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 linear regression cannot be applied to categorical outcomes with two levels
3. Identify the simple 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 GLMs
5. Determine coefficient estimates using maximum likelihood estimation (MLE) and apply it in R

OLS in lm()
MLE in glm()

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 linear regression cannot be applied to categorical outcomes with two levels
3. Identify the simple 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) and apply it in R

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 to better serve people of color who develop breast cancer
  - 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)

Outcome: Breast Cancer Diagnosis	
Early Stage (N = 7,152)	Late Stage (N = 2,848)
Cancer is localized	Cancer is regional or distant

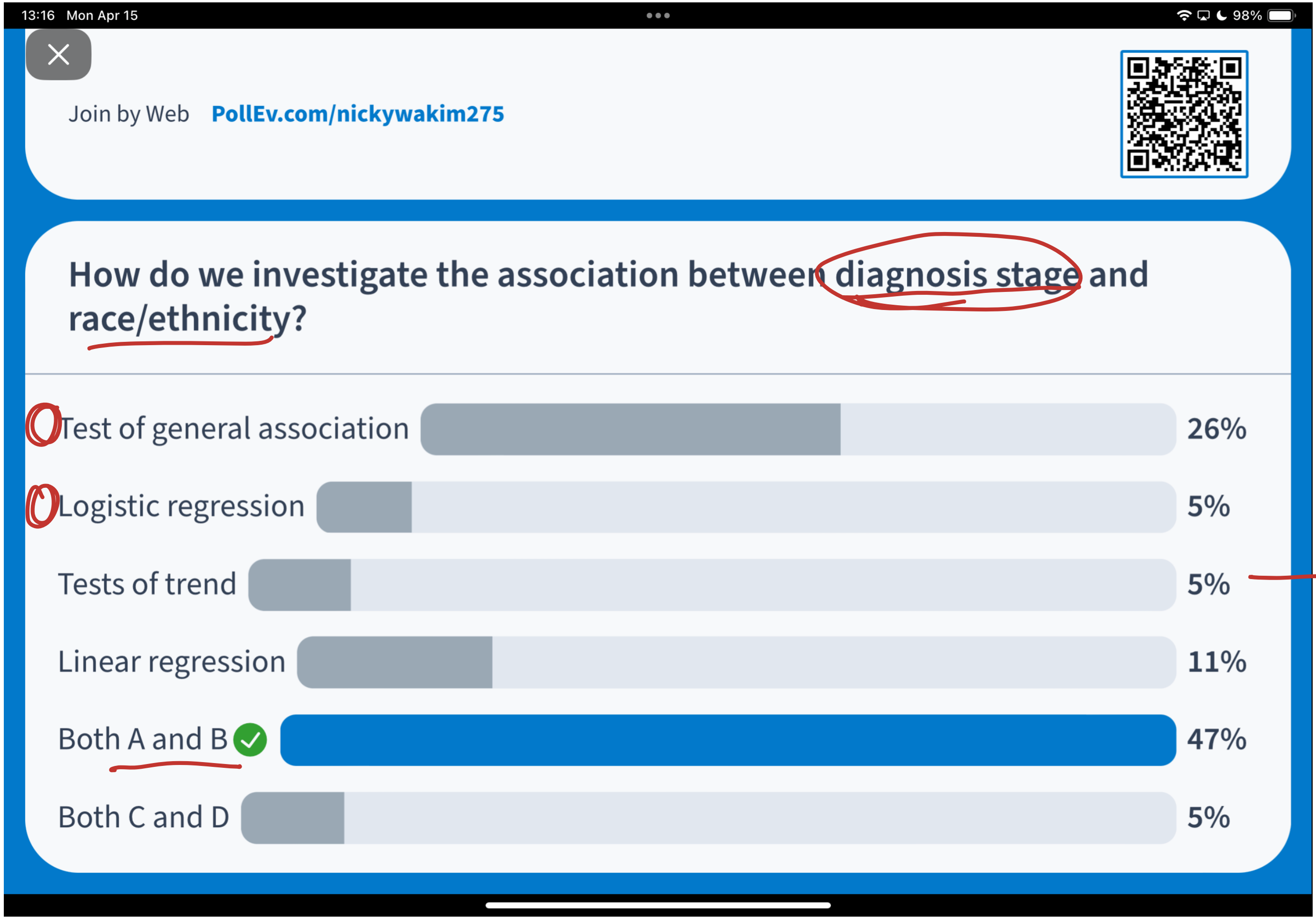
Example: Health disparities in breast cancer diagnosis (2/2)

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

    >3

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

Individual #	Diagnosis stage	Race/Ethnicity	Age
1	Early	Non-Hispanic Black	71
2	Early	Non-Hispanic White	35
3	Late	Non-Hispanic Asian/Pacific Islander	59
4	Early	Hispanic-Latinx	68
…			

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	Age
1	Early	Non-Hispanic Black	71
2	Early	Non-Hispanic White	35
3	Late	Non-Hispanic Asian/Pacific Islander	59
4	Early	Hispanic-Latinx	68
…			

↓

Logistic Regression Model

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 linear regression cannot be applied to categorical outcomes with two levels

3. Identify the simple 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) and apply it in R

Reference: individual components

General form	Breast Cancer Diagnosis Example
i	Individual who has been diagnosed with breast cancer in the United States
$Y_i = 1$	Individual i received a late-stage diagnosis of breast cancer
$Y_i = 0$	Individual i did not receive a late-stage diagnosis of breast cancer (early-stage diagnosis)
$P(Y_i = 1 \| X_i) = \boxed{\pi(X_i)}$	Probability that individual receives a late-stage diagnosis of breast cancer given their observed covariates
$P(Y_i = 0 \| X_i) = 1 - \pi(X_i)$	Probability that individual does not receive a late-stage diagnosis of breast cancer given their observed covariates
$X_{1,i}$	$R\_E_{1,i}$ (race/ethnicity) for individual i -or- Age_i for individual i

