Multiple logistic regression and interpretations

β_0

April 26, 2023

Class 8 Notes

Class 8 Learning Objectives

1. From last class: Interpret coefficient estimates for continuous variables in *simple* logistic regression

2. Construct and fit a multiple logistic regression model

3. Test for significance of individual coefficients or sets of coefficients

4. Estimate the predicted probability of our outcome w/ mult variables

5. Interpret odds ratios for coefficients while adjusting for other variables

controlling for

Class 8 Learning Objectives

1. From last class: Interpret coefficient estimates for continuous variables in *simple* logistic regression

2. Construct and fit a multiple logistic regression model

3. Test for significance of individual coefficients or sets of coefficients

4. Estimate the predicted probability of our outcome

5. Interpret odds ratios for coefficients while adjusting for other variables

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

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

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

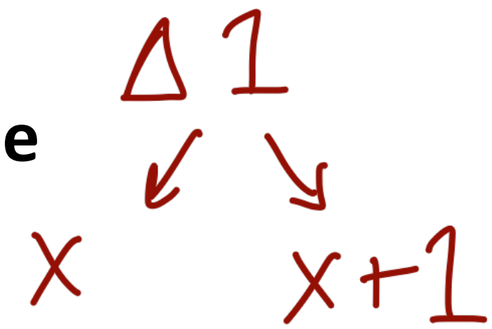
$$F = \beta_0 + \beta_1 X$$

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

systematic component

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

- 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

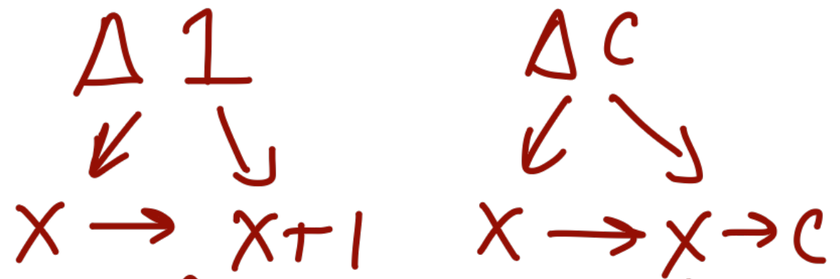
- We should be able to compute and interpret coefficients for a continuous independent covariate x for an arbitrary change of “c” units in x

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:

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

↳ if c=1

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

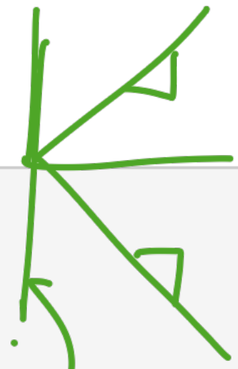
- The c is chosen to be a clinically meaningful unit change in x
- The value of c should be clearly specified in all tables and calculations
  - Because the estimated OR and the corresponding CI depends on the choice of c value

Example: Age and Late Stage Diagnosis (I)

continuous

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

Example: Age and Late Stage Diagnosis (I)



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

flipped version ↳ $P(Y_i = 0 \mid x_i)$

```
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 ***
## ---
```

log(odds ratio) for 1 ~~unit~~ year increase in age.

Example: Age and Late Stage Diagnosis (II)

epiDisplay package

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

Example: Age and Late Stage Diagnosis (II)

$$logit(\pi(x_i)) = logit(p_i) = \ln\left(\frac{p_i}{1-p_i}\right)$$

```
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
```

odds of late stage breast cancer diag is 1.06 times more for an individual 1 year older.

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

$1.06 - 1 = 0.06$ $0.94 - 1 = 6\%$ decrease

- 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
```

77% increase in odds for every 10 yr increase in age.

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

$10 \times se(\hat{\beta}_1)$

