PUBH 8342 Examples

Regression Models

Observed Regression Model

$Yi = B0 + B1Xi + Ei$

$Ei \sim N(0, \sigma^2)$ is the variance in the predicted error in the model, defined as residuals (observed value – expected value)

Expected Regression Model

$E(Yi) = E(B0 + B1Xi) + E(Ei)$

Since $Ei \sim N(0, \sigma^2)$, $E(Ei) = 0$ and $E(Yi) = E(B0 + B1Xi)$

Therefore, Observed = Predicted + Error

Variable Selection Approach

- Draw your DAG
- Fit the crude model
- Fit the fully adjusted model -> Calculate (crude-adjusted/crude)
- Do a backwards cumulative adjustment, only keeping variables that change the point estimate from a fully adjusted model by more than 10% (cumulatively)

Understanding Additive vs Multiplicative Scales

****Logistic in epidemiology is the log base e, the natural logarithm (ln)

Linear Scale

$Pr[Y=1] = B0 + B1X$

Logistic Scale

$Odds (Y=1) = Pr[Y=1]/Pr[Y=0] = P / (1-P)$

$Logit (P) = log (odds) = log (P/1-P)$

$Log (Odds) = B0 + B1X$

$Odds = e^{(B0 + B1X)}$

Switching from odds to probabilities

$Logit (P) = log (odds)$

$Expit function = logistic function = exp(odds)/ 1 + exp (odds) = P$

$E^{(logit(P))} = P = expit (odds)$

Calculating Coefficient Estimates by hand

- Consider Pr(Y) = B0 + B1X + B2X2 + B3X1X2 = 0.1 (when X1 = 0 and X2 = 0)
- Consider Pr(Y) = B0 + B1X + B2X2 + B3X1X2 = 0.2 (when X1 = 0 and X2 = 1)
- Consider Pr(Y) = B0 + B1X + B2X2 + B3X1X2 = 0.3 (when X1 = 1 and X2 = 0)
- Consider Pr(Y) = B0 + B1X + B2X2 + B3X1X2 = 0.4 (when X1 = 1 and X2 = 1)

B0 (When X1 = 0 and X2 = 0) = 0.1

B1 (When X1 = 1 and X2 = 0 minus X1 = 0 and X2 = 0) = 0.3 – 0.1 = 0.2

B2 (When X2 = 1 and X1 = 0 minus X2 = 0 and X1 = 0) = 0.2 – 0.1 = 0.1

B3 (When X2 = 1 and X1 = 1 minus all other B, not P!) = 0.4 – (0.2+0.1+0.1) = 0

Calculating Coefficient Estimates by hand

- Logit (P) = log (odds) = log (P/1-P) = B0 + B1X, therefore Odds = e^(B0 + B1X)
- Consider Logit [Pr(Y)] = B0 + B1X + B2X2 + B3X1X2 ; Pr(Y) = 0.1 (when X1 = 0 and X2 = 0)
- Consider Logit [Pr(Y)] = B0 + B1X + B2X2 + B3X1X2; Pr(Y) = 0.2 (when X1 = 0 and X2 = 1)
- Consider Logit [Pr(Y)] = B0 + B1X + B2X2 + B3X1X2; Pr(Y) =0.3 (when X1 = 1 and X2 = 0)
- Consider Logit [Pr(Y)] = B0 + B1X + B2X2 + B3X1X2; Pr(Y) = 0.4 (when X1 = 1 and X2 = 1)

B0: (When X1 = 0 and X2 = 0) logit P = logit (0.1) = Odds. OR for Bo = e^(logit P) = e^(logit(0.1))

B1: (When X1 = 1 and X2 = 0 minus X1 = 0 and X2 = 0) = logit(0.3) – logit(0.1). OR for B1 = e^ [logit(0.3) – logit(0.1)]

B2 (When X2 = 1 and X1 = 0 minus X2 = 0 and X1 = 0) = logit(0.2) – logit(0.1). OR for B2 = e^ [logit(0.2) – logit(0.1)]

B3 (When X2 = 1 and X1 = 1 minus all other B, not P!) = logit (0.4) – (B2+B1+B0)

Difference between odds and logistic function

- Odds (Y) = Pr(Y=1) / Pr(Y=0) = P / (1-P)
- Logit function = log (odds) = log (P/1-P)
- Logistic function = expit function = exp (Odds) / 1 + exp (Odds) = P

The logit function gives us odds. The logistic function is the inverse of the logit function and gives us probabilities. We use the logit link in logistic regression though

- Log (odds) = Bo + BiX
- Odds = e^(Bo + BiX)

Poisson Regression

- Pr(Y) > Log (Pr(Y)). For our count outcome, E(Y)/N = Count/Total > log(E(Y)/N)
- Since log(a/b) = log(a) – log(b): log(E(Y)/N) = Log (E(Y)) – Log(N) = Bo + BiX
- Therefore we have: Log(E(Y)) = Bo + BiX + Log(N) where Log(N) is our offset term

Non collapsibility of the data

If our stratum specific ORs were both 2.4, we would expect that the overall OR should be the average of both, thus being 2.4. Otherwise, having Crude =/= Stratum specific would mean that we have confounding in our data. However, for Odds Ratio, due to the nature of the math, we often have different odds when we collapse stratum specific values into a crude one, and we have no way of telling if there is confounding or it's just odds ratios being odds ratios. This means that our odds are non-collapsible.

Correlated Data Analysis

Hierarchical data

- Correlation among observations within units
- Predictor variables at different level (Level 1 or level 2)
- Wide format: 1 row = 1 person with correlated observations (repeated measures are new columns of variable)
- Long format: Each person/cluster has multiple rows with a cluster variable to identify the correlation.

Correlation in our data

- Correlation in nested or clustered data (0.01 to 0.1 correlation)
- Correlation in longitudinal data (0.6 to 0.8 correlation)

How to address your correlation in your research

- Cite your assumptions (clustered or longitudinal data)
- Pick the correlation values that you will use in your analysis
- Justify them to avoid comments: explicitly state that there is correlation in the data, that you tested for it, and how it impacted the analysis

How to deal with correlation when there is confounding in the analysis:

- If the outcomes and predictors are level 1 variables, and your correlation creates clusters, you can treat the correlation as a nuisance to keep to a minimum.
- If the outcomes and/or predictors are level 2 variables (same level as the clusters), then the correlation isn't just a nuisance and needs to be dealt with.
- At the most, our models can only handle 3 levels of correlation (repeated observations > within an individual > spatial clusters)

How to model for clusters: in order to ensure sufficient power, make sure that if your outcome is a level 2 variable, you maximize the number of clusters. (Level 1 * Level 2 = N)

- Proc mixed = x time x*time; random intercept time / subject = id / random slope
- GEE vs mixed: in proc mixed, our correlation variable Z is used to fit a trajectory over time between the clusters. In GEE, the model treats each cluster as a repeated cross section of the same sample in order to account for the correlation.

Repeated	Time
1	1
2	1
3	1
4	2
5	2
6	2
7	3
8	3
9	3

In the above dataset, 1, 4 and 7 would be correlated repeated variables at time 1, 2 and 3. GEE will remove the correlation by treating clusters (T1, T2, T3) as repeated cross sections. Since it removes the correlation you can't fit a trajectory between times 1, 2 and 3, since we won't know which data point to link to whom, unless you have a separate clustered ID.

- Fixed effect: predictor variables that has a measured/defined value in our dataset. B describes the relationship between the predictor and the outcome.
- Random effect: people in our dataset do not have an assigned value, but can be assigned one from a distribution extrapolated from calculated values. Residuals in our models are a random effect because we assume that they come from a normal distribution N(0, sigma^2) which can be generalized beyond our sample (CLT).
- Random intercept: there is correlation in the outcome due to the clusters. Random slopes: the effect changes by each cluster, so this is similar to an interaction between the cluster and Beta.
- Degrees of Freedom for correlated data: within cluster variables (Level1): Sum (1 to G) (m-1) – (# of level 1 variables without the intercept) / between cluster variables (Level2): G-(# of between cluster variables + intercept)

There are usually two main concerns when beginning a data analysis project

- First Question: Is there a violation of the normality assumption? (continuous / binary / ordinal or count outcome)
- Second Question: Is there a violation in the independence between observations assumption? (Correlated Data)

Why does correlated data matter? If you compare the same data with independent vs. correlated outcomes, the coefficients (betas) will be similar, but the standard errors / residuals (SE) around these coefficients will be smaller, because the model thinks you have *more* independent observations than you actually have. This leads to narrower confidence intervals (smaller SE) and smaller p values (Larger T or Z statistic). Your coefficients will erroneously be more significant.

Correlated Data = non independence in the outcomes: Correlated Data Analysis is about specifying the correct SE for the correlation in our data (avoid biased SE)

- Can be introduced with longitudinal data (repeated observations over time) -> Examine the zero order correlation for different time points (pearson correlation)
- Can be introduced with nested data (clustered over space or a common setting) -> Use the intraclass correlation to assess the proportion of the variance in the outcome due to the cluster (Between cluster / Within Cluster + Between Cluster)

If you are interested in the individual level effect, you need to determine how much correlation you can have in your data until it severely biases your SE-> Depends on the number of people you have in each cluster

If you are interested in the group level effect, you have to adjust the number of clusters and individuals per clusters in your analysis by using the design (also called variance inflation) effect = 1 + [(m-1)*ICC] where M is the average number of people per group.

Random Effects Models

With nested data, we have individuals grouped into clusters (students in classrooms/patients in clinic/siblings in a family). This creates a correlated data problem because outcomes/residuals from the same clusters are more related than those from different clusters when looking at our outcome parameter. One way to deal with the correlation during parameterization is to use the random intercept approach:

- The residuals in our models represent the difference between the expected value and the observed values in our outcome of interest. For independent data, the residuals are iid: they have ~N(0, σ^2). However, with clustered data, these residuals are not centered around 0 but have different means based on the cluster they're from.
- In the random intercept mode, we fit an overall slope in your model, then use the intercept from each cluster specific equation to create a random intercept parameter which will account from the correlation in each cluster.
- Fixed effects can have variability, but each participant comes in with already predicted values (dependent variables). In contrast, random effects are not predicted values, but rather a distribution of a variable that what assigned after the participants were included in the study. In regular linear models, this usually applies only to the residuals, but in random intercept models, we let the intercepts from each cluster become a random variable too, so that on average they take on a distribution with mean 0. This will allow us to generalize the parameter to a whole population, since we use a distribution rather than predicted values.

Random slopes models

In this example, we are looking at one continuous outcome with time invariant predictors.

- Put the data in long format (where each repeated observation has its own row, and there is a separate variable indicating the cluster).
- Look at the outcome, then look at the predictors by running descriptive statistics (means and tabulations)
- Lastly, look at the clustering (correlation coefficients ant graphs).

Calculate the intra class correlation by specifying a mixed model with only the outcome and clustering variable in the outcome. Under the random effects output, you will find the variance due to the residuals as expected (the within group variance), as well as the variance due to the clustering variable (the between group variance). You can use both to calculate the Intra Class Correlation (ICC).

Not only do we care about nesting because we need to account for the correlation due to cluster in order to properly estimate the standard error for our test statistics (p-value and confidence interval), but we may have a level 2 dependent variables that we need to include when parameterizing our level 1 predictor. However, we cannot use the n-1 degree of freedom to estimate their effect (assuming independence) so we have to use the impact on the between group variance from the random effect by evaluating then using g-1 degrees of freedom. In practice, this would be done by specifying the level 2 variable after the cluster variable in the mixed model.

We collect longitudinal data on individuals because we are interested in the change over time in our outcome. Rather than doing multiple pre/post evaluations between time points to evaluate the change in our outcome over time, the random slopes model allows us to explicitly model this change.

- The "philosophical" advantage of the random slopes model is that by allowing the slope to vary in each cluster instead of using fixed effects, we can generalize our findings to the whole population from which these clusters are sampled from.
- The "mathematical" advantage of the random slopes model is that since the clusters are not in the model as a fixed effect, rather than spending m-1 degrees of freedom (where m is the number of clusters) for our model, we only need to spend 1 degree of freedom using a random variable.
- Not only are we able to properly account for the correlation in the data with the random effects model when looking at the change in our outcome over time, but we can also look at differences in our outcome between clusters.

Survival Analysis

There are four methods to deals with incidence density data:

- Tabular Methods
- Life Tables
- Poisson Regression
- Cox Regression

Survival analysis is a collective term for statistical methods that are used to study binary outcomes such as dead and alive while taking into account the time when the event of interest occurs. We chose not to use logistic regression for this type of analysis for two main reasons: logistic regression doesn't take into account varying person-time at risk, and the measure of association from logistic regression is an odds ration, not a relative risk.

There are 3 types of time at risk to consider:

- Observation start – Event observed
- Observation start – lost to follow up
- Observation start – end of the study

Time to events have two components

- Dichotomous Outcome (Yes/No event) -> We could not use logistic regression because we miss time to event
- Continuous Outcome (Time to event) -> We could not use linear regression because we would lose the dichotomous outcome.

Lifetables

- N is the number of people at the start of period t, d is the number of people who died during the time interval, and t is the person-time accrued over the interval t.
- Conditional rate: Death / N @ the beginning of the time interval
- Conditional Survival ($P[T>t|T\geq t]$): 1- Conditional Rate
- Cumulative survival : Product of all conditional rates from previous time intervals.

Cox regression

Rate is defined as #events/person-time at risk. The denominator can be conceptualized as a single person who is observed (at risk) for some period of time where time is measured in some specified unit. Now assume that we are interested in the instantaneous rate of the disease at a time point t: it is often referred to as the hazard rate, and its related measure of association is called the hazard rate. Cox regression is modeling the log(rate) but does not assume a constant rate over time, and allows for a continuously changing hazard function. Cox regression can be considered a semi-parametric method (parametric methods assume that there is some distribution from which the calculations are derived)

The reference hazard is a function of time that could change continuously (ie over the follow up time)

- h(t) = [reference hazard]* e ^[B1 * X1]
- h(t) = ho(t) * e ^[B1 * X1]
- log [h(t)] = log [ho(t)] + [B1 * X1]
- log [h(t) Person a/h(t) Person b] = [B1 * X1]
- B1 = Predicted Difference in log(hazard) per one unit increment in X.

Since log [h(t) Person a/h(t) Person b] = [B1 * X1], this is equivalent to =log [h(t) Person a] – log [h(t) Person b] = [B1 * X1] which is a constant, the two time functions should be parallel over time.

SAS Code

Proc PHREG;

Model follow-uptime*outcome(censored value) = exposure;

Run;

```
/*Example*/

PROC PHREG data = framinghammkI;

MODEL STRKYEARS*STROKE(0) = AGE / RL TIES=EFRON;

RUN;


PROC PHREG data = framinghammkI;

MODEL STRKYEARS*STROKE(0) = AGE / RL TIES=BRESLOW;

RUN;


PROC PHREG data = framinghammkI;

MODEL STRKYEARS*STROKE(0) = AGE / RL TIES=EXACT;

RUN;
```

Review of the output

- Dependent Variable: SAS will label the follow-up time as the dependent variable but that is not correct (this is actually the time at risk). Remember that the dependent variable in cox regression is log(rate of disease) so just like in logistic regression, the real dependent variable is incident disease (aka the censoring variable)
- Ties handling:means that two or more observations have the same event time (follow up time) the default is ties = breslow. The ties = efron option is often recommended

- Testing Global Null Hypothesis: similar to proc logistic, score test here is the same as the log rank test
- The maximum likelihood estimates does not have an intercept term (characteristic of Cox regression)
- 1 – the percent censored = failure rate

Testing the proportional hazards assumption

We can test the assumption by modeling an interaction between exposure variable and time at risk

log [h(t)] = log [ho(t)] + [B1 * X1] + B2*t*X1

- if B2 is not statistically significant, we do not have enough evidence against the proportional hazards ratio
- if B2 is statistically significant we have statistical evidence that the proportional hazards ratio assumption does not hold
- if B2 is positive, the hazard ratio increases over time
- if B2 is negative, the hazard ratio decrease over time

SAS code

Proc PHREG;

Model follow-uptime*outcome(censored value) = exposure;

Phazard = exposure*[log(follow-uptime)]

Run;

Example

```
PROC PHREG data = framinghammkI;

MODEL STRKYEARS*STROKE(0) = AGE PHAZARD / RL;

PHAZARD = AGE*(LOG(STRKYEARS));

TITLE1 'PHAZARD IS THE TEST OF THE PROPORTIONAL HAZARDS ASSUMPTION';

RUN;
```

The Interpretation of phazard is similar to that of test for homogeneity of the hazard ratio over time B2

How to stratify the cohort follow up time if the proportional hazard assumption is violated

Divide the follow-up time at the midpoint (0-12, >12-24 years) with 0-12 years as the reference group, then create variables reflecting the interaction of age and time interval. As with the interaction term, above, you will need to create the indicator variables for follow-up time within the PHREG procedure itself. Please note that this method of dividing follow-up time into discrete windows will not work if variables are created within the DATA step

Example

```
PROC PHREG;
MODEL STRKYEARs*STROKE(0) = AGE PH12_24;
IF 0 LT STRKYEARS LE 12 THEN TIME1=1; ELSE TIME1=0;
IF STRKYEARS GT 12 THEN TIME2=1; ELSE TIME2=0;
PH0_12 = AGE*TIME1;
PH12_24 = AGE*TIME2;
RUN;
```

```
PROC PHREG;
MODEL STRKYEARS*STROKE(0) = AGE PH9_20 PH20_24;
IF 0 LT STRKYEARS LE 9 THEN TIME1=1; ELSE TIME1=0;
IF 9 LT STRKYEARS LE 20 THEN TIME2=1; ELSE TIME2=0;
IF STRKYEARS GT 20 THEN TIME3=1; ELSE TIME3=0;
PH0_9 = AGE*TIME1;
PH9_20 = AGE*TIME2;
PH20_24 = AGE*TIME3;
RUN;
```

How to stratify follow-up time using Cox regression to create different follow up times for subcohorts

1. Cohort that has less or equal than 10 years of follow-up:

```
*End of follow at 10 years;
End10d=basedate+10;

*Censoring those with fu >10 years at 10 years;
if cancerdate gt End10d or cancerdate= .  then cancer1=0;
else if . <cancerdate le End10d   then cancer1=1;

*Creating new person years;
if tpyrs le 10 then tpyrs1=tpyrs;
else if tpyrs gt 10 then tpyrs1=10;
```

2. Cohort that has more than 10 years of follow-up:

```
 if  . lt cancerdate le End10d  then delete;

if tpyrs le 10 then delete;
else if tpyrs gt 10 then tpyrs2=tpyrs-10;
```

The downside of this approach is that those with person- years less than 10 in the second cohort are deleted. However, we could code as tpyrs2=0 if they were followed for less than 10.

