

Lesson 6: Interpretations and Visualizations of Odds Ratios

Nicky Wakim

2026-04-13

Learning Objectives

1. Interpret odds ratios from fitted simple logistic regression model for a continuous explanatory variable.
2. Interpret odds ratios from fitted simple logistic regression model for a binary explanatory variable.
3. Interpret odds ratios from fitted simple logistic regression model for a multi-level categorical explanatory variable.
4. Report the odds ratio using a table and/or a forest plot.

Age

Recall our example: Late stage breast cancer diagnosis

- Recall that we fit a simple logistic regression for late stage breast cancer diagnosis using the predictor, age:

```
1 bc_reg = bc %>% glm(formula = Late_stage_diag ~ Age_c, family = binomial)
2 tidy(bc_reg, conf.int=T) %>% gt() %>% tab_options(table.font.size = 38) %>%
3   fmt_number(decimals = 3)
```

term	estimate	std.error	statistic	p.value	conf.low	conf.high
(Intercept)	-0.989	0.023	-42.637	0.000	-1.035	-0.944
Age_c	0.057	0.003	17.780	0.000	0.051	0.063

Population logistic regression model

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

Fitted logistic regression model

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

Coefficients on the log-odds scale

Population logistic regression model

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

- β_0 : log-odds of $Y = 1$ when X is 0
- β_1 : increase in log-odds of $Y = 1$ for every 1 unit increase in X

Fitted logistic regression model

$$\text{logit}(\hat{\pi}(X)) = \hat{\beta}_0 + \hat{\beta}_1 \cdot X$$

- $\hat{\beta}_0$: estimated log-odds of $Y = 1$ when X is 0
- $\hat{\beta}_1$: estimated increase in log-odds of $Y = 1$ for every 1 unit increase in X
- Can use expected instead of estimated

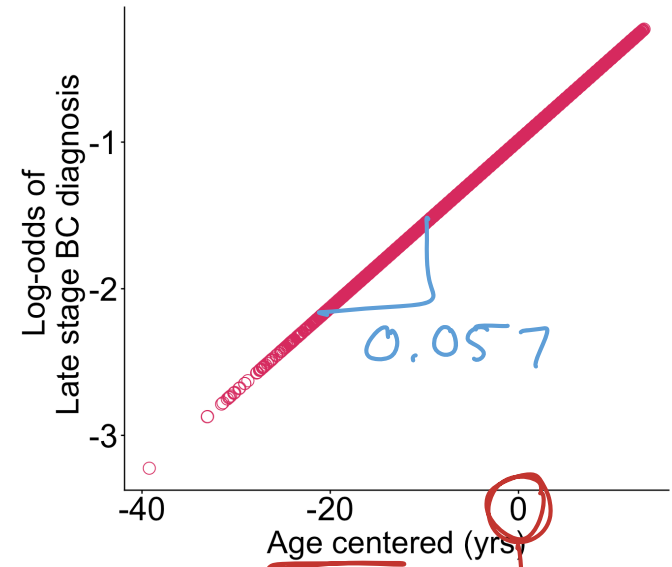
Recall our example: Late stage breast cancer diagnosis

- Fitted logistic regression model:

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

of LSBC diagnosis

- $\hat{\beta}_0$: The estimated log-odds is -0.989 for someone who is 61.71 years (95% CI: -1.035, -0.944)
- $\hat{\beta}_1$: The estimated increase in log-odds is 0.057 for every 1 year increase in age (95% CI: 0.051, 0.063).
- What does a log-odds even mean?? Or an increase in log-odds??
- We will need to calculate the **odds ratio** and its confidence interval
 - Then we will visualize the odds ratio



*mean age
61.71 yrs
(centered
age = 0)*

Poll Everywhere Question 1

13:42 Mon Apr 13

94%

Join by Web [PollEv.com/nickywakim275](https://poll-ev.com/nickywakim275)



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

~~No transformation needed~~

0%

~~Take the difference between the estimates~~

6%

Take the log of the estimates

26%

Take the exponential of the estimates ✓

68%

$$P(Y=1)$$

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

$$\log\left(\frac{P(Y=1)}{1-P(Y=1)}\right)$$

$$\log\left(\log\left(\frac{P(Y=1)}{1-P(Y=1)}\right)\right)$$

$$\exp\left(\log\left(\frac{P(Y=1)}{1-P(Y=1)}\right)\right)$$

odds

We typically interpret our results using odds ratios

For our *fitted* simple logistic regression model with a continuous predictor

$$\text{logit}(\hat{\pi}(X)) = \hat{\beta}_0 + \hat{\beta}_1 \cdot X$$

- How do we go from interpretations of $\hat{\beta}_0$ and $\hat{\beta}_1$ using log odds to odds ratios?
- We will need to take the exponential of our model:
 - $\exp(\hat{\beta}_0)$: expected odds that $Y = 1$ when X is 0.
 - $\exp(\hat{\beta}_1)$: expected odds ratio that $Y = 1$ for every 1 unit increase in X
- Important distinction:
 - We take the inverse logit to find our predicted probability →
 - We take the exponential to interpret the odds/odds ratios

Intro/Recap of Interpreting Fitted Model

- Interpret coefficients from fitted logistic regression model
 - **Goodness-of-fit of model should be assessed before summarizing findings** (have not covered yet)
 - In this lecture: assume model fits data well
- The interpretation of the coefficients involves two issues:
 - The functional relationship between the dependent variable and the independent variable (*link function*)
 - Unit of change for the independent variable
units for X
- We will learn the interpretation for
 - Binary independent variable
 - Categorical independent variable with multiple groups
 - We looked at this for our race and ethnicity variable
 - Continuous independent variable

Learning Objectives

1. Interpret odds ratios from fitted simple logistic regression model for a continuous explanatory variable.
2. Interpret odds ratios from fitted simple logistic regression model for a binary explanatory variable.
3. Interpret odds ratios from fitted simple logistic regression model for a multi-level categorical explanatory variable.
4. Report the odds ratio using a table and/or a forest plot.

