Lecture 6: Epi commands and tables

BIOSTAT 212, Summer 2017

Kristen Aiemjoy
Department of Epidemiology and Biostatistics
University of California San Francisco

P=0.000?

- P-value is a probability so it can't = 0
- Report as:
  - P<0.0005 [gives the most information]
    - P<0.001 [also ok]
    - P<0.01 [also ok]
- To get the exact p-value use the test statistic:
  - `display 1 - chi2(df, chi2)`
  - `display 2*ttail(df, abs(t))`

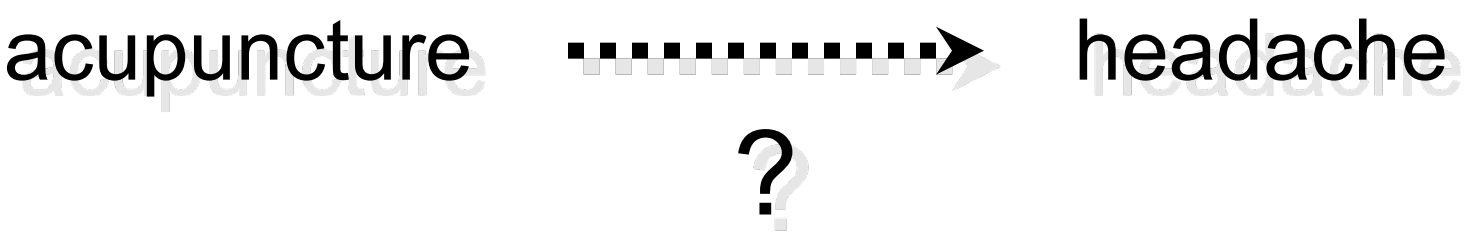
  - `df = degrees of freedom`

Outline for today

- **Introduction to basic epidemiology concepts**
- Estimating Measures of Association in Stata
  - Risk ratios & differences
  - Rate ratios & differences
  - Odds ratios
  - Logistic regression
- Creating and exporting tables
  - `outreg2`
  - `tabout`