How to look at attenuations in an association using Cox proportional hazards

When we have divided person time in the past, we usually take the entire person-time from that study's baseline (or last assessment if updated) to the end of follow-up and divide follow-up time at the median.

In this example:
censortime = follow-up time from baseline for study to 12/31/12
mediantime = median follow-up time for the analytic cohort
censortime_new = new follow-up time variable for the =<median time strata
cancer = case status, where 1 is a case

For the =<median follow-up time strata, we only change the censor date for ppts with follow-up greater than the median time.
If censortime > mediantime, then censortime_new=mediantime;
Else censortime_new=censortime.

For the > median follow-up strata, we only change cases that occur before median to non-cases.
If censortime =< mediantime and cancer=1, then cancer=0.

For the first strata, we are just looking at cases that occur in the near-term. We don't have to change anyone's cancer status, because later cases are automatically censored before they occur. In the second strata, we are just looking at cases that happen later in follow-up. Cases that occur in the near-term are still censored when they occur, but they are no longer considered a case. In both strata, everyone is included. Then we can see whether our exposure has the same association with cases that occur closer in time to its measurement and with cases that occur further in the future.

Data analysis for survival analysis and cox regression in STATA

- Define your person time variable and failure variable: stset (timevar), failure(failurevar)
- Look at the Kaplan Meier Curves from the lifetable: sts graph
- Run the crude cox model: sts cox treatmentvar. The cox model has an innate proportional hazard assumption for H(t) = Ho(t)*exp(B*t). The baseline hazard ho(t) is model but no value is directly calculated

Check the equality of the survival curves: Visual inspection / log rank test ~ chi square test / Include Interactions with time

- Visual Inspection of the KM curves: Plot the ln(-ln) of the exposed vs the ln(-ln) curves of the unexposed. We want the curves to be parallel to support the proportional hazard assumption. Stphplot by (treatmentvar)
- Goodness of fit test: log rank test ~ chi square test of the expected vs observed survival times in both conditions. Estat phtest is the STATA command for proportionality assumption test
- Model the treatment as time varying coefficient: sts cox treatmentvar, tvc (treatment var). the tvc command is the equivalent of x##y interaction in STATA, and the texp(_t) is the link function indicating how time is modeled (linear, log, ect….). If the interaction term is significant, then the proportional hazard assumption is violated.

Interactions with time

$$HR(t) = \frac{h(t, x_1 = 1)}{h(t, x_1 = 0)}$$
$$= \frac{h_0(t)\exp(\beta_1 + \theta \times t)}{h_0(t)}$$
$$= \exp(\beta_1 + \theta \times t)$$

Stratification on covariates that violate the proportional hazard assumption with time

- An easy approach to deal with non-proportional hazards for confounders
- Extends the normal Cox model by allowing a different baseline hazard for each stratum
- S=1 does not need to be proportion to S=0 anymore, but we cannot estimate a HR for S
- However, when the proportional hazard assumption is violated for your main exposure or you have a time varying exposure, you have to model time varying effects.

Modeling time varying effects

Categorizing follow-up times

The data may suggest the HR is constant within time segments

For instance: HR(0<t<8) and HR(8<=t<35)

For this, we need to “split” the data, either using (stsplit) manually, or letting stata do it

STATA analysis steps

stset stime, f(died) id(id)

stcox treatment

failure _d: failure

analysis time _t: time

stsplit post,at(0) after(wait)

*recode the new variable "post" so 1 is post-transplant

recode post -1=0 0=1

*Wait is coded as 0 if the person never got a tranplant. this fixes it

replace post=0 if wait==0

(34 real changes made)

Summary for adjusting the model for confounders and multiple time points

- Model the confounders in the stcox model, then model them as time varying covariates. If the interaction with time is significant, you can stratify the analysis on the confounder levels.
- Use the lincom command in STATA to get interpretable interaction values between the main effect and its time varying coefficient: it should be (e^B1* e^B3)^time, but since lincom uses the log scale, you have to input lincom (B1 + B3) * time, eform > exponentiates the final results.
- If you choose to use multiple time points, texp(time point cut), STATA will create time point cut as an indicator variable (0/1) so it is easier to lincom the interaction values
- Time varying exposures can also happen in your data, as someone can become exposed over the followup time. You need to account for that person going from unexposed to exposed by splitting the observation so that they are censored at the end of their unexposed time, then have new follow up start until they experience the event. Use stsplit newvarname, at (0) after (follow-up change var) where 0 is the value of the new starting follow up time, and follow-up change var is the variable that indicates when during follow up the person went from unexposed to exposed. You will have to recode the -1 to 0 and 0 to 1 for the newvarname, as well as replace newvarname == 0 if followupvarchange ==0 to indicate that a person never changed exposure status. Then you can repeat the cox analysis as usual, unsing the new var name as your treatmentvar.

SUMMARY OF COX REGRESSION MODELS & PROPORTIONAL HAZARDS ASSUMPTIONS IN SAS

INTRODUCTION – COX MODEL DEFINITION

The first semi-parametric model was proposed **by Cox (1972)** who assumed that the **covariates-related component is distributed exponentially**

The covariates-related component is expressed as exp(β**x**), thus the model has the following formula:

$$\lambda(t, \boldsymbol{x}) = \lambda_0(t)\exp(\beta \boldsymbol{x})$$

Where:

- $\lambda(t, \boldsymbol{x})$ – hazard function that depends on timepoint t and vector of covariates **x**
- $\lambda_0(t)$ – baseline hazard function that depends on time only
- $\exp(\beta \boldsymbol{x})$ – covariates-related component

PROPORTIONAL HAZARD ASSUMPTION

Comparing hazard between two subjects at time t via **HAZARD RATIO:**

Subject (1) – covariates: **x** = $\mathbf{x}_1$
Subject (2) – covariates: **x** = $\mathbf{x}_2$

$$HR = \frac{\lambda(t, \boldsymbol{x_1})}{\lambda(t, \boldsymbol{x_2})} = \frac{\lambda_0(t)\exp(\beta \boldsymbol{x_1})}{\lambda_0(t)\exp(\beta \boldsymbol{x_2})} = \frac{\exp(\beta \boldsymbol{x_1})}{\exp(\beta \boldsymbol{x_2})} = \exp\left[\beta(\boldsymbol{x_1} - \boldsymbol{x_2})\right]$$

HR (hazard ratio) – proportion of hazard function value for two subjects wit different values of covariate(s) at the given timepoint t

HR does not depend on time (on covariates only)
λ_0(t) – baseline hazard function does not have defined mathematical formula

PROPORTIONAL HAZARD ASSUMPTION

Proportional hazard assumption – discussion

Violation does not cause serious problems as in such cases parameter estimate can be interpreted as 'average effect' of the covariate (e.g. Allison, 1995)

Violation should be taken into account and appropriate modification of the model should be used to enable more precise interpretation (e.g. Hosmer, Lemeshow, 1999)

example of study site in clinical trial for which it is very likely that the assumption will be violated

SAMPLE DATASET

Data for 60 patients from open-label clinical trial on safety of newly invented therapy for brain cancer

AGE = Age at screening

SITE = Number of study site (SITE = 1 stands for Site B, SITE = 2 stands for Site A)

TIME = Time (in days) from the beginning of therapy till death (if patient dies) or till the end of the observational period (if patient survives)

CENSOR = Indicator of the event (CENSOR = 1 stands for death of patient, CENSOR = 0 stands for survival till the end of the observational period)

VERIFICATION OF PH ASSUMPTION

Proprotional hazard assumption –
methods of verification

- plot of 'log-negative-log' of the Kaplan-Meier estimator of survival function:
  curves on the plot should be pararell with distance that is constant over time

- plot of Schoenfeld residuals as a function of time:
  residuals should not show any trend

- adding interaction of a covariate with function of time variable:
  newly added variable should not be statistically significant

VERIFICATION OF PH ASSUMPTION

Proportional hazard assumption for the Cox model estimated for 60 subjects from the open-label study:

Step 1
Model estimation – time to death is being analyzed, AGE and SITE included as covariates

```
/*Initial Cox model estimation - site and age as covariates*/
proc phreg data = a.site;
      format site site.;
      model time*censor(0) =  age site  / ties = exact;
      output out = schoen ressch = age_s site_s ;
run;
```

Analysis of Maximum Likelihood Estimates

Parameter	DF	Parameter Estimate	Standard Error	Chi-Square	Pr > ChiSq	Hazard Ratio	95% Hazard Ratio Confidence Limits	
AGE	1	0.20690	0.07405	7.8069	0.0052	1.230	1.064	1.422
SITE	1	-0.74290	0.38926	3.6423	0.0563	0.476	0.222	1.020

Both variables statistically significant, confidence interval for hazard ratio for AGE does not include 1, for SITE however confidence interval includes 1, which means that there might be no difference between two study sites in terms of risk of dying

VERIFICATION OF PH ASSUMPTION

Step 2

Verification of proportional hazard assumption for AGE – cont.

- Adding interaction of AGE by TIME to the model

```
/*Cox model estimation - age and site as covariates, with interaction of site
and time added*/
proc phreg data = a.site;
      format site site.;
      model time*censor(0) = age site age _t/ ties = exact;
      age _t = age *time;
run;
```

Analysis of Maximum Likelihood Estimates

Parameter	DF	Parameter Estimate	Standard Error	Chi-Square	Pr > ChiSq	Hazard Ratio
AGE	1	0.01692	0.38357	0.0019	0.9648	1.017
SITE	1	-0.70788	0.39323	3.2406	0.0718	0.493
age_t	1	0.00149	0.00296	0.2516	0.6160	1.001

Newly added variable is not statistically significant which indicates that proportional hazard assumption is satisfied for AGE

VERIFICATION OF PH ASSUMPTION

Step 3

Verification of proportional hazard assumption for SITE – cont.

- Adding interaction of SITE by TIME to the model

Analysis of Maximum Likelihood Estimates

Parameter	DF	Parameter Estimate	Standard Error	Chi-Square	Pr > ChiSq	Hazard Ratio
AGE	1	0.23313	0.07399	9.9268	0.0016	1.263
SITE	1	-5.90164	2.80362	4.4311	0.0353	0.003
site_t	1	0.03985	0.02123	3.5233	0.0605	1.041

Interaction of SITE by TIME is significant at the level of 0.1 which may lead to the conclusion that proportional hazard assumption is likely to be violated for SITE

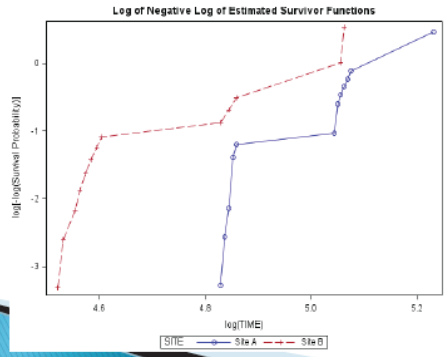
VERIFICATION OF PH ASSUMPTION

Step 3

Verification of proportional hazard assumption for SITE – cont.

➢ Plot of 'log-negative-log' of survival function

```
/*Lifetables estimation according to Kaplan-Meier formula;
generation of the plot of 'log-negative-log' of survival function
separately for each site -> STRATA statement*/
proc lifetest data = a.site plots = (s, lls);
      format site site.;
      strata site;
      time time*censor(0);
run;
```



Two lines corresponding to log[-log[S(t)]] are not distributed parallelly, the distance is changing over time which suggests violation from PH assumption for SITE

VERIFICATION OF PH ASSUMPTION

Conclusions:

➢ proportional hazard assumption satisfied for AGE =>
impact of AGE on risk of event experience is constant over time
➢ proportional hazard assumption not satisfied for SITE =>
impact of SITE on risk of event experience is not constant over time

Modification of the model for non-proportional hazard purpose

Adding interaction of covariate(s) with function of time

Stratification model

INTERACTIONS WITH FUNCTION OF TIME

The idea:

- add interaction of a covariate for which proportional hazard assumption is violated with time variable (or some function of time)
- if the interaction is statistically significant -> the effect of the given covariate is not constant over time
- including interaction in the model enables to interpret parameter estimate taking this fact into account
- interaction with time: both method of PH assumption verification and solution to the problem of its violation

INTERACTIONS WITH FUNCTION OF TIME

Initial model:

Analysis of Maximum Likelihood Estimates

Parameter	DF	Parameter Estimate	Standard Error	Chi-Square	Pr > ChiSq	Hazard Ratio
AGE	1	0.20690	0.07405	7.8069	0.0052	1.230
SITE	1	-0.74290	0.38926	3.6423	0.0563	0.476

vs model with SITE by TIME interaction:

Analysis of Maximum Likelihood Estimates

Parameter	DF	Parameter Estimate	Standard Error	Chi-Square	Pr > ChiSq	Hazard Ratio
AGE	1	0.23313	0.07399	9.9268	0.0016	1.263
SITE	1	-5.90164	2.80362	4.4311	0.0353	0.003
site_t	1	0.03985	0.02123	3.5233	0.0605	1.041

INTERACTIONS WITH FUNCTION OF TIME

Difference in interpretation:

Initial model:
HR = 0.476 => Subjects from Site A (SITE = 2) are approximately 100*(1-0.476)% = 52.4% less likely to die than subjects treated in Site B (SITE = 1)

Model with SITE by TIME interaction:
HR between subjects from Site A and Site B depends on time as follows:

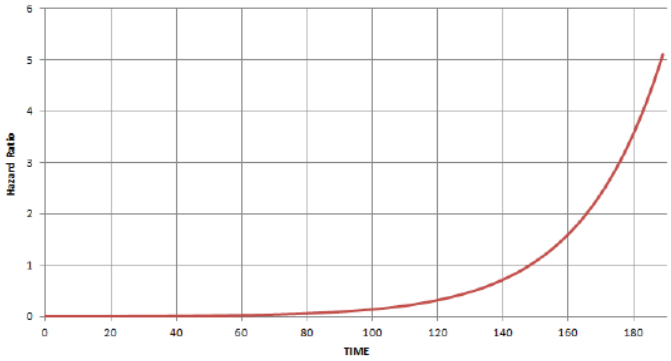
$$HR = \frac{\lambda(t, age = a, site = 2)}{\lambda(t, age = a, site = 1)} = \frac{\lambda_0(t)\exp\left[\beta_1 a + 2\beta_2 + 2\beta_3 t\right]}{\lambda_0(t)\exp\left[\beta_1 a + \beta_2 + \beta_3 t\right]} = \exp[\beta_2 + \beta_3 t]$$

where:
β_1 – parameter estimate for age
β_2 – parameter estimate for site
β_3 – parameter estimate for interaction of site and time

INTERACTIONS WITH FUNCTION OF TIME

PAREXEL. Right where you need us™ CELEBRATING 30 YEARS

Hazard ratio as function of time:



- *for relatively low values of time: subjects from Site B are much more likely to die than subjects from Site A (HR very low),*
- *HR increases over time reaching value of 1 on 148th day which means that chances of dying on 148th day are approximately equal for subjects from both sites,*
- *after 148th day HR exceeds value of 1 which means that subjects from Site A are more likely to die than subjects from Site B (even 3 times – after 160 days)*

STRATIFIED MODEL

The idea:

- split the whole sample into subgroups on the basis of categorical variable (here: stratification variable) and estimate the model, letting the baseline hazard function differ between subsamples

- stratification variable should be chosen so that it interacted with time (i.e. PH assumption is violated for this variable) and is not of primary interest as stratification of the model automatically excludes the stratification variable from set of explanatory variables

- coefficient estimates: equal across strata for all explanatory variables

STRATIFIED MODEL

Estimation:

Model formula for stratum s:

$$\lambda_s(t, \boldsymbol{x}) = \lambda_{s0}(t) \exp(\beta \boldsymbol{x})$$

where s = 1, 2, …, S – number of stratum

Partial likelihood function = product of partial likelihood functions for each stratum
For details, please refer to Hosmer, Lemeshow 1999

Baseline survival function and covariates-adjusted survival function estimates might be obtained for stratification model (e.g. in SAS, BASELINE statement in PHREG procedure performs appropriate calculations)

STRATIFIED MODEL

Initial model:

Analysis of Maximum Likelihood Estimates

Parameter	DF	Parameter Estimate	Standard Error	Chi-Square	Pr > ChiSq	Hazard Ratio
AGE	1	0.20690	0.07405	7.8069	0.0052	1.230
SITE	1	-0.74290	0.38926	3.6423	0.0563	0.476

vs stratified model:

```
/*Cox model estimation - AGE as a covariate, SITE as stratification variable;
saving estimates of survival and cumulative hazard functions in BASE dataset*/
proc phreg data = a.site;
      baseline out = base  survival = surv cumhaz = cumhaz;
      format site site.;
      strata site;
      model time*censor(0) = age / ties = exact;
run;
```

Summary of the Number of Event and Censored Values

Stratum	SITE	Total	Event	Censored	Percent Censored
1	Site A	30	17	13	43.33
2	Site B	30	14	16	53.33
Total		60	31	29	48.33

Analysis of Maximum Likelihood Estimates

Parameter	DF	Parameter Estimate	Standard Error	Chi-Square	Pr > ChiSq	Hazard Ratio
AGE	1	0.20959	0.07534	7.7389	0.0054	1.233

CONCLUSIONS

In general:

Accounting for the fact that proportional hazard assumption is violated provides more detailed results as compared with initial model

Interaction with time:

- enables to analyze how HR changes over time
- provides parameter estimates for variable for which PH assumption is violated
- might require more computational resources than stratified model (Allison, 1995)

Stratified model:

- requires less computational resources
- enables to obtain baseline and covariates-adjusted survival function estimate for each stratum
  - does not provide parameter estimate for stratification variable

Statistical Models

Statistical Models, Briefly

What is a Model?

