Lecture 6: Tables in Stata

BIOSTAT 212

Amanda Irish | amanda.irish@ucsf.edu
Department of Epidemiology and Biostatistics
University of California San Francisco

Agenda:

1. Stored results & matrices
2. putdocx
3. Descriptive tables
4. Regression/results tables
5. Other table-making options

Stored Results

Stored Results

- Many commands in Stata save output in memory
- Output useful to us is stored as **scalars** (individual numbers), **matrices** (tables or data-frames that only include numbers), or **functions**
- **Scalars** and **matrices** useful for creating tables

Stored Results

- Results of commands are stored as different "classes" according to the type of command
- Useful to us are `r()` and `e()`
  - r-class results are produced by general statistical commands like `summarize`
  - e-class results are produced by estimating statistical commands like `regress`
- Stay in Stata's memory until the same command is run again

View Stored Results: `return list`

```
. summ price

    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
       price |         74    6165.257    2949.496       3291      15906

. return list

scalars:
                  r(N) =  74
              r(sum_w) =  74
               r(mean) =  6165.256756756757
                r(Var) =  8699525.974268788
                 r(sd) =  2949.495884768919
                r(min) =  3291
                r(max) =  15906
                r(sum) =  456229
```

What is a matrix?

- Rectangular array (or table) of numbers
- Dimensions *m* x *n*

$$
\begin{array}{c}
 \\
1 \\
2 \\
3 \\
\vdots \\
m
\end{array}
\begin{array}{c}
\begin{array}{cccc} 1 & 2 & \dots & n \end{array} \\
\begin{bmatrix}
a_{11} & a_{12} & \dots & a_{1n} \\
a_{21} & a_{22} & \dots & a_{2n} \\
a_{31} & a_{32} & \dots & a_{3n} \\
\vdots & \vdots & \vdots & \vdots \\
a_{m1} & a_{m2} & \dots & a_{mn}
\end{bmatrix}
\end{array}
$$

Store results in a matrix

```
. return list

scalars:
                  r(N) =  74
              r(sum_w) =  74
               r(mean) =  6165.256756756757
                r(Var) =  8699525.974268788
                 r(sd) =  2949.495884768919
                r(min) =  3291
                r(max) =  15906
                r(sum) =  456229

. matrix A=r(mean)

. matrix B=r(Var)

. matlist A

             |        c1
-------------+-----------
          r1 |  6165.257
```

Useful Matrix Operations

- `matrix C = A \ B`
  - Add rows to matrix
- `matrix D = A , B`
  - Add columns to matrix
- `matrix E = D'`
  - Transpose matrix

```
. matrix C=A\B

. matlist C
```

	c1
r1	6165.257
r1	8699526

```
. matrix D=A,B

. matlist D
```

	c1	c1
r1	6165.257	8699526

```
. matrix E=D'

. matlist E
```

	r1
c1	6165.257
c1	8699526

help matrix

Description

Comments are provided below under the following headings.

	Subject	Also see help
1.	Inputting matrices by hand	**matrix define**
2.	Matrix expressions	
	Matrix operators	matrix operators
	Matrix functions	matrix functions
3.	Matrix subscripting	matrix subscripting
4.	Submatrix extraction	matrix extraction
5.	Submatrix substitution	matrix substitution
6.	Data <——> Matrix conversion	**mkmat**

putdocx

Export matrices to word with `putdocx`

1. `putdocx clear` Closes any open putdocx
2. `putdocx begin` Creates document to export
3. `putdocx table tablename=matrix(matrixname)`
4. `putdocx save filename [, replace append]`

putdocx table (options)

table_options	Description
memtable	keep table in memory rather than add it to document
width(#[*unit*|%])	set table width
halign(*hvalue*)	set table horizontal alignment
indent(#[*unit*])	set table indentation
layout(*layouttype*)	adjust column width
cellmargin(*cmarg*, #[*unit*])	set margins for each table cell
cellspacing(#[*unit*])	set spacing between adjacent cells and the edges of the table
border(*bspec*)	set pattern, color, and width for border
headerrow(#)	set number of the top rows that constitute the table header
title(*string*)	add a title to the table
note(*string*)	add notes to the table