```
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
```

For a 10 year increase in age, the odds of late stage BC diagnosis increase by 1.77 times.

Class 8 Learning Objectives

1. From last class: Interpret coefficient estimates for continuous variables in *simple* logistic regression

2. Construct and fit a multiple logistic regression model

3. Test for significance of individual coefficients or sets of coefficients

4. Estimate the predicted probability of our outcome

5. Interpret odds ratios for coefficients while adjusting for other variables

Introduction to Multiple Logistic Regression

- In multiple logistic regression model, we **have > 1 independent variable**
  - Sometimes referred to as the "multivariable regression".
  - The independent variable can be any type:
    - Continuous
    - Categorical (ordinal or nominal)

~~multivariate~~ → multiple outcomes

- We will follow similar procedures as we did for simple logistic regression
  - But we need to **change our interpretation of estimates**

Multiple Logistic Regression Model (I)

- Assume we have a collection of k independent variables, denoted by $\boldsymbol{x}' = (x_1, x_2, \ldots x_k)$

- The conditional probability is $\Pr(\mathrm{Y} = 1|\mathbf{x}) = \pi(\mathbf{x})$ $P(Y_i = 1 \mid x_1, x_2, x_3, \ldots x_k)$

$$\mathrm{logit}\big(\pi(\mathbf{x})\big) = g(\boldsymbol{x}) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k$$

link function same

systematic component changing

Fitting the Multiple Logistic Regression Model (I)

- For a multiple logistic regression model with k independent variables , the **vector of estimates** can be denoted by

$$\boldsymbol{\beta}' = (\beta_0, \beta_1, \ldots, \beta_k)$$

- As with the simple logistic regression, we **use maximum likelihood method** for estimating coefficients

- For a model with k independent variables, there is $k + 1$ coefficients to estimate

(helpful df's)

Breast Cancer Example

- For breast cancer diagnosis example, recall:
  - **Outcome:** early or late stage breast cancer diagnosis
  - **Primary Covariate:** Race/ethnicity
    - While non-Hispanic white individuals are more likely to be diagnosed with breast cancer, non-Hispanic black individuals have the highest death rate
    - Racism affecting health care quality
  - **Additional covariate:** Age
    - Also a risk factor for cancer diagnosis

- We want to fit a multiple logistic regression model with both risk factors included as dependent variables

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

Breast Cancer Example: Model

We can fit the logistic regression model with both race and ethnicity and age:

$$\begin{aligned}&\text{logit}(\pi(x_i))\\&= \beta_0 + \beta_1 I(R\_E_i = H{-}L) + \beta_2 I(R\_E_i = NH\ AIAN)\\&+ \beta_3 I(R\_E_i = NH\ API) + \beta_4 I(R\_E_i = NH\ B) + \beta_5 Age_i\end{aligned}$$

- 6 total coefficients ($\beta_0 - \beta_6$)

Fitting Multiple Logistic Regression Model

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

```
##
## Call:
## glm(formula = Late_stage_diag ~ Race_Ethnicity + Age, family = binomial(),
##     data = bc_diagnosis)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.1996  -0.8609  -0.7100   1.3523   2.4407
##
## Coefficients:
##                                                  Estimate Std. Error z value
## (Intercept)                                     -4.564946   0.205093 -22.258
## Race_EthnicityHispanic-Latino                   -0.015424   0.083653  -0.184
## Race_EthnicityNH American Indian/Alaskan Native -0.085704   0.484110  -0.177
## Race_EthnicityNH Asian/Pacific Islander          0.133965   0.083797   1.599
## Race_EthnicityNH Black                           0.357692   0.071789   4.983
## Age                                              0.057151   0.003209  17.811
##                                                 Pr(>|z|)
## (Intercept)                                      < 2e-16 ***
## Race_EthnicityHispanic-Latino                     0.854
```

Breast Cancer Example: Fitted Model

We can fit the logistic regression model with both race and ethnicity and age:

$$\begin{aligned}&\text{logit}(\pi(x_i))\\&= -4.56 - 0.02I\left(R_{Ei} = H{-}L\right) - 0.09I\left(R_{Ei} = NH\ AIAN\right)\\&+ 0.13I\left(R_{Ei} = NH\ API\right) + 0.36I\left(R_{Ei} = NH\ B\right) + 0.06Age_i\end{aligned}$$

- 6 total coefficients ($\beta_0 - \beta_6$)

Class 8 Learning Objectives

1. From last class: Interpret coefficient estimates for continuous variables in *simple* logistic regression

2. Construct and fit a multiple logistic regression model

3. Test for significance of individual coefficients or sets of coefficients

4. Estimate the predicted probability of our outcome

5. Interpret odds ratios for coefficients while adjusting for other variables

Testing Significance of the Coefficients

- Refer to Class 6 slides 16-38 and Class 7 slides 6-10

- We use the same three tests that we discussed in Simple Logistic Regression to test individual coefficients

1. Wald test
2. ~~Score test~~
3. Likelihood ratio test → can be expanded to sets.

- Textbook and our class focuses on Wald and LRT only

Wald Test

- **Wald test in Multiple Logistic Regression is the exact same thing as in Simple Logistic Regression**

- Assumes test statistic W follows a standard normal distribution under the null hypothesis

$$W = \frac{\hat{\beta}_j}{se(\hat{\beta}_j)} \sim \mathrm{N}(0,1)$$

where $\hat{\beta}_j$ is a MLE and $j = 0, \ldots, k$

- The test is a routine output in R

Wald Test in R

