Lesson 5: Simple Logistic Regression

Nicky Wakim

2024-04-15

Learning Objectives

1. Recognize why the tests we've learned so far are not flexible enough for continuous covariates or multiple covariates.
2. Recognize why ordinary linear regression cannot be applied to categorical outcomes with two levels
3. Identify the logistic regression model and define key notation in statistics language
4. Connect linear and logistic regression to the larger group of models, generalized linear model
5. Determine coefficient estimates using maximum likelihood estimation (MLE)

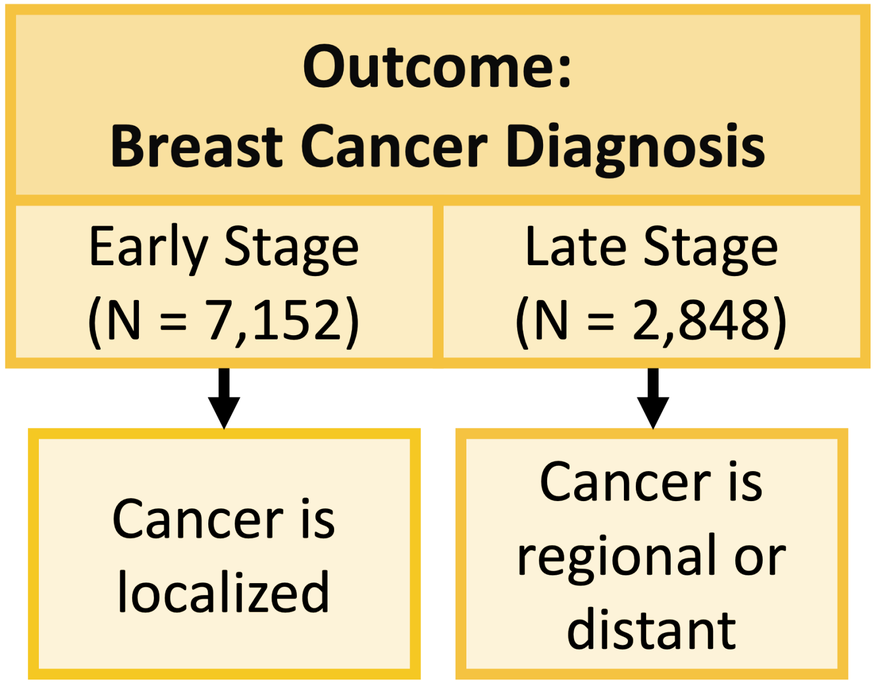
Health disparities in breast cancer diagnosis: working example

- **Question:** Is race/ethnicity and/or age associated with an individual's diagnosed stage of breast cancer?
  - For now, consider each covariate separately
- **Population:** individuals who are assigned female at birth who have been diagnosed with breast cancer in the United States
- Data from the Surveillance, Epidemiology, and End Results (SEER) Program (2014-2018)

Please note that this question has been answered

- You can take a look at the Breast Cancer Research Foundation's page: Understanding Breast Cancer Racial Disparities
- Big contributors to racial disparities include:
  - Underrepresentation in clinical trials
  - Access to healthcare
  - More aggressive cancers more likely in people of Native American, African, Hispanic, and Latin American descent
- Our analysis will not be new, but this kind of work has shed light on the importance of focused research on people of color
  - Dr. Davis focuses research on genomics and tumor microenvironment in African and African American patients
  - Dr. Ambrosone focuses research on how immune cells differ between patients. Specifically on the DARC gene, which is an evolved gene that helps fight malaria, that is found at a higher rate in people with African descent.

Example: Health disparities in breast cancer diagnosis (1/2)



Example: Health disparities in breast cancer diagnosis (2/2)

Outcome: Breast Cancer Diagnosis

Early Stage (N = 7,152)	Late Stage (N = 2,848)
↓	↓
Cancer is localized	Cancer is regional or distant

Covariate: Age

Continuous,
range: 22-75 years old

Covariate: Race/Ethnicity

Hispanic-Latinx (N = 846)	Non-Hispanic American Indian/Alaska Native (N = 23)
Non-Hispanic Asian/Pacific Islander (N = 790)	Non-Hispanic Black (N = 1,040)
Non-Hispanic White (N = 7,301)	

Poll Everywhere Question 1

How do we determine differences in diagnosis? (1/2)

- Breast cancer diagnosis study: two variables that are categorical
- We could use a contingency table (or two-way table)

	Breast Cancer Diagnosis		
Race/Ethnicity	**Early Stage**	**Late Stage**	**Total**
Non-Hispanic White	5,321	1,980	7,301
Non-Hispanic Black	683	357	1,040
Non-Hispanic Asian/Pacific Islander	556	234	790
Hispanic-Latinx	575	271	846
Non-Hispanic American Indian/Alaska Native	17	6	23
Total	7,152	2,848	10,000