- A formal way of describing how some variable(s) behave
- By "formal" I mean that a model uses mathematics/statistics
- A model makes assumptions that limit the possible states of nature we will consider

What Goes in a Model?

- What model do we choose? What assumptions do we need to make some sort of inference?
- The Model needs to account for our positivity, consistency and exchangeability assumptions that are needed to go from association to causation. These are the fundamental assumptions that should not be violated
- Furthermore, there are modeling assumptions that we need to consider: for example, exchangeability is only realistic in large RCT's so we would have to rely on conditional exchangeability and positivity within strata to satisfy our modeling assumptions

Exchangeability

- That is, the probability of observing the outcome does not depend on whether it was the X=0 or X=1 group that was assigned the treatment
- The probability of mesothelioma that was observed among those employed 10 years is that which we would have observed had the non-workers (counter to the fact) worked for 10 years
- This is what epidemiologists have tried to express in decades of literature on 'no confounding' assumptions

Positivity

- Every exposure level is possible
- We need this in order to have a well defined causal effect
- How can I estimate the effect of 10 years of taconite exposure if no one could possibly be exposed?

Is Exchangeability Plausible?

- In large RCTs, it seems very plausible!
- In observational epidemiology, less so
- Instead, we hope that if we condition on a number of covariates we have exchangeability within those strata
- Perhaps within strata of age, gender, and education the two exposure groups are (more) exchangeable

Conditional Exchangeability

- We can account for these Z's using traditional methods like Mantel Haenzel or standardization
- But what if there are a large number of Z or they are continuous?

More Assumptions!

- Exchangeability and Positivity are the basic assumptions we always need to make
- But to identify causal effects from observational data, we will often need to add more assumptions
- In this class, those additional assumptions will be regression ***models***
- Regression ***models*** are powerful and their assumptions can/should be laid out clearly and carefully
- Much of this class will involve explaining/evaluating these assumptions

Regression Model

The 'truth' is too complicated and we won't usually have enough data to estimate the function non parametrically

Instead, we specify a much simpler regression model to approximate the truth

The form of the regression model (linear, logistic, log-linear, etc) is a question of what simplification seems like an adequate approximation of the regression truth

Regression Assumptions

Many regression assumptions are extensions of the basic assumptions:

exchangeability —> conditional exchangeability

positivity —> positivity in each stratum

Its often easier to assess the additional assumptions of a regression model specification

dose a plot of the data look more like a linear relationship or a threshold, for example?

is there heteroskedasticity?

Progression through Regression

Any good analysis starts with careful consideration of the study and its design

DAGs!!!

Careful uni/bivariate analyses (and DAGs!) help to assess the plausibility of the 'basic' assumptions of exchangeability and positivity.

On top of all these basic checks, we need to pay careful attention to the assumptions of our regression model

Linear Regression
Logistic Regression
Stata

Linear Regression Model

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 z_i + \varepsilon_i$$

- i→ individual observation
- B_0 → Slope (effect at baseline / in population)
- B_1 → Change in yi due to a 1 unit change in x
- ε_i is the predicted error in the model.
- We often assume that ε_i is normally distributed
- ε_i ~ N (0, σ^2) σ is the variance

Linear Regression Assumptions

- What are they? How do you check for them?

1. Linearity: the relationship between X and Y is linear (check with a scatter plot)
2. Independent observations: the Yi's conditional on X and Z are independent (check for correlation and confounding)
3. Normally Distributed: The residuals around the best fitted line should be normally distributed (check using a QQ plot)
4. Equal variance and Homoscedasticity (Equal variance in the ***residuals***)

How Important are these assumptions?

- Linearity: Very important
- Independence: Important, but can be addressed by increasing the sample size for correlation, or adjusting for confounding
- Normality: Not too important as we can use the Central Limit Theorem to guarantee a normal distribution of the outcomes, regardless of the distribution of the predictors with a large enough samples
- Homoscedasticity: Important, otherwise we need to adjust for it in the model

Logistic Regression

$$\text{logit}(\Pr[Y_i = 1]) = \beta_0 + \beta_1 x_i + \beta_2 x_i$$

- What is logit( )? The Logit function of a number p between 0 and 1 is defined by logit (p) = log (p/1-p)
- What is expit( )? The Expit function of any number & is also known as the logistic function and is defined by expit (&) = exp(&)/exp(&) + 1
- If p is a probability, then p/1-p is the corresponding odds, and the logit is the logarithm of the odds
- Where's the distributional assumption? It is the binomial distribution for 1 trial, or the Bernoulli distribution for multiple trials
- Where's the error term? There is no error tem, because we are predicting an outcome based on our observed data, rather than fitting the best line to the observed data.

Logistic Regression Assumptions

- Independence
- Model fit
  - All the necessary variable are included
  - Have the right form (interactions, non-linearities)
- No collinearity

$$\text{logit}(\Pr[Y_i = 1]) = \beta_0 + \beta_1 x_i + \beta_2 x_i$$

Estimating Coefficients using Maximum Likelihood

- How do we estimate $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2$?
- Likelihood: the 'probability' of the data we saw given a set of parameter values
  - Only…the our data is fixed. So the likelihood looks at the probability of the data at different parameter values.
  - The maximum likelihood estimates are the parameter values that return the highest likelihood

Maximum Likelihood Example

- Flip a coin 10 times, get 7 heads

$$L = \binom{10}{7} p^7 (1-p)^3$$

- Our job is to find the value of p that makes L as big as possible
  - What value of p would make the observed data most likely

Maximum Likelihood

- Regression models are a slightly more complicated version of the same problem
- The model we specify implies a likelihood.

$$\prod_{i=1}^{N} \left(\frac{e^{\beta_0+\beta_1 x_i+\beta_2 z_i}}{1+e^{\beta_0+\beta_1 x_i+\beta_2 z_i}} \right)^{y_i} \left(1 - \frac{e^{\beta_0+\beta_1 x_i+\beta_2 z_i}}{1+e^{\beta_0+\beta_1 x_i+\beta_2 z_i}} \right)^{1-y_i}$$

- Software packages have very clever algorithms to find the maximum value of the parameters
- The interpretation is the same:
  - These are the parameters that maximize the likelihood of the observed data

Variable Selection in Epidemiology

►Richard MacLehose, PhD

Error Terms in Regression

For linear regression: Yi = Bo + B1xi + Ei -> This Error term is normally Distributed with N(0, σ^2)

E(Yi) = E(Bo + B1xi + Ei) but since the mean of Ei is 0, we can just write

E(Yi) = Bo + B1xi

In other words , Observed = Predicted + Error

Confounding

In this case there are a number of variables that are possible confounders

Failure to control for these makes the assumption of exchangeability tenuous

- adjusting for them does not make the assumption correct. It makes the assumption more plausible

There are a slew of possible confounders. We'll consider 5:

- Race (white, black, hispanic, Asian, other); parent education (HS, some college, college, advanced); age (years); depression score (ces-d); employment status (full time, part time, stay at home, unemployed);

How do we decide if these are confounders?

Confounder Selection

Consider the regression model:

$$E(Y) = \beta_0 + \beta_1 x + \beta_2 black + \beta_3 asian + \beta_4 hispanic + \beta_5 other + \beta_6 age + \beta_7 hscollege + \beta_8 college + \beta_9 advanced + \beta_{10} unemployed + \beta_{11} parttime$$

It is traditional in epidemiology to consider whether each possible confounder need be in the model

- Does age need to be in the model? Could we leave it out?

Essentially a bias - variance tradeoff

- Including additional variables in the model will typically increase the standard error of the coefficients leading to wider confidence intervals (CI)
- If a variable isn't a confounder, we could exclude it without worrying about bias AND doing so would narrow our CIs.
- In the extreme, including all the variables might cause the model to become so unstable that it can't be fit or the CIs to be laughably wide.

Stepwise Confounder Selection

The statistics world has developed a huge number of methods for ***variable selection***

The simplest of these, stepwise selection, is seldom appropriated for use in epidemiology

- it is still commonly taught in biostatistics courses, which is why i feel obliged to cover it here

There are two varieties of stepwise selection: forward and backward

Backwards Stepwise Algorithm

1. Fit the full model
2. Choose the least significant variable
3. If its p-value is > some arbitrary cut point (often 0.10 or 0.20), exclude that variable from the model
4. Fit model excluding this variable
5. Repeat 2-4 until no variables with p-val > cut point are found

Stepwise Algorithms

These algorithms perform very poorly in estimating causal effects

Why?

1. Confounding is not something that can be assessed by p-values. A variable with a low p-value can be a very unimportant confounder. A variable with a large p-value can be a very important confounder
2. We are trying to answer the question of how best to estimate the causal contrast. Stepwise is answering a different question: how can I build a parsimonious model that predicts the outcome as well as possible
   - importantly, stepwise isn't even very good at answering the question it proposes to be answering (because p-values also don't imply much about predictive ability)

Change in Estimate Criteria for model Selection

This is probably the most common variable selection technique in epidemiology

As with stepwise, there is a forward and backward version

In fact, it is exactly like stepwise regression except that it replaces the inclusion/exclusion criteria

- exclude if p-val > cutoff —> exclude if effect of interest changes by <10%

Backward Change in estimate algorithm

1. Fit full model and retain effect estimate 1
2. Choose a (group of) variable(s) to exclude
3. Refit model without this (these) variable(s) and retain the effect estimate 2
4. If the two estimates differ by 10% or less, exclude the variable(s) from the final model
5. Repeat steps 2-4 until no variable is left that changes the estimates by <10%

Forward Change in Estimate Algorithm

1. Fit crude model and retain the effect estimate 1
2. Choose a (group of) variable(s) to include
3. Refit model with this (these) variable(s) and retain the effect estimate 2
4. If the estimates differ by 10% or more, retain the variable(s) from the final model
5. Repeat steps 2-4 until no variable is left that changes the estimates differ by >10%

Change in Estimate Algorithm Problems, I

Problem: lots and lots of ambiguity

Why 10%? This is the amount of bias we're willing to accept. Who chose that and why?

Forward and backward may get different results

Change in Estimate Algorithm Problems, II

Problem: more serious ambiguity

Is a change from RD=0 to RD=0.1 the same as RD=0.4 to RD=0.5?

Is a change from RR=.8 to RR=1.0 the same as RR=2.2 to RR=2.4?

Relative change? On the log scale?

Change in Estimate Algorithm- Summary

The change in estimate approach isn't *bad*. It's certainly an improvement over stepwise

However, I think we can do better as a discipline

While I do occasionally use the change in estimate algorithm, it is never the first thing I turn to

I'll demonstrate change in estimate algorithm a little later in the lecture

A Broader View of Selection

Before we embark on the change-of-estimate approach, we're already doing variable selection

How did we decide which variables (potential confounders) to include in the full model?

Hopefully, we thought carefully about how our exposure and outcome are causally related

Last semester, you learned about DAGs

Drawing a DAG is almost always the first thing I do when starting a new analysis. I draw a DAG before I fire up Stata

DAG Variable Selection

DAGs give us a formal foundation to decide what should be under consideration as a confounder

- Race, education, age, U

And what should not be considered as a potential confounder

- fast food consumption

Further, DAGs tell us how to find sufficient sets of variables to control for confounding

- Race, education, age in this case

To Select or Not To Select

After we find our minimal set of confounders that is sufficient for adjustment

- if our DAG is correct…
- we have no measurement error…
- we have the proper functional form…
- etc, etc

Controlling for this set of confounders will get us an unbiased estimate (in expectation)

Is it really worth the reduction in variance (gain in precision) to get rid of a confounder and introduce bias?

Crude vs Adjusted

There is a substantial change in the point estimate between the crude and adjusted:

$$\left[\frac{-0.611 - (-0.341)}{-0.341}\right]_{=.205} \quad \left[\frac{-0.071 - (-0.341)}{-0.341}\right]_{=.205}$$

Clearly, we should include at least some of these confounders

But is it worth the effort, given the modest gain in precision?

I would not proceed with a confounder selection approach in this case

In my experience, most datasets with moderate sample sizes have not warranted the extra effort

But if we did…

Model	Point Estimate	Lower CI	Upper CI	Width	Change in Est
Full	-0.341	-0.611	-0.071	0.540	-
-age	-0.339	-0.606	-0.072	0.534	0.6%
-age, race	-0.329	-0.596	-0.062	0.534	3.5%
-age, race, ces-d	-0.395	-0.649	-0.140	0.509	-15.8%

Final Model?

Perhaps you could make a case for the results of the full model minus age and race in terms of parsimony and slightly narrow CIs

(though I'm skeptical)

But you could never convince a reviewer that you result shouldn't be adjusted for age and race!

My Current Model Selection Algorithm

1. Draw the DAG
2. Figure out what sets of variables could control confounding (under idealized circumstances)
3. Compare regression results under the "fully adjusted" model to the "crude" model.
4. If there is no appreciable gain in precision to be found, stop here
5. If there is some appreciable gain in precision, consider backward elimination using the change-in-estimate approach

Limitations

All of the model selection algorithms are cheating!

They look at the data numerous times, running multiple models and conducing multiple tests/comparisons

There should be a penalty for this!!!!

In fact, you can show that the CIs you get from these procedures are too narrow (because they do not force you to pay a penalty)

This, too, argues strongly in favor of the "Draw the DAG; Adjust for what needs to be adjusted for; Stop" Algorithm

Biological vs Statistical Interactions

- Richard MacLehose, PhD

Interactions - DAGs

What do people mean when they say effect modification and interaction

1. effect modification is what you say when communicating to an epidemiologist; interaction, to a statistician
2. with effect modification, there is emphasis primarily on 1 variable; interaction keeps the 2 variables on equal footing
3. DAGs are agnostic about interaction or EM

Statistical View

Consider 3 models

$$\Pr(Y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2$$
$$\log[\Pr(Y)] = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2$$
$$\text{logit}[\Pr(Y)] = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2$$

Each of these models includes an interaction term between x_1 and x_2

If β_3 is non-zero, then there is an interaction

if β_3 is non-zero x_1 modifies the effect of x_2 on Y

if β_3 is non-zero x_2 modifies the effect of x_1 on Y

…on that scale (additive, log or logit)

Interactions are Scale Dependent

If an interaction is present on one scale, it may or may not be present on another

Logistic-scale interaction doesn't imply an additive interaction

However, lack of an interaction on one scale DOES imply interaction on every other scale provided both variables have a main effect

Lack of an additive interaction, implies a logistic interaction

Effect Modification is Scale Dependent

To be clear, when we are talking about effect modification we should specify the scale as well

There is effect modification of the effect of asbestos on lung cancer by smoking on the RD scale, for example

This has led to the clearer term "effect measure modification" (EMM)

There is some discrepancy between how a statistician and epidemiologists would think about EMM.

Statistician: the interaction depends on the scale of the outcome in the regression model

Epidemiologist: the interaction depends on the scale of the effect measure

Thought Experiment*

Imagine there are two exposures, x1 and x2, and one outcome, y.

Pretend there are 4 types of people in the world:

1. those who get y regardless of x1 or x2 (doomed)
2. those who get y if they are exposed to x1, regardless of x2 (x1 causative)
3. those who get y if they are exposed to x2, regardless of x1 (x2 causative)
4. those who are will not get y regardless of x1 or x2 (immune)

Notice, there are no types of people for whom x1 and x2 interact to cause disease

- (5.) those who get y only if they are exposed to x1 and x2 (Causal Synergism)

*Greenland, Poole 1988. Scand J. Work Environment & Health
(and course work with Poole)

Omnipotence…

	Doomed (y)	x1 causal (Y)	x2 causal (Y)	Immune (Y)
x1=1 x2=1	1	1	1	0
x1=1 x2=0	1	1	0	0
x1=0 x2=1	1	0	1	0
x1=0 x2=0	1	0	0	0

$Pr(Y=1|Set[x_1=1],Set[x_2=1])=.75$
$Pr(Y=1|Set[x_1=0],Set[x_2=1])=.50$
$Pr(Y=1|Set[x_1=1],Set[x_2=0])=.50$
$Pr(Y=1|Set[x_1=0],Set[x_2=0])=.25$

$RD|Set[X_2=0]=0.50-0.25=0.25$
$RD|Set[X_2=1]=0.75-0.50=0.25$
$RR|Set[X_2=0]=0.50/0.25=2.0$
$RR|Set[X_2=1]=0.75/0.50=1.5$

	Doomed (y)	causal (Y)	x2 causal (Y)	Immune (Y)
x1=1 x2=1	1	1	1	0
x1=1 x2=0	1	1	0	0
x1=0 x2=1	1	0	1	0
x1=0 x2=0	1	0	0	0
N	100	100	100	100

Interactions and Effect Measures (1)

If there are no 'interacting types' in the population (and both exposures have an effect) the stratum specific RD's will be homogenous. RR's and OR's can be heterogenous even in the absence of 'casual interaction.'

Omnipotence (2)…

	Doomed (y)	x1 causal (Y)	x2 causal (Y)	Causal Synergism (Y)	Immune (Y)
x1=1 x2=1	1	1	1	1	0
x1=1 x2=0	1	1	0	0	0
x1=0 x2=1	1	0	1	0	0
x1=0 x2=0	1	0	0	0	0

$Pr(Y=1|Set[x_1=1],Set[x_2=1])=.80$
$Pr(Y=1|Set[x_1=0],Set[x_2=1])=.40$
$Pr(Y=1|Set[x_1=1],Set[x_2=0])=.40$
$Pr(Y=1|Set[x_1=0],Set[x_2=0])=.20$

$RD|Set[X_2=0]=0.40-0.20=0.20$
$RD|Set[X_2=1]=0.80-0.40=0.40$
$RR|Set[X_2=0]=0.40/0.20=2.0$
$RR|Set[X_2=1]=0.80/0.40=2.0$

	Doomed (y)	causal (Y)	x2 causal (Y)	Causal Synergism (y)	Immune (Y)
x1=1 x2=1	1	1	1	1	0
x1=1 x2=0	1	1	0	0	0
x1=0 x2=1	1	0	1	0	0
x1=0 x2=0	1	0	0	0	0
N	100	100	100	100	100

Interactions and Effect Measures (2)

In this example, if there are 'interacting types' in the population (and both exposures have an effect) the stratum specific RD's will be heterogenous. RR's and OR's may be heterogenous or homogenous in the presence of 'casual interaction.'

