Modelos de Variable Dependiente Cualitativa o Limitada

Práctica 3

José Burgos 25-0140

2025-08-16

Intrucciones Generales

Esta práctica tiene como objetivo aplicar diversos modelos econométricos cuando la variable dependiente es cualitativa o limitada. Utilizaremos bases de datos extraídas de los libros de texto:

Wooldridge, Jeffrey M. Introducción a la Econometría.

Hill, Carter; Griffiths, William; Lim, Guay. Introduction to Econometrics.

Cada ejercicio corresponde a un tipo distinto de modelo.

Ejercicio 1: Modelo de Respuesta Binaria (Probit o Logit)

Base de datos: attend.csv (Wooldridge)

- Variable dependiente: `attend = 1` si asistió a la universidad, 0 si no. Variables
- explicativas sugeridas: `faminc`, `parcoll`, `urban`, `income`.

Tareas:

1. Estimar un modelo Probit para `attend`.

```
attend <- read.csv("attend.csv")
mod_probit <- glm(attend ~ faminc + parcoll + urban + income,
                  data = attend, family = binomial(link = "probit"))
summary(mod_probit)$coefficients |>
  kable(digits = 3,
               col.names = c("Coeficiente", "Error Estándar", "Valor t", "Valor p"),
               caption = "Coeficientes del modelo Probit para attend")
```

Table 1: Coeficientes del modelo Probit para attend

	Coeficiente	Error Estándar	Valor t	Valor p
(Intercept)	-1.923	0.254	-7.585	0.000
faminc	0.015	0.006	2.687	0.007

	Coeficiente	Error Estándar	Valor t	Valor p
parcoll	0.800	0.123	6.524	0.000
urban	0.674	0.124	5.442	0.000
income	0.006	0.004	1.537	0.124

2. Interpretar los coeficientes.

Los coeficientes del modelo Probit indican cómo cada variable explicativa afecta la probabilidad de asistir a la universidad (`attend = 1`).

`faminc`: Por cada unidad de aumento en el ingreso familiar (`faminc`), la probabilidad de asistir a la universidad aumenta en aproximadamente 0.015, manteniendo las demás variables constantes.

`parcoll`: Si al menos uno de los padres asistió a la universidad (`parcoll = 1`), la probabilidad de que el individuo asista a la universidad aumenta en aproximadamente 0.800, en comparación con aquellos cuyos padres no asistieron.

`urban`: Vivir en un área urbana (`urban = 1`) aumenta la probabilidad de asistir a la universidad en aproximadamente 0.674, en comparación con aquellos que viven en áreas rurales.

`income`: Por cada unidad de aumento en el ingreso personal (`income`), la probabilidad de asistir a la universidad aumenta en aproximadamente 0.006, manteniendo las demás variables constantes.

3. Calcular los efectos marginales de faminc y parcoll.

```
margins_probit <- margins(mod_probit, variables = c("faminc", "parcoll"))
summary(margins_probit) |>
  kable(digits = 3, caption = "Efectos marginales")
```

Table 2: Efectos marginales

factor	AME	SE	z	p	lower	upper
faminc	0.005	0.002	2.744	0.006	0.001	0.009
parcoll	0.269	0.036	7.521	0.000	0.199	0.339

Ingreso familiar (`faminc`): un incremento de una unidad en el ingreso familiar aumenta la probabilidad del evento en 0.5 puntos porcentuales (`p = 0.006`), efecto pequeño pero estadísticamente significativo al 1 %.

Padres con estudios universitarios (`parcoll`): tener padres con educación universitaria incrementa la probabilidad del evento en 26.9 puntos porcentuales (`p < 0.001`), efecto grande y altamente significativo, lo que sugiere un impacto sustancial de la educación parental sobre la probabilidad analizada.

Ejercicio 2: Modelo Logit Multinomial

Base de datos: travelmode (Hill et al.)

