Lecture 15: Effect modification, and confounding in logistic regression

Sandy Eckel
seckel@jhsph.edu

16 May 2008
Today’s logistic regression topics

- Including categorical predictor
  - create dummy/indicator variables
  - just like for linear regression
- Comparing nested models that differ by two or more variables for logistic regression
  - Chi-square ($X^2$) Test of Deviance
    - i.e., likelihood ratio test
  - analogous to the F-test for nested models in linear regression
- Effect Modification and Confounding
Mean SAT scores were compared for the 50 US states. The goal of the study was to compare overall SAT scores using state-wide predictors such as

- per-pupil expenditures
- average teachers’ salary
Variables

- **Outcome**
  - Total SAT score [sat_low]
    - 1=low, 0=high

- **Primary predictor**
  - Average expenditures per pupil [expen] in thousands
    - Continuous, range: 3.65-9.77, mean: 5.9
    - Doesn’t include 0: center at $5,000 per pupil

- **Secondary predictor**
  - Mean teacher salary in thousands, in quartiles
    - salary1 – lowest quartile
    - salary2 – 2\textsuperscript{nd} quartile
    - salary3 – 3\textsuperscript{rd} quartile
    - salary4 – highest quartile
  - four dummy variables for four categories; **must exclude one category** to create a reference group
Analysis Plan

- Assess primary relationship (parent model)
- Add secondary predictor in separate model (extended model)
  - Determine if secondary predictor is statistically significant
  - How? Use the Chi-square test of deviance
Models and Results
(note that only exponentiated slopes are shown)

\[
\log \left( \frac{p}{1-p} \right) = \beta_0 + \beta_1 (\text{Expenditure} - 5)
\]

**Model 1 (Parent): Only primary predictor**

| sat_low | Odds Ratio | Std. Err. | z    | P>|z| | [95% Conf. Interval] |
|---------|------------|-----------|------|------|---------------------|
| expenc  | 2.484706   | .8246782  | 2.74 | 0.006 | 1.296462 4.76201    |

\[
\log \left( \frac{p}{1-p} \right) = \beta_0 + \beta_1 (\text{Expenditure} - 5) + \beta_2 (\text{Salary} = 2) + \beta_3 (\text{Salary} = 3) + \beta_4 (\text{Salary} = 4)
\]

**Model 2 (Extended): Primary Predictor and Secondary Predictor**

| sat_low | Odds Ratio | Std. Err. | z    | P>|z| | [95% Conf. Interval] |
|---------|------------|-----------|------|------|---------------------|
| expenc  | 1.796861   | .7982988  | 1.32 | 0.187 | .7522251 4.292213   |
| salary2 | 2.783137   | 2.815949  | 1.01 | 0.312 | .3830872 20.21955   |
| salary3 | 2.923654   | 3.2716    | 0.96 | 0.338 | .326154 26.20773    |
| salary4 | 4.362678   | 6.147015  | 1.05 | 0.296 | .2756828 69.03933   |

The $X^2$ Test of Deviance

- We want to compare the parent model to an extended model, which differs by the three dummy variables for the four salary quartiles.
- The $X^2$ test of deviance compares nested logistic regression models
  - We use it for nested models that differ by two or more variables because the Wald test cannot be used in that situation.
Performing the Chi-square test of deviance for nested logistic regression

1. Get the **log likelihood** (LL) from both models
   - Parent model: \( LL = -28.94 \)
   - Extended model: \( LL = -28.25 \)

2. Find the **deviance** for both models
   - Deviance = \(-2(\text{log likelihood})\)
     - Parent model: \( \text{Deviance} = -2(-28.94) = 57.88 \)
     - Extended model: \( \text{Deviance} = -2(-28.25) = 56.50 \)
   - Deviance is analogous to residual sums of squares (RSS) in linear regression; it measures the ‘deviation’ still available in the model
     - A **saturated** model is one in which every \( Y \) is perfectly predicted
Performing the Chi-square test of deviance for nested logistic regression, cont...

