

# SOSC 5340: Generalized Linear Model

Han Zhang

# Outline

Binary Outcomes

Assumptions

MLE

Interpretations

GLM

Multinomial and Ordered Logit

Poisson, Negative Binomial, and Zero-inflated Poisson

Model Selection

Bootstrap

Today's Review

# 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

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

- When  $Y$  is binary:

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

- $P(Y = 1|X)$  is the conditional probability of  $Y = 1$  given  $X$
- Conditional probability must be between 0 and 1 by definition
  - But  $\alpha + X\beta$  is not always between 0 and 1
  - So Assumption 3 is very likely to be violated

## How do we model binary outcome?

- **Linear Probability Model (LPM)**
  - Just pretend this problem does not exist; still run OLS regression with binary outcome.
- Alternatively: we can apply a function  $F$  onto  $\alpha + X\beta$  to ensure

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

# Logistic regression

- Two useful functions:

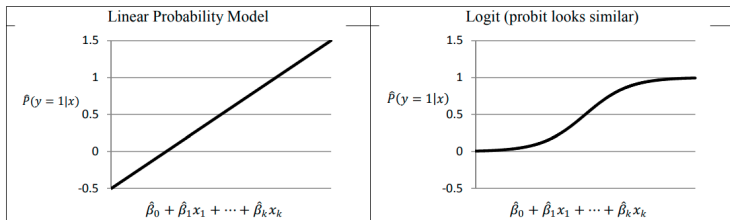
- $\text{logit}(X) = \log\left(\frac{X}{1-X}\right)$
- $\text{logit}^{-1}(X) = \frac{\exp(X)}{1+\exp(X)}$

- Logistic Regression**

- We use the **inverse-logit** function as  $F$

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

# Logistic Regression vs Linear Probability Model



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

## Probit regression

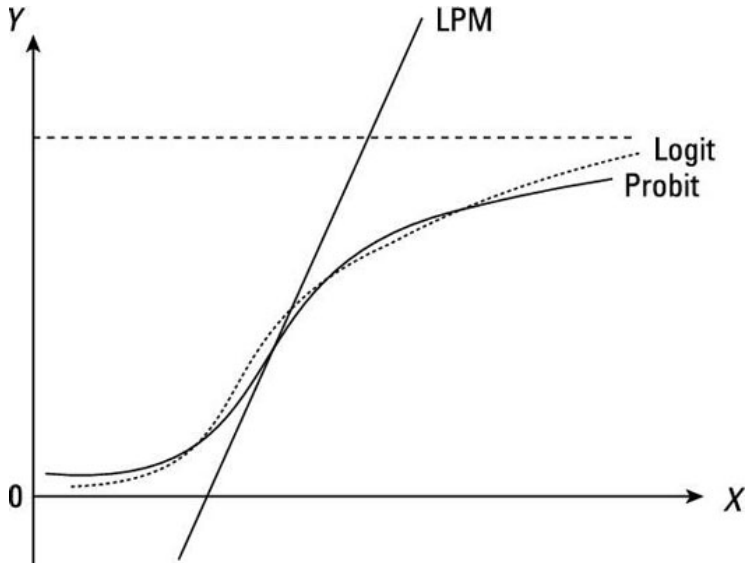
- We can also “squash”  $\alpha + X\beta$  using **standard normal CDF** (normal cumulative density function)

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

- Statistical model using normal CDF 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
  - $\Phi$  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

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

- 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 (**linear model**)

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

- 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: linear model + **non-linear** transformation

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

$$Y = \begin{cases} 1 & \text{if } y^* > 0 \\ 0 & \text{otherwise} \end{cases} \quad (2)$$

- $Y^*$  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  
 $\epsilon$  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  $\Phi$ ; 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  $\beta$  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  $\beta$  estimate for OLS regression

$$\hat{\beta} = [\mathbf{X}^T \mathbf{X}]^{-1} \mathbf{X}^T \mathbf{Y} \quad (3)$$