Variable dependiente: mode (auto, bus, tren, avion)

Variables explicativas sugeridas: income, time, cost

Tareas:

1. Estimar un modelo logit multinomial, con automóvil como base.

```
# Convertir variable dependiente a factor
travelmode <- travelmode |>
  mutate(mode = factor(mode))

# modelo multinomial
modelo_multinomial <- multinom(
  mode ~ income + wait + gcost, data = travelmode, trace = FALSE)

summary(modelo_multinomial)$coefficients |>
  kable(digits = 3,
              col.names = c("Coeficiente", "Error Estándar", "Valor z", "Valor p"),
              caption = "Coeficientes del modelo Logit Multinomial")
```

Table 3: Coeficientes del modelo Logit Multinomial

	Coeficiente	Error Estándar	Valor z	Valor p
train	5.065	-0.007	-0.145	0.018
bus	4.587	-0.006	-0.105	0.010
car	-4.797	0.456	-14.957	0.352

2. Interpretar los coeficientes.

En el modelo logit multinomial, tomando al automóvil como la categoría base, los resultados muestran que un mayor nivel de ingreso aumenta de manera significativa la probabilidad relativa de elegir tanto el tren como el bus frente al automóvil, mientras que en el caso del avión, aunque el coeficiente estimado es negativo, no resulta estadísticamente significativo. En síntesis, el ingreso se relaciona positivamente con la elección de transporte alternativo al automóvil, excepto en el caso del avión, donde no se observa evidencia clara de efecto.

3. Calcular probabilidades predichas para un individuo con ingreso medio.

```
# Calcular probabilidades predichas
inc_mean <- mean(travelmode$income, na.rm = TRUE)
newdata  <- travelmode
newdata$income <- inc_mean

probs_mat <- predict(modelo_multinomial, newdata = newdata, type = "probs")

colMeans(as.matrix(probs_mat)) |>
  kable(
    caption = "Probabilidades predichas \n para un individuo con ingreso medio",
    digits = 3)
```

Table 4: Probabilidades predichas para un individuo con ingreso medio

	x
air	0.249
train	0.250
bus	0.250
car	0.251

Ejercicio 3: Modelo Logit Ordenado

Base de datos: jobsat.dta (`Wooldridge`)

Variable dependiente: satisfact = 1 (`baja`), 2 (`media`), 3 (`alta satisfacción`)

Variables explicativas sugeridas: age, educ, tenure, union Tareas:

```
jobsat <- read.csv("jobsat.csv") |> # 2.  Convertir satisfact en factor ordenado
  mutate(
    satisfact = factor(satisfact, levels = c(1, 2, 3),
                       ordered = TRUE, labels = c("Baja", "Media", "Alta")))
```

1. Estimar un modelo logit ordenado.

```
mod_logit_ordenado <- polr(satisfact ~ age + educ + tenure + union,
                   data = jobsat,
                   method = "logistic",
                   Hess = TRUE)

summary(mod_logit_ordenado)$coefficients |>
  kable(digits = 3,
               col.names = c("Coeficiente", "Error Estándar", "Valor z", "Valor p"),
               caption = "Coeficientes del modelo Logit Ordenado")
```

Table 5: Coeficientes del modelo Logit Ordenado

Coeficiente	Error Estándar	Valor z	Valor p
age	0.074	0.010	7.551
educ	0.386	0.048	8.034
tenure	0.142	0.021	6.858
union	0.884	0.202	4.367
Baja|Media	8.790	0.877	10.026
Media|Alta	10.602	0.919	11.531

2. Interpretar el efecto de union.

El coeficiente estimado para la variable union es 0.884 ($p < 0.01$), lo que indica que pertenecer a un sindicato incrementa significativamente la probabilidad de ubicarse en una categoría más alta de satisfacción laboral. En modelos logit ordenados, los coeficientes se interpretan como un cambio en la utilidad latente: un valor positivo implica que la pertenencia sindical desplaza la distribución hacia niveles más altos de satisfacción laboral.

3. Calcular probabilidades predichas por nivel de educación.

```r
# Calcular probabilidades predichas
newdata <- data.frame(
  age = mean(jobsat$age, na.rm = TRUE),
  educ = c(8, 12, 16),  # Ejemplo: baja, media y alta educación
  tenure = mean(jobsat$tenure, na.rm = TRUE),
  union = 0
)

predict(mod_logit_ordenado, newdata, type = "probs") |>
  kable(digits = 3,
               col.names = c("Baja", "Media", "Alta"),
               caption = "Probabilidades predichas por nivel de educación")
```

Table 6: Probabilidades predichas por nivel de educación

Baja	Media	Alta
0.832	0.136	0.032
0.513	0.353	0.134
0.183	0.395	0.421

Los resultados de las probabilidades predichas muestran que, a medida que aumenta el nivel educativo, la probabilidad de ubicarse en la categoría de baja satisfacción laboral disminuye considerablemente, mientras que crecen las probabilidades de estar en niveles medios y, especialmente, altos de satisfacción. En particular, con educación alta la mayor probabilidad corresponde a la categoría de alta satisfacción, lo que confirma el efecto positivo de la educación sobre la satisfacción laboral.