3. Find the **change in deviance** between the nested models
   
   \[ \text{deviance}_{\text{parent}} - \text{deviance}_{\text{extended}} = 57.88 - 56.50 = 1.38 = \text{Test Statistic (X}^2\text{)} \]

4. **Evaluate** the change in deviance
   - The change in deviance is an observed Chi-square statistic
   - \( df = \# \) of variables added
   - \( H_0: \) all new \( \beta \)'s are 0 in the population
     i.e., \( H_0: \) the parent model is better
The Chi-square test of deviance for our nested logistic regression example

- **H₀**: After adjusting for per-pupil expenditures, all the slopes on salary indicators are 0 \((β_2 = β_3 = β_4 = 0)\)

- \(X^2_{obs} = 1.38\)
  - df = 3

- With 3 df and \(α=0.05\), \(X^2_{cr}\) is 7.81
  - \(X^2_{obs} < X^2_{cr}\)
  - Fail to reject \(H₀\)

- **Conclude**: After adjusting for per-pupil expenditure, teachers’ salary is not a statistically significant predictor of low SAT scores
Notes about Chi-square deviance test

- The deviance test gives us a framework in which to add several predictors to a model simultaneously.
- Can only handle nested models.
- Analogous to F-test for linear regression.
- Also known as "likelihood ratio test"
1. Fit parent model
   
   ```r
   fit.parent <- glm(y~x1, family=binomial())
   ```

2. Fit the extended model (parent model is nested within the extended model)
   
   ```r
   fit.extended <- glm(y~x1+x2+x3, family=binomial())
   ```

3. Perform the Chi-square deviance test
   
   ```r
   anova(fit.parent, fit.extended, test="Chi")
   ```

Example output:

```
Analysis of Deviance Table

Model 1: y ~ x1
Model 2: y ~ x1 + x2 + x3

  Resid. Df Resid. Dev Df Deviance  P(>|Chi|)
 1     48     64.250                      2     46     48.821  2   15.429 0.0004464
```

Chi-square Test Statistic

- P-value
- Degrees of freedom
Effect Modification and Confounding in Logistic Regression

Heart Disease
Smoking and Coffee
Example
Effect modification in logistic regression

- Just like with linear regression, we may want to allow different relationships between the primary predictor and outcome across levels of another covariate
- We can model such relationships by fitting interaction terms in logistic regressions
- Modelling effect modification will require dealing with two or more covariates
Logistic models with two covariates

- \( \text{logit}(p) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 \)

Then:

\[ \text{logit}(p \mid X_1 = X_1 + 1, X_2 = X_2) = \beta_0 + \beta_1 (X_1 + 1) + \beta_2 X_2 \]

\[ \text{logit}(p \mid X_1 = X_1, X_2 = X_2) = \beta_0 + \beta_1 (X_1) + \beta_2 X_2 \]

\[ \Delta \text{ in log-odds} = \beta_1 \]

- \( \beta_1 \) is the **change in log-odds** for a 1 unit change in \( X_1 \) *provided \( X_2 \) is held constant.*
Interpretation in General

- Also: \[ \log \left( \frac{\text{odds}(Y = 1| X_1 + 1, X_2)}{\text{odds}(Y = 1| X_1, X_2)} \right) = \beta_1 \]

- And: \[ \text{OR} = \exp(\beta_1) \]

- \( \exp(\beta_1) \) is the multiplicative change in odds for a 1 unit increase in \( X_1 \) provided \( X_2 \) is held constant.