Coefficient interpretation: Continuous Independent Variable

- For simplicity, we assume the linear relationship between logit and continuous variable X
- Again using simple logistic regression model to illustrate the interpretation of $\hat{\beta}$ for a continuous variable X

$$\text{logit}(\hat{\pi}(X)) = \hat{\beta}_0 + \hat{\beta}_1 \cdot X$$

- The estimated slope coefficient, $\hat{\beta}_1$, is the **expected change in the log odds for 1 unit increase in X**
 - Additional attention should be paid to picking a meaningful units of change in X

How do we get the odds ratio for X's coefficient?

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

For $\exp(\hat{\beta}_1)$

• Thus,

- We compare $X_1 = x$ and $X_2 = x + 1$,
- So we have

$$X_1: \text{logit}(\hat{\pi}(X = x)) = \hat{\beta}_0 + \hat{\beta}_1 \cdot x$$

and
 $X_2:$

$$\text{logit}(\hat{\pi}(X = x + 1)) = \hat{\beta}_0 + \hat{\beta}_1 \cdot (x + 1)$$

• And...

$$\begin{aligned} & \left[\text{logit}(\hat{\pi}(X = x + 1)) - \text{logit}(\hat{\pi}(X = x)) \right] \exp[\hat{\beta}_1] \\ &= \hat{\beta}_0 + \hat{\beta}_1 \cdot (x + 1) - [\hat{\beta}_0 + \hat{\beta}_1 \cdot x] \\ &= \hat{\beta}_0 + \hat{\beta}_1 x + \hat{\beta}_1 - \hat{\beta}_0 - \hat{\beta}_1 x \\ &= \hat{\beta}_1 \end{aligned}$$

log odds $\hat{\beta}_1$

$$\begin{aligned} \hat{\beta}_1 &= \text{logit}(\hat{\pi}(X = x + 1)) - \text{logit}(\hat{\pi}(X = x)) \\ \hat{\beta}_1 &= \log\left(\frac{\hat{\pi}(X = x + 1)}{1 - \hat{\pi}(X = x + 1)}\right) - \log\left(\frac{\hat{\pi}(X = x)}{1 - \hat{\pi}(X = x)}\right) \\ &\text{odds} \quad \text{odds} \\ \hat{\beta}_1 &= \log\left(\frac{\frac{\hat{\pi}(X = x + 1)}{1 - \hat{\pi}(X = x + 1)}}{\frac{\hat{\pi}(X = x)}{1 - \hat{\pi}(X = x)}}\right) \\ &\text{odds} \quad \text{odds} \\ &= \log(b) - \log(c) \\ &= \log\left(\frac{b}{c}\right) \\ &\text{odds} = \frac{p}{1-p} \\ \text{OR} &= 1.06 \end{aligned}$$

Example: Interpretation of Age Coefficient/OR

In log-odds scale:

$$\hat{\beta}_1 = 0.057$$

One year increase in age is associated with expected 0.057 increase in **log-odds** of receiving a late stage breast cancer diagnosis

$$100\% \cdot (\exp(\hat{\beta}_1) - 1)$$

In odds scale:

$$\exp(\hat{\beta}_1) = 1.06$$

- One year increase in age is associated with expected 1.06 **times the odds** of receiving a late stage breast cancer diagnosis
 - Can also say: For every one year increase in age, the expected odds of late stage breast cancer diagnosis increases 1.06 times.
- Can also subtract 1 from the odds ratio and multiply by 100 to obtain the percentage change in odds for 1-unit increase.
 - 1-year increase in age is associated with an **6% increase in the predicted odds** of late stage diagnosis

Example: Age and Late Stage Diagnosis (1/5)

Odds ratio from logistic regression

Compute the estimate and 95% confidence interval for odds ratio for late stage breast cancer diagnosis for every 1 year increase in age.

Needed steps:

1. Fit the regression model
2. Transform the coefficients into odds ratios
3. Interpret the odds ratio

Example: Age and Late Stage Diagnosis (2/5)

Odds ratio from logistic regression

Compute the estimate and 95% confidence interval for odds ratio for late stage breast cancer diagnosis for every 1 year increase in age.

1. Fit the regression model

```
1 bc_reg = bc %>%  
2   glm(formula = Late_stage_diag ~ Age_c, family = binomial)
```

Example: Age and Late Stage Diagnosis (3/5)

Odds ratio from logistic regression

Compute the estimate and 95% confidence interval for odds ratio for late stage breast cancer diagnosis for every 1 year increase in age.

2. Transform the coefficients into odds ratios • Option 1: `tidy()`

```
1 tidy_bc_reg = tidy(bc_reg, CI'S conf.int=T, exponentiate = T)
2 tidy_bc_reg %>% gt() %>% tab_options(table.font.size = 35) %>%
3   fmt_number(decimals = 3)
```

	term	estimate	std.error	statistic	p.value	conf.low	conf.high
$\exp(\hat{\beta}_0) \leftarrow$	(Intercept)	0.372	0.023	-42.637	0.000	0.355	0.389
$\exp(\hat{\beta}_1) \leftarrow$	Age_c	1.059	0.003	17.780	0.000	1.052	1.065

```
1 tidy_bc_reg$conf.low # I prefer tidy() bc now I can grab each component
[1] 0.3551931 1.0520321
```

Example: Age and Late Stage Diagnosis (4/5)

Odds ratio from logistic regression

Compute the estimate and 95% confidence interval for odds ratio for late stage breast cancer diagnosis for every 1 year increase in age.

2. Transform the coefficients into odds ratios • Option 2: `logistic.display()`

```
1 logistic.display(bc_reg) # Cannot grab each component in this
```

Logistic regression predicting Late_stage_diag : 1 vs 0

	$\exp(\hat{\beta}_1)$ OR(95%CI)	P(Wald's test)	P(LR-test)
Age_c (cont. var.)	1.06 (1.05,1.07)	< 0.001	< 0.001

Log-likelihood = -5754.8442

No. of observations = 10000

AIC value = 11513.6884

Poll Everywhere Question 2

14:07 Mon Apr 13

87%



Join by Web PollEv.com/nickywakim275



Please fill in the blanks for the following interpretation of our $\hat{OR}$: A 1 year increase in age is associated with 1.06 (95% CI: (1.05, 1.07)) times the odds of a late stage breast cancer diagnosis.