## 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,  $P(Y = 1|X) = \text{logit}^{-1}(X\beta)$ , 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  $\text{logit}^{-1}(X_i\beta)$  or  $1 - \text{logit}^{-1}(X_i\beta)$  (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**:  $L$

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

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

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

## MLE inference

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

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

- $\frac{\partial^2 \log L}{\partial \beta^2}$  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:
  - $P(Y_i = 1|X_i) = \text{logit}^{-1}(X_i\beta)$ , and  
 $P(Y_i = 0|X_i) = 1 - P(Y_i = 1|X_i)$
  - We can write this in a single equation:

$$P(Y_i|X_i) = [\text{logit}^{-1}(X_i\beta)]^{Y_i} [1 - \text{logit}^{-1}(X_i\beta)]^{1-Y_i} \quad (6)$$

- Step 2: for all  $n$  points:

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

## MLE in practice: logistic regression

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

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

- And remember that  $\text{logit}^{-1}(X\beta) = \frac{\exp(X\beta)}{1 + \exp(X\beta)}$
- With some math, you will find that

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

# Optimization

- We want to select  $\beta$  that makes  $\log L$  the largest
- How? Two solutions
- Standard calculus
  - Find  $\beta$  that makes the partial derivative  $\frac{\partial L}{\partial \beta} = 0$ .
  - For logistic regression, in general, you cannot analytically solve  $\beta$  that makes the partial derivative zero.
- Numerical optimization:
  - Try many  $\beta$ ; calculate their  $\log L$
  - choose one that gives the largest  $\log L$ .
  - How? There are infinite number of choices of  $\beta$
  - There are many mature optimization algorithms that help you find  $\beta$  quicker

# Optimization

- One commonly used optimization method: **gradient descent**

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

- $\eta$  is called learning rate; try different options
- You need to choose an starting  $\beta$ ; try several random guess

# Optimization

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



## Odds and Log Odds

- Let us move on to interpreting regression coefficients

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

- $\frac{P(Y=1|X)}{P(Y=0|X)}$  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  $Y=0$  given  $X$
- $\log\left[\frac{P(Y=1|X)}{P(Y=0|X)}\right]$  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  $\beta$  increase in **log-odds**
  - Problem: it is very intuitive to think about what  $\beta$  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 male. Income is in thousand dollars

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

- A simple rule of thumb (based on Gelman and Hill, *Data Analysis using Regression and Multilevel Hierarchical Models*, 2007.)
  - Divide your  $\beta$  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  $\beta$  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;

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

- In other words, one unit increase in  $X$  leads to  $\beta(\text{logit}^{-1}X\beta)(1 - \text{logit}^{-1}X\beta)$  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  $\beta$  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  $\beta$  in probit regressions (it is not an odds ratio)
- Probit just makes math calculation easier, but it lacks a natural interpretation.

## Limited Dependent Variable

- Beyond binary outcomes,  $Y \in \{0, 1\}$
- Categorical:
  - e.g., major choices;
- Integer (count):  $Y \in \{0, 1, 2, \dots\}$ 
  - e.g., event counts
- Censored: observed  $Y$  is in a certain range, but we know in reality they should not be
  - e.g., US census write anyone who report their age  $> 90$  as 90; so in census, age is between  $[0, 90]$
- The common problem is that the outcome  $Y$  is limited to some regions, not in  $(-\infty, \infty)$ 
  - so economists sometimes call them as **limited dependent variable**

# Generalized Linear Model

- To model limited dependent variables, we use **generalized linear model** (GLM)
- GLM looks like:
  - $h(E(Y|X)) = X\beta$
  - or,  $E(Y|X) = h^{-1}(X\beta)$
- $h()$  is called **link** function
- Linear regression is a kind of GLM, where  $h(X) = X$
- Logistic regression is a kind of GLM, where  $h(X) = \text{logit}(X)$
- Other GLM choose different  $h()$  to model different types of  $Y$