- The result is similar for \( X_2 \)

- What if the effects of each of \( X_1 \) and \( X_2 \) depend on the presence of the other?
  - Effect modification!
Data: Coronary Heart Disease (CHD), Smoking and Coffee

<table>
<thead>
<tr>
<th>CHD</th>
<th>Smoke</th>
<th>Coffee</th>
<th>n</th>
</tr>
</thead>
<tbody>
<tr>
<td>yes</td>
<td>no</td>
<td>no</td>
<td>15</td>
</tr>
<tr>
<td>no</td>
<td>no</td>
<td>no</td>
<td>42</td>
</tr>
<tr>
<td>yes</td>
<td>yes</td>
<td>no</td>
<td>11</td>
</tr>
<tr>
<td>no</td>
<td>yes</td>
<td>no</td>
<td>8</td>
</tr>
<tr>
<td>yes</td>
<td>no</td>
<td>yes</td>
<td>15</td>
</tr>
<tr>
<td>no</td>
<td>no</td>
<td>yes</td>
<td>21</td>
</tr>
<tr>
<td>yes</td>
<td>yes</td>
<td>yes</td>
<td>25</td>
</tr>
<tr>
<td>no</td>
<td>yes</td>
<td>yes</td>
<td>14</td>
</tr>
</tbody>
</table>

n = 151
Study Information

- **Study Facts:**
  - Case-Control study (disease = CHD)
  - 40-50 year-old males previously in good health

- **Study questions:**
  - Is smoking and/or coffee related to an increased odds of CHD?
  - Is the association of coffee with CHD higher among smokers? That is, is smoking an **effect modifier** of the coffee-CHD associations?
Fraction with CHD by smoking and coffee

Number in each cell is the proportion of the total number of individuals with that smoking/coffee combination that have CHD

<table>
<thead>
<tr>
<th>Coffee</th>
<th>No</th>
<th>Yes</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>No</td>
<td>15/57= .26</td>
<td>11/19= .58</td>
<td>26/76= .34</td>
</tr>
<tr>
<td>Yes</td>
<td>15/36= .42</td>
<td>25/39= .64</td>
<td>40/75= .53</td>
</tr>
<tr>
<td>Total</td>
<td>30/93= .32</td>
<td>36/58= .62</td>
<td>66/151= .44</td>
</tr>
</tbody>
</table>
Pooled data (ignoring smoking)

<table>
<thead>
<tr>
<th>Coffee</th>
<th>CHD Case</th>
<th>Cntl</th>
<th>% CHD Case</th>
</tr>
</thead>
<tbody>
<tr>
<td>No</td>
<td>26</td>
<td>50</td>
<td>.34</td>
</tr>
<tr>
<td>Yes</td>
<td>40</td>
<td>35</td>
<td>.53</td>
</tr>
</tbody>
</table>

Odds ratio of CHD comparing coffee to non-coffee drinkers

\[
\frac{.53/(1-.53)}{.34/(1-.34)} = 2.2
\]

95% CI = (1.14, 4.24)
Among **Non-Smokers**

<table>
<thead>
<tr>
<th></th>
<th>Coffee</th>
<th>No coffee</th>
</tr>
</thead>
<tbody>
<tr>
<td>CHD</td>
<td>15</td>
<td>15</td>
</tr>
<tr>
<td>No CHD</td>
<td>21</td>
<td>42</td>
</tr>
</tbody>
</table>

\[
P(\text{CHD} | \text{Coffee drinker}) = \frac{15}{(15+21)} = 0.42
\]

\[
P(\text{CHD} | \text{Not Coffee drinker}) = \frac{15}{(15+42)} = 0.26
\]

**Odds ratio of CHD comparing coffee to non-coffee drinkers**

\[
\frac{0.42/(1-0.42)}{0.26/(1-0.26)} = 2.06
\]

**95% CI = (0.82, 4.9)**
Among Smokers

<table>
<thead>
<tr>
<th></th>
<th>Coffee</th>
<th>No coffee</th>
</tr>
</thead>
<tbody>
<tr>
<td>CHD</td>
<td>25</td>
<td>11</td>
</tr>
<tr>
<td>No CHD</td>
<td>14</td>
<td>8</td>
</tr>
</tbody>
</table>

P(CHD| Coffee drinker) = 25/(25+14) = 0.64
P(CHD| Not Coffee drinker) = 11/(11+8) = 0.58

Odds ratio of CHD comparing coffee to non-coffee drinkers

\[
\frac{.64/(1-.64)}{.58/(1-.58)} = 1.29
\]