Example: Age and Late Stage Diagnosis (5/5)

Odds ratio from logistic regression

Compute the estimate and 95% confidence interval for odds ratio for late stage breast cancer diagnosis for every 1 year increase in age.

3. Interpret the odds ratio

For every one year increase in age, there is an expected 5.86% increase in the odds of late stage breast cancer diagnosis (95% CI: 5.2%, 6.53%).

$$\begin{aligned} & (1.0586 - 1) \times 100\% \\ & \quad \quad \quad \underline{5.86\%} \\ \text{odds} = \underline{1} & \rightarrow \underline{1.0586} \\ \text{6\% of } 1 & \rightarrow \frac{(0.0586 \times 1)}{+ 1} \end{aligned}$$

6% of 1.0586

$$\begin{aligned} & (0.0586 \times) \\ & \quad \quad \quad 1.0586 \\ & + \underline{1.0586} \end{aligned}$$

Transformations of continuous variable to make more interpretable

- Sometimes a change in “1” unit may not be considered clinically interesting
- For example, a 1 year increase in age or a 1 mm Hg increase in systolic blood pressure may be too small for a meaningful change in log odds
 - Instead, we may be interested to find out the log odds change for a increase of 10 years in age or 10 mm Hg in systolic blood pressure
- On the other hand, if the range of X is small (say 0-1), then a change in 1 unit of X is too large to be meaningful
- We should be able to compute and interpret coefficients for a continuous independent covariate x for an arbitrary change of “c” units in x

Transformations of continuous variable to make more interpretable

- The estimated log odds ratio for a change of c units in x can be obtained from

$$\hat{g}(x+c) - \hat{g}(x) = c\hat{\beta}_1$$

Handwritten notes below the equation:
Under $x+c$: $x+1$ and $x+10$
Under x : x and x
Under $c\hat{\beta}_1$: $10\hat{\beta}_1$ (with an arrow pointing from c to 10)

- $\widehat{OR}(c) = \exp(c\hat{\beta}_1)$

- The 95% CI for $\widehat{OR}(c)$ is:

$$\exp\left(c\hat{\beta}_1 \pm 1.96 \cdot c \cdot SE_{\hat{\beta}_1}\right)$$

- The c is chosen to be a clinically meaningful unit change in x
- The value of c should be clearly specified in all tables and calculations
 - Because the estimated OR and the corresponding CI depends on the choice of c value

Example: 10 year increase in age and Late Stage Diagnosis

- What if we are interested in learning the OR corresponding to 10-year increase in age?

```
1 bc_reg = glm(Late_stage_diag ~ Age_c, data = bc, family = binomial)
2 tidy(bc_reg, conf.int=T) %>% gt() %>% tab_options(table.font.size = 35) %>%
3   fmt_number(decimals = 3)
```

term	estimate	std.error	statistic	p.value	conf.low	conf.high
(Intercept)	-0.989	0.023	-42.637	0.000	-1.035	-0.944
Age_c	0.057	0.003	17.780	0.000	0.051	0.063

1 yr unit

Example: 10 year increase in age and Late Stage Diagnosis

- What if we are interested in learning the OR corresponding to 10-year increase in age?

$$\widehat{OR}(10) = \exp \left(\underbrace{10 \cdot \hat{\beta}_1}_{10 \cdot 0.057} \right) = \exp \left(\underbrace{0.56965} \right) = \underbrace{1.767}$$

- The 95% CI for $\widehat{OR}(10)$ is:

$$\begin{aligned}\widehat{OR}(10) &= \exp \left(10 \cdot \hat{\beta}_1 \pm 1.96 \cdot 10 \cdot SE_{\hat{\beta}_1} \right) \\ &= \exp \left(10 \cdot 0.056965 \pm 1.96 \cdot 10 \cdot 0.003204 \right) \\ &= (1.66, 1.88)\end{aligned}$$

Conclusion: For every 10-year increase in age, the expected odds of late stage breast cancer diagnosis increases 1.77 times (95% CI: 1.66, 1.88).

Example: 10 year increase in age and Late Stage Diagnosis

- What if we are interested in learning the OR corresponding to 10-year increase in age?

```
1 bc2 = bc %>% mutate(Age_c_10 = Age_c/10)
2 bc_reg_10 = glm(Late_stage_diag ~ Age_c_10, data = bc2, family = binomial)
3 tidy(bc_reg_10, conf.int=T, exponentiate = T) %>% gt() %>% tab_options(table.font.size = 35)
4   fmt_number(decimals = 3)
```

term	estimate	std.error	statistic	p.value	conf.low	conf.high
(Intercept)	0.372	0.023	-42.637	0.000	0.355	0.389
Age_c_10	1.768	0.032	17.780	0.000	1.661	1.883

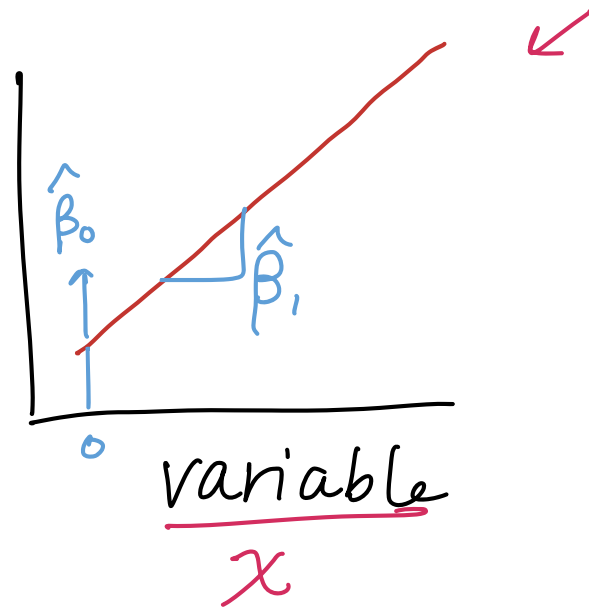
every 1 unit is actually 10 yrs

Last Note About Continuous Independent Variable

- Notice that the logistic regression model suggests that logit is linear in the covariate
- The model implies the additional odds of late stage breast cancer diagnosis for a 40 year-old compared to a 30 year-old is the same as the additional odds of late stage breast cancer diagnosis for a 60 year-old compared to a 50-year-old
- This assumption may not be realistic
- To address this, we may consider using higher order terms (e.g., x^2 , x^3 , ...) or other nonlinear transformation (e.g., $\log(x)$)
- Categorize the continuous variable may be another option
- See Lesson 8 and Lesson 14 in our [BSTA 512/612 class](#)

$$\text{logit}(\pi(x)) = \beta_0 + \beta_1 x + \beta_2 x^2$$

log odds
of $Y=1$



How do we get the odds for the intercept?

