Class 7: Simple logistic regression and its coefficient interpretations

Announcements

Virtual OH on Wed.

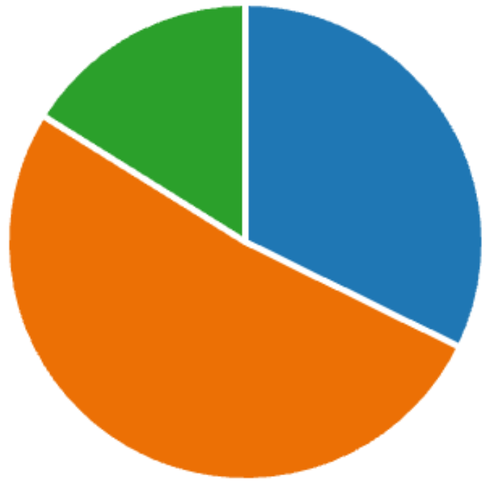
- Office hours tomorrow (4/25) are canceled
- Out of office: will not be responding to emails/Slack
  - Tuesday – flying
  - Thursday – graduation
- Keep posting in the #redos_xcn channel!
  - Really helpful for the poster and others to see different ways to do problems

Announcements

1. I want to get a sense of how the presentation of R code went today. Do you prefer this live coding presentation to previous annotations of code?

More Details

Prefer live coding	10
Prefer annotating code in slides	16
No preference	5



Homework 1 Redo

- Question 2c: Make sure to include distribution of test statistic
  - This is on me
  - If you chose a **chi-squared** test statistic, what distribution does it follow?

np or np(1-p)

Common mistakes

- Question 5: Criteria for different tests in table
  - Think about the **expected cell counts**

`lower.tail` logical; if TRUE (default), probabilities are $P[X \leq x]$, otherwise $P[X > x]$.

- How to use pbinom to calculate P(X>=x)? What does **lower.tail=FALSE** do?
- What to include in a short summary/conclusion of results.
  - Usually, three things need to be covered in your conclusion:
    - Is your test result significant? Compare p-value or test statistic to your designated alpha or critical value
    - Do you reject or fail to reject the null hypothesis?
    - What is the overall conclusion (clinically meaningful statement)?
- For the Cochran-Armitage test for trend, you can state the specific trend (decreased/increased) based on the table

pick dec OR inc

Class 7 Learning Objective

1. Test the significance of estimated coefficients and create confidence interval for each using three tests (Finish up coding for Wald and LRT)

2. Predict the probability of "success" for an individual with specific covariate values

3. Interpret coefficient estimates for binary independent variables in simple logistic regression

4. Interpret coefficient estimates for categorial independent variables with multiple groups in simple logistic regression

5. Interpret coefficient estimates for continuous variables in simple logistic regression

Class 7 Learning Objective

1. Test the significance of estimated coefficients and create confidence interval for each using three tests (Finish up coding for Wald and LRT)

2. Predict the probability of “success” for an individual with specific covariate values

3. Interpret coefficient estimates for binary independent variables in simple logistic regression

4. Interpret coefficient estimates for categorial independent variables with multiple groups in simple logistic regression

5. Interpret coefficient estimates for continuous variables in simple logistic regression

Which test to use?

- All three tests are asymptotically equivalent
  - As sample approaches infinity
- For testing significance of single covariate coefficient:
  - LRT: most people's recommendation
    - Wald and score are only approximations of LRT
      - For smaller samples, LRT better
  - Wald test is very convenient
    - Automatically performed in R
    - Does not need to estimate two models (LRT does)
    - Good for testing separate coefficients for a categorical covariate
  - Score test
    - Does not need to estimate two models (LRT does)
    - I don't really see people use this…

np or $np(1-p)$ > 5 or 10

Wald Test in R

```
summary(r_e_glm)
```

```
##
## Call:
## glm(formula = Late_stage_diag ~ Race_Ethnicity, family = binomial(),
##     data = bc_diagnosis)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.9170  -0.7954  -0.7954   1.4624   1.6394
##
## Coefficients:
##                                                 Estimate Std. Error z value
## (Intercept)                                      -0.98856    0.02632 -37.553
## Race_EthnicityHispanic-Latino                    -0.03273    0.08225  -0.398
## Race_EthnicityNH American Indian/Alaskan Native  -0.05289    0.47559  -0.111
## Race_EthnicityNH Asian/Pacific Islander           0.12312    0.08225   1.497
## Race_EthnicityNH Black                            0.33981    0.07041   4.826
##                                                 Pr(>|z|)
## (Intercept)                                      < 2e-16 ***
## Race_EthnicityHispanic-Latino                      0.691
## Race_EthnicityNH American Indian/Alaskan Native    0.911
## Race_EthnicityNH Asian/Pacific Islander            0.134
## Race_EthnicityNH Black                          1.39e-06 ***
```

Wald test statistic

two sided

Likelihood Ratio Test in R (I)

likelihood of model w/ R/E is sig. greater than model w/out R/E

```
r_e_glm = glm(Late_stage_diag ~ Race_Ethnicity, data = bc_diagnosis,
              family = binomial())
int_glm = glm(Late_stage_diag ~ 1, data = bc_diagnosis,
              family = binomial())
logLik(r_e_glm)
```

$\text{logit}(\pi(x)) = \beta_0$

```
## 'log Lik.' -5918.134 (df=5)
```

```
logLik(int_glm)
```

```
## 'log Lik.' -5930.477 (df=1)
```

Likelihood Ratio Test in R (II)

$$-2 \ln\left(\frac{L(\beta_0)}{L(\beta_0, \beta_1 - \beta_4)}\right) \rightarrow -2\left[\ln(L(\beta_0)) - \ln(L(\beta_0 - \beta_4))\right]$$

```
G = as.numeric(-2*(logLik(int_glm) - logLik(r_e_glm)))
G
```

```
## [1] 24.68761
```

```
p_val = pchisq(G, df = 4, lower.tail = FALSE)
p_val
```

```
## [1] 5.813435e-05
```

$$G \sim \chi^2_{df=4}$$

Δ coef

Confidence Interval for the Coefficients (I)

- The confidence interval for the estimated coefficients (slope and intercept) reported in statistical software is based on the **Wald test**

- The 95% CI for $\hat{\beta}_j$:

$$\hat{\beta}_j \pm 1.96 \text{se}(\hat{\beta}_j)$$

in R output

- The $se(\hat{\beta}_0)$ and $se(\hat{\beta}_1)$ can be found in the output with $\hat{\beta}_0$ and $\hat{\beta}_1$

