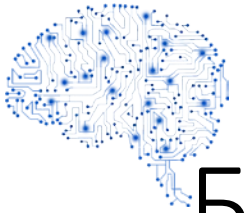
Моделі з бінарними залежними змінними

Професор, д.е.н. Ставицький А.В.

План

- Бінарні залежні змінні
- Probit/Logit моделі
- Оцінка Probit/Logit моделей
- Вивід Probit/Logit моделей

1. Бінарні залежні змінні



Бінарні залежні змінні

- Раніше залежні змінні могли приймати довільні значення
- Що робити, якщо цікавить результат рекламної компанії?
- Залежна змінна приймає лише два значення «успіх» та «невдача».

Приклад

- Припустимо, що необхідно передбачити результат футбольного матчу на основі котирувань букмекерів.

Приклад: модель – 1

- Можна розглянути модель:

$$D_i^{Win} = \beta_0 + \beta_1 Spread_i + \varepsilon_i$$

де i позначає номер гри

- Як інтерпретувати коефіцієнти регресії?
- Як робити прогноз?

Приклад: модель - 2

$$D_i^{Win} = \beta_0 + \beta_1 Spread_i + \varepsilon_i$$

- D_i^{Win} примає значення 0 або 1. Немає сенсу говорити, що зростання котирування на 1 збільшує залежну змінну на β_1, оскільки D_i^{Win} може змінюватися лише лише від 0 до 1 та від 1 до 0.
- Замість прогнозування D_i^{Win}, спрогнозуємо ймовірність того, що $D_i^{Win} = 1$.

Приклад: модель – 3

$$D_i^{Win} = \beta_0 + \beta_1 Spread_i + \varepsilon_i$$

- Тоді збільшення на 1 котирування збільшить ймовірність виграшу на β_1.
- Потім будемо передбачати значення D_i^{Win} на основі ймовірності виграшу.

Приклад: модель – 4

$$D_i^{Win} = \beta_0 + \beta_1 Spread_i + \varepsilon_i$$

- Якщо використовувати лінійну регресійну модель для оцінки ймовірностей, то можна назвати модель лінійна ймовірнісна модель (LPM).

Проблеми з LPM

- Похибки не розподілені нормально
- Похибки гетероскедастичні
- Прогнозні значення залежної змінної можуть бути за межами 0 та 1.

Що котирування показують щодо ймовірності виграшу?

Dependent Variable: WIN
Method: Least Squares
Sample: 1 644
Included observations: 644
White Heteroskedasticity-Consistent Standard Errors & Covariance

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	**0.500000**	**0.018849**	**26.52593**	**0.0000**
SPREAD	**−0.025180**	**0.003068**	**−8.26065**	**0.0000**

R-squared	0.087582	Mean dependent var	0.500000
Adjusted R-squared	0.086161	S.D. dependent var	0.500389
S.E. of regression	0.478346	Akaike info criterion	1.366137
Sum squared resid	146.8993	Schwarz criterion	1.380012
Log likelihood	−437.8960	*F*-statistic	61.62496
Durbin–Watson stat	2.034242	Prob(*F*-statistic)	0.000000

Характеристики моделі

- Розраховано робастні оцінки стандартних похибок у формі Уйта.
- Залишки гетероскедастичні.
- Гетероскедастичність є єдиним порушенням теореми Гауса-Маркова.

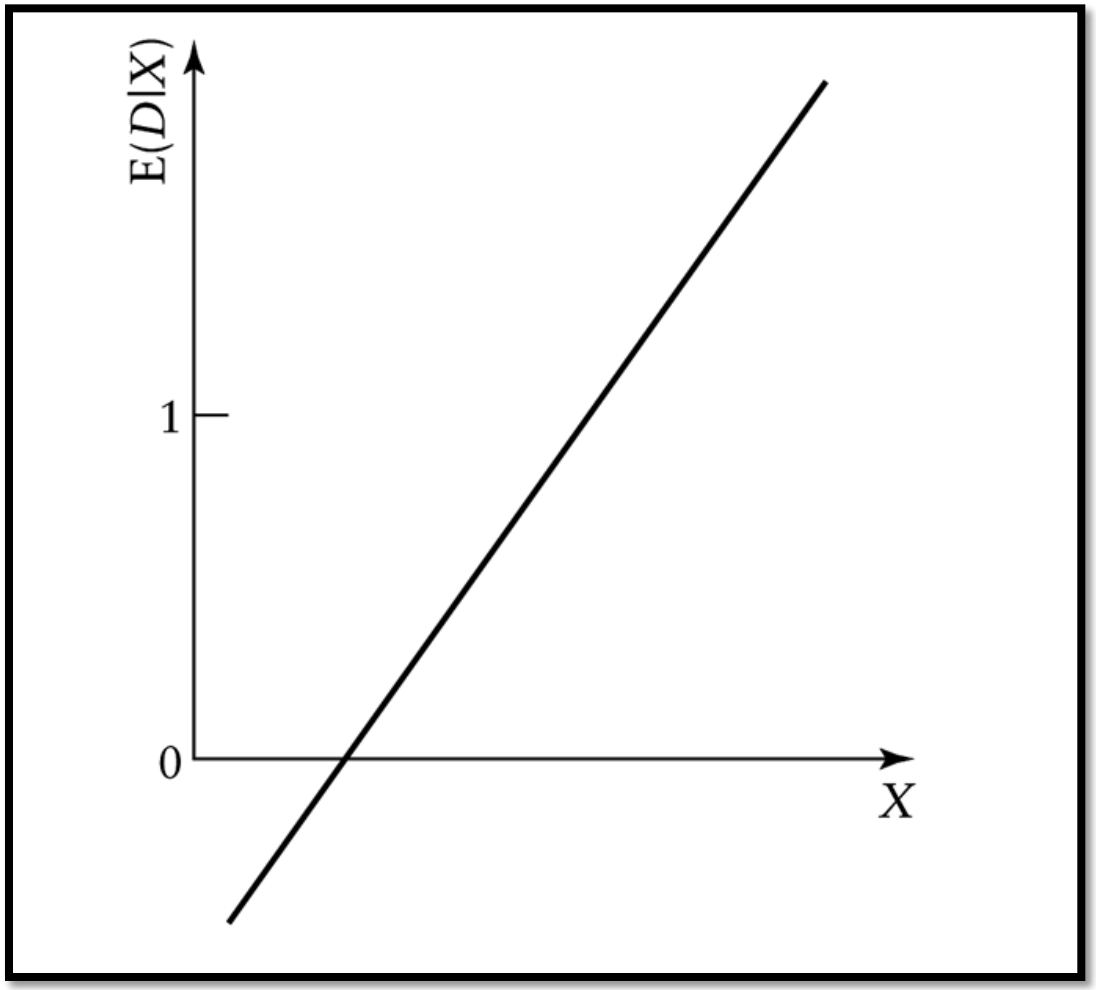
Аналіз моделі - 1

- Лінійна модель працює гарно математично.
- Існує серйозна проблема у інтепретації.
- Якщо котирування дорівнює 21, то ймовірність команди виграти: $0.5 - 0.025 \bullet 21 = -0.025 < 0$

Аналіз моделі – 2

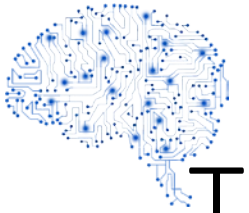
- Якщо X = 21, E(Y | X) = -0.025
- Це означає ймовірність у -2.5% виграшу.
- Якщо X = -21, ймовірність виграти 102.5%>100%.

