MIKROEKONOMETRIA
ZAJĘCIA 6

WIKTOR
BUDZIŃSKI

ZMIENNE BINARNE

W wielu przypadkach zmienna objaśniana może przyjmować tylko dwie wartości

- Zazwyczaj oznaczane przez 0 i 1
- W takim przypadku estymator MNK nie będzie zgodny

Dla takich danych łatwo można jednak skonstruować estymator **Metody Największej Wiarygodności**

- Dla przypomnienia: w MNW musimy jakoś zdefiniować prawdopodobieństwo zaobserwowania zmiennej objaśnianej jako funkcję parametrów

ZADANIE 1

Skonstruuj estymator MNW dla rzutu monetą

1. Oznaczmy reszkę jako 1, a orła jako 0
2. Jako parametr rozkładu przyjmij prawdopodobieństwo „sukcesu”
3. „Zaimplementuj” estymator w STATA’cie
4. Przetestuj hipotezę czy moneta jest „uczciwa”

Modele binarne

Model dla zmiennej binarnej nie różni się specjalnie od tego dla rzutu monetą

- Różnica polega przede wszystkim na tym, że chcemy sformułować prawdopodobieństwo sukcesu jako funkcję jakichś zmiennych objaśniających

Najczęściej stosuje się tzw. model funkcji wskaźnikowej

$$y^* = \boldsymbol{\beta}'\mathbf{X} + \varepsilon$$

Gdzie obserwujemy tylko

$$y = \begin{cases} 1 & jeśli \quad y^* > 0 \\ 0 & jeśli \quad y^* \leq 0 \end{cases}$$

Modele binarne

Dla tak zdefiniowanego modelu możemy zapisać prawdopodobieństwo sukcesu jako

$$\begin{aligned}\Pr\left[y=1|\mathbf{X}\right] &= \Pr\left[y^{*}>0\right] \\ &= \Pr\left[\boldsymbol{\beta}'\mathbf{X}+\varepsilon>0\right] \\ &= \Pr\left[-\varepsilon<\boldsymbol{\beta}'\mathbf{X}\right] \\ &= F\left(\boldsymbol{\beta}'\mathbf{X}\right)\end{aligned}$$

Forma funkcyjna modelu będzie zależała od tego jaki rozkład przyjmiemy dla błędu losowego

NORMALIZACJA SKŁADNIKA LOSOWEGO

W przeciwieństwie do regresji liniowej, w modelu dla zmiennej binarnej nie da się zidentyfikować wariancji składnika losowego

- Intuicyjnie, wynika to z tego, że funkcja wskaźnikowa jest nieobserwowalna, a jej skala nie ma żadnej interpretacji
- Prostym rozwiązaniem jest po prostu przyjęcie, że ta wariancja równa się 1
- Sprawia to, że parametry modelu nie mają niestety ilościowej interpretacji

