Questions from last class

1. Does the order you put your variables in, in the R table, determine which is the explanatory, response and confounding variables?

NOT
→ does ^ matter for (c)MH test (three way table)

2. Why are we using Breslow-Day test or what is the rule of thumb for significant difference of crude vs adjusted ratio?

3. Is kappa coefficient lower case k or upper case K? We used two K's today and I am curious if the written way can make it more noticeable to differentiate the two K's

K K K

Mistake in last class!!

method = "midp"

```
oddsratio(heart_data, rev="b")$measure
```

Computes median estimate!

```
##                             NA
## odds ratio with 95% C.I.  estimate      lower    upper
##                 Control   1.0000000         NA       NA
##                 Treatment 0.6744415 0.4440882 1.014767
```

```
oddsratio(heart_data, rev="b", method = "wald")$measure
```

```
##                             NA
## odds ratio with 95% C.I.  estimate      lower    upper
##                 Control   1.0000000         NA       NA
##                 Treatment 0.6734586 0.4465044 1.015772
```

Announcements

- Virtual class on **Wednesday 4/19 AND Wednesday 4/26**
  - Virtual office hours on Wednesday for Nicky
  - Will send appropriate link for class

- Mistake in HW 2 – Question 3, part g
  - Corrected homework assignment now available

Class 5: Simple Logistic Regression

Class 5 Learning Objectives

1. Recognize why ordinary linear regression cannot be applied to categorical outcomes with two levels

2. Identify the logistic regression model and define key notation in statistics language

3. Connect linear and logistic regression to the larger group of models, generalized linear model

4. Interpret coefficient estimates for a categorical covariate in logistic regression model using odds ratios

5. Determine coefficient estimates using maximum likelihood estimation (MLE)

Class 5 Learning Objectives

1. Recognize why ordinary linear regression cannot be applied to categorical outcomes with two levels

2. Identify the logistic regression model and define key notation in statistics language

3. Connect linear and logistic regression to the larger group of models, generalized linear model

4. Interpret coefficient estimates for a categorical covariate in logistic regression model using odds ratios

5. Determine coefficient estimates using maximum likelihood estimation (MLE)

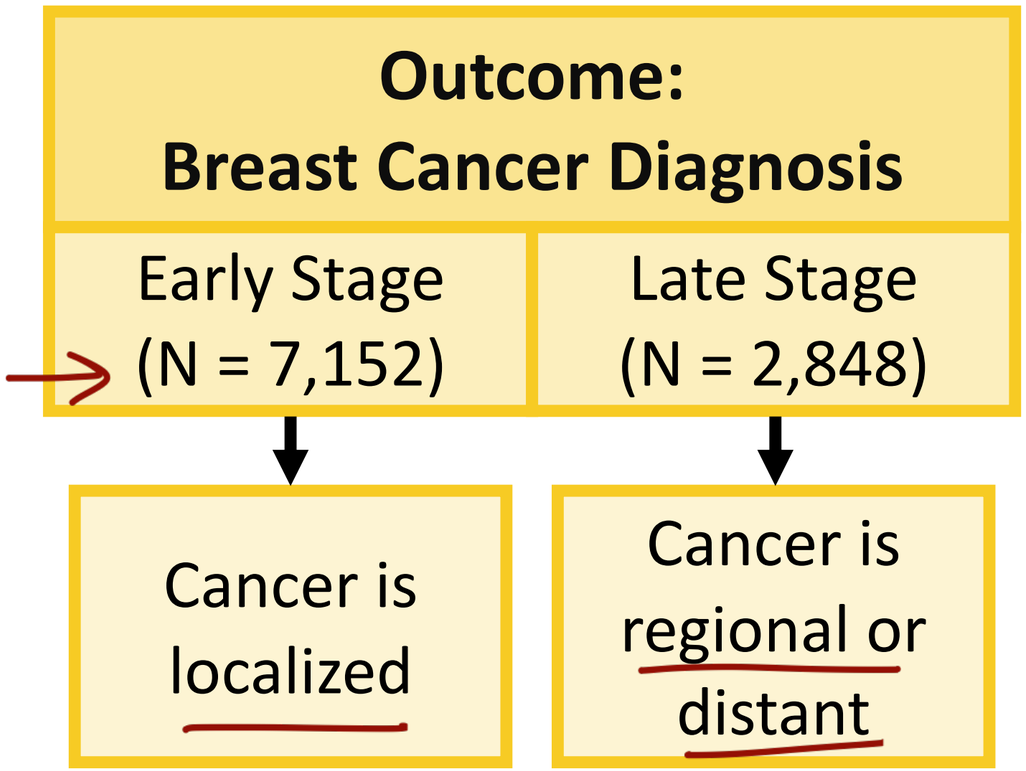
Example: Health disparities in breast cancer diagnosis (I)

- **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)

Example: Health disparities in breast cancer diagnosis (II)



Example: Health disparities in breast cancer diagnosis (II)