Для деяких X $E(D|X_i) > 1$,
для деяких X $E(D|X_i) < 0$



Вимоги

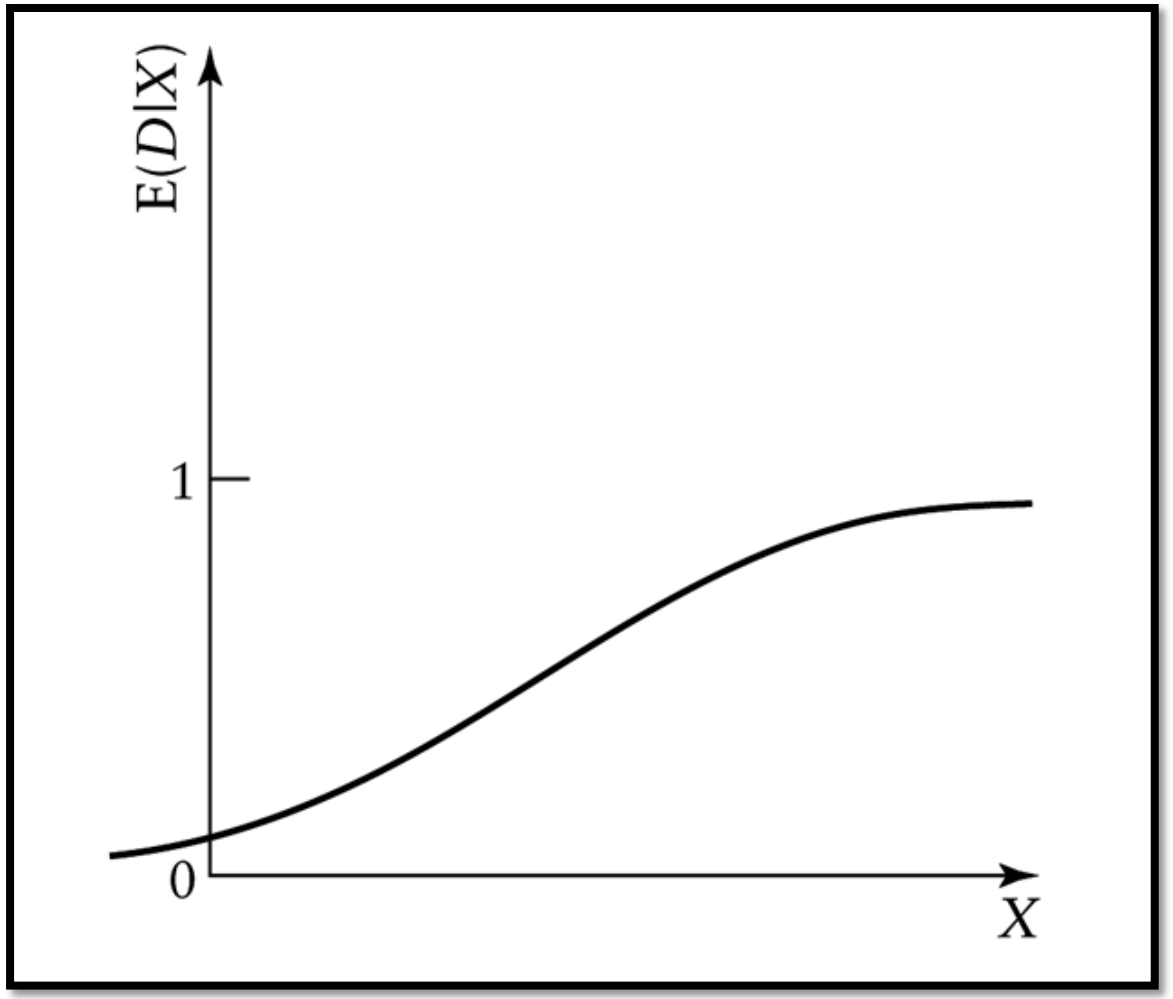
- Лінійна регресія теоретично дає прогнози від -∞ до +∞.
- Ймовірності мають бути між 0 та 1.
- Лінійна ймовірнісна модель не зможе гарантувати адекватні прогнози.



Транслятор

- Необхідно розробити транслятор, який:
  - При наближенні прогнозу до -∞ ймовірність має наближатися до 0.
  - При наближенні прогнозу до +∞ ймовірність має наближатися до 1.
  - Не існує ймовірностей менших 0 та більших 1.

Графік ймовірності перемоги



Питання

- Як побудувати такий транслятор?
- Як його оцінити?