Where, if anywhere, does the below summary present the standard error for the coefficient corresponding to Non-Hispanic Black individuals?

https://www.polleverywhere.com/clickable_images/U62OjysKlwYuU5YPTQsUa

When poll is active, respond at **PollEv.com/nickywakim275**

Where, if anywhere, does the below summary present the standard error for the coefficient corresponding to Non-Hispanic Black individuals?

```
> summary(r_e_glm)

Call:
glm(formula = Late_stage_diag ~ Race_Ethnicity, family = binomial(),
    data = bc_diagnosis)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.9170  -0.7954  -0.7954   1.4624   1.6394

Coefficients:
                                          Estimate Std. Error z value Pr(>|z|)
(Intercept)                               -0.98856    0.02632 -37.553  < 2e-16 ***
Race_EthnicityHispanic-Latino             -0.03273    0.08225  -0.398    0.691
Race_EthnicityNH American Indian/Alaskan Native -0.05289    0.47559  -0.111    0.911
Race_EthnicityNH Asian/Pacific Islander    0.12312    0.08225   1.497    0.134
Race_EthnicityNH Black                     0.33981    0.07041   4.826 1.39e-06 ***
---
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: 11836  on 9995  degrees of freedom
AIC: 11846

Number of Fisher Scoring iterations: 4
```

Confidence Interval for the Coefficients (II)

- Note: confint() uses profile likelihood method for CI
- For Wald confidence interval, should use confint.default()

```
confint.default(r_e_glm) %>% round(., 3)
```

```
##                                                    2.5 % 97.5 %
## (Intercept)                                       -1.040 -0.937
## Race_EthnicityHispanic-Latino                     -0.194  0.128
## Race_EthnicityNH American Indian/Alaskan Native   -0.985  0.879
## Race_EthnicityNH Asian/Pacific Islander           -0.038  0.284
## Race_EthnicityNH Black                             0.202  0.478
```

Confidence Interval for the Odds Ratio

- As we saw in last class, we take the exponential of the coefficient to get the odds ratio
- We can do the same for the confidence interval

exp(LL_CI), exp(UL_CI)

```
confint.default(r_e_glm) %>% exp(.) %>% round(., 3)
```

```
##                                                    2.5 % 97.5 %
## (Intercept)                                        0.353  0.392
## Race_EthnicityHispanic-Latino                      0.824  1.137
## Race_EthnicityNH American Indian/Alaskan Native    0.373  2.409
## Race_EthnicityNH Asian/Pacific Islander            0.963  1.329
## Race_EthnicityNH Black                             1.224  1.613
```

odds for NH White ind.

Confidence Interval for the Odds Ratio (II)

- Other function for estimated odds ratios

```
library(epiDisplay)
logistic.display(r_e_glm)
```

```
##
## Logistic regression predicting Late_stage_diag : 1 vs 0
##
##                                                 OR(95%CI)          P(Wald's test)
## Race_Ethnicity: ref.=NH White
##    Hispanic-Latino                              0.97 (0.82,1.14)   0.691
##    NH American Indian/Alaskan Native            0.95 (0.37,2.41)   0.911
##    NH Asian/Pacific Islander                    1.13 (0.96,1.33)   0.134
##    NH Black                                     1.4 (1.22,1.61)    < 0.001
##
##                                                 P(LR-test)
## Race_Ethnicity: ref.=NH White                   < 0.001
```

estimate is gone → Never get OR for Reference

vs. NH white individuals

Confidence Interval for the Odds Ratio (III)

- As we saw in last class, we take inverse OR
- We can do the same for the confidence interval

Now we compare: NH White 1.03x the odds of H/L

OR_CI

```
OR = confint.default(r_e_glm) %>% exp(.)
round(1/OR, 2)
```

1/OR

```
##                                                    2.5 % 97.5 %
## (Intercept)                                         2.83   2.55
## Race_EthnicityHispanic-Latino                       1.21   0.88   1.03
## Race_EthnicityNH American Indian/Alaskan Native     2.68   0.42   1.06
## Race_EthnicityNH Asian/Pacific Islander             1.04   0.75
## Race_EthnicityNH Black                              0.82   0.62
```

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 **(95% CI: (0.35, 0.39))**
$\exp(\hat{\beta}_1) = 0.97$	For Hispanic-Latinx individuals, the **odds of diagnosis** of late-stage breast cancer is **1.03 times lower (95% CI: (0.88, 1.21))** 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 (95% CI: (0.42, 2.68))** 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 (95% CI: (0.96 1.33))** 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 (95% CI: (1.22 1.61))** than Non-Hispanic White individuals.

Class 7 Learning Objective

1. Test the significance of estimated coefficients and create confidence interval for each using three tests (Finish up coding for Wald and LRT)

2. Predict the probability of “success” for an individual with specific covariate values

3. Interpret coefficient estimates for binary independent variables in simple logistic regression

4. Interpret coefficient estimates for categorial independent variables with multiple groups in simple logistic regression

5. Interpret coefficient estimates for continuous variables in simple logistic regression

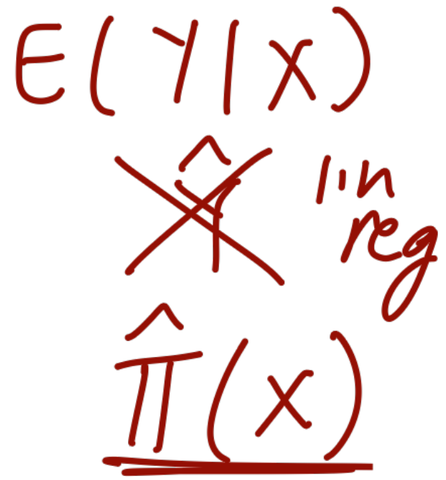
Predicted Probability (I)

- We may be interested in predicting probability of having a late stage breast cancer diagnosis for Non-Hispanic Asian and Pacific Islander individuals.
- The **predicted probability** is the estimated probability of having the event for given values of covariate(s)
- In **simple logistic regression**, the predicted probability is:

$$\hat{\pi}(x_i) = \frac{\exp(\hat{\beta}_0 + \hat{\beta}_1 x_{1,i})}{1 + \exp(\hat{\beta}_0 + \hat{\beta}_1 x_{1,i})} = \frac{\exp(\hat{g}(x))}{1 + \exp(\hat{g}(x))}$$

So we use $\hat{g}(x)$ to represent the fitted *systematic component*:

$$\hat{g}(x) = \hat{\beta}_0 + \hat{\beta}_1 x_{1,i}$$



Predicted Probability (II)

