Stats for Clinical Trials, Math 150
Jo Hardin, Spring 2023
Simpson's Paradox

Simpson's Paradox

Simpson's paradox is when the association between two variables is opposite the partial association between the same two variables after controlling for one or more other variables.

Consider the example on smoking and 20-year mortality (case) from section 3.4 of *Regression Methods in Biostatistics*, pg 52-53.

age		smoker	nonsmoker	$\widehat{\text{prob smoke}}$	$\widehat{\text{odds smoke}}$	$\widehat{OR}$
all	case	139	230	0.377	0.604	0.685
	control	443	502	0.469	0.882	
18-44	case	61	32	0.656	1.906	1.627
	control	375	320	0.540	1.172	
45-64	case	34	66	0.340	0.515	1.308
	control	50	127	0.282	0.394	
65+	case	44	132	0.250	0.333	1.019
	control	18	55	0.247	0.327	

What we see is that the vast majority of the controls were young, and they had a high rate of smoking. A good chunk of the cases were older, and the rate of smoking was substantially lower in the oldest group. However, within each group, the cases were more likely to smoke than the controls.

R code / logistic regression on Simpson's Paradox smoking data

```
> glm( death ~ smoke, family="binomial") %>% tidy()
#> # A tibble: 2 5
#>   term        estimate std.error statistic  p.value
#>   <chr>          <dbl>     <dbl>     <dbl>    <dbl>
#> 1 (Intercept)   -0.781    0.0796     -9.80 1.10e-22
#> 2 smoke         -0.379    0.126      -3.01 2.59e- 3

> glm( death ~ smoke + age, family="binomial") %>% tidy()
#> # A tibble: 4 5
#>   term        estimate std.error statistic  p.value
#>   <chr>          <dbl>     <dbl>     <dbl>    <dbl>
#> 1 (Intercept)   -0.668     0.135     -4.96 7.03e- 7
#> 2 smoke          0.312     0.154      2.03 4.25e- 2
#> 3 ageold         1.47      0.188      7.84 4.59e-15
#> 4 ageyoung      -1.52      0.173     -8.81 1.26e-18

> glm( death ~ smoke * age, family="binomial") %>% tidy()
#> # A tibble: 6 5
#>   term           estimate std.error statistic  p.value
#>   <chr>             <dbl>     <dbl>     <dbl>    <dbl>
#> 1 (Intercept)      -0.655     0.152    -4.31  1.61e- 5
#> 2 smoke             0.269     0.269     0.999 3.18e- 1
#> 3 ageold            1.53      0.221     6.93  4.29e-12
#> 4 ageyoung         -1.65      0.240    -6.88  6.00e-12
#> 5 smoke:ageold     -0.251     0.420    -0.596 5.51e- 1
#> 6 smoke:ageyoung    0.218     0.355     0.614 5.40e- 1
```

What we see is that the vast majority of the controls were young, and they had a high rate of smoking. A good chunk of the cases were older, and the rate of smoking was substantially lower in the oldest group. However, within each group, the cases were more likely to smoke than the controls.

After *adjusting* for age, smoking is no longer significant. But more importantly, age is a variable that reverses the effect of smoking on cancer - Simpson's Paradox. Note that the effect is not due to the observational nature of the study, and so it is important to adjust for possible influential variables regardless of the study at hand.

- What does it mean to *adjust* for age in this context? It means that we have to include it in the model.

- What does it mean that the interaction terms are not significant in the last model? It means that the value of the interaction coefficients (above, b_4 and b_5) are within the range of values we would have gotten just by chance (if $\beta_4 = 0$ and $\beta_5 = 0$).
- We can estimate any of the OR (of dying for smoke vs not smoke) from the given coefficients:

$$\begin{aligned}
&\textsc{simple model}\\
\text{overall OR} &= e^{-0.37858} = 0.685\\
&\textsc{additive model}\\
\text{young, middle, old OR} &= e^{0.3122} = 1.366\\
&\textsc{interaction model}\\
\text{old OR} &= e^{0.2689+0.2177} = 1.627\\
\text{middle OR} &= e^{0.2689} = 1.308\\
\text{old OR} &= e^{0.2689+-0.2505} = 1.019
\end{aligned}$$