Building towards simple logistic regression

1st covariate

- **Goal:** model the probability of our outcome ($\pi(X)$) with the covariate (X_1)

not for i

- In simple linear regression, we use the model in its various forms:

$$Y = \beta_0 + \beta_1 X_1 + \epsilon$$

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

$$E[Y|X] = \beta_0 + \beta_1 X_1$$

$$\widehat{Y} = \beta_0 + \beta_1 X_1$$

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

$[0, 1]$

Simple Logistic Regression Model: Components

- Outcome: Y - binary (two-level) categorical variable
  - $Y = 1$
  - $Y = 0$

Y= Late stage

Y= early stage

- Covariate: X_1
  - For today: simple logistic regression with one covariate
  - X_1 can be continuous or categorical

- Probability of outcome for individual with observed covariates
  - $P\left(Y = 1|X\right) = \pi\left(X\right)$
  - $P\left(Y = 0|X\right) = 1 - \pi(X)$

- Because the expected value is a weighted average, we can say:

$$E(Y|X) = \sum_{j=1}^{2} P(Y|X)\cdot Y$$

$$\begin{aligned} E(Y|X) &= \overset{0.5}{P(Y = 1|X)} \cdot 1 + \overset{0.5}{P(Y = 0|X)} \cdot 0 \\ &= P(Y = 1|X) \cdot 1 \\ &= P(Y = 1|X) \\ &= \pi(X) \end{aligned}$$

$P(Y=0) = 1 - P(Y=1)$

$P(Y=0) + P(Y=1) = 1$

  - For categorical outcomes, $\pi(X)$ (or π for shorthand), is more widely used than $E(Y|X)$

$\pi \cdot 1 + (1-\pi)0$

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 OL~~R~~S?
- Let's review OL~~R~~S 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

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

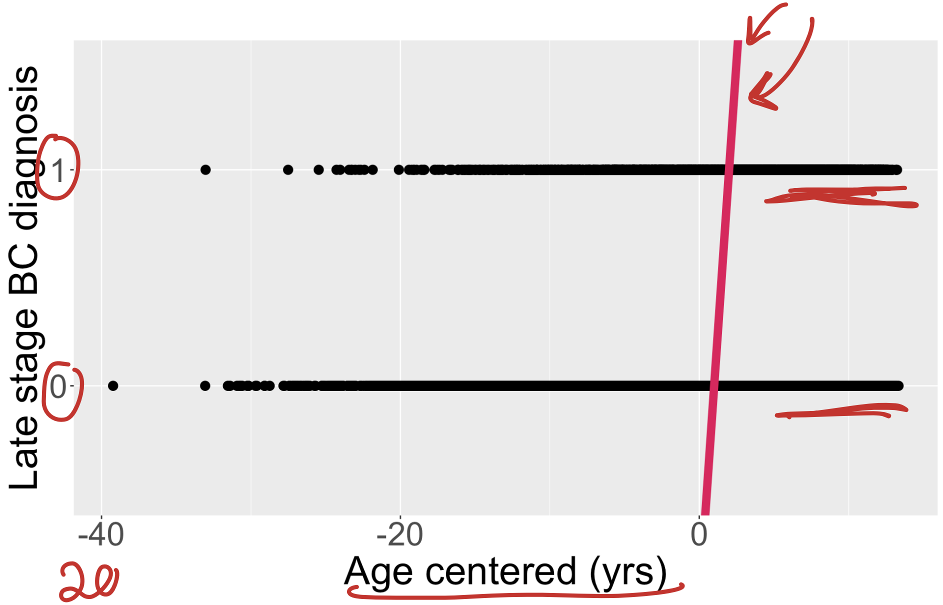
  - Normality and homoscedasticity assumption for residuals (ϵ):
    - Normality: residuals are normally distributed $\longrightarrow$ $\varepsilon \sim N(0, \sigma^2)$
    - Homoscedasticity (equal variance): Variance of Y given X ($\sigma^2_{Y|X}$), is the same for any X

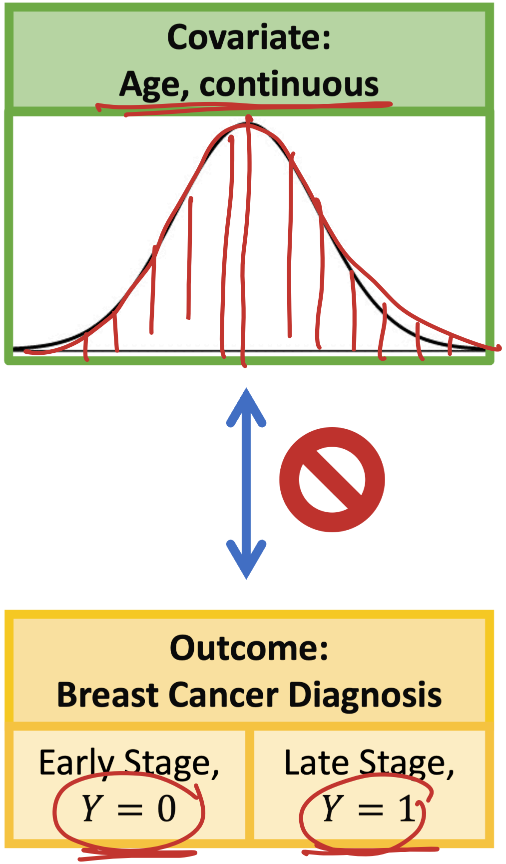
is not a fn of X

- 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

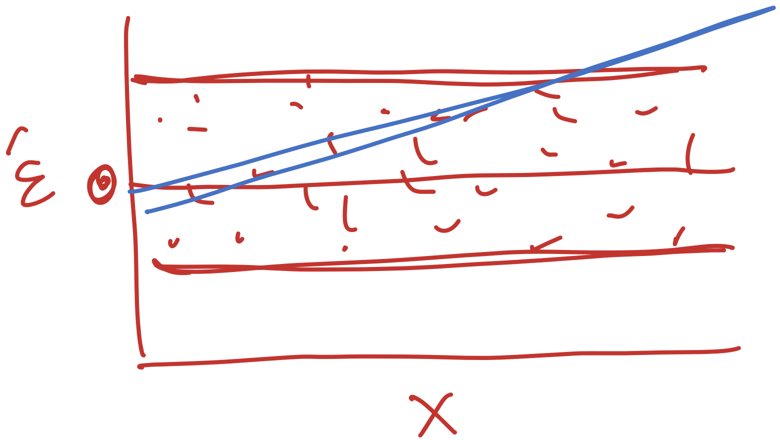
obs Y	fitted $\hat{Y}$	
	0	1
0	$\epsilon = 0 - 0 = 0$	$\epsilon = 0 - 1 = -1$
1	$\epsilon = 1 - 0 = 1$	$\epsilon = 1 - 1 = 0$

Violated: Homoscedasticity

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

- When Y is a binary outcome

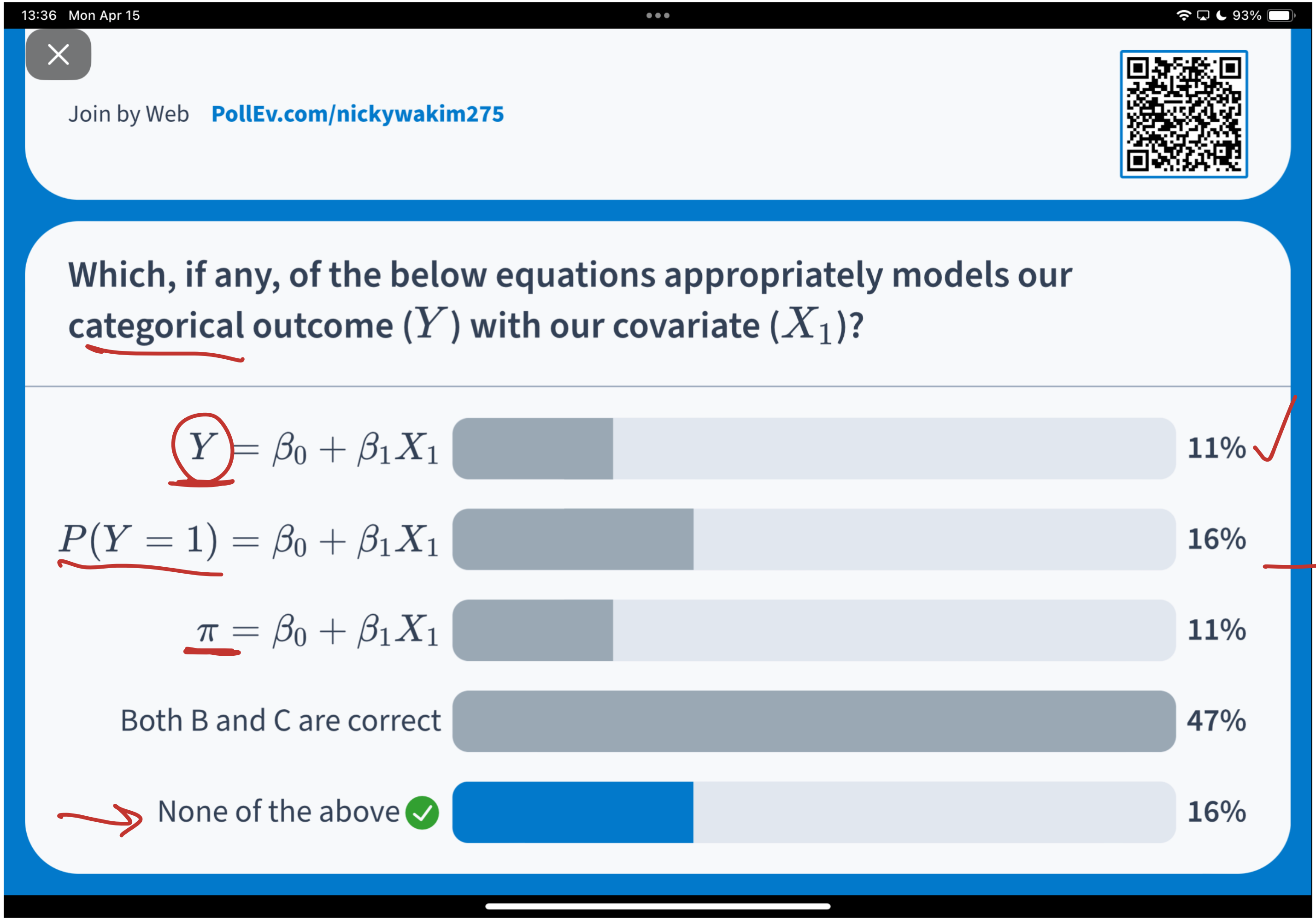
$$\begin{aligned} \mathrm{var}\,(Y) &= \pi(1-\pi) \\ &= \underbrace{(\beta_0 + \beta_1 X)}_{\pi}\,\underbrace{(1 - \beta_0 - \beta_1 X)}_{\pi} \end{aligned}$$



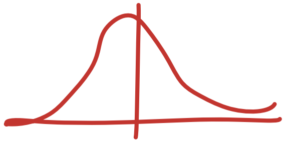
  - Variance depends on X

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

Poll Everywhere Question 2



$\rightarrow [0,1]$



$-\infty$ ∞

Need a transform

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 linear regression cannot be applied to categorical outcomes with two levels
3. Identify the simple 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) and apply it in R

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? (1/2)

log-odds

Stage at diagnosis
Y

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

Logit probability
$\text{logit}(\pi) = \underset{\text{log}}{\ln}\Big(\underbrace{\frac{\pi}{1-\pi}}_{\text{odds}}\Big)$

Two levels:

$Y = 0$

$Y = 1$

Range of probabilities:

$0 \leq \pi \leq 1$

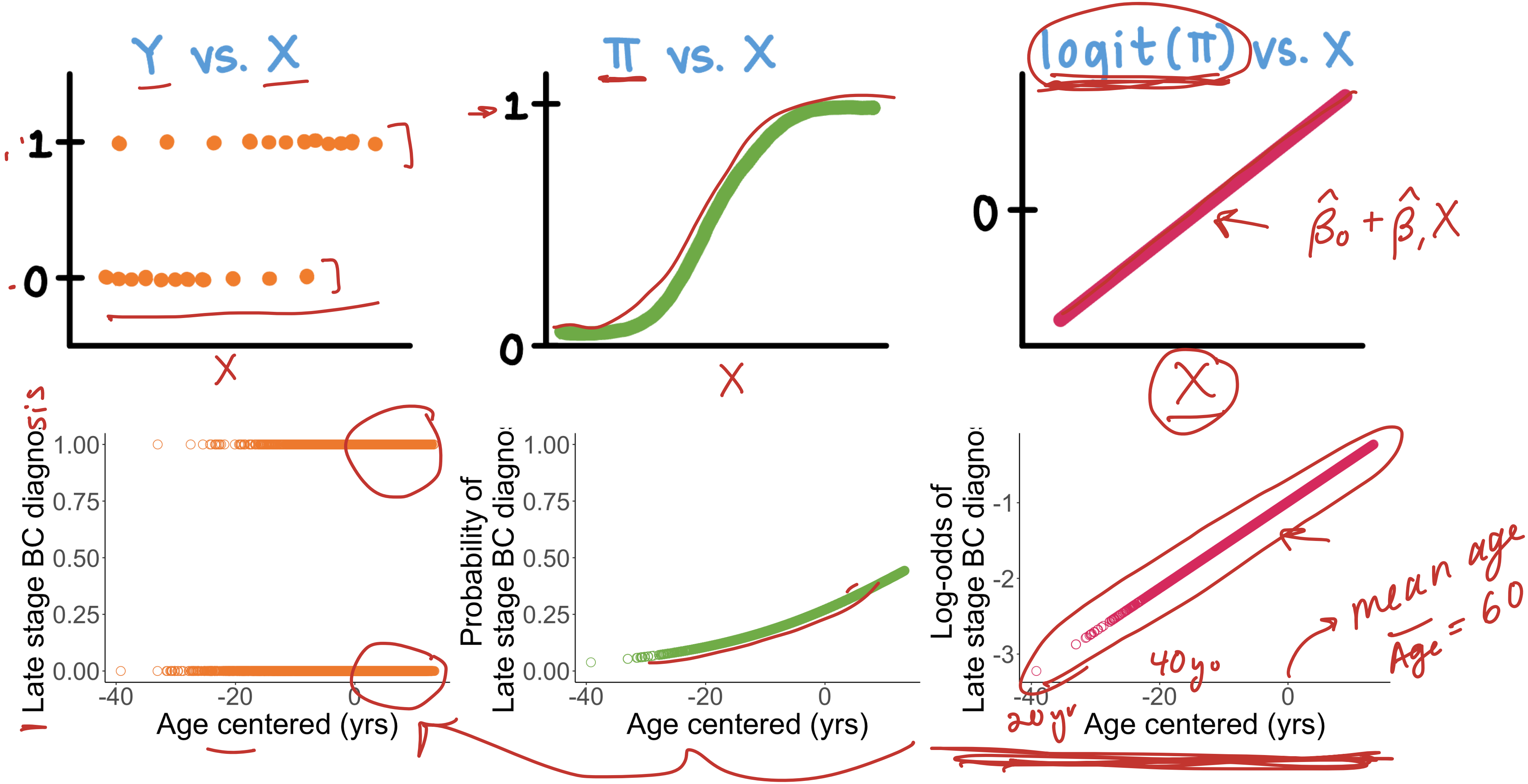
Range of logit values:

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

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

$\log(\pi)$ $(-\infty, 0]$

How do we transform our outcome? (2/2)



Simple Logistic Regression Model

The (population) regression model is denoted by:

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

Components

π	probability that the outcome occurs ($Y = 1$) given X
β_0	intercept
β_1	slope
X	predictor, covariate, independent variable
$\cancel{\epsilon}$	~~residuals, error term~~

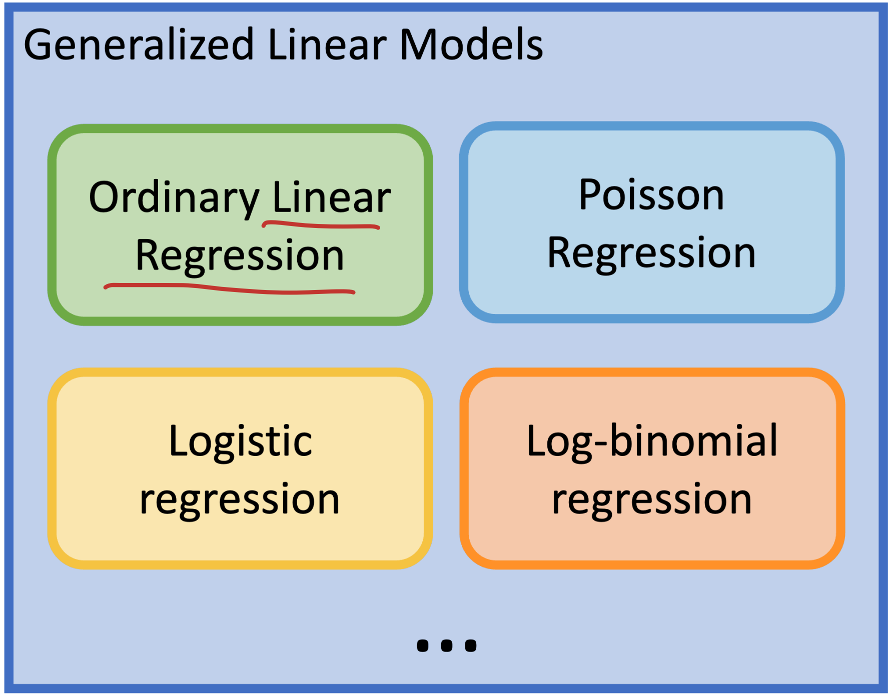
Learning Objectives

1. Recognize why the tests we've learned so far are not flexible enough for continuous covariates or multiple covariates.
2. Recognize why linear regression cannot be applied to categorical outcomes with two levels
3. Identify the simple 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) and apply it in R

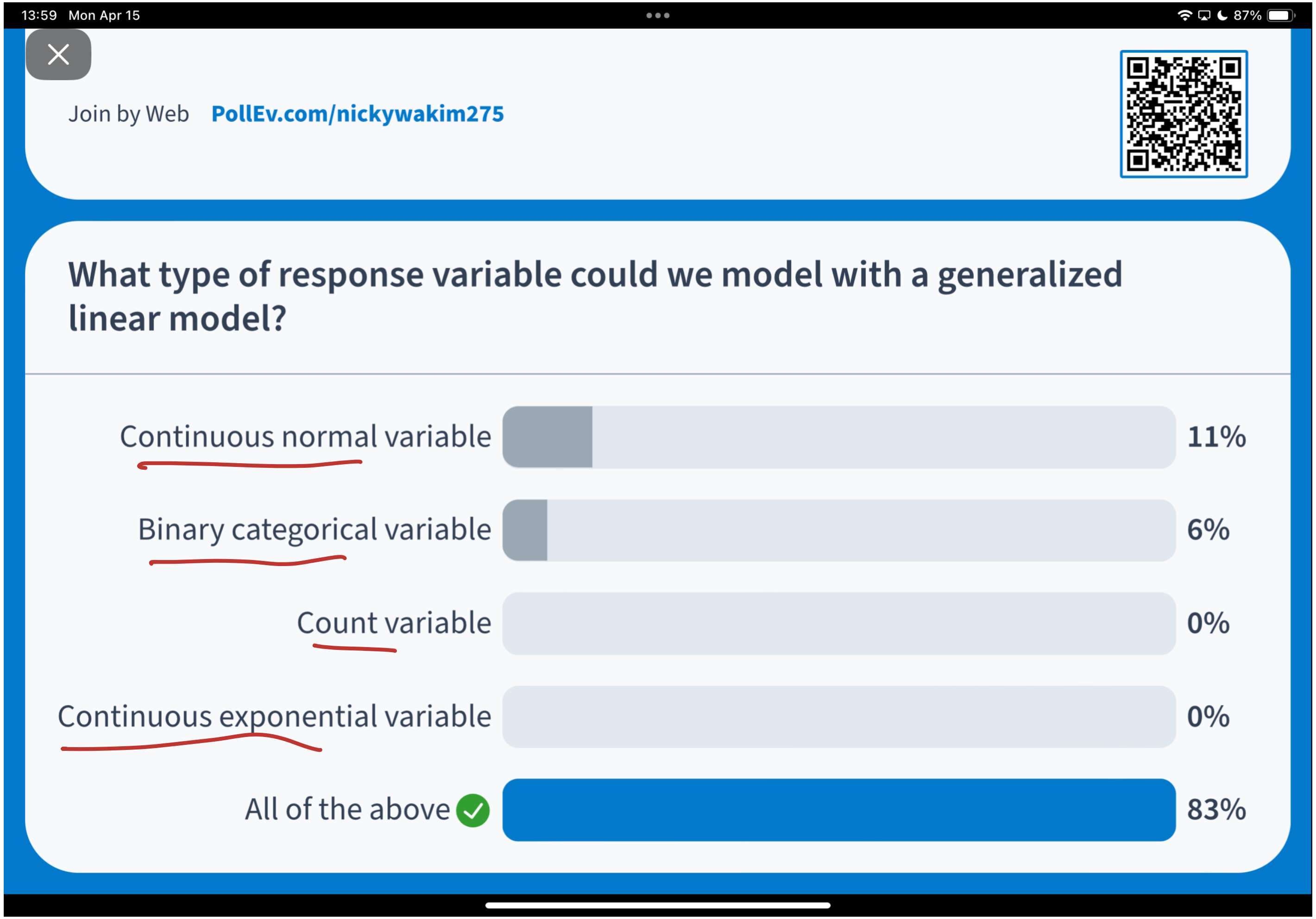
Generalized Linear Models (GLMs) (1/2)

- **Generalized Linear Models** are a class of models that includes regression models for **continuous and categorical responses**
  - Responses follow *exponential family distribution*
  - Helps us set up other types of regressions using each outcome's needed transformations

- 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	Systematic component	Link function
• Identify the response variable Y • Specify a suitable (presumably) distribution for it	• Specify the explanatory variable(s) for the model	• 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~~ identifies 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

$$\boxed{} = \beta_0 + \beta_1 X_1 + \ldots + \beta_k X_k$$

- 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:

full

link

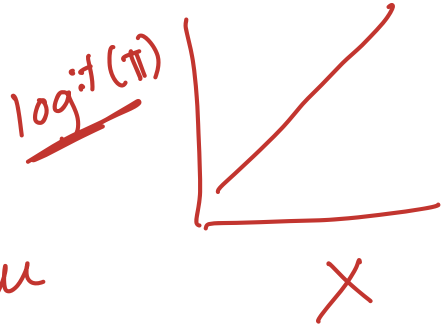
$$g(\mu) = \beta_0 + \beta_1 X_1 + \ldots + \beta_k X_k$$

  - $g(\mu)$ 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

(π)

- The link function connects the random component with the systematic component
- Can also think of this as:

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



g(μ)   identity link for cont Y ($12)   g(μ) = 1 · μ

g(μ)/g(π)   logit link   g(π) = logit(π)

GLM: Link Function

left side of model

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	$g(\mu) = \text{logit}(\mu) = \log\boxed{\left[\frac{\mu}{1-\mu}\right]}$	Categorical response variable	Logistic regression
Log link	$g(\mu) = \log(\mu)$	Categorical response variable	Log-binomial regression

Risk ratio

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 linear regression cannot be applied to categorical outcomes with two levels
3. Identify the simple 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) and apply it in R

Reminder: 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~~

Estimation for Logistic Regression Model

- Same as linear regression model: we need to estimate the values of β_0 and β_1

- Maximum likelihood: yields values for the unknown parameters that *maximize the probability of obtaining observed set of data*
  - In linear regression, this leads to least squares estimation
  - Maximum likelihood estimators (MLE): values of parameters that maximize likelihood

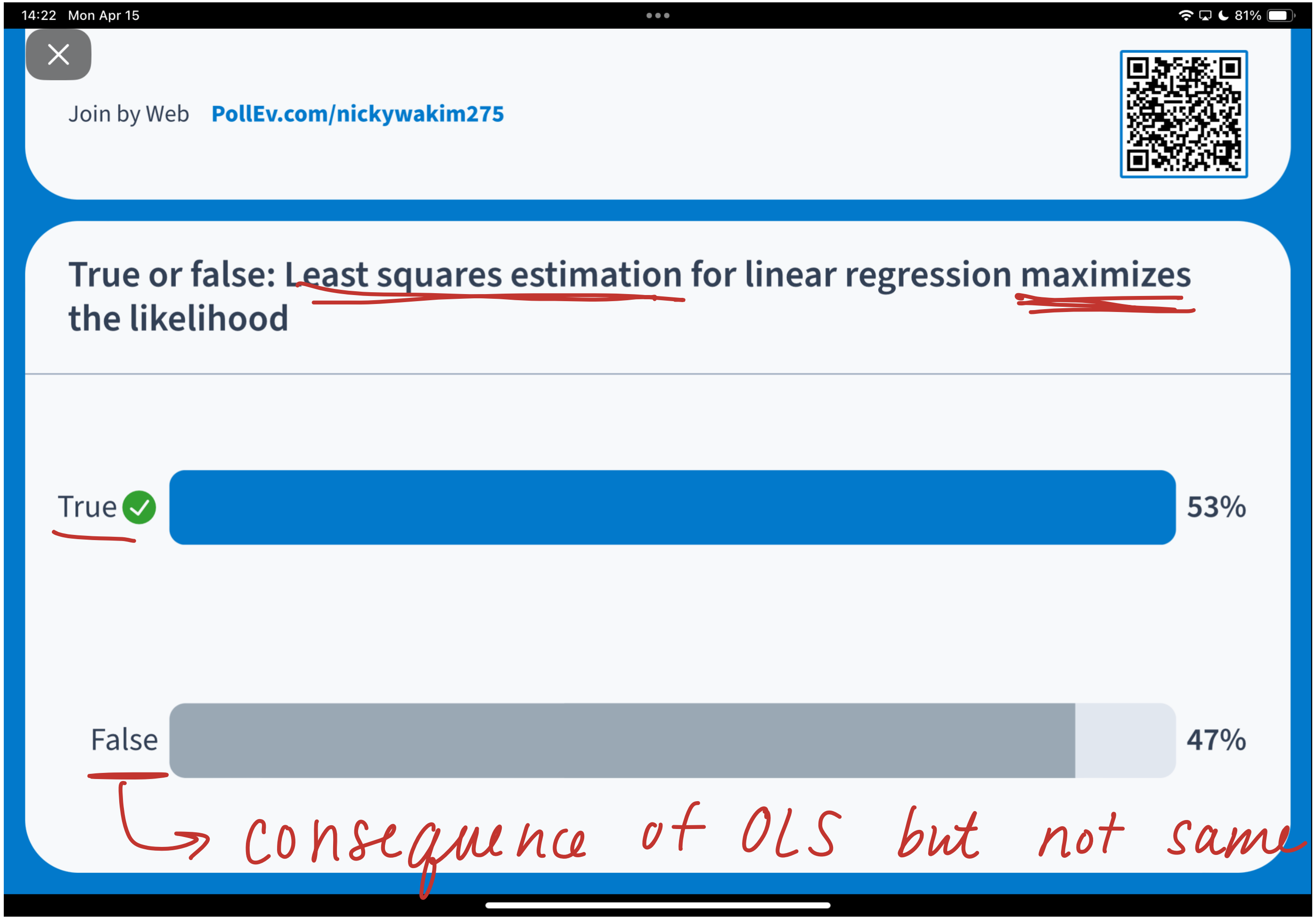
$\to \hat{\beta}_0, \hat{\beta}_1$

512: $\hat{\beta}_{0,OLS}$ $\hat{\beta}_{1,OLS}$

513: $\hat{\beta}_{0,MLE}$ $\hat{\beta}_{1,MLE}$

- Likelihood function: expresses the probability of the observed data as a function of the unknown parameters

Poll Everywhere Question 4



consequence of OLS but not same procedure
(IN LIN REG ONLY)

How to find Maximum Likelihood Estimator (MLE)?