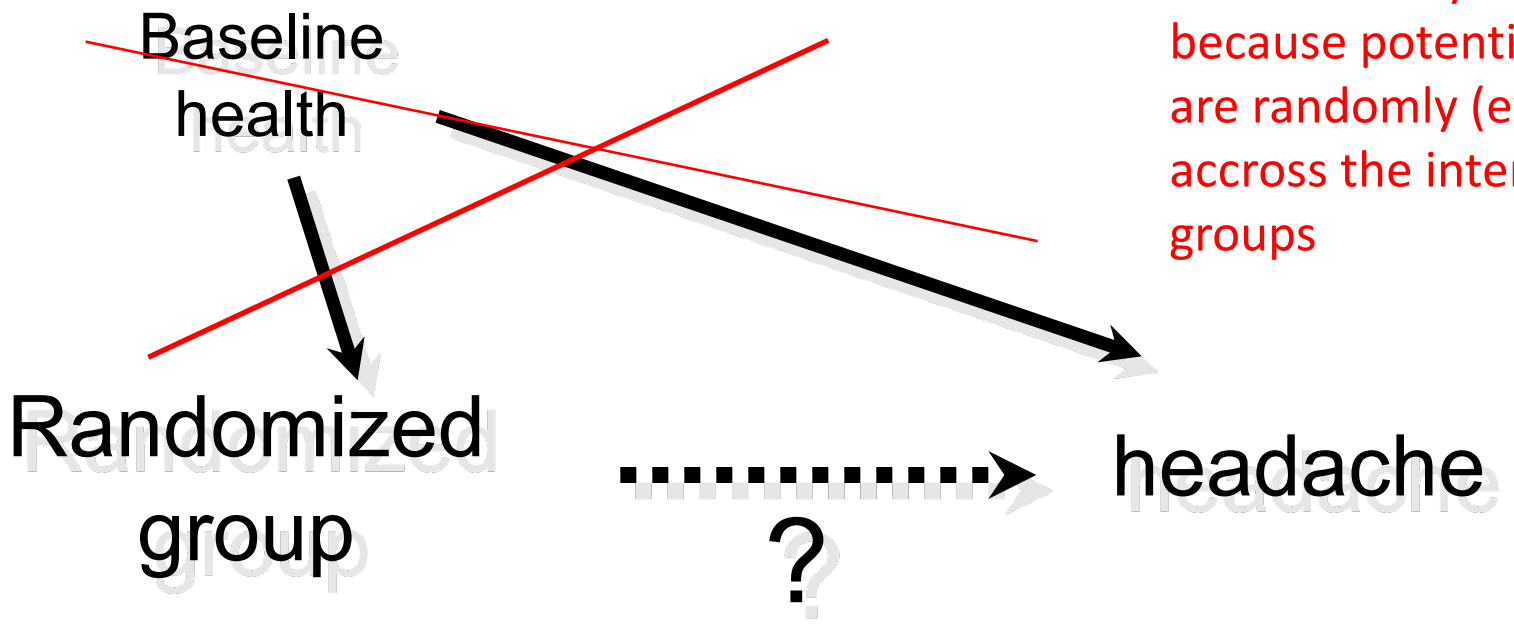
Exposures and Outcomes

- Does “exposure” cause “outcome”?



Confounding

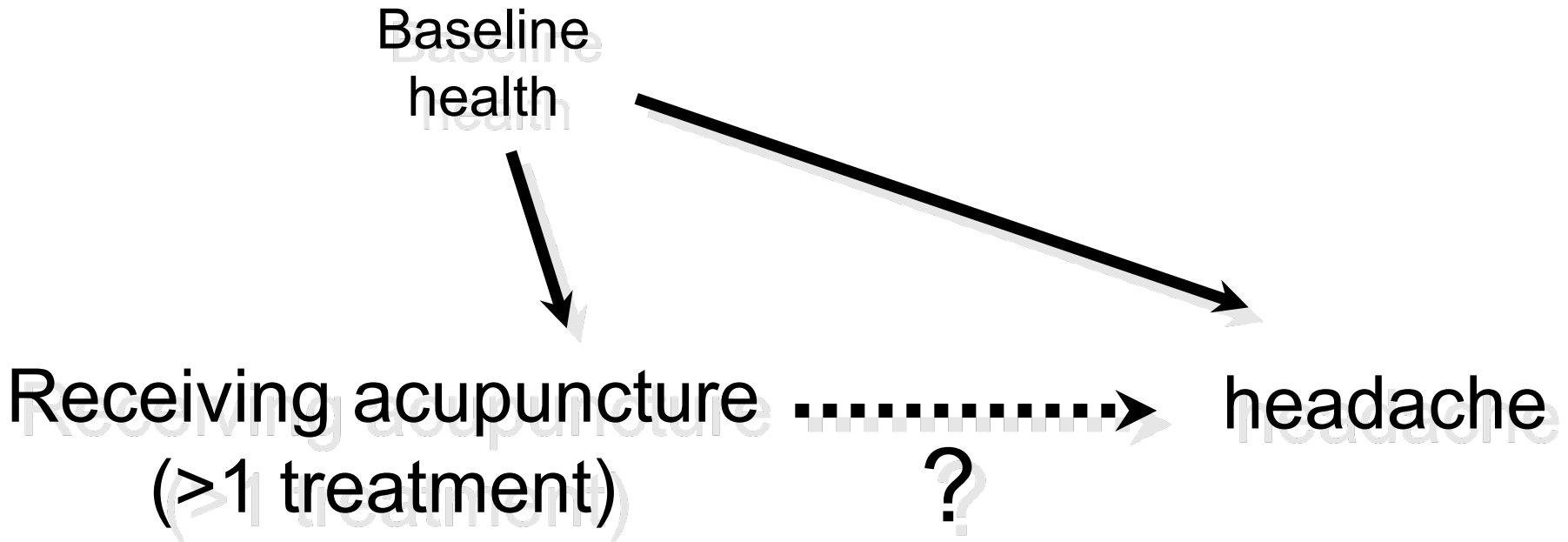
- Do other variables mask or exaggerate the association between exposure and outcome?



