Lesson 15: Other types of categorical regression

Nicky Wakim

2025-05-21

Learning Objectives

1. Review Generalized Linear Models and how we can branch to other types of regression.
2. Identify outcome, examples, population model, and interpretations for different generalized linear models.

Learning Objectives

1. Review Generalized Linear Models and how we can branch to other types of regression.

2. Identify outcome, examples, population model, and interpretations for different generalized linear models.

Review: Generalized Linear Models (GLMs)

Generalized Linear Models

Random component

- Identify the response variable Y
- Specify a suitable (presumably) distribution for it

Systematic component

- Specify the explanatory variable(s) for the model

X's & β's

will mean something different

Link function

- Specify a functional form of $E(Y)$ that is related to the explanatory variables through a prediction equation in linear form

if $Y \sim \text{binom}(p)$

$Y \neq X\beta$

$\text{logit}(P(Y=1))$

GLM: Random Component

- The random component specifies the response variable Y and selects a probability distribution for it

- Basically, we are just identifying the distribution for our outcome
  - If Y is binary: assumes a binomial distribution of Y
  - If Y is count: assumes Poisson or negative binomial distribution of Y
  - If Y is continuous: assumea Normal distribution of Y

GLM: Systematic Component

- The systematic component specifies the explanatory variables, which enter linearly as predictors

cont Y

Y

$$\beta_0 + \beta_1 X_1 + \ldots + \beta_k X_k$$

- Above equation includes:
  - Centered variables ✓ $X - \mu_x$
  - Interactions ✓ $X_1 \times X_2$
  - Transformations of variables (like squares) ✓ $(X - \mu_x)^2$ or X^2

- Systematic component is the **same** as what we learned in Linear Models

GLM: Link Function

- If $\mu = E(Y)$, then the link function specifies a function $g(.)$ that relates μ to the linear predictors as:

$$g(\mu) = \beta_0 + \beta_1 X_1 + \ldots + \beta_k X_k$$

  - $g(\mu)$ is the transformation we make to $E(Y)$ (aka μ) so that the linear predictors (right side of equation) can be linked to the outcome
- The link function connects the random component with the systematic component
- Can also think of this as:

$$\mu = g^{-1}(\beta_0 + \beta_1 X_1 + \ldots + \beta_k X_k)$$

★ It's basically like saying $g(\mu)$ **IS** $\text{logit}(\mu)$ and thus

$g(\mu) = \text{logit}(\mu) = \text{logit}(P(Y=1|X))$

$$\underbrace{\text{logit}(\mu)}_{\text{link (informed by random)}} = \underbrace{\beta_0 + \beta_1 X_1 + \ldots + \beta_k X_k}_{\text{sys}}$$

link fn IS logit (μ)

GLM: Link Function

numeric data / cont

Dist'n of Y	Typical uses	Link name	Link function	Common name
512 Normal	Linear-response data	Identity	$g(\mu) = \mu$	Linear regression
Bernoulli / Binomial	outcome of single yes/no occurrence	Logit	$g(\mu) = \text{logit}(\mu)$	Logistic regression
Poisson	count of occurrences in fixed amount of time/space or rate	Log	$g(\mu) = \log(\mu)$	Poisson regression
Bernoulli / Binomial	outcome of single yes/no occurrence	Log	$g(\mu) = \log(\mu)$	Log-binomial regression
Multinomial	outcome of single occurence with K > 2 options, *nominal*	Logit	$g(\mu) = \text{logit}(\mu)$	Multinomial logistic regression
Multinomial	outcome of single occurence with K > 2 options, *ordinal*	Logit	$g(\mu) = \text{logit}(\mu)$	Ordinal logistic regression

Poll Everywhere Question 1

Search 09:41 Tue Jan 9 100%

Join by Web **PollEv.com/nickywakim275**

What is the purpose of a link function?

X's

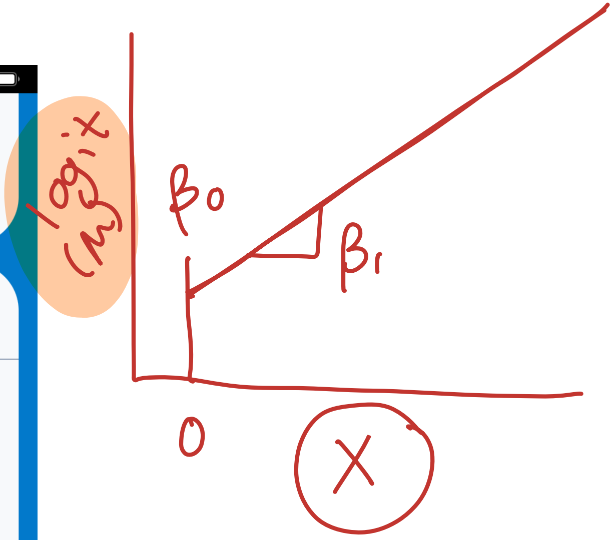
To transform the predictors/covaraites to a linear scale 0%

