Case-Control Data and Logistic Regression

Questions from Last Class?

Resources for the Curious

- *Categorical Data Analysis Using SAS* by Stokes, Davis and Koch
  - Somewhat staid book, and in the wrong language, but SAS Press books are often very accessible treatments of a topic
  - I have this book in my library
- *Logistic Regression* by Klein and Kleinbaum
- There are *tons* of logistic regression tutorials online, on Coursera, etc. Be aware that many of these are machine learning focused, but the principles are the same

Types of Observational Data

- Most of these got covered in Module 1
- Binary/Continuous
- People/Time/Events
- The decision on *how* to model is largely dictated by *what* you are modeling

Why Regression?

- Module 1 taught you how to calculate measures of association with 2x2 tables, long division, etc.
- Adjusting for confounding using stratification
- All of this seemed to work well enough, can be done in Excel, on a whiteboard, etc. – why bother with regression?

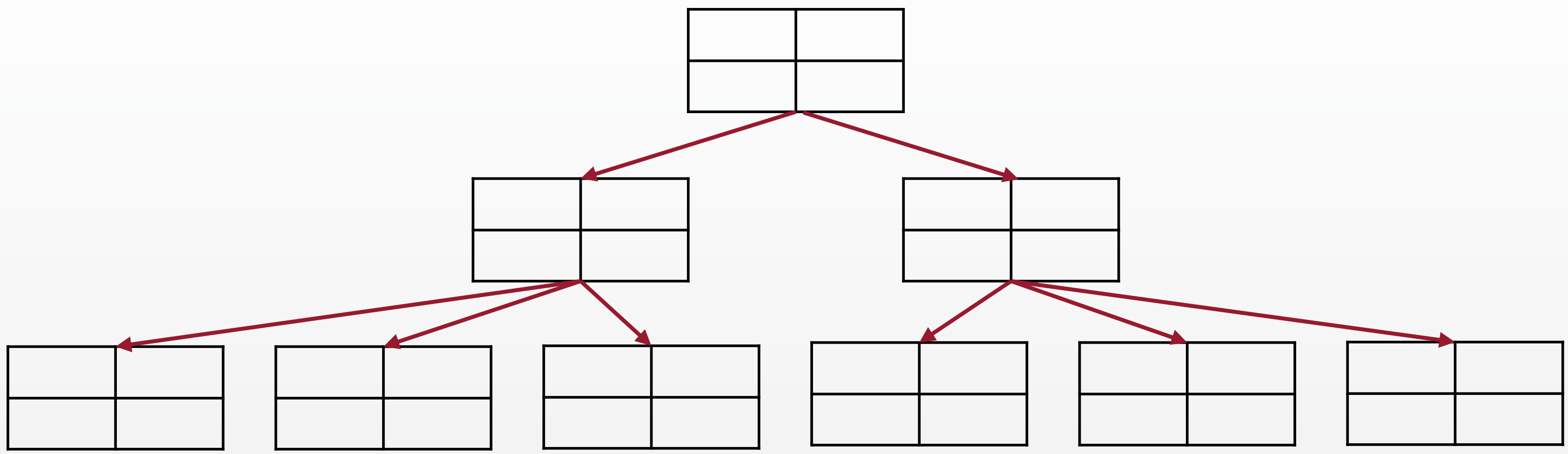
Stratification

100	50
40	220

60	20
20	120

40	30
20	100

Stratification



And this is just for two variables…what about 16?

