This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike License. Your use of this material constitutes acceptance of that license and the conditions of use of materials on this site.
Section D

Incorporating Sampling Variability: Simple Logistic Regression
The method used to estimate the regression coefficients in logistic regression is called the method of maximum likelihood.

Basically, the resulting estimates of the slope and intercept are the values that make the observed data most likely among all choices of values for $\hat{\beta}_o$ and $\hat{\beta}_1$.

This method can be computationally intensive and is of course best handled by the computer.

Along with the estimates of $\hat{\beta}_o$ and $\hat{\beta}_1$ this method yields estimates of the standard error for each: these pieces of information can be used to create confidence intervals and do hypothesis tests.
Random sampling behavior of estimated regression coefficients is normal for “large samples,” and centered at true population value.

As such, we can use same ideas to create 95% CIs and get p-values that we have been using all along.
One reminder: the coefficients from logistic regression are on the \( \ln \text{odds} \) (intercept) and \( \ln \text{odds ratio} \) (slope) scale.

For insight/a reminder into why the sampling distribution of odds and odds ratios is not necessarily normal, but the sampling distribution of the \( \ln \) of such quantities is, check out the last optional section from lecture 6 in SR1.

We will create confidence intervals on the coefficient scale and will need to exponentiate the results to get corresponding CIs on the odds (ratio) scale.

Hypothesis testing and p-value will also be obtained on the coefficient scale.
Example: CHD and Age

- Recall the results from logistic regression of log odds of CHD evidence on age:

\[ \hat{\beta}_1 \]

<table>
<thead>
<tr>
<th>Variable</th>
<th>Estimated Coefficient</th>
<th>Standard Error</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age</td>
<td>0.135</td>
<td>0.036</td>
</tr>
<tr>
<td>Constant</td>
<td>-6.54</td>
<td>1.73</td>
</tr>
</tbody>
</table>

- Recall, \( \hat{\beta}_1 \) is the estimated \( \ln \) odds ratio of CHD evidence for two groups who differ by one year in age: the corresponding odds ratio estimate is \( \hat{O}R = e^{\hat{\beta}_1} = e^{0.135} \approx 1.14 \)
Example: CHD and Age

How to get a 95% CI for $\beta_1$, the population value of the \textit{ln odds} ratio?

Same old approach: $\hat{\beta}_1 \pm 2 \times SE(\hat{\beta}_1)$
- For this example: $0.135 \pm 2 \times 0.036$
  $(0.06, 0.21)$
- Notice, the 95% CI does not include 0, which would indicate no relationship between CHD and age: however this 95% confidence interval is for an \textit{ln odds} ratio

To get the corresponding 95% CI for the odds ratio relating CHD to age, exponentiate the endpoints of the 95% for $\beta_1$ $(e^{0.06}, e^{0.21})$
  $(1.06, 1.23)$
- Notice this 95% CI does not include one, which would be indicative of no CHD/age relationship on the odds ratio scale
Example: CHD and Age

- **p-value for testing:**
  - $H_0: \beta_1 = 0$ \hspace{1cm} $H_0: e^{\beta_1} = 1$ (OR = 1)
  - $H_0: \beta_1 \neq 0$ \hspace{1cm} $H_0: e^{\beta_1} \neq 1$ (OR \neq 1)

- Assume null true and calculate standardized “distance” of $\hat{\beta}_1$ from 0

  $$z = \frac{\hat{\beta}_1 - 0}{SE(\hat{\beta}_1)} = \frac{\hat{\beta}_1}{SE(\hat{\beta}_1)} = \frac{0.135}{0.036} \approx 3.75$$

- **p-value** is the probability of being 3.75 or more standard errors away from 0 on a normal curve: very low in this example, $p < .001$
Example: CHD and Age: Finally Stata!