To make sure the outcome follows a normal distribution 0%

To make a transformation of the outcome so we can map our predictors to it ✓ 0%

To minimize the residual sum of squares. 0%

Powered by **Poll Everywhere**



Learning Objectives

1. Review Generalized Linear Models and how we can branch to other types of regression.

2. Identify outcome, examples, population model, and interpretations for different generalized linear models.

Linear regression

- **Outcome type:** continuous , numeric
- **Example outcomes:**
  - Height ✓
  - IAT score ✓
  - Heart rate ✓

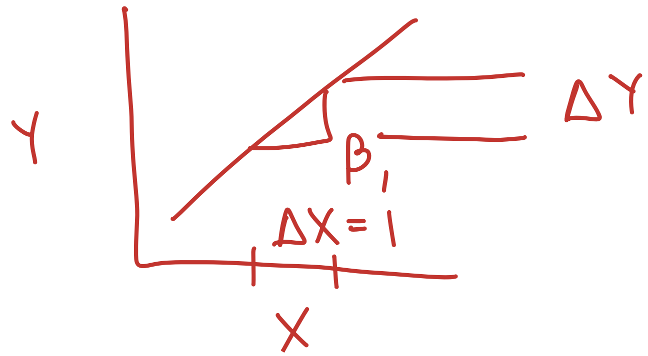
normal distribution

$Y \sim N(\mu, \sigma^2)$

- **Population model**

$$E(Y \mid X) = \mu = \beta_0 + \beta_1 X$$

- **Interpretations** of coeff. β_1
  - The change in average Y for every 1 unit increase in X



Linear regression: Process for data analysis

Model Selection

- Building a model
- Selecting variables
- Prediction vs interpretation
- Comparing potential models

Model Fitting

- Find best fit line
- Using OLS in this class
- Parameter estimation
- Categorical covariates
- Interactions

as X's

Model Evaluation

- Evaluation of model fit
- Testing model assumptions
- Residuals ✓
- Transformations
- Influential points ✓
- Multicollinearity ✓

Model Use (Inference)

- Inference for coefficients
- Hypothesis testing for coefficients
- Inference for expected Y given X
- Prediction of new Y given X

Logistic regression

- **Outcome type:** binary, yes or no
- **Example outcomes:**
  - Food insecurity ✓
  - Disease diagnosis for patient ✓
  - Fracture ✓

$Y \sim \text{binom}(p)$

- **Population model**

$$\text{logit}(\mu) = \text{logit}(\pi(X)) = \beta_0 + \beta_1 X$$

- **Interpretations** of β_1
  - The log-odds ratio for every 1 unit increase in X
  - $\exp(\beta_1)$ is odds ratio for every 1 unit increase in X

we can get risk ratio from this!

Logistic regression: Process for data analysis

Model Selection

- Build a model
- Select variables
- Prediction vs association
- Comparing potential models

Model Fitting

- Find model that maximizes likelihood function
- Parameter estimation (MLEs)
- Categorical covariates
- Interactions

Model Evaluation

- Evaluation of model fit
- Check model assumptions
- Transformations
- Influential points
- Multicollinearity
- Overdispersion

Model Use (Inference)

- Inference for odds ratios
- Hypothesis testing for odds ratios
- Inference for expected π given X
- Prediction of new Y given X

Log-binomial Regression

- Outcome type: binary, yes or no
- Example outcomes:
  - Food insecurity
  - Disease diagnosis for patient
  - Fracture

$Y \sim \text{binom}(p)$

prob can be > 1

$\mu \in [0, \infty)$

- Population model

$$\log(\mu) = \log(\pi(X)) = \beta_0 + \beta_1 X$$

- Interpretations of β_1
  - We have log of probability on the left
  - So exponential of our coefficients will be **risk ratio**

$\exp(\beta_1)$: risk ratio for every 1 unit inc in X

Poisson Regression

- **Outcome type:** Counts or rates

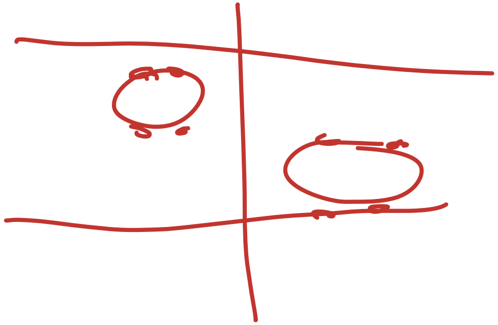
- **Example outcomes:**
  - Number of children in household
  - Number of hospital admissions
  - Rate of incidence for COVID in US counties