In the presence of interacting causal types, it is possible for these types to 'cancel out' such that the stratum specific RD's can be homogenous even in the pretense of interactions

- Presence of RD heterogeneity implies interacting causal types
- Presence of RD homogeneity doesn't imply anything

Why Use Non-Linear Models?

The Point: The additive measure is more relevant when assessing interactions

We can't always run linear regressions

Study design may dictate the model we run

- Case-control studies often require logistic regression

A linear regression may provide a terrible fit

Logistic or log-linear models may be far easier to fit to the data

Assessing Additive Interactions in non-Additive Models

Interaction contrast (IC)

$$IC = R_{11} - R_{10} - R_{01} + R_{00}$$
$$IC = (R_{11} - R_{10}) - (R_{01} - R_{00})$$
$$IC = RD_1 - RD_0$$

Interaction contrast ratio (ICR), aka RERI

$$ICR = IC / R_{00}$$
$$ICR = R_{11} / R_{00} - R_{10} / R_{00} - R_{01} / R_{00} + R_{00} / R_{00}$$
$$ICR = RR_{11} - RR_{10} - RR_{01} + 1$$

	Smoke No	Smoke Yes
Asbestos No	0.0012	0.0129
Asbestos Yes	0.0066	0.0602

$$R_{00} = 0.0012$$
$$R_{01} = 0.0129$$
$$R_{10} = 0.0066$$
$$R_{11} = 0.0602$$

IC=0.0602-0.0066-0.0129+0.0012=0.0419

IC>0, means "super-additive" interaction

- The two exposures together yields a larger effect than if they acted independently

ICR=0.0602/0.0012-0.0066/0.0012-0.0129/0.0012+1=34.92

- ICR>0 means "super-additive" interaction
- The exact size of that "super-additivity" is not known with ICR unless R_{00} is known

IC and ICR: Regression

More often, we'll want to control for variables or use a regression to calculate IC and ICR

IC & CI are easily accomplished via a regression model

ICR is easy, but the CI are a bit more complex

```
. logistic lung i.smoke##i.asbestos

Logistic regression                               Number of obs   =      15022
                                                  LR chi2(3)      =     262.48
                                                  Prob > chi2     =     0.0000
Log likelihood = -1003.6958                       Pseudo R2       =     0.1156

------------------------------------------------------------------------------
        lung | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      1.smoke |   11.02119   4.643674     5.70   0.000     4.825948    25.16947
  1.asbestos |    5.626375   3.414051     2.85   0.004     1.712871    18.48131
             |
smoke#asbestos |
         1 1 |    .8700761   .5422806    -0.23   0.823     .2564722    2.951713
             |
       _cons |    .0011865   .0004847   -16.49   0.000     .0005328    .0026422
------------------------------------------------------------------------------

. di exp(_b[1.smoke])*exp(_b[1.asbestos])*exp(_b[1.smoke#1.asbestos])-exp(_b[1.smok
> e])-exp(_b[1.asbestos])+1
38.305276
```

ICR

ICR Confidence Intervals

This is a little bit trickier

The ICR is a non-linear function of the regression coefficients

A few options

1. Wald-style intervals (see Hosmer, Lemeshow 1992)
2. Bootstrap intervals (see Assmann, Hosmer, Lemeshow, 1996)
3. Likelihood based intervals (Richardson, Kaufman, 2008)

Trends and Splines

Richard MacLehose, PhD

Dose-Response

Dose response curves are very important in public health

Often used as an indication of "causality"

Criteria #7 in the much misunderstood Doll-Hill list of causal criteria

For many exposure-disease relationships, if x amount of an exposure leads to some probability of disease, 2x amount of exposure may lead to even higher probability of disease

Important in exposure modeling

Should we dichotomize the exposure? if so, where?

Is there a threshold? if so, where?

Is the dose-response linear?

Linear?

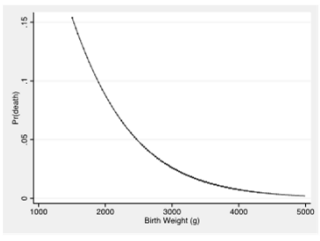
```
. logistic death bweight [pw=wt] ,cformat(%9.3f)

Logistic regression
Number of obs   =       2257

Wald chi2(1)    =      89.65

Prob > chi2     =     0.0000
Log pseudolikelihood = -31765.241
Pseudo R2       =     0.0540

------------------------------------------------------------------------------
             |               Robust
       death | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     bweight |       0.999      0.000    -9.47   0.000         0.998       0.999
       _cons |       1.237      0.507     0.52   0.604         0.554       2.760
------------------------------------------------------------------------------
```



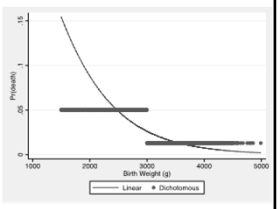
Binary?

What if we dichotomized

We are epidemiologists after all!

birthweight <3kg vs >=3kg?

```
. logistic death i.dbw [pw=wt]
------------------------------------------------------------------------------
             |               Robust
       death | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         dbw |
      >=3kg  |    .248184   .0240418   -14.39   0.000     .2052661    .3000755
       _cons |   .0528926   .0038729   -40.15   0.000     .0458214    .0610549
------------------------------------------------------------------------------
```



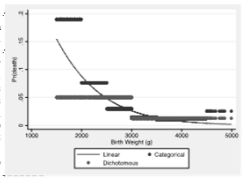
How can we check the model fit?

What if we categorize the exposure and examine the crude probability of death in each strata of the exposure?

- catbw: 1.5-2kg; 2-2.5kg;2.5-3kg; 3-3.5kg; 3.5-4kg; 4-4.5kg; >4.5kg

```
. logistic death i.catbw [pw=wt]
```

catbw	death 0	death 1
1.5-2kg	.8103	.1897
2-2.5kg	.9237	.0763
2.5-3kg	.9702	.0298
3-3.5kg	.9852	.0148
3.5-4kg	.99	.01
4-4.5kg	.9881	.0119
>4.5kg	.9744	.0256
Total	.9781	.0219



Perils

Linear (perhaps in the log-odds) trends may not be realistic and could miss very important changes

- such as very heavy infants being at increased risk of death

Dichotomization may be easy to interpret but can drastically over/understate the real risk

More discrete categorization can pick up interesting trends but:

- assumes everyone in a group has the same risk
- sudden jump in risk between categories
- requires a lot of data

Splines

Splines are a very flexible tool to estimate dose-response curves

Splines extend the idea of categorizing data

However, rather than have categories with sudden jumps in risk, we force the risk to be continuous between segments

Linear Splines

These are the easiest type of splines

Unrealistic, as we'll see

Define categories as before

- 2.0; 2.5; 3.0; 3.5; 4.0; 4.5
- These are now called "knots"

We will model the risk (or log-odds) as straight lines between these knots, but will force the end of one line segment to meet the end of the next (no jumps)

Linear Spline Creation

$Knot_1$=2, $knot_2$=3

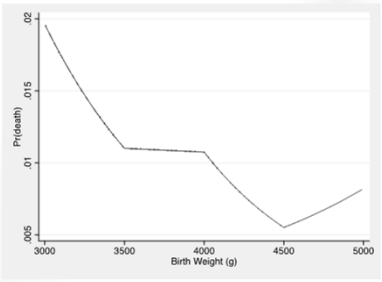
$spline_k=(bweight-knot_k)_+$

- The + notation means $(bweight-knot_k)=0$ if $(bweight-knot_k)<0$
- $spline_1$=(bweight-2.0) if bweight>2.0, else 0
- $spline_2$=(bweight-3.0) if bweight>3.0, else 0

```
. logistic death bweight spline1-spline6 [pw=wt]

Logistic regression                               Number of obs   =       2257
                                                  Wald chi2(7)    =     286.35
                                                  Prob > chi2     =     0.0000
Log pseudolikelihood = -30554.405                 Pseudo R2       =     0.0900

------------------------------------------------------------------------------
             |               Robust
       death | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     bweight |   .9985702   .0011225    -1.27   0.203     .9963725    1.000773
     spline1 |    .999434    .001691    -0.33   0.738     .9961251    1.002754
     spline2 |   .9999153   .0011209    -0.08   0.940     .9977208    1.002115
     spline3 |    1.00091   .0007862     1.16   0.247     .9993703    1.002452
     spline4 |   1.001124   .0008214     1.37   0.171     .9995154    1.002735
     spline5 |   .9987053    .001277    -1.01   0.311     .9962054    1.001211
     spline6 |   1.002145   .0010417     2.06   0.039     1.000106    1.004189
       _cons |   2.689758   5.559798     0.48   0.632     .0468013    154.5854
------------------------------------------------------------------------------
```



Zooming in on the linear spline prediction

Quadratic Splines

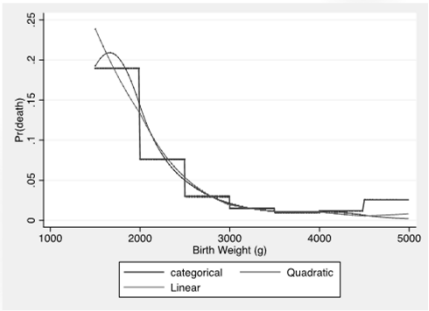
Linear splines cure the problem of large jumps in probability between categories. However, the transition at the knots is too jagged

- If we allow the line segments in each category to be quadratic curves in each segment, we can cure this
- Linear Spline: $\log it[\Pr(Y_i)] = \beta_0 + \beta_1 bweight_i + \beta_2 spline_{1,i} + \beta_3 spline_{2,i} \ldots$
- Quadratic: $\log it[\Pr(Y_i)] = \beta_0 + \beta_1 bweight_i + \beta_2 bweight_i^2 + \beta_3 spline_{1,i}^2 + \beta_4 spline_{2,i}^2 \ldots$

```
. logistic death bweight bweight_sq spline1_sq spline2_sq spline3_sq spline4_sq spl
> ine5_sq spline6_sq [pw=wt]

Logistic regression                               Number of obs   =       2257
                                                  Wald chi2(8)    =     282.03
                                                  Prob > chi2     =     0.0000
Log pseudolikelihood =  -30539.81                 Pseudo R2       =     0.0904

------------------------------------------------------------------------------
             |               Robust
       death | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     bweight |   1.013535   .0193658     0.70   0.482     .9762809    1.052211
  bweight_sq |    .999996   5.10e-06    -0.79   0.427      .999986    1.000006
  spline1_sq |   1.000005   6.88e-06     0.71   0.476     .9999914    1.000018
  spline2_sq |   .9999993   3.12e-06    -0.22   0.830     .9999932    1.000005
  spline3_sq |   1.000001   2.27e-06     0.28   0.780     .9999962    1.000005
  spline4_sq |   1.000001   2.67e-06     0.24   0.808     .9999954    1.000006
  spline5_sq |   .9999953   5.56e-06    -0.84   0.402     .9999844    1.000006
  spline6_sq |   1.000004   4.77e-06     0.80   0.425     .9999945    1.000013
       _cons |   3.72e-06   .0000658    -0.71   0.480     3.30e-21    4.20e+09
------------------------------------------------------------------------------
```



Spline Problems

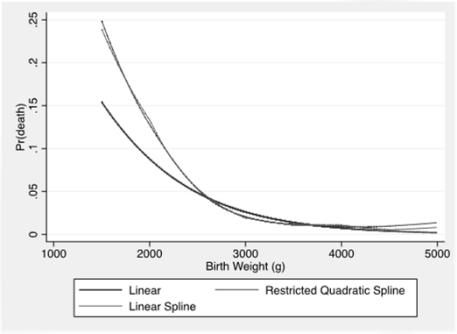
The tail end of splines are often VERY unstable

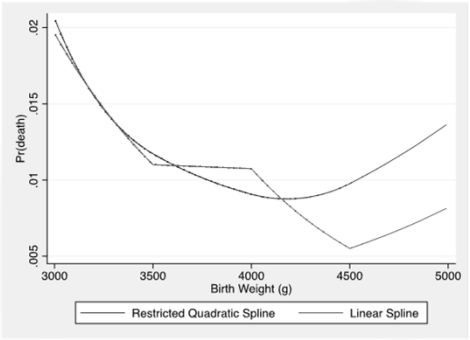
Little data exists in at either end to estimate these curves

The quadratic term can lead to very odd results

- As a result we usually restrict the ends of our spline curve to be linear
  - Delete the first quadratic term
  - subtract the last quadratic term from the remaining spline terms

$$\log it[\Pr(Y_i)] = \beta_0 + \beta_1 bweight_i + \beta_2 (spline_{1,i}^2 - spline_{1,i}^2) + \beta_3 (spline_{2,i}^2 - spline_{1,i}^2) \ldots$$





If Quadratic Splines are Better than Linear Ones…

- Can we keep increasing the power and improve fit?
- Yes
- Cubic splines are very common
- They are a bit smoother than quadratic splines
- Cubic splines have nice theoretical properties
  - They minimize mean square error
- In my experience, its hard to tell the difference between cubic and quadratic splines
- Restricted cubic splines are called "natural cubic splines"

Choosing Knots

- Spline results can be very sensitive to knot location
- There's no great rule for choosing knot locations
  - Often choose equally space percentiles
    - 20, 40, 60, 80 for example
- There are methods that allow for random knot location
  - hard to implement
- Wise to rerun the analysis with different knots to see if you get different results

Spline Problems

- Interpretation can be trickier
- Plots give a lovely visual interpretation
- No effect measures = hard for clinicians
- Could calculate RR's, RD's at various exposure patterns

Ordinal & Multinomial Logistic Regression

Richard MacLehose, PhD

Introduction

Ordinary logistic regression: dichotomous outcome

yes/no; dead/alive; infected/not infected

What if we have more than 2 outcome choices:

- unordered: Yes/No/Maybe; No Cancer/lung cancer/pancreatic cancer;
- ordered: CD4 count 0-100/ 101-200/ 201-300/ 301+ ; Strongly disagree/ disagree/neutral/agree/strongly agree

Multinomial Logistic Regression

Multinomial logistic regression is an extension of logistic regression to >2 outcome categories (typically unordered outcome categories)

- Outcomes must be mutually exclusive
- Also called polytomous logistic regression

Odds model

$$odds(Y) = \frac{\Pr(Y=1)}{\Pr(Y=0)} = \exp(\alpha + \beta_1 x)$$

Probilities

$$\Pr(Y=0) = \frac{1}{1+\exp(\alpha+\beta_1 x)}$$

$$\Pr(Y=1) = \frac{\exp(\alpha+\beta_1 x)}{1+\exp(\alpha+\beta_1 x)}$$

Effect Measure

$$OR = \frac{Odds(Y \mid X=1)}{Odds(Y \mid X=0)} = \frac{\exp(\alpha+\beta_1)}{\exp(\alpha)} = \exp(\beta_1)$$

Odds model

$$odds(Y=1 vs Y=0) = \frac{\Pr(Y=1)}{\Pr(Y=0)} = \exp(\alpha_1 + \beta_1 x)$$

$$odds(Y=2 vs Y=0) = \frac{\Pr(Y=2)}{\Pr(Y=0)} = \exp(\alpha_2 + \beta_2 x)$$

Effect Measure

$$OR = \frac{Odds(Y=1 vs Y=0 \mid X=1)}{Odds(Y=1 vs Y=0 \mid X=0)} = \frac{\exp(\alpha_1+\beta_1)}{\exp(\alpha_1)} = \exp(\beta_1)$$

$$OR = \frac{Odds(Y=2 vs Y=0 \mid X=1)}{Odds(Y=2 vs Y=0 \mid X=0)} = \frac{\exp(\alpha_2+\beta_2)}{\exp(\alpha_2)} = \exp(\beta_2)$$

Multinomial Regression

We have more coefficients (extra set for each outcome contrast)

We need to be careful which exposure contrast we're talking about but also which outcome contrasts

"Exposed individuals have 1.2 times the odds of Y=1 than Y=0, relative to unexposed individuals"

$$odds(Y=1 vs Y=0) = \frac{\Pr(Y=1)}{\Pr(Y=0)} = \exp(\alpha_1 + \beta_1 x)$$

$$odds(Y=2 vs Y=0) = \frac{\Pr(Y=2)}{\Pr(Y=0)} = \exp(\alpha_2 + \beta_2 x)$$

It Quacks Like a Duck…

There are those who do not like calling these quantities "odds"

For them: odds=Pr(Y=1)/(1-Pr(Y=1))

But…Mirriam Webster: the ratio of the probability of one event to that of an alternative event

They prefer to call these quantities "risk ratios"

$\exp(\beta_1)$ then a "relative risk ratio"

I really don't like this nomenclature, but its what Stata uses

$$odds(Y=1 vs Y=0) = \frac{\Pr(Y=1)}{\Pr(Y=0)} = \exp(\alpha_1 + \beta_1 x)$$

$$odds(Y=2 vs Y=0) = \frac{\Pr(Y=2)}{\Pr(Y=0)} = \exp(\alpha_2 + \beta_2 x)$$

Interpretation

Smokers have 0.57 times the odds of being overweight (relative normal weight) compared to non smokers

Smokers have 0.40 times the odds of being obese (relative to normal) compared to non-smokers

Is there a simpler alternative?

Indeed. You could run two logistic regression models:

1. Outcome=(Normal,overweight); exclude obese
   - ↖ logit wtcat cursmoke age if wtcat!=2
2. Outcome=(Normal,Obese); exclude overweight
   - ↖ logit wtcat cursmoke age if wtcat !=1

Multinomial Cautions

Whether you want to call exponentiated multinomial coefficients ORs or not, they behave like ORs.

- Hard to interpret
- Not collapsible, not a good measure of effect unless the risk is low

You can use 'margins' after mlogit in stata to avoid interpreting Ors…more on this later in the semester

Independence of Irrelevant Alternatives

The odds of choosing one outcome relative to another should be independent of other outcomes

Say you could choose to go to work by a red bus or a car. If you're equally likely to go by car vs red bus, the odds=.5/.5=1

Now, introduce another outcome, a blue bus that travels the same path to work as the red bus.