# GLM

- In practice, scholars use MLE to make statistical estimation and inference for GLM
- Recall that to use MLE, we need to make assumptions about what  $p(Y|X)$  looks like

# Estimation and Inference of MLE

- Steps are standard
  1. write down  $P(Y|X)$
  2. write down  $\log L$ : the log-likelihood function
  3. obtain coefficient estimates that maximize log-likelihood
    - and use Hessian matrix to calculate confidence interval

## Extending Logistic Regression

- Suppose we have categorical outcome with more than two values
- Sometimes, these categories have no intrinsic orders
  - E.g., majors choices between ( Economics = 1, Political Science = 2, Sociology = 3, Public Policy = 4 )
- Other times, these categories are **ordinal**
  - E.g., a survey ask whether you think religion deters economic growth, on a 1-7 scale.
  - 1 means strongly disagree, and 7 means strongly agree
  - Order gives more information than pure categories
  - Why not use continuous outcome models?
    - Don't want to assume equal distances between levels
    - Say, moving from 1-4 is different from 4-7
    - Assuming continuous  $Y$  does not distinguish these two

## Ordered Logit: ordered outcome

- Peter McCullough, *Regression Models for Ordinal Data*, 1980
- Recall that logistic regression assumes a generating process based on latent variables

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

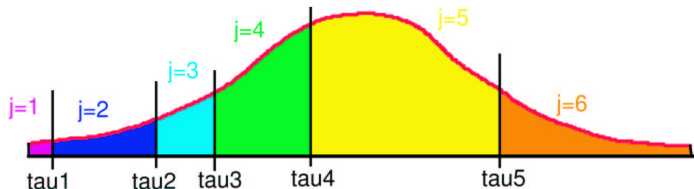
$$Y = \begin{cases} 1 & \text{if } y^* > 0 \\ 0 & \text{otherwise} \end{cases} \quad (10)$$

- $Y^*$  is an unobserved latent variable
- if the latent variable is bigger than a pre-determined **cutoff** (here 0),  $Y = 1$
- Otherwise,  $Y = 0$

## Ordered Logit

- We can borrow the same intuition to derive ordered logit regression, with  $J > 2$  ordinal categories
- We create  $J - 1$  latent cutoffs

$$Y = \begin{cases} 1 & \text{if } Y^* \leq \tau_1 \\ 2 & \text{if } \tau_1 < Y^* \leq \tau_2 \\ 3 & \text{if } \tau_2 < Y^* \leq \tau_3 \\ \cdot & \\ J & \text{if } \tau_{J-1} \leq Y^* \end{cases} \quad (11)$$





## Ordered Logit

- So now the Assumption 3 for ordered logit becomes:

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

$$Y = \begin{cases} 1 & \text{if } Y^* \leq \tau_1 \\ 2 & \text{if } \tau_1 < Y^* \leq \tau_2 \\ 3 & \text{if } \tau_2 < Y^* \leq \tau_3 \\ \vdots & \\ J & \text{if } \tau_{J-1} \leq Y^* \end{cases} \quad (12)$$

- And the error  $\epsilon$  follows a standard logistic distribution (the same as logistic regression)

## Ordered Logit vs Linear Regression

- It may be easier to change from “very unlikely” (1) to “unlikely” (2), but it is more difficult to change from “unlikely” to “neutral” (3)
- For linear regression
  - It takes the same amount of changes in  $X$  to turn  $Y$  from 1 to 2 versus  $Y$  from 2 to 3
- For ordered logit
  - $Y$  changing from 1 to 2 means latent  $Y^*$  changes from below  $\tau_1$  to  $(\tau_1, \tau_2)$
  - $Y$  changing from 2 to 3 means latent  $Y^*$  changes from  $(\tau_1, \tau_2)$  to  $(\tau_2, \tau_3)$
  - It often requires a different amount a change in  $X$  to move  $Y$  from 1 to 2 versus from 2 to 3. That's what we want to capture

