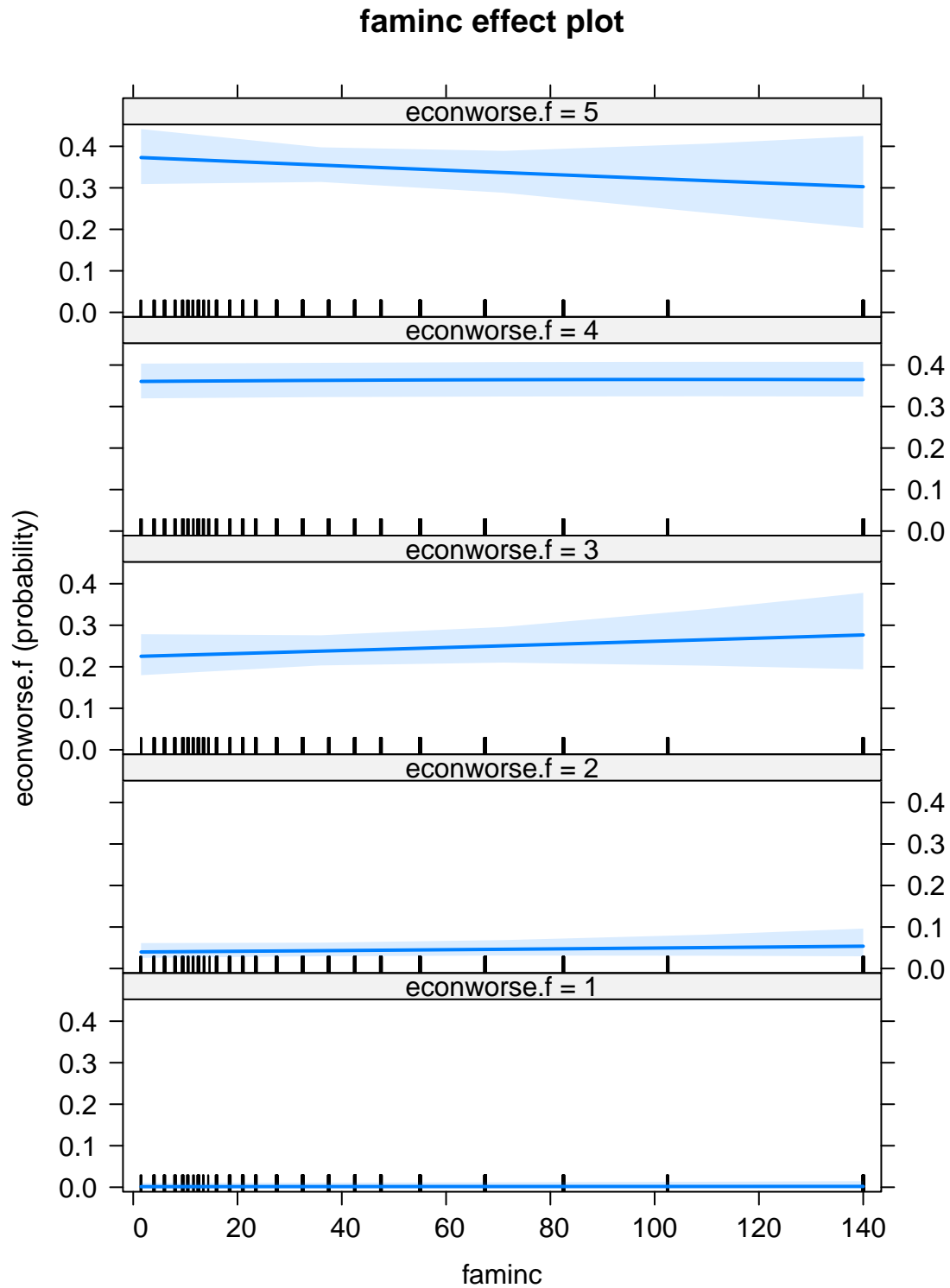
```
p_load(effects)
plot(effect("ideology", o.logit))
```



```
plot(effect("educyears", o.logit))
```

educyears effect plot

```
plot(effect("faminc", o.logit))
```



```
knitr::purl('Diagnosticos.Rnw')

## Error in file(con, "r"): no se puede abrir la conexi'on

Stangle('Diagnosticos.Rnw')

## Error in SweaveReadFile(file, syntax, encoding = encoding): no Sweave file with
name 'Diagnosticos.Rnw' found
```