Outcome: Breast Cancer Diagnosis

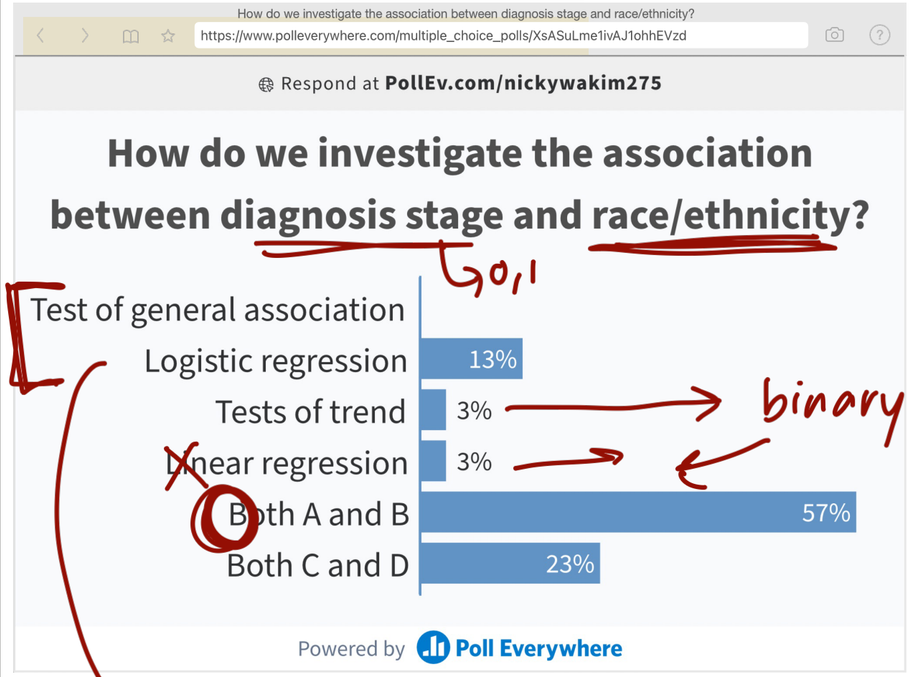
Early Stage (N = 7,152)	Late Stage (N = 2,848)
Cancer is localized	Cancer is regional or distant

Covariate: Age

Continuous,
range: 22-75 years old

Covariate: Race/Ethnicity categorical

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)	



0,1

binary response : not test trend

simple logistic reg

X^2 test, LRT

How do we determine differences in diagnosis? (I)

- 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? (II)

- 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? (II)

- 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

Class 5 Learning Objectives

1. Recognize why ordinary linear regression cannot be applied to categorical outcomes with two levels

2. Identify the logistic regression model and define key notation in statistics language

3. Connect linear and logistic regression to the larger group of models, generalized linear model

4. Interpret coefficient estimates for a categorical covariate in logistic regression model using odds ratios

5. Determine coefficient estimates using maximum likelihood estimation (MLE)

Example: Data at the individual level

similar $E(Y_i|X_i)$ $\pi(X_i)$

Outcome: Breast Cancer Diagnosis

Early Stage, $Y_i = 0$	Late Stage, $Y_i = 1$
$P(Y_i = 0\|x_i) = 1 - \pi(x_i)$	$P(Y_i = 1\|x_i) = \pi(x_i)$

$1-\pi(X_i)$ $\pi(x_i)$

Covariate: Age

Continuous, Age_i

Covariate: Race/Ethnicity

Hispanic-Latinx, $I(R\_E_i = H\text{-}L) = 1$ if not $= 0$	Non-Hispanic American Indian/Alaska Native, $I(R\_E_i = NH\ AIAN)$
Non-Hispanic Asian/Pacific Islander, $I(R\_E_i = NH\ API)$	Non-Hispanic Black, $I(R\_E_i = NH\ B)$
Non-Hispanic White, $I(R\_E_i = NH\ W)$	

Reference

Reference for individual overview

General form	Breast Cancer Diagnosis Example
i	Individual who was assigned female at birth 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)
$P(Y_i = 1\|x_i) = \pi(x_i)$	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

- **Goal:** model the probability of our outcome ($\pi(x_i)$) with the covariate ($x_{1,i}$)

- In simple linear regression, we use the model:

$$Y_i = \beta_0 + \beta_1 x_{1,i} + \epsilon_i$$

-or-

$$\hat{Y}_i = E(Y_i|x_i) = \beta_0 + \beta_1 x_{1,i}$$

$X_{(1)i}$ → 1 means 1st cov. i : ind.

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

$Y_i = 0$ or $Y_i = 1$ → $\pi(x_i)$ in $[0,1]$

Simple Logistic Regression Model: Components

- Outcome: Y_i - two-level categorical variable
  - $Y_i = 1$
  - $Y_i = 0$

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

Y value

$E(Y|X) = P(Y=1|X) \cdot 1 + P(Y=0|X) \cdot 0$

$= P(Y=1|X) = \pi(x)$

- Covariate: $x_{1,i}$
  - For now: simple logistic regression with one covariate
  - $x_{1,i}$ can be continuous or categorical

Can we apply OLR to our binary outcome?

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

Review of Ordinary Linear Regression (OLR) (I)

- For simplicity, consider simple linear regression model (where there is a single independent variable x).
- The ordinary linear regression can be written as:

$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i \qquad i = 1, \ldots, n.$$

Y_i: dependent variable → CONTINUOUS OUTCOME

x_i: independent variable

β_0 : intercept

β_1 : slope (regression coefficient)

ϵ_i: error

Review of Ordinary Linear Regression (II)

- Assumptions of ordinary linear regression model:
  - **Independence:** ϵ_i is independent of ϵ_i, i≠j.
    - i.e., knowing the error of one observation tells you nothing about the error of another observation.
  - **Linearity:** linear relationship between $\mu(= E(Y|x))$ and X
    - $\mu_i = \beta_0 + \beta_1 x_i$
  - **Normality and homoscedasticity** assumption for ϵ_i:
    - $\epsilon_i \sim N(0, \sigma^2)$