```
wald_test = summary(age_glm)$coefficients
rownames(wald_test) = c("Intercept", "R_E: HL", "R_E: NH AIAN",
                        "R_E: NH API","R_E: NH B", "Age")
round(wald_test, 3)
```

```
##                Estimate Std. Error z value Pr(>|z|)
## Intercept        -4.565      0.205 -22.258    0.000
## R_E: HL          -0.015      0.084  -0.184    0.854
## R_E: NH AIAN     -0.086      0.484  -0.177    0.859
## R_E: NH API       0.134      0.084   1.599    0.110
## R_E: NH B         0.358      0.072   4.983    0.000
## Age               0.057      0.003  17.811    0.000
```

Wald Test Process

1. State hypothesis
   - Are we testing a single coefficient? What is the null?

2. Run the Wald test
   - Look at summary from model

3. Use the values from the test to translate into hypothesis statement
   - What is the coefficient estimate?
   - Either state the test statistic value, p-value, or confidence interval
     - If test statistic: compare to the critical value for distribution at specific α level
     - If p–value: compare to α level
     - If confidence interval: show CI does or does not include value in the null (include your $1 - \alpha\%$ level)

4. Translate your statement with hypothesis into clinically meaningful statement

Likelihood Ratio Test (I)

- Likelihood ratio test answers the question:

  - For a specific covariate, **which model tell us more about the outcome variable**: the **model including** the covariate or the **model omitting** the covariate?

  - For a set of covariates, **which model tell us more about the outcome variable**: the **model including** the set of covariates or the **model omitting** the set of covariates?

Likelihood Ratio Test (II)

- If testing **single variable** and it's **continuous or binary**, still use this hypothesis test:

$$H_0: \beta_j = 0$$
$$H_1: \beta_j \neq 0$$

- If testing **single variable** and it's **categorical with i groups** ($i \geq 2$), use this hypothesis test:

$$H_0: \beta_j = \beta_{j+1} = \cdots = \beta_{j+i-1} = 0$$
H_1: at least one of the above β's is not equal to 0

↳ $\beta_0 \neq 0$ or $\beta_1 \neq 0$ or $\beta_2 \neq 0$ ...

- If testing a **set of variables**, use this hypothesis test:

$$H_0: \beta_1 = \beta_2 = \cdots = \beta_k = 0$$
H_1: at least one of the above β's is not equal to 0

Likelihood Ratio Test (III)

- We compute the value of G, which is the difference of deviances of two models:

set can be of variable

$$\rightarrow \quad G = -2ln\left[\frac{\text{likelihood without set of variables}}{\text{likelihoodwith set of variables}}\right]$$

$$\hookrightarrow \quad X^2 \quad ln(X^2) = 2\,ln(X)$$

- Under the null hypothesis, $G \sim X^2_{df}$
  - where df = #coefficients in model with set of variables − #coefficients in model without set of variables

Likelihood Ratio Test Process

1. State hypothesis
   - Are we testing a single coefficient or a set of variables? What is the null?

2. Calculate the likelihood ratio test statistic (G)
   - Fit the full model (includes the single or set of variables)
   - Fit the reduced model (excludes the single or set of variables)

3. Use the calculated G to translate into hypothesis statement
   - Either state the test statistic value or p-value
     - If test statistic: compare to the critical value for distribution at specific α level
     - If p–value: compare to α level

CI ~~of coefficients~~

$\sim X^2_{df}$

4. Translate your statement with hypothesis into clinically meaningful statement

LRT Examples: Breast Cancer Diagnosis

$$\text{logit}(\pi(x_i)) = \beta_0 + \beta_1 I(R\_E_i = H{-}L) + \beta_2 I(R\_E_i = NH\ AIAN) + \beta_3 I(R\_E_i = NH\ API) + \beta_4 I(R\_E_i = NH\ B) + \beta_5 Age_i$$

- **Example 1:** Single, **continuous variable**: Age
- **Example 2:** Single, **>2 categorical variable**: Race and Ethnicity → Set of Coeff.
- **Example 3:** **Set of variables**: Race and Ethnicity, and Age

LRT Example 1: Breast Cancer Diagnosis (I)

- Single, continuous variable: Age

1. Hypothesis test:
    - H_0: $\beta_5 = 0$
    - H_1: $\beta_5 \neq 0$

LRT Example 1: Breast Cancer Diagnosis (II)

- Single, continuous variable: Age

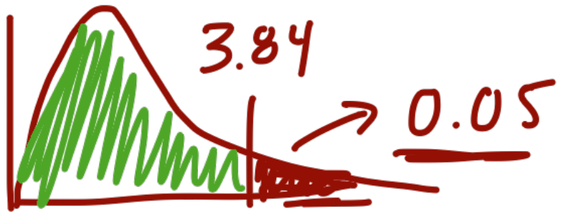
2. Calculate G value