For the multinomial model to hold, the introduction of this new colored bus must not alter the previous odds=1

- But it likely would make the car vs red bus odds =.5/.25=2

Think of the iia assumption as probiting nested decision making: car vs bus then red vs blue

Ordinal Regression

Sometimes we'll have categorical outcomes but they'll have a clear order

bad; neutral; good

underweight ;normal; overweight; obese

0-100; 101-200; 201-300; 301+ CD4 counts

We could model them with a multinomial regression, but might get better power out of an ordinal regression that respect the rank order (and limits the number of coefficients)

Two types:

1. Proportional odds (cumulative odds) models
2. Continuation ratio models

Proportional Odds Model

Compare lower to higher at various cut-points

Low | Medium-Low | Medium-High | High

Low | Medium-Low | Medium-High | High

Low | Medium-Low | Medium-High | High

Proportional Odds Model

$$\Pr(Y \geq z) = \frac{\exp(\alpha_z + \beta x)}{1 + \exp(\alpha_z + \beta x)}$$

$$\Pr(Y < z) = \frac{1}{1 + \exp(\alpha_z + \beta x)}$$

$$odds(Y \geq z) = \frac{\Pr(Y \geq z)}{\Pr(Y < z)} = \exp(\alpha_z + \beta x)$$

What's the P-Odds Model Assuming?

The intercepts are telling us about the different baseline odds

α_z is the odds of the zth or greater outcome

While the odds of the outcomes vary by level, the odds ratios are constant

- X has the same OR if we're talking about Y>=2 vs Y<2 or y>=4 vs y<4

$$odds(Y \geq z) = \frac{\Pr(Y \geq z)}{\Pr(Y < z)} = \exp(\alpha_z + \beta x)$$

Interpretation

Smokers have 0.53 times the odds of being overweight or obese (vs normal weight) than non-smokers.

- Smokers have 0.53 times the odds of being obese (vs overweight or normal weight) than non-smokers.

Problems and Checks

The constant OR between categories seems unrealistic in many settings

I don't think I've ever used a proportional odds model in an analysis before

Can check the proportional odds model by running separate logistic models

1. Normal vs (Over/Obese)
2. (normal & over) vs obese

Continuation Ratio Models

Continuation Ratio $\frac{\Pr(Y > z)}{\Pr(Y = z)} = \exp(\alpha_z + \beta x)$

- Inverse Continuation Ratio $\frac{\Pr(Y = z)}{\Pr(Y < z)} = \exp(\alpha_z + \beta x)$
- Stata can implement inverse continuation ratio with "ocratio"

Correlated Data

What is Correlated Data?

- Data that "cluster" or are grouped along one or more dimensions
- In practice, numerous dimensions along which our data can cluster:
  - Time: Repeated Measures
  - Familial Unit: Couples, Siblings
  - Setting: Classroom, Hospital
  - Space: Block group, ZIP code
- This clustering often invokes correlation in the data, since those in the same cluster (family, school, time) are often more similar than those in different clusters
- Violates the 'residuals are independent' assumption of most regression models

Quantifying the Correlation

- With longitudinal data, examine the zero-order correlation for different time points
- With nested data, typically use the intraclass correlation to assess the level of similarity of subjects in the same cluster (vs. those in different clusters)
  - "proportion of variance accounted for by the cluster"
  - ICC = variance due to clustering/total variance
  - DEFF or VIF = 1 + ((m-1)*ICC)

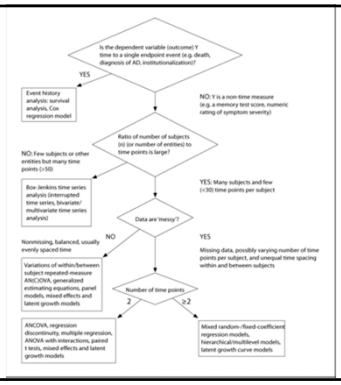
Many (MANY, MANY!) Approaches to Analyze Longitudinal Data

- "Change Score"
- Baseline Adjusted ANCOVA
- Repeated Measures ANOVA
- MANOVA
- Fixed Effects (e.g., Difference in Differences)
- Random Effects (e.g., Multilevel Model)
- GEE

Marginal vs. Structural Approaches

- Structural models explicitly model the correlation due to clustering using fixed or random effects.
- There are many situations where one is not particularly interested in the estimates for the "clustered" part of the data, but just want to have accurate estimates and standard errors for the fixed effect predictors.
- Marginal models are good at controlling for the effects of nonindependence, and in many instances are preferable to "structural" models (e.g., when the random part of the model is unknown, more robust to model misspecification).
- We'll directly compare both approaches when we get to longitudinal data examples.

- Covariance Pattern Model
- Mixed Model Regression
  - Random Intercept (Nested Data)
  - Random Slopes (Longitudinal Data)
- Discrete Outcome
  - Generalized Estimating Equations (GEE)
  - Generalized Mixed Modeling



Covariance Pattern Modeling

Covariance Pattern Model

- Proposed by Jennrich and Schluchter (1986)
- Conceptually an extension of MANOVA (Multiple ANOVA)
  - MANOVA -> CPM
  - Repeated Measures ANOVA -> Mixed Model Regression
- Time is treated categorically and Observations are assumed measured at same points (but can have missingness)
- Model the variances and covariances of the repeated observations directly without regard for between and within-variance
- Use PROC MIXED in SAS and xtmixed/mixed in STATA, but not technically a mixed model

Covariance Pattern Model

- $Y_{it} = b_0 + b_1 Time_t + b_2 X_i + e_{it}$
  - $e_{it} \sim N(0,S)$
    - where S specifies the pattern of variances and covariances (Usually Independent, Compound, Autoregressive, Toeplitz, Unstructured Correlation Matrices)
- Assumptions: like OLS, associations are linear and residuals are normal. Unlike OLS, residuals can be heteroscedastic and correlated

General Linear Model

- $y_i = b_0 + b_1 x_i + e_i$

  where i=1 to N study participants, y is the outcome, x is a predictor, b_0 is an intercept, b_1 is the regression coefficient, and e is residual
- One outcome, one or more predictors, one source of random variation
- Assumptions: associations are linear, residuals are normal, homoscedastic, and independent

General Link Functions Used in Data Analysis

Common distributions with typical uses and canonical link functions

Distribution	Support of distribution	Typical uses	Link name	Link function	Mean function
Normal	real: $(-\infty, +\infty)$	Linear-response data	Identity	$\mathbf{X}\boldsymbol{\beta} = \mu$	$\mu = \mathbf{X}\boldsymbol{\beta}$
Exponential	real: $(0, +\infty)$	Exponential-response data, scale parameters	Inverse	$\mathbf{X}\boldsymbol{\beta} = -\mu^{-1}$	$\mu = -(\mathbf{X}\boldsymbol{\beta})^{-1}$
Gamma					
Inverse Gaussian	real: $(0, +\infty)$		Inverse squared	$\mathbf{X}\boldsymbol{\beta} = -\mu^{-2}$	$\mu = (-\mathbf{X}\boldsymbol{\beta})^{-1/2}$
Poisson	integer: $0, 1, 2, \ldots$	count of occurrences in fixed amount of time/space	Log	$\mathbf{X}\boldsymbol{\beta} = \ln(\mu)$	$\mu = \exp(\mathbf{X}\boldsymbol{\beta})$
Bernoulli	integer: $\{0, 1\}$	outcome of single yes/no occurrence	Logit	$\mathbf{X}\boldsymbol{\beta} = \ln\left(\frac{\mu}{1-\mu}\right)$	$\mu = \frac{\exp(\mathbf{X}\boldsymbol{\beta})}{1+\exp(\mathbf{X}\boldsymbol{\beta})} = \frac{1}{1+\exp(-\mathbf{X}\boldsymbol{\beta})}$
Binomial	integer: $0, 1, \ldots, N$	count of # of "yes" occurrences out of N yes/no occurrences			
Categorical	integer: $[0, K)$ K-vector of integer: $[0, 1]$, where exactly one element in the vector has the value 1	outcome of single K-way occurrence			
Multinomial	K-vector of integer: $[0, N]$	count of occurrences of different types (1 .. K) out of N total K-way occurrences			

Common Covariance Patterns

- Compound Symmetry
  - Common variance at each time
  - Common covariance for each $cov_{i,j}$
  - 2 parameters: s^2 and s_1

$$\begin{bmatrix} \sigma^2+\sigma_1 & \sigma_1 & \sigma_1 & \sigma_1 \\ \sigma_1 & \sigma^2+\sigma_1 & \sigma_1 & \sigma_1 \\ \sigma_1 & \sigma_1 & \sigma^2+\sigma_1 & \sigma_1 \\ \sigma_1 & \sigma_1 & \sigma_1 & \sigma^2+\sigma_1 \end{bmatrix}$$

Common Covariance Patterns

- First order Autoregressive
  - Common variance at each time
  - Covariance for each $cov_{i,j}$ decays exponentially with distance apart in time
  - 2 parameters: s^2 and r

$$\sigma^2 \begin{bmatrix} 1 & \rho & \rho^2 & \rho^3 \\ \rho & 1 & \rho & \rho^2 \\ \rho^2 & \rho & 1 & \rho \\ \rho^3 & \rho^2 & \rho & 1 \end{bmatrix}$$

Common Covariance Patterns

- Toeplitz
  - Common variance at each time
  - Unique covariance for each $cov_{i,j}$ 'band'
  - t parameters: s^2 and t-1 covariances

$$\begin{bmatrix} \sigma^2 & \sigma_1 & \sigma_2 & \sigma_3 \\ \sigma_1 & \sigma^2 & \sigma_1 & \sigma_2 \\ \sigma_2 & \sigma_1 & \sigma^2 & \sigma_1 \\ \sigma_3 & \sigma_2 & \sigma_1 & \sigma^2 \end{bmatrix}$$

Common Covariance Patterns

- Unstructured
  - Unique variance at each time
  - Unique covariance for each $cov_{i,j}$ pair
  - t(t+1)/2 parameters
  - Quickly becomes a large number !

$$\begin{bmatrix} \sigma_1^2 & \sigma_{21} & \sigma_{31} & \sigma_{41} \\ \sigma_{21} & \sigma_2^2 & \sigma_{32} & \sigma_{42} \\ \sigma_{31} & \sigma_{32} & \sigma_3^2 & \sigma_{43} \\ \sigma_{41} & \sigma_{42} & \sigma_{43} & \sigma_4^2 \end{bmatrix}$$

Preparing for the analysis

- Do your literature review to see how the outcome of interests / similar analyses were reported in the literature
- Make Blank tables for your analysis in an excel spreadsheet (Table 1 for demographics, Table 2 for main model, Table 3 for secondary/stratified analyses, and figures when appropriate i.e. interaction, clustered data)

Analysis Steps: EDA

- Put the data in wide format (Each person has one row)
  - Run descriptive statistics (mean, variance, histogram, boxplots) on the outcome
  - Run descriptive statistics of your exposure variables (for your Table 1, make sure that covariates are equally distributed across Txt arms or case/control status)
  - Observe the correlation in the outcomes (pearson correlation coefficients/scatterplots of repeated measures/ Calculate the ICC and Covariance Matrices)
  - Cross tabulate or run simple t-tests for your outcome vs your exposures

Analysis steps: Modeling

- Put the data in long format (each correlated observation has its own row)
- Based on your EDA and outcome type (continuous, binary, categorical or count) pick your regression model
- Run crude model between outcome and main exposure -> Run adjusted models, and properly specify correlation if needed
- Look at stratified analyses (Breslow day for stratum specific estimates) -> Look at pooled estimates interactions -> Look for potential confounding
- Make figures and graphs where appropriate

Model Troubleshooting

- If your model does not converge, check your outcome variables to make sure you are using the appropriate link function for the data (linear, logistic, poisson, Hazard, cox, ect…)
- The second step is to check for complete separation in your data (on the continuous scale, when your predictor perfectly separate your outcome values: more common with binomial outcomes, try to make bigger categories in your predictors to remediate that)
- The last step is to change the convergence criterion: Least square means > Maximum Likelihood > Restricted Maximum Likelihood

A Four-Wave Repeated Measures Data Example

- N=300 14-year olds
- 4 waves of annual data (baseline, followup1, followup 2, followup 3)
- Outcome = Smoking Stage (coded 1 [nonsmoker] to 6 [daily smoker])
- Predictors = Age, Sex, Race (and Time)

Regression Formula

$smoke_{it} = b_0 + b_1(male_i) + b_2(white_i) + b_3(baseline\_age_i) + b_4(followup1_i) + b_5(followup2_i) + b_6(followup3_i) + e_{it}$

***The Follow up variable here indicates when each cluster was made for each observation

Analysis

Wide dataset (first 10 observations)

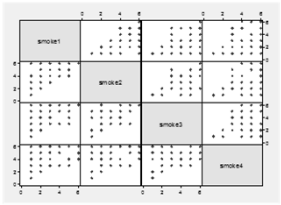
. list in 1/10

Means, variances and correlations

. corr smoke*, means

Plot correlations

. graph matrix smoke*



Convert dataset from wide to long

. reshape long smoke, i(id) j(time)

Create new variables for repeated measures timepoints

. gen age14 = age_bl – 14

. gen baseline = time == 1

. gen followup1 = time == 2

. gen followup2 = time == 3

. gen followup3 = time == 4

"Long" Dataset (first 12 obs)

```
. list in 1/12, sep(4)

     +-------------------------------------------------------------------------------------+
     |     id   time   white   male   smoke   age_bl   age14   baseline   follow~1   follow~2   follow~3 |
     |-------------------------------------------------------------------------------------|
  1. | 0000026      1       1      1       1    14.08      .08          1          0          0          0 |
  2. | 0000026      2       1      1       1    14.08      .08          0          1          0          0 |
  3. | 0000026      3       1      1       2    14.08      .08          0          0          1          0 |
  4. | 0000026      4       1      1       2    14.08      .08          0          0          0          1 |
     |-------------------------------------------------------------------------------------|
  5. | 0000031      1       1      1       1    14.25      .25          1          0          0          0 |
  6. | 0000031      2       1      1       1    14.25      .25          0          1          0          0 |
  7. | 0000031      3       1      1       1    14.25      .25          0          0          1          0 |
  8. | 0000031      4       1      1       1    14.25      .25          0          0          0          1 |
     |-------------------------------------------------------------------------------------|
  9. | 0000041      1       1      0       2    14.58      .58          1          0          0          0 |
 10. | 0000041      2       1      0       2    14.58      .58          0          1          0          0 |
 11. | 0000041      3       1      0       2    14.58      .58          0          0          1          0 |
 12. | 0000041      4       1      0       5    14.58      .58          0          0          0          1 |
     +-------------------------------------------------------------------------------------+
```

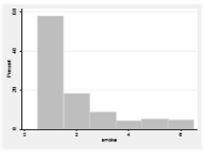
Summarize Outcome

```
. summarize smoke, detail

                            smoke
-------------------------------------------------------------
      Percentiles      Smallest
 1%            1              1
 5%            1              1
10%            1              1       Obs                1200
25%            1              1       Sum of Wgt.        1200

50%            1                      Mean           1.949167
                      Largest       Std. Dev.      1.451438
75%            2              6
90%            5              6       Variance       2.106672
95%            5              6       Skewness       1.157132
99%            6              6       Kurtosis       4.331571

. histogram smoke, discrete percent
```



Model 1: Independent

```
. mixed smoke male white age14 followup1 followup2 followup3 || id:, noconstant residuals(independent, t(time))
------------------------------------------------------------------------------
 Random-effects Parameters  |   Estimate   Std. Err.     [95% Conf. Interval]
----------------------------+-------------------------------------------------
                 var(Residual) |   1.988166   .0811665      1.83528    2.153787
------------------------------------------------------------------------------
```

Model 2: Compound Symmetry

```
. mixed smoke i.male i.white age14 ib1.time || id:, noconstant residuals(exchangeable, t(time)) nofetable nolog stddev
------------------------------------------------------------------------------
 Random-effects Parameters |   Estimate   Std. Err.      [95% Conf. Interval]
---------------------------+--------------------------------------------------
id:                (empty) |
---------------------------+--------------------------------------------------
Residual: Exchangeable     |
                  sd(e) |   1.410023   .0465142      1.321742    1.504201
                corr(e) |   .7329713   .0201337      .6909875    .7700268
------------------------------------------------------------------------------
LR test vs. linear regression:       chi2(1) =   839.52   Prob > chi2 = 0.0000
```

Model 3: Autoregressive

```
. mixed smoke i.male i.white age14 ib1.time || id:, noconstant residuals(ar 1, t(time)) nofetable nolog reml

------------------------------------------------------------------------------
 Random-effects Parameters |   Estimate   Std. Err.      [95% Conf. Interval]
---------------------------+--------------------------------------------------
id:                (empty) |
---------------------------+--------------------------------------------------
Residual: AR(1)            |
                  rho |   .8202728   .0133451      .7923448    .8447699
               var(e) |   1.976441   .128285      1.740343    2.244569
------------------------------------------------------------------------------
LR test vs. linear regression:       chi2(1) =  1019.66   Prob > chi2 = 0.0000
```

Model 4: Toeplitz

```
. mixed smoke i.male i.white age14 ib1.time || id:, noconstant residuals(toeplitz, t(time)) nofetable nolog stddev

------------------------------------------------------------------------------
 Random-effects Parameters |   Estimate   Std. Err.      [95% Conf. Interval]
---------------------------+--------------------------------------------------
id:                (empty) |
---------------------------+--------------------------------------------------
Residual: Toeplitz(3)      |
                 rho1 |  .8179375   .0132776      .7901862    .8423414
                 rho2 |  .6712907   .0257688      .6176239    .7187349
                 rho3 |  .5205164   .0425774      .4321892    .5989512
                 sd(e) |   1.395543   .0445516      1.310899    1.485652
------------------------------------------------------------------------------
LR test vs. linear regression:       chi2(3) =  1022.48   Prob > chi2 = 0.0000
```

Model 5: Unstructured

```
. mixed smoke c.male c.white age14 ib1.time || id:, noconstant residuals(un, t(time)) nofetable nolog stddev
------------------------------------------------------------------------------
 Random-effects Parameters |   Estimate   Std. Err.      [95% Conf. Interval]
---------------------------+--------------------------------------------------
id:                (empty) |
---------------------------+--------------------------------------------------
Residual: Unstructured     |
                 sd(e1) |   .9945137   .0406028      .9180344    1.077364
                 sd(e2) |   1.350744   .0551849      1.246801    1.463353
                 sd(e3) |   1.520149   .062538      1.402389    1.647799
                 sd(e4) |   1.694587   .0694044      1.563711    1.836417
            corr(e1,e2) |   .8207485   .0188618      .780182    .8544407
            corr(e1,e3) |   .7025175   .0295408      .6398591    .7558972
            corr(e1,e4) |   .6110963   .0364239      .5347191    .6775699
            corr(e2,e3) |   .8585169   .0152453      .8255636    .8836348
            corr(e2,e4) |   .7498723   .0253502      .6958175    .7954798
            corr(e3,e4) |   .8287791   .0181331      .7897335    .8611367
------------------------------------------------------------------------------
LR test vs. linear regression:       chi2(9) =  1170.26   Prob > chi2 = 0.0000
```

Which Covariance Pattern to Choose?