1. Construct a likelihood function for an individual

2. Construct the likelihood function across the sample

3. Convert to log-likelihood

estimates

4. Find parameter ~~values~~ that maximize log-likelihood (MLEs)

$\hat{\beta}_0, \hat{\beta}_1$

1. Construct a likelihood function for an individual

- Within a dataset with n subjects, for the ith subject:
  - if $Y_i = 1$, the contribution to the likelihood function is $\pi\left(X_i\right)$
  - if $Y_i = 0$, the contribution to the likelihood function is $1 - \pi\left(X_i\right)$

$P(Y=0)$

- The **contribution from the ith subject** to the likelihood function can be expressed as:

$$\pi(X_i)^{Y_i}\left[1 - \pi\left(X_i\right)\right]^{1-Y_i}$$

$P(Y=1)^{Y}\,P(Y=0)^{1-Y}$

if $Y=1$: $P(Y=1)^{(1)}\,\underbrace{P(Y=0)^{1-1}}_{P(Y=0)^{0}=1}$

if $Y=0$ $P(Y=1)^{0}\,P(Y=0)^{1-0} \rightarrow 1-\pi(X)$

Recall

- Y_i: Response variable of the ith subject
- X_i: Independent variable for the ith subject
- $\pi\left(X_i\right) = \Pr\left(Y_i = 1 \mid X_i\right)$
- $1 - \pi\left(X_i\right) = \Pr(Y_i = 0 \mid X_i)$

2. Construct the likelihood function across the sample

- Since there are n subjects in the data, and **each subject is considered independent of each other**, the likelihood function for the whole data can be expressed as:

$$l(\beta_0, \beta_1) = \prod_{i=1}^{n} \pi(X_i)^{Y_i}(1 - \pi(X_i))^{1-Y_i}$$

$n = 60$

$20\ Y = 1$ $\quad$ $40\ Y = 0$

$\pi(X)^{20}\,(1 - \pi(X))^{40}$

Recall

- Y_i: Response variable of the ith subject
- X_i: Independent variable for the ith subject
- $\pi(X_i) = \Pr(Y_i = 1 | X_i)$
- $1 - \pi(X_i) = \Pr(Y_i = 0 | X_i)$

$$\log(a^b c^d) = b\log(a) + d\log(c)$$

3. Convert to log-likelihood

- Mathematically, it is easier to work with the **log likelihood** function for maximization
- The log likelihood function is:

$$L(\beta_0, \beta_1) = \ln(l(\beta_0, \beta_1))$$
$$= \sum_{i=1}^{n} \left[Y_i \cdot \ln[\pi(X_i)] + (1 - Y_i) \cdot \ln[1 - \pi(X_i)] \right]$$

Recall

- Y_i: Response variable of the ith subject
- X_i: Independent variable for the ith subject
- $\pi(X_i) = \Pr(Y_i = 1 | X_i)$
- $1 - \pi(X_i) = \Pr(Y_i = 0 | X_i)$

4. Find MLEs that maximize log-likelihood

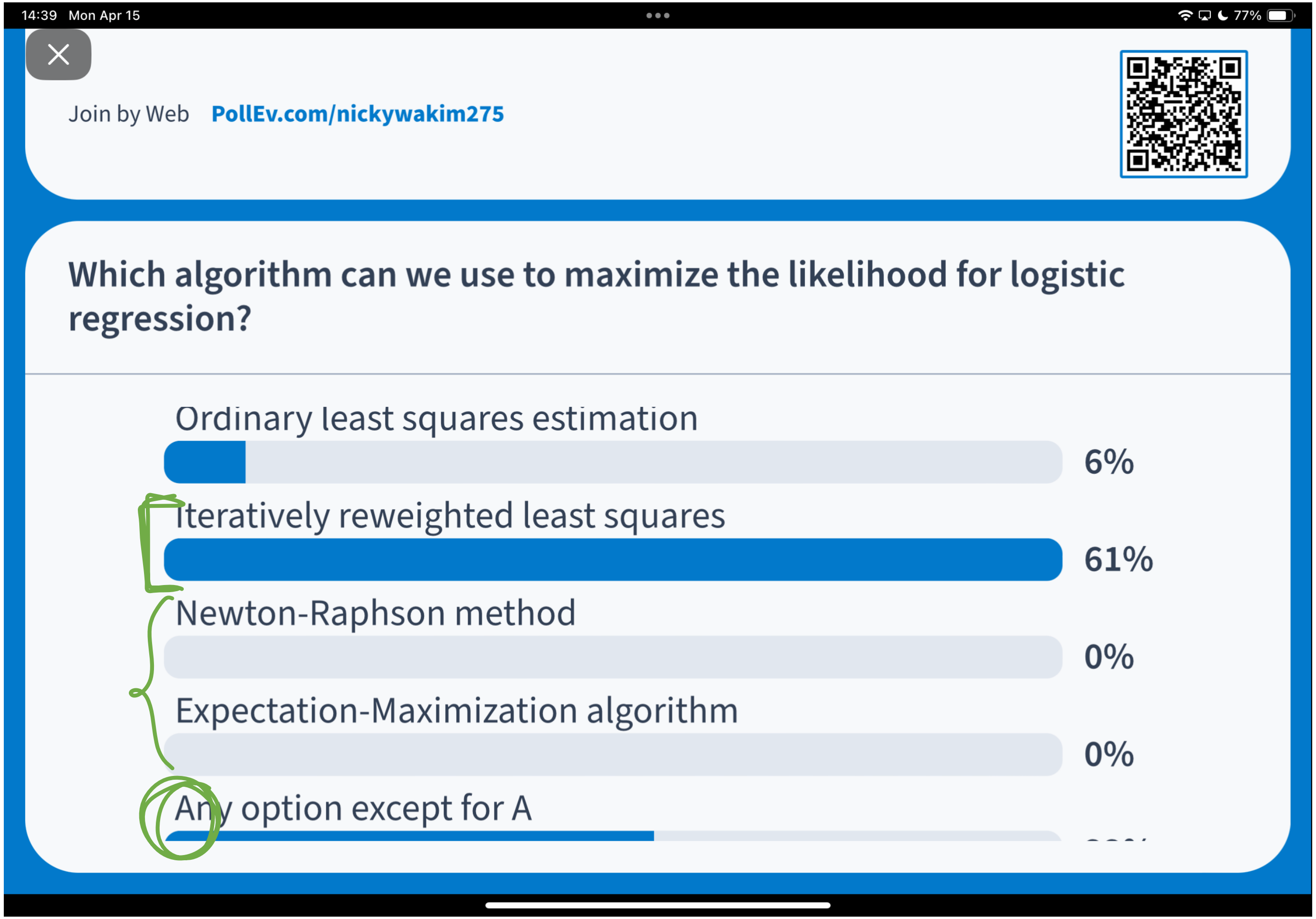
- To find β_0 and β_1 that maximizes $L(\beta_0, \beta_1)$:
  - We **differentiate** $L(\beta_0, \beta_1)$ with respect to β_0 and β_1...
  - And **set the resulting expression to zero**