How do we determine differences in diagnosis? (2/2)

- Contingency table does not work for…
  - Continuous covariates
  - Multiple covariates

- **Logistic regression models can handle multiple covariates that are continuous or categorical**

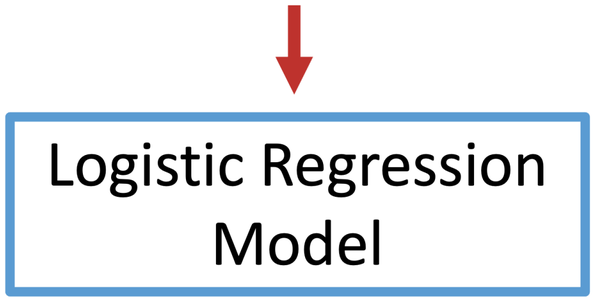
Individual #	Diagnosis stage	Race/Ethnicity
1	Early	Non-Hispanic Black
2	Early	Non-Hispanic White
3	Late	Non-Hispanic Asian/Pacific Islander
4	Early	Hispanic-Latinx
…		

How do we determine differences in diagnosis? (2/2)

- Contingency table does not work for…
  - Continuous covariates
  - Multiple covariates

- **Logistic regression models can handle multiple covariates that are continuous or categorical**

Individual #	Diagnosis stage	Race/Ethnicity
1	Early	Non-Hispanic Black
2	Early	Non-Hispanic White
3	Late	Non-Hispanic Asian/Pacific Islander
4	Early	Hispanic-Latinx
…		



Learning Objectives

Reference for individual overview

Building towards simple logistic regression

- **Goal:** model the probability of our outcome ($\pi(X)$) with the covariate (X_1)
- In simple linear regression, we use the model in its various forms:

$$
\begin{aligned}
Y &= \beta_0 + \beta_1 X_1 + \epsilon \\
E[Y|X] &= \beta_0 + \beta_1 X_1 \\
\widehat{Y} &= \beta_0 + \beta_1 X_1
\end{aligned}
$$

- Potential problem? Probabilities can only take values from 0 to 1

Simple Logistic Regression Model: Components

Can we apply OLS to our binary outcome?

- Let’s see if we can apply OLS/linear regression to our binary outcome
- What assumptions do our data need to meet in order to use OLR?
- Let’s review OLR assumptions!

Review of simple linear regression(1/2)

The (population) regression model is denoted by:

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

Components

Y	response, outcome, dependent variable
β_0	intercept
β_1	slope
X	predictor, covariate, independent variable
ϵ	residuals, error term

Review of simple linear regression (2/2)

- Assumptions of the linear regression model:
  - Independence: observations are independent
  - Linearity: linear relationship between $E[Y|X]$ and X