In randomized trials you typically don't have to worry about confounding because potential confounding variables are randomly (evenly) distributed accross the intervention and control groups

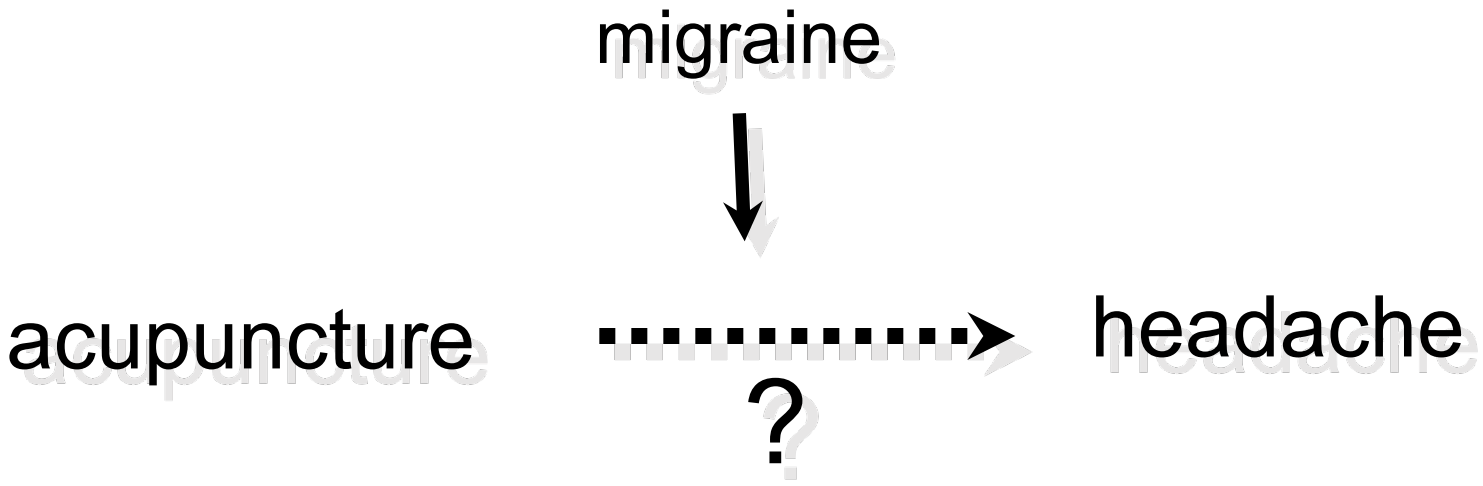
Confounding

- Do other variables mask or exaggerate the association between exposure and outcome?



Interaction

- Is the effect of exposure on outcome different across strata (categories) of a third variable?



Outline for today

- Introduction to basic epidemiology concepts
- **Estimating Measures of Association in Stata**
  - Risk ratios & differences
  - Rate ratios & differences
  - Odds ratios
  - Logistic regression
- Creating and exporting tables
  - `outreg2`
  - `tabout`

Measures of Association

- An association between an exposure and outcome can be assessed by comparing the probability of outcome in the exposed group to the probability of outcome in unexposed group.

- **Ratio (division)**

  [Pr(outcome) in exposed]

  [Pr(outcome) in unexposed]

- **Difference (subtraction)**

  [Pr(outcome) in exposed] – [Pr(outcome) in unexposed]

Ratio Measures of Association

Study Design	Ratio Measure	Stata command
Cross-sectional study	Prevalence Ratio (PR)	`cs O E, by(C)` `logisitic O E C`

Cohort study	Risk Ratio (RR)	`cs O E, by(C)`
		`binreg O E C, rr`
		`regress O E C`
		`logistic O E C`

	Rate Ratio (RR)	`ir O E T` `poisson O E C, irr`
	Hazard Ratio (HR)	`stcox O E C`

Case-control study	Odds Ratio (OR)	`cc O E, by(C)`
		`mhodds O E, by(C)`
		`logistic O E C, or`
		`tabodds O E, adjust(C)`

O = outcomevar E = exposurevar C = covariates T = time

Difference Measures of Association

Study Design	Ratio Measure	Stata command

Cross-sectional study	Prevalence Difference	`cs O E`
		`logistic O E C`
		`regress O E C`