## Ordered Logit

- For MLE, we have to explicitly write down  $P(Y|X)$

$$\begin{aligned}
 P(Y = 1|X) &= \Pr(\beta X + \epsilon \leq \tau_1|X) \\
 &= P(\epsilon \leq \tau_1 - \beta X|X) \\
 &= F(\tau_1 - \beta X), (\text{definition of cumulative probability } F) \\
 &= \text{logit}^{-1}(\tau_1 - \beta X)
 \end{aligned}
 \tag{13}$$

$$\begin{aligned}
 P(Y = 2|X) &= \Pr(\tau_1 < \beta X + \epsilon \leq \tau_2|X) \\
 &= \Pr(\tau_1 - \beta X < \epsilon \leq \tau_2 - \beta X|X) \\
 &= F(\tau_2 - \beta X) - F(\tau_1 - \beta X) \\
 &= \text{logit}^{-1}(\tau_2 - \beta X) - \text{logit}^{-1}(\tau_1 - \beta X)
 \end{aligned}
 \tag{14}$$

And so on and so forth, for  $j$  up to  $J - 1$

## Ordered Logit

The last category  $J$

$$\begin{aligned} P(Y = J|X) &= P(\tau_{J-1} \leq \beta X + \epsilon|X) \\ &= P(\epsilon \geq \tau_{J-1} - \beta X|X) \\ &= 1 - P(\epsilon < \tau_{J-1} - \beta X) \\ &= 1 - F(\tau_{J-1} - \beta X) \\ &= 1 - \text{logit}^{-1}(\tau_{J-1} - \beta X) \end{aligned} \tag{15}$$

- We have written down  $P(Y|X)$  for every possible value of  $Y$ .
- Now we can use MLE to estimate parameters
- Now, there are regression coefficients  $\beta$ , as well as cutoffs  $\tau$
- Statistical software will return estimates for both

## Ordered Logit

- What do the cutoffs  $\tau$  mean?
- Recall that  $P(Y = 1|X) = \text{logit}^{-1}(\tau_1 - \beta X)$
- And  $P(Y = 2|X) = \text{logit}^{-1}(\tau_2 - \beta X) - \text{logit}^{-1}(\tau_1 - \beta X)$
- We add then together:

$$P(Y = 1|X) + P(Y = 2|X) = P(Y \leq 2|X) = \text{logit}^{-1}(\tau_2 - \beta X) \quad (16)$$

- And take the logit:

$$\text{logit}(P(Y \leq 2)) = \tau_2 - \beta X \quad (17)$$

- The rest is similar

$$\text{logit}(P(Y \leq j)) = \tau_j - \beta X$$

- In this way,  $\tau$  looks like intercepts in normal regressions; so some other software (R) call them intercepts

## Multinomial Logit: categorical outcome

- Multinomial logit: for categorical outcomes that have **no intrinsic** order
- We extend logistic regression in a different way
- $Y$  has  $J$  levels, from 0 to  $J - 1$
- For logistic regression,  $P(Y = 1|X) = \text{logit}^{-1}X\beta = \frac{\exp(X\beta)}{1 + \exp(X\beta)}$
- For multinomial logit, we make similar assumptions about  $P(Y = j|X)$

$$P(Y = j|X) = \text{logit}^{-1}X\beta_j = \frac{\exp(X\beta_j)}{1 + \sum_{j=1}^J \exp(X\beta_j)} \quad (18)$$

- And for reference group, its

$$P(Y = 0|X) = \text{logit}^{-1}X\beta_j = \frac{1}{1 + \sum_{j=1}^J \exp(X\beta_j)} \quad (19)$$

## Multinomial Logit

- For all levels except the reference group, it has its own regression coefficients
- Say we have 7 categories and 4 predictors (each of them is continuous), then in total we will have  $6 * 5 = 30$  coefficients
  - $6 = 7 - 1$
  - $5 = 4 + 1$  (plus intercepts)
- Also because we know what  $P(Y = j|X)$  looks like for every possible value of  $Y$ , we can use MLE to estimate  $\beta_j$