95% CI = (0.42, 4.0)
Plot Odds Ratios and 95% CIs
Define Variables

- \( Y_i = 1 \) if CHD case, 0 if control
- \( \text{coffee}_i = 1 \) if Coffee Drinker, 0 if not
- \( \text{smoke}_i = 1 \) if Smoker, 0 if not
- \( p_i = \text{Pr} (Y_i = 1) \)
- \( n_i = \text{Number observed at pattern}_i \) of Xs
Logistic Regression Model

- $Y_i$ are independent
- **Random part**
  $Y_i$ are from a Binomial $(n_i, p_i)$ distribution
- **Systematic part**
  log odds ($Y_i=1$) (or logit($Y_i=1$)) is a function of
  - Coffee
  - Smoking
  - and coffee-smoking interaction
  \[
  \log\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 \text{coffee}_i + \beta_2 \text{smoke}_i + \beta_3 \text{coffee} \times_i \text{smoke}_i
  \]
Interpretations – stratify by smoking status

\[
\log\left( \frac{p_i}{1-p_i} \right) = \beta_0 + \beta_1 \text{coffee}_i + \beta_2 \text{smoke}_i + \beta_3 \text{coffee}_i \times \text{smoke}_i
\]

If smoke = 0

\[
\log\left( \frac{p_i}{1-p_i} \right) = \beta_0 + \beta_1 \text{coffee}_i
\]

If smoke = 1

\[
\log\left( \frac{p_i}{1-p_i} \right) = \beta_0 + \beta_1 \text{coffee}_i + \beta_2 \times 1 + \beta_3 \text{coffee}_i \times 1 = (\beta_0 + \beta_2) + (\beta_1 + \beta_3) \text{coffee}_i
\]

- \( \exp(\beta_1) \): odds ratio of being a CHD case for coffee drinkers -vs- non-drinkers among non-smokers
- \( \exp(\beta_1 + \beta_3) \): odds ratio of being a CHD case for coffee drinkers -vs- non-drinkers among smokers
Interpretations – stratify by coffee drinking

\[
\log\left( \frac{p_i}{1 - p_i} \right) = \beta_0 + \beta_1 \text{coffee}_i + \beta_2 \text{smoke}_i + \beta_3 \text{coffee}_i \times \text{smoke}_i
\]

If coffee = 0
\[
\log\left( \frac{p_i}{1 - p_i} \right) = \beta_0 + \beta_2 \text{smoke}_i
\]

If coffee = 1
\[
\log\left( \frac{p_i}{1 - p_i} \right) = \beta_0 + \beta_1 \times 1 + \beta_2 \text{smoke}_i + \beta_3 1 \times \text{smoke}_i = (\beta_0 + \beta_1) + (\beta_2 + \beta_3) \text{smoke}_i
\]

- \(\exp(\beta_2)\): odds ratio of being a CHD case for smokers -vs- non-smokers among non-coffee drinkers
- \(\exp(\beta_2 + \beta_3)\): odds ratio of being a CHD case for smokers -vs- non-smokers among coffee drinkers
Interpretations

$\log \left( \frac{p_i}{1 - p_i} \right) = \beta_0 + \beta_1\text{coffee}_i + \beta_2\text{smoke}_i + \beta_3\text{coffee}_i \times \text{smoke}_i$

- $\frac{e^{\beta_0}}{1 + e^{\beta_0}}$ Probability of CHD if all X’s are zero
  - i.e., fraction of cases among non-smoking non-coffee drinking individuals in the sample (determined by sampling plan)

- $\exp(\beta_3)$: ratio of odds ratios

What do we mean by this?
exp(\(\beta_3\)) Interpretations

\[
\log \left( \frac{p_i}{1 - p_i} \right) = \beta_0 + \beta_1 \text{coffee}_i + \beta_2 \text{smoke}_i + \beta_3 \text{coffee}_i \times \text{smoke}_i
\]