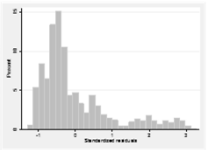
- Many ways to decide:
  - Parsimony
  - Experience/Knowledge
  - Model Fit – (what FLW recommend)
    - LR Test – nested models
    - Information Criteria – nonnested models
  - Visually/Graphically

Model Fit

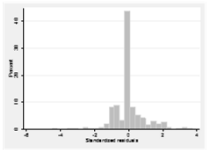
	Ind	CS	AR(1)	TOEP	UN
LL	-2115.054	-1695.296	-1615.315	-1603.815	-1529.923
df	8	9	9	11	17
AIC	4246.107	3408.591	3248.63	3229.63	3093.846
BIC	4286.828	3454.402	3294.441	3285.621	3180.377

Model Diagnostics

Standardized Residuals from Independent Model



Standardized Residuals from Unstructured Model



Comparing Model Results

parameter	Ind		CS		AR(1)		TOEP		UN	
Intercept	**1.96**	**.17**	**1.96**	**.28**	**1.88**	**.27**	**1.86**	**.27**	**1.73**	**.21**
Male	**-.24**	**.08**	-.24	.15	-.24	.14	-.24	.14	-.21	.11
White	**-.35**	**.14**	-.35	.25	-.28	.24	-.27	.24	-.04	.19
Age14	.05	.13	.05	.24	.10	.23	.11	.23	-.08	.19
Followup1	**.31**	**.12**	**.31**	**.06**	**.31**	**.05**	**.31**	**.05**	**.31**	**.05**
Followup2	**.53**	**.12**	**.53**	**.06**	**.53**	**.07**	**.53**	**.07**	**.53**	**.06**
Followup3	**.82**	**.12**	**.82**	**.06**	**.82**	**.08**	**.82**	**.08**	**.82**	**.08**

Model Interpretation

- Just like OLS!
- Intercept = predicted value of outcome when all predictors = 0
- Beta = change in predicted value of outcome associated with a unit change in that predictor
- Easy for a main effects linear regression like this

What if We Have Interactions?

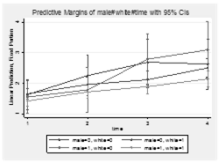
.mixed smoke i.male#i.white#ib1.time age14 || id:, noconstant residuals(un, t(time))

You could calculate margins for 3 way interactions:

. margins male#white#time

And then plot the margins

. marginsplot, xdimension(time) plotdimension(male white)



Or Calculate Margins Separately for Male*Time and White*Time

You could calculate Margins Separately for Male*Time and White*Time

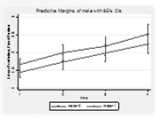
. margins male, over(time)

Get P-values

margins r.male, over(time)

And Plots

. marginsplot, xdimension(time) plotdimension(male)



You could calculate Margins Separately for Male*Time and White*Time

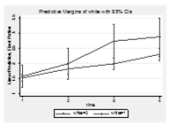
. margins white, over(time)

Get P-values

. margins r.white, over(time)

And Plots

. marginsplot, xdimension(time) plotdimension(white)



Summary

- Data that are clustered- nested- hierarchical-longitudinal often have nonzero correlation of the outcome within cluster
- This correlation creates correlated residuals, and failure to account for this correlation will produce biased standard errors
- Covariance pattern models relax the OLS assumption of independent (and homoscedastic) residuals
- Good model for longitudinal data
- Very few convergence issues
- Can accommodate interactions and time-varying exposures

Hierarchical Regression

PubH 8342
Feb 23, 2016

Hierarchical Data

- Individuals are nested in groups
- Individuals within a group are more similar to each other than individuals in different groups
- Two orientations to interpreting the effect of nesting on modeling:
  - Nuisance: standard errors are underestimated
  - Contextual phenomenon: between-group heterogeneity is important (but concealed if ignored in model)
- These two are often inter-twined in the real world

Hierarchical Linear Model

- Family of models with many variants/names – random effects model, multilevel model, mixed model, empirical bayes, etc.
- Level 1 = individuals
- Level 2 = groups/clusters
- Random intercept is a simple (simplest?) version
- Adds a 2nd random variable (i.e., random intercept) that models level 2 variability

The Statistical Model

$$Y_{ij} = b_0 + b_1(X_{ij}) + u_{0j} + e_{ij}$$

where i represents the n subjects per cluster and j represents the k clusters, b_0 is the overall intercept, b_1 is the slope describing the association between the predictor X and the outcome Y, and e_{ij} is the random error. The new part is u_{0j}, which is a random variable which represents a cluster-level intercept.

Fixed vs. Random Effects

- Fixed Effect
  - Most predictors and covariates are treated as fixed
  - Randomly assigned treatments or conditions typically considered fixed
  - All predictors and covariates in general linear model
- Random Effect
  - Predictors or covariates where specific level is not assigned, but based on a distribution of values
  - "Clusters randomly sampled from a population of clusters"
  - Want to generalize beyond specific levels in study

A Two-Level Clustered Data Example

- k=30 colleges
- M=39 students per college
- N=1170 students
- Outcome = BMI
- Predictor = Sex and Race

Analysis

Long Dataset
. list student college male white public bmi in 1/25
Look at the Outcome
. sum bmi
. sum bmi, detail
. histogram bmi
(bin=30, start=15.348812, width=.97986688)
Look at the predictors
. tab1 white male
Look at the clustering
. tab college

Calculate Intraclass Correlation

mixed bmi || college:

Log likelihood = -3352.4849 Prob > chi2 = .

bmi	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
_cons	23.58258	.1761361	133.89	0.000	23.23736	23.9278

Random-effects Parameters	Estimate	Std. Err.	[95% Conf.	Interval]
college: Identity				
var(_cons)	.4764582	.2410624	.1767511	1.284362
var(Residual)	17.71611	.7420461	16.31982	19.23186

LR test vs. linear regression: chibar2(01) = 9.53 Prob >= chibar2 = 0.0010

ICC = variance due to clustering/total variance
= .4765 / (.4765 + 17.716)
= .026

Between 2 and 3% of the variability in BMI among this sample of college students can be explained by or attributed to the college they attend.

DEFF = 1 + ((m-1)*ICC)
= 1 + (38 * .026)
= 1.99

Calculate Intraclass Correlation

. mixed bmi || college:

. estat icc

Intraclass correlation

Level	ICC	Std. Err.	[95% Conf.	Interval]
college	**.0261897**	.0130315	**.0097826**	**.0682184**

Ignore Clustering

bmi_i = intercept + beta1($male_i$) + beta2($white_i$) + e_i

. mixed bmi male white

Random-effects Parameters	Estimate	Std. Err.	[95% Conf.	Interval]
var(Residual)	17.45518	.7311171	16.07946	18.94861

Model Clustering as a Random Effect

bmi_{ij} = intercept + beta1($male_{ij}$) + beta2($white_{ij}$) + c_j($college_j$) + e_{ij}

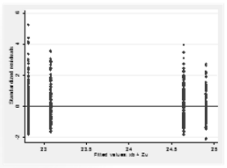
. mixed bmi i.male i.white || college:

Random-effects Parameters	Estimate	Std. Err.	[95% Conf.	Interval]
college: Identity				
var(_cons)	.5976011	.2700622	.2464589	1.449033
var(Residual)	16.85889	.7156552	15.51299	18.32155

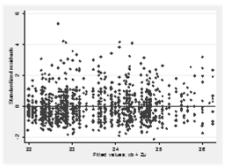
LR test vs. linear regression: chibar2(01) = 14.03 Prob >= chibar2 = 0.0001

Compare Residuals of 2 Models

Model ignoring Nesting



Model with Random Intercept for College



A Two-Level Clustered Data Example with Cluster-level Predictor

- k=30 colleges
- M=39 students per college
- N=1170 students
- Outcome = BMI
- Predictor = Sex and Race (Individual) and **Public/Private (Cluster)**

The Statistical Model

Level 1: $Y_{ij} = b_{0j} + b_{1j}(X_{ij}) + e_{ij}$
Level 2: $b_{0j} = g_{00} + g_{01}(Z_j) + u_{0j}$

where i represents the n subjects per cluster and j represents the k clusters, b_{0j} is the intercept for each of the j clusters, b_{1j} is the regression coefficient describing the association between the predictor X and the outcome Y (for each of the j clusters), e_{ij} is the random error for each individual i, g_{00} is the overall mean intercept and g_{01} is the regression coefficient describing the association between the level-2 predictor Z and the intercept, and u_{0j} is a random variable for the cluster.

Add Level-2 Predictor

```
. mixed bmi i.male i.white || college: public

------------------------------------------------------------------------------
  Random-effects Parameters |   Estimate   Std. Err.     [95% Conf. Interval]
----------------------------+-------------------------------------------------
college: Independent        |
                 var(public) |  1.401582   .779758      .4710421    4.170394
                 var(_cons) |  .0043475   .1051127     1.14e-23    1.65e+18
----------------------------+-------------------------------------------------
                 var(Residual) |  16.85622   .7140219     15.51328    18.31541
------------------------------------------------------------------------------
LR test vs. linear regression:       chi2(2) =    18.24   Prob > chi2 = 0.0001

Note: LR test is conservative and provided only for reference.
```

Re-Calculate ICC

```
*original model with just male and white
Intraclass correlation
------------------------------------------------------------------------------
                       Level |        ICC   Std. Err.     [95% Conf. Interval]
-----------------------------+------------------------------------------------
                     college |  .0261897   .0130315      .0097826    .0682184
------------------------------------------------------------------------------

*expanded model with male, white and public/private
Conditional intraclass correlation
------------------------------------------------------------------------------
                       Level |        ICC   Std. Err.     [95% Conf. Interval]
-----------------------------+------------------------------------------------
                     college |  .0002579   .006234     6.72e-25           1
------------------------------------------------------------------------------
Note: ICC is conditional on zero values of random-effects covariates.
Ratio = .00026/.026 = .98
Including type of school funding (public vs private) explained approximately 98% of the cluster-specific variability in BMI.
```

Summary

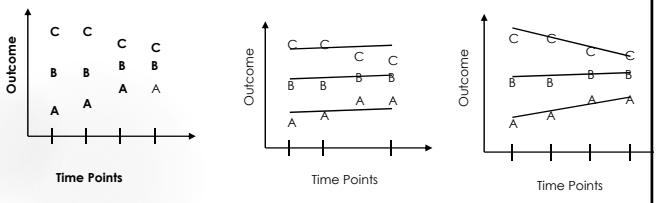
- General linear mixed model, under all of its names, is an extension of the general linear model via random effects
- Appropriate for analyzing data with individuals nested within groups
- ICC, calculable using MIXED and ESTAT ICC, is a measure of the similarity of those in clusters
- Model building must consider the level of predictors
- In practice, the simple level-1 model is often extended, including having slope differences (not just intercept differences) by cluster, and including cluster-level predictors and cross-level interactions among predictors – primary issue is using appropriate degrees of freedom
- Many use "multilevel" regression models to refer to models with cross-level interactions

Mixed Model Regression

Random Slopes Regression

PubH 8342
Mar 3, 2016

Mixed Models Rational



Take a longitudinal study with 3 subjects

Fit a Random intercept model

Let the slopes vary for each cluster

Terminology

- Once again, numerous names for these models
  - Random coefficient model
  - Linear mixed effects model
  - Multilevel model of change
  - Growth curve model
  - Empirical Bayes model
  - Hierarchical linear model

Random Coefficient Model

- Use the multilevel aspect of these models to fit individual change at one level and differences across individuals at another
- Model the intercepts and slopes for each individual subject as *random coefficients*
- Random coefficients assume that these coefficients come from a random sample from some population of possible coefficients
- Use maximum likelihood or restricted maximum likelihood to estimate model parameters (as opposed to least squares)

The Statistical Model

- Specifically,
  - Level 1: $Y_{it} = a_i + b_i (T_t) + e_{it}$
  - Level 2: $a_i = g_{00} + g_{01}(X) + d_i$
    $b_i = g_{10} + g_{11}(X) + d_i$

  where g_{00} is the intercept and g_{01} is the slope describing the association between the predictor X and the intercept of the growth curve and g_{10} is the intercept and g_{11} is the slope describing the association between the predictor X and the slope of the growth curve

  These two levels are combined into one large, multilevel regression model

When do we use this?

- Examples of Research Questions amenable to random coefficient modeling
  - How does alcohol use change over the lifespan?
  - Compared to controls, do participants in a weight-loss intervention program lose more weight over a 24-month period?
  - Are changes in physical activity associated with changes in blood pressure?
  - Are there different trajectories of changes in cholesterol following Lipitor use and, if so, how many?

Growth curve modeling

- There are 2 common and distinct approaches for estimating change over time
  - Random coefficient modeling as implemented by the General Linear Mixed Model uses random coefficients to model change over time
  - Structural Equation Modeling (SEM) uses latent variables to model change over time

Two Conceptual Phases of a Growth Curve Model

- We also split the process of conducting growth curve models into two conceptual phases
  - Phase 1 – Model the "growth" or change over time
  - Phase 2 – Explain variability in change over time by regressing the slope parameter(s) on a set of predictors/covariates

Phase 1: Intra-individual Change

- One of the great strengths of growth curve modeling is that the nature of change over time can be examined, and is not limited to linear change
  - It may be that there is very little change over time
  - It may be that there is change but it is flat over the first couple of time points and then suddenly increases
  - It may be that there is a drop from wave 1 to wave 2, and then a gradual increase (return to baseline) over the next 3 waves
- We can examine these possibilities using model testing and fit criteria

Phase 1: Intra-individual Change

- Step 1: How many time points do you have?
  - The number of time points places an upper limit on the complexity of the line that can be estimated
  - 3 time points → straight line (linear)
  - 4 time points → curving, monotonic line (quadratic)
  - 5 time points → curving, nonmonotonic line (cubic)
- Step 2: Is your outcome (Y) measured the same at each time point?
  - Important to verify that the outcome to which you are fitting a growth curve was measured the same at each time point
  - This is usually not an issue for studies of relatively short duration originally designed as longitudinal trials
  - Does tend to become an issue when the data were not originally collected with the intention of a longitudinal study and/or the time period is 10-20 years and the measures were "improved" over time
  - There are available options when measures change over time

Phase 1: Intra-individual Change

- Step 3: Look at the data
  - Little substitute for descriptive statistics and plots
  - "Connect the dots" plots
- Step 4: Estimate a number of unconditional models
  - Estimate the max number of models given the # of time points
- Step 5: Compare models
  - Multiple criteria including AIC, BIC, SBIC, LR test
  - Also examine model parameters; should be general agreement between model fit criteria and significance of parameters

Phase 1: Intra-individual Change

- Step 6: Select a best-fitting model and interpret
  - Ideally fit criteria will be consistent
  - This model is retained as best describing change in the outcome over time in your sample
  - Plot predicted scores
    - individual growth curves
    - average growth curve

Phase 2: Inter-individual Differences

- The second phase of conducting a growth curve analysis is to attempt to explain *differences* in the intercept and slopes
- Take the best-fitting model from phase 1 and regress the intercept and slope(s) from this model (e.g., a_i and b_i) on a set of potential predictors in another regression model

Phase 2: Inter-individual Differences

- Step 1: Add a number of predictors to the best-fitting model from phase 1
  - Do NOT estimate conditional models for all of the different growth curve models from phase 1 – only the retained model
- Step 2: Trim and/or refine model
  - At this point in the process, the growth curve model is no different than any statistical model you may estimate
  - Specific fields differ in how they prefer to deal with covariates/model fit/model parsimony/ confounders/etc.
  - I would recommend using whatever techniques are considered appropriate in your area for how best to determine what variables to include in the conditional model

Phase 2: Inter-individual Differences

- Step 3: Interpret results
  - For each predictor in the model there will be a regression coefficient associated with each of the growth curve parameters
  - If the model is a linear growth curve, there will be two regression coefficients for each predictor: one associated with the intercept and one associated with the linear slope
  - If the model is a quadratic growth curve, there will be three regression coefficients for each predictor: one associated with the intercept, one associated with the linear slope, and one associated with the quadratic slope
  - It is quite possible that a predictor will be associated with the linear slope, for example, but not with the intercept or quadratic slope

Phase 2: Inter-individual Differences

- Step 4: Plot results
  - As with phase 1 results, plotting of results is often a particularly good way of presenting results
  - For categorical predictors, typically are interested in plotting the average growth curve for each level
  - For example, can use predicted scores from the model to plot the average growth curve for men vs. women, intervention vs. control, etc.
  - For continuous predictors, can re-estimate model using quartiles/quintile/etc. for that variable and create average plots for those categories
  - Margins works pretty well in Stata

Illustrative Example

- Same as Covariance Pattern Model
- N=553
- Measured at 4 times: 14, 15, 16, 17 years old
- Outcome: Smokestage (range 1-6)
- Predictors: Sex & Race (dropped age)
- Research Questions:
  - Is smoking changing over time during adolescence?
  - Are sex and race associated with smoking, both at baseline (age 14) and change in smoking over time (14-17)?