```
multi_bc = glm(Late_stage_diag ~ Race_Ethnicity + Age, data = bc_diagnosis,
               family = binomial())
r_e_bc = glm(Late_stage_diag ~ Race_Ethnicity, data = bc_diagnosis,
             family = binomial())
library(lmtest)
lrtest(multi_bc, r_e_bc)
```

```
## Likelihood ratio test
##
## Model 1: Late_stage_diag ~ Race_Ethnicity + Age
## Model 2: Late_stage_diag ~ Race_Ethnicity
##   #Df  LogLik Df  Chisq Pr(>Chisq)
## 1   6 -5741.8
## 2   5 -5918.1 -1 352.63 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

LRT Example 1: Breast Cancer Diagnosis (III)

- Single, continuous variable: Age



3. Translate to hypothesis

The test statistic is 352.6 under a $\chi^2(df = 1)$ distribution, which is much greater than 3.84 (critical value for 5% significance level). We reject the null hypothesis that the coefficient corresponding to age is 0.

p-value << 0.05

qchisq(p=0.95, df=1)

4. Clinically meaningful statement

The results suggest that the coefficient corresponding to age does not equal to 0. Thus, the model including age helps us fit the breast cancer diagnosis outcome.

LRT Example 2: Breast Cancer Diagnosis (I)

- Single, categorical variable: Race and Ethnicity

1. Hypothesis test
   - H_0: $\beta_1 = \beta_2 = \beta_3 = \beta_4 = 0$
   - H_1: at least one of the above β's is not equal to 0

SLR

old full: $\text{logit}(p) = \beta_0 + \beta_1 RE1 + \beta_2 RE2 + \ldots$

old red: $\text{logit}(p) = \beta_0$

current reduced

MLR

current full: $\text{logit}(p) = \beta_0 + \beta_{1\text{-}4} RE1\text{-}4 + \beta_5 age$

$\text{logit}(p) = \beta_0 + \beta_1 age$

☆ differ from F-test??

LRT Example 2: Breast Cancer Diagnosis (II)

- Single, categorical variable: Race and Ethnicity

2. Calculate G

```
multi_bc = glm(Late_stage_diag ~ Race_Ethnicity + Age, data = bc_diagnosis,
               family = binomial())
age_bc = glm(Late_stage_diag ~ Age, data = bc_diagnosis,
               family = binomial())
library(lmtest)
lrtest(multi_bc, age_bc)
```

```
## Likelihood ratio test
##
## Model 1: Late_stage_diag ~ Race_Ethnicity + Age
## Model 2: Late_stage_diag ~ Age
##   #Df  LogLik Df  Chisq Pr(>Chisq)
## 1   6 -5741.8
## 2   2 -5754.8 -4 26.053  3.087e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

LRT Example 2: Breast Cancer Diagnosis (III)

- Single, categorical variable: Race and Ethnicity

3. Translate to hypothesis

   The test statistic is 26.02 under a $\chi^2(df = 4)$ distribution, which is much greater than 9.49 (critical value for 5% significance level). We reject the null hypothesis that the coefficients corresponding to race and ethnicity are all 0.

4. Clinically meaningful statement

   The results suggest that the model including race and ethnicity is more likely than the model without. Thus, the model including race and ethnicity helps us fit the breast cancer diagnosis outcome.

LRT Example 3: Breast Cancer Diagnosis (I)

- Single, categorical variable: Race and Ethnicity, and age

1. Hypothesis test
   - H_0: $\beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = 0$
   - H_1: at least one of the above β's is not equal to 0

LRT Example 3: Breast Cancer Diagnosis (II)

- Single, categorical variable: Race and Ethnicity

2. Calculate G