2. PROBIT/LOGIT моделі

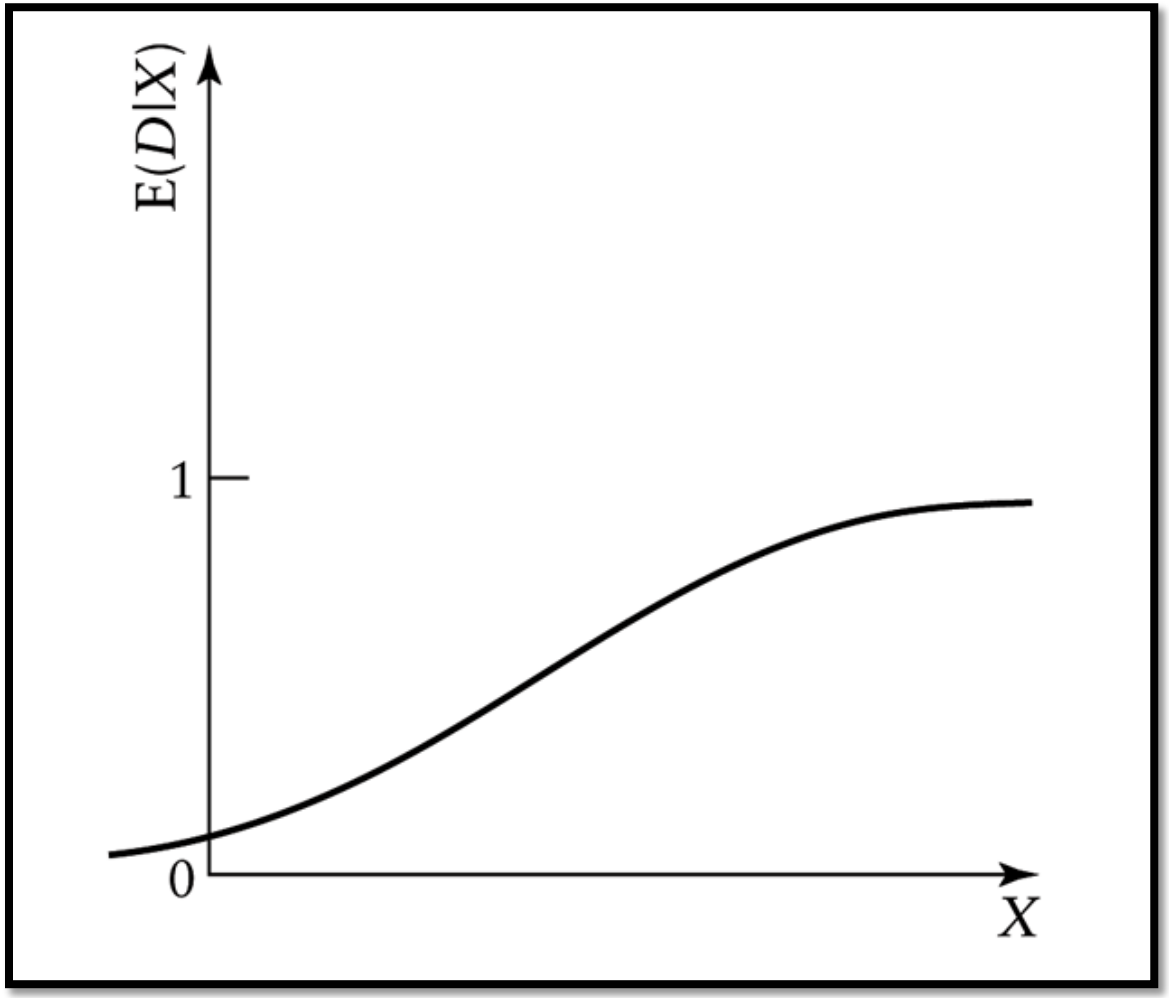
Probit/Logit моделі

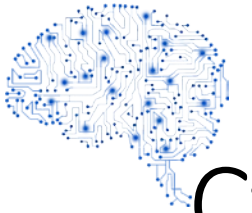
- На практиці використовуються ДВА таких транслятори:
  - probit
  - logit
- Різниці між ними дуже незначна.

Передумови

- Значення коефіцієнта регресії може змінювати ймовірність по-різному.
- Якщо якась команда дуже-дуже ймовірно виграє або дуже-дуже ймовірно програє, маленька зміна котирування має незначний вплив.
- Якщо команда має шанси виграти 50/50, то незначна зміна котирувань може призвести до значних змін у ймовірності.

Графік ймовірності перемоги





Стуктура Probit/Logit моделей

- У Probit і Logit моделей однакова структура.
  - Розрахувати допоміжну змінну Z за допомогою лінійної регресії. Z може приймати значення від -∞ до +∞.
  - Використати нелінійну трансформацію Z в ймовірність між 0 and 1.
  - Чим вище значення E(Z), тим більш ймовірною є перемога команди.

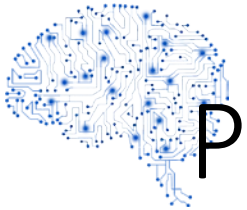
Допоміжна змінна

- Допоміжна змінна Z – це лінійна функція незалежних змінних:

$$Z = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \ldots + \beta_K X_{Ki} + \varepsilon_i$$

- Наша мета – оцінити ці коефіцієнти β_i.
- Ще більш важливим оцінити E(Z):

$$E(Z) = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \ldots + \beta_K X_{Ki}$$



Різниця між Probit/Logit моделями

- Ймовірність Y – це нелінійна функція від E(Z).
  - probit модель використовує кумулятивну функцію щільності стандартного нормального розподілу.
  - logit модель використовує кумулятивну функцію щільності логістичного розподілу.

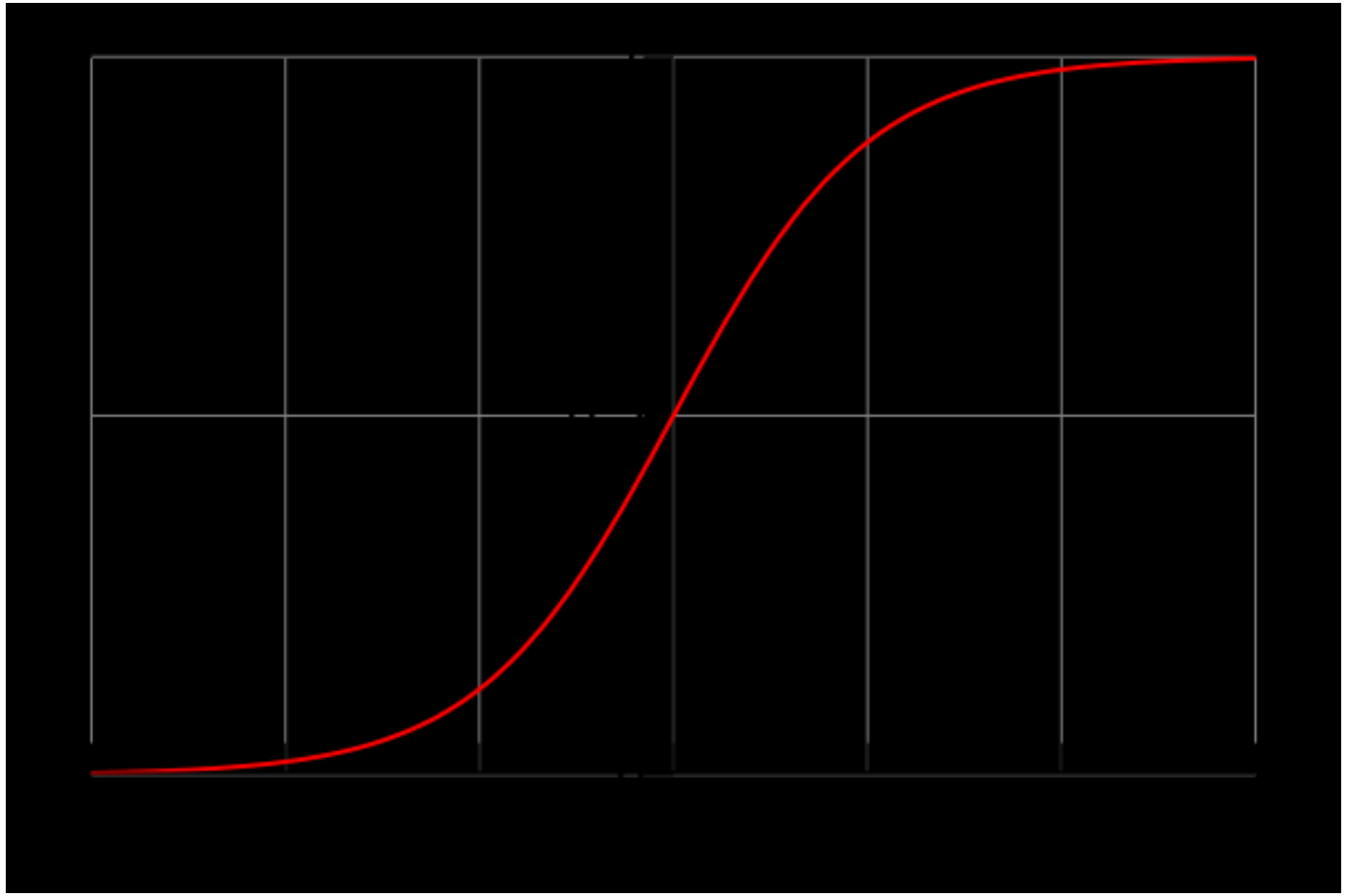
Logit модель

- Розраховує ймовірності за функції:

$$Y_i = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X_i)}}$$

- Чому дорівнює Yi при:
  - Xi = +∞
  - Xi = -∞?

Сигмовидна крива



Ймовірність

- Формула

$$P_i = \frac{1}{1+e^{-(\beta_0+\beta_1 X_i)}} = \frac{1}{1+e^{-Z_i}} = \frac{e^Z}{1+e^Z}$$

- де

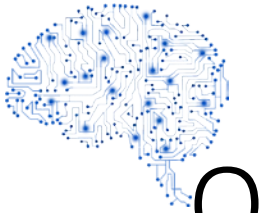
$$Z_i = \beta_0 + \beta_1 X_i$$

Логарифм відношення шансів

- logit модель можна представити так:

$$L_i = \ln\left(\frac{P_i}{1-P_i}\right) = Z_i = \beta_0 + \beta_1 X_i$$

- Ця модель дає можливість інтерпретувати вплив зміни коефіцієнтів на шанси перемогти.