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

- When $X = 0$, we have

$$\text{logit}(\hat{\pi}(X = 0)) = \hat{\beta}_0$$

- Thus,

$$\hat{\beta}_0 = \text{logit}(\hat{\pi}(X = 0))$$

$$\exp[\hat{\beta}_0] = \exp[\text{logit}(\hat{\pi}(X = 0))]$$

$$\exp[\hat{\beta}_0] = \exp\left[\log\left(\frac{\hat{\pi}(X = 0)}{1 - \hat{\pi}(X = 0)}\right)\right]$$

$$\exp[\hat{\beta}_0] = \frac{\hat{\pi}(X = 0)}{1 - \hat{\pi}(X = 0)}$$

Learning Objectives

1. Interpret odds ratios from fitted simple logistic regression model for a continuous explanatory variable.
2. Interpret odds ratios from fitted simple logistic regression model for a binary explanatory variable.
3. Interpret odds ratios from fitted simple logistic regression model for a multi-level categorical explanatory variable.
4. Report the odds ratio using a table and/or a forest plot.

Coefficient Interpretation: Binary Independent Variable

- Independent variable X is a binary variable (X can take values: 0 or 1)

- We are fitting the population simple logistic regression model:

$$\text{logit}(\pi(X)) = \beta_0 + \beta_1 \cdot I(X = 1)$$

$$\begin{aligned} \hookrightarrow I(X=1) &= 1 && \text{if } X=1 \\ I(X=1) &= 0 && \text{if } X \neq 1 \end{aligned}$$

- The logit difference is β_1 for binary independent variable
 - β_1 represents the change/difference in the logit for $X = 1$ vs. $X = 0$
 - AKA log odds ratio comparing $X = 1$ to $X = 0$
- It will be much easier to understand if we can interpret the coefficient using odds ratio (OR)

Binary predictor: How do we interpret the coefficient? (1/2)

- For individuals with $X = 0$:

$$\text{logit}(\pi(X = 0)) = \beta_0 + \beta_1 \times (0) = \beta_0$$

- For individuals with $X = 1$:

$$\text{logit}(\pi(X = 1)) = \beta_0 + \beta_1 \times (1) = \beta_0 + \beta_1$$

- To solve for β_1 , we take the difference of the logits:

$$\text{logit}(\pi(X = 1)) - \text{logit}(\pi(X = 0)) = (\cancel{\beta_0} + \beta_1) - (\cancel{\beta_0}) = \beta_1$$

β_1 is diff of logits comparing $X=1$ to $X=0$

Binary predictor: How do we interpret the coefficient? (2/2)

$$\text{logit}(\pi(X=1)) - \text{logit}(\pi(X=0)) = (\beta_0 + \beta_1) - (\beta_0) = \beta_1$$

$$\beta_1 = \text{logit}(\pi(X=1)) - \text{logit}(\pi(X=0))$$

$$\beta_1 = \log\left(\frac{\pi(X=1)}{1 - \pi(X=1)}\right) - \log\left(\frac{\pi(X=0)}{1 - \pi(X=0)}\right)$$

$$\exp(\beta_1) = \frac{\frac{\pi(X=1)}{1 - \pi(X=1)}}{\frac{\pi(X=0)}{1 - \pi(X=0)}}$$

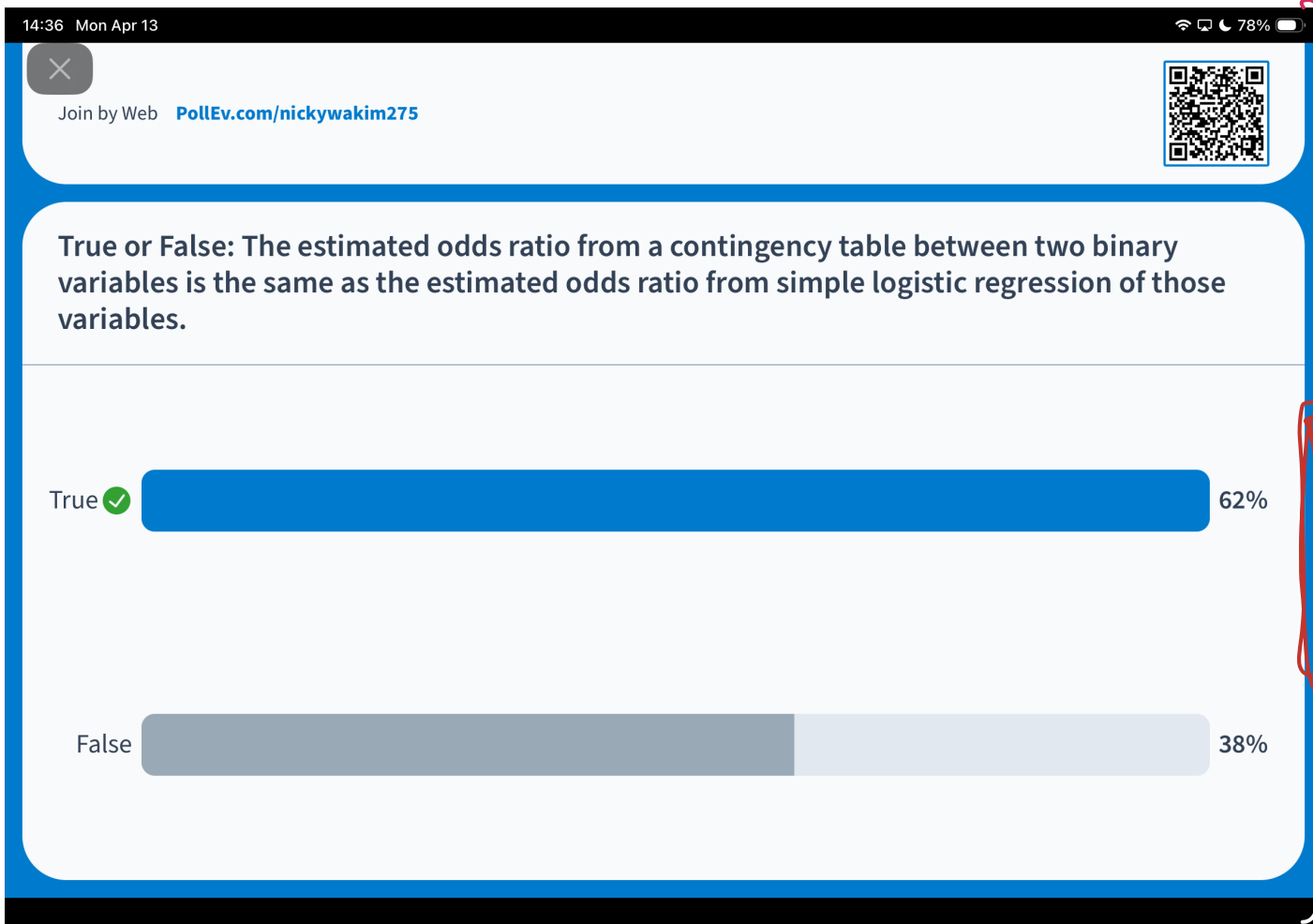
$$\exp(\beta_1) = \frac{\frac{\pi(X=1)}{1 - \pi(X=1)}}{\frac{\pi(X=0)}{1 - \pi(X=0)}}$$

