Cohort Data

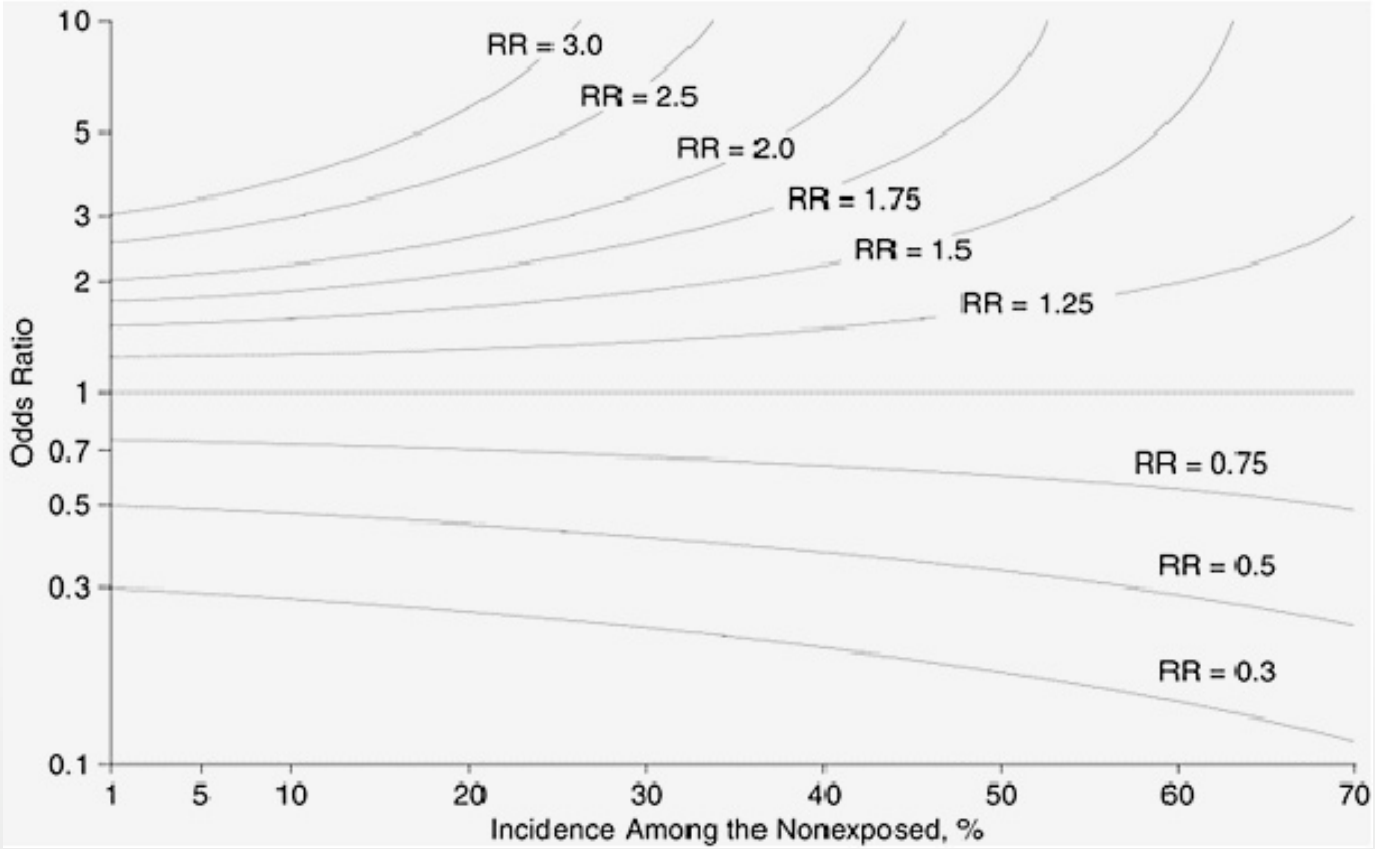
Questions on the Homework/Last Class?

This Week's Focus

- Cohort data
- Single outcome measurement at the end of the study
- Also applicable to other studies that produce categorical data (i.e. studies of prevalence)

The Rare Disease Assumption

- Logistic regression isn't the ideal – it's a compromise
  - But it's a very *good* compromise
- But it's utility (and the utility of the case-control study) is build on the assumption that the outcome is rare
- In this case, two things are true:
  - The OR approximates an RR
  - A case-control study is more efficient


RR = 3.0
RR = 2.5
RR = 2.0
RR = 1.75
RR = 1.5
RR = 1.25
RR = 0.75
RR = 0.5
RR = 0.3
Odds Ratio
10
5
3
2
1
0.7
0.5
0.3
0.1
1
5
10
20
30
40
50
60
70
Incidence Among the Nonexposed, %

For the *Extremely* Common Outcome

- Reverse the coding of your outcome
- Model the probability of *not* having the outcome

For Prevalence > 10% and < 90%

- OR isn't a good approximation of what we actually want to estimate
- Estimate the RR directly
- We can do this with Binomial regression