```
multi_bc = glm(Late_stage_diag ~ Race_Ethnicity + Age, data = bc_diagnosis,
               family = binomial())
int_bc = glm(Late_stage_diag ~ 1, data = bc_diagnosis,
               family = binomial())
library(lmtest)
lrtest(multi_bc, int_bc)
```

```
## Likelihood ratio test
##
## Model 1: Late_stage_diag ~ Race_Ethnicity + Age
## Model 2: Late_stage_diag ~ 1
##   #Df  LogLik Df  Chisq Pr(>Chisq)
## 1   6 -5741.8
## 2   1 -5930.5 -5 377.32  < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

LRT Example 3: Breast Cancer Diagnosis (III)

- Single, categorical variable: Race and Ethnicity

3. Translate to hypothesis

The test statistic is 377.32 under a $\chi^2(df = 5)$ distribution, which is much greater than 11.7 (critical value for 5% significance level). We reject the null hypothesis that the coefficients corresponding to age and race and ethnicity are all 0.

4. Clinically meaningful statement

The results suggest that the model including age and race and ethnicity is more likely than the model without. Thus, the model including age and race and ethnicity helps us fit the breast cancer diagnosis outcome.

Notes on Likelihood Ratio Test

- Likelihood ratio test is only suitable to test "nested" models
  - "Nested" models means the bigger model (full model) contains all the independent variables of the smaller model (reduced model)
  - We **cannot** compare the following two models using LRT:
    - **Model 1:** $\text{logit}(\pi(x_i)) = \beta_0 + \beta_1 I(R\_E_i = H{-}L) + \beta_2 I(R\_E_i = NH\ AIAN) + \beta_3 I(R\_E_i = NH\ API) + \beta_4 I(R\_E_i = NH\ B)$
    - **Model 2:** $\text{logit}(\pi(x_i)) = \beta_0 + \beta_1 Age_i$

SLR

- If the two models to be compared are not nested, likelihood ratio test should not be used

Class 8 Learning Objectives

1. From last class: Interpret coefficient estimates for continuous variables in *simple* logistic regression

2. Construct and fit a multiple logistic regression model

3. Test for significance of individual coefficients or sets of coefficients

4. Estimate the predicted probability of our outcome

5. Interpret odds ratios for coefficients while adjusting for other variables

Estimated/Predicted Probability for MLR

- Basic idea for predicting/estimating probability stays the same

- Calculations will be slightly different
  - Especially for the confidence interval

- Recall our fitted model for late stage breast cancer diagnosis:

$$\begin{aligned}&\text{logit}(\pi(x_i))\\&= -4.56 - 0.02I\left(R_{E_i} = H{-}L\right) - 0.09I\left(R_{E_i} = NH\ AIAN\right)\\&+ 0.13I\left(R_{E_i} = NH\ API\right) + 0.36I\left(R_{E_i} = NH\ B\right) + 0.06Age_i\end{aligned}$$

Confidence Interval for Predicted Probability (I)

- Consider a multiple logistic regression model with k independent variables.

$$\hat{g}(\boldsymbol{x}) = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \cdots + \hat{\beta}_k x_k$$

- We can express it more concisely using vector notation:

$$\hat{g}(\boldsymbol{x}) = \boldsymbol{x}'\widehat{\boldsymbol{\beta}}$$

where $\widehat{\boldsymbol{\beta}}' = \left(\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_k\right)$, and $\mathbf{x}' = (\mathrm{x}_0, \mathrm{x}_1, \ldots, \mathrm{x}_\mathrm{k})$.

- Notice that $x_0 = 1$ so that $\hat{\beta}_0$ corresponds to the intercept term

Confidence Interval for Predicted Probability (II)

- The estimated variance of the estimator of the logit is:

$$\text{var}[\hat{g}(\boldsymbol{x})] = \sum_{j=0}^{k} x_j^2 \text{var}(\hat{\beta}_j) + \sum_{j=0}^{k} \sum_{l=j+1}^{k} 2x_j x_l \text{cov}(\hat{\beta}_j, \hat{\beta}_l)$$

- Similarly, the equation can be expressed using matrix

$$var[\hat{g}(\boldsymbol{x})] = \mathbf{x}'(\mathbf{X}'\mathbf{V}\mathbf{X})^{-1}\mathbf{x}$$

where $\boldsymbol{x} = \begin{pmatrix} 1 & x_{11} & \cdots & x_{1k} \\ 1 & x_{21} & \cdots & x_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \cdots & x_{nk} \end{pmatrix}$

and $\boldsymbol{v} = \begin{pmatrix} \hat{\pi}_1(1-\hat{\pi}_1) & 0 & \cdots & 0 \\ 0 & \hat{\pi}_2(1-\hat{\pi}_2) & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \hat{\pi}_n(1-\hat{\pi}_n) \end{pmatrix}$

in SLR: $\hat{g}(x) = \beta_0 + \beta_1 x$

$\beta_1 x_1 + \beta_2 x_2 \cdots$

var(X + Y + Z) =

var(X) + var(Y+Z)

+ 2 cov(X, Y+Z)

DO NOT just add SE's from R glm() output

Confidence Interval for Predicted Probability (III)

- 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:

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

Prediction and Confidence Interval in R

get predicted logit

```
newdata = with(bc_diagnosis,
               data.frame(Race_Ethnicity = "NH Asian/Pacific Islander",
                          Age = 40))
pred = predict(multi_bc, 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)
```

if type = "link"

probability

$var(\hat{g}(x))$

```
##      Pred.1        LL.1        UL.1
## 0.10480578 0.08479657 0.12481499
```

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

response $\hat{\pi}(x)$

For NHAPI individual who is 40 yrs old the predicted probability of late stage bc is 0.10.

Class 8 Learning Objectives

1. From last class: Interpret coefficient estimates for continuous variables in *simple* logistic regression

2. Construct and fit a multiple logistic regression model

3. Test for significance of individual coefficients or sets of coefficients

4. Estimate the predicted probability of our outcome

5. Interpret odds ratios for coefficients while adjusting for other variables

Multivariable Logistic Regression Model

- The multivariable model of logistic regression (called multiple logistic regression) is useful in that it *statistically adjusts* the estimated effect of each variable in the model

  - Each estimated coefficient provides an estimate of the log odds **adjusting for all other variables included in the model**

  - The **adjusted odds ratio** can be different from or similar to the unadjusted odds ratio

  - Comparing adjusted vs. unadjusted odds ratios is a useful activity.

Interpretation of Coefficients in MLR

- The interpretation of coefficients in multiple logistic regression is essentially the same as the interpretation of coefficients in simple logistic regression

- For interpretation, we need to
  - point out that these are adjusted estimates
  - provide a list of other variables in the model

Example: Race and Ethnicity and Age model fit

- Let's look at the fitted model:

$$\begin{aligned}&\text{logit}(\pi(x_i))\\&= -4.56 - 0.02I\left(R_{E\,i} = H{-}L\right) - 0.09I\left(R_{E\,i} = NH\ AIAN\right)\\&+ 0.13I\left(R_{E\,i} = NH\ API\right) + 0.36I\left(R_{E\,i} = NH\ B\right) + 0.06Age_i\end{aligned}$$

- Let's take the coefficient for Non-Hispanic Black individuals.
- $\hat{\beta}_4 = 0.36$: Estimated log-OR of late stage breast cancer diagnosis comparing Non-Hispanic Black individuals to Non-Hispanic White individuals is 0.36, **controlling for age**. $\exp(\hat{\beta}_4) = 1.43$
- **Better interpretation:** The estimated odds of late stage breast cancer diagnosis for Non-Hispanic Black individuals is 1.43 (95% CI: (1.24, 1.65))times that of Non-Hispanic White individuals, **controlling for age.**

What does "*controlling for*" mean? (I)

- Consider a multivariable model with two independent variables: one dichotomous (the risk factor) and one continuous (say, age)
  - **Primary interest:** to estimate the effect of the risk factor on the outcome variable
  - But we want to assess whether to adjust for age

- Adjusting for age **may not be necessary** if age distribution is quite similar for the exposed and unexposed groups

- But if the **age distribution does differs for the two groups** (for example, the exposed group are older than the unexposed group), **we need to adjust for age** using multivariable regression modeling

Class 8 Learning Objectives

1. From last class: Interpret coefficient estimates for continuous variables in *simple* logistic regression

2. Construct and fit a multiple logistic regression model

3. Test for significance of individual coefficients or sets of coefficients

4. Estimate the predicted probability of our outcome

5. Interpret odds ratios for coefficients while adjusting for other variables

Wrap-up

- No exit ticket today!! → free attendance!

- **Next class**
  - More about coefficients and interactions next class!
  - More about model building!