Ejercicio 4: Modelo Tobit (Censura)

Base de datos: healthexp.dta (Hill et al.)

Variable dependiente: exph (gasto en salud, puede ser cero)

Variables explicativas sugeridas: age, income, insured, chronic

Tareas:

1. Estimar un modelo Tobit.

```
healthexp <- read.csv("healthexp.csv")
mod_tobit <- tobit(exph ~ age + income + insured + chronic,
                   left = 0, data = healthexp)

stargazer(mod_tobit, type = "text", digits = 3,
          title = "Modelo Tobit para gasto en salud (exph)")
```

```
##
## Modelo Tobit para gasto en salud (exph)
## ========================================
##                    Dependent variable:
##                ----------------------------
##                           exph
## ----------------------------------------
## age                       21.108
##                          (14.034)
##
## income                    -0.010
##                           (0.021)
##
## insured                 1,043.982**
##                          (452.646)
##
## chronic                   676.775
##                          (525.189)
##
## Constant                  28.350
##                          (982.752)
##
## ----------------------------------------
## Observations                500
## Log Likelihood          -3,235.766
## Wald Test            9.726** (df = 4)
## ========================================
## Note:          *p<0.1; **p<0.05; ***p<0.01
```

2. Comparar resultados con OLS.

```
mod_ols_tobit<- lm(exph ~ age + income + insured + chronic,
              data = healthexp)
stargazer(mod_tobit, mod_ols_tobit, type = "text", digits = 3,
          title = "Comparación entre Modelo Tobit y OLS")
```

```
##
## Comparación entre Modelo Tobit y OLS
## ===============================================================
##                                 Dependent variable:
##                     ------------------------------------------
##                                         exph
##                             Tobit                 OLS
##                              (1)                  (2)
## ---------------------------------------------------------------
## age                         21.108               10.939
##                            (14.034)             (9.222)
##
## income                      -0.010               -0.009
##                            (0.021)              (0.014)
##
## insured                  1,043.982**           591.267**
##                           (452.646)            (293.537)
##
## chronic                    676.775              412.913
##                           (525.189)            (347.197)
##
## Constant                   28.350             1,966.409***
##                           (982.752)            (639.151)
##
## ---------------------------------------------------------------
## Observations                500                   500
## R2                                               0.015
## Adjusted R2                                      0.007
## Log Likelihood           -3,235.766
## Residual Std. Error                     3,101.442 (df = 495)
## F Statistic                             1.921 (df = 4; 495)
## Wald Test             9.726** (df = 4)
## ===============================================================
## Note:                             *p<0.1; **p<0.05; ***p<0.01
```

En el modelo Tobit con censura en cero, los resultados muestran que únicamente la variable `insured` *presenta un efecto estadísticamente significativo sobre el gasto en salud, con un coeficiente estimado cercano a 1,044 ($p < 0.05$). Esto sugiere que contar con seguro incrementa de manera sustancial el gasto esperado en salud, incluso considerando la presencia de observaciones con gasto nulo. En contraste, las variables* `age`, `income` *y* `chronic` *no exhiben efectos significativos en este modelo. Además, la prueba de Wald ($\chi^2 = 9.73$; $p < 0.05$) confirma una significancia conjunta moderada de los regresores.*

Por su parte, la estimación mediante Mínimos Cuadrados Ordinarios (OLS) también reporta un efecto positivo de `insured`*, aunque de menor magnitud (aproximadamente 591), y mantiene la no significancia del resto de las variables. Esta comparación evidencia que, al omitir la naturaleza censurada de los datos, el modelo OLS tiende a subestimar el verdadero efecto de estar asegurado sobre el gasto en salud, reforzando la pertinencia del uso del modelo Tobit en este contexto.*