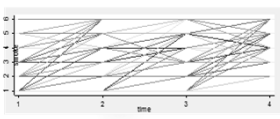
Illustrative Example

- Step 1: How many time points?
  - 4 – we'll limit to quadratic change
- Step 2: Outcome measured the same across time?
  - Yes
- Step 3a: Prep the data

```
gen id = real(mastered)
reshape long smoke, i(id) j(time)
tabulate time, generate(t)
xtset id time, yearly
```

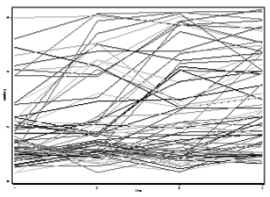
Step 3b: Look at the data

```
. xtline smoke, overlay
```



Select a random subset of your data

```
. xtline smokej if id<1000000, overlay legend(off)
  scheme(s1mono) scale(.5)
```



Step 4: Estimate Models

- Random Intercept, No Time Effect

```
. mixed smoke || id:

------------------------------------------------------------------------------
 Random-effects Parameters |   Estimate   Std. Err.      [95% Conf. Interval]
---------------------------+--------------------------------------------------
id: Identity               |
                 var(_cons)|   1.454083   .1322331      1.216695    1.737788
---------------------------+--------------------------------------------------
                var(Residual)|   .6508333   .0106806      .6303948    .6718316
------------------------------------------------------------------------------
LR test vs. linear regression: chibar2(01) =   719.66 Prob >= chibar2 = 0.0000
```

Random intercept, fixed time then quadratic time

```
. mixed smoke time || id:

. gen time2 = time^2

. mixed smoke time time2 || id:

------------------------------------------------------------------------------
 Random-effects Parameters |   Estimate   Std. Err.      [95% Conf. Interval]
---------------------------+--------------------------------------------------
id: Identity               |
                 var(_cons)|   1.483947   .1321589      1.246267    1.766955
---------------------------+--------------------------------------------------
                var(Residual)|   .5313789   .0250494      .4844827    .5828144
------------------------------------------------------------------------------
LR test vs. linear regression: chibar2(01) =   849.97 Prob >= chibar2 = 0.0000
```

Step 4: Estimate Models

Random Intercept, Random Linear Time

```
. mixed smoke time || id: time

------------------------------------------------------------------------------
 Random-effects Parameters |   Estimate   Std. Err.      [95% Conf. Interval]
---------------------------+--------------------------------------------------
id: Independent            |
                 var(time) |   .1235812   .0138766      .0991684    .1540038
                 var(_cons)|   .8540762   .0988716      .6807033    1.071607
---------------------------+--------------------------------------------------
                var(Residual)|   .315476   .0177432      .2825482    .3522412
------------------------------------------------------------------------------
```

Random Intercept, Random Linear Time

```
. mixed smoke time || id: time

------------------------------------------------------------------------------
 Random-effects Parameters |   Estimate   Std. Err.      [95% Conf. Interval]
---------------------------+--------------------------------------------------
id: Independent            |
                 var(time) |   .1235812   .0138766      .0991684    .1540038
                 var(_cons)|   .8540762   .0988716      .6807033    1.071607
---------------------------+--------------------------------------------------
                var(Residual)|   .315476   .0177432      .2825482    .3522412
------------------------------------------------------------------------------
LR test vs. linear regression:       chi2(2) =  1084.76  Prob > chi2 = 0.0000
```

Random Intercept, Random Linear Time

Random Intercept, Random Linear Time

```
. mixed smoke time || id: time, cov(unstr)

------------------------------------------------------------------------------
 Random-effects Parameters |   Estimate   Std. Err.      [95% Conf. Interval]
---------------------------+--------------------------------------------------
id: Unstructured           |
                 var(time) |   .1341546   .0163711      .1056165    .1704037
                 var(_cons)|   .9298501   .1167093      .7270683    1.189189
          cov(time, _cons) |  -.0456967   .0334615     -.111128     .0198867
---------------------------+--------------------------------------------------
                var(Residual)|   .3078333   .0177728      .2748979    .3447147
------------------------------------------------------------------------------
LR test vs. linear regression:       chi2(3) =  1086.79  Prob > chi2 = 0.0000
```

Random Intercept, Linear Slope, & Quadratic Slope for Time

```
. mixed smoke time timesq || id: time timesq, cov(unstr)

------------------------------------------------------------------------------
 Random-effects Parameters |   Estimate   Std. Err.      [95% Conf. Interval]
---------------------------+--------------------------------------------------
id: Unstructured           |
                 var(time) |   1.025289   .2366657      .6521758    1.611861
              var(timesq) |   .0397993   .0091772      .0253278    .0625391
                 var(_cons)|   .6523312   .2447744      .3126595    1.361021
        cov(time,timesq) |  -.1870182   .045737     -.2766612   -.0973752
           cov(time, _cons) |  -.3246209   .2180787     -.7520474    .1028055
        cov(timesq, _cons) |   .0597963   .0414812     -.0215054    .1410979
---------------------------+--------------------------------------------------
                var(Residual)|   .2281667   .0186301      .1944243    .2677652
------------------------------------------------------------------------------
```

Compare Fit

Model	-2LL	# random parameters
Random I	-1789.4611	2
Random I, Fixed S	-1698.2493	2
Random I, Fixed S & Q	-1698.2109	2
Random I and S	-1579.8103	4
Random I, S, & Q	-1545.3817	7

Conditional Model

- Step 1: Bring in Predictors
- Step 2: Trim/Refine Model

```
. mixed smoke time i.white i.male i.white#c.time i.male#c.time|| id: time, cov(unstr)

------------------------------------------------------------------------------
 Random-effects Parameters |   Estimate   Std. Err.      [95% Conf. Interval]
---------------------------+--------------------------------------------------
id: Unstructured           |
                 var(time) |   .1309002   .0161118      .1028419    .1666135
                 var(_cons)|   .9178713   .1157573      .7168571    1.175252
          cov(time, _cons) |  -.0445466   .0330963     -.1094142    .0203209
---------------------------+--------------------------------------------------
                var(Residual)|   .3078333   .0177728      .2748979    .3447147
------------------------------------------------------------------------------
```

Interpret and Plot

- Step 3: Interpret Results
  - Intercept = predicted smokestage when all predictors = 0 (including Time=0)
  - Time = change in smokestage for each 1-unit increase in Time
  - Main Effects = effect on intereept when predictor = 1 (vs predictor = 0)
  - Time Interactions = difference in slopes (per year) between levels of predictor
- Step 4: Plot Results
  - Margins can help with this

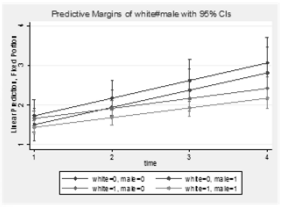
Calculate Margins

```
. margins white#male, over(time)
```

time#white#male		Delta-method Margin	Std. Err.	z	P>|z|	[95% Conf.	Interval]
	1 0 0	1.730661	.2090753	8.28	0.000	1.320881	2.140441
	1 0 1	1.508417	.212966	7.08	0.000	1.091012	1.925823
	1 1 0	1.662579	.0962517	17.27	0.000	1.473929	1.851228
	1 1 1	1.440335	.0863057	16.69	0.000	1.271179	1.609491
	2 0 0	2.176573	.2245564	9.69	0.000	1.736451	2.616696
	2 0 1	1.944218	.2287351	8.50	0.000	1.495905	2.39253
	2 1 0	1.91729	.1033787	18.55	0.000	1.714672	2.119909
	2 1 1	1.684935	.0926962	18.18	0.000	1.503254	1.866616
	3 0 0	2.622485	.2673145	9.81	0.000	2.098558	3.146412
	3 0 1	2.380018	.2722889	8.74	0.000	1.846342	2.913695
	3 1 0	2.172002	.1230632	17.65	0.000	1.930803	2.413201
	3 1 1	1.929535	.1103467	17.49	0.000	1.713259	2.14581
	4 0 0	3.068397	.3268132	9.39	0.000	2.427855	3.70894
	4 0 1	2.815819	.3328949	8.46	0.000	2.163357	3.468281
	4 1 0	2.426714	.1504545	16.13	0.000	2.131828	2.721599
	4 1 1	2.174135	.1349076	16.12	0.000	1.909721	2.438549

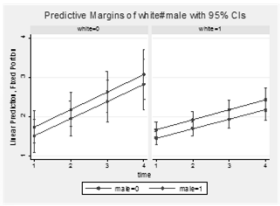
Plot Margins

```
. marginsplot, xdimension(time) plotdimension(white male)
```



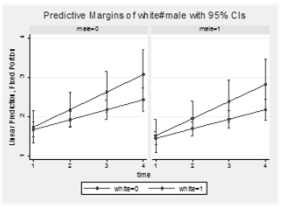
Plot Margins

```
. marginsplot, xdimension(time) plotdimension(male)
```



Plot Margins

```
. marginsplot, xdimension(time) plotdimension(white)
```



Generalized Linear Mixed Models

PubH 8342
March 22, 2016

GeneralizedLMM in STATA

- meglm
  - Defaults to mixed effects general linear model, but we usually use mixed for these models (has a few more capabilities)
  - Can specify a variety of distributions (family) and link functions
  - Also, a lot of the common models have their own statement
    - melogit (family = bernoulli, link = logit)
    - meprobit (family = bernoulli, link = probit)
    - meologit (family = ordinal, link = logit)
    - mepoisson (family = poisson, link = log)
    - menbreg (family = nbinomial, link = log)
- Can also use gsem to fit multilevel generalized regression models using a structural equation modeling framework

Multilevel Logistic Regression

- Similar to Nested Example used earlier
- N=1170 students nested in 39 colleges
- Predictors:
  - Sex and Race (student-level)
  - Public (college-level)
- Outcome: Dichotomized BMI
  - "Normal" – coded 0 – BMI <= 25
  - Overweight/Obese – coded 1 – BMI > 25

Means and Correlations

```
. corr smokec*, means
(obs=300)

    Variable |         Mean     Std. Dev.          Min           Max
-------------+--------------------------------------------------------
     smokec1 |     .0633333     .2439685            0             1
     smokec2 |     .1333333     .3405026            0             1
     smokec3 |     .1566667     .3640935            0             1
     smokec4 |     .2266667     .4193747            0             1


             |  smokec1  smokec2  smokec3  smokec4
-------------+------------------------------------
     smokec1 |   1.0000
     smokec2 |   0.6629   1.0000
     smokec3 |   0.4150   0.5863   1.0000
     smokec4 |   0.3169   0.5605   0.6647   1.0000
```

Print First 20 Obs

```
. list student college male white public bmi overweight in 1/20

     +-----------------------------------------------------------------+
     | student   college   male   white   public         bmi   overwe~t |
     |-----------------------------------------------------------------|
  1. |    4002         4      1       1        0   21.92687484        0 |
  2. |    4003         4      0       1        0   25.10464676        1 |
  3. |    4004         4      0       1        0   22.46494515        0 |
  4. |    4005         4      0       1        0   22.13213115        0 |
  5. |    4006         4      0       1        0   21.79927652        0 |
     |-----------------------------------------------------------------|
  6. |    4007         4      0       1        0   23.77706695        0 |
  7. |    4008         4      0       1        0   21.11268513        0 |
  8. |    4009         4      0       1        0   25.82441064        1 |
  9. |    4010         4      0       1        0   19.76622953        0 |
 10. |    4011         4      0       1        0   22.11215161        0 |
     |-----------------------------------------------------------------|
 11. |    4012         4      0       1        0   27.45715518        1 |
 12. |    4013         4      0       1        0   24.21038497        0 |
 13. |    4014         4      0       1        0   20.46806114        0 |
 14. |    4015         4      0       1        0   27.34159476        1 |
 15. |    4016         4      0       1        0   22.67356881        0 |
     |-----------------------------------------------------------------|
 16. |    4017         4      1       1        0   32.28380723        1 |
 17. |    4018         4      0       1        0   22.80718792        0 |
 18. |    4019         4      1       1        0   23.70994417        0 |
 19. |    4020         4      0       1        0   26.57103476        1 |
 20. |    4021         4      0       1        0   19.76622953        0 |
     +-----------------------------------------------------------------+
```

Calculate Intraclass Correlation

. meglm overweight || college: , family(binomial) link(logit)

. estat icc

OR

. melogit overweight || college:

. estat icc

```
Intraclass correlation
------------------------------------------------------------------------------
                        Level |        ICC   Std. Err.      [95% Conf. Interval]
------------------------------+-----------------------------------------------
                      college |   .0363983   .0185863       .013193    .096431
------------------------------------------------------------------------------
```

Analysis

Ignore Clustering

. melogit overweight i.male i.white

Model Clustering as a Random Effect

. meglm overweight i.male i.white || college: , family(binomial) link(logit)

Calculate Odds Ratios

. melogit overweight i.male i.white || college: , or

Add a Level 2 predictor

. melogit overweight i.male i.white public || college:

Odds ratios

. melogit overweight i.male i.white public || college: , or

Compare a few Models

. estimates table indmodel memodel geemodel, equation(1) se

```
-------------------------------------------------------
    Variable |  indmodel     memodel     geemodel
-------------+-----------------------------------------
#1           |
        male |
           1 |   .88362337   .91747644   .88362337
             |   .13275517   .13791474   .11239004
             |
       white |
           1 |  -.0912401   -.0505846   -.0912401
             |   .14466445   .15506228   .17894548
             |
      public |   .27782041   .28385236   .27782041
             |    .132259    .1849862   .18398093
       _cons |  -1.2634259  -1.3316216  -1.2634259
             |   .15577535   .18826046   .21172928
-------------+-----------------------------------------
var(_cons[~]) |
       _cons |               .1252482
             |               .0687358
-------------------------------------------------------
                                        legend: b/se
```

Longitudinal Logistic Example

- Very similar to CP, Random Coefficient and GEE example
- N=2212
- 4 waves (age 14, 15, 16, 17)
- Predictors: Sex and Race
- Outcome: Dichotomized Smoking
  - Nonsmokers – coded 0 – smokestages 1-3
  - Smokers – coded 1 – smokestages 4-6

Print First 12 Obs

```
. list id time white male smoke smokec in 1/12, sep(4)

     +-----------------------------------------------------+
     | id   time   white   male   smoke   smokec |
     |-----------------------------------------------------|
  1. | 26      1       1      1       1       0 |
  2. | 26      2       1      1       1       0 |
  3. | 26      3       1      1       2       0 |
  4. | 26      4       1      1       2       0 |
     |-----------------------------------------------------|
  5. | 31      1       1      1       1       0 |
  6. | 31      2       1      1       1       0 |
  7. | 31      3       1      1       1       0 |
  8. | 31      4       1      1       1       0 |
     |-----------------------------------------------------|
  9. | 41      1       1      0       2       0 |
 10. | 41      2       1      0       2       0 |
 11. | 41      3       1      0       2       0 |
 12. | 41      4       1      0       5       1 |
     +-----------------------------------------------------+
```

Estimate ICC

```
. meglm smokec || id: , fam(binomial) link(logit)
. estat icc

Intraclass correlation

------------------------------------------------------------------------------
                       Level |        ICC   Std. Err.      [95% Conf. Interval]
-----------------------------+------------------------------------------------
                          id |   .8120036   .0326569      .7395807    .8678835
------------------------------------------------------------------------------
```

Random Intercept Model

```
. meglm smokec i.time i.white i.male i.white#i.time i.male#i.time || id: , fam(binomial) link(logit)

------------------------------------------------------------------------------
      smokec |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        time |
          2  |   3.680046   1.784544     2.06   0.039     .1824036    7.177688
          3  |   6.373001   1.880172     3.22   0.001     2.382052    9.74395
          4  |   6.168935   1.881738     3.28   0.001     2.480797    9.857074
             |
      1.white |   2.104823   1.87396      1.12   0.261    -1.568071    5.777717
      1.male  |  -2.941913   .9313322    -3.16   0.002    -4.76729    -1.116535
              |
white#time    |
        1 2   |  -2.257507   1.806767    -1.25   0.211    -5.798706    1.283691
        1 3   |  -4.015634   1.891405    -2.12   0.034    -7.72272    -.3085498
        1 4   |  -3.117722   1.879233    -1.66   0.097    -6.801892    .5654474
              |
male#time     |
        1 2   |   1.080521   .6572109     1.26   0.207    -.1975813    2.740624
        1 3   |   1.53134    .8755761     1.75   0.080    -.1847379    3.247457
        1 4   |   1.9076     .9029539     2.11   0.035     .1378427    3.677357
              |
        _cons |  -3.62462    2.044085    -4.71   0.000    -13.63095    -5.618287
--------------+----------------------------------------------------------------
id            |
   var(_cons) |   31.43984   7.741859                      19.40338    50.94285
------------------------------------------------------------------------------
LR test vs. logistic regression: chibar2(01) =   448.71 Prob>=chibar2 = 0.0000
```

Random Intercept Model with OR

```
. meglm, or

------------------------------------------------------------------------------
      smokec | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        time |
          2  |   39.64821   70.75398     2.06   0.039     1.200098    1309.976
          3  |   432.9812   837.2412     3.22   0.001     10.82771    17395.21
          4  |   477.4773   898.8436     3.28   0.001     11.95078    19092.94
             |
      1.white |   8.205451   15.37706     1.12   0.261     .2084469    323.0209
      1.male  |   .0527447   .0491415    -3.16   0.002     .0085034    .3274123
              |
white#time    |
        1 2   |   .1041109   .1880074    -1.25   0.211     .0030315    3.609941
        1 3   |   .0180315   .0341049    -2.12   0.034     .0004427    .7345121
        1 4   |   .0442579   .0831495    -1.66   0.097     .0011128    1.760235
              |
male#time     |
        1 2   |   2.946215   2.525527     1.26   0.207     .5490415    15.8097
        1 3   |   4.624466   4.049067     1.75   0.080     .8313221    25.72465
        1 4   |   6.7349     6.08311      2.11   0.035     1.147795    39.54175
              |
        _cons |   .0000661   .0001351    -4.71   0.000     1.20e-06    .0036309
--------------+----------------------------------------------------------------
id            |
   var(_cons) |   31.43984   7.741859                      19.40338    50.94285
------------------------------------------------------------------------------
```