- **Population model**

mean
rate

$$\log(\mu) = \log(\lambda) = \beta_0 + \beta_1 X$$

- **Interpretations**
  - The count (or rate) ratio for every 1 unit increase in X



Count cars in 10 min
OR get rate cars/min

Multinomial logistic regression

- **Outcome type:** multi-level categorical, no inherent order

binary $\text{logit}(\pi) = \log\left(\frac{\pi}{1-\pi}\right)$ π yes π no

$\log\left(\frac{\mu_2}{\mu_1}\right) = \log\left(\frac{\pi_2}{\pi_1}\right)$

- **Example outcomes:**
  - Blood type —
  - US region (from WBNS)
  - Primary site of lung cancer (upper lobe, lower lobe, overlapped, etc.)
- We have additional restriction that the multiple group probabilities sum to 1

$P_{AB} + P_B + P_A + P_0 = 1$

- **Population models** 3 grp

$$\log\left(\frac{\mu_{\text{group 2}}}{\mu_{\text{group 1}}}\right) = \beta_0 + \beta_1 X$$

$$\log\left(\frac{\mu_{\text{group 3}}}{\mu_{\text{group 1}}}\right) = \beta_2 + \beta_3 X$$

$\log\left(\frac{\mu_{gp3}}{\mu_{gp2}}\right) = \beta_4 + \beta_5 X$

- **Interpretations**
  - Basically fitting two binary logistic regressions at same time!
  - β_1 First equation: how a one unit change in X changes the log-odds of going from group 1 to group 2
  - β_3 Second equation: how a one unit change in X changes the log-odds of going from group 1 to group 3

$\exp \beta_3 =$

Ordinal logistic regression

- Outcome type: multi-level categorical, with inherent order
- Example outcomes:
  - Satisfaction level (likert scale)
  - Pain level
  - Stages of cancer

logit(P(Y≤1))

P(Y>1)=
1−P(Y≤1)

- Population models , with levels $k = 1, 2, 3, \ldots, K$

$$\log\left(\frac{P(Y \leq 1)}{P(Y > 1)}\right) = \beta_0 + \beta_1 X$$

$$\log\left(\frac{P(Y \leq k)}{P(Y > k)}\right) = \beta_0 + \beta_1 X$$

- When these variables are predictors, we are pretty lenient about treating them as continuous
  - We must be VERY STRICT when they are outcomes
  - They do not meet the assumptions we place on continuous outcomes in linear regression!!
- We have additional restriction that the multiple group probabilities sum to 1

- Interpretations
  - Basically fitting K binary logistic regressions at same time!
  - First equation: how a one unit change in X changes the log-odds of going from group 1 to any other group
  - Second equation: how a one unit change in X changes the log-odds of going from group 1 or 2 to group 3 or above

Resources

Linear regression resources

- 512/612 class site!!
- Online textbook by Dr. Nahhas

Logistic regression resources

- Online textbook by Dr. Nahhas

Log-binomial Regression resources

- Online textbook by Dr. Nahhas
- Article on `logbin` package that is used to fit log-binomial regression

Ordinal logistic regression resources

- Online textbook by Dr. Nahhas
- Online textbook by Dr. Werth with data and R script

Poisson Regression resources

- PennState 504 website
- Online textbook by Dr. Nahhas
- YouTube video on R tutorial for Poisson Regression
  - Dr. Fogerty is a professor in Political Science, so just beware they may not have formal statistical training
- Guided R tutorial page on Poisson regression
- Online textbook by Dr. Werth
  - Social scientist, so just beware they may not have formal statistical training

Multinomial logistic regression resources

- YouTube video on R tutorial for Poisson Regression
  - Again, Dr. Fogerty is a professor in Political Science
- R-bloggers post with guided R code
- Online textbook by Dr. Werth with data and R script

Even more regressions...

Dist'n of Y	Typical uses	Link name	Link function	Common name
Bernoulli / Binomial	outcome of single yes/no occurrence	Probit	$g(\mu) = \Phi^{-1}(\mu)$	Probit regression
Bernoulli / Binomial	outcome of single yes/no occurrence	Complementary log-log	$g(\mu) = \log(-\log(1-\mu))$	Complementary log-log regression
Multinomial	outcome of single occurence with K > 2 options, *nominal*	Probit	$g(\mu) = \Phi^{-1}(\mu)$	Multinomial probit regression
Multinomial	outcome of single occurence with K > 2 options, *ordinal*	Probit	$g(\mu) = \Phi^{-1}(\mu)$	Ordered probit regression

More regression resources

- Probit regression
- Complementary log-log
- Multinomial probit
- Ordered probit

General resources

- Dr. Fogerty's YouTube series
- Dr. Werth's Categorical Book
- Dr. Nahhas' Book
- The Epidemiologist R Handbook
  - Analysis AND R work