- Let's calculate the predicted probability of having a late stage breast cancer diagnosis for Non-Hispanic Asian and Pacific Islander individuals:

$$\hat{\pi}(x_i) = \frac{\exp(-0.99 - 0.03I(R\_E_i = H{-}L) - 0.05I(R\_E_i = NH\ AIAN) + 0.12I(R\_E_i = NH\ API) + 0.34I(R\_E_i = NH\ B))}{1 + \exp(-0.99 - 0.03I(R\_E_i = H{-}L) - 0.05I(R\_E_i = NH\ AIAN) + 0.12I(R\_E_i = NH\ API) + 0.34I(R\_E_i = NH\ B))}$$

$$\hat{\pi}(x_i) = \frac{\exp(-0.99 - 0.03\times0 - 0.05\times0 + 0.12\times1 + 0.34\times0)}{1 + \exp(-0.99 - 0.03\times0 - 0.05\times0 + 0.12\times1 + 0.34\times0)}$$

$\beta_0 + \beta_{NHAPI}$

$$\hat{\pi}(x_i) = \frac{\exp(-0.99 + 0.12\times1)}{1 + \exp(-0.99 + 0.12\times1)} = 0.296$$

Predicted Probability (III)

- Predicted probability is NOT our predicted outcome
    - We cannot interpret it as the predicted Y for individuals with certain covariate values
    - Example: our predicted probability does not tell us that one individual is or is not diagnosed with late stage breast cancer
- The 0.296 is the estimate of the mean (i.e., proportion) of Non-Hispanic Asian and Pacific Islander individuals who are diagnosed with late stage breast cancer

Confidence Interval for Predicted Probability (I)

- We first construct the **95% confidence interval for $\hat{g}(x)$**:

$$\hat{g}(x) \pm 1.96se[\hat{g}(x)] = [\hat{g}(x)_L, \hat{g}(x)_U]$$

- Once the confidence intervals for $\hat{g}(x)$ is obtained, we can **transform it to get the 95% CI for $\hat{\pi}(x)$**

- Since $\hat{\pi}(x) = \frac{\exp(\hat{g}(x))}{1+\exp(\hat{g}(x))}$, its 95% CI is:

$$\left(\frac{\exp(\hat{g}(x)_L)}{1+\exp(\hat{g}(x)_L)}, \frac{\exp(\hat{g}(x)_U)}{1+\exp(\hat{g}(x)_U)}\right)$$

Confidence Interval for Predicted Probability (II)

- To get a confidence interval for $\hat{g}(x)$, need to compute $\text{se}[\hat{g}(x)]$

- Notice that $\hat{g}(x) = \hat{\beta}_0 + \hat{\beta}_1 x_{1,i}$
  - If we assume $\hat{\beta}_0$ and $\hat{\beta}_1$ both follow normal distribution, then $\hat{g}(x)$ also follows normal distribution

- The variance of $\hat{g}(x)$ for a given x is:

$$\begin{aligned}\text{var}\big(\hat{g}(x)\big) &= \text{var}\big(\hat{\beta}_0 + \hat{\beta}_1 x\big) \\ &= \text{var}\big(\hat{\beta}_0\big) + \text{var}\big(\hat{\beta}_1 x\big) + 2\text{cov}\big(\hat{\beta}_0, \hat{\beta}_1 x\big) \\ &= \text{var}\big(\hat{\beta}_0\big) + \text{x}^2\text{var}\big(\hat{\beta}_1\big) + 2\text{cov}\big(\hat{\beta}_0, \hat{\beta}_1\big)\end{aligned}$$

- The standard error for $\hat{g}(x)$ is:

$$\text{se}[\hat{g}(x)] = \sqrt{\text{var}[\hat{g}(x)]}$$

Important facts: $\text{var}(X + Y) = \text{var}(X) + \text{var}(Y) + 2\text{cov}(X, Y)$
-and- $\text{var}(aX) = a^2\text{var}(X)$

3 9

Confidence Interval for Predicted Probability (III)

- Then where do we get the variance and covariance estimates for βs?
- From analysis output (after fitting the simple logistic regression model)

```
age_glm = r_e_glm = glm(Late_stage_diag ~ Age, data = bc_diagnosis,
                        family = binomial())
vcov(age_glm) %>% round(., 6)
```

```
##                (Intercept)      Age
## (Intercept)       0.041616 -0.00065
## Age              -0.000650  0.00001
```

$\text{var}(\hat{\beta}_0)$ $\text{var}(\hat{\beta}_1)$ $\text{cov}(\hat{\beta}_0, \hat{\beta}_1)$

Where in this variance-covariance matrix is the variance for the coefficient corresponding to Hispanic-Latinx individuals? (This is a hard output to understand!)

https://www.polleverywhere.com/clickable_images/VWijEOMr8GfOTbTAXQAaw

Respond at **PollEv.com/nickywakim275**

Where in this variance-covariance matrix is the variance for the coefficient corresponding to Hispanic-Latinx individuals? (This is a hard output to understand!)

6 × 6 5 × 5

```
> vcov(r_e_glm) %>% round(., 6)
                                                (Intercept) Race_EthnicityHispanic-Latino
(Intercept)                                        0.000693                     -0.000693
Race_EthnicityHispanic-Latino                     -0.000693                      0.006765
Race_EthnicityNH American Indian/Alaskan Native   -0.000693                      0.000693
Race_EthnicityNH Asian/Pacific Islander           -0.000693                      0.000693
Race_EthnicityNH Black                            -0.000693                      0.000693
                                                Race_EthnicityNH American Indian/Alaskan Native
(Intercept)                                                                           -0.000693
Race_EthnicityHispanic-Latino                                                          0.000693
Race_EthnicityNH American Indian/Alaskan Native                                        0.226183
Race_EthnicityNH Asian/Pacific Islander                                                0.000693
Race_EthnicityNH Black                                                                 0.000693
                                                Race_EthnicityNH Asian/Pacific Islander
(Intercept)                                                                   -0.000693
Race_EthnicityHispanic-Latino                                                  0.000693
Race_EthnicityNH American Indian/Alaskan Native                                0.000693
Race_EthnicityNH Asian/Pacific Islander                                        0.006765
Race_EthnicityNH Black                                                         0.000693
                                                Race_EthnicityNH Black
(Intercept)                                                 -0.000693
Race_EthnicityHispanic-Latino                                0.000693
Race_EthnicityNH American Indian/Alaskan Native              0.000693
Race_EthnicityNH Asian/Pacific Islander                      0.000693
Race_EthnicityNH Black                                       0.004958
```