Strengths of Regression

- Can handle adjustment by *many* variables
- Can handle non-categorical data
- Can smooth/spackle over empty cells
  - If you know what happens to 26 year olds and 28 year olds, you can guess what happens to 27 year olds
- You can predict
  - Given m, X and b, solve for Y
  - Regression is the foundational toolset of machine learning/data science

What Regression Isn't

Regression Can't…

- Automatically fix bad data collection
- Control for bias that it (or you) don't know about
- Solve your sample size problems for you
- …solve *any* of your problems **for** you – regression is a tool, and a dumb one at that

Assumptions and Problems of Regression

- Positivity: An individual has a non-zero probability of having any combination of parameter values
  - Regression assumes cells with 0's happened by chance – what if those cells are impossible?
- Model misspecification: Missing confounders, the wrong distribution, etc. will give you the wrong answer
  - This is, I would argue, the biggest problem in observational epi
- Nonidentifiability: Two (or more) combinations of parameters are equally supported by the data, and there is no "best fit"
- Others we will discuss as the class goes on
- There are *more* assumptions necessary for causal inference, which is beyond the scope of this module

Reading a Regression Equation

- Regression is essentially progressively more complex versions of y = mx +b

$$Y = \underbrace{\beta_0 + \beta_1 A + \varepsilon}_{\text{Linear Predictor}}$$

$$Y = \alpha + \beta_1 A + \gamma \boldsymbol{Z}$$

What's a Link Function?

- A link function is a function that describes the relationship between Y and the rest of the equation
- Linear Predictor: $\mathbf{X}\boldsymbol{\beta}$
- Link function: $g(\mathrm{Y}) = \mathbf{X}\boldsymbol{\beta}$
- Identity: $\mathrm{Y} = \mathbf{X}\boldsymbol{\beta}$
- Log: $\ln(\mathrm{Y}) = \mathbf{X}\boldsymbol{\beta}$
- Logit: $\frac{Y}{1-Y} = \mathbf{X}\boldsymbol{\beta}$

What is a Distribution?

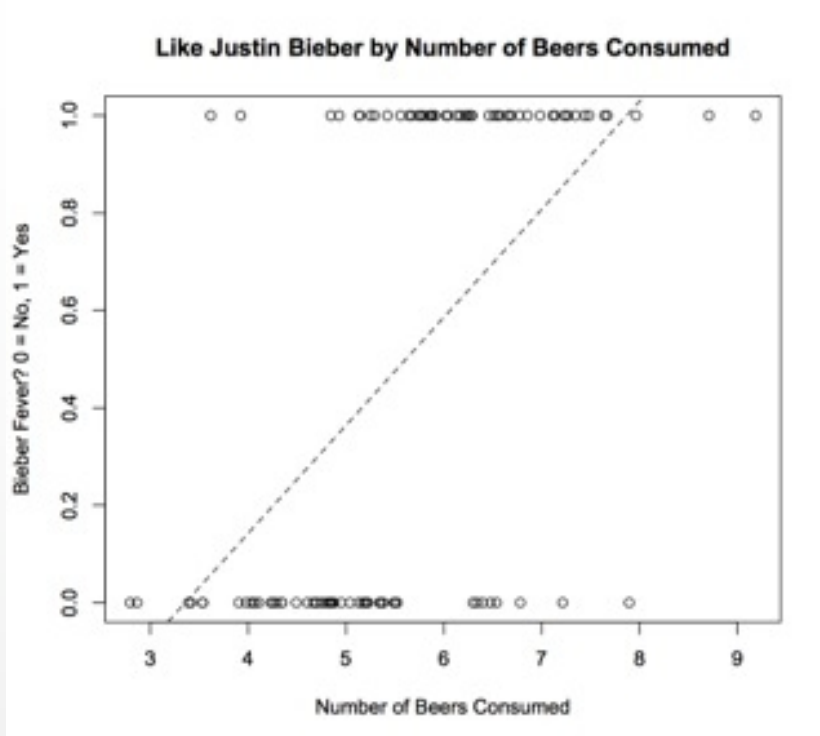
- Linear regression assumes things came from a normal distribution
- This is *often* not true
- Other distributions are common
- Binomial: Binary data
- Poisson/Negative Binomial: Counts and rates
- Exponential/Weibull/Gamma: Time
- When unspecified, it is often assumed to be normal

Least Squares and Maximum Likelihood

- Two ways to estimate the best fitting parameter
- Linear regression often uses least squares
- Most of the other models we will discuss use some form of maximum likelihood

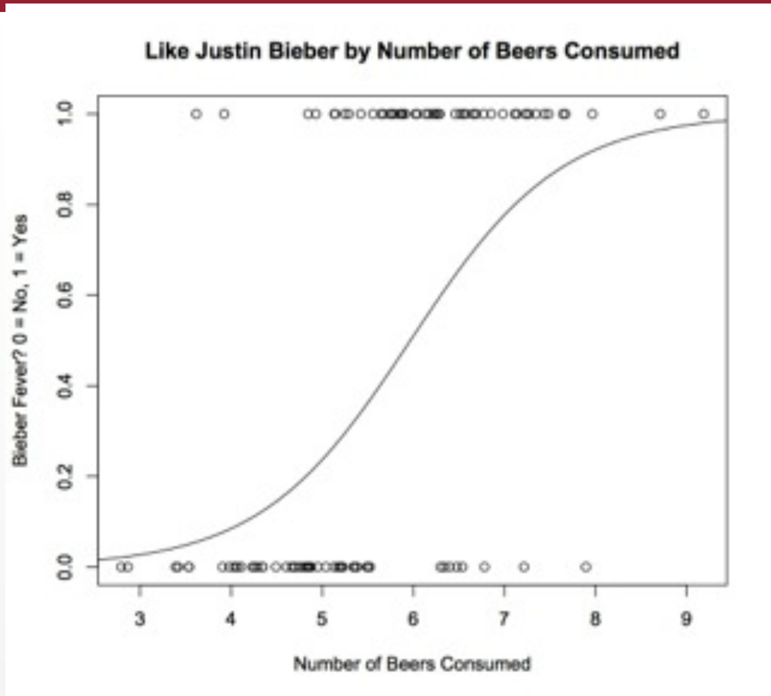
Categorical Data

- Linear regression works really well for continuous, normally distributed outcomes
- But binary outcomes don't really work well in the linear regression context



- May predict values >1 (which doesn't make any sense for a 0/1 value…)
- Assumes a linear function between exposure and the value of the outcome
- Undesirable properties of the residuals

- So we turn to logistic regression
- Instead of predicting a *value* of Y, lets predict the probability that Y = 1



- S-shaped curve bounded at 0 and 1
- Allows for different levels of change over levels of exposure, especially high or low
- Residuals are better behaved

Formal Equation

$$\log\left(\frac{p(Y=1)}{1-p(Y=1)}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 \ldots$$

$$p(Y=1) = \frac{e^{\beta_0+\beta_1 X_1}}{1+e^{\beta_0+\beta_1 X_1}}$$

In R

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

Console ~/Dropbox/Work/Classes/VETPATH571/

```
> summary(glm(Survived ~ Age + Sex, family=binomial(link='logit'),data=train))

Call:
glm(formula = Survived ~ Age + Sex, family = binomial(link = "logit"),
    data = train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.7405  -0.6885  -0.6558   0.7533   1.8989

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  1.277273   0.230169   5.549 2.87e-08 ***
Age         -0.005426   0.006310  -0.860     0.39
Sexmale     -2.465920   0.185384 -13.302  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 964.52  on 713  degrees of freedom
Residual deviance: 749.96  on 711  degrees of freedom
  (177 observations deleted due to missingness)
AIC: 755.96

Number of Fisher Scoring iterations: 4

>
```

An aside…

- In some fields (like economics) you will encounter something called ‘probit’ regression
- This is meant to model the same things as logistic regression
- Uses a different link function (probit instead of logit)
- The reason logistic regression is popular in epidemiology is because the output can be interpreted as an odds ratio
- Probit models are handy for some advanced applications in econometrics

Variable Selection

- One of the biggest hurdles in *all* regression is variable selection
- The challenge is not only to choose *what* variables, but what *form*
  - Recall from previous lectures the different between linear and quadratic forms of a variable
  - We'll pick this thread up in a few slides