- ~~Such equations are called likelihood equations.~~
  - $\cancel{\sum [Y_i - \pi(X_i)] = 0}$
  - $\cancel{\sum X_i [Y_i - \pi(X_i)] = 0}$

we think result of
differentiation

- In logistic regression, there is no “closed form” solution to the above equations
- Need to use iterative algorithm, such as iteratively reweighted least squares (IWLS) algorithm, should be used to find the MLEs for logistic regression

Poll Everywhere Question 5



How do perform MLE in R?

- `glm()` function automatically does MLE for you
  - For logistic regression with a binary outcome, we need to set the `family` within `glm()` to "binomial" which will automatically set the logit link
- You can explore other algorithms (other than IWLS) to maximize the likelihood
  - Pretty good Cross Validated post on algorithms in `glm()`

Example: Breast cancer diagnosis (1/3)

- Let's start with simple logistic regression with late stage breast cancer diagnosis as the outcome and age as our independent variable

- We want to fit:

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

  - Don't forget: $\pi(Age) = P(Y = 1|Age) = P(\text{Late stage BC diagnosis}|Age)$

Example: Breast cancer diagnosis (2/3)

logit

```
1  bc_reg = glm(Late_stage_diag ~ Age_c, data = bc, family = binomial)
2  summary(bc_reg)
```

```
Call:
glm(formula = Late_stage_diag ~ Age_c, family = binomial, data = bc)

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.989422   0.023205  -42.64   <2e-16 ***
Age_c        0.056965   0.003204   17.78   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 11861  on 9999  degrees of freedom
Residual deviance: 11510  on 9998  degrees of freedom
AIC: 11514

Number of Fisher Scoring iterations: 4
```

IWLS

Really big if convergence

Example: Breast cancer diagnosis (3/3)

- Translate the results back to an equation!
- Just going to pull the coefficients so I have a reference as I create the fitted regression model:

```
summary(bc_reg)$coefficients
```

```
              Estimate Std. Error   z value     Pr(>|z|)
(Intercept) -0.9894225  0.0232055 -42.63742 0.000000e+00
Age_c        0.0569645  0.0032039  17.77974 1.014557e-70
```

- Fitted logistic regression model:

$$\text{logit}(\pi(Age)) = -0.989 + 0.057 \cdot Age$$

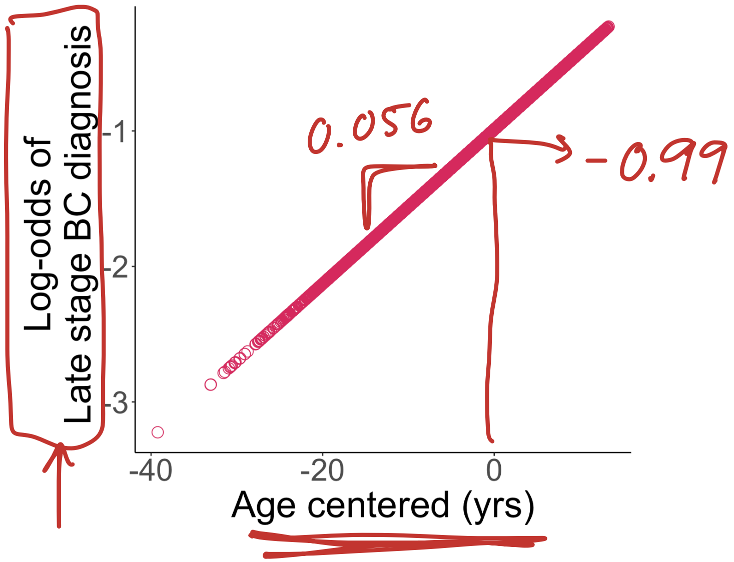
Example: Breast cancer diagnosis (3/3)

- Translate the results back to an equation!
- Just going to pull the coefficients so I have a reference as I create the fitted regression model:

```
summary(bc_reg)$coefficients
```

```
               Estimate  Std. Error   z value      Pr(>|z|)
(Intercept) -0.9894225  0.0232055 -42.63742 0.000000e+00
Age_c        0.0569645  0.0032039  17.77974 1.014557e-70
```

- This is the fitted line:



- Fitted logistic regression model:

$$\text{logit}(\pi(Age)) = -0.989 + 0.057 \cdot Age$$

We will need to reverse the transformation process in slide 24-25 to find the odds ratios

- Will do in next week's lessons

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 linear regression cannot be applied to categorical outcomes with two levels
3. Identify the simple 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) and apply it in R