0.006765 = $var(\beta_{H/L})$

Powered by Poll Everywhere

Confidence Interval for Predicted Probability (I)

- We first construct the 95% confidence interval for $\hat{g}(x)$:

$$\hat{g}(x) \pm 1.96se[\hat{g}(x)] = [\hat{g}(x)_L, \hat{g}(x)_U]$$

- Once the confidence intervals for $\hat{g}(x)$ is obtained, we can transform it to get the 95% CI for $\hat{\pi}(x)$
- Since $\hat{\pi}(x) = \frac{\exp(\hat{g}(x))}{1+\exp(\hat{g}(x))}$, its 95% CI is:

$$\left(\frac{\exp(\hat{g}(x)_L)}{1+\exp(\hat{g}(x)_L)}, \frac{\exp(\hat{g}(x)_U)}{1+\exp(\hat{g}(x)_U)}\right)$$

Confidence Interval for the Predicted Probability (IV)

```
newdata = with(bc_diagnosis, data.frame(Race_Ethnicity = "NH Asian/Pacific Islander"))
pred = predict(r_e_glm, newdata, se.fit = T, type="response")

LL_CI = pred$fit - qnorm(1-0.05/2) * pred$se.fit
UL_CI = pred$fit + qnorm(1-0.05/2) * pred$se.fit
c(Pred = pred$fit, LL = LL_CI, UL = UL_CI)
```

predicted prob

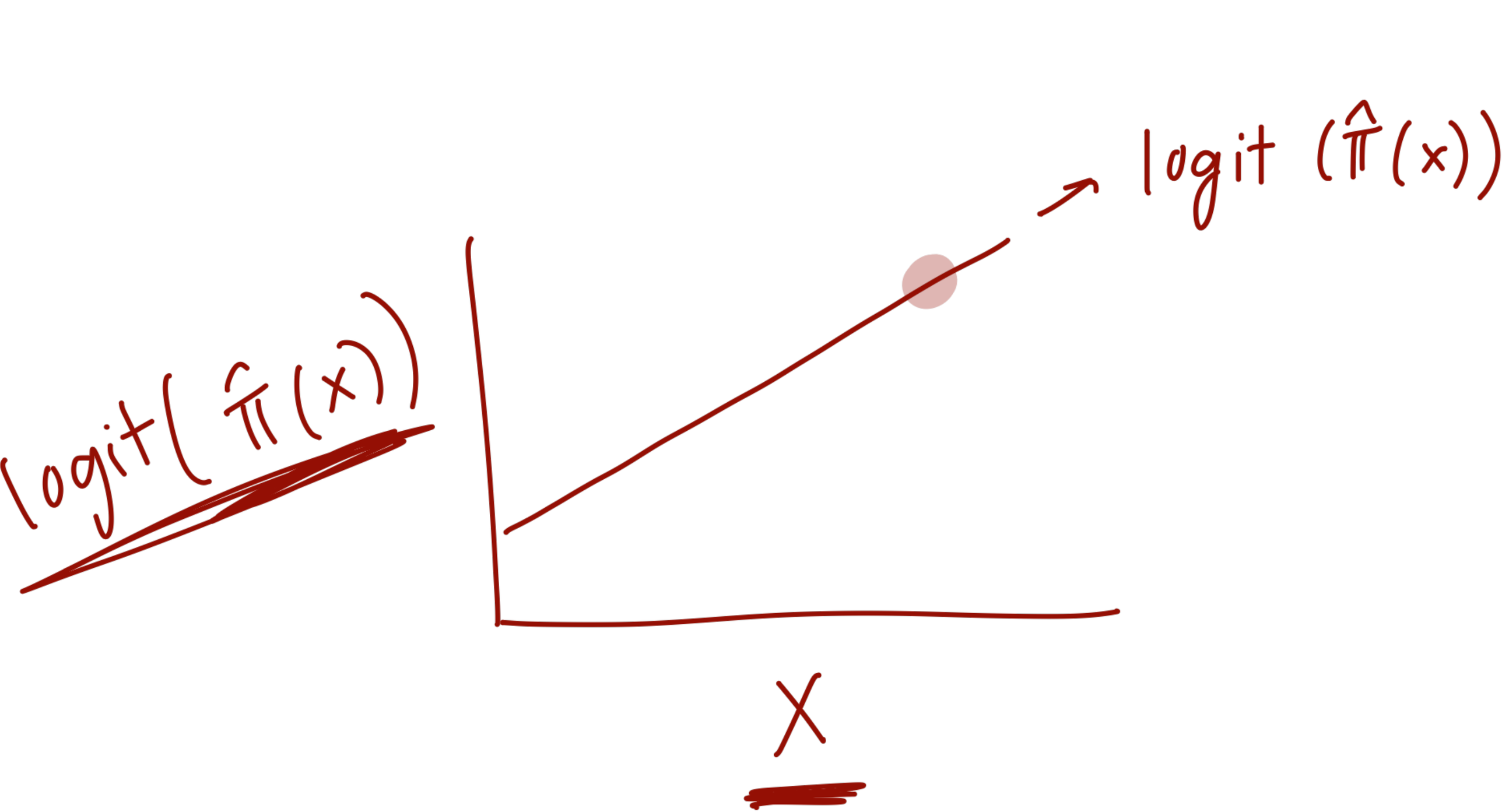
1.96

Make more notes on this.

```
##     Pred.1       LL.1       UL.1
## 0.2962025 0.2643640 0.3280410
```

95% confident pop. proportion is b/w

The predicted probability of late stage breast cancer diag. is 0.30 (95% CI: 0.26, 0.33) for NH API individuals.


logit $(\hat{\pi}(x))$
$\rightarrow$ logit $(\hat{\pi}(x))$
X

Class 7 Learning Objective

★ interpret coeff: log oddr ★ 〉

1. Test the significance of estimated coefficients and create confidence interval for each using three tests (Finish up coding for Wald and LRT)

2. Predict the probability of “success” for an individual with specific covariate values

3. Interpret coefficient estimates for binary independent variables in simple logistic regression

4. Interpret coefficient estimates for categorial independent variables with multiple groups in simple logistic regression

5. Interpret coefficient estimates for continuous variables in simple logistic regression

Intro/Recap of Interpreting Fitted Model

- Interpret coefficients from fitted logistic regression model
  - **Goodness-of-fit of model should be assessed before summarizing findings**
  - 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
- 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

Coefficient Interpretation: **Binary** Independent Variable

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

- The logit difference is β_1 for binary independent variable
  - β_1 represents the change/difference in the logit for $x = 1$ vs. $x = 0$