Binomial Regression

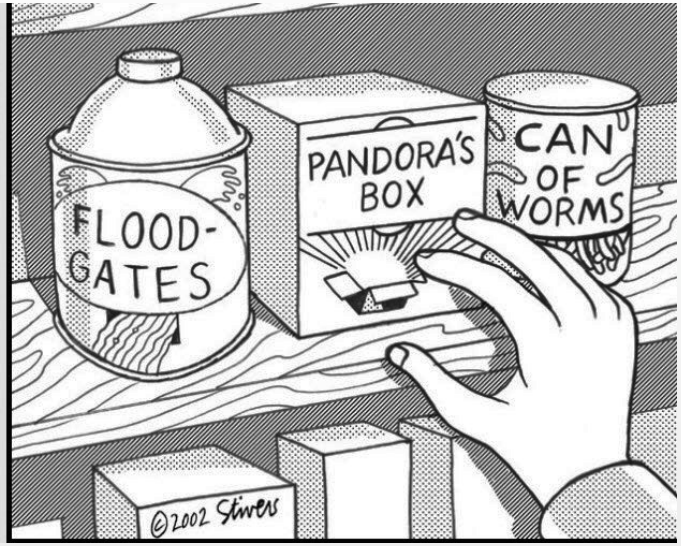
$$\log(p(Y = 1)) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 \ldots$$

In R

```
model <- glm(Y ~ X1 + X2,family=binomial(link='log'), data=data)
```

What Does This Do?

- Changes the link function from logit to log
- We're now directly estimating the RR
- But we're working with a much less clean likelihood function when we do it – the log function is not restricted to 0/1, as just one example



What Are the Problems?

- #1: Convergence Issues
- Potential Solutions:
  - Provide starting values
    - start=c(Coef1,Coef2,Coef3,etc.)
  - More iterations
    - maxit=X (default X = 25)
  - Relax convergence criteria
    - epsilon=1e-X (default X = 8)
    - USE WITH CAUTION
- It's still not working! Now what?

Alternate Regression Approaches

- Bayesian estimation using MCMC
  - May get around ML convergence issues
  - This is relatively difficult, involves different software packages, has its own model diagnostic difficulties
- Poisson Regression
  - We'll cover Poisson regression more later, but typically used for count data
  - Has been shown, with robust variance, to also estimate the RR
  - Zou G. A modified Poisson regression approach to prospective studies with binary data. *Am J Epidemiol.* 2004;159(7):702-6

In R

```
library(sandwich)

library(sandwich)

library(lmtest)

a <-glm(Survived~Sex,data=titanic,family=poisson)

b.out <- NULL

se.out <- NULL

b.out <- rbind(b.out,a$coef)

se.out <- rbind(se.out, coeftest(a,vcov=sandwich)[,2])
```

Comparison

- Titanic Survival (Male vs. Female)
- OR: 0.099 (0.078,0.12)
- RR (Binomial): 0.29 (0.26,0.32)
- RR (Poisson): 0.29 (0.26,0.32)

General Threats to Cohort Studies

- Cohort Selection
  - Should not have experienced the outcome
  - Should be *at risk* of experiencing the outcome
  - "Healthy Worker Effect"
    - Employed people are generally healthier than unemployed people, so a cohort recruited from an occupation (exposed) + the general public (unexposed) may be biased
  - When do you define exposure?
    - Ever vs. Never?
    - At study start?

Selection Bias

- Your exposure and your outcome both influence an individual's probability of being in your study
- The same as conditioning on a collider
- Instead of controlling (stratifying) by a variable as you do when you condition on a collider, you *only look at one stratum* (the people in your study)

Loss to Follow-up

- Cohort studies generally follow people over some period of time
- Attrition in your cohort is inevitable
- We'll talk about ways to handle this in future lectures
- *Differential* loss to follow-up based on exposure status is a major threat to study validity
  - This is also a challenge in RCTs

The Interaction Between Study Types

- People often view studies as mutually exclusive, clear and distinct categories
- This often isn't the case practically
- This *isn't* the case philosophically
  - It's useful to consider a case-control study as a study nested within a hypothetical (or real) cohort study
- These studies, in turn, are trying to estimate a *counterfactual*
- Make sure you've defined your question well

"TROHOC Fallacy"

- We talked a lot about case-control studies last week, and then I've talked to some other people working on study designs
- There is a preoccupation in case-control studies with defining the cases and controls as a group of people who are sick, and a group of people who are healthy
- Instead, view case-control studies as an efficient means of estimating the results of a much more expensive cohort study
- Your controls should match *the source population*, not be as healthy as possible

When to Do What Study

- Clearly, if the data came to you in a certain way, you're stuck
- Case-control studies are efficient for rare outcomes – you don't have to follow massive population groups to get enough cases to generate inference
- This is useful if your study involves expensive or intrusive data collection

The Role of Time

- If you are at all interested in the *time* it takes for an outcome to occur, cohort studies are the way to go
- We'll discuss survival analysis next Monday
- This is helpful for inevitable outcomes (i.e. death) or where the potential impact of an exposure is slowing an outcome (i.e. increasing time until an AIDS defining illness, first MI, etc.)
- Temporal relationships become clear

Natural Cohorts

- Cohorts are also good for following a natural grouping of people and examining multiple outcomes
- Framingham Study, Nurses Health Study, etc. are famous examples, but there are lots of these
- These types of cohorts are often quite expensive, but invaluable