Cohort study	Risk Difference	`cs O E`
		`regress O E C`
		`binreg O E C, rd`

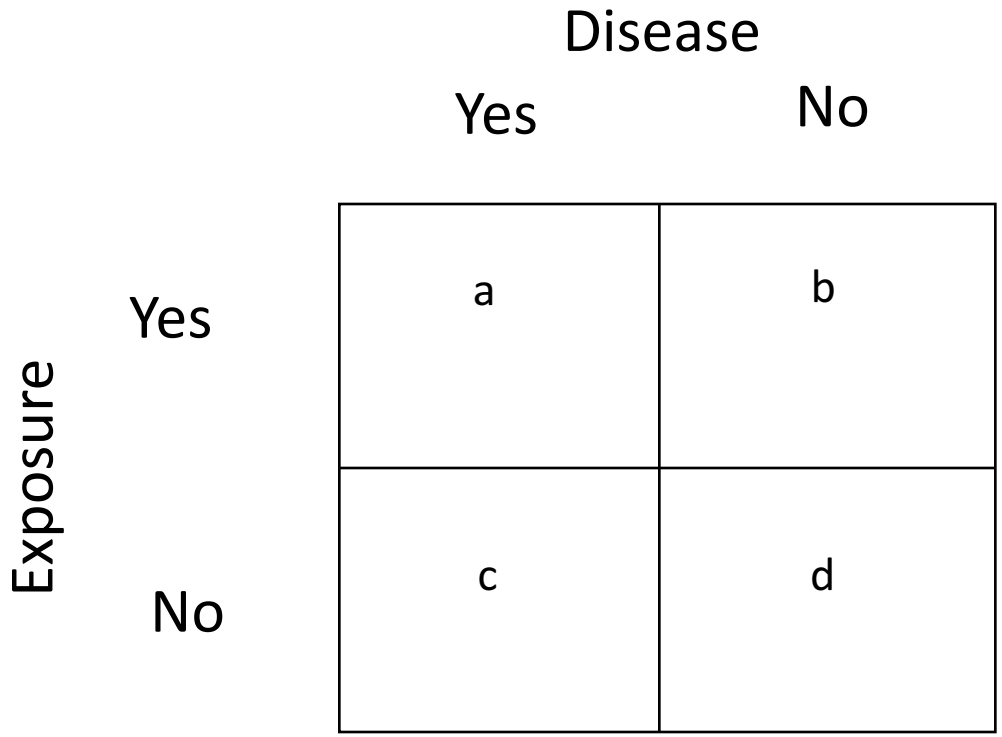
	Rate Difference	`ir O E T, ird` `poisson O E C`
Case-control study	N/A	

`O = outcomevar E = exposurevar C = covariates T = time`

Outline for today

- Introduction to basic epidemiology concepts
- Estimating Measures of Association in Stata
  - **Risk ratios & differences**
  - Rate ratios & differences
  - Odds ratios
  - Logistic regression
- Creating and exporting tables
  - `outreg2`
  - `tabout`

Risk Ratio and Risk Difference



Risk Ratio =

$$\frac{a/(a+b)}{c/(c+d)}$$

Risk Difference =

$$a/(a+b) - c/(c+d)$$

Risk = Probability

cohort studies & RCTs

*similar concept for prevalence ratios, used in cross-sectional studies

Risk Ratios/Differences: `cs`

`cs outcomevar predictorvar`

– calculates risk ratios and risk differences

Options

`,by (varname)` -stratify by a third variable [to look for interaction]

`[, missing]` -include missing data as a category of the stratified variable

Limitations

- Cannot adjust for more than 1 covariate

Practice

- Generate a new dichotomous variable for severe post-study headache.
  - `gen posths_sev = (posths>20)`
- Be careful of missing data
  - `codebook posths posths_sev`
- Instead:

```
1. gen posths_sev = (posths>20)if posths!=.
   label define posths_label 0 “no” 1 “yes”
   label values posths_sev posths_label
   label variable posths_sev “severe headache”
```

-OR-

```
2. recode posths (min/19.99 = 0 ”no") (20/max = 1 ”yes")
   (.=.), gen(posths_sev)
   label variable posths_sev “severe headache”
```

Practice

```
cs posths_sev no_acu
```

- What is the risk of headache in the exposed group (no acupuncture)?

```
di (76/(76+84))
```

- What is the risk of headache in the unexposed group (acupuncture)?

```
di (58/(58+108))
```

- What is the RR?

```
di .48/.35
```

- What is the risk difference?

```
di .48-.35
```

Number of acupuncture treatments

Severe headache (posths>20)