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

$$\log(b) - \log(c) = \log\left(\frac{b}{c}\right)$$

OR of $Y=1$
comparing
 $X=1$ to $X=0$

Poll Everywhere Question 3



Cont table

Diab

gluc

yes

no

imp.

88

20

normal

21

90

$$\text{logit}(\pi(x)) =$$

$$\beta_0 + \beta_1 I(\text{gluc} = \text{imp})$$

$$\exp(\beta_1) =$$

OR for
cont table

Example: Binary age and Late Stage Diagnosis (1/4)

Odds ratio from logistic regression

What is the odds ratio of late stage breast cancer diagnosis for older individuals (>65 years old) compared to younger individuals (≤ 65 years old)?

- Two options to calculate this value:
 - **Option 1:** Calculate $\widehat{OR}$ from 2x2 contingency table
 - Try at home: Refer to [Lesson 3](#) for this process
 - **Option 2:** Calculate $\widehat{OR}$ from logistic regression

Needed steps for **Option 2:**

1. Fit the regression model
2. Transform the coefficients into odds ratios
3. Interpret the odds ratio

	L S B C	
older?	0	1
0	☐	☐
1	☐	☐

Example: Binary age and Late Stage Diagnosis (2/4)

Odds ratio from logistic regression

What is the odds ratio of late stage breast cancer diagnosis for older individuals (>65 years old) compared to younger individuals (≤ 65 years old)?

1. Fit the regression model

```
1 bc3 = bc %>% mutate(Age_binary = ifelse(Age > 65, 1, 0))
2 age_bin_glm = bc3 %>% glm(formula = Late_stage_diag ~ Age_binary, family = binomial)
```

Handwritten notes:
if Age > 65
Age ≤ 65

Example: Binary age and Late Stage Diagnosis (3/4)

Odds ratio from logistic regression

What is the odds ratio of late stage breast cancer diagnosis for older individuals (>65 years old) compared to younger individuals (≤ 65 years old)?

2. Transform the coefficients into odds ratios

```
1 age_bin_tidy = tidy(age_bin_glm, conf.int=T, exponentiate = T)
2 age_bin_tidy %>% gt() %>%
3   tab_options(table.font.size = 35) %>%
4   fmt_number(decimals = 3)
```

term	estimate	std.error	statistic	p.value	conf.low	conf.high
(Intercept)	0.297	0.031	-39.608	0.000	0.280	0.315
Age_binary	1.875	0.045	13.928	0.000	1.716	2.048

↳ OR of LSBC comparing older to younger individuals

Poll Everywhere Question 4

14:45 Mon Apr 13

75%



Join by Web PollEv.com/nickywakim275



Please fill in the blanks for the following interpretation of our $\hat{OR}$: The estimated odds of late stage breast cancer among individuals over 65 years old is 1.87 (95% CI: (1.72, 2.05)) times higher than individuals 65 years or younger.

0.29, higher

20%

0.29, lower

0%

1.72, higher ✓

80%

1.72, lower

0%

Example: Binary age and Late Stage Diagnosis (4/4)

Odds ratio from logistic regression

What is the odds ratio of late stage breast cancer diagnosis for older individuals (>65 years old) compared to younger individuals (≤ 65 years old)?

3. Interpret the odds ratio

The estimated odds of late stage breast cancer among individuals over 65 years old is 1.87 (95% CI: (1.72, 2.05)) times that of individuals 65 years or younger.

1.87 times higher than (1)

$$1.87 \times 1 = 1.87$$

1.87 times higher than 2

$$1.87 \times 2 = 3.74$$

times
no X, higher

2.87

times
no X, higher:

3.87

Learning Objectives

1. Interpret odds ratios from fitted simple logistic regression model for a continuous explanatory variable.
2. Interpret odds ratios from fitted simple logistic regression model for a binary explanatory variable.
3. Interpret odds ratios from fitted simple logistic regression model for a multi-level categorical explanatory variable.
4. Report the odds ratio using a table and/or a forest plot.