Оцінка моделі

- Модель оцінюється методом максимальної правдоподібності
- Значимість коефіцієнтів здійснюється за допомогою z-statistics
- Для аналізу якості моделі застосовується Pseudo R^2
- Для перевірки адекватності моделі застосовується Likelihood Ratio test (LR), який має χ2-розподіл з k степенями свободи

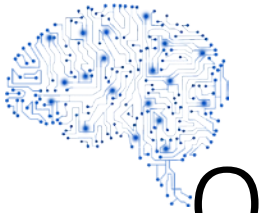
Кумулятивна функція щільності для нормального розподілу

- Функція щільності:

$$f(X) = \frac{1}{\sqrt{2\sigma^2\pi}} e^{-\left(\frac{(X-\mu)^2}{2\sigma^2}\right)}$$

- Кумулятивна функція щільності:

$$F(X) = \int_{-\infty}^{X_0} \frac{1}{\sqrt{2\sigma^2\pi}} e^{-\left(\frac{(X-\mu)^2}{2\sigma^2}\right)}$$



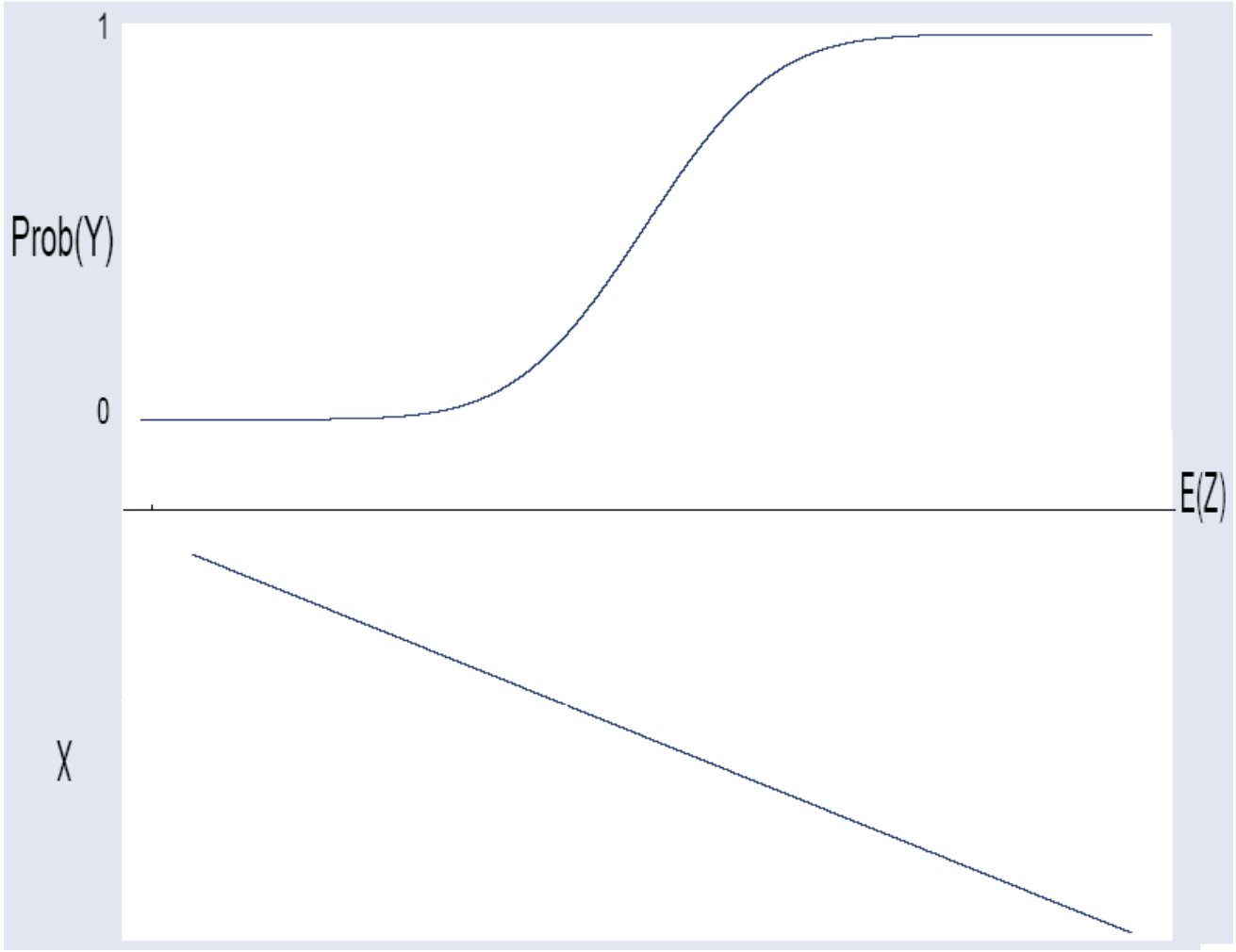
Оцінка probit

- Метод максимальної правдоподібності
- Якість моделі - Pseudo R^2
- Адекватність моделі - LR ratio з k степенями свободи

Припущення моделей

- Всі Y знаходяться у множині {0,1}
- Вони статистично незалежні
- Немає мультиколінеарності

Графік

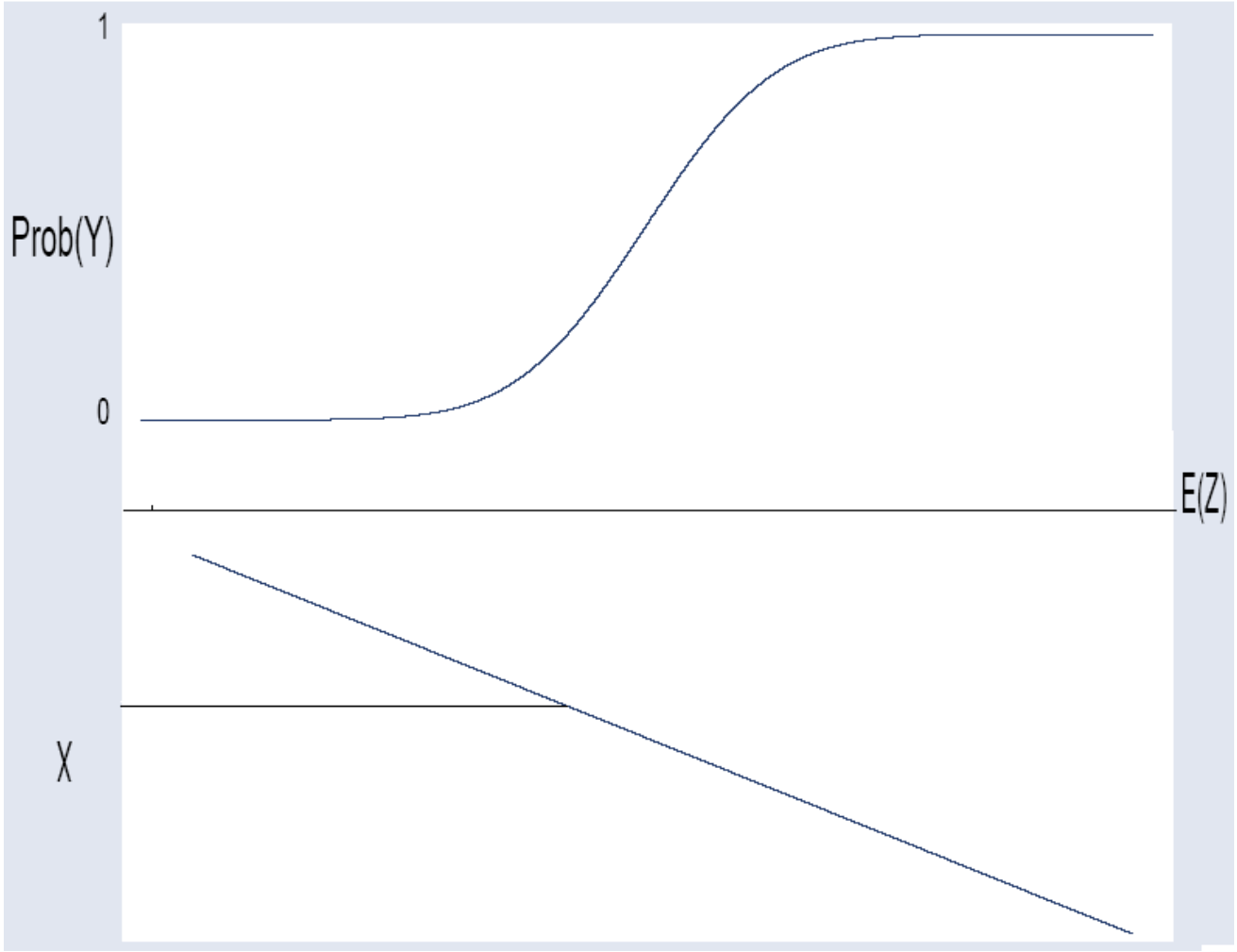


Прогноз – 1

- Для прогнозу ймовірності Prob(Y) для заданого значення X спочатку обраховується Z за допомогою регресії:

$$E(Z) = \hat{Z}_i = \beta_0 + \beta_1 \hat{X}_i$$

Графік-прогноз – 1

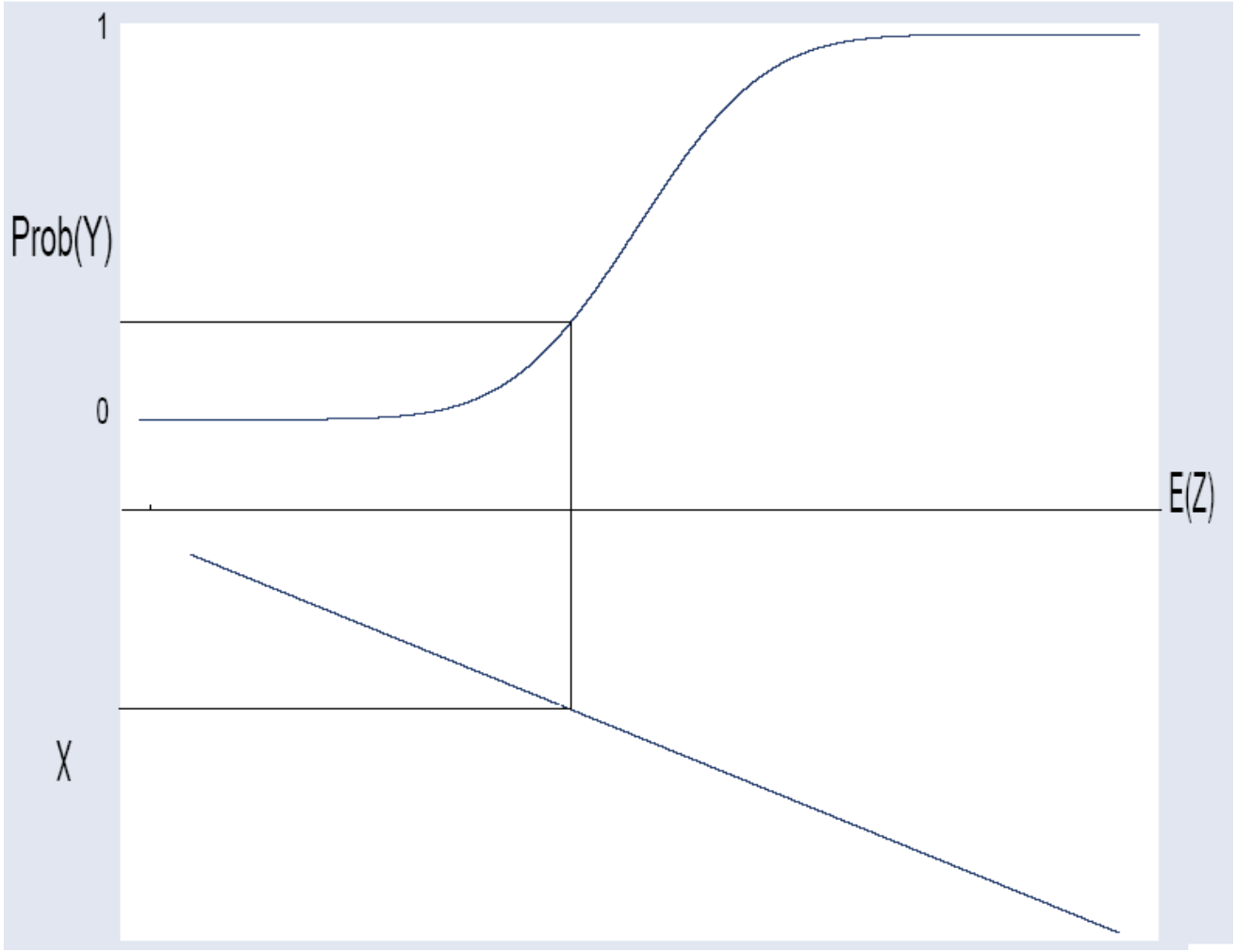


Прогноз – 2

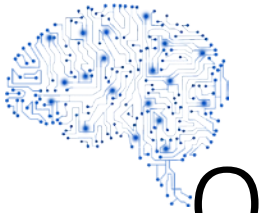
- Потім використовуємо нелінійну функцію для трансформації Z у Prob(Y):

$$\widehat{Prob(Y)} = F(\hat{Z})$$

Графік-прогноз – 2

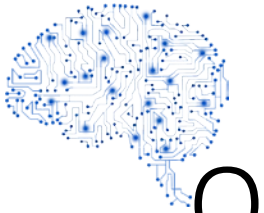


3. Оцінка та аналіз PROBIT/LOGIT моделей



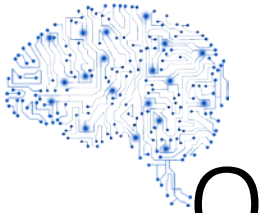
Оцінка Probit/Logit моделей – 1

- Здійснюється за допомогою методу максимальної правдоподібності.



Оцінка Probit/Logit моделей – 2

- Необхідно зазначити такі елементи:
  - Залежна фіктивна змінна (результат і-ї гри команди (1 – виграш, 0 – програш))
  - Пояснювальну незалежну змінну (рівень котирувань на і-й матч)
  - Яку нелінійну функцію використовувати як транслятор (logit чи probit)



Оцінка Probit/Logit моделей – 3

- Комп’ютер розраховує коефіцієнти β_i
- Аналітик їх аналізує та інтерпретує.