- Stata can be used to estimate logistic regression models

- For CHD/age example, the data is in Stata in the following format

```
+--------+
<table>
<thead>
<tr>
<th>chd   age</th>
</tr>
</thead>
</table>
1. | 0     20 |
2. | 0     26 |
3. | 0     28 |
4. | 1     29 |
5. | 0     29 |
6. | 0     30 |
7. | 0     30 |
8. | 0     30 |
9. | 0     30 |
10. | 0    32 |
+--------+
```
The “logit” command will give the resulting model on the coefficient scale: syntax

- \textit{logit} \ y \ x_1
- \ y \ must \ take \ on \ values \ of \ 0 \ and \ 1 \ only
The “logit” command will give the resulting model on the coefficient scale: syntax

- \texttt{logit y x1}
- \(y\) must take on values of 0 and 1 only

\[
\text{. logit chd age}
\]

<table>
<thead>
<tr>
<th>Iteration</th>
<th>Log likelihood</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>-39.649049</td>
</tr>
<tr>
<td>1</td>
<td>-29.512908</td>
</tr>
<tr>
<td>2</td>
<td>-28.980636</td>
</tr>
<tr>
<td>3</td>
<td>-28.968083</td>
</tr>
<tr>
<td>4</td>
<td>-28.968073</td>
</tr>
</tbody>
</table>

Logistic regression

\[
\text{Log likelihood} = -28.968073
\]

|             | chd | Coef. | Std. Err. | z     | P>|z| | [95% Conf. Interval] |
|-------------|-----|-------|-----------|-------|------|----------------------|
| age         | .1353143 | .0359736 | 3.76  | 0.000 | .0648072 | .2058213 |
| _cons       | -6.540226  | 1.733787  | -3.77 | 0.000 | -9.938387 | -3.142065 |
Example: CHD and Age: Finally Stata!

- The “logit” command will give the resulting model on the coefficient scale: syntax
  - *logit* $y x_1$
  - ($y$ must take on values of 0 and 1 only)

```
. logit chd age
```

<table>
<thead>
<tr>
<th>Iteration</th>
<th>Log likelihood</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>-39.649049</td>
</tr>
<tr>
<td>1</td>
<td>-29.512908</td>
</tr>
<tr>
<td>2</td>
<td>-28.980636</td>
</tr>
<tr>
<td>3</td>
<td>-28.968083</td>
</tr>
<tr>
<td>4</td>
<td>-28.968073</td>
</tr>
</tbody>
</table>

Logistic regression

<table>
<thead>
<tr>
<th>Number of obs</th>
<th>58</th>
</tr>
</thead>
<tbody>
<tr>
<td>LR chi2(1)</td>
<td>21.36</td>
</tr>
<tr>
<td>Prob &gt; chi2</td>
<td>0.0000</td>
</tr>
<tr>
<td>Pseudo R2</td>
<td>0.2694</td>
</tr>
</tbody>
</table>

Log likelihood = -28.968073

|        | Coef.   | Std. Err. | z       | P>|z|   | [95% Conf. Interval] |
|--------|---------|-----------|---------|-------|----------------------|
| chd    |         |           |         |       |                      |
| age    | .1353143| .0359736  | 3.76    | 0.000 | .0648072 .2058213    |
| _cons  | -6.540226| 1.733787  | -3.77   | 0.000 | -9.938387 -3.142065  |
The “logit” command will give the resulting model on the coefficient scale: syntax

- `logit y x_1`
- `(y must take on values of 0 and 1 only)`

```
. logit chd age
```

Iteration 0:  log likelihood = -39.649049
Iteration 1:  log likelihood = -29.512908
Iteration 2:  log likelihood = -28.980636
Iteration 3:  log likelihood = -28.968083
Iteration 4:  log likelihood = -28.968073

Logistic regression                       Number of obs   =       58
LR chi2(1)       =    21.36
Prob > chi2     =  0.0000
Pseudo R2        = 0.2694

Log likelihood = -28.968073

```
  chd |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
---------+-------------------------------------------------------------
   age |   0.1353143   0.0359736    3.76   0.000     0.0648072    0.2058213
    _cons |  -6.5402261   1.7337873   -3.77   0.000    -9.9383867   -3.142065
---------+-------------------------------------------------------------
```
Example: CHD and Age: Finally Stata!