Can also use xtlogit

```
. xtset id time, yearly
. xtlogit smokec i.time i.white i.male i.white#i.time i.male#i.time, re

------------------------------------------------------------------------------
      smokec |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        time |
          2  |   3.111735   1.474609     2.11   0.035     .2215553    6.001915
          3  |   5.278168   1.523407     3.46   0.001     2.292346    8.263991
          4  |   5.405977   1.524271     3.55   0.000     2.418462    8.393493
             |
      1.white |   1.317545   1.538437     0.86   0.392    -1.697727    4.332827
      1.male  |  -2.738538   .8037797    -3.41   0.001    -4.3139     -1.163177
              |
white#time    |
        1 2   |  -1.764872   1.500239    -1.18   0.239    -4.705287    1.173543
        1 3   |  -3.342714   1.53997     -2.17   0.030    -6.361      -.3244292
        1 4   |  -2.542218   1.537574    -1.65   0.098    -5.555821    .4712747
              |
male#time     |
        1 2   |   .9327547   .7723359     1.08   0.281    -.6809958    2.346505
        1 3   |   1.181753   .7743753     1.53   0.127    -.3358746    2.699521
        1 4   |   1.455921   .7859531     1.85   0.064    -.0845186    2.996361
              |
        _cons |  -7.779738   1.549364    -5.02   0.000    -10.81444    -4.74304
--------------+----------------------------------------------------------------
      /lnsig2u |   2.988659   .1427588                      2.708857    3.268461
--------------+----------------------------------------------------------------
      sigma_u |   4.454348   .3180915                      3.874546    5.125513
          rho |   .8578823   .0174052                      .8202452    .8897082
------------------------------------------------------------------------------
```

and xtlogit will also do GEE!!!

```
. xtlogit smokec i.time i.white i.male i.white#i.time i.male#i.time, pa

------------------------------------------------------------------------------
      smokec |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        time |
          2  |   1.249031   .6204414     2.01   0.044     .0338879    2.465973
          3  |   2.064412   .6196381     3.33   0.001     .8519434    3.28688
          4  |   2.134135   .6200449     3.44   0.001     .9188689    3.3494
             |
      1.white |   .4675226   .7323927     0.64   0.524    -.9679007    1.904946
      1.male  |  -1.011971   .4061577    -2.49   0.013    -1.808025    -.2159164
              |
white#time    |
        1 2   |  -.6082766   .6377601    -0.95   0.340    -1.858264    .6417102
        1 3   |  -1.174353   .6367042    -1.85   0.065    -2.424271    .075564
        1 4   |  -.8622918   .6367797    -1.36   0.173    -2.116357    .3917734
              |
male#time     |
        1 2   |   .2931971   .358589      0.82   0.414    -.4095244    .9959186
        1 3   |   .4093424   .353603      1.16   0.247    -.2837669    1.102452
        1 4   |   .4755712   .3519829     1.35   0.177    -.2143026    1.165445
              |
        _cons |  -2.82387    .7153423    -4.00   0.000    -4.265915    -1.461825
------------------------------------------------------------------------------
```

Compare results from the three

```
. estimates table memodel xtmodel geemodel, star equation(1)
```

Variable	memodel	xtmodel	geemodel
#1			
time			
2	3.6800458*	3.1117354*	1.2499307*
3	6.0730012**	5.2781683***	2.0464117***
4	6.1689355**	5.4059776***	2.1341345***
white			
1	2.1048231	1.317545	.46752264
male			
1	-2.9419127**	-2.7385381***	-1.0119709*
white#time			
1 2	-2.2575571	-1.7648723	-.60827443
1 3	-4.0156343*	-3.3427144*	-1.1763533
1 4	-3.1177222	-2.5423179	-.86829174
male#time			
1 2	1.0805212	.83275448	.29319711
1 3	1.5313597	1.1817732	.40914236
1 4	1.9075999*	1.4559211	.47557123
_cons	-9.6246201***	-7.7797383***	-2.84387***
var(_cons[-])			
_cons	31.439837***		
lnsig2u			
_cons		2.988459***	

legend: * p<0.05; ** p<0.01; *** p<0.001

Survival Analysis, pt1

Rich MacLehose, PhD

Time to Event Example

21 individuals were given a new treatment for leukemia; 21 individuals were given a placebo

Patients were followed for up to 35 weeks *or until they relapse* or are lost to follow-up

- Gehan-Freirich survival data
- Analyzed by Cox in original 1972 proportional hazards paper

42.9% of treated relapse; 100% of placebo relapse

mean survival time: 17 weeks for treated; 9 weeks for placebo

How should we analyze these data?

Survival Analysis-When?

Think of survival analysis when

- The event of interest is dichotomous
- There is a time-to-event component that you measured
- The data are censored

Why not use logistic regression? (Loose on time to event data using just the Yes/No outcome)

Why not use linear regression? (By using time to event as a continuous outcome, both cases and controls will get predicted times, but these values would be biased for censored people since we do not know when the outcome happened)

Censoring

Censoring=missing data

You do not know the survival time for some individuals under study

A "complete case" analysis could lead to selection bias

Key Assumption: the censoring time is independent of (potential) survival time

It's necessary to consider the reason a person is missing!

Independent Censoring?

1. The study is designed to end at 35 weeks and everyone still alive at 35 weeks is considered censored
2. People who are very ill have trouble coming into the clinic to be monitored and are more likely to be lost to followup
3. People who die of car accidents are censored
4. People who die as a result of treatment (but not the disease of interest) are censored
5. People who die of heart attacks are censored

Censoring

Types of Censoring

- Right censoring: Person is lost to followup before they have the event of interest
- Left censoring: Event occurs before followup begins
- Interval censoring: Event occurs between two time points (clinical exams, perhaps)

Coping with Censoring

- Right censoring: This is what survival analysis is built for
- Interval censoring: Typically, dealing with this involves discretizing follow-up time into blocks
- Left censoring: Typically more difficult and ignored

Truncation

- Incomplete data due to study design issues
- Left truncation: delayed entry into study
  - Example: association between smoking and age at menopause
  - Start survival time at age 35
  - A woman who tries to enroll at age 40 has truncated person time from 35-40
  - Account for this in analysis (easy)
- Right truncation: only subjects with event are included in study
  - Cancer registries: to be in the registry, you had to get cancer

Life Tables

Many modern survival analysis techniques are extensions of life tables

in the same way that logistic regression is an extension of contingency tables

- Life tables are still very useful (and underused)
- Life tables estimate the cumulative probability of survival past some time
- Kaplan-Meier estimates of survival are based on life-tables
  - KM: exact follow-up times
  - Life tables: categorical follow up times

Survival Estimators

If T is the RV: time to event

- Conditional Rate $\Pr(t \le T < t+\Delta \mid T \ge t)/\Delta$
- Conditional Survival $\Pr(T > t \mid T \ge t)$
- Cumulative Survival (Kaplan Meier) $S(t) = \Pr(T > t) = \prod_{i=1}^{t} \Pr(T > i \mid T \ge i)$

Computing Life Table Estimates

The number of people who start time period t

The number of people who die in time period t

The person-time accrued in time period t

n_t $\Pr(t \le T < t+\Delta \mid T \ge t)/\Delta = d_t/(n_t \times t_t)$

d_t $\Pr(T > t \mid T \ge t) = (n_t - d_t)/n_t$

t_t $\Pr(T > t) = \prod \Pr(T > i \mid T \ge i)$

Life Table (for the treated)

Time (week)	N Beginning	Deaths	Lost	Conditional Rate	Conditional Survival	Cumulative Survival
6-7	21	4	1	4/21	17/21	17/21
8-9	16	0	1	0/16	16/16	16/16 * 17/21
10-11	15	1	2	1/15	14/15	14/15* 16/16 * 17/21
12-13	12	1	0	1/12	11/12	11/12* 14/15* 16/16 * 17/21
etc						

Why not just: 11/21? Because it ignores censoring

Life Tables con't

- Can be calculated different ways
  - Allow censored people to contribute half the survival time, for example
  - Stata: example!
- Standard errors calculated via Greenwood formula
  - Can underestimate variance
  - Bootstrap

Survival Analysis, pt2

Rich MacLehose, PhD

Cox PH Models

- Cox, 1972, proposed semi-parametric proportional hazards models
- Overwhelming choice for survival analysis
- Richard Peto (in a written discussion on the paper): "…it is very pretty"
- Cox in a later article suggested that parametric survival models were underutilized

Cox PH Models

- Cox formulated the PH model in terms of relative hazards

$$h(t \mid x_1,....,x_k) = h_0(t)\exp(\beta_1 x_1 + \cdots + \beta_k x_k)$$

- Effects are ratios of hazards:

$$HR = \frac{h(t \mid x_1 = 1, x_2 = a)}{h(t \mid x_1 = 0, x_2 = a)} = \frac{h_0(t)\exp(\beta_1 + \beta_2 a)}{h_0(t)\exp(\beta_2 a)} = \exp(\beta_1)$$

- Cox realized most people would only want to estimate the HR's, which don't depend on the baseline hazard

Partial Likelihood

All regression until now is based on the likelihood: The 'probability' of the data, given the parameters

Partial likelihood is calculated at each failure time: the hazard that person j died at this time from among those still at risk

$$\frac{h_0(t)\exp(\beta x_j)}{\sum_{i \in R_t} h_0(t)\exp(\beta x_i)} = \frac{\exp(\beta x_j)}{\sum_{i \in R_t} \exp(\beta x_i)}$$

This is exactly the same way that conditional logistic regression models for 1:n matched data are calculated

Cox PH Model: Key Assumptions

- Independent observations
- Independent censoring
- Hazard in the exposed is proportional to the hazard in the unexposed and that proportion is constant over time

Interpretation

stcox treatment

failure _d: failure

analysis time _t: time

- The hazard of relapse among the untreated is 4.5 times the hazard of relapse among the treated.
  - ↖at all times during follow-up

Assessing Proportionality

1. Visual inspection
2. Goodness of fit tests
3. Including interactions with time

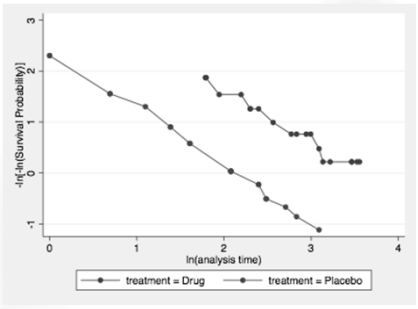
1. Visual Inspection

Compare the KM (or regression based) estimate of S(t) for the exposed and unexposed over time

Cox PH model implies:

$$S(t,X) = S_0(t)^{\exp(\beta x)}$$
$$\ln[S(t,X)] = \ln[S_0(t)]\exp(\beta x)$$
$$-\ln[S(t,X)] = -\ln[S_0(t)]\exp(\beta x)$$
$$\ln\{-\ln[S(t,X)]\} = \ln\{-\ln[S_0(t)]\} + (\beta x)$$

- ln(-ln(S)) in the exposed should be beta units higher than ln(-ln(baseline)) at all times




```
stphplot, by(treatment)
```

2. Goodness of Fit Test

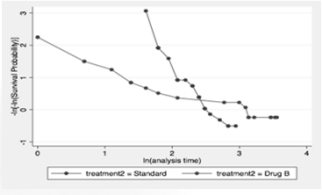
Uses residuals

If PH assumption is met, residuals should be independent of time

The p-value gives a somewhat easier interpretation than the subjective visual inspection

- stcox treatment
- estat phtest

3. Interactions with Time




```
. stcox treatment, tvc(treatment)
texp(ln(_t))

         failure _d:  failure
   analysis time _t:  time
```

```
------------------------------------------------------------------------------
             _t | Haz. Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
----------------+-------------------------------------------------------------
main            |
   treatment    |   9.242166   12.13469     1.69   0.090     .7049706    121.1648
----------------+-------------------------------------------------------------
tvc             |
   treatment    |    .7168046   .4074961    -0.59   0.558     .2352347    2.184239
------------------------------------------------------------------------------
Note: variables in tvc equation interacted with ln(_t)
```

Survival Analysis, pt3

Rich MacLehose, PhD

Interactions with time

We've seen this when we tested for non proportional hazards

$$h(t,x_1) = h_0(t)\exp(\beta_1 x_1 + \theta x_1 \times t)$$

$$HR(t) = \frac{h(t,x_1=1)}{h(t,x_1=0)} = \frac{h_0(t)\exp(\beta_1 + \theta \times t)}{h_0(t)} = \exp(\beta_1 + \theta \times t)$$

Interactions with time

- This is called the Extended Cox Model
- We now have HR(t) rather than HR
- How do we calculate HR(t)?
- Ln(time) or time?

Time Varying Coefficients

Why would hazards be non-proportional?

What if we find evidence of non-proportional hazards?

- First, non PH in confounders (ancillary variables)
- Second, non pH in main effect

Stratification

An easy approach to deal with non-proportional hazards for confounders

Extends the normal Cox model by allowing a different baseline hazard for each stratum

$$h_{1,s}(t,x_1,\ldots,x_k) = h_{0,s}(t)\exp(\beta_1 x_1 + \cdots + \beta_k x_k)$$

- S=1 does not need to be proportion to S=0 anymore
- Cannot estimate a HR for S

Stratification

$$h_{1,s}(t,x_1,\ldots,x_k) = h_{0,s}(t)\exp(\beta_1 x_1 + \cdots + \beta_k x_k)$$

The hazards for X1 … Xk must still be proportional

The model assumes that the effect of X1…Xk is the same for all strata

- No interaction assumption (this can be relaxed)
- The no interaction assumption can be tested by adding a strata*time term in the model

Time Varying Coefficients

What if our main effect is the variable that exhibits a departure for proportionality?

Is stratification a good option?

Lets look at an example:

- Similar Leukemia data but with a different treatment
- WBC count is a confounder

New Dataset

- Leukemia.dta
- Different treatment
- WBC count is a confounder
- In-Class project: assess PH assumption

Modeling Time Varying Effects

We have a few options here

- Model the interaction explicitly: covariate*time interaction
- categorize followup time and allow different HR's in each time category

Categorizing Follow-up Time

A different HR at each time may be unappealing

The data may suggest the HR is constant within time segments

- For instance: HR(0<t<8) and HR(8<=t<35)
- For this, we need to "split" the data
  - manually (stsplit), or let stata do it

Time Varying Exposures

Its not uncommon to have exposure that vary over time

- People transition from unexposed to exposed, for example

Consider the "Stanford Heart Transplant" example

- Individuals are deemed eligible for a heart transplant
- Wait until suitable donor (unexposed)
- have transplant (exposed)
- Monitor time to death (before or after transplant)

Treat follow-up as an "exposed" and an "unexposed" period

id	Transplant	Wait	Died	stime	Year
1	1	10	1	40	70
2	0	20	1	20	74

id	Transplant	Wait	Died	stime	Year	_t0	_t
1	1	10	1	40	70	0	40
2	0	20	1	20	74	0	20

id	Transplant	Wait	Died	stime	Year	_t0	_t
1	0	10	0	40	70	0	10
1	1	10	1	40	70	10	40
2	0	20	1	20	74	0	20

```
. stset stime, f(died) id(id)

                id:  id
     failure event:  died != 0 & died < .
obs. time interval:  (stime[_n-1], stime]
 exit on or before:  failure

--------------------------------------------------------------------------
      103  total observations
        0  exclusions
--------------------------------------------------------------------------
      103  observations remaining, representing
      103  subjects
       75  failures in single-failure-per-subject data
  31938.1  total analysis time at risk and under observation
                                        at risk from t =         0
                              earliest observed entry t =         0
                                   last observed exit t =      1799

. list id _* wait age if  id==50

     +----------------------------------------+
     | id   _st   _d    _t   _t0   wait   age |
     |----------------------------------------|
 93. | 50     1    1   979     0     83    45 |
     +----------------------------------------+
```

```
. stsplit post,at(0) after(wait)
(69 observations (episodes) created)

. *recode the new variable "post" so 1 is post transplant
. recode post -1=0 0=1
(post: 172 changes made)

. *Wait is coded as 0 if the person never got a tranplant. this fixes it
. replace post=0 if wait==0
(34 real changes made)

. list id _* wait age if  id==50

 +-----------------------------------------------------------------+
 | stime   wait   died   transp~t    _st    _d     _t    _t0   post |
 |-----------------------------------------------------------------|
 |    83     83      .          1     1     0     83     0      0 |
 |   979     83      1          1     1     1    979    83      1 |
 +-----------------------------------------------------------------+
```

```
.  stcox post

        failure _d:  died
  analysis time _t:  stime
                id:  id

Iteration 0:   log likelihood = -298.31514
Iteration 1:   log likelihood = -298.25194
Iteration 2:   log likelihood = -298.25193
Refining estimates:
Iteration 0:   log likelihood = -298.25193

Cox regression -- Breslow method for ties

No. of subjects =          103                     Number of obs   =       172
No. of failures =           75
Time at risk    =      31938.1
                                                   LR chi2(1)      =      0.13
Log likelihood  =   -298.25193                     Prob > chi2     =    0.7222

------------------------------------------------------------------------------
          _t | Haz. Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        post |   1.111523    .331695     0.35   0.723     .6193088    1.994939
------------------------------------------------------------------------------
```

Principal Component Analysis

PUBH 8342

APRIL 7, 2016

Principal Components Analysis

- Often have variables in our dataset that are related to one another
- This relation leads to redundancy, and it may be beneficial to remove this redundancy by creating a small number of indexes that account for most of the variance in the observed variables
- These new indexes can then be used in a model, as either predictors or an outcome

Principal Components Analysis

- PCA finds patterns in the data (variances/covariances)
- It is hard to find patterns using traditional or informal techniques (e.g., graphical) when the data are highly dimensional
- PCA is an analytic tool for identifying these patterns and then compressing the data from the larger number of dimensions to this smaller number of dimensions, without losing much information

Principal Components Analysis

- PCA will derive a small number of linear combinations (aka principal components) of a set of variables that retains a maximum amount of information in those original variables
- Each variable gets a weight that can be used to develop an overall score for that component
- Good technique when the goal is to develop a linear combination and/or data reduction

Principal Components Analysis

$C_1 = b_{11} (X_1) + b_{12} (X_2) + \ldots. b_{1p} (X_p)$

C_1 = score on principal component 1

b_{1p} = weight (regression coefficient) for observed variable p

X_p = score on observed variable p

Some Definitions

- PCA uses some mathematical terminology that is popular in linear algebra/matrix algebra
- Eigen – German. Translates to "unique" in English.
- Eigenvector - A column (vector) of numbers for each component. Each number in the column is the coefficient/weight for that component for that variