**Data Management**

Gen [new var] = [formula] eg: gen diff = x1 – x2 gen logvl= log(vl)

egen [new var] = [formula] eg: egen time = sum(var)

replace [var = new value] eg: replace x = 5 if y == 20

recode [old var name] coding, gen [new var name]

eg: recode oldY min/5=1 6/10=2 11/max=3, gen (newY)

reshape long/wide [var], i(id\_var) j(time-var) eg: reshape long dep, i(id) j(week}

hist [cont var] eg: hist Y

sktest [cont var] eg; sktest Y

swilk [cont var] eg: swilk Y

qnorm [cont var] eq: qnorm Y

disp [formula] eg: disp 20\*8

[command…] if [condition] eg: sum xxxx if group == 1

**Descriptive statistics = describe sample stats**

tab [cat var]

sum [cont var] , detail

**Descriptvie study = Parameter estimation = descriptive x-sect survey – to get statistics + SE or 95%CI**

prop [ cat var]

mean [cont var]

means [cont var]

**Analytic study designs = hypothesis testing for all designs**

1. **Comparison between groups**

1 Y (outcome) and 1 X (exposure)

Compare 2 groups : Y = continuous ; X = 2 independent groups

Y = normally distributed

sdtest Y, by(X)

ttest Y, by (X) [unequal]

Y = NOT normally distributed

ranksum Y, by (X)

or

gen newY = log(Y)

sdtest newY, by(X)

ttest newY, by (X) [unequal]

Test changes of Y (pre-post): Y = continuous measured 2 time points ; X = 1 group

Y = normally distributed (paired ttest)

ttest Y time 1 = Y time 2

Y = NOT normally distributed

signrank Y time 1 = Y time 2

or

gen newYtime1 = log(Y time 1)

gen newYtime2 = log(Y time 2)

ttest newY time 1 = newY time 2

Compare >= 2 groups : Y = continuous ; X >= 2 independent groups

Y = normally distributed (ANOVA)

anova Y X

or

oneway Y X , [tab] [bonf, schef, etc.]

mean Y, over(X)

Y = NOT normally distributed

gen newY = log(Y)

oneway newY X, [tab] [bonf, schef, etc.]

Compare >= 2 groups (chi-square test): Y = categorical, X >= 2 groups

Y = categorical

tab [row var] [col var], [row or col] [chi expected exact]

Compare discordant of exposures between matched case-control (McNemar Chi-square)

tab exp\_grp1=exp\_grp2, cell

symmetry exp\_grp1=exp\_grp2

1. **Regression (multivariables analysis)**

Simple regression - 1 Y (outcome) and 1 X (exposure)

Multiple regression - 1 Y (outcome) and >= 2 Xs (exposures)

2.1 **Linear regression: (all designs)**

Simple regression - 1 Y = continuous, normally distributed; 1 X = cont or cat

Multiple regression - 1 Y = continuous, normally distributed; >= 2 Xs = cont or cat

regress Y X

regress Y X1 X2 X3

2.2 **Logistic regression: (all designs)**

Simple regression - 1 Y = cat var; 1 X = cont or cat

Multiple regression - 1 Y = cat var; >= 2 Xs = cont or cat

logistic Y X or logit Y X

logistic Y X1 X2 X3 or logit Y X1 X2 X3

*Note: exponential(beta) = Odds ratio*

**Conditional Logistic**

clogit Y X1 X2 X3, group(grp\_var) or

**Multinomial Logistic**

mlogit Y X1 X2 X3, base(#) rr

**Ordinal Logistic**

ologit Y X1 X2 X3, or

2.3 **poisson regression: (cohort or experimental study designs)**

Simple regression - 1 Y = cat var (incidence) or count; 1 X = cont or cat

Multiple regression - 1 Y = cat var (incidence) or count; >= 2 Xs = cont or cat

poisson Y X , [irr] robust

poisson Y X1 X2 X3, [irr] [exp(time var)]

poisgof

nbreg Y X1 X2 X3, [irr] [exp(time var)]

zip/zinb/ztp/ztnb Y X1 X2 X3, [irr] [exp(time var)]

*Note: exponential(beta) = risk ratio or relative risk*

*Note: Cross-sectional design is OK to use Poisson regression (with robust option) to get*