3. Calcular efecto marginal esperado del ingreso.

```r
# --- AME de income en Tobit (censura a la izquierda en 0) ---
# Extraer beta y sigma
beta <- coef(mod_tobit)
beta_inc <- beta["income"]

# Escala (sigma) del modelo tobit
sigma <- if (!is.null(mod_tobit$scale)) mod_tobit$scale else summary(mod_tobit)$scale

# x por observación
X  <- model.matrix(mod_tobit)          # incluye el intercepto
xb <- as.numeric(X %*% beta)

# z y Phi(z)
z   <- xb / sigma
Phi <- pnorm(z)

# Efecto marginal individual y promedio (AME)
me_income_i <- Phi * beta_inc
AME_income  <- mean(me_income_i, na.rm = TRUE)

AME_income
```

```
## [1] -0.006195267
```

El efecto marginal promedio estimado para la variable income en el modelo Tobit resulta ser de aproximadamente –0.006. Esto implica que, en promedio, un aumento de una unidad en el ingreso se asocia con una reducción de 0.006 unidades monetarias en el gasto esperado en salud, considerando tanto a los individuos con gasto positivo como a aquellos con gasto censurado en cero. Dado que el valor es muy pequeño y cercano a cero, la evidencia empírica sugiere que el ingreso no ejerce un efecto económicamente relevante sobre el gasto en salud en esta muestra.

Ejercicio 5: Modelo de Selección Muestral (Heckman)

Base de datos: mroz.dta (Wooldridge)

Variable dependiente: `lwage` (log salario horario), sólo si `inlf = 1`

Variables explicativas (participación): `age`, `kidslt6`, `kidsge6`, `educ`, `nwifeinc`

Variables explicativas (salario): `educ`, `exper`, `expersq`, `city`

Tareas:

1. Estimar el modelo de Heckman de dos etapas.

```
data("mroz")

mod_heckman <- heckit(
  selection = inlf ~ age + kidslt6 + kidsge6 + educ + nwifeinc,
  outcome   = lwage ~ educ + exper + I(exper^2) + city,
  data = mroz,
  method = "2step"
)

summary(mod_heckman)
```

```
## --------------------------------------------
## Tobit 2 model (sample selection model)
## 2-step Heckman / heckit estimation
## 753 observations (325 censored and 428 observed)
## 14 free parameters (df = 740)
## Probit selection equation:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.422400   0.472640   0.894    0.372
## age         -0.034439   0.007577  -4.545 6.41e-06 ***
## kidslt6     -0.892175   0.114432  -7.797 2.16e-14 ***
## kidsge6     -0.037700   0.040432  -0.932    0.351
## educ         0.155838   0.023900   6.520 1.30e-10 ***
## nwifeinc    -0.020923   0.004583  -4.565 5.84e-06 ***
## Outcome equation:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.6105375  0.2761219  -2.211  0.02733 *
## educ         0.1092879  0.0166707   6.556 1.04e-10 ***
## exper        0.0416296  0.0131874   3.157  0.00166 **
## I(exper^2)  -0.0008153  0.0003938  -2.071  0.03874 *
## city         0.0497915  0.0685207   0.727  0.46766
## Multiple R-Squared:0.1584,   Adjusted R-Squared:0.1484
##    Error terms:
##               Estimate Std. Error t value Pr(>|t|)
## invMillsRatio  0.05625    0.13583   0.414    0.679
## sigma          0.66395         NA      NA       NA
## rho            0.08472         NA      NA       NA
## --------------------------------------------
```

2. Comparar con estimación `OLS` directa.

```
mod_ols <- lm(lwage ~ educ + exper + I(exper^2) + city,
              data = subset(mroz, inlf == 1))

stargazer(mod_heckman, mod_ols, type = "text", digits = 3,
          title = "Comparación entre Heckman y OLS",
          dep.var.labels = c("Heckman: Log Salario", "OLS: Log Salario"))
```

```
##
## Comparación entre Heckman y OLS
## ===========================================================
##                                 Dependent variable:
##                     ------------------------------------
##                                Heckman: Log Salario
##                           Heckman              OLS
##                          selection
##                             (1)                (2)
## -----------------------------------------------------------
## educ                     0.109***           0.106***
##                           (0.017)            (0.014)
##
## exper                    0.042***           0.041***
##                           (0.013)            (0.013)
##
## I(exper2)                -0.001**           -0.001**
##                          (0.0004)           (0.0004)
##
## city                      0.050              0.054
##                           (0.069)            (0.068)
##
## Constant                 -0.611**           -0.531***
##                           (0.276)            (0.199)
##
## -----------------------------------------------------------
## Observations               753                428
## R2                        0.158              0.158
## Adjusted R2               0.148              0.150
## rho                       0.085
## Inverse Mills Ratio 0.056 (0.136)
## Residual Std. Error                   0.667 (df = 423)
## F Statistic                       19.856*** (df = 4; 423)
## ===========================================================
## Note:                           *p<0.1; **p<0.05; ***p<0.01
```