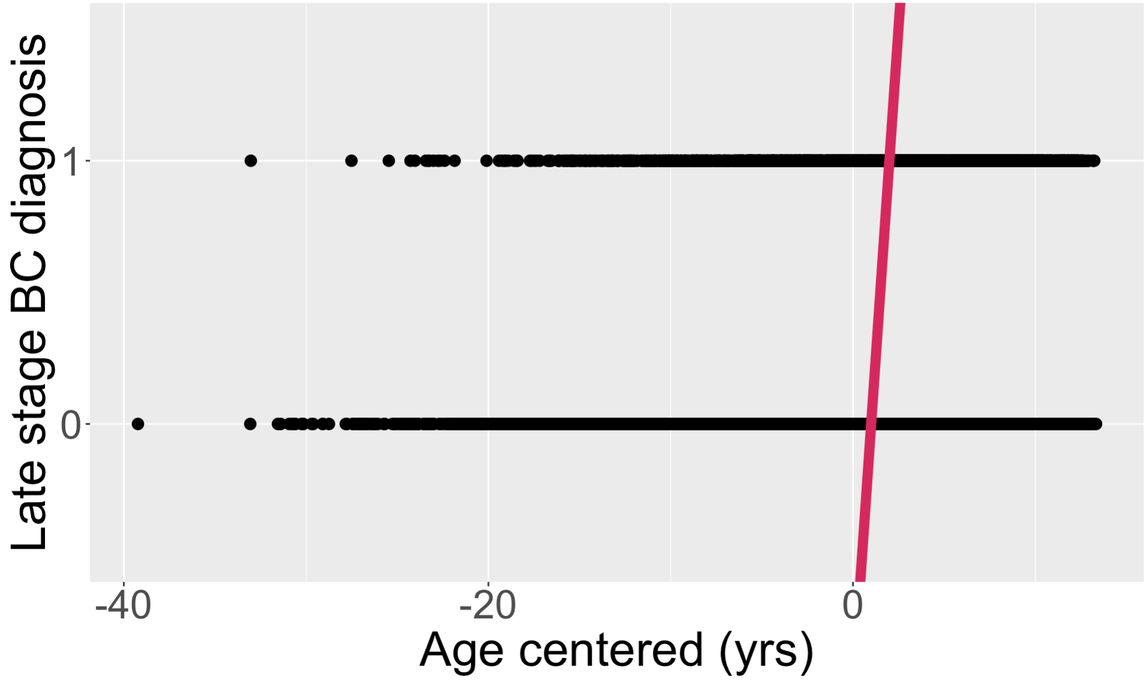
$$E[Y|X] = \beta_0 + \beta_1 \cdot X$$

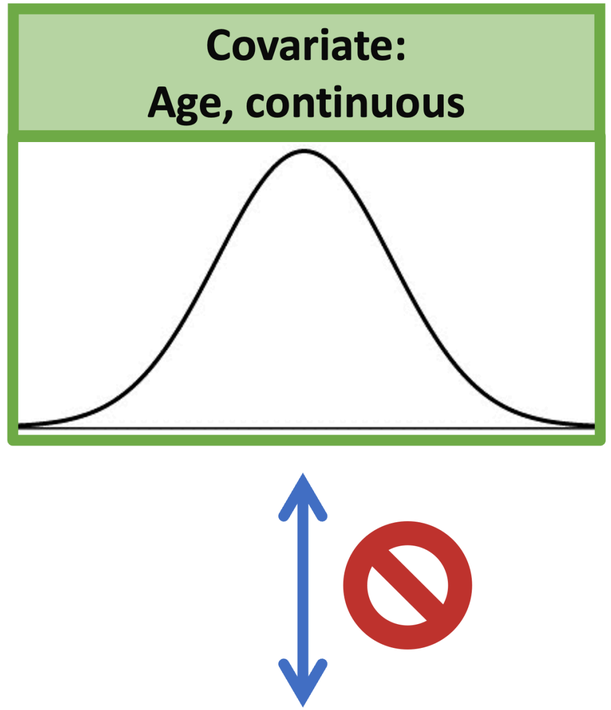
  - Normality and homoscedasticity assumption for residuals (ϵ):
    - Normality: residuals are normally distributed
    - Homoscedasticity (equal variance): Variance of Y given X ($\sigma^2_{Y|X}$), is the same for any X

- Which assumptions are violated if dependent variable is categorical?
  - Think in terms of binary dependent variable

Violated: Linearity

- The relationship between the variables is linear (a straight line):
  - $E[Y|X]$ or $\pi(X)$, is a straight-line function of X
- The independent variable X can take any value, while $\pi(X)$ is a probability that should be bounded by [0,1]
  - We cannot use linear mapping to translate X to $\pi(X)$





Violated: Normality

- In linear regression, ϵ is distributed normally

- Recall that Y can take only one of the two values: 0 or 1
- And the fitted Y, $\widehat{Y}$ can also only take values 0 or 1
- Thus, $\epsilon = Y - \widehat{Y}$ can only take values -1, 0, or 1

- Then ϵ **cannot follow a normal distribution**, which would require ϵ to have a continuum of values and no upper or lower bound

Y	$\widehat{Y}$	
	0	1
0	$\epsilon = 0 - 0 = 0$	$\epsilon = 0 - 1 = -1$
1	$\epsilon = 1 - 0 = 1$	$\epsilon = 1 - 1 = 0$

Violated: Homoscedasticity

- In linear regression, $\text{var}(\epsilon) = \sigma^2$
  - Variance does not depend on X

- When Y is a binary outcome

$$\begin{aligned}\text{var}\left(Y\right) &= \pi\left(1-\pi\right) \\ &= \left(\beta_0 + \beta_1 X\right)\left(1-\beta_0-\beta_1 X\right)\end{aligned}$$

  - Variance depends on X

- Because variance depends on X: no homoscedasticity
  - Variance will not be equal across X-values

What happens if we use OLR for categorical responses?

Simple Logistic Regression Model: Components

Poll Everywhere Question 2

How do we fix these violations?

- **Question:** How do we manipulate our response variable so that we fix these violations?

- **Answer:** We need to *transform the outcome* so we can map differences in covariates to the two levels
  - Will discuss in a few slides: called link function

How do we transform our outcome?