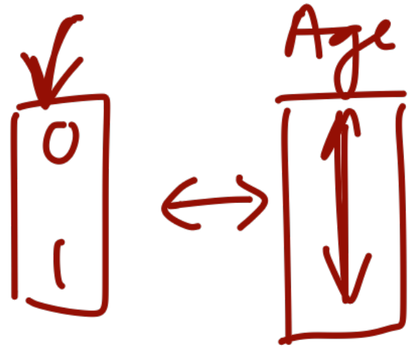
σ^2 does not depend on x

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

Violated: Linearity

- Y is a binary outcome with two possible values: 1 and 0

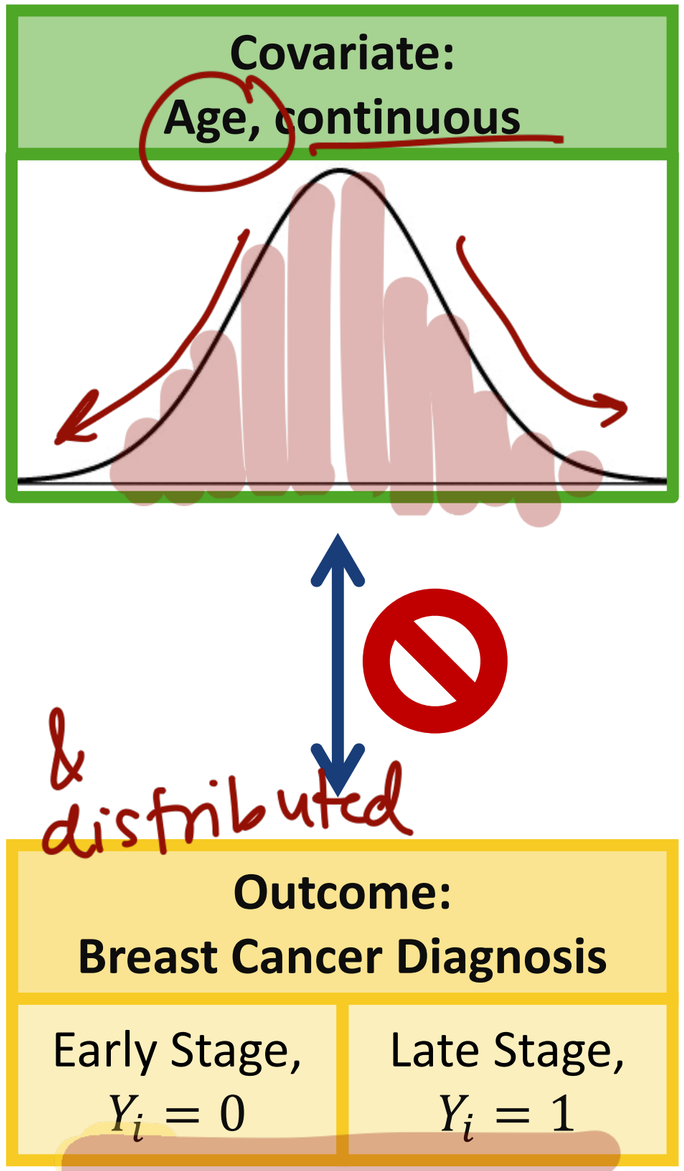
$$\mu_i = E(Y_i|x_i) = 1 \times \mathrm{P}(Y_i = 1|x_i) + 0 \times \Pr(Y_i = 0|x_i)$$
$$= \mathrm{P}(Y_i = 1|x_i)$$
$$= \pi_i = \pi(x_i)$$



- Based on Linear relationship assumption:

$$\mu_i = \pi_i = \beta_0 + \beta_1 x_i$$

continuous & normally distributed

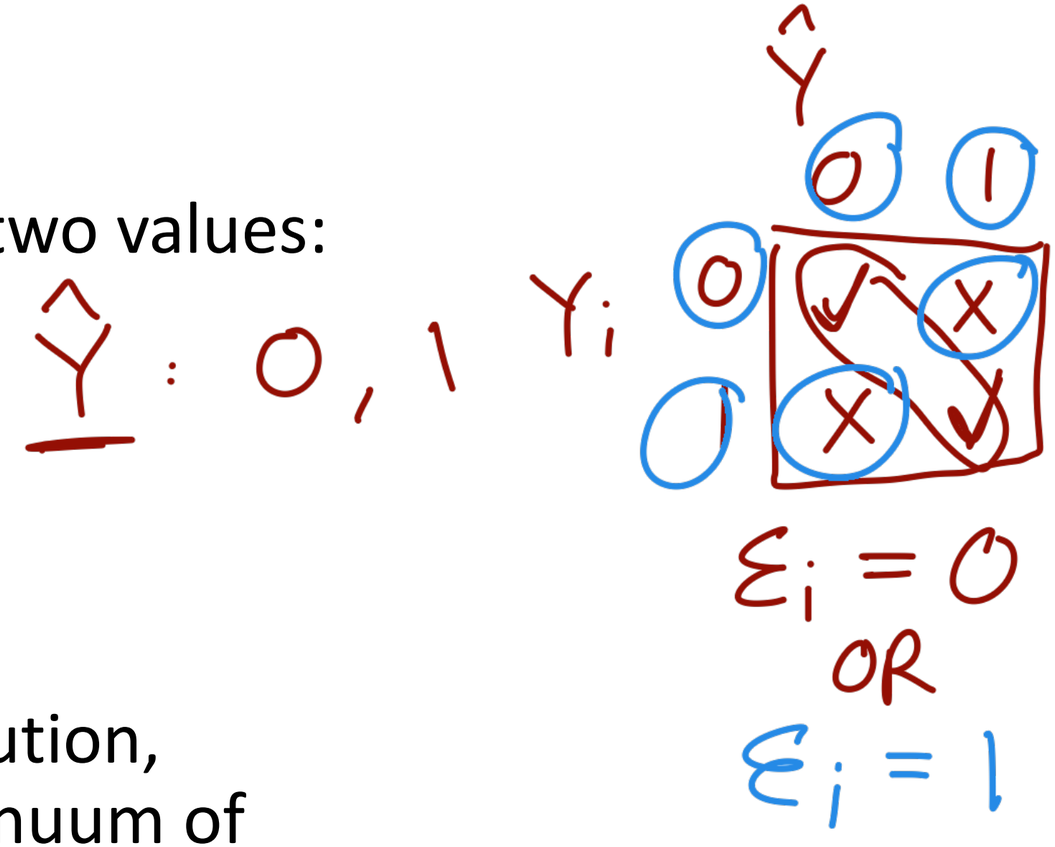


- The independent variable x_i can take any value, while π_i is a probability that should be bounded by (0,1)
  - We cannot use linear mapping to translate x_i to π_i