- The “logistic” command will give the resulting model on the exponentiated scale: syntax
  - `logistic y x1`
  - `(y must take on values of 0 and 1 only)`

  . logistic chd age

  Logistic regression
  Number of obs = 58
  LR chi2(1)    = 21.36
  Prob > chi2   = 0.0000
  Pseudo R2     = 0.2694

  Log likelihood = -28.968073

  +--------------------------------------------------+
  | chd       | Odds Ratio | Std. Err. |    z  |   P>|z| | [95% Conf. Interval] |
  +--------------------------------------------------+
  age | 1.144897   | 0.0411861 | 3.76  | 0.000 | 1.066953   | 1.228534 |
  +--------------------------------------------------+

- Notice, there is no intercept information in this table—why?
- Also, standard error given here not helpful for computing 95% CI for odds ratio
Example: CHD and Age

- How about confidence intervals for the odds ratio when the comparison is on two groups who differ by more than one unit of $x_1$?

- For example, what does the CHD/age results estimate as the odds ratio of CHD evidence for 60 year olds compared to 50 olds? What is a 95% CI for this odds ratio?

- Well, from last section we know the estimated odds ratio is found by taking $e^{10\hat{\beta}_1} = e^{10 \times 1.35} = e^{1.35} \approx 3.9$ (again notice that this is the same as taking $(\hat{OR})^{10} = (1.14)^{10}$ or would be if not for minor differences because of rounding)

- Properties of 95% CI similar on a coefficient scale: a 95% CI for $10\beta_1$:

$$10\hat{\beta}_1 \pm 2 \times SE(10\hat{\beta}_1)$$

$$10\hat{\beta}_1 \pm 2 \times 10 \times SE(\hat{\beta}_1) \rightarrow 10\left(\hat{\beta}_1 \pm 2 \times SE(\hat{\beta}_1)\right)$$
Example: CHD and Age

- So on odds ratio scale, a 95% CI for $e^{10\beta_1}$ will be given by:

$$e^{10(\hat{\beta}_1 \pm 2 \times SE(\hat{\beta}_1))}$$

$$\left( e^{10(\hat{\beta}_1 - 2 \times SE(\hat{\beta}_1))}, e^{10(\hat{\beta}_1 + 2 \times SE(\hat{\beta}_1))} \right)$$

- But notice, the above can be expressed as:

$$\left( \left[ e^{(\hat{\beta}_1 - 2 \times SE(\hat{\beta}_1))} \right]^{10}, e^{\left[ (\hat{\beta}_1 + 2 \times SE(\hat{\beta}_1)) \right]^{10}} \right)$$

- Which is just: $(L^{10}, U^{10})$

- Where L and U are the lower and upper endpoints respectively for the 95% CI for $e^{\beta_1}$, the odds ratio of CHD evidence for a one year difference in age
Example: CHD and Age: Finally Stata!

- Again, from logistic output:

```
. logistic chd age

Logistic regression                          Number of obs   =       50
LR chi2(1)        =     21.36
Prob > chi2       =     0.0000
Pseudo R2         =     0.2694

Log likelihood = -28.968073

+------------------------------------------------------------------+
|     chd | Odds Ratio |     Std. Err. |      z  |     P>|z| |     [95% Conf. Interval] |  
|---------|------------|---------------|---------|--------|------------------------|---
| age     | 1.14497    | 0.0411861     | 3.76    | 0.000  | 1.066953              | 1.220534 |
+------------------------------------------------------------------+
```

- The estimated odds ratio of CHD for 60 year olds compared to 50 years olds (10 year difference in age) is 3.9
- A 95% confidence interval for the true (population) odds ratio is given by $(1.06^{10}, 1.23^{10}) \rightarrow (1.8, 7.9)$
Death in the ICU: Patients with Sepsis

- Recall the results from logistic regression of log odds of death on shock status at the time of ICU admission:

\[ \hat{\beta}_1 \]

<table>
<thead>
<tr>
<th>Variable</th>
<th>Estimated Coefficient</th>
<th>Standard Error</th>
</tr>
</thead>
<tbody>
<tr>
<td>Shock (1 = “yes”)</td>
<td>2.61</td>
<td>0.75</td>
</tr>
<tr>
<td>Constant</td>
<td>-1.77</td>
<td>0.29</td>
</tr>
</tbody>
</table>

- Here \( \hat{OR} = e^{\hat{\beta}_1} = e^{2.61} \approx 13.75 \)
How to get a 95% CI for $\beta_1$, the population value of the ln odds ratio?

Same old approach: $\hat{\beta}_1 \pm 2 \times SE(\hat{\beta}_1)$
- For this example: $2.61 \pm 2 \times 0.75$
  $(1.11, 4.11)$
- Notice, the 95% does not include 0, which would indicate no relationship between death and shock: however this 95% confidence interval is for a ln odds ratio.

To get the corresponding 95% CI for the odds ratio relating death to shock, exponentiate the endpoints of the 95% for $\beta_1$
- $(e^{1.11}, e^{4.11})$
- $(3.0, 61.0)$
- Notice this 95% CI does not include 1, which would be indicative of no CHD/age relationship on the odds ratio scale.
Death in the ICU: Patients with Sepsis

- p-value for testing:
  - $H_0$: $\beta_1 = 0$    $H_0$: $e^{\beta_1} = 1$ (OR = 1)
  - $H_0$: $\beta_1 \neq 0$    $H_0$: $e^{\beta_1} \neq 0$ (OR $\neq 1$)

- Assume null true and calculate standardized “distance” of $\hat{\beta}_1$ from 0
  $$z = \frac{\hat{\beta}_1 - 0}{SE(\beta_1)} = \frac{\hat{\beta}_1}{SE(\beta_1)} = \frac{2.61}{.75} \approx 3.5$$

- p-value is probability of being 3.5 or more standard errors away from 0 on a normal curve: very low in this example, $p < .001$
Results from *logit* command

```
. logit death shock

Iteration 0:  log likelihood = -52.764216
Iteration 1:  log likelihood = -46.750687
Iteration 2:  log likelihood = -46.094137
Iteration 3:  log likelihood = -45.98851
Iteration 4:  log likelihood = -45.988289

Logistic regression                     Number of obs =  106
LR chi2(1) =  13.55
Prob > chi2 =  0.0002
Pseudo R2 =  0.1284

Log likelihood = -45.988289

------------------------------------------------------------------------------
dead  |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
------|------------------|--------|--------|---------|------------------------
      |      shock      |        |        |         |                       |
      |      _cons      |        |        |         |                       |
------------------------------------------------------------------------------
    dead |  2.61496     0.748202  3.49   0.000     1.148511    4.081409
      | -1.767662    2.891776 -0.61   0.540    -2.33444    -1.200884
------------------------------------------------------------------------------
```
Results from *logistic* command

```
. logistic death shock

Logistic regression
Number of obs = 106
LR chi2(1)    = 13.55
Prob > chi2   = 0.0002
Pseudo R2     = 0.1284

Log likelihood = -45.980289

|            | Odds Ratio | Std. Err. |    z  |   P>|z|  |   [95% Conf. Interval] |
|-------------|------------|-----------|-------|-------|------------------------|
| shock       | 13.6667    | 10.22543  | 3.49  | 0.000 | 3.153493               | 59.22885  |
```
One last thing: there is no equivalent to $R^2$ from linear regression in logistic regression; variability and percentage of variability is harder to compute/conceptualize with binary outcomes.

Don’t be fooled by something called Pseudo $R^2$!

```
.logistic death shock

Logistic regression                           Number of obs  =  106
LR chi2(1)         =  13.55
Prob > chi2       =  0.0002
Pseudo R2         =  0.1284

Log likelihood = -45.988289

------------------------------------------------------------------------------
dead   |   Odds Ratio   Std. Err.     z    P>|z|     [95% Conf. Interval]
------------------------------------------------------------------------------
shock  |  13.66667     10.22543    3.49   0.000      3.153493    59.22885
------------------------------------------------------------------------------
```