- Like last week, it will be much easier if we can interpret the coefficient using odds ratio (OR)

Binary: How do we interpret the coefficient? (I)

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

For individuals with $x_i = 0$:

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

($\beta_1 \times (0) = 0$)

For individuals with $x_i = 1$:

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

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

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

Binary: How do we interpret the coefficient? (II)

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

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

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

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

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

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

$\log(F) - \log(E)$

$\log\left(\frac{F}{E}\right)$

OR $x=1$ vs. $x=0$

Review of Odds Ratio

- **Odds** for a subject with $x = 1$:

$$\frac{\pi(x_i = 1)}{1 - \pi(x_i = 1)}$$

- **Odds** for a subject with $x = 0$

$$\frac{\pi(x_i = 0)}{1 - \pi(x_i = 0)}$$

- **Odds Ratio** for $x = 1$ vs. $x = 0$

$$OR = \frac{\frac{\pi(x_i = 1)}{1 - \pi(x_i = 1)}}{\frac{\pi(x_i = 0)}{1 - \pi(x_i = 0)}}$$

How does this relate to a 2x2 table?

- 2x2 table with the respective logistic functions in each cell

Outcome Variable (Y)	Independent Variable (X)	
	$x = 1$	$x = 0$
$y = 1$	$\pi(1) = \frac{\exp(\beta_0 + \beta_1)}{1 + \exp(\beta_0 + \beta_1)}$ a	$\pi(0) = \frac{\exp(\beta_0)}{1 + \exp(\beta_0)}$ b
$y = 0$	$1 - \pi(1) = \frac{1}{1 + \exp(\beta_0 + \beta_1)}$ c	$1 - \pi(0) = \frac{1}{1 + \exp(\beta_0)}$ d
Total	1.0	1.0

Recall:

$\pi(1) = \pi(x_i = 1)$
$= P(Y_i = 1 | x_i = 1)$

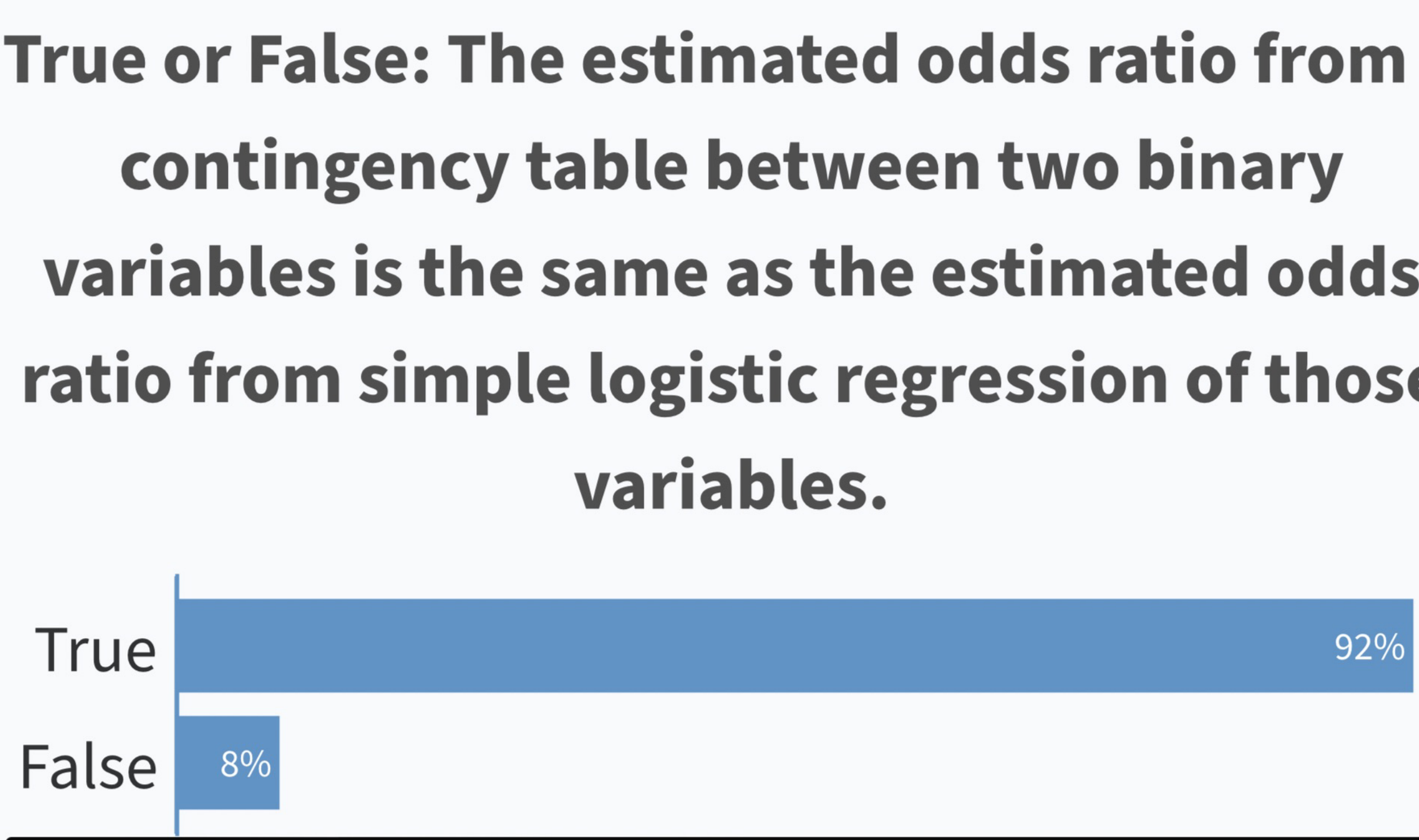
$\pi(0) = \pi(x_i = 0)$
$= P(Y_i = 1 | x_i = 0)$

True or False: The estimated odds ratio from a contingency table between two binary variables is the same as the estimated odds ratio from simple logistic regres…

https://www.polleverywhere.com/multiple_choice_polls/5YFUWk0zCZp0iuyq221ta

Respond at **PollEv.com/nickywakim275**

True or False: The estimated odds ratio from a contingency table between two binary variables is the same as the estimated odds ratio from simple logistic regression of those variables.