- exp(\(\beta_3\)): factor by which odds ratio of being a CHD case for coffee drinkers -vs- nondrinkers is multiplied for smokers as compared to non-smokers

or

- exp(\(\beta_3\)): factor by which odds ratio of being a CHD case for smokers -vs- non-smokers is multiplied for coffee drinkers as compared to non-coffee drinkers

COMMON IDEA: Additional multiplicative change in the odds ratio beyond the smoking or coffee drinking effect alone when you have both of these risk factors present
Some Special Cases: No smoking or coffee drinking effects

- Given

\[
\log \left( \frac{p_i}{1 - p_i} \right) = \beta_0 + \beta_1 \text{coffee}_i + \beta_2 \text{smoke}_i + \beta_3 \text{coffee}_i \times \text{smoke}_i
\]

- If \( \beta_1 = \beta_2 = \beta_3 = 0 \)

- Neither smoking nor coffee drinking is associated with increased risk of CHD
Some Special Cases: Only one effect

- Given

\[
\log\left( \frac{p_i}{1 - p_i} \right) = \beta_0 + \beta_1 coffee_i + \beta_2 smoke_i + \beta_3 coffee_i \times smoke_i
\]

- If \( \beta_2 = \beta_3 = 0 \)
  - Coffee drinking, but not smoking, is associated with increased risk of CHD

- If \( \beta_1 = \beta_3 = 0 \)
  - Smoking, but not coffee drinking, is associated with increased risk of CHD
Some Special Cases

\[ \log\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_{1\text{coffee}_i} + \beta_{2\text{smoke}_i} + \beta_{3\text{coffee}_i \times \text{smoke}_i} \]

- If \( \beta_3 = 0 \)
  - Smoking and coffee drinking are both associated with risk of CHD but the odds ratio of CHD-smoking is the same at both levels of coffee
  - Smoking and coffee drinking are both associated with risk of CHD but the odds ratio of CHD-coffee is the same at both levels of smoking
- Common idea: the effects of each of these risk factors is purely additive (on the log-odds scale), there is no interaction
Model 1: main effect of coffee

\[
\log\left(\frac{p_i}{1 - p_i}\right) = \beta_0 + \beta_1 \text{coffee}_i
\]

Logit estimates

Number of obs = 151
LR chi2(1) = 5.65
Prob > chi2 = 0.0175
Pseudo R2 = 0.0273

Log likelihood = -100.64332

| chd   | Coef.      | Std. Err. | z     | P>|z|   | [95% Conf. Interval] |
|-------|------------|-----------|-------|-------|----------------------|
| coffee | 0.7874579  | 0.3347123 | 2.35  | 0.019 | 0.1314338 - 1.443482 |
| (Intercept) | -0.6539265 | 0.2417869 | -2.70 | 0.007 | -1.12782 - 0.180329 |
Model 2: main effects of coffee and smoke

\[
\log\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 coffee_i + \beta_2 smoke_i
\]

Logit estimates

|       | Coef. | Std. Err. | z    | P>|z| | [95% Conf. Interval] |
|-------|-------|-----------|------|------|---------------------|
| coffee | 0.5269764 | 0.3541932 | 1.49 | 0.137 | -0.1672295 - 1.221182 |
| smoke  | 1.101978  | 0.3609954 | 3.05 | 0.002 | 0.3944404 - 1.809516 |
| (Intercept) | -0.9572328 | 0.2703086 | -3.54 | 0.000 | -1.487028 - 0.4274377 |

Number of obs = 151
LR chi2(2) = 15.19
Prob > chi2 = 0.0005
Pseudo R2 = 0.0734
Log likelihood = -95.869718
Model 3: main effects of coffee and smoke AND their interaction

\[
\log\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_{1\text{coffee}_i} + \beta_{2\text{smoke}_i} + \beta_{3\text{coffee}_i \times \text{smoke}_i}
\]