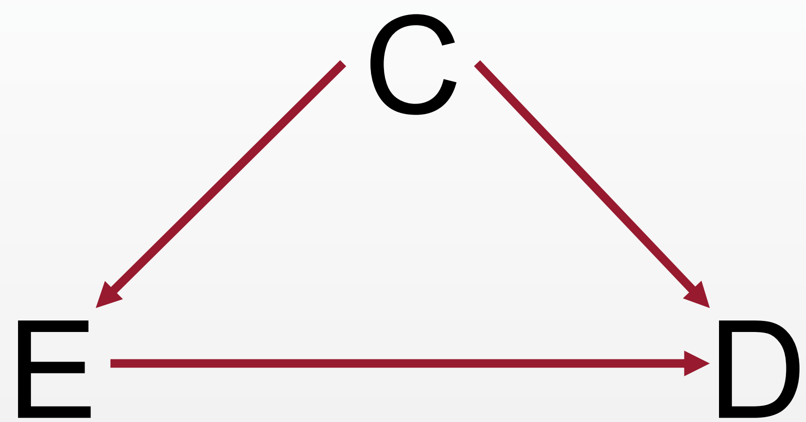
How to Choose Variables

- Realistically, there will be some variables you have to control for, depending on your subject area.
- Use the literature to help inform your choices
- DAGS!

How *Not* To Choose Variables

- Automated algorithms (forward selection, backward selection, stepwise selection)
  - These all add or drop variables based on a significance threshold
  - This can drop variables that are confounders
- *If you must*, $p = 0.05$ is an anticonservative threshold
  - You want to protect yourself from accidentally dropping a needed variable
  - Use a much more generous threshold, like $p = 0.20$

What if I Just Don't Know?



A variable "C" should have a meaningful impact on the E -> D relationship

10% Change-in-Estimate

- If you include the variable vs. if you don't, does the estimate you're interested in change by 10% or more?
- If so, keep the variable
- Hint: log(Estimate 1) – log(Estimate 2) > 0.10

What Shape Should A Variable Be In?

- Just having a linear term assumes a linear response
- Quadratic terms are also possible
- Cubic, quartic etc. are possible, but rare and difficult to interpret or argue that they have biological meaning
- Splines
- You can use AIC to evaluate between potential forms
  - $AIC = -2 \times \ln(likelihood) + 2k$; k = # of paramerers

What if My Variable is Categorical?

- Is it **ordered** and are the steps the same?
  - Percentiles etc.
  - Convert it to a continuous variable
  - If you can't, you can still use it as one
- If not, use indicator variables
  - A series of 0/1 variables for each possible "state" of the original variable - 1
  - **Pick sensible names**

Logistic Regression on Paired or Matched Data

- Unconditional logistic regression overestimates the OR in a matched study
- You can use conditional logistic regression to estimate the conditional likelihood within each strata (each case : *k* controls set)
- Intercept is not estimated, so no way to estimate direct probabilities
  - We don't care, but again, the people who want to use logistic regression for prediction do
- Use *clogit()* in the *survival* package
  - There are statistical reasons why it lives in a package for survival analysis

Exact Logistic Regression

- Useful when you have *very* small samples and lots of cells with zeros
- Computationally very burdensome
- Maximum likelihood relies on asymptotic results, not valid with small sample sizes
- Use *elrm()* in the *elrm* package
- Note that you can get *very* funny answers from exact logistic regression

Logistic Regression on Multinomial Data

- Does it *need* to be multinomial, or can you collapse some categories?
- You can do essentially pairwise comparisons to a fixed baseline, or to the "neighboring" categories
- Use *multinom()* in the *nnet* package
- Good tutorial here: https://stats.idre.ucla.edu/r/dae/multinomial-logistic-regression/

Errors in Logistic Regression

- Convergence issues
- Complete or Quasi-complete seperation