Приклад

Dependent Variable: WIN
Method: ML - Binary Logit (Quadratic hill climbing)
Time: 17:21
Sample: 1 644
Included observations: 644
Convergence achieved after 3 iterations
Covariance matrix computed using second derivatives

Variable	Coefficient	Std. Error	z-Statistic	Prob.
C	**6.55E-17**	**0.082525**	**7.94E-16**	**1.0000**
SPREAD	**−0.109814**	**0.015211**	**−7.219175**	**0.0000**

Mean dependent var	0.500000	S.D. dependent var	0.500389
S.E. of regression	0.477754	Akaike info criterion	1.301076
Sum squared resid	146.5357	Schwarz criterion	1.314951
Log likelihood	−416.9466	Hannan-Quinn criter.	1.306460
Restr. log likelihood	−446.3868	Avg. log likelihood	−0.647433
LR statistic (1 df)	58.88031	McFadden R-squared	0.065952
Probability(LR stat)	1.68E-14		
Obs with Dep = 0	322	Total obs	644
Obs with Dep = 1	322		

Аналіз Probit/Logit моделей

- Оцінка коефіцієнта дорівнює -0.1098
- Збільшення на 1 котирувань зменшує E(Z) на 0.1098.
- Як інтерпретувати нахил dZ/dX ?

Аналіз статистичної значимості

- Для аналізу нахилу dZ/dX використовуються z-статистики аналогічно до t-статистик у МНК.

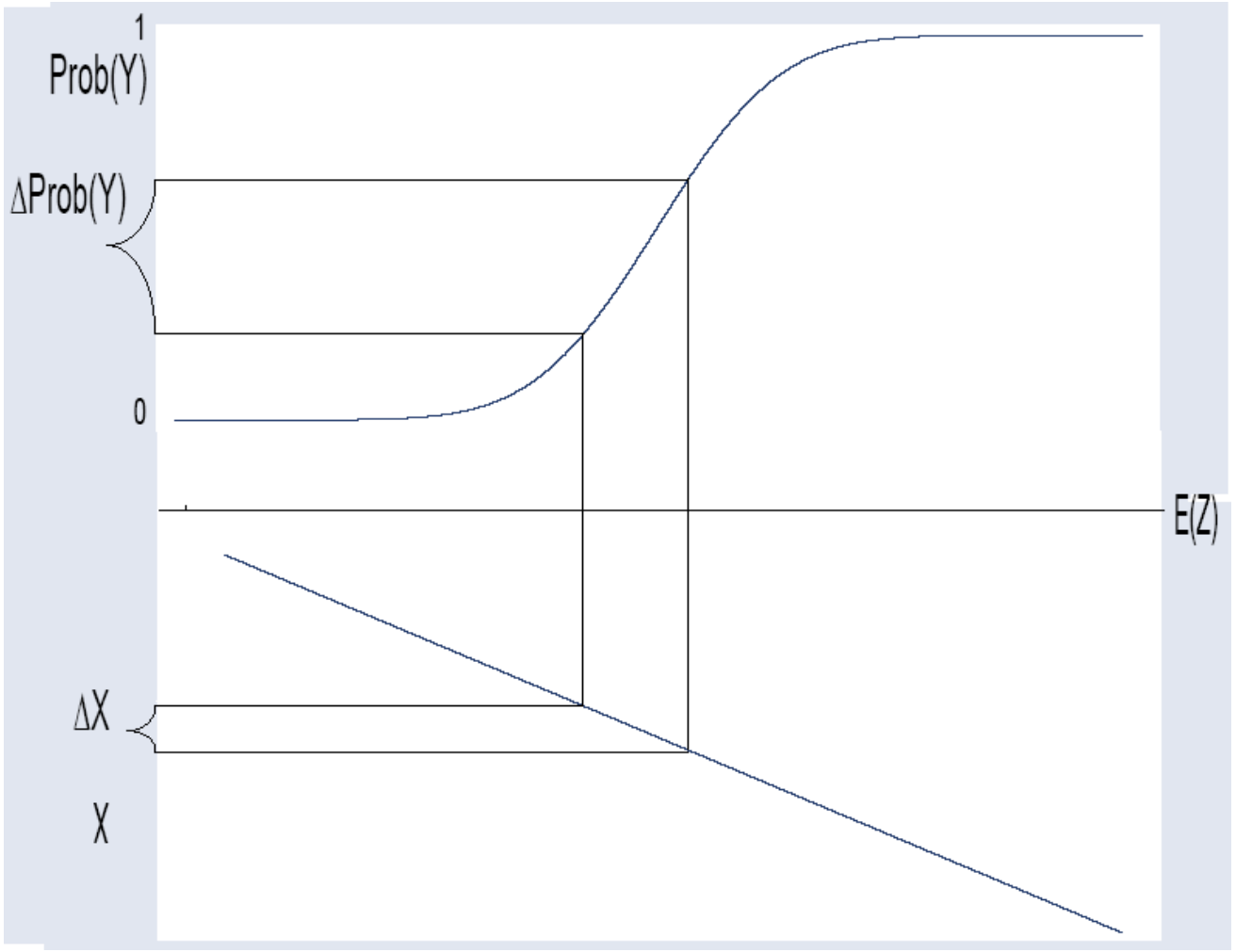
Аналіз знаків

- Якщо dZ/dX>0, то dProb(Y)/dX>0.
- z-statistic=-7.22, перевищуючи 5% критичне значення 1.96. Змінна котирувань статистично значима для моделі.
- Коефіцієнт від’ємний. Більше значення котирувань призводить до нижчих шансів на перемогу команди.

Підхід оцінки величини зміни впливу

- Спрогнозувати Prob(Y) для різних значень X, розрахувати зміну ймовірності.