	Yes	No
No acu.	76	84
1+ acu.	58	108

Outline for today

- Introduction to basic epidemiology concepts
- Estimating Measures of Association in Stata
  - Risk ratios & differences
  - **Rate ratios & differences**
  - Odds ratios
  - Logistic regression
- Creating and exporting tables
  - `outreg2`
  - `tabout`

Rate Ratio and Rate Difference

Rate Ratio =

[Cases in exposed/exposed person-time]
[Cases in unexposed/unexposed person-time]

Rate Difference =

[Cases in exposed/exposed person-time] -[Cases in unexposed/unexposed person-time]

Use:

```
ir O E T
poisson O E C, irr
cox O E C
```

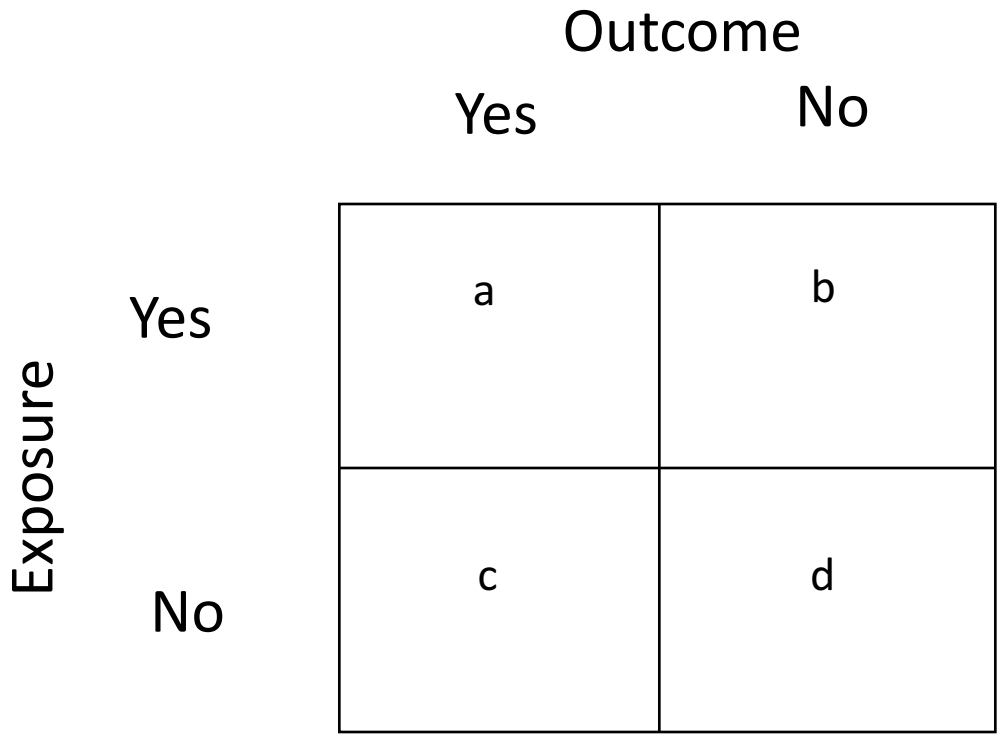
```
O = outcomevar
E = exposurevar
C = covariates
T = time
```

For cohort studies

Outline for today

- Introduction to basic epidemiology concepts
- Estimating Measures of Association in Stata
  - Risk ratios & differences
  - Rate ratios & differences
  - **Odds ratios**
  - Logistic regression
- Creating and exporting tables
  - `outreg2`
  - `tabout`