## Interpreting multinomial logit

- Based on the assumptions of multinomial, it is easy to see:

$$\frac{P(Y = j|X)}{P(Y = 0|X)} = \exp(X\beta_j) \quad (20)$$

- Therefore, one unit increase in  $X$  leads to  $\exp(\beta_j)$  increase in odds ratio of  $Y = j$  occurring, relative to  $Y = 0$



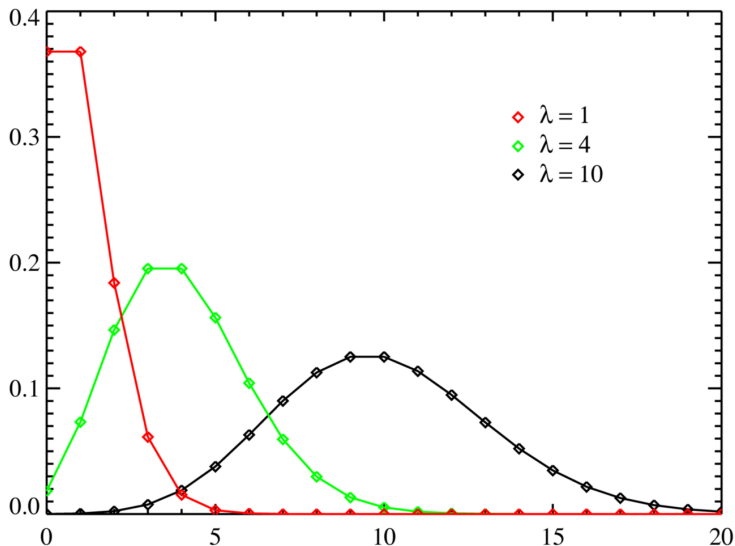
## Poisson Distribution

- Example:  $Y$  is event count
  - e.g., number of times each person visit a physician)
  - Number of new born / decease in a country
  - Usually small counts are more likely than large counts
- Key difference:  $Y$  are **non-negative integers**; in linear regression  $Y$  is assumed to be continuous variable between  $(-\infty, \infty)$
- Event count usually follows Poisson distribution

$$Pr(X = k) = \frac{\tau^k e^{-\tau}}{k!}$$

- $k! = k(k-1)(k-2) \cdots 1$  is factorial
- Property:  $E(X) = V(X) = \tau$

# Poisson Distribution



## Poisson Regression

- The conditional probability  $P(Y|X)$  is assumed to be distributed according to Poisson:

$$P(Y = y|X) = \frac{\exp(-\tau) \tau^y}{y!}, \quad y = 0, 1, 2, \dots \quad (21)$$

$$\tau = \exp(X\beta)$$

- And the conditional expectation  $E(Y|X)$  is given by:

$$E(Y|X) = \tau = \exp(X\beta) \quad (22)$$

## Poisson Regression (cont'd)

- Why don't we explicitly write  $E(\epsilon) = 0$  and  $E(\epsilon X) = 0$  as in the Assumption 1 and 2 of linear, logistic and probit regressions?
  - Hint: our assumption of the form of  $P(Y|X)$  is very strong
  - It directly gives what  $E(Y|X)$  should look like
  - And  $E(\epsilon) = 0$  and  $E(\epsilon X) = 0$  are essentially the property of  $\epsilon = Y - E(Y|X)$
  - So in many textbooks, when introducing generalized linear models, they will omit Assumptions 1 and 2, since it is implied by the assumption of the function form of  $P(Y|X)$
- Poisson assumption implies that the data is **heteroskedastic**:

$$\begin{aligned}
 V(\epsilon|X) &= V(Y - E(Y|X)|X) \\
 &= V(Y|X) \\
 &= \exp(X\beta)
 \end{aligned}
 \tag{23}$$

## Poisson and Log-Linear model

- Poisson regressions:

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

- An alternative way is to take **log** at both side of the equation

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

- It means that the **link function** of Poisson regression is **log**
- Sociologists and demographers call  $\log E(Y|X) = \log(\tau) = X\beta$  as **log-linear** model

## MLE for Poisson regression

$$1. P(Y = y|X) = \frac{\exp(-\tau)\tau^y}{y!}; \tau = \exp(X\beta)$$

$$2. \text{Likelihood is: } L = \prod_{i=1}^N \frac{\exp(-\tau_i)\tau_i^{y_i}}{y_i!}$$

- and log-likelihood is :

$$\sum_{i=1}^n y_i X_i' \beta - \exp(X_i' \beta) - \log y_i!$$

3. try to maximize by setting the derivative to be 0

$$\frac{\partial l}{\partial \beta} = \sum_{i=1}^n (y_i - \exp(X_i' \beta)) X_i = 0$$

- There is no closed-form solution, unfortunately. Numerical optimization is required.

## Interpretation of Poisson Regression

- In log-linear model format:

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

- One unit increase in  $X$  leads to  $\beta$  increase of the average of  $y$  in its **log scale**
- In Poisson regression format:

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

- One unit increase in  $X$  leads to  $\exp(\beta) - 1$  increase in  $Y$
- One unit increase in  $X$  multiplies the mean of  $Y$  by a factor  $\exp(\beta)$
- The ratio between the new  $Y$  and old  $Y$  is  $\exp(\beta)$ , on average

## Over-dispersion of Count Data

- Poisson regression assumes that  $P(Y|X)$  follows a Poisson distribution
- Recall that Poisson distribution assumes that the mean and the variance is the same
- Sometimes we have data whose variance is bigger than mean
- E.g., Long, J. Scott. 1990. *The Origins of Sex Differences in Science*. Social Forces. 68(3):1297-1316.
- The outcome is the number of published articles by a Ph.D. student in biochemistry
- The mean number of articles is 1.69 and the variance is 3.71, a bit more than twice the mean.
- Why? There are always super-starts :) and people who publish nothing : (



## Zero-inflated Poisson Regression

- One common situation of over-dispersion: there are a lot of zeros in the outcome  $Y$  and a few big values, which boosts the variance of outcome
- Example: civil war as outcome.
- Zero-inflated Poisson Regression is designed to address this issue
- It assumes that data has two generating processes
  1. With probability  $1 - \lambda$ , the data is generated according to Poisson with mean  $\tau$
  2. With probability  $\lambda$ , we generate excess zeros.
- The final conditional probability is

$$P(Y = y|X) = \lambda + (1 - \lambda) \frac{\exp(-\tau) \tau^y}{y!}$$

## Zero-inflated Poisson Regression (cont'd)

- With the assumptions in the previous slide
- $E(Y|X) = (1 - \lambda)\tau$
- $V(Y|X) = (1 - \lambda)\tau(1 + \tau\lambda)$
- $V$  is bigger than  $E$ , of a ratio of  $1 + \tau\lambda$
- Essentially, zero-inflated Poisson regression is the mix of two regressions:
  - One Poisson regression, with prob  $1 - \lambda$
  - One logistic regressions (0 and all others), with prob  $\lambda$
  - Each regression has its own coefficients
- So it is a more complex model than negative binomial regression, which adds only one additional parameter

## Negative binomial regression

- Another way to deal with over-dispersion: choose a different functional form about  $P(Y|X)$

$$P(Y = y|X) = \frac{\Gamma(\alpha + y)}{y! \Gamma(\alpha) (\tau + \alpha)^{\alpha+y}} \quad (24)$$

- And  $\tau = \exp(X\beta)$
- $\Gamma$  is Gamma function, an extension of factorial
- With this more complex parametric assumption
- $E(Y|X) = \tau$  (similar to Poisson regression)
- $V(Y|X) = \tau(1 + \frac{1}{\alpha}\tau)$
- Positive  $\alpha$  ensures that variance is bigger than the mean