LOGIT VS. PROBIT

Zdecydowanie najczęściej przyjmowanymi rozkładami błędów losowych są

1. Logistyczny
2. Normalny

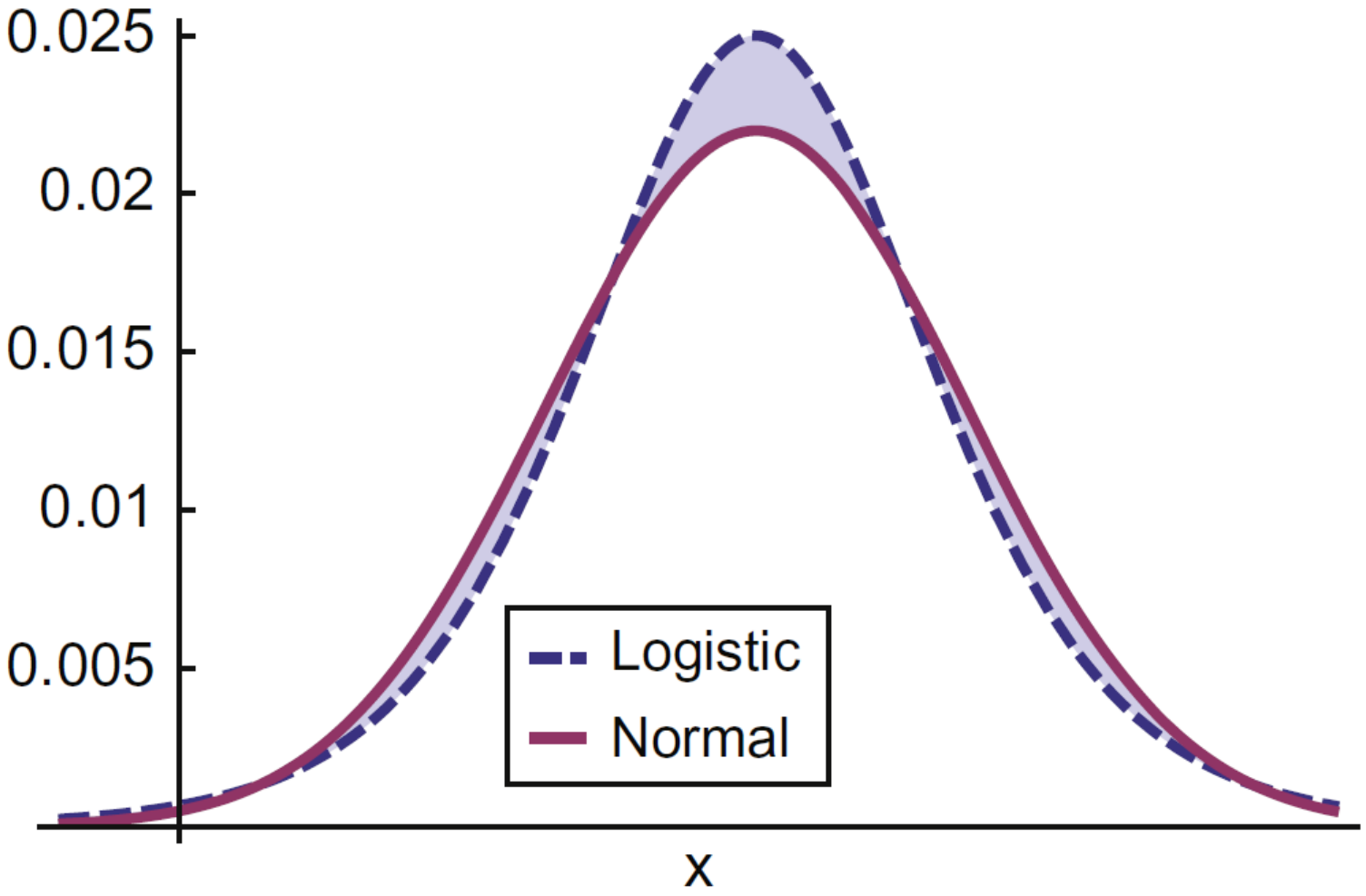
Prowadzi to odpowiednio do modelu logitowego i probitowego

W logicie dystrybuanta jest dana przez

$$\Lambda(x) = \frac{\exp(x)}{1+\exp(x)}$$

W probicie nie ma ona analitycznej formy funkcyjnej

- Po prostu dystrybuanta rozkładu normalnego


0.025
0.02
0.015
0.01
0.005
Logistic
Normal
x

LOGIT VS. PROBIT

W praktyce oba modele będą dawać bardzo zbliżone wyniki

- Główną zaletą logitu jest prosty wzór na dystrybuantę
- Wielkości parametrów mogą się różnić, ale nie ma to znaczenia
  - I tak nie mają one absolutnej interpretacji
  - Efekty krańcowe powinny być do siebie zbliżone

