SOSC 5340: Logistic Regression and MLE

Han Zhang

Feb 16, 2021

Outline

How you wish I can be improved?

- *It is a little bit slow as most of us have already had some quantitative backgrounds. I hope we can move to causal inference as soon as possible.*
- *Personally, I think it is a little too fast and sometimes I have problems catching up*

- The class has mixed background
- If you find class content too easy
  - Try to prove everything we mentioned (sometime I skipped the proofs). You will find that it's not that easy if you start doing it by yourself
    - This will lay out a solid foundation of what we are going to learn next (especially machine learning and causal inference); when things get complex, it's not as easy as you think
    - For instance, are regression estimate of IV and DID consistent? unbiased? asymptotically normal?
  - Come to me and I can suggest you more things to read
  - Maybe you have talent to do methodology research

How you wish I can be improved

- *Too abstract!*
- *Bit theory and not too down-to-earth; too much math for non-MATH/ECON background*

- Purely teaching you how to run regression will let you start early, but won't let you go far
- We will see how the abstract knowledge help us in applied work
  - lots of presentations coming soon.
  - Also our first tutorial and assignment
- If you cannot follow the proof, try to follow the logic

What I expect from you

- Give me more feedback
    - Like asking questions more often :)
- If you are afraid of peer pressure:
    - send a private question to me; I will answer it but will not mention you rname
    - click the "too slow" button if you find it hard to follow

Logistics: Assigned Papers

- The assigned paper list has been posted
- Choose one you are interested to present
  - The first presentation starts on Mar 2, two weeks later
- Two kinds of articles:
  - Applied
  - Methodological: several of them are not easy; if you believe the content is too easy, go for it. I will give you bonus points if you are able to grasp the contents.
- All are expected to read the article and ask questions

Logistics: Exercise 1

- Will post it by the end of today
- Due in two weeks (Mar 2)

Recommended readings for today

- If you want to see some formal proofs:
- Wooldridge, *Introductory Econometrics: A Modern Approach*, 2015. Chapter 17
- Hansen, *Econometrics*, 2020. Chapter 4, 5, 23. Free at the author's website `https://www.ssc.wisc.edu/~bhansen/econometrics/`

Binary Outcome

- Binary outcome variable:
  - $Y_i \in \{0, 1\}$
- Examples in social science: numerous!
  - Higher education: 1 = has college education; 0 = does not have college education
  - Conflict: 1 = civil war; 0 = no civil war
  - Voting: 1 = vote; 0 = abstain

How do we model binary outcome?

- We already know that conditional expectation $E(Y|X)$ is the best predictor
- Linear regression: with assumptions 1,2 and especially 3

$$E(Y|X) = X\beta$$

- When Y is binary:

$$E(Y|X) = P(Y = 1|X)$$

- $P(Y = 1|X)$ is the conditional probability of $\mathrm{Y} = 1$ given X
- What is different here: conditional probability must be between 0 and 1 by definition!

How do we model binary outcome?

- $E(Y|X) = X\beta$ can be bigger than 1 or smaller than 0
- Linear Probability Model: just tolerate this problem; still run OLS regression with binary outcome.
- Alternatively: we can apply a function F onto $X\beta$ to ensure

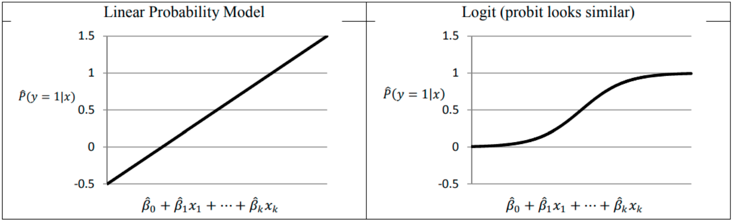
$$0 \leq E(Y|X) = F(X\beta) \leq 1$$

Logistic regression

- Two useful functions:
  - $logit(X) = log(\frac{X}{1-X})$
  - $logit^{-1}(X) = \frac{exp(X)}{1+exp(X)}$
- Logistic Regression
  - We use the inverse-logit function as F

$$E(Y|X) = logit^{-1}(X\beta) = \frac{exp(X\beta)}{1 + exp(X\beta)} = \frac{1}{1 + exp(-X\beta)}$$