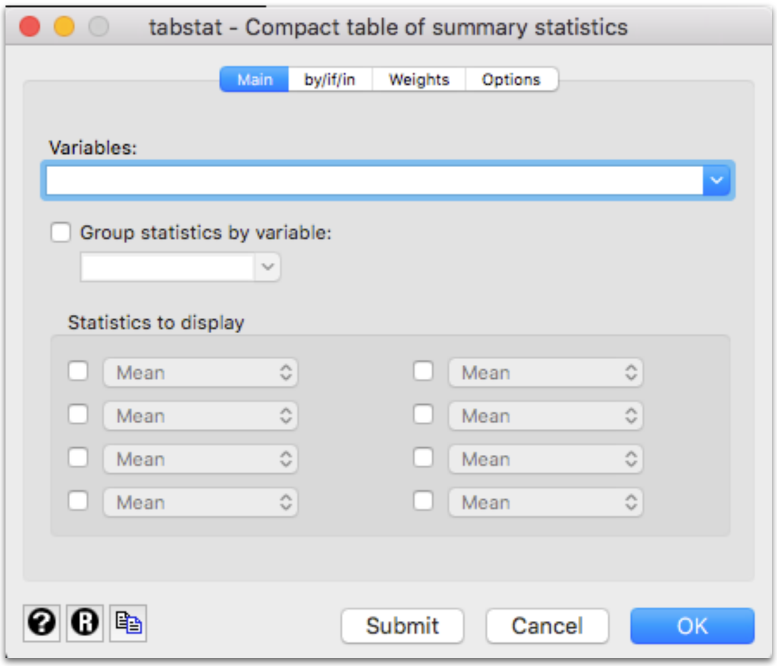
cell_options	Description
append	append objects to current content of cell
rowspan(#)	merge cells vertically
colspan(#)	merge cells horizontally
span(#1, #2)	merge cells both horizontally and vertically
linebreak[(#)]	add line breaks into the cell
cell_fmt_options	options that control the look of cell contents

Useful putdocx options

```
, rownames
, colnames
, title("table title")
, layout(autofitcontents)
```

Descriptive Tables

Table of summary statistics: `tabstat`



foreign	variable	mean	sd	min	max
Domestic	price	6072.423	3097.104	3291	15906
	weight	3317.115	695.3637	1800	4840
	mpg	19.82692	4.743297	12	34
	rep78	3.020833	.837666	1	5
Foreign	price	6384.682	2621.915	3748	12990
	weight	2315.909	433.0035	1760	3420
	mpg	24.77273	6.611187	14	41
	rep78	4.285714	.7171372	3	5

```
sysuse auto, clear
tabstat price weight mpg rep78, by(foreign)
     stat(mean sd min max) nototal long col(stat)
```

Statistics > Summaries, tables, and tests > Other tables > Compact table of summary statistics

`tabstat` syntax

`tabstat varlist, statistics(______)`

-produces a summary table of statistics for all variables in `varlist`

Statistics

`mean`	-mean
`n`	-count of nonmissing observations
`sum`	-sum
`max`	-maximum
`min`	-minimum
`range`	-range = max - min
`sd`	-standard deviation
`variance`	-variance
`median`	median

*more options-check help file

`tabstat` syntax – useful options

`, by(varname)`	
`, save`	-stores results in r(table)
`, col(stats)`	-var in rows, stats in columns
`, nototal`	-with by(): suppresses output for “total”
`, long`	-with by(): shows all var names

Sample workflow for `tabstat`

1. `tabstat varlist, statistics(______) save`
2. `return list`
3. `matrix matname = r(Stat___)`
4. `matlist matname`
5. +/- manipulate matrix if needed
6. `putdocx clear` (if needed)
7. `putdocx begin`
8. `putdocx table tablename=matrix(matrixname)`
9. `putdocx save docname`

Load acupuncture data set

```
use "acupunture.dta", clear
```

Practice

- Produce a table of [mean, standard deviation, min and max] for post-study headache score by intervention group

- What is the mean post-study headache score in the intervention group?

- What is the mean post-study headache score in the control group?

