Cohort Data, Missing Data and Non-Convergence

Questions from Last Class?

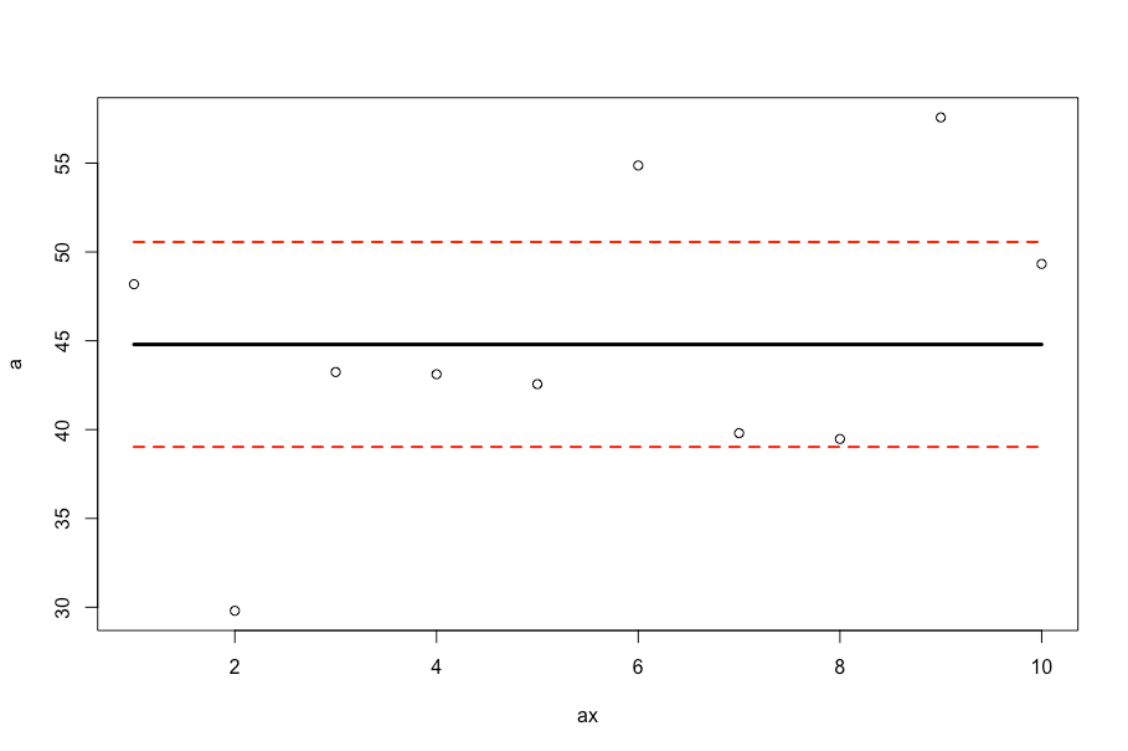
PS 2

- "Is X appropriate" – Remember to consider *all* assumptions behind a given model
- Confidence Intervals
  - Confidence intervals need not cover all points
  - Indeed, with lots of data they *should* not cover all points
  - The confidence interval is the uncertainty about the *fit*, not the variability inherent in the data