Violated: Normality

- In OLR, ϵ_i is distributed normally

- Recall that Y_i can take only one of the two values: 0 or 1
  - If Y_i=0 then $\epsilon_i = -\beta_0 - \beta_1 x_i$
  - If Y_i=1 then $\epsilon_i = 1 - \beta_0 - \beta_1 x_i$



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

Violated: Homoscedasticity

- In OLR, $\text{var}(\epsilon_i) = \sigma^2$
  - Variance does not depend on x_i

- When Y_i is a binary outcome

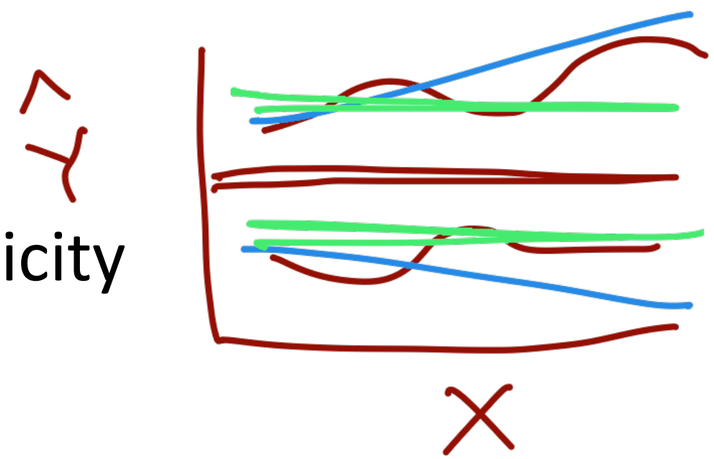
$$\text{var}(Y_i) = \pi_i(1-\pi_i)$$
$$= (\beta_0 + \beta_1 x_i)(1 - \beta_0 - \beta_1 x_i)$$

$\sum Y_i \sim$ binomial

$\hookrightarrow E(Y) = np$

$\text{var}(Y) = np(1-p)$

  - The $\text{var}(\epsilon_i)$ depends on x_i



- Because variance depends on x_i - no homoscedasticity

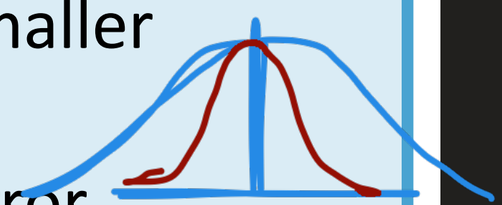
What happens if we use OLR for categorical responses?

Violation of Normality/linearity assumption

- Very serious when sample size is small
- Less serious when sample size is large
  - The coefficient estimate will have a distribution approximately normal even when ϵ_i is not normally distributed, according to central limit theorem.

Violation of homoscedasticity

- Inefficient coefficient estimates
  - i.e., there can be other methods of estimation with smaller standard error
- Inconsistent standard error estimates
  - i.e., the estimated standard errors could be biased (either upward or downward) to unknown degree



→ coefficient estimate mean okay, SE overestimated

Simple Logistic Regression Model: Components

- Outcome: Y_i - two-level categorical variable
  - $Y_i = 1$
  - $Y_i = 0$

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

- Covariate: $x_{1,i}$
  - For now: simple logistic regression with one covariate
  - $x_{1,i}$ can be continuous or categorical

Which, if any, of the below equations appropriately models our categorical outcome ($$Y_i$$) with our covariate ($$x_{1,i}$$)?

https://www.polleverywhere.com/multiple_choice_polls/xrijeTEsxzZrpajtYJf2U

Respond at **PollEv.com/nickywakim275**

Which, if any, of the below equations appropriately models our categorical outcome (Y_i) with our covariate ($x_{1,i}$)?

$Y_i = \beta_0 + \beta_1 x_{1,i}$ — 10%

$P(Y_i = 1) = \beta_0 + \beta_1 x_{1,i}$ — 31%

$p_i = \beta_0 + \beta_1 x_{1,i}$ — 3%

Both B and C are correct — 41%

None of the above — 14%

Powered by Poll Everywhere

0 OR 1

for linear regression only, we break assump.

$p_i = \pi(x_i) = P(Y_i = 1 \mid X_i)$ = my lazy $P(Y_i = 1)$

range $[0, 1]$

$X_i \sim N(\mu, \sigma^2)$ $(-\infty, \infty)$

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?

logit fn = $\log_e\left(\frac{\pi(x_i)}{1-\pi(x_i)}\right)$

Stage at diagnosis
Y_i

→

Probability of Late-stage diagnosis
$\pi_i = P(Y_i = 1 | x_i)$

→

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

Two levels:

$Y_i = 0$

$Y_i = 1$

Range of probabilities:

$0 \leq \pi_i \leq 1$

Range of logit values:

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

Note: people use π_i (or p_i) to mean $\pi(x_i)$

Simple Logistic Regression Model

$$\text{logit}(\pi_i) = \beta_0 + \beta_1 x_{1,i}$$

π_i: Probability that the outcome occurs ($Y_i = 1$) given $x_{1,i}$

β_0: Coefficient for the intercept

β_1: Coefficient for the independent variable (covariate)

$x_{1,i}$: Independent variable (covariate/predictor)

Simple Logistic Regression Model

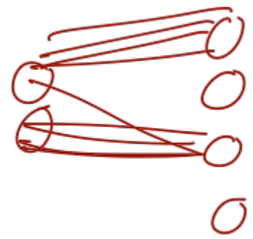
$$\text{logit}(\pi_i) = \beta_0 + \beta_1 x_{1,i}$$

π_i: Probability that the outcome occurs ($Y_i = 1$) given $x_{1,i}$

β_0: Coefficient for the intercept

β_1: Coefficient for the independent variable (covariate)

$x_{1,i}$: Independent variable (covariate/predictor)



Class 5 Learning Objectives

1. Recognize why ordinary linear regression cannot be applied to categorical outcomes with two levels

2. Identify the logistic regression model and define key notation in statistics language

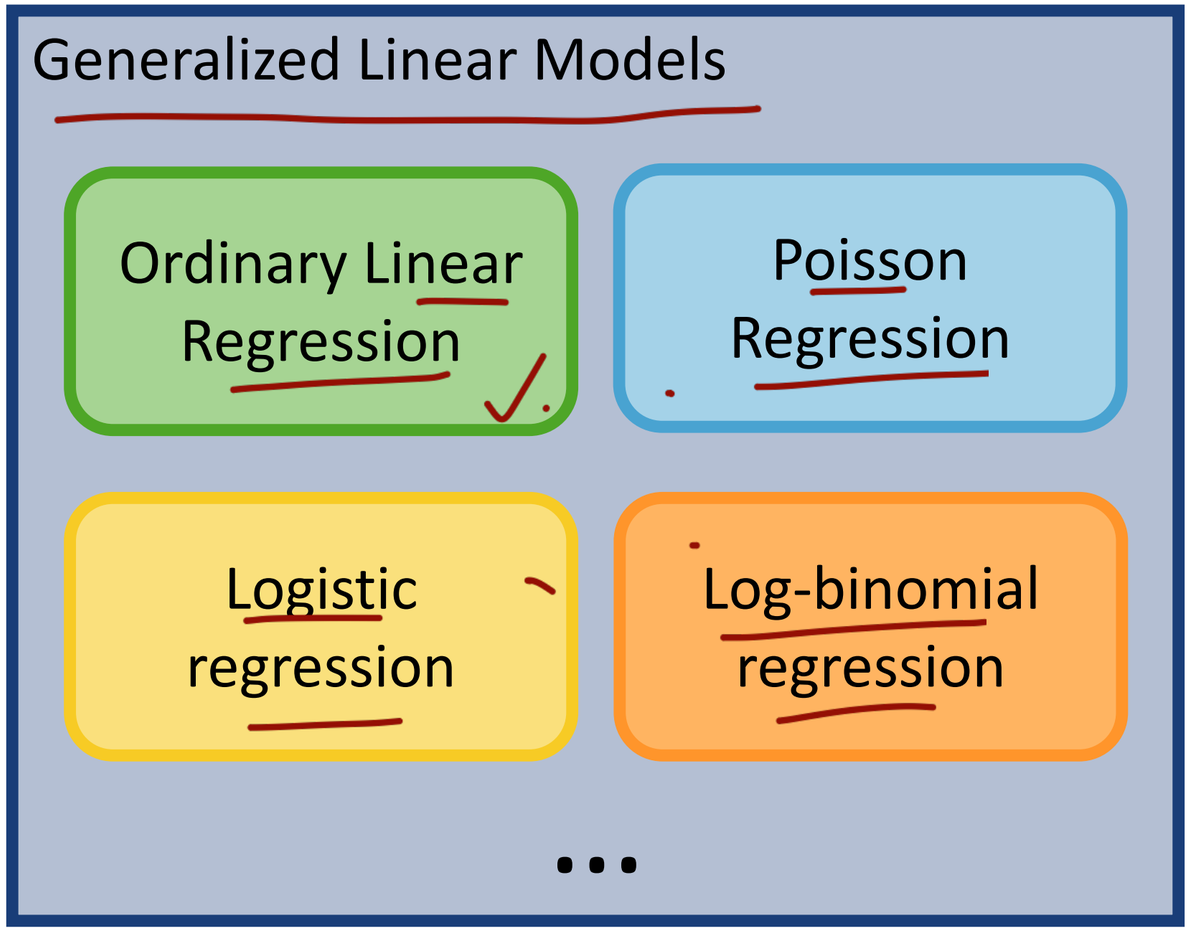
3. Connect linear and logistic regression to the larger group of models, generalized linear model

4. Interpret coefficient estimates for a categorical covariate in logistic regression model using odds ratios

5. Determine coefficient estimates using maximum likelihood estimation (MLE)

Generalized Linear Models (GLM) (I)

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



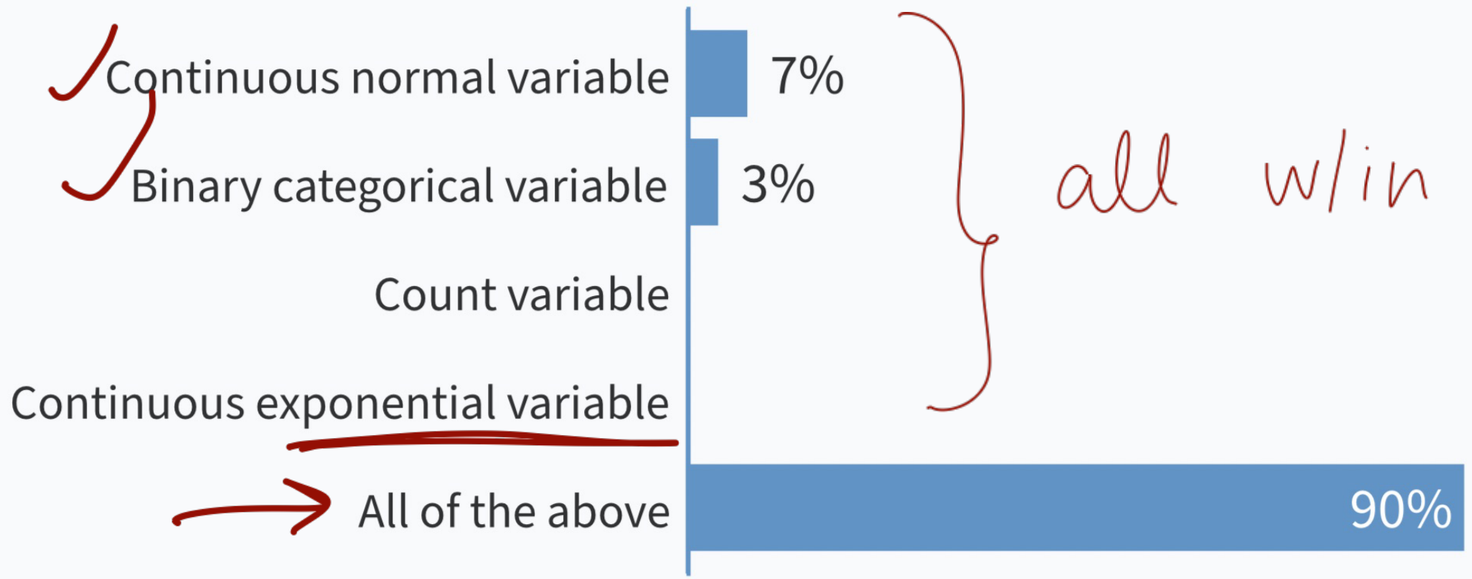
- Here we will focus on the GLMs for **categorical data**
  - **Logistic regression** is just a one type of GLM
  - **Poisson regression** – for counts
  - **Log-binomial** can be used to focus on risk ratio

What type of response variable could we model with a generalized linear model?

https://www.polleverywhere.com/multiple_choice_polls/V3plbaYe90nnzqB3luWiZ

Respond at **PollEv.com/nickywakim275**

What type of response variable could we model with a generalized linear model?



all w/in exp. family of distribution

Powered by Poll Everywhere

Generalized Linear Models (GLM) (II)

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

logit

Random Component

- The random component specifies the response variable Y ($Y_1, Y_2, \ldots, Y_n$) and selects a probability distribution for it

- Basically, we are just identifying the distribution for our outcome

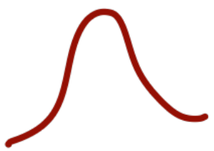
  - If Y is **binary**: assumes a **binomial** distribution of Y
  - If Y is **count**: assumes **Poisson** or negative binomial distribution of Y
  - If Y is **continuous**: assume **Normal** distribution of Y

Systematic Component

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

$$\beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k$$

- Here the x_i can be based on other explanatory variables.
  - For example, $x_3 = x_1 x_2$, or $x_3 = x_1^2$
  - Aka, interactions and squared-covariates maintain linearity

β_k^2 $\log(x) \rightarrow$ 

Link Function

- Let $\boldsymbol{\mu} = \boldsymbol{E}(\boldsymbol{Y})$, the link function specifies a function $g(.)$ that relates μ to the linear predictor as

$$g(\mu) = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k$$

  - $\boldsymbol{g}(\boldsymbol{\mu})$ is the transformation we make to $\boldsymbol{E}(\boldsymbol{Y})$ (aka $\boldsymbol{\mu}$) so that the linear predictors (right side of equation) can be linked to the outcome

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

$$\mu = g^{-1}(\beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k)$$

$g(\cdot)$

$\mu \longleftrightarrow \beta_0 + \beta_1 x_1$

$g(\mu) =$ logit(μ)

$Y \sim$ Normal

$g(\mu) = 1 \cdot \mu$

$\mu = g^{-1}(\beta_0 + \beta_1 x_1)$

$\mu = \beta_0 + \beta_1 x_1$

Link Function: Common types

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

Simple Logistic Regression Model

$$\text{logit}(p_i) = \beta_0 + \beta_1 x_{1,i}$$

p_i: Probability that the outcome occurs ($Y_i = 1$)

β_0: Coefficient for the intercept

β_1: Coefficient for the independent variable (covariate)

$x_{1,i}$: Independent variable (covariate/predictor)

Class 5 Learning Objectives

1. Recognize why ordinary linear regression cannot be applied to categorical outcomes with two levels

2. Identify the logistic regression model and define key notation in statistics language

3. Connect linear and logistic regression to the larger group of models, generalized linear model

4. Interpret coefficient estimates for a categorical covariate in logistic regression model using odds ratios

5. Determine coefficient estimates using maximum likelihood estimation (MLE)

Reference

Simple Logistic Regression Model: Breast cancer

+ Race / Ethnicity

$$\text{logit}(\pi(x_i)) = \beta_0 + \beta_1 I(R\_E_i = H{-}L) + \beta_2 I(R\_E_i = NH\ AIAN) + \beta_3 I(R\_E_i = NH\ API) + \beta_4 I(R\_E_i = NH\ B)$$

$\pi(x_i)$: probability of late-stage breast cancer diagnosis given individual's covariate values

β_0: intercept ($\text{logit}(\pi(x_i))$ of the reference group, Non-Hispanic White individuals)

β_1: increase in $\text{logit}(\pi(x_i))$ of Hispanic-Latinx individuals compared to Non-Hispanic White individuals

β_2: increase in $\text{logit}(\pi(x_i))$ of Non-Hispanic American Indians/Alaska Native individual compared to Non-Hispanic White individuals

β_3: increase in $\text{logit}(\pi(x_i))$ of Non-Hispanic Asian/Pacific Islander individuals compared to Non-Hispanic White individuals

β_4: increase in $\text{logit}(\pi(x_i))$ of Non-Hispanic Black individuals compared to Non-Hispanic White individuals

$I(\dots)$: Indicator that race/ethnicity of individual i is specified race/ethnicity

How do we interpret the coefficients?

$$\begin{aligned}&\text{logit}(\pi(x_i))\\&= \beta_0 + \beta_1 I(R\_E_i = H{-}L) + \beta_2 I(R\_E_i = NH\ AIAN)\\&+ \beta_3 I(R\_E_i = NH\ API) + \beta_4 I(R\_E_i = NH\ B)\end{aligned}$$

- $\hat{\beta}_0$: intercept $\text{logit}(\pi(x_i))$ of the reference group, Non-Hispanic White individuals)
- $\hat{\beta}_1$: increase in $\text{logit}(\pi(x_i))$ of Hispanic-Latinx individuals compared to Non-Hispanic White individuals

f

How do we interpret the coefficients?

$$\begin{aligned}&\text{logit}(\pi(x_i))\\&= \beta_0 + \beta_1 I(R\_E_i = H{-}L) + \beta_2 I(R\_E_i = NH\ AIAN)\\&+ \beta_3 I(R\_E_i = NH\ API) + \beta_4 I(R\_E_i = NH\ B)\end{aligned}$$

- $\hat{\beta}_0$: intercept $\text{logit}(\pi(x_i))$ of the reference group, Non-Hispanic White individuals)
- $\hat{\beta}_1$: increase in $\text{logit}(\pi(x_i))$ of Hispanic-Latinx individuals compared to Non-Hispanic White individuals

What does $\text{logit}\big(\pi(x_i)\big)$ mean?