Coefficient Interpretation: Multi-group Categorical Variable

- Independent variable X is a multi-level categorical variable

- Let's say X takes values: a, b, c, or d

- We are fitting the simple ^{ish} logistic regression model:

$$\text{logit}(\pi(X)) = \beta_0 + \beta_1 \cdot I(X = b) + \beta_2 \cdot I(X = c) + \beta_3 \cdot I(X = d)$$

- Where a is our reference group

- β_1 represents the change/difference in the log-odds for $X = b$ vs. $X = a$ ^{of $\gamma = 1$}

- AKA log odds ratio comparing $X = b$ to $X = a$

Coefficient Interpretation: Multi-group Categorical Variable

We are fitting the simple logistic regression model with reference group a :

$$\text{logit}(\pi(X)) = \beta_0 + \beta_1 \cdot I(X = b) + \beta_2 \cdot I(X = c) + \beta_3 \cdot I(X = d)$$

- β_0 : the log-odds of event $Y = 1$ for group a
- β_1 : the difference in log-odds of event $Y = 1$ comparing group b to group a
- β_2 : the difference in log-odds of event $Y = 1$ comparing group c to group a
- β_3 : the difference in log-odds of event $Y = 1$ comparing group d to group a

Multi-level categorical: How do we interpret the coefficient? (II)

$$\text{logit}(\pi(X=c)) - \text{logit}(\pi(X=a)) = (\cancel{\beta_0} + \beta_1 \cdot 0 + \beta_2 \cdot 1 + \beta_3 \cdot 0) - (\cancel{\beta_0} + \beta_1 \cdot 0 + \beta_2 \cdot 0 + \beta_3 \cdot 0) = \beta_2$$

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

$$\beta_2 = \text{logit}(\pi(X=c)) - \text{logit}(\pi(X=a))$$

$$\beta_2 = \log\left(\frac{\pi(X=c)}{1 - \pi(X=c)}\right) - \log\left(\frac{\pi(X=a)}{1 - \pi(X=a)}\right)$$

$$\beta_2 = \log\left(\frac{\frac{\pi(X=c)}{1 - \pi(X=c)}}{\frac{\pi(X=a)}{1 - \pi(X=a)}}\right)$$

$$\exp(\beta_2) = \frac{\frac{\pi(X=c)}{1 - \pi(X=c)}}{\frac{\pi(X=a)}{1 - \pi(X=a)}}$$

$$\frac{\text{odds}_c}{\text{odds}_a}$$

odds ratio of $Y=1$
comparing group c
to group a

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

$$\log(j) - \log(i) \\ = \log\left(\frac{j}{i}\right)$$

Coefficient Interpretation: Multi-group Categorical Variable

We are fitting the simple logistic regression model with reference group a :

$$\rightarrow \text{logit}(\pi(X)) = \underline{\beta_0} + \beta_1 \cdot I(X = b) + \beta_2 \cdot I(X = c) + \beta_3 \cdot I(X = d)$$

- $\exp(\beta_0)$: the odds of event $Y = 1$ for group a
- $\exp(\beta_1)$: the odds of event $Y = 1$ for group b is $\exp(\beta_1)$ times the odds of event $Y = 1$ for group a
- $\exp(\beta_2)$: the odds of event $Y = 1$ for group c is $\exp(\beta_2)$ times the odds of event $Y = 1$ for group a
- $\exp(\beta_3)$: the odds of event $Y = 1$ for group d is $\exp(\beta_3)$ times the odds of event $Y = 1$ for group a

- Remember, as soon as we fit the regression model, we talk about the “expected odds” or “estimated odds”

$\exp(\hat{\beta}_i)$: expected or estimated OR

How do we pick the reference group?

- The choice can be more apparent for multi-group categorical independent variables within studies
- For example, if we want to evaluate the association between clinical response and four treatments.
 - The treatment variable has 4 categories: “active treatment A”, “active treatment B”, “active treatment C” and “Placebo treatment”
 - The investigator is interested in comparing each of the three active treatment with the placebo treatment
 - Then the placebo treatment should be picked as the reference group
- In this class, we will try our best to pick a reference from our own knowledge...
 - **but** in general this is a decision I make with someone who has expertise in the field of study

below 300% FPL
50% FPL

Yes →
No

OR of Food inc
for someone
below 300% FPL
compared above 300%

Example: Late stage diagnosis and race and ethnicity

- We chose Non-Hispanic White individuals as reference group

- Underlying health disparities linked to racism in healthcare and in clinical studies

- There is evidence that white individuals receive a certain standard of care that is not paralleled for POC (Mateo and Williams (2021))

Race/Ethnicity	Breast Cancer Diagnosis		Total
	Early Stage	Late Stage	
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

- Something else to consider: if your reference group does not have a lot of individuals, then you may not see any statistical significance!

Example: Late stage diagnosis and race and ethnicity

Odds ratio from logistic regression

What is the odds ratio of late stage breast cancer diagnosis for Non-Hispanic Asian/Pacific Islander individuals compared to Non-Hispanic White individuals?

Needed steps:

1. Fit the regression model
2. Transform the coefficients into odds ratios
3. Interpret the odds ratio

Example: Late stage diagnosis and race and ethnicity

Odds ratio from logistic regression

What is the odds ratio of late stage breast cancer diagnosis for Non-Hispanic Asian/Pacific Islander individuals compared to Non-Hispanic White individuals?

1. Fit the regression model

```
1 RE_glm = bc %>%  
2 glm(formula = Late_stage_diag ~ Race_Ethnicity, family = binomial)
```

$$\text{logit}(\pi(X)) = \beta_0 + \beta_1 \cdot I(X = \text{Hispanic-Latino}) + \beta_2 \cdot I(X = \text{NH American Indian/Alaskan Native}) + \beta_3 \cdot I(X = \text{NH Asian/Pacific Islander}) + \beta_4 \cdot I(X = \text{NH Black})$$

Example: Late stage diagnosis and race and ethnicity

Odds ratio from logistic regression

What is the odds ratio of late stage breast cancer diagnosis for Non-Hispanic Asian/Pacific Islander individuals compared to Non-Hispanic White individuals?

2. Transform the coefficients into odds ratios

```
1 RE_tidy = tidy(RE_glm, conf.int=T, exponentiate = T)  
2 RE_tidy %>% gt() %>%  
3   tab_options(table.font.size = 35) %>%  
4   fmt_number(decimals = 3)
```

↳ to get OR

term	estimate	std.error	statistic	p.value	conf.low	conf.high
(Intercept) <i>odds of Y=1 for NH White</i>	0.372	0.026	-37.553	0.000	0.353	0.392
Race_EthnicityHispanic-Latino	0.968	0.082	-0.398	0.691	0.822	1.135
Race_EthnicityNH American Indian/Alaskan Native	0.948	0.476	-0.111	0.911	0.342	2.287
→ Race_EthnicityNH Asian/Pacific Islander	1.131	0.082	1.497	0.134	0.961	1.327
Race_EthnicityNH Black	1.405	0.070	4.826	0.000	1.223	1.611

Example: Late stage diagnosis and race and ethnicity

Odds ratio from logistic regression

What is the odds ratio of late stage breast cancer diagnosis for Non-Hispanic Asian/Pacific Islander individuals compared to Non-Hispanic White individuals?