Odds Ratio



Odds Ratio =

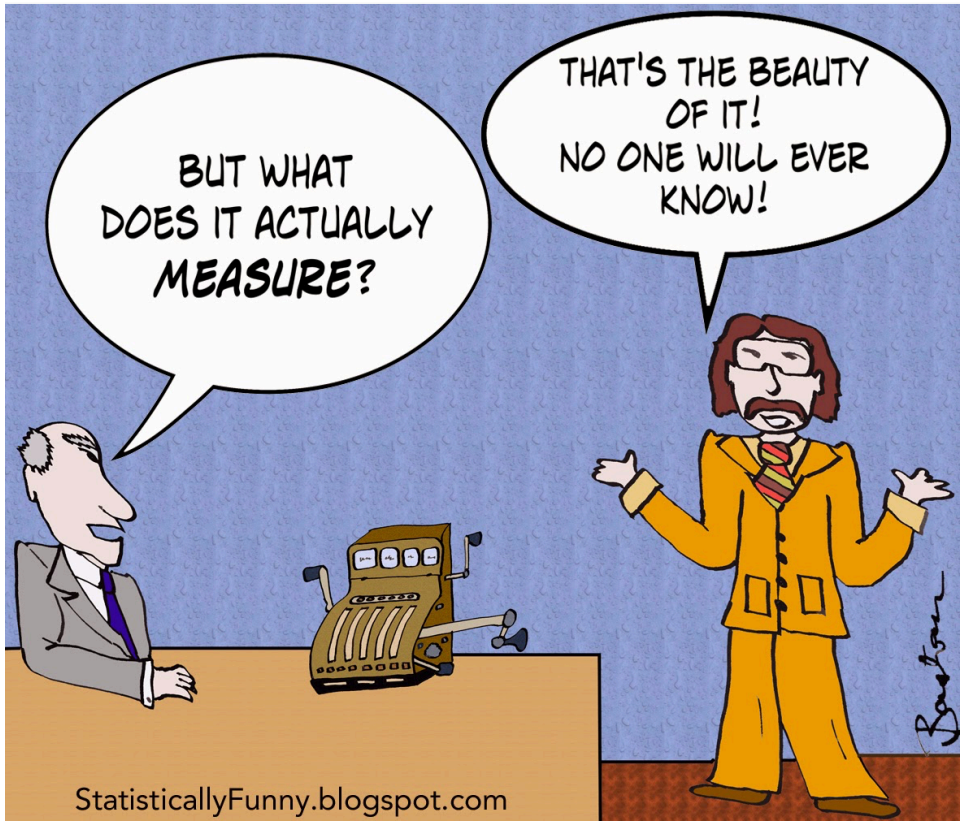
$$\frac{a/b}{c/d}$$

For cross-control studies

The odd thing about Odds Ratios….

- Odds are not-intuitive, difficult to communicate and interpret

- They are farther from the null than the risk ratio
  - exaggerate the association if interpreted as a probability



Odds Ratios: `cc`

`cc outcomevar predictorvar`

– calculates odds ratios

Options

`,by (varname)`	-stratify by a third variable [to look for interaction]
`[, missing]`	-include missing data as a category of the stratified variable

Limitations

Cannot adjust for more than 1 covariate

Practice

- Fill out cells of table

```
tab no_acu posths_sev
```

- Calculate the OR manually using the di function

```
di (76/84)/(58/108)
di (76/58)/(84/108)
```

- Compare OR with `cc` command

```
cc posths_sev no_acu
```

Number of acupuncture treatments	Severe headache (posths>20) Yes	Severe headache (posths>20) No
No acu.	76	84
1+ Acu	58	108

Odds Ratios: `mhodds`

`mhodds outcomevar predictorvar [covariates]`

– calculates odds ratios and can adjust for covariates

Options

`,by (varname)`	-stratify by a third variable [to look for interaction]
`[, missing]`	-include missing data as a category of the stratified variable

Limitations

- Uses stratification to adjust for covariates
  - with several covariates, strata start to get “thin”

Outline for today

- Introduction to basic epidemiology concepts
- Estimating Measures of Association in Stata
  - Risk ratios & differences
  - Rate ratios & differences
  - Odds ratios
  - **Logistic regression**
- Creating and exporting tables
  - `outreg2`
  - `tabout`

Logistic regression

```
logistic outcomevar predictorvar covariates

logit outcomevar predictorvar covariates, or

glm outcomevar predictorvar covariates, family(binomial)
     link(logit) eform
```

Options

```
##
```
-Specify an interaction between 2 variables

If using any categorical predictors with >2 categories, use `i.catvar`

Practice

- What is the OR for the effect of acupuncture on headache?

```
logistic posths_sev no_acu
```

- What is the OR adjusted for baseline health?

```
logistic posths_sev no_acu health_bl
```