In the same way that we transformed Y_i to get to $\text{logit}\big(\pi(x_i)\big)$, we need to transform $\text{logit}\big(\pi(x_i)\big)$ to a clinically meaningful value

How do we compare one group to another? (I)

For non-Hispanic White individuals,

$$\text{logit}(\pi(R\_E_i = NH\ W)) = \beta_0$$

For Hispanic-Latinx individuals,

$$\text{logit}(\pi(R\_E_i = H{-}L)) = \beta_0 + \beta_1 I(H\text{-}L)$$

The difference between the two groups:

$$\text{logit}(\pi(R\_E_i = H{-}L)) - \text{logit}(\pi(R\_E_i = NH\ \text{W}))$$

How do we compare one group to another? (II)

$$\text{logit}(\pi(R\_E_i = H{-}L)) - \text{logit}(\pi(R\_E_i = NH\,\text{W})) = \beta_0 + \beta_1 - \beta_0$$

$$= \beta_1$$

$$\beta_1 = \text{logit}(\pi(R\_E_i = H{-}L)) - \text{logit}(\pi(R\_E_i = NH\,\text{W}))$$

$$\beta_1 = \log\left(\frac{\pi_{HL}}{1-\pi_{HL}}\right) - \log\left(\frac{\pi_{NHW}}{1-\pi_{NHW}}\right)$$

$$\beta_1 = \log\left(\frac{\frac{\pi_{HL}}{1-\pi_{HL}}}{\frac{\pi_{NHW}}{1-\pi_{NHW}}}\right)$$

$$\text{logit}(\pi) = \log\left(\frac{\pi}{1-\pi}\right)$$

$$\log(t) - \log(e) = \log\left(\frac{t}{e}\right)$$

odds of late stage diagnosis for H-L

odds of late diagnosis for NH W.

exp()

$$\exp(\beta_1) = \frac{\frac{\pi_{HL}}{1-\pi_{HL}}}{\frac{\pi_{NHW}}{1-\pi_{NHW}}}$$

$\exp(\beta_1)$ = ?

Can we find odds of H-L from this??

What does $\text{logit}(\pi_i(x_i))/\text{logit}(\pi_i)/\text{logit}(p_i)$ mean?

$p_i = P(Y_i = 1|x_i)$
$1 - p_i = P(Y_i = 0|x_i)$

Logit probability

$$\text{logit}(p_i) = \log_e\left(\frac{p_i}{1-p_i}\right)$$

For Hispanic-Latinx individuals, the $\boldsymbol{\text{logit}(p_i)}$ value is $\hat{\beta}_1$ higher compared to Non-Hispanic White individuals.

Take exponential (exp) of log

Log odds ratio

$$\log_e\left(\frac{P(Y_i = 1|x_i)}{P(Y_i = 0|x_i)}\right)$$

For Hispanic-Latinx individuals, the **log odds ratio** is $\hat{\beta}_1$ higher compared to Non-Hispanic White individuals.

Take exponential (exp) of log

Odds ratio

For Hispanic-Latinx individuals, the **odds** of $Y_i = 1$ is $\exp(\hat{\beta}_1)$ times higher than Non-Hispanic White individuals.

How do we interpret the coefficients?

$$\begin{aligned}&\text{logit}(p_i)\\&= \beta_0 + \beta_1 I(R\_E_i = H{-}L) + \beta_2 I(R\_E_i = NH\ AIAN)\\&+ \beta_3 I(R\_E_i = NH\ API) + \beta_4 I(R\_E_i = NH\ B)\end{aligned}$$

$\exp(\hat{\beta}_0)$: The **odds of diagnosis** of late-stage breast cancer for a Non-Hispanic White individuals is $\exp(\hat{\beta}_0)$.

$\exp(\hat{\beta}_1)$: **If $\boldsymbol{exp(\hat{\beta}_1) > 1}$:** For Hispanic-Latinx individuals, the **odds of diagnosis** of late-stage breast cancer is $\boldsymbol{exp(\hat{\beta}_1)}$ **times higher** than Non-Hispanic White individuals.

$\exp(\hat{\beta}_1)$: **If $\boldsymbol{exp(\hat{\beta}_1) < 1}$:** For Hispanic-Latinx individuals, the **odds of diagnosis** of late-stage breast cancer is $\boldsymbol{(1/exp(\hat{\beta}_1))}$ **times lower** than Non-Hispanic White individuals **OR**

For Non-Hispanic White individuals, the **odds of diagnosis** of late-stage breast cancer is $\boldsymbol{(1/exp(\hat{\beta}_1))}$ **times higher** than Hispanic-Latinx individuals

Let's run the model!

```
r_e_glm = glm(Late_stage_diag ~ Race_Ethnicity, data = bc_diagnosis,
              family = binomial())
summary(r_e_glm)$coefficients %>% round(., 2)
```

outcome

explanatory

generalized linear regression lm()

glm (family= Gaussian)