3. Discutir la existencia de sesgo de selección.

La comparación entre la estimación mediante el modelo de Heckman y la estimación por OLS directo muestra que los coeficientes de las variables explicativas principales (`educ`*,* `exper` *e* `exper`^{2}*) son consistentes en signo, magnitud y significancia estadística. En ambos modelos, la educación mantiene un efecto positivo y significativo sobre el logaritmo del salario (0.11), al igual que la experiencia, mientras que la variable cuadrática de experiencia refleja rendimientos decrecientes.*

La diferencia fundamental radica en el término de selección, conocido como el Inverse Mills Ratio (λ), incluido en la corrección de Heckman. En este caso, el coeficiente estimado para λ no resulta estadísticamente significativo ($\rho \approx 0.085,\ p > 0.1$), lo cual sugiere que no existe evidencia sólida de sesgo de selección en la muestra analizada.

En consecuencia, los resultados indican que, aunque el modelo de Heckman es más general y permite corregir potenciales problemas de autoselección en la participación laboral, en este caso específico las estimaciones obtenidas mediante OLS no presentan un sesgo sistemático relevante.

Ejercicio 6: Modelo de Conteo (Poisson / Binomial Negativa)

Base de datos: `health.csv` (Wooldridge)

Variable dependiente: `numvisit` (número de visitas al médico)

Variables explicativas sugeridas: `age`, `income`, `educ`, `insured`, `health`

Tareas:

1. Estimar un modelo Poisson

```
health <- read.csv("health.csv")

# 3. Modelo Poisson
mod_pois <- glm(numvisit ~ age + income + educ + insured + health,
                family = poisson(link = "log"),
                data = health)

stargazer(mod_pois, type = "text",
          title = "Modelo Poisson - Visitas al médico")
```

```
##
## Modelo Poisson - Visitas al médico
## ================================================
##                         Dependent variable:
##                     ---------------------------
##                               numvisit
## ------------------------------------------------
## age                            0.016***
##                                 (0.003)
##
## income                        -0.00002***
##                                 (0.00001)
##
## educ                           -0.099***
##                                 (0.028)
##
## insured                        -0.419***
##                                 (0.106)
##
## health                         -0.263***
##                                 (0.043)
##
## Constant                       1.917***
##                                 (0.471)
##
## ------------------------------------------------
## Observations                      500
## Log Likelihood                 -551.562
## Akaike Inf. Crit.              1,115.124
## ================================================
## Note:                *p<0.1; **p<0.05; ***p<0.01
```

2. Interpretar el efecto de insured y calcular el efecto marginal. 3

```
marg_insured <- margins(mod_pois, variables = "insured")
summary(marg_insured) |>
  kable()
```

factor	AME	SE	z	p	lower	upper
insured	-0.3227605	0.0834412	-3.868118	0.0001097	-0.4863023	-0.1592187

Interpretación del efecto de `insured`

En el modelo de conteo Poisson estimado, el coeficiente de la variable `insured` *es* -0.419 *y resulta altamente significativo* ($p < 0.01$). *Dado que los coeficientes en el modelo Poisson se interpretan en términos log-lineales, este valor indica que, manteniendo las demás variables constantes, estar asegurado se asocia con una reducción en el número esperado de visitas médicas. En términos relativos, el efecto puede expresarse como*:

$$\%\Delta E[\text{numvisit}] \approx \left(e^{-0.419} - 1\right) \times 100 \approx -34.2\%$$

Es decir, las personas aseguradas presentan, en promedio, un 34% menos visitas al médico respecto a aquellas no aseguradas.

Efecto marginal esperado

El cálculo del efecto marginal arroja un valor de $\boldsymbol{AME = -0.323}$ ($p < 0.001$). *Esto significa que, en promedio, estar asegurado reduce en aproximadamente 0.32 el número esperado de visitas médicas por individuo, manteniendo constantes las demás variables.*