## Other count data model

- Zero truncated regressions
  - Say, the outcome of the length of stay in a hospital, which is at least 1 day
  - Zero-truncated Poisson:
    - Remove the probability  $P(y = 0)$  because it's not possible
    - Re-scale the rest of the probability distribution to make it sums to 1

## How do we choose between models?

- Let us use our example of number of published articles by Ph.D. biochemists
- We can choose between three models:
  - Poisson regression
  - Negative binomial regression
  - Zero-inflated Poisson regression
- Decide whether or not to use Poisson regression is relative easier: (Cameron and Trivedi, "Regression-based tests for overdispersion in the Poisson model", *Journal of Econometrics*, 1990)
- Assume  $E(Y|X) = \tau$ , then
- Null Hypothesis:  $V(Y|X) = E(Y|X) = \tau$
- Alternative Hypothesis:  $V(Y|X) = \tau + c\tau$
- Cameron and Trivedi's overdispersion test just seeks to examine whether  $c = 0$
- (For R users: `dispersiontest` in AER package)

## Use Likelihood for Hypothesis Testing

- But how can we compare negative binomial regression vs zero-inflated Poisson regression?
- We can compare **Likelihood** among similar models to choose the best one
- Intuition:
  - Likelihood  $L$  represents the joint probability that we observe the entire data, given our parameters
  - Assume we have two models
  - A better model should have larger likelihood

## Likelihood Ratio Test

- Define Likelihood Ratio Test Statistics  $D$  as:

$$\begin{aligned} D &= -2 \log \frac{L_{\text{null}}}{L_{\text{alternative}}} \\ &= 2(\log L_{\text{alternative}} - \log L_{\text{null}}) \end{aligned} \tag{25}$$

- For comparing models, null model is often the simpler model, and alternative model is often the more complex model
- Null Hypothesis:  $D = 0$
- Alternative Hypothesis:  $D > 0$
- The bigger the  $D$ , the more evidence for the alternative model

## Likelihood Ratio Test (cont'd)

- Wilk's Theorem (1938):  $D$  has an  $\chi^2$ -distribution, with degrees of freedom equal to the difference in number of parameters between alternative model and the null model, if the null model is **nested** within the alternative model
- Nested basically means that the null model can be viewed as a simple case of the alternative model
  - e.g., null is logistic regression with 5 variables; alternative adds another variable
  - null is Poisson; alternative is negative binomial or zero-inflated Poisson
- For non-nested models, Wilk's Theorem does not hold; we need something else (shortly)



## Likelihood Ratio Test (cont'd)

- How do express Wilk's Theorem in the p-value language?
  - Say we get a  $D = 12$ , and the degree of freedom is 2
  - Definition: the probability of obtaining a test statistics that equals to  $D$  or higher is approximately  $p \iff$  p-value is  $p$
  - $P(D < 12, d.f. = 2) = 0.9975$ 
    - in R, just type `pchisq(12, 2)`, which is the cumulative probability distribution of  $D$
    - It means that the probability of observing a  $D$  smaller than 12 is 0.9975
    - So the probability we observe a  $D$  equal to or larger than 12 is  $1 - 0.9975 = 0.0025$ , which is our **p-value**)

## Bias-Variance Trade-Off and Likelihood Ratio Test

- But, a more complex model (adding more parameters) usually can predict more accurately and thus often always have larger likelihood
- AIC: Akaike information criterion (named after Hirotugu Akaike, 1974); reaching balance between predictive power and model complexity
- $k$  is the number of parameters in a model

$$AIC = 2k - 2 \log L \quad (26)$$

## Bootstrap

- So far we have used normal confidence interval to obtain confidence intervals and p-values
- These calculations requires asymptotically normal estimator
- The Bootstrap is an alternative approach to construct confidence intervals; one of the most important modern statistical concept (Efron, 1979)
  - reply on computer resampling; no math formula needed