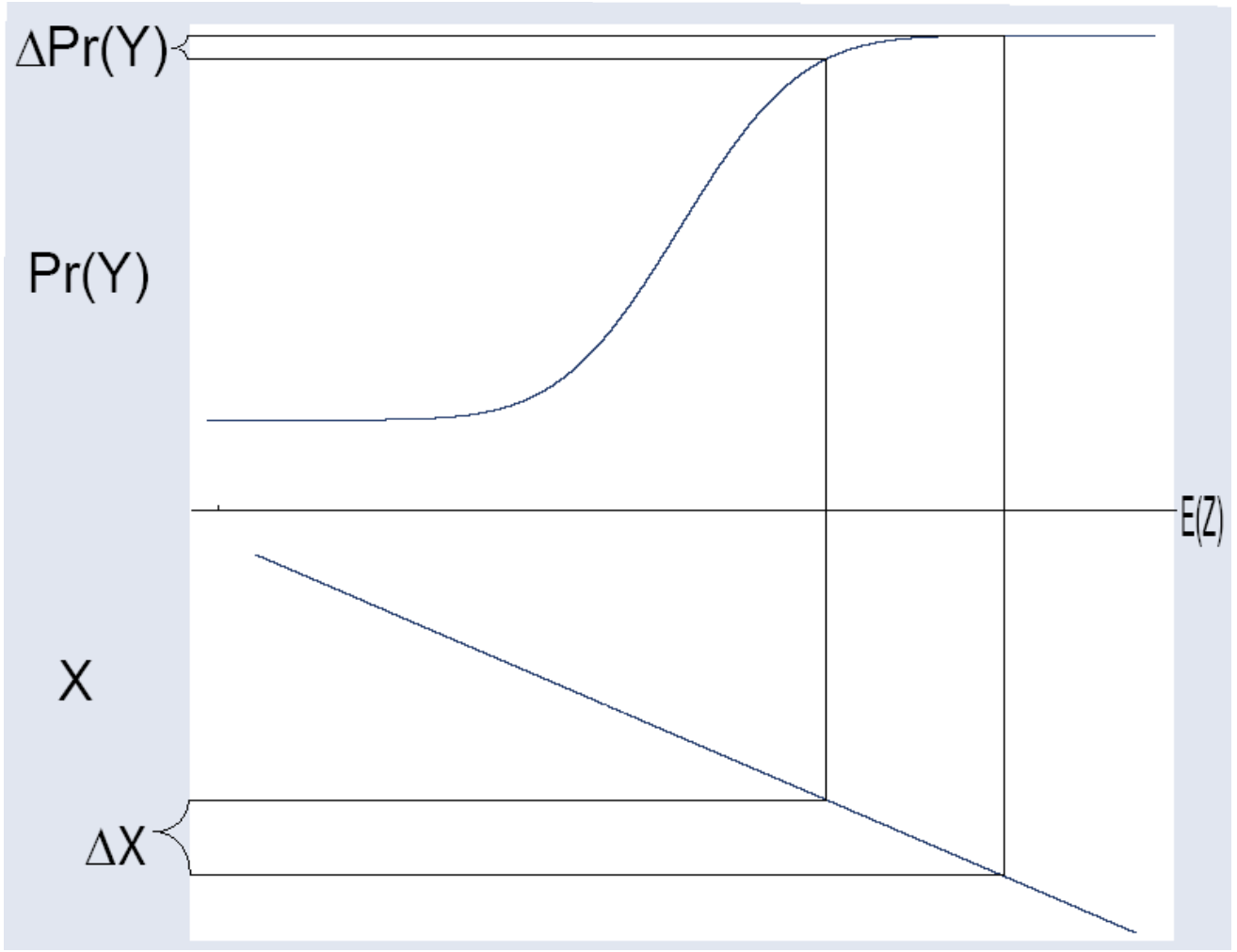
Підхід – 2



Але…

- Ефект зміни X на 1 сильно міняється, залежачи від початкового значення E(Z).
- E(Z) залежить від значень всіх незалежних змінних.

Підхід – 2



Приклад - 1

- Розглянемо зміну на 1, якщо SPREAD = 5.88, причому за умови, що інші змінні не є релевантними.

Приклад – 2

- Крок 1: Розраховуємо E(Z) для X = 5.88 та X = 6.88, використовуючи відповідну регресійну функцію.
- Крок 2: Підставляємо E(Z) у формулу логістичної функції щільності (logit).

Приклад - регресія

Dependent Variable: WIN
Method: ML - Binary Logit (Quadratic hill climbing)
Time: 17:21
Sample: 1 644
Included observations: 644
Convergence achieved after 3 iterations
Covariance matrix computed using second derivatives

Variable	Coefficient	Std. Error	z-Statistic	Prob.
C	**6.55E-17**	**0.082525**	**7.94E-16**	**1.0000**
SPREAD	**−0.109814**	**0.015211**	**−7.219175**	**0.0000**

Mean dependent var	0.500000	S.D. dependent var	0.500389
S.E. of regression	0.477754	Akaike info criterion	1.301076
Sum squared resid	146.5357	Schwarz criterion	1.314951
Log likelihood	−416.9466	Hannan-Quinn criter.	1.306460
Restr. log likelihood	−446.3868	Avg. log likelihood	−0.647433
LR statistic (1 df)	58.88031	McFadden R-squared	0.065952
Probability(LR stat)	1.68E-14		
Obs with Dep = 0	322	Total obs	644
Obs with Dep = 1	322		

Приклад – 3

$$Z(5.88) = 0 - 0.1098 \bullet 5.88 = 0.6456$$

$$Z(6.88) = 0 - 0.1098 \bullet 6.88 = 0.7554$$

$$F(\hat{Z}) = \frac{\exp(\hat{Z})}{1 + \exp(\hat{Z})}$$

$$F(0.7554) - F(0.6456) = 3.20 - 3.44 = -0.024.$$

Приклад – 4

- Зміна котирувань з 5.88 до 6.88 призводить до зниження шансів на виграш на 2.4%.
- А ось зміна котирувань з 8.88 до 9.88 дає зменшення лише на 2.1%.

4. Аналіз шоків PROBIT/LOGIT

Припущення для Y

Y_i визначається змінною Z.

$$Z_i = \beta_0 + \beta_1 X_{1i} + .. + \beta_K X_{Ki} + \varepsilon_i$$

$Y_i = 1$ якщо $Z_i > 0$

$Y_i = 0$ якщо $Z_i \leq 0$

Припущення для залишків

- Ми знаємо розподіл ε_i.
- У probit моделі залишки ε_i мають стандартний нормальний розподіл.
- У logit моделі залишки ε_i мають логістичний розподіл.

Шоки – 1

- Оскільки ми знаємо розподіл залишків ε_i , то можемо розрахувати ймовірність того, що конкретне спостереження отримає шок, який змусить перейти від додатного значення $Z > 0$ до від’ємного ($Z<0$).

Шоки – 2

- Розраховуємо

$$E(Z_i) = \beta_0 + \beta_1 X_{1i} + \ldots + \beta_K X_{Ki}$$

- Визначаємо області, для яких

$$E(Z_i) + \varepsilon_i < 0 \text{ or } E(Z_i) + \varepsilon_i > 0$$

- Використовуючи розподіл ε_i , рахуємо ймовірність попадання єі до кожної області.

Приклад – 1

Нехай $E(Z_i) = 1$

Якщо $\varepsilon_i > -1$, то $E(Z_i) + \varepsilon_i > 0$

Якщо $E(Z_i) + \varepsilon_i > 0$, то $Y_i = 1$

Якою буде $Prob(\varepsilon_i > -1)$

для стандартного

нормального розподулу?

Приклад – 2

- $\text{Prob}(\varepsilon_i > -1) \approx 0.83$
- Якщо $Z_i = 1$, то з імовірністю 83% Y = 1.

Приклад – 3

- Нехай оцінюємо знову probit модель та $E(Z_i) = -2$. Для яких значень ε_i буде $Z_i > 0$ (тобто $Y = 1$)?
- Якщо $\varepsilon_i > 2$, $Z_i > 0$ (то $Y = 1$).
- Для стандартного нормального розподілу $Prob(\varepsilon_i > 2) \approx 0.025$. Ймовірність виграшу лише 2.5%.