- Save r(StatTotal) in a matrix called “overall”

- Use putdocx to export the matrix to Word

Practice - answers

- Produce a table of [mean, standard deviation, min and max] for post-study headache score by intervention group

```
tabstat posths, by(group) stat(mean sd min max) save
```

- What is the mean post-study headache score in the intervention group?
  - 19.05

- What is the mean post-study headache score in the control group?
  - 24.48

- Save r(StatTotal) in a matrix called "overall"
  - `matrix overall=r(StatTotal)`

Create summary dataset: `collapse`

`collapse` [(stat)] `newvarname1=varname1` [[(stat)] ...]

Where stat can be:

`mean`	-mean
`median`	-median
`count`	-counts non-missing values
`semean`	-standard error of the mean
`sum`	-sum

IMPORTANT: collapse creates a summary dataset. ALWAYS use "`preserve`" before using collapse and "`restore`" afterwards to recover the original dataset!

`collapse` example

```
preserve // California vs. non_FLP

drop if dob_y < 1997

collapse (mean) mhi_all, by(dob_y flp_region postpolicy)

twoway  (line mhi_all dob_y if flp_region==0 & postpolicy==0, lcolor(gs4) lpattern(dash)) ///
        (line mhi_all dob_y if flp_region==0 & postpolicy==1 lcolor(gs4) lpattern(dash))  ///
        (line mhi_all dob_y if flp_region==1 & postpolicy==0, lcolor(blue))               ///
        (line mhi_all dob_y if flp_region==1 & postpolicy==1, lcolor(blue)),              ///
        xtitle("Child's year of birth") ytitle("Mean mental health indicator score")      ///
        xline(2004, lcolor(black) lpattern(dot))                                          ///
        legend(on order (1 "Non-FLP states" 3 "California"))                             ///
        graphregion(color(white)) xsize(6) ysize(4)                                       ///
        name(MHI_CA)
restore
```

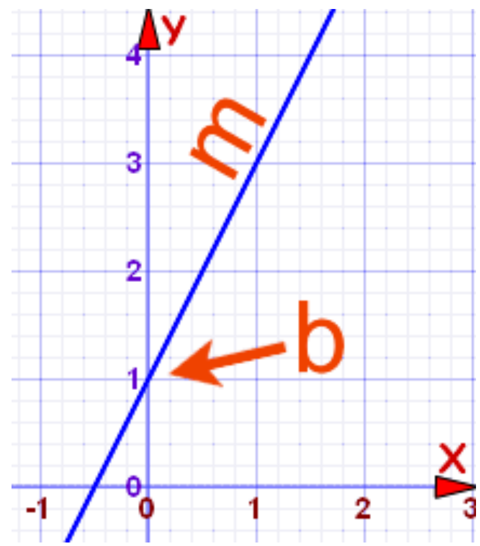
2 x 2 tables: tabulate

```
. tabulate sex group, matcell(sexgroup)

           |  intervention group
       sex |   control  acupunctu |     Total
-----------+----------------------+----------
      male |        31         33 |        64
    female |       165        172 |       337
-----------+----------------------+----------
     Total |       196        205 |       401

. matlist sexgroup

             |        c1         c2
-------------+----------------------
          r1 |        31         33
          r2 |       165        172
```