Logit estimates

|             | Coef.    | Std. Err. | z     | P>|z|   | [95% Conf. Interval] |
|-------------|----------|-----------|-------|-------|---------------------|
| coffee      | .6931472 | .4525062  | 1.53  | 0.126 | -.1937487 to 1.580043 |
| smoke       | 1.348073 | .5535208  | 2.44  | 0.015 | .2631923 to 2.432954 |
| coffee_smoke| -.4317824| .7294515  | -.59  | 0.554 | -.1861481 to .9979163 |
| (Intercept) | -1.029619| .3007926  | -3.42 | 0.001 | -1.619162 to -.4400768 |

Number of obs = 151
LR chi2(3) = 15.55
Prob > chi2 = 0.0014
Log likelihood = -95.694169

Pseudo R2 = 0.0751
Comparing Models 1 & 2

Question: Is smoking a confounder?

<table>
<thead>
<tr>
<th>Variable</th>
<th>Model1</th>
<th></th>
<th>Model2</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>se</td>
<td></td>
<td>se</td>
</tr>
<tr>
<td>Intercept</td>
<td>-.65</td>
<td>.24</td>
<td>-.96</td>
<td>.27</td>
</tr>
<tr>
<td>Coffee</td>
<td>.79</td>
<td>.33</td>
<td>.53</td>
<td>.35</td>
</tr>
<tr>
<td>Smoking</td>
<td>1.10</td>
<td>.36</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Look at Confidence Intervals

- **Without Smoking**
  \[ OR = e^{0.79} = 2.2 \]
  - 95% CI for log(OR): \( 0.79 \pm 1.96(0.33) \)
    \[ = (0.13, 1.44) \]
  - 95% CI for OR: \( (e^{0.13}, e^{1.44}) \)
    \[ = (1.14, 4.24) \]

- **With Smoking** (adjusting for smoking)
  \[ OR = e^{0.53} = 1.7 \]

**Smoking does not confound** the relationship between coffee drinking and CHD
- since 1.7 is in the 95% CI from the model without smoking
Conclusion regarding confounding

- So, ignoring smoking, the CHD and coffee OR is 2.2 (95% CI: 1.14 - 4.26)
- Adjusting for smoking, gives more modest evidence for a coffee effect
- However, smoking does not appear to be an important confounder
Interaction Model

Question: Is smoking an effect modifier of CHD-coffee association?

<table>
<thead>
<tr>
<th>Variable</th>
<th>Est</th>
<th>se</th>
<th>z</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Model 3</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Intercept</td>
<td>-1.0</td>
<td>.30</td>
<td>-3.4</td>
</tr>
<tr>
<td>Coffee</td>
<td>.69</td>
<td>.45</td>
<td>1.5</td>
</tr>
<tr>
<td>Smoking</td>
<td>1.3</td>
<td>.55</td>
<td>2.4</td>
</tr>
<tr>
<td>Coffee*Smoking</td>
<td>-.43</td>
<td>.73</td>
<td>-.59</td>
</tr>
</tbody>
</table>
Testing Interaction Term

- $Z = -0.59$, p-value = 0.554

- We fail to reject $H_0$: interaction slope = 0

- And we conclude there is little evidence that smoking is an effect modifier!
Question: Model selection

What model should we choose to describe the relationship of coffee and smoking with CHD?
Fitted Values

- We can use transform to get fitted probabilities and compare with observed proportions using each of the three models.

- **Model 1:**
  \[
  \hat{p} = \frac{e^{-.65+.79\text{Coffee}}}{1 + e^{-.65+.79\text{Coffee}}}
  \]

- **Model 2:**
  \[
  \hat{p} = \frac{e^{-.96+.53\text{Coffee}+1.1\text{Smoking}}}{1 + e^{-.96+.53\text{Coffee}+1.1\text{Smoking}}}
  \]

- **Model 3:**
  \[
  \hat{p} = \frac{e^{-.103+.69\text{Coffee}+1.3\text{Smoking}-.43(\text{Coffee}^*\text{Smoking})}}{1 + e^{-.103+.69\text{Coffee}+1.3\text{Smoking}-.43(\text{Coffee}^*\text{Smoking})}}
  \]
# Observed vs Fitted Values

<table>
<thead>
<tr>
<th>Coffee</th>
<th>Model</th>
<th>Smoking</th>
<th>Pr (case)</th>
<th>Pr (case)</th>
</tr>
</thead>
<tbody>
<tr>
<td>No</td>
<td>observed</td>
<td>No</td>
<td>.26</td>
<td>.58</td>
</tr>
<tr>
<td></td>
<td>Model 1</td>
<td>Yes</td>
<td>.34</td>
<td>.34</td>
</tr>
<tr>
<td></td>
<td>Model 2</td>
<td>No</td>
<td>.28</td>
<td>.54</td>
</tr>
<tr>
<td></td>
<td>Model 3</td>
<td>Yes</td>
<td>.26</td>
<td>.58</td>
</tr>
<tr>
<td>Yes</td>
<td>observed</td>
<td>No</td>
<td>.42</td>
<td>.64</td>
</tr>
<tr>
<td></td>
<td>model 1</td>
<td>Yes</td>
<td>.53</td>
<td>.53</td>
</tr>
<tr>
<td></td>
<td>model 2</td>
<td>No</td>
<td>.39</td>
<td>.66</td>
</tr>
<tr>
<td></td>
<td>model 3</td>
<td>Yes</td>
<td>.42</td>
<td>.64</td>
</tr>
</tbody>
</table>
Saturated Model

- Note that fitted values from Model 3 exactly match the observed values indicating a "saturated" model that gives perfect predictions.

- Although the saturated model will always result in a perfect fit, it is usually not the best model (e.g., when there are continuous covariates or many covariates).
The Likelihood Ratio Test will help decide whether or not additional term(s) “significantly” improve the model fit.

Likelihood Ratio Test (LRT) statistic for comparing nested models is:
- -2 times the difference between the log likelihoods (LLs) for the Null -vs- Extended models.

We’ve already done this earlier in today’s lecture!!
- Chi-square ($X^2$) Test of Deviance is the same thing as the Likelihood Ratio Test.
- Used to compare any pair of nested logistic regression models and get a p-value associated with the $H_0$: the ‘new’ $\beta$’s all=0.
A case-control study was conducted with 151 subjects, 66 (44%) of whom had CHD, to assess the relative importance of smoking and coffee drinking as risk factors. The observed fractions of CHD cases by smoking, coffee strata are

<table>
<thead>
<tr>
<th>Coffee</th>
<th>No</th>
<th>Yes</th>
</tr>
</thead>
<tbody>
<tr>
<td>No</td>
<td>.26</td>
<td>.58</td>
</tr>
<tr>
<td>Yes</td>
<td>.42</td>
<td>.64</td>
</tr>
</tbody>
</table>
Example Summary: Unadjusted ORs

- The odds of CHD was estimated to be 3.4 times higher among smokers compared to non-smokers
  - 95% CI: (1.7, 7.9)

- The odds of CHD was estimated to be 2.2 times higher among coffee drinkers compared to non-coffee drinkers
  - 95% CI: (1.1, 4.3)
Controlling for the potential confounding of smoking, the coffee odds ratio was estimated to be 1.7 with 95% CI: (.85, 3.4).

Hence, the evidence in these data are insufficient to conclude coffee has an independent effect on CHD beyond that of smoking.
Finally, we estimated the coffee odds ratio separately for smokers and non-smokers to assess whether smoking is an effect modifier of the coffee-CHD relationship. For the smokers and non-smokers, the coffee odds ratio was estimated to be 1.3 (95% CI: .42, 4.0) and 2.0 (95% CI: .82, 4.9) respectively. There is little evidence of effect modification in these data.
Including categorical predictors in logistic regression
- create dummy/indicator variables
- just like for linear regression

Comparing nested models that differ by two or more variables for logistic regression
- Chi-square (X^2) Test of Deviance
  - i.e., likelihood ratio test
  - analogous to the F-test for nested models in linear regression

Effect Modification and Confounding in logistic regression