Stage at diagnosis
Y

→

Probability of Late-stage diagnosis
$\pi = P(Y = 1)$

→

Logit probability
$\text{logit}(\pi) = \ln\left(\frac{\pi}{1-\pi}\right)$

Two levels:

$Y = 0$

$Y = 1$

Range of probabilities:

$0 \leq \pi \leq 1$

Range of logit values:

$-\infty \leq \text{logit}(\pi) \leq \infty$

Note: people use π (or p) to mean $\pi(X)$

Simple Logistic Regression Model

The (population) regression model is denoted by:

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

Components

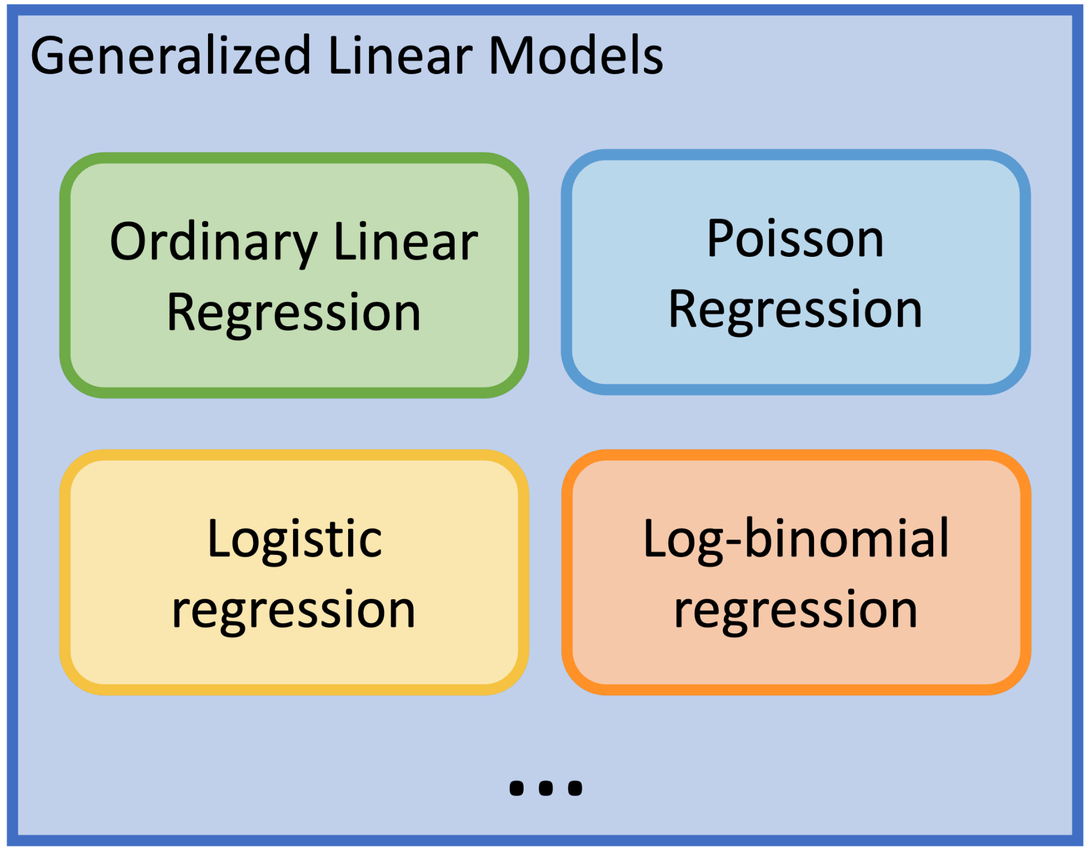
π	probability that the outcome occurs ($Y = 1$) given X
β_0	intercept
β_1	slope
X	predictor, covariate, independent variable
ϵ	residuals, error term

Learning Objectives

Generalized Linear Models (GLMs) (1/2)

- **Generalized Linear Models** are a class of models that includes regression models for **continuous and categorical responses**
  - Responses follow *exponential family distribution*

- Here we will focus on the GLMs for **categorical/count data**
  - **Logistic regression** is just a one type of GLM
  - **Poisson regression** – for counts
  - **Log-binomial** can be used to focus on risk ratio



Poll Everywhere Question 3

Generalized Linear Models (GLMs) (2/2)

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

Link function

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

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

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

- Above equation includes:
  - Centered variables
  - Interactions
  - Transformations of variables (like squares)

- 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 predictor 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:

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

GLM: Link Function

Link	Link function	Type of response variable	Type of regression
Identity link	$g(\mu) = 1 \times \mu$	Continuous response variables	Linear regression
Log link	$g(\mu) = \log(\mu)$	Discrete count response variable	Poisson regression
Logit link	$g(\mu) = \text{logit}(\mu) = \log\left[\frac{\mu}{1-\mu}\right]$	Categorical response variable	Logistic regression
Log link	$g(\mu) = \log(\mu)$	Categorical response variable	Log-binomial regression

Simple Logistic Regression Model

Learning Objective

Estimation for Logistic Regression Model

Poll Everywhere Question 5

How to find Maximum Likelihood Estimator (MLE)?

Construct a likelihood function for an individual

Construct the likelihood function across the sample

Convert to log-likelihood

Find MLEs that maximize log-likelihood

How do we do this in R?

Learning Objectives