ALTERNATYWNE MODELE

- Inne rozkłady składnika losowego, które nie zakładają symetryczności:
  - Gumbel, Gompit, Gompertz (ε ma rozkład wartości ekstremalnych)
  - *Complementary log log:* $CDF(x) = \exp(-\exp(-x))$
  - I inne (np. arctangens, burr (scobit), …)
  - Wyniki mogą się czasem znacznie różnić od wyników logitu i probitu
  - Przydatne, gdy jeden z wyników jest rzadki
    - Np. mały odsetek 'tak' w próbie

FUNKCJA NAJWIĘKSZEJ WIARYGODNOŚCI

- Funkcja prawdopodobieństwa masy

$$P(Y_i \mid \mathbf{X}_i) = F(\boldsymbol{\beta}'\mathbf{X}_i)^{Y_i} \left(1 - F(\boldsymbol{\beta}'\mathbf{X}_i)\right)^{1-Y_i}$$

- Obserwacje są niezależne. Jaka jest szansa, że dostaniemy taki zbiór obserwacji, jaki mamy?

$$L(\boldsymbol{\beta} \mid \text{dane}) = \prod_{i=1}^{N} F(\boldsymbol{\beta}'\mathbf{X}_i)^{Y_i} \left(1 - F(\boldsymbol{\beta}'\mathbf{X}_i)\right)^{1-Y_i}$$

  - Funkcja wiarygodności (ang. *likelihood function*)
- Chcemy znaleźć takie β, żeby zmaksymalizować tę funkcję
  - Czyli, żeby nasz model miał takie parametry, żeby przewidywał takie prawdopodobieństwa, które dają największą szansę takiego wyniku, jaki obserwujemy

EFEKTY KRAŃCOWE

- Parametry w modelach dla zmiennych binarnych mają tylko interpretacje jakościową
- Aby uzyskać oszacowania ilościowe należy policzyć efekty krańcowe:

$$\frac{\partial E(\mathbf{Y}|\mathbf{X})}{\partial \mathbf{X}} = \frac{\partial F(\mathbf{X}\boldsymbol{\beta})}{\partial \mathbf{X}} = \frac{dF(\mathbf{X}\boldsymbol{\beta})}{d(\mathbf{X}\boldsymbol{\beta})}\boldsymbol{\beta} = f(\mathbf{X}\boldsymbol{\beta})\boldsymbol{\beta}$$

- Wzrost X o jednostkę średnio zwiększa p-stwo sukcesu o…

ZADANIE 2

1. Wczytaj projekt *me_apples*.dta
2. Oszacuj model, w którym gotowość do zakupu dodatnich ilości eko-jabłek wyjaśniana jest przez:
   - Poziom edukacji respondenta
   - Cenę zwykłych jabłek
   - Cenę eko-jabłek
   - To, czy wybór dokonywany jest w sezonie
   - To czy respondent był mężczyzną
   - Dochód rodziny
   - Liczbę członków rodziny w każdej kategorii wiekowej
3. Co dzieje się z modelem, jeśli dodatkowo uwzględnić wielkość rodziny?
4. Policz efekty krańcowe

'DOPASOWANIE' MODELU

- Po estymacji modelu zwykle raportowane są:
  - Wartość funkcji LL w optimum (w punkcie konwergencji)
  - Wartość funkcji LL 'w 0' (model wykorzystujący tylko stałą)
    - Odpowiada założeniu, że wszystkie parametry w modelu nieistotne
- Nie ma miary R^2 z MNK
- Pewnym rozwiązaniem jest Pseudo-R^2 (kilka wersji)
  - Miary oparte na wartości funkcji LL (wiele wersji):
    - $$\text{McFadden's pseudo-}R^2 = \frac{1-\ln L(\boldsymbol{\beta})}{\ln L(\mathbf{Y})} = 1-\ln L/\ln L_0$$
    - Im model lepiej dopasowany tym R2 bliższe 1
      - Ale przedział między 0 a 1 nie ma interpretacji naturalnej
        - Nie wiemy, gdzie jest zero
        - Z powodu istnienia składnika losowego model nie osiąga 1