Загальний розв’язок – 1

Нехай ε_i має кумулятивну функцію щільності F

Тобто $Prob(\varepsilon_i < a) = F(a)$

$Prob(\varepsilon_i > a) = 1 - F(a)$

Якщо F симетрична, то $1 - F(a) = F(-a)$

Загальний розв’язок – 2

$$Prob(Y_i = 1)$$

$$= Prob(\varepsilon_i > -E(Z_i))$$

$$= Prob(\varepsilon_i > -\hat{\beta}_0 - \hat{\beta}_1 X_{1i} - \ldots \hat{\beta}_K X_{Ki})$$

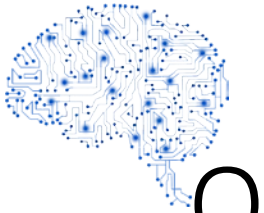
$$= 1 - Prob(\varepsilon_i < -\hat{\beta}_0 - \hat{\beta}_1 X_{1i} - \ldots \hat{\beta}_K X_{Ki})$$

$$= 1 - F(-\hat{\beta}_0 - \hat{\beta}_1 X_{1i} - .. \hat{\beta}_K X_{Ki})$$

$$= F(\hat{\beta}_0 + \hat{\beta}_1 X_{1i} + .. \hat{\beta}_K X_{Ki})$$

(для симетричних розподілів)

Огляд



Основна мета

- Оцінити ймовірність настання певної події.
- Розглядаємо лінійну ймовірнісну модель, у якій залежною змінною є бінарна змінна.

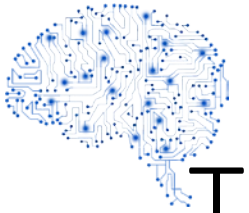
Проблеми з LPM

- Похибки не розподілені нормально
- Похибки гетероскедастичні
- Прогнозні значення залежної змінної можуть бути за межами 0 та 1.

Вимоги

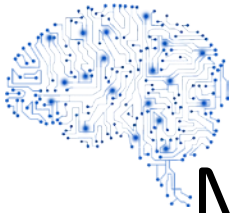
- Лінійна регресія теоретично дає прогнози від -∞ до +∞.
- Ймовірності мають бути між 0 та 1.
- Лінійна ймовірнісна модель не зможе гарантувати адекватні прогнози.

- Probit або Logit



Транслятор

- Необхідно розробити транслятор, який:
  - При наближенні прогнозу до -∞ ймовірність має наближатися до 0.
  - При наближенні прогнозу до +∞ ймовірність має наближатися до 1.
  - Не існує ймовірностей менших 0 та більших 1.



Model

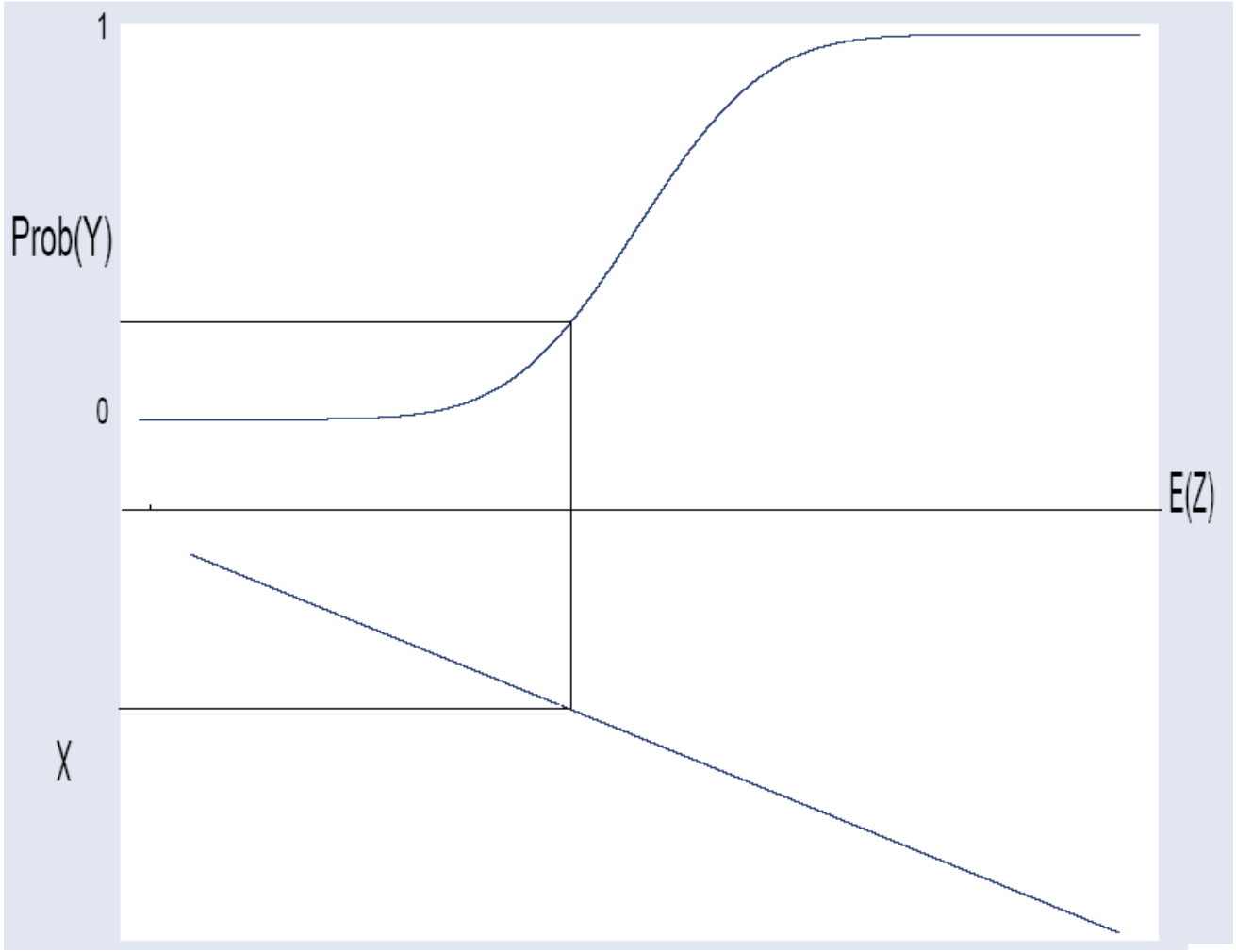
- Для прогнозу Prob(Y) для певного X, розраховуємо Z за допомогою регресії :

$$E(Z) = \hat{Z}_i = \beta_0 + \beta_1 \hat{X}_i$$

- Потім використовуємо нелінійну функцію для трансформації Z у Prob(Y):

$$\widehat{Prob(Y)} = F(\hat{Z})$$

Графік прогнозу



Припущення для Y

Y_i визначається змінною Z.

$$Z_i = \beta_0 + \beta_1 X_{1i} + .. + \beta_K X_{Ki} + \varepsilon_i$$

$Y_i = 1$ якщо $Z_i > 0$

$Y_i = 0$ якщо $Z_i \leq 0$

Припущення для залишків

- Ми знаємо розподіл ε_i.
- У probit моделі залишки ε_i мають стандартний нормальний розподіл.
- У logit моделі залишки ε_i мають логістичний розподіл.

Загальний розв’язок

$$Prob(Y_i = 1)$$

$$= Prob(\varepsilon_i > -E(Z_i))$$

$$= Prob(\varepsilon_i > -\hat{\beta}_0 - \hat{\beta}_1 X_{1i} - \ldots \hat{\beta}_K X_{Ki})$$

$$= 1 - Prob(\varepsilon_i < -\hat{\beta}_0 - \hat{\beta}_1 X_{1i} - \ldots \hat{\beta}_K X_{Ki})$$

$$= 1 - F(-\hat{\beta}_0 - \hat{\beta}_1 X_{1i} - .. \hat{\beta}_K X_{Ki})$$

$$= F(\hat{\beta}_0 + \hat{\beta}_1 X_{1i} + .. \hat{\beta}_K X_{Ki})$$

(для симетричних розподілів)

Питання?