Binary Dependent Variable

GV 903 Week 16

Winnie Xia Muzhou Zhang

January 18, 2023

Binary Data

- A variable is binary if it only has two values, 0 or 1 ("No" or "Yes", etc.)
    - Did you vote or not?
    - Did a country adopt this new policy or not?
    - Did the war or protest end or not?

Why not linear model

- A typical OLS equation looks like:

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

and assumes that the error term, ε, is normal.

Why not linear model

- A typical OLS equation looks like:

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

  and assumes that the error term, ε, is normal.
- Running OLS with a binary dependent variable is called **linear probability model**.

Why not linear model

- A typical OLS equation looks like:

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

  and assumes that the error term, ε, is normal.
- Running OLS with a binary dependent variable is called **linear probability model**.
- The interpretation is the exact same as regular OLS. The only difference is that our interpretation of the dependent variable is now in probability terms.

Why not linear model

- A typical OLS equation looks like:

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

  and assumes that the error term, ε, is normal.
- Running OLS with a binary dependent variable is called **linear probability model**.
- The interpretation is the exact same as regular OLS. The only difference is that our interpretation of the dependent variable is now in probability terms.
- So we say, **a one-unit increase in X is associated with a three percentage point increase in the probability**.

Why not linear model

- We will get prediction outside the interval between 0 and 1.

Why not linear model

- We will get prediction outside the interval between 0 and 1.
- Violate the homoscedasticity assumption of OLS.

Why not linear model

- We will get prediction outside the interval between 0 and 1.
- Violate the homoscedasticity assumption of OLS.
- By using OLS estimator, we assume linear trend in probabilities.

Generalized Linear Model

- A GLM equation looks like:

$$E(Y|X) = F(\beta_0 + \beta_1 X)$$

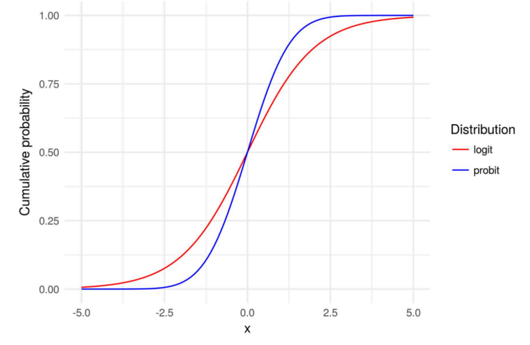
- Key difference:
    - **estimated by maximum likelihood**, rather than OLS.
    - **binomial distribution for binary data**, rather than normal distribution.
    - In R, **glm()**, rather than lm().

Probit vs. Logit

- Logit model is a form of a statistical model that is used to predict the probability of an event occurring
- Probit model is similar to logit model, but it determines the likelihood that an item or event will fall into one of a range of categories by estimating the probability that observation with specific features will belong to a particular category.
- So dependent variable for probit model can only take on **one of the two values, such as yes or no, true or false**.

Probit vs. Logit

Logit models are used to **model logistic distribution** while probit models are used to model the **cumulative standard normal distribution**.



Interpretation

- Coefficients are log odd-ratios.

Interpretation

- Coefficients are log odd-ratios.
- From the coefficients themselves we can get direction (positive/negative) and significance, but not **the size of effect**.

Interpretation

- Coefficients are log odd-ratios.
- From the coefficients themselves we can get direction (positive/negative) and significance, but not **the size of effect**.
- So we convert them into odds-ratio by exponentiating: expo(coef(mode))

How to interpret odds ratio

- 1 is the reference point.

How to interpret odds ratio

- 1 is the reference point.
- Odds ratio above 1 increased occurrence of an event.

How to interpret odds ratio

- 1 is the reference point.
- Odds ratio above 1 increased occurrence of an event.
- Odds ratio below 1 decreased occurrence of an event.

Predicted Probability/ Marginal effect

- To interpret the effects more clearly, we need to calculate the predicted probability.

Predicted Probability/ Marginal effect

- To interpret the effects more clearly, we need to calculate the predicted probability.
    - predict(model, newdata, type="response")
    - And we write, **the predicted probability for the occurrence of an event is *****.