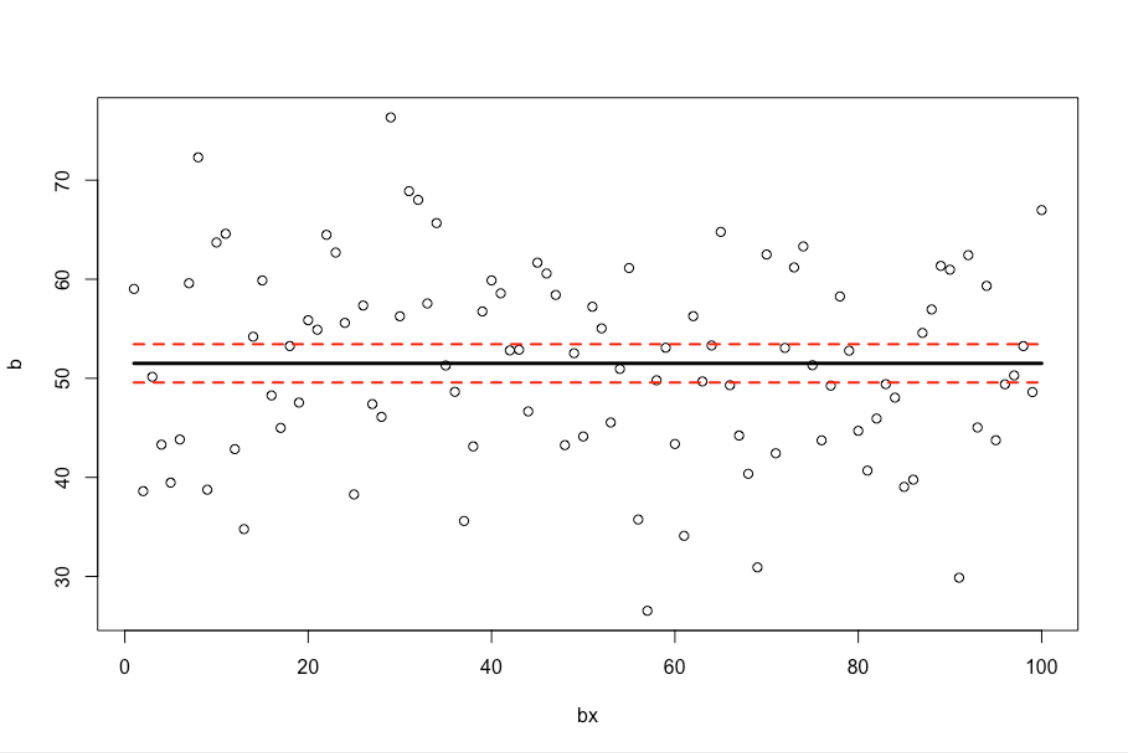
Example

- Consider randomly generated data from a normal distribution with a mean of 50 and a standard deviation of 10
- lm(random_data~1) is a *perfect* model

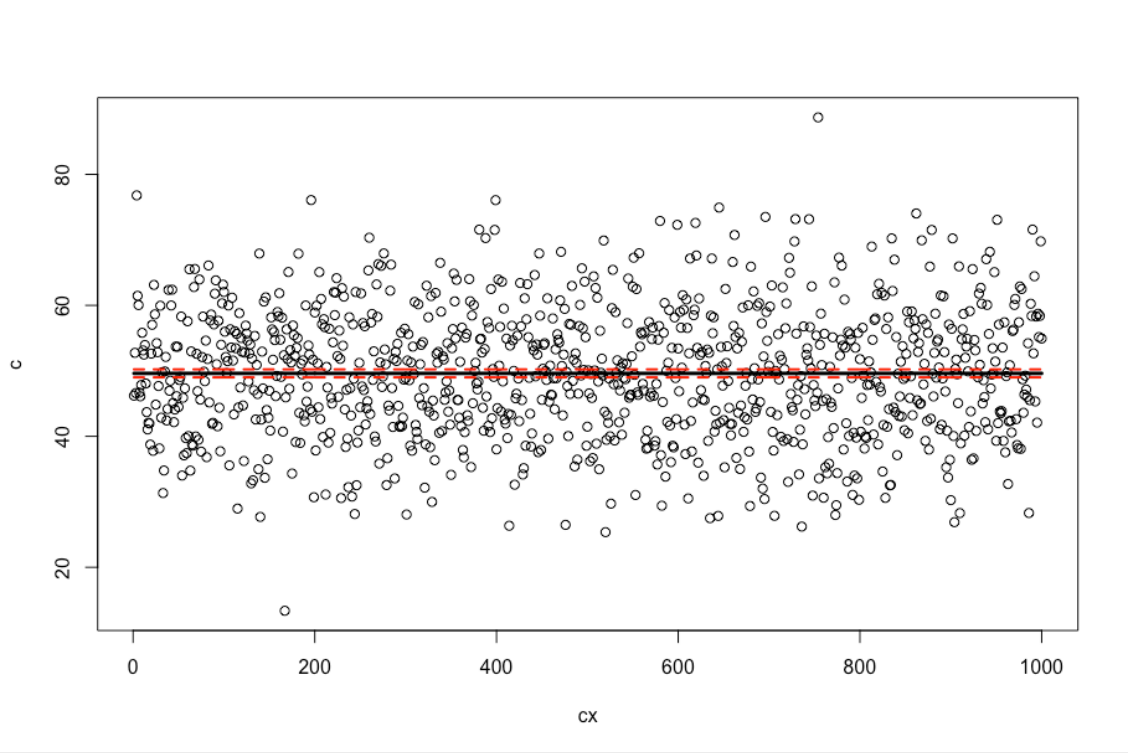
10 Samples



100 Samples



1000 Samples

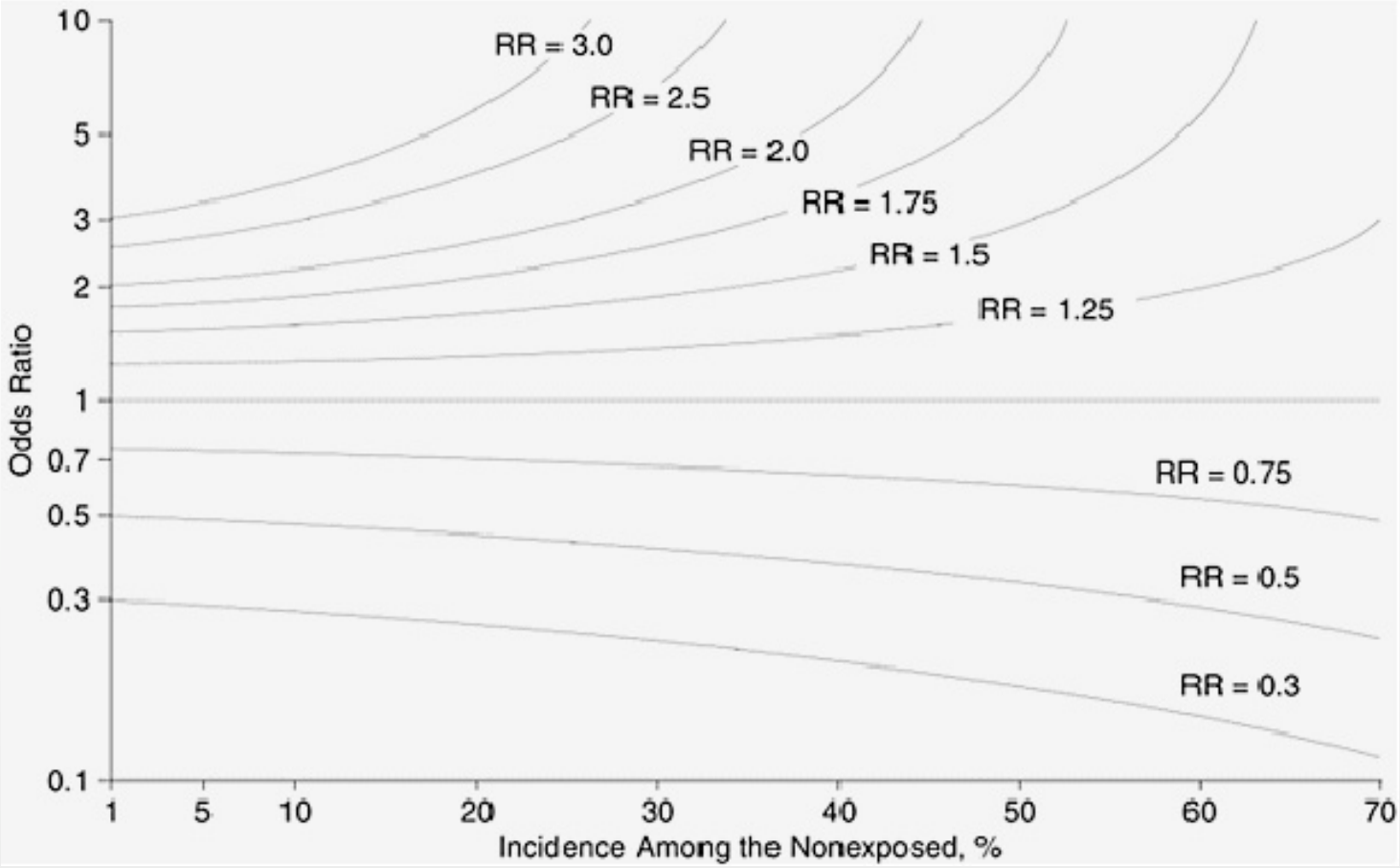


This Week's Focus

- 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

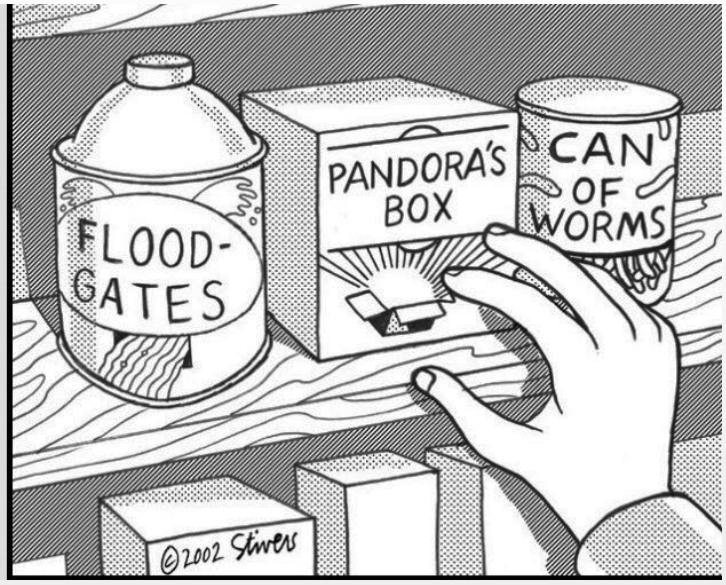
$$\log(p(Y = 1)) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 \ldots$$

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(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

Missing Data

Missing Data

- Missing data is inevitable
- People drop out of studies
- People skip questions
- People mess up their answers to questions
- Samples get ruined
- Lab tests have errors, are below the threshold of detection, etc.

Three “Types” of Missingness

- Missing *Completely* at Random
- Missing at Random
- Missing Not at Random

Missing Completely at Random

- The probability of a data value being missing is independent of all other other data (observed and unobserved)
- Missing data is a random sample of your data
- Common Examples:
  - Low-level data corruption (a bit of memory gets hit by stellar radiation – yes this happens rarely)
  - The power goes out while running a set of samples through a test
  - Someone spills coffee on some paper records

Missing at Random

- The probability of a data value being missing is dependent only on observed data
- *Conditional on all other variables* a missing data point is random
- Examples:
  - Healthcare workers are more likely to have detailed case histories
  - Eligible patients with advanced disease less likely to have complete data
    - May miss appointments, have medical contraindications for some tests, etc.

Missing Not at Random

- The probability of a data value being missing is dependent on some unknown factor
- Examples
  - Low values may be harder to detect by an assay
  - The probability of the result of the assay being missing is related to the value of the assay – which cannot be observed

Dealing with Missing Data

- Complete case analysis
  - Only analyze the data sets with complete data
  - This is the default in many/most software packages
- This can dramatically reduce your sample size
- This is only valid if your missing data is MCAR
- That's a really strong assumption

Indicator Variables

- Assigning missing values an indicator that they're missing, and then treat that as a value of the variable
  - i.e. Yes/No/Missing
- This has been extensively studied and is a very bad idea that will produce biased values

Imputation

- Assign values to the missing variables
- Believe it or not, this is more conservative than complete case analysis
- Mean Imputation: All missing values get the mean value of the non-missing data
- Single Imputation: Build a model to predict the value of missing variables, each missing value takes a predicted value
- Other methods to decide values

Multiple Imputation

- Create many versions of the data
- Build a model to predict the value of missing variables, each missing data is probabilistically assigned
- Example:
  - 40% chance of a binary variable being 1 results in 4 data sets with a value of 1 and 6 with a value of zero
- Run the model on each data set, pool the results
- Details in R.A. Little and D.B. Rubin. 1989. The Analysis of Social Science Data with Missing Values. *Sociological Methods and Research*.

Typical Assumptions

- Missing at Random
- Multivariate normal distribution for missing data
- More elaborate models are possible, even for MNAR
  - This involves modeling the process by which the data is generated