Let's run the model!

```
r_e_glm = glm(Late_stage_diag ~ Race_Ethnicity, data = bc_diagnosis,
              family = binomial())
summary(r_e_glm)$coefficients %>% round(., 2)
```

```
##                                                  Estimate Std. Error z value
## (Intercept)                                         -0.99       0.03  -37.55
## Race_EthnicityHispanic-Latino                       -0.03       0.08   -0.40
## Race_EthnicityNH American Indian/Alaskan Native     -0.05       0.48   -0.11
## Race_EthnicityNH Asian/Pacific Islander              0.12       0.08    1.50
## Race_EthnicityNH Black                               0.34       0.07    4.83
##                                                  Pr(>|z|)
## (Intercept)                                          0.00
## Race_EthnicityHispanic-Latino                        0.69
## Race_EthnicityNH American Indian/Alaskan Native      0.91
## Race_EthnicityNH Asian/Pacific Islander              0.13
## Race_EthnicityNH Black                               0.00
```

How do we transform our coefficient estimates to get the estimated odds ratios?

https://www.polleverywhere.com/multiple_choice_polls/962OEBE3wbeLezq2BxN0F

Respond at **PollEv.com/nickywakim275**

How do we transform our coefficient estimates to get the estimated odds ratios?

- No transformation needed
- Take the difference between the estimates
- Take the log of the estimates
- Take the exponential of the estimates

Total Results: 0

Powered by Poll Everywhere

Transform the coefficients to ORs

```
exp(r_e_glm$coefficients)
```

Transform the coefficients to ORs

```
exp(r_e_glm$coefficients)
```

```
##                                                 (Intercept)
##                                                   0.3721105
##                                 Race_EthnicityHispanic-Latino
##                                                   0.9678002
## Race_EthnicityNH American Indian/Alaskan Native
##                                                   0.9484848
##                       Race_EthnicityNH Asian/Pacific Islander
##                                                   1.1310170
##                                     Race_EthnicityNH Black
##                                                   1.4046741
```

$= \exp(\hat{\beta}_0)$

$= \exp(\hat{\beta}_2)$

Reference

Interpretation of odds ratios (categorical covariate)

Odds ratio	Interpretation
$\exp(\hat{\beta}_0) = 0.37$	The **odds of diagnosis** of late-stage breast cancer for a Non-Hispanic White individuals is 0.37.
$\exp(\hat{\beta}_1) = 0.97$	For Hispanic-Latinx individuals, the **odds of diagnosis** of late-stage breast cancer is **0.97 times higher** than Non-Hispanic White individuals.
$\exp(\hat{\beta}_2) = 0.94$	For Non-Hispanic American Indian and Alaska Native individuals, the **odds of diagnosis** of late-stage breast cancer is **1.06 times lower** than Non-Hispanic White individuals.
$\exp(\hat{\beta}_3) = 1.13$	For Non-Hispanic Asian and Pacific Islander individuals, the **odds of diagnosis** of late-stage breast cancer is **1.13 times higher** than Non-Hispanic White individuals.
$\exp(\hat{\beta}_4) = 1.40$	For Non-Hispanic Black individuals, the **odds of diagnosis** of late-stage breast cancer is **1.4 times higher** than Non-Hispanic White individuals.

Class 5 Learning Objectives

1. Recognize why ordinary linear regression cannot be applied to categorical outcomes with two levels

2. Identify the logistic regression model and define key notation in statistics language

3. Connect linear and logistic regression to the larger group of models, generalized linear model

4. Interpret coefficient estimates for a categorical covariate in logistic regression model using odds ratios

5. Determine coefficient estimates using maximum likelihood estimation (MLE)

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

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

True or false: Least squares estimation for linear regression maximizes the likelihood

https://www.polleverywhere.com/multiple_choice_polls/blj7aLnO1LpNSEoxVz70a

Respond at **PollEv.com/nickywakim275**

True or false: Least squares estimation for linear regression maximizes the likelihood

True

False

Total Results: 0

Powered by Poll Everywhere

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

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

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 $\pi(x_i)$.
  - if $Y_i = 0$, the contribution to the likelihood function is $1 - \pi(x_i)$

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

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

$$\pi(x_i)^{y_i}[1 - \pi(x_i)]^{1-y_i}$$

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:

$$l(\beta_0, \beta_1) = \prod_{i=1}^{n} \pi(x_i)^{y_i} \left(1 - \pi(x_i)\right)^{1-y_i}$$

Recall:

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

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:

$$l(\beta_0, \beta_1) = \prod_{i=1}^{n} \pi(x_i)^{y_i} \left(1 - \pi(x_i)\right)^{1-y_i}$$

Recall:

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

3. Convert to log-likelihood

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

- The log likelihood function is:

$$L(\beta_0, \beta_1) = \ln\big(l(\beta_0, \beta_1)\big)$$

$$= \sum_{i=1}^{n} \{y_i \ln[\pi(x_i)] + (1 - y_i)\ln[1 - \pi(x_i)]\}$$

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.
  - $\sum[y_i - \pi(x_i)] = 0$
  - $\sum x_i[y_i - \pi(x_i)] = 0$

- Iterative algorithm, such as iteratively reweighted least squares (IWLS) algorithm, should be used to find the MLEs for logistic regression

How do we do this in R?

- glm() function automatically does MLE for you
- You can explore other algorithms (other than IWLS) to maximize the likelihood
- Let’s take a peak in R

Class 5 Learning Objectives

1. Recognize why ordinary linear regression cannot be applied to categorical outcomes with two levels

2. Identify the logistic regression model and define key notation in statistics language

3. Connect linear and logistic regression to the larger group of models, generalized linear model

4. Interpret coefficient estimates for a categorical covariate in logistic regression model using odds ratios

5. Determine coefficient estimates using maximum likelihood estimation (MLE)

Wrap-up

- 4-minute exit ticket

- **Next class**
  - Continue with simple logistic regression
  - More on the coefficient estimate
    - Testing significance
    - Confidence interval
- Predicted probability
- Multiple logistic regression

Class 5 Exit Ticket

https://forms.office.com/r/GLwsuqsnqt