*Prevalence Ratio (PR) – the output is RR but considered it as PR.*

2.4 **cox’s regression: (cohort or experimental study designs)**

Simple regression - 1 Y = time to event; 1 X = cont or cat

Multiple regression - 1 Y = time to event; >= 2 Xs = cont or cat

stset [ time var], fail [endpoint var]

sts [graph] [list] [, by(group var)]

stsum [, by(group var)]

sts test [group var]

stcox X1 X2 X3

stphtest

*Note: exponential(beta) = hazard ratio which is somewhat like relative risk*

**Parametric survival model:**

streg X1,X2, X3, dist( expo/ weib)

**Multiple failures model – repeated outcomes (= extended cox model)**

1. ***Independent events (AG)***

stset [ time\_ var], fail [endpoint\_var] exit [time .] id [id\_var]

or

stset [ time\_stop\_var], fail [endpoint\_var] time0[time\_start\_var] exit [time .] id [id\_var]

sts [graph] [list] [, by(group\_var)]

stcox X1 X2 X3

1. ***Conditional events (PWP)***

stset [ time\_ var], fail [endpoint\_var] exit [time .] id [id\_var]

or

stset [ time\_stop\_var], fail [endpoint\_var] time0[time\_start\_var] exit [time .] id [id\_var]

sts [graph] [list] [, by(group\_var)]

stcox X1 X2 X3 , strata [event\_number\_var]

stcox [time\_point\_vars – eg., t1 t2 t3] , strata [event\_number\_var]

1. ***Competing risks (WLW)***

stset [ time\_var], fail [endpoint\_var]

stcox X1 X2 X3 , strata [event\_number\_var] cluster[risk\_set\_var]

**Analytic study designs = repeated measures / correlated data / clustered data**

1. **Basic stats**

Use basic stats commands – ttest/F-test/Anova/Chi-square at each time point

1. **Repeated measures ANOVA**

***No intervention (One group)***

anova Y [id\_var] [time\_var], repeat(time\_var)

**Intervention study (>= 2 groups) (nested design)**

anov Y X1/id\_var|X1 time\_var time\_var#X1 /, repeat(time\_var)

X1 (drug)

( drug=1) (drug=2)

Pt 1 2 3 4 5 6 7 8 9 20 11 12

Time 1 x x x x x x x x x x x x

Time 2 x x x x x x x x x x x x

Time 3 x x x x x x x x x x x x

1. **Comparison of slopes**

gen slope=0

statsby slope= \_b[timepoint\_var], by(grp\_var id\_var) clear: regr Y [timepoint\_var]

ttest slope, by(grp\_var)

1. **Generalized estimation equation (extended linear/logistic/poisson regression)**

***Linear regression model***

xtgee Y X1 X2 X3, i(id-var) corr(exchange/indep/unstr/ar) t(timepoint\_var)

xtcorr

***Logistic/poisson regression model***

xtgee Y X1 X2 X3, i(id-var) family(bin/poi) corr(exchange/indep/unstr/ar)

t(timepoint\_var) eform

1. **Generalized/multilevel mixed model(extended linear/logistic/poisson regression)**

***Linear random intercept model***

xtmixed Y X1 X2 X3 || [id-var]:

***Linear random slope model***

xtmixed Y X1 X2 X3 || [id-var]:[X1 X2 for random slope]

***Logistic/poisson random slope model***

xtmelogit Y X1 X2 X3 || [id-var]:[X1 X2 for random slope] , or

xtmepoisson Y X1 X2 X3 || [id-var]:[X1 X2 for random slope] , irr

***Multilevel linear/logistic/poisson random slope model***

xtmixed Y X1 X2 X3 || [id-var\_level 1]: || [id-var\_level 2]:[X3 for random slope]

xtmixed Y X1 X2 X3 || [id-var\_level 1]:[X1 X2 for random slope]

|| [id-var\_level 2]:[X3 for random slope]

xtmelogit Y X1 X2 X3 || [id-var\_level 1]: || [id\_var\_level 2]: [X1 X2 for random slope] , or

xtmepoisson Y X1 X2 X3 || [id-var\_level 1]:[X1 X2 for random slope]

|| [id\_var\_level 2]:, irr

jaranit.kae@mahidol.ac.th