3 / 6 Instructions Responses Correctness More EXIT

How does this relate to a 2x2 table?

$$\text{OR} = \frac{a/c}{b/d} = \frac{\left(\frac{\exp(\beta_0+\beta_1)}{1+\exp(\beta_0+\beta_1)}\right)}{\left(\frac{1}{1+\exp(\beta_0+\beta_1)}\right)} \Big/ \frac{\left(\frac{\exp(\beta_0)}{1+\exp(\beta_0)}\right)}{\left(\frac{1}{1+\exp(\beta_0)}\right)} = \frac{\exp(\beta_0+\beta_1)}{\exp(\beta_0)} = e^{\beta_1}$$

$$\exp(\beta_0+\beta_1-\beta_0)$$

- Simple relationship between coefficient and odds ratio is a primary reason why we report OR for categorical data analysis.
- For binary independent variable x, **OR computed in logistic regression** model is the **same as** **OR computed using contingency table**

Example: Binary age and Late Stage Diagnosis (I)

- 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

- **Option 2:** Calculate $\widehat{OR}$ from logistic regression

Example: Binary age and Late Stage Diagnosis (II)

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

Option 1: Calculate $\widehat{OR}$ from 2x2 contingency table

```
bc_diagnosis2 = bc_diagnosis %>% mutate(Age_binary = ifelse(Age > 65, 1, 0))
bc_table = table(cancer = bc_diagnosis2$Late_stage_diag, age = bc_diagnosis2$Age_bi
nary)

epitools::oddsratio(bc_table, method="wald")$measure
```

Example: Binary age and Late Stage Diagnosis (II)

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

Option 1: Calculate $\widehat{OR}$ from 2x2 contingency table

```
bc_diagnosis2 = bc_diagnosis %>% mutate(Age_binary = ifelse(Age > 65, 1, 0))
bc_table = table(cancer = bc_diagnosis2$Late_stage_diag, age = bc_diagnosis2$Age_bi
nary)

epitools::oddsratio(bc_table, method="wald")$measure
```

```
##        odds ratio with 95% C.I.
## cancer estimate    lower    upper
##      0 1.000000       NA       NA
##      1 1.874623 1.715971 2.047944
```

Example: Binary age and Late Stage Diagnosis (III)

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

Option 2: Calculate $\widehat{OR}$ from logistic regression

```
age_bin_glm = glm(Late_stage_diag ~ Age_binary, data = bc_diagnosis2,
                  family = binomial)
exp(cbind(OR = coef(age_bin_glm), confint(age_bin_glm)))
```

Example: Binary age and Late Stage Diagnosis (III)

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

Option 2: Calculate $\widehat{OR}$ from logistic regression

```
age_bin_glm = glm(Late_stage_diag ~ Age_binary, data = bc_diagnosis2,
                  family = binomial)
exp(cbind(OR = coef(age_bin_glm), confint(age_bin_glm)))
```

```
##                OR      2.5 %    97.5 %
## (Intercept) 0.2970531 0.2796429 0.3153414
## Age_binary  1.8746234 1.7160349 2.0480433
```

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…

https://www.polleverywhere.com/multiple_choice_polls/sLqk9CDjSUGDBJ6MjEBa5

Respond at **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: (_____, 2.05)) times _____than individuals 65 years or younger.

0.29 = log OR

- 0.29, higher
- 0.29, lower
- 1.72, higher
- 1.72, lower

Total Results: 0

Powered by Poll Everywhere

$\frac{1}{1.72}$ → lower

65 yrs younger vs. ≥ 65 yrs

Example: Binary age and Late Stage Diagnosis (I)

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

Interpretation: 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.

*__Note:__ because we are not asked to test whether or not the estimate OR is significant, we can leave our conclusion as above.

Computing OR from β (I)

- $\widehat{OR} = \exp(\hat{\beta}_1)$ if binary independent variable is **coded as 0 or 1**
- Consider a general case that the x variable can take one of two values: $x = a$ or $x = b$

$$\ln\left[\widehat{OR}(a,b)\right] = \hat{g}(x=a) - \hat{g}(x=b)$$
$$= \left(\hat{\beta}_0 + \hat{\beta}_1 \times a\right) - \left(\hat{\beta}_0 + \hat{\beta}_1 \times b\right) = \hat{\beta}_1 \times (a-b)$$

$$\widehat{OR}(a,b) = \exp\left[\hat{\beta}_1 \times (a-b)\right]$$

- If $(a-b) = 1$, then $\widehat{OR}(a,b) = \exp(\hat{\beta}_1)$

- Most health sciences literature will use the **reference cell coding** $(x = 0)$
  - **Always justify your choice of referent category!**

Class 7 Learning Objective

1. Test the significance of estimated coefficients and create confidence interval for each using three tests (Finish up coding for Wald and LRT)

2. Predict the probability of “success” for an individual with specific covariate values

3. Interpret coefficient estimates for binary independent variables in simple logistic regression

4. Interpret coefficient estimates for categorial independent variables with multiple groups in simple logistic regression

5. Interpret coefficient estimates for continuous variables in simple logistic regression

Coefficient Interpretation: **Multi-group Categorical** Variable

- Categorical variable with more than two categories

- We discussed this in Class 5, Slides 42-54

- How to pick the reference group?

	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 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

Example: Late stage diagnosis and race and ethnicity

- Chose Non-Hispanic White individuals as reference group
- Rooted in underlying health disparities linked to racism
- There is evidence that white individuals receive a certain standard of care

	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

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6941147/
https://www.nature.com/articles/s41572-021-00258-1

Example: Late stage diagnosis and race and ethnicity

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

Example: Late stage diagnosis and race and ethnicity

```
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
```

Example: Late stage diagnosis and race and ethnicity

- Let's convert to estimate OR with the confidence interval

```
exp(cbind(OR = coef(r_e_glm), confint(r_e_glm)))
```

Example: Late stage diagnosis and race and ethnicity

- Let's convert to estimate OR with the confidence interval

```
exp(cbind(OR = coef(r_e_glm), confint(r_e_glm)))
```

```
##                                                        OR      2.5 %     97.5 %
## (Intercept)                                     0.3721105 0.3533253 0.3917351
## Race_EthnicityHispanic-Latino                   0.9678002 0.8223138 1.1353316
## Race_EthnicityNH American Indian/Alaskan Native 0.9484848 0.3417844 2.2865956
## Race_EthnicityNH Asian/Pacific Islander         1.1310170 0.9612074 1.3270918
## Race_EthnicityNH Black                          1.4046741 1.2226824 1.6114661
```

odd

What if you want to compare other groups?

- What if we want to estimate OR of Non-Hispanic Asian Pacific Islander and Non-Hispanic Black individuals?

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

  - **Option 1:** Change reference group and refit the model (maybe the easiest option)
  - **Option 2:** Estimate OR using $\hat{\beta}$s provided in the model

$$\ln[\text{OR}(\text{NH API}, \text{NH Black})] = \hat{g}(\text{NH API}) - \hat{g}(\text{NH Black})$$
$$= [\beta_0 + \beta_3 \times 1] - [\beta_0 + \beta_4 \times 1] = \beta_3$$
$$\ln\left[\widehat{\text{OR}}(\text{NH API}, \text{NH Black})\right] = \hat{\beta}_3 - \hat{\beta}_4$$

$$\widehat{\text{OR}}(\text{NH API}, \text{NH Black}) = \exp\left(\hat{\beta}_3 - \hat{\beta}_4\right)$$

What if you want to compare other groups?

$$\widehat{\text{OR}}(\text{NH API}, \text{NH Black}) = \exp(\hat{\beta}_3 - \hat{\beta}_4)$$

var(X-Y) → var($\hat{\beta}_3 - \hat{\beta}_4$)

var($\hat{\beta}_3$) = var($\hat{\beta}_{NHAPI}$)

var($\hat{\beta}_4$) = var($\hat{\beta}_{NHB}$)

```
coefs = coef(r_e_glm)
var_cov = vcov(r_e_glm)
se_diff = sqrt(var_cov[4,4] + var_cov[5,5] - 2*var_cov[4,5])
OR_diff = round(exp(cbind(OR = (coefs[4]-coefs[5]),
          LL_CI = coefs[4]-coefs[5] - 1.96*se_diff,
          UL_CI = coefs[4]-coefs[5] + 1.96*se_diff)), 3)
rownames(OR_diff) = "OR between NH API and NH B  "
OR_diff
```

cov($\hat{\beta}_3, \hat{\beta}_4$)

var($\hat{\beta}_3 - \hat{\beta}_4$) = var($\hat{\beta}_3$) + var($\hat{\beta}_4$) − 2cov($\hat{\beta}_3, \hat{\beta}_4$)

What if you want to compare other groups?

$$\widehat{\text{OR}}(\text{NH API}, \text{NH Black}) = \exp\left(\hat{\beta}_3 - \hat{\beta}_4\right)$$

```
coefs = coef(r_e_glm)
var_cov = vcov(r_e_glm)
se_diff = sqrt(var_cov[4,4] + var_cov[5,5] - 2*var_cov[4,5])
OR_diff = round(exp(cbind(OR = (coefs[4]-coefs[5]),
            LL_CI = coefs[4]-coefs[5] - 1.96*se_diff,
            LL_CI = coefs[4]-coefs[5] + 1.96*se_diff)), 3)
rownames(OR_diff) = "OR between NH API and NH B  "
OR_diff
```

```
##                                  OR LL_CI LL_CI
## OR between NH API and NH B    0.805  0.66 0.983
```

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

releve l ( )

```
bc_diagnosis3 = bc_diagnosis %>%
  mutate(Race_Ethnicity = relevel(Race_Ethnicity, ref = "NH Black"))
r_e_glm_2 = glm(Late_stage_diag ~ Race_Ethnicity, data = bc_diagnosis3,
                family = binomial())
logistic.display(r_e_glm_2)
```

- Yes
- No

Total Results: 24

Poll Everywhere Question 5

```
bc_diagnosis3 = bc_diagnosis %>%
  mutate(Race_Ethnicity = relevel(Race_Ethnicity, ref = "NH Black"))
r_e_glm_2 = glm(Late_stage_diag ~ Race_Ethnicity, data = bc_diagnosis3,
                family = binomial())
logistic.display(r_e_glm_2)
```

```
##
## Logistic regression predicting Late_stage_diag : 1 vs 0
##
##                                          OR(95%CI)          P(Wald's test)
## Race_Ethnicity: ref.=NH Black
##    NH White                              0.71 (0.62,0.82)   < 0.001
##    Hispanic-Latino                       0.69 (0.56,0.84)   < 0.001
##    NH American Indian/Alaskan Native     0.68 (0.26,1.73)   0.413
##    NH Asian/Pacific Islander             0.81 (0.66,0.98)   0.033
##
##                                          P(LR-test)
## Race_Ethnicity: ref.=NH Black            < 0.001
##    NH White
##    Hispanic-Latino
##    NH American Indian/Alaskan Native
##    NH Asian/Pacific Islander
##
## Log-likelihood = -5918.1336
## No. of observations = 10000
## AIC value = 11846.2672
```

Class 7 Learning Objective

1. Test the significance of estimated coefficients and create confidence interval for each using three tests (Finish up coding for Wald and LRT)

2. Predict the probability of “success” for an individual with specific covariate values

3. Interpret coefficient estimates for binary independent variables in simple logistic regression

4. Interpret coefficient estimates for categorial independent variables with multiple groups in simple logistic regression

5. Interpret coefficient estimates for continuous variables in simple logistic regression

Coefficient Interpretation: **Continuous** Independent Variable (I)

- For simplicity, we assume the linear relationship between logit and continuous variable x

- Again using simple logistic regression model to illustrate the interpretation of β for a continuous variable x

$$\hat{g}(x) = \hat{\beta}_0 + \hat{\beta}_1 x$$

- 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

Coefficient Interpretation: **Continuous** Independent Variable (II)

- 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), than 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

Coefficient Interpretation: **Continuous** Independent Variable (III)

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

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

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

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

$$\exp(c\hat{\beta}_1 \pm 1.96 * c * se(\hat{\beta}_1))$$

- 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: Age and Late Stage Diagnosis (I)

```
age_glm = glm(Late_stage_diag ~ Age, data = bc_diagnosis,
              family = binomial())
summary(age_glm)
```

Example: Age and Late Stage Diagnosis (I)

```
age_glm = glm(Late_stage_diag ~ Age, data = bc_diagnosis,
              family = binomial())
summary(age_glm)
```

```
##
## Call:
## glm(formula = Late_stage_diag ~ Age, family = binomial(), data = bc_diagnosis)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.0816  -0.8637  -0.7140   1.3629   2.4192
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -4.504457   0.204000  -22.08   <2e-16 ***
## Age          0.056965   0.003204   17.78   <2e-16 ***
## ---
```

Example: Age and Late Stage Diagnosis (II)