3. Interpret the odds ratio

The estimated odds of late stage breast cancer among Non-Hispanic Asian/Pacific Islander individuals is 1.13 (95% CI: (0.96, 1.33)) times that of Non-Hispanic White individuals.

$Y = 1$
↓
odds of
 $Y = 1$

↪ $X = \text{NH White}$

$X = \text{NH Asian/Pacific Islander}$

What if you want to compare other groups?

- What if we want to estimate OR comparing Non-Hispanic Asian Pacific Islander to Non-Hispanic Black individuals?
- **Option 1:** Change reference group and refit the model (maybe the easiest option)
- **Option 2:** Estimate OR using fitted coefficients ($\hat{\beta}$'s) in the current model:

$$\begin{aligned}\log(OR(\text{NH API}, \text{NH B})) &= \text{logit}(\pi(X = \text{NH API})) - \text{logit}(\pi(X = \text{NH B})) \\ &= [\beta_0 + \beta_3 \cdot 1] - [\beta_0 + \beta_4 \cdot 1] \\ \log(\widehat{OR}(\text{NH API}, \text{NH B})) &= \hat{\beta}_3 - \hat{\beta}_4 \\ \widehat{OR}(\text{NH API}, \text{NH B}) &= \exp(\hat{\beta}_3 - \hat{\beta}_4)\end{aligned}$$

Reference: Using `estimable()` in Option 2

- We could use `estimable()` to calculate this linear combination of coefficients (see [BSTA 512 Lesson](#))

```
1 library(gmodels)
2 RE_glm %>% estimable(
3   c("(Intercept)"
4     "Race_EthnicityHispanic-Latino"
5     "Race_EthnicityNH American Indian/Alaskan Native"
6     "Race_EthnicityNH Asian/Pacific Islander"
7     "Race_EthnicityNH Black"
8     conf.int = 0.95) %>%
9   exp(.)
```

$c(0, 0)$

$= 0, \quad \# \text{beta0}$
 $= 0, \quad \# \text{beta1}$
 $= 0, \quad \# \text{beta2}$
 $= 1, \quad \# \text{beta3}$
 $= -1, \quad \# \text{beta4}$

$\rightarrow 1\hat{\beta}_3 + (-1)\hat{\beta}_4$

	Estimate	Std. Error	X^2 value	DF	Pr(> X^2)	Lower.CI
$\rightarrow (0 \ 0 \ 0 \ 1 \ -1)$	0.8051811	1.107021	93.89366	2.718282	1.033623	0.6580966
	Upper.CI					
$\rightarrow (0 \ 0 \ 0 \ 1 \ -1)$	0.9851389					

Poll Everywhere Question 5

13:40 Wed Apr 15

92%

Join by Web PollEv.com/nickywakim275



Will the following code give us the odds ratio between Non-Hispanic Asian/Pacific Islander individuals and Non-Hispanic Black individuals?

$$\beta_2 \text{ I}(x = \text{NH Asian/Pacific Islander})$$
$$\rightarrow \exp(\hat{\beta}_2)$$

Yes ✓

74%

```
1 bc3 = bc %>%
2   mutate(Race_Ethnicity = relevel(Race_Ethnicity, ref = "NH Black"))
3 RE_glm2 = glm(Late_stage_diag ~ Race_Ethnicity, data = bc3,
4               family = binomial)
```

```
1 bc3 = bc %>%
2   mutate(Race_Ethnicity = relevel(Race_Ethnicity, ref = "NH Black"))
3 RE_glm2 = glm(Late_stage_diag ~ Race_Ethnicity, data = bc3,
4               family = binomial)
5 tidy(RE_glm2, conf.int=T, exponentiate = T) %>% gt() %>%
6   tab_options(table.font.size = 38) %>%
7   fmt_number(decimals = 3)
```

$$\text{OR}_{\text{NHAPI vs NHB}} = \exp(\hat{\beta}_2)$$
$$\text{OR}_{\text{NHB v NHAPI}} = \frac{1}{\exp(\hat{\beta}_2)}$$

What if you want to compare other groups? Option 1

```
1 bc3 = bc %>%
2   mutate(Race_Ethnicity = fct_relevel(Race_Ethnicity, "NH Black"))
3 RE_glm2 = glm(Late_stage_diag ~ Race_Ethnicity, data = bc3,
4               family = binomial)
5 tidy(RE_glm2, conf.int=T, exponentiate = T) %>% gt() %>%
6   tab_options(table.font.size = 38) %>%
7   fmt_number(decimals = 3)
```

term	estimate	std.error	statistic	p.value	conf.low	conf.high
(Intercept)	0.523	0.065	-9.934	0.000	0.459	0.594
Race_EthnicityNH White	0.712	0.070	-4.826	0.000	0.621	0.818
Race_EthnicityHispanic-Latino	0.689	0.102	-3.664	0.000	0.564	0.840
Race_EthnicityNH American Indian/Alaskan Native	0.675	0.479	-0.819	0.413	0.242	1.641
Race_EthnicityNH Asian/Pacific Islander	0.805	0.102	-2.131	0.033	0.659	0.982

OR NHAPI vs NHB

Learning Objectives

1. Interpret odds ratios from fitted simple logistic regression model for a continuous explanatory variable.
2. Interpret odds ratios from fitted simple logistic regression model for a binary explanatory variable.
3. Interpret odds ratios from fitted simple logistic regression model for a multi-level categorical explanatory variable.
4. Report the odds ratio using a table and/or a forest plot.