### *Principle:*

1. *use the sample as if it is the population*
2. *draw samples from the (pseudo) population, and calculate quantity of interest*
3. *repeate 2 multiple times; we have a sampling distribution of the quantity of interest*

## Bootstrap example: confidence interval for mean

Assume we already have  $X_1, \dots, X_n$  be i.i.d. random samples of random variable  $X$ ). We are interested in estimating 95% confidence interval for sample mean  $\bar{X}$

1. Take a **with replacement** sample of size  $n$  from  $X_1, \dots, X_n$
2. Calculate the sample mean of the new sample,  $\bar{X}_1$
3. Repeat 1 and 2 for  $m$  times. We end up having  $m$  estimated means
4. Take the 2.5% and 97.5% quantile of the  $m$  estimated means. These two quantiles give us the bootstrap confidence intervals.

## Bootstrap example: logistic regression coefficients

- Example for calculating the confidence interval for logistic regression coefficients
  1. Take a **with replacement** sample of size  $n$  of the original data
  2. Estimate regression coefficient with this sample, and save it
  3. Repeat 1 and 2 for  $m$  times. We end up having  $m$  estimates of  $\beta$ ,  $(\hat{\beta}_1, \dots, \hat{\beta}_m)$
- With  $m$  estimated  $\beta$ , we can essentially approximate its probability density. It becomes easier to calculate every quantity:
  - $E(\hat{\beta})$  is approximated by the mean of  $(\hat{\beta}_1, \dots, \hat{\beta}_m)$
  - Standard error of  $\hat{\beta}$  is approximated by the standard error of  $(\hat{\beta}_1, \dots, \hat{\beta}_m)$
  - Take the  $\frac{1-\alpha}{2}$  and  $\frac{1+\alpha}{2}$  quantile of the values  $(\hat{\theta}_1, \dots, \hat{\theta}_m)$ . These two quantiles give us the bootstrap confidence intervals.

## Bootstrap example: more complex properties

- Sometimes we want to calculate confidence interval for more general quantity of interests that are a function of  $\beta$
- E.g., we want to get confidence interval for predicted probability:  $\hat{P}(Y = 1|X) = \frac{\exp(X\hat{\beta})}{1+\exp(X\hat{\beta})}$
- We know
  - lower bound of  $\beta$  is  $\hat{\beta}_{lower} = \hat{\beta} - 1.96\hat{se}$
  - upper bound of  $\beta$  is  $\hat{\beta}_{upper} = \hat{\beta} + 1.96\hat{se}$
- Is the confidence interval for predicted probability given by the normal approximated confidence interval?

$$\left( \frac{\exp(X\hat{\beta}_{lower})}{1 + \exp(X\hat{\beta}_{lower})}, \frac{\exp(X\hat{\beta}_{upper})}{1 + \exp(X\hat{\beta}_{upper})} \right)$$

- NO

## Bootstrap simulations

See `lec4_codes.Rmd` and PDF

## When do we use bootstrap?

- Bootstrap method is a very general method, can be used to calculate confidence intervals for most models you have seen in this class
- Bootstrap is often your last resort
  - It is slow (not possible until 80s)
  - The estimates are random
  - Less theoretical guarantee
  - But it nearly always work if you do not know how to calculate standard errors for some quantity of interests



## Today's Review

Type of Y	Regression to use
Continuous	linear
Binary	logit/probit
Categorical	multinomial logit / ordered logit
Count (integer)	Poisson, negative binomial and zero-inflated

## Recommended Readings

- More 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/>
- There are many other GLMs (e.g., **censored** outcome).
  - <https://data.princeton.edu/wws509>, Generalized Linear Models course by Germán Rodríguez
  - Powers, Daniel, and Yu Xie. *Statistical methods for categorical data analysis*. Emerald Group Publishing, 2008.