`logistic` annotated output

```
logistic posths_sev no_acu health_bl

Logistic regression                             Number of obs[b]    =        326
                                                LR chi2(2)[c]       =      16.31
                                                Prob > chi2[d]      =     0.0003
Log likelihood =  -212.62392[a]                 Pseudo R2[e]        =     0.0369

------------------------------------------------------------------------------
  headache[f] | Odds Ratio[j]   Std. Err.[k]      z[l]    P>|z|[l]     [95% Conf. Interval][m]
--------------+---------------------------------------------------------------
      no_acu |   1.80966[g]    .4209703     2.55   0.011     1.147063    2.855005
  health_bl |   .9824079[h]    .0053527    -3.26   0.001     .9719726    .9929552
       _cons |   1.514117[i]    .5353516     1.17`  0.241     .757177     3.027762
------------------------------------------------------------------------------
```

Logistic regression annotated output, cont.

a Log likelihood - This is the log likelihood of the model. The value `-212.62392` has no meaning in and of itself; rather, this number can be used to help compare nested models.

b Number of obs - This is the number of observations that were used in the analysis. This number may be smaller than the total number of observations in your data set if you have missing values for any of the variables used in the logistic regression.

c LR chi2(2) - This is the likelihood ratio (LR) chi-square test. The number in the parenthesis indicates the number of degrees of freedom. In this model, there are 2 predictors, so there are 2 degrees of freedom.

d Prob > chi2 - This is the probability of obtaining the chi-square statistic given that the null hypothesis is true. In other words, this is the probability of obtaining this chi-square statistic (`16.31`) if there is in fact no effect of the independent variables, taken together, on the dependent variable. In this case, the model is statistically significant because the p-value is `0.0003`

e Pseudo R2 - This is the pseudo R-squared. Logistic regression does not have an equivalent to the R-squared that is found in linear regression; however, many people have tried to come up with one. There are a wide variety of pseudo-R-square statistics. Because this statistic does not mean what R-square means in linear regression (the proportion of variance explained by the predictors), interpret this statistic with great caution.

Logistic regression annotated output, cont.

f headache - This is the outcome in our logistic regression. The variables listed below it are the predictors (exposure and covariates)

g Odds Ratio – These are the odds ratios for the logistic regression equation for predicting the outcome variable (headache) from the exposure (acupuncture) and covariate (baseline health score). They are converted from the log-odds units used for the logistic regression equation.

log(p/1-p) = b0 + b1*no_acu + b2*health_bl

g no_acu – Patients who received no acupuncture had 2.78 times higher odds of severe headache (post headache score >= 30) compared to patients who had at least 1 acupuncture treatment, holding baseline health constant.

h health_bl – For every one-unit increase in baseline health score the odds of severe headache (post headache score >= 30) decrease by .98, holding acupuncture treatment constant.

i constant - This is the expected value of the log-odds of **headache** when all of the predictor variables equal zero.

k Std. Err. - These are the standard errors associated with the ORs. The standard error is used for testing whether the parameter is significantly different from 0; by dividing the parameter estimate by the standard error you obtain a z-value (see the column with z-values and p-values). The standard errors can also be used to form a confidence interval for the parameter, as shown in the last two columns of this table.

l z and **P>|z|** - These columns provide the z-value and 2-tailed p-value used in testing the null hypothesis that the OR is 0.

m [95% Conf. Interval] - This shows a 95% confidence interval for the OR.

Outline for today

- Introduction to basic epidemiology concepts
- Estimating Measures of Association in Stata
  - Risk ratios & differences
  - Rate ratios & differences
  - Odds ratios
  - Logistic regression
- Creating and exporting tables
  - **`outreg2`**
  - `tabout`

outreg2

```
ssc install outreg2
```

- An add-on package to make nice looking regression output and summary tables
- export into .txt, word(.rtf), excel(.xml), LaTeX(.tex), or stata data file(.dta)
- Many options for selecting the statistics you want to include
- Easy to add title, notes
- Specify how you want numbers formatted (2 decimal points for example)

outreg2 example ~ word

```
logistic posths_sev no_acu health_bl
outreg2 using myreg, eform word
```

Will export a **.rtf** file you can open in word

File name, you choose. Will save in whatever working directory you are in

myreg.rtf

VARIABLES	(1) posths_sev
posths_sev	
no_acu	1.810**
	(0.421)
health_bl	0.982***
	(0.00535)
Constant	1.514
	(0.535)
Observations	326

seEform in parentheses
*** p<0.01, ** p<0.05, * p<0.1

outreg2 example ~ excel

```
logistic posths_sev no_acu health_bl
outreg2 using myreg, eform replace excel ctitle(odds
    ratio)
```