'DOPASOWANIE' MODELU

- Kryterium informacyjne Akaike (AIC)

$$AIC = 2k - 2\ln L(\boldsymbol{\beta})$$

- AIC skorygowane (dla skończonej próby)

$$AICc = AIC + \frac{2k(k+1)}{N-k-1}$$

- Bayesowskie kryterium informacyjne (BIC, Schwarza)

$$BIC = k\ln N - 2\ln L(\boldsymbol{\beta})$$

- Jak w modelu liniowym – nie jest to kryterium 'statystycznego' porównania, ale często używane
  - Mówi, który model lepiej pasuje do danych
  - Nie mówi czy dobrze, o ile lepiej i czy istotnie lepiej
  - W porównaniu z pseudo-R^2 przynajmniej bierze pod uwagę liczbę zmiennych objaśniających

'DOPASOWANIE' MODELU

- Miary oparte na przewidywaniu wyników
  - Tabela poprawnych i niepoprawnych predykcji
    - Błędy I i II typu – trade-off
    - Ile % poprawnych predykcji ma naiwna reguła $Y_i = 1$?
- Miary oparte na przewidywaniu prawdopodobieństw
  - BenAkiva i Lerman's pseudo-$R^2 = \frac{1}{N}\sum_{i=1}^{N}\left(Y_i\hat{F}\left(\mathbf{X}_i\boldsymbol{\beta}\right)+\left(1-Y_i\right)\left(1-\hat{F}\left(\mathbf{X}_i\boldsymbol{\beta}\right)\right)\right)$
    - Średnie prawdopodobieństwo poprawnych predykcji
    - Problem gdy obserwacje w próbie są niesymetryczne

ZADANIE 2 c.d.

5. Porównaj oszacowany model z modelem probitowym
   - Porównaj dopasowanie wykorzystując różne miary

ZAŁOŻENIA

- Pominięte zmienne
  - Jeśli zmienne są nieskorelowane z tymi, które obserwujemy to estymator będzie nieobciążony, ale zmieni się jego skala
  - Jeśli są skorelowane to wystąpi endogeniczność i nasz estymator będzie obciążony
- Heteroskedastyczność
  - Model dla zmiennej binarnej nie będzie zgodny
  - A to duży problem, bo dane mikro często są heteroskedastyczne
- Forma funkcyjna
  - Niepoprawna forma funkcyjna również prowadzi do obciążenia i niezgodności

TESTY

- Analogicznie jak w regresji liniowej można testować poprawność formy funkcyjnej
  - Test zamiast RESET nazywa się LINKTEST
  - Estymujemy model: $y^* = \mathbf{X\beta} + \varepsilon$
  - Liczymy wartości dopasowane: $\hat{y} = \mathbf{X\hat{\beta}}$
  - Liczymy regresję pomocniczą:

$$y^* = \alpha_0 + \alpha_1 \hat{y} + \alpha_2 \hat{y}^2 + \alpha_3 \hat{y}^3 + \ldots + \varepsilon$$

  - Testujemy łączną istotność $\alpha_2, \alpha_3, \ldots$

TESTY

- Heteroskedastyczność można analizować jedynie poprzez jej bezpośrednie modelowanie
  - Możemy wykorzystać probit z heteroskedastycznością
  - Liczymy model:

$$y_i^* = \mathbf{X}_i\boldsymbol{\beta} + \sigma_i\varepsilon$$

  - Wcześniej zakładaliśmy, że $\forall_i \quad \sigma_i = 1$
  - Teraz

$$\sigma_i = \exp\left(\mathbf{Z}_i\boldsymbol{\alpha}\right)$$

ZADANIE 2 c.d.

6. Sprawdź poprawność formy funkcyjnej analizowanego modelu logistycznego
   - Wykorzystaj polecenie wbudowane w STATĘ
   - Policz 'na piechotę'
7. Sprawdź czy w analizowanym modelu występuje dająca się objaśnić heteroskedastyczność.