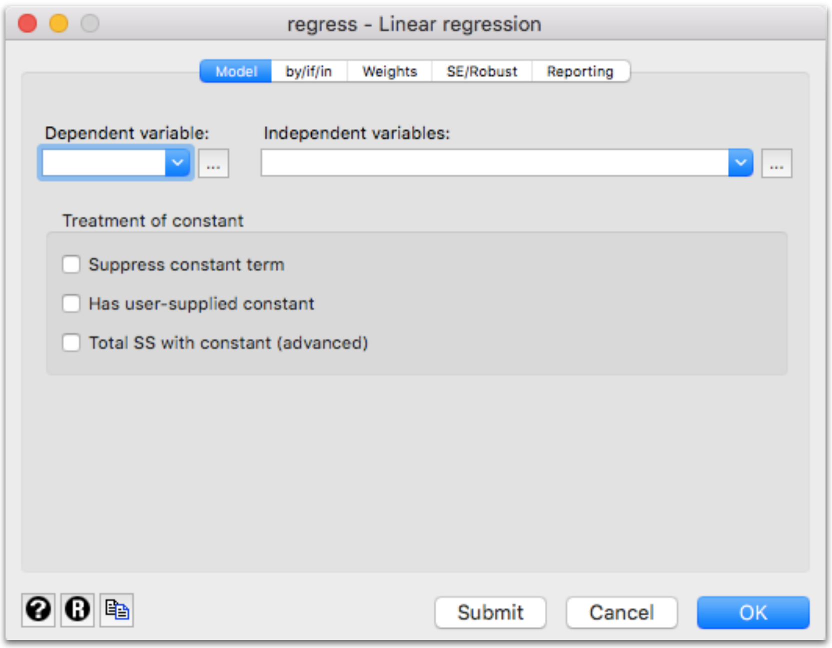
Regression/Results Tables

What is a linear **regression**?

- Fit a linear equation to your data
- $y = mx + b$
  - m=slope
  - b=intercept
  - y=outcome / dependent
  - x=predictor / independent



Linear regression




```
. regress mpg weight

      Source |       SS           df       MS      Number of obs   =        74
-------------+----------------------------------   F(1, 72)        =    134.62
       Model |   1591.9902         1   1591.9902   Prob > F        =    0.0000
    Residual |  851.469256        72  11.8259619   R-squared       =    0.6515
-------------+----------------------------------   Adj R-squared   =    0.6467
       Total |  2443.45946        73  33.4720474   Root MSE        =    3.4389

------------------------------------------------------------------------------
         mpg |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      weight |  -.0060087   .0005179   -11.60   0.000    -.0070411   -.0049763
       _cons |   39.44028   1.614003    24.44   0.000     36.22283    42.65774
------------------------------------------------------------------------------
```

```
sysuse auto
regress mpg weight
```

Statistics > Linear models and related > Linear regression

Practice

- Do age and chronicity (years with headache) have a linear relationship?

  - Step 1: Visualize (hint: use a scatterplot and linear fit line)

  - Step 2: Linear regression

Practice

- Do age and chronicity (years with headache) have a linear relationship?

  - Step 1: Visualize

```
scatter chronicity age || lfit  chronicity age
```

  - Step 2: Linear regression

```
regress chronicity age
```

Linear regression annotated output

regress chronicity age

Source[a]	SS[b]	df[c]	MS[d]
Model	18021.6443	1	18021.6443
Residual	57783.9268	399	144.821872
Total	75805.5711	400	189.513928

Number of obs[e]	=	401
F(1, 399)[f]	=	124.44
Prob > F[f]	=	0.0000
R-squared[g]	=	0.2377
Adj R-squared[h]	=	0.2358
Root MSE[i]	=	12.034

chronicity[j]	Coef.[k]	Std. Err.[l]	t[m]	P>|t|[m]	[95% Conf.	Interval][n]
age	.6074257	.0544519	11.16	0.000	.5003772	.7144742
_cons	-6.202495	2.55145	-2.43	0.015	-11.21846	-1.18653

Linear regression annotated output

a. Source – This is the source of variance, Model, Residual, and Total. The Total variance is partitioned into the variance which can be explained by the independent variables (Model) and the variance which is not explained by the independent variables (Residual, sometimes called Error). Note that the Sums of Squares for the Model and Residual add up to the Total Variance, reflecting the fact that the Total Variance is partitioned into Model and Residual variance.

b. SS – These are the Sum of Squares associated with the three sources of variance, Total, Model and Residual. These can be computed in many ways. Conceptually, these formulas can be expressed as:
SSTotal The total variability around the mean.
SSResidual The sum of squared errors in prediction.
SSModel The improvement in prediction by using the predicted value of Y over just using the mean of Y.

- Note that the SSTotal = SSModel + SSResidual.
- Note that SSModel / SSTotal is equal to 0.2377, the value of R-Squared.

c. df – These are the degrees of freedom associated with the sources of variance. The total variance has N-1 degrees of freedom. In this case, there were N=401 study participants, so the DF for total is 400. The model degrees of freedom corresponds to the number of predictors + the intercept minus 1 (K-1). Including the intercept, there are 2 predictors, so the model has 2-1=1 degrees of freedom. The residual degrees of freedom is the DF total minus the DF model, 400 – 1 is 399.

d. MS – These are the Mean Squares, the Sum of Squares divided by their respective DF. For the Model, 18021.6443 / 1 = 18021.6443. These are computed so you can compute the F ratio, dividing the Mean Square Model by the Mean Square Residual to test the significance of the predictors in the model.

https://stats.idre.ucla.edu/stata/output/regression-analysis/

Linear regression annotated output, cont.

e. **Number of obs** – This is the number of observations used in the regression analysis.

f. **F** and **Prob > F** – The F-value is the Mean Square Model divided by the Mean Square Residual. The p-value is compared to your alpha level (typically 0.05) and, if smaller, you can conclude "Yes, the independent variables reliably predict the dependent variable". Note that this is an overall significance test assessing whether the group of independent variables when used together reliably predict the dependent variable, and does not address the ability of any of the particular independent variables to predict the dependent variable.

g. **R-squared** – R-Squared is the proportion of variance in the dependent variable (**chronicity**) which can be predicted from the independent variable (**age**). This value indicates that 23.77% of the variance in chronicity can be predicted from the variables **age**. Note that this is an overall measure of the strength of association, and does not reflect the extent to which any particular independent variable is associated with the dependent variable.

h. **Adj R-squared** – Adjusted R-square. As predictors are added to the model, each predictor will explain some of the variance in the dependent variable simply due to chance. One could continue to add predictors to the model which would continue to improve the ability of the predictors to explain the dependent variable, although some of this increase in R-square would be simply due to chance variation in that particular sample. The adjusted R-square attempts to yield a more honest value to estimate the R-squared for the population.

i. **Root MSE** – Root MSE is the standard deviation of the error term, and is the square root of the Mean Square Residual (or Error).

https://stats.idre.ucla.edu/stata/output/regression-analysis/

Linear regression annotated output, cont.

j. chronicity – This column shows the dependent variable at the top (**chronicity**) with the predictor variables below it (**age** and **_cons**). The last variable (**_cons**) represents the constant, also referred to in textbooks as the Y intercept, the height of the regression line when it crosses the Y axis. In other words, this is the predicted value of **chronicity** when all other variables are 0.

k. Coef. – These are the values for the regression equation for predicting the dependent variable from the independent variable. These estimates tell the amount of increase in chronicity that would be predicted by a 1 unit increase in the predictor.

Chronicity_Predicted = -6.2 +.61*age

For every year increase in age, there is a predicted .61 increase in chronicity

l. Std. Err. – These are the standard errors associated with the coefficients. The standard error is used for testing whether the parameter is significantly different from 0 by dividing the parameter estimate by the standard error to obtain a t-value. 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.

m. t and **P>|t|** – These columns provide the t-value and 2-tailed p-value used in testing the null hypothesis that the coefficient (parameter) is 0.

n. [95% Conf. Interval] – This shows a 95% confidence interval for the coefficient.

https://stats.idre.ucla.edu/stata/output/regression-analysis/

Linear regression output interpretation

Coefficient for age (*m*) = 0.61

Intercept (*b*) = -6.2

Equation: chronicity = 0.61**age* – 6.2

For every year increase in age, there is a predicted .61 increase in chronicity.

For 50 year olds, the predicted chronicity is:

y = (0.6074257*50) -6.202495

= 24.16879

```
. sum chronicity if age==50
```

Variable	Obs	Mean	Std. Dev.	Min	Max
chronicity	15	18.93333	14.22506	2	38

Making results tables: estimates table

1. Label your variables and values!

2. Run regressions and store estimates
   (`estimates store modelname`)

3. Create estimates table
   (`estimates table modelnamelist`)

4. Export to Word
   (`putdocx table tablename=etable`)

Other Table Making Options

putexcel

- Similar to putdocx, will export results/tables directly into excel so you can format them there.

```
/*
    putexcel: export matrix created above to excel
*/

putexcel set "output/Lecture6Notes", sheet(example) modify
putexcel A1=matrix(controlres), names nformat(number_d2)
putexcel A10=matrix(acures), names nformat(number_d2)
```

User-written packages (`findit packagename`)

- `table1`: good for creating table1 in excel
- `tabout`: exports summary tables
- `outreg2`: makes nice looking regression output and summary tables (using LaTeX)

table1

```
ssc install table1
```

`table1` generates a "table 1" of characteristics for a manuscript, which may be either continuous or categorical. The observations are often grouped, with a "p-value" column on the right comparing the characteristics between groups.

```
table1, vars(var1 type \ var2 type)
```

<u>variable `type` options</u>

`contn` - continuous, normally distributed

`conts` - continuous, skew

`cat` - categorical, groups compared using Pearson's chi-squared

`cate` - categorical, groups compared using Fisher's exact test

`bin` - binary, groups compared using Pearson's chi-squared

`bine` - binary, groups compared using Fisher's exact test

table1, cont.

```
table1, by(varname) vars(var1 type \ var2 type)
```

<u>additional options</u>

`, by(varname)`	- report summary by a group variable
`test`	- include column describing the significance test
`pdp(#)`	- max number of decimal places in p-value
`onecol`	- report categorical variable levels underneath variable name instead of in separate column
`format(%fmt)`	- default display format for contn and conts variables
`cformat(%fmt)`	- default display format for cat and bin variables
`missing`	- don't exclude missing values from categorical variables
`saving(filename.xls)`	- save table to Excel file

Example

```
table1, by(heartatk) vars(bmi contn \ bpsystol contn \ sex cat) saving(sumtable.xls)
```

Factor	Level	heartatk = 0	heartatk = 1	p-value
N		9873	476	
Body Mass Index (BMI), mean (SD)		25.49638 (4.913941)	26.39891 (4.870273)	<0.001
systolic blood pressure, mean (SD)		130.4081 (23.23776)	140.813 (23.0926)	<0.001
1=male, 2=female				

	A	B	C	D	E
1	Factor	Level	heartatk = 0	heartatk = 1	p-value
2	N		9873	476	
3	Body Mass Index (BMI), mean (SD)		25.49638 (4.913941)	26.39891 (4.870273)	<0.001
4	systolic blood pressure, mean (SD)		130.4081 (23.23776)	140.813 (23.0926)	<0.001
5	1=male, 2=female	Male	4597 (46.6%)	318 (66.8%)	<0.001
6		Female	5276 (53.4%)	158 (33.2%)	
7					
8					

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 (acupuncture.dta)

```
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					

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 (e.g. 2 decimal places)

Example

```
logistic heartatk highbp sex
outreg2 using OR_Table, 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

	(1)
VARIABLES	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 heartatk highbp sex
outreg2 using OR_Table, 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 excel

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 heartatk highbp sex
outreg2 using OR_Table, eform replace tex(pretty)
ctitle(odds ratio)
```

Will export a **.tex** file you can open in your LaTeX writing program or on overleaf website

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 and TeX Studio (Both free)
   https://www.latex-project.org/get/
   https://www.texstudio.org

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

Extra slides

`logistic` annotated output

`logistic heartatk highbp sex`

```
Logistic regression                               Number of obs[a]    =     10,349
                                                  LR chi2(2)[b]       =     119.17
                                                  Prob > chi2[c]      =     0.0000
Log likelihood = -1871.0093[a]                    Pseudo R2[d]        =     0.0309

------------------------------------------------------------------------------
    heartatk[f] | Odds Ratio[j]   Std. Err.[k]      z[l]    P>|z|[m]     [95% Conf. Interval][n]
-------------+----------------------------------------------------------------
      highbp |   1.872771[g]   .1794883     6.55    5.9e-11     1.552046    2.259772
         sex |   .4562092[h]   .0455626    -7.86    3.9e-15     .3751048    .5548498
       _cons |   .1082779[i]   .0168897   -14.25    4.4e-46     .0797563     .146999
------------------------------------------------------------------------------
Note: _cons estimates baseline odds.
```

Logistic regression annotated output, cont.

a **Log likelihood** - This is the log likelihood of the model. The value `-1871.0093` 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 heartatk - 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*highbp+ b2*sex

g highbp– Study participants with high blood pressure had 1.7 times higher odds of heartatk compared to study partipants with lower blood pressure, controlling for sex

h sex– Females have .48 times lower odds of heat attack compared to males, controlling for blood pressure.

i constant - This is the expected value of the log-odds of **heart attack** 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.