Logistic Regression vs Linear Probability Model



- inverse-logit function “squashs” $X\beta$ to $[0, 1]$

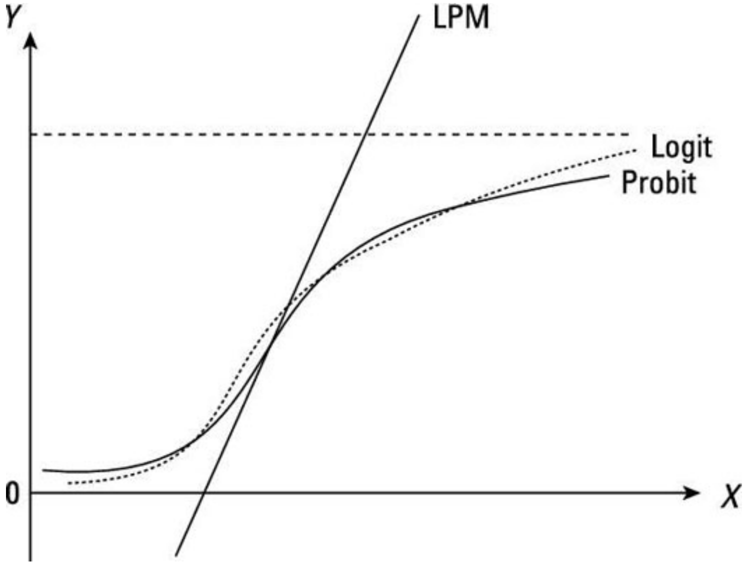
Probit regression

- We can also "squash" $X\beta$ using standard normal CDF (normal cumulative density function)

$$E(Y|X) = \Phi(X\beta)$$

- Statistical model using normal CDF to squash $X\beta$ is known as probit regression
- In general, any CDF can be used as F to squash $X\beta$ to $[0, 1]$
  - inverse-logit is the CDF of standard logistic distribution
  - Φ is the CDF of standard normal distribution

Probit vs Logit vs Linear Probability



More on linear probability model

- Binary data (and more general, most categorical data) always exhibit heteroscedasticity

$$\begin{aligned} V(\epsilon|X) &= V(Y - X\beta|X) \\ &= V(Y|X) \\ &= P(Y=1|X)\big[1 - P(Y=1|X)\big] \end{aligned} \quad (1)$$

- The above equation shows that variance of error changes based on the value of X! It is always heteroscedastic.
- So always use robust standard error if you decide to use OLS regression to model binary outcomes (linear probability model).

Assumptions of OLS regression

- Assumption 1: the expected error is 0

$$E(\epsilon) = 0$$

- Assumption 2: mean independent between X and the error

$$E(\epsilon|X) = 0$$

- Assumption 3 of OLS (data generating process)

$$Y = X\beta + \epsilon$$

- Assumption 5: normal error (which implies Assumption 4, homoscedastic error)

$$\epsilon \sim N(0, \sigma^2)$$

Assumptions of Logistic/Probit regressions

- Assumption 1 and 2: shared by logit/probit regressions
- Assumption 3 of logit/probit: data generating process

$$
\begin{aligned}
Y^* &= X\beta + \epsilon \\
Y &= \begin{cases} 1 & \text{if } y^* > 0 \\ 0 & \text{otherwise} \end{cases}
\end{aligned} \tag{2}
$$

- Y^* is an unobserved latent variable
- if the latent variable is bigger than a pre-determined cutoff (here 0), we get $Y = 1$
- We only observe samples of Y
    - economists may say that Y^* is the underlying preference, and Y is revealed preference

Assumptions about error of logit/probit

- Assumption 5 of Logistic/Probit regressions
  ϵ is distributed according to the probability density distribution of a CDF function F
  - F is inverse-logit function; the error follows standard logistic distribution
  - F is Φ; the error follows standard normal distribution

Assumptions 3 and 5 together lead to

$$E(Y|X) = F(X\beta)$$

Estimation of parameters in OLS regressions: review

- There are two ways to estimate β in linear regression
- We can write some population equations, plug-in the sample analog, and solve these sample equations
- We can also directly minimize empirical MSE
- Both solutions result in the same β estimate for OLS regression

$$\hat{\beta} = \left[\mathbf{X}^T\mathbf{X}\right]^{-1}\mathbf{X}^T\mathbf{Y} \tag{3}$$

Maximum Likelihood Estimation

- There is no way to write down a closed-form solution for logistic regression coefficients.
- We use Maximum Likelihood Estimation (MLE)
- MLE is a general methods for estimating parameters in parametric statistical models and making statistical inference.
- Requirement: assumptions about functional form of conditional probability $P(Y|X)$
- Say, in logistic regression, $P(Y = 1|X) = logit^{-1}(X\beta)$, and $P(Y = 0|X) = 1 - P(Y = 1|X)$
- For a single data point, the probability we observe Y_i is exactly given by $logit^{-1}(X_i\beta)$ or $1 - logit^{-1}(X_i\beta)$ (depending on observed Y_i)

Maximum Likelihood Estimation

- Because we have i.i.d. samples, we can multiple these empirical probabilities together, as the probability that we observe the entire sample.
- The probability we observe the entire sample is called likelihood: L

$$L = \prod_{i=1}^{n} P(Y_i|X_i) \tag{4}$$

- L is a function of unknown β
- Naturally, we say that a good β is the one that makes the likelihood the largest.
  - Intuitively, it says that our chosen β should make the probability to observe the entire sample the largest.
- Put it differently, our estimate of β should maximize the likelihood function.