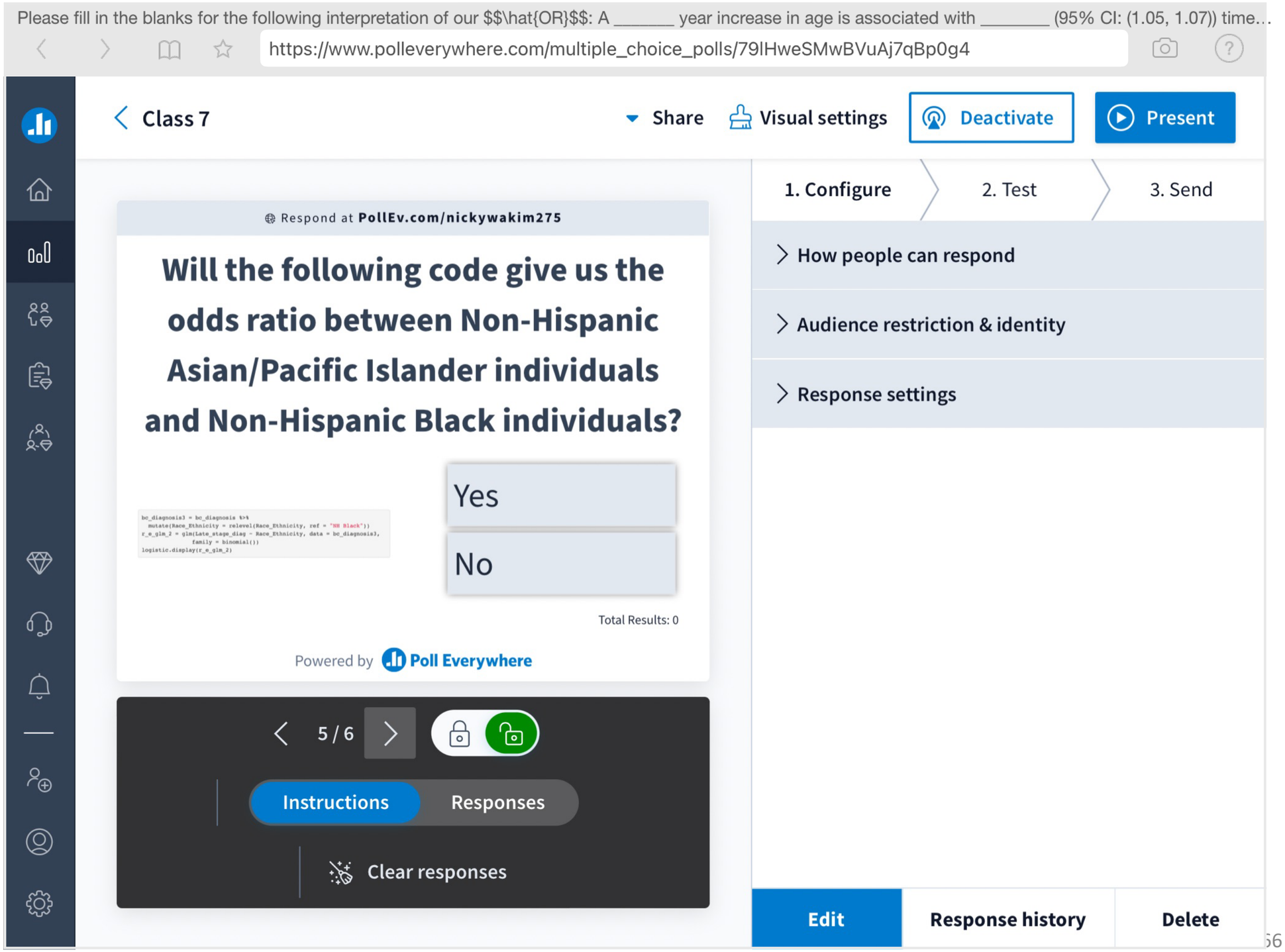
```
logistic.display(age_glm)
```

Example: Age and Late Stage Diagnosis (II)

```
logistic.display(age_glm)
```

```
##
## Logistic regression predicting Late_stage_diag : 1 vs 0
##
##                     OR(95%CI)          P(Wald's test) P(LR-test)
## Age (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 6



Example: Interpretation of Age Coefficient/OR (I)

- $\hat{\beta}_1$ is 0.057, suggesting that one year increase in age is associated with 0.057 increase in log odds of receiving a late stage breast cancer diagnosis

- $\exp(\hat{\beta}_1)$ is 1.06, suggesting that one year increase in age is associated with 1.06 times the odds of receiving a late stage breast cancer diagnosis

- For continuous covariates in logistic regression model, it is helpful to subtract 1 from the odds ratio and multiply by 100 to obtain the percentage change in odds for 1-unit increase.

- The estimated OR for age is 1.06, suggesting that a 1-year increase in age is associated with a 6% increase in the predicted odds of late stage diagnosis in the patient population

Example: Interpretation of Age Coefficient/OR (II)

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

$$\widehat{OR}(10) = \exp\left(10 * \hat{\beta}_1\right) = \exp(0.56965) = 1.767$$

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

$$\exp\left(10 * \hat{\beta}_1 \pm 1.96 * 10 * se\left(\hat{\beta}_1\right)\right)$$
$$= \exp(10 * 0.056965 \pm 1.96 * 10 * 0.003204)$$
$$= (1.66, 1.88)$$

Example: Interpretation of Age Coefficient/OR (III)

```
age_10 = estimable(age_glm, c(0,10))
age_10
```

```
##         Estimate Std. Error X^2 value DF Pr(>|X^2|)
## (0 10) 0.569645   0.032039   316.119  1           0
```

```
OR_age_10 = exp(c(OR = age_10$Estimate,
                  L_CI = age_10$Estimate - 1.96*age_10$`Std. Error`,
                  U_CI = age_10$Estimate + 1.96*age_10$`Std. Error`))
OR_age_10
```

```
##       OR     L_CI     U_CI
## 1.767639 1.660051 1.882200
```

Class 7 Learning Objective

1. Test the significance of estimated coefficients and create confidence interval for each using three tests (Finish up coding for Wald and LRT)

2. Predict the probability of “success” for an individual with specific covariate values

3. Interpret coefficient estimates for binary independent variables in simple logistic regression

4. Interpret coefficient estimates for categorial independent variables with multiple groups in simple logistic regression

5. Interpret coefficient estimates for continuous variables in simple logistic regression

Wrap-up

*don't worry too much about Q5!!

- 4-minute exit ticket

- Links to further information on health disparities among different race and ethnicity groups
  - https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6941147/
  - https://www.nature.com/articles/s41572-021-00258-1

- **Next class**
  - Multiple logistic regression and coefficient interpretations!!

Class 7 Exit Ticket

https://forms.office.com/r/Ri4ukc1XyH