Replace last version with the same name, otherwise it will add-on as an additional column

Will export a **.xml** file you can open in word

Change column title

myreg.xml

	A	B	C
1			
2		(1)	
3	VARIABLES	odds ratio	
4			
5	posths_sev		
6			
7	no_acu	1.810**	
8		(0.421)	
9	health_bl	0.982***	
10		(0.00535)	
11	Constant	1.514	
12		(0.535)	
13			
14	Observations	326	
15	seEform in parentheses		
16	*** p<0.01, ** p<0.05, * p<0.1		
17			

outreg2 example ~ LaTeX

```
logistic posths_sev no_acu health_bl
outreg2 using myreg, eform replace tex(pretty)
ctitle(odds ratio)
```

Will export a **.tex** file you can open in word

Formatting option which adds bolds and italics

Opening .tex files

1. Use a cloud-based reader
   https://www.overleaf.com
   (*Lots of cool templates*)

1. Download TeX (Free)
   https://www.latex-project.org/get/

2. R Studio can open and format .tex files (but you need to have TeX installed

1. Open with your text editor, copy and paste to overleaf.com
2. Open directly with TeX (if installed)

myreg.tex

myreg.tex

```
\documentclass[]{article}
\setlength{\pdfpagewidth}{8.5in} \setlength{\pdfpageheight}{11in}
\begin{document}
\begin{center}
\begin{tabular}{lc} \hline
 & (1) \\
VARIABLES & odds ratio \\ \hline
\vspace{4pt} & \begin{footnotesize}\end{footnotesize} \\
posths\_sev &  \\
\vspace{4pt} & \begin{footnotesize}\end{footnotesize} \\
no\_acu & 1.810** \\
\vspace{4pt} & \begin{footnotesize}(0.421)\end{footnotesize} \\
health\_bl & 0.982*** \\
\vspace{4pt} & \begin{footnotesize}(0.00535)\end{footnotesize} \\
Constant & 1.514 \\
 & \begin{footnotesize}(0.535)\end{footnotesize} \\
\vspace{4pt} & \begin{footnotesize}\end{footnotesize} \\
 Observations & 326 \\ \hline
\multicolumn{2}{c}{\begin{footnotesize} seEform in parentheses\end{footnotesize}} \\
\multicolumn{2}{c}{\begin{footnotesize} *** p$<$0.01, ** p$<$0.05, *
p$<$0.1\end{footnotesize}} \\
\end{tabular}
\end{center}
\end{document}
```

VARIABLES	(1) odds ratio
posths_sev	
no_acu	1.810**
	(0.421)
health_bl	0.982***
	(0.00535)
Constant	1.514
	(0.535)
Observations	326

seEform in parentheses

*** p<0.01, ** p<0.05, * p<0.1

Overleaf demo

Free & Interactive Online Introduction to LaTeX

https://www.overleaf.com/latex/learn/free-online-introduction-to-latex-part-1#.WaTwo62ZNjQ

Outline for today

- Introduction to basic epidemiology concepts
- Estimating Measures of Association in Stata
  - Risk ratios & differences
  - Rate ratios & differences
  - Odds ratios
  - Logistic regression
- Creating and exporting tables
  - `outreg2`
  - **`tabout`**

tabout

```
ssc install tabout
```

- An add-on package to export summary tables
- Currently you can only export .txt or .csv files
- Beta version can export to docX and excel

  http://tabout.net.au/docs/home.php
- Include several variables, summary statistics and test statistics of your choice

`tabout` example

```
tabout group using table1.csv, replace ///
cell(mean basehs mean oneyrhs mean posths) format(1p 1p 1p) ///
style(csv) sum clab("Baseline" "Post"  "1yrPost") npos(col)
```

	A	B	C	D	E
1	intervention group	Mean	Mean	Mean	N
2		Baseline	Post	1yrPost	N
3	control	27.50%	22.30%	24.50%	196
4	acupuncture	25.60%	16.20%	19.10%	205
5	Total	26.50%	19.10%	21.60%	401
6					