How to present odds ratios: Table

- `tbl_regression()` in the `gtsummary` package is helpful for presenting the odds ratios in a clean way

```
1 library(gtsummary)
2 tbl_regression(RE_glm, exponentiate = TRUE) %>%
3   as_gt() %>% # allows us to use tab_options()
4   tab_options(table.font.size = 38)
```

Characteristic	OR	95% CI	p-value
Race_Ethnicity			
NH White	—	—	
Hispanic-Latino	0.97	0.82, 1.14	0.7
NH American Indian/Alaskan Native	0.95	0.34, 2.29	>0.9
NH Asian/Pacific Islander	1.13	0.96, 1.33	0.13
NH Black	1.40	1.22, 1.61	<0.001

Abbreviations: CI = Confidence Interval, OR = Odds Ratio

How to present odds ratios: Forest Plot Setup

```
1 library(broom.helpers)
2 RE_tidy = tidy_and_attach(RE_glm, conf.int=T, exponentiate = T) %>%
3   tidy_remove_intercept() %>%
4   tidy_add_reference_rows() %>% tidy_add_estimate_to_reference_rows() %>%
5   tidy_add_term_labels()
6 glimpse(RE_tidy)
```

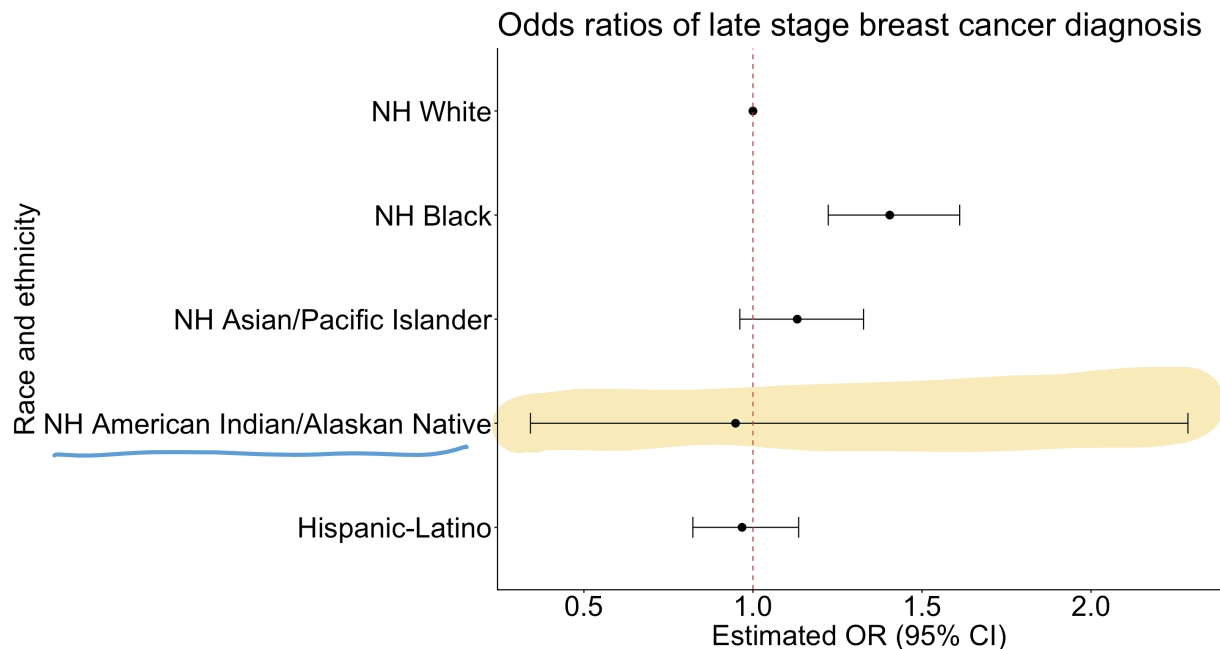
Rows: 5

Columns: 16

\$ term	<chr> "Race_EthnicityNH White", "Race_EthnicityHispanic-Latin...
\$ variable	<chr> "Race_Ethnicity", "Race_Ethnicity", "Race_Ethnicity", "...
\$ var_label	<chr> "Race_Ethnicity", "Race_Ethnicity", "Race_Ethnicity", "...
\$ var_class	<chr> "factor", "factor", "factor", "factor", "factor"
\$ var_type	<chr> "categorical", "categorical", "categorical", "categoric...
\$ var_nlevels	<int> 5, 5, 5, 5, 5
\$ contrasts	<chr> "contr.treatment", "contr.treatment", "contr.treatment"...
\$ contrasts_type	<chr> "treatment", "treatment", "treatment", "treatment", "tr...
\$ reference_row	<lgl> TRUE, FALSE, FALSE, FALSE, FALSE
\$ label	<chr> "NH White", "Hispanic-Latino", "NH American Indian/Alas...
\$ estimate	<dbl> 1.0000000, 0.9678002, 0.9484848, 1.1310170, 1.4046741
\$ std.error	<dbl> NA, 0.08224948, 0.47558680, 0.08224988, 0.07041472
\$ statistic	<dbl> NA, -0.3979312, -0.1112089, 1.4968682, 4.8257715
\$ p.value	<dbl> NA, 6.906809e-01, 9.114507e-01, 1.344276e-01, 1.394623e...
\$ conf.low	<dbl> NA, 0.8223138, 0.3417844, 0.9612074, 1.2226824
\$ conf.high	<dbl> NA, 1.135332, 2.286596, 1.327092, 1.611466

How to present odds ratios: Forest Plot

```
1 ggplot(data=RE_tidy, aes(y=label, x=estimate, xmin=conf.low, xmax=conf.high)) +  
2   geom_point(size = 3) + geom_errorbarh(height=.2) +  
3   geom_vline(xintercept=1, color='#C2352F', linetype='dashed', alpha=1) +  
4   theme_classic() +  
5   labs(x = "Estimated OR (95% CI)", y = "Race and ethnicity",  
6         title = "Odds ratios of late stage breast cancer diagnosis") +  
7   theme(axis.title = element_text(size = 25), axis.text = element_text(size = 25), title =
```



References

- Mateo, Camila M., and David R. Williams. 2021. "Racism: A Fundamental Driver of Racial Disparities in Health-Care Quality." *Nature Reviews Disease Primers* 7 (1): 1–2. <https://doi.org/10.1038/s41572-021-00258-1>.
- Yedjou, Clement G., Jennifer N. Sims, Lucio Miele, et al. 2019. "Health and Racial Disparity in Breast Cancer." *Advances in Experimental Medicine and Biology* 1152: 31–49. https://doi.org/10.1007/978-3-030-20301-6_3.