MLE estimate

- In practice, it is easier to work with log of likelihood, called log-likelihood
- $\log L = \sum_{i=1}^{n} \log P(Y_i|X_i)$
- We try to find β that maximize log-likelihood

$$\hat{\beta}_{MLE} = \arg\max_{\beta} \log L$$

MLE inference

- And estimated variance of $\hat{\beta}_{MLE}$ is given by

$$\widehat{V}(\hat{\beta}_{MLE}) = \left(\mathbb{E}_{\beta}\left(\frac{\partial^2 \log L}{\partial \beta^2}\right)\right)^{-1} \tag{5}$$

- $\frac{\partial^2 \log L}{\partial \beta^2}$ is called Hessian matrix.
- Last, we can use normal approximated intervals for confidence interval (below is an example for 95% confidence interval)

$$\left(\hat{\beta}_{MLE} - 1.96 * \hat{\sigma}(\hat{\beta}_{MLE}), \hat{\beta}_{MLE} - 1.96 * \hat{\sigma}(\hat{\beta}_{MLE})\right)$$

MLE properties

- MLE estimate has some good properties:
- It is consistent
- It is asymptotically normal (so we can use normal-approximated confidence interval)
- Unbiaseness? No guarantee

MLE in practice: logistic regression

- Step 1: write single point probability distribution; this case it is easy:
  - $P(Y_i = 1|X_i) = logit^{-1}(X_i\beta)$, and $P(Y_i = 0|X_i) = 1 - P(Y_i = 1|X)$
  - We can write this in a single equation:

$$P(Y_i|X_i) = \left[logit^{-1}(X_i\beta)\right]^{Y_i}\left[1 - logit^{-1}(X_i\beta)\right]^{1-Y_i} \tag{6}$$

- Step 2: for all n points:

$$L = \prod_{i=1}^{n} P(Y_i|X_i) = \prod_{i=1}^{n} \left[logit^{-1}(X_i\beta)\right]^{Y_i}\left[1 - logit^{-1}(X_i\beta)\right]^{1-Y_i} \tag{7}$$

MLE in practice: logistic regression

- Step 2 (cont'd): the log-likelihood is

$$\log L = \sum_{i=1}^{n} Y_i \log \left(logit^{-1}(X_i) + (1 - Y_i) \right) \log \left[1 - logit^{-1}(X_i) \right] \quad (8)$$

- And remember that $logit^{-1}(X\beta) = \frac{exp(X\beta)}{1+exp(X\beta)}$
- We want to select β that makes $\log L$ the largest

Optimization

- How can we find β that minimize $\log L$? Two solutions
- Standard calculus
  - Find β that makes the partial derivative $\frac{\partial L}{\partial \beta} = 0$.
  - In logistic regression, you cannot analytically solve β that makes the partial derivative zero.
- Optimization:
  - Try many β and choose one that minimize $\log L$.
  - How? There may be infinite choices of β
  - There are many mature optimization algorithms that help you find β quicker

Optimization

- There are many many more optimization methods
- They basically follow the similar idea: makes some initial guesses of β and gradually improve on older estimates
- in R, use `optim` package

Optimization

- One commonly used optimization method: gradient descent
  - It's not used in `optim` package in R but widely used in more advanced algorithms

$$\beta_{new} = \beta_{old} + \eta \cdot \frac{\partial \log L}{\partial \beta} \tag{9}$$

- With some math, you will find that

$\frac{\partial \log L}{\partial \beta} = \sum_{i=1}^{n} \left[Y_i - logit^{-1}(X\beta) \right] X_i$

- η is called learning rate; try different options
- You need to choose an starting β; try several random guess

Odds and Log Odds

- Let us move on to interpreting regression coefficients

$$X\beta = logit(E(Y|X)) = log\Big[\frac{P(Y=1|X)}{1-P(Y=1|X)}\Big] = log\Big[\frac{P(Y=1|X)}{P(Y=0|X)}\Big]$$

- $\frac{P(Y=1|X)}{P(Y=0|X)}$ is called odds; it is the ratio between two conditional probabilities: $Y=1$ vs $Y=0$, given X.
    - Odds >1 means $Y=1$ is more likely than $Y=0$ give X
- $log\Big[\frac{P(Y=1|X)}{P(Y=0|X)}\Big]$ is the log of odds; we call it log-odds
- Following the interpretation of OLS regression, we can interpret logistic regression coefficient in this way:
    - One unit increase in X will lead to β increase in log-odds
    - Problem: it is very intuitive to think about what β increase in log-odds means

Logistic Regression Interpretations: Approach 1

- Example, we are interested in the effect of income and gender on whether a person vote or not. For gender, 1 is female and 0 is female. Income is in thousand dollars

$$P(Y=1|X) = logit^{-1}(-1.92 + 0.032 * \text{income} + 0.67 * \text{gender})$$

- A simple rule of thumb (based on Gelman and Hill, *Data Analysis using Regression and Multilevel Hierarchical Models*, 2007.)
  - Divide your β by 4, and this is roughly the upper bound of the change in probability
  - For income, we divide 0.032 by 4. It means that one unit (a thousand) increase in income predicts no more than 0.8% increase in the probability of voting.
  - For gender, $0.67/4 = 0.168$. This suggests that female's voting probability is 16.7% more than that of male's
  - Do not write this in formal paper!

Logistic regression interpretations: Approach 2

- Remember one unit increase in X lead to β increase in log-odds.
- Write the conditional probability $P(Y=1|X)$ before change as p_b, and the condition probability $P(Y=1|X)$ after increasing X for one unit as p_a

$$log\frac{p_a}{1-p_a} - log\frac{p_b}{1-p_b} = \beta \implies \frac{\frac{p_a}{1-p_a}}{\frac{p_b}{1-p_b}} = exp(\beta)$$

- $\frac{\frac{p_a}{1-p_a}}{\frac{p_b}{1-p_b}}$ is called odds ratio
- One unit increase in X leads to $exp(\beta)$ change in odds ratio
- For income, $exp(0.032) = 1.03$
  - This means that odds is 1.03 times higher for one unit increase in income
  - Or in other words, odds ratio increase by 3%
- For gender, $exp(0.67) = 1.95$
  - This means that odds of voting is 1.95 times higher among females compared with males

Logistic regression interpretations: Approach 3

- We can always calculate the marginal effect: how conditional probability changes for one unit increase in X: $\frac{\partial P(Y=1|X)}{\partial X}$
- After some calculations, you will find that;

$$\frac{\partial P(Y = 1|X)}{\partial X} = \beta(logit^{-1}X\beta)(1 - logit^{-1}X\beta)$$

- In other words, one unit increase in X leads to $\beta(logit^{-1}X\beta)(1 - logit^{-1}X\beta)$ changes in predicted probability
- It is easy to see that the marginal effect will change depending on exact values of X
- The marginal effect is generally bigger, when X is around the mean

Logistic regression interpretations: Approach 3

- Typically there are two ways to visualize/show marginal effect
- Marginal effect at the mean (MEM)
  - Set all other variable at their mean value
  - MEM is the change in predicted probability when the focal independent variable change for one unit
  - Cons: setting categorical variables at their means are not meaningful
    - e.g., 0 is female and 1 is male; what is gender = 0.45 means?
- Average marginal effect (AME)
  - For each observation, holding other variables at their observed value; calculate marginal effect for one focal variable
  - Take the average of marginal effects of the focal variable for each observation
- R package `margins` and stata command `margins` will return AME by default; has to explicit set parameters to calculate marginal effect at the mean
- `https://cran.r-project.org/web/packages/margins/vignettes/TechnicalDetails.pdf`

Logistic regression interpretations: Approach 4

- Just plot predicted probability versus one focal variable you are mainly interested in
- And holding other X at a fixed level.
  - say, holding others at the mean
  - or at a particular value that are theoretically interesting
- This is especially useful if you have interaction terms

Predicted probability (example)

See RMarkdown codes and files.

What are practical recommendations?

- Use the divide by 4 rule and make an intuitive sense of how large the effect is
- Then calculate AME or MEM
- Or plot the predicted probabilities versus the key independent variables
- You can state that
  - One unit increase in X leads to β change in log-odds
  - Or, one unit increase in X leads to $exp(\beta)$ change in odds ratio
  - (but I personally find them hard to grasp; and I am sure I am not the only one)

How to interpret probit regressions?

- No direct substantive interpretation of β in probit regressions (it is not an odds ratio)
- Probit just makes math calculation easier, but it lacks a natural interpretation.

Today's Review

- What are the assumptions of logistic/probit regressions?
- What is MLE?
- Different views to interpret logistic regression results
  - divide by 4 rule
  - marginal effect
  - plot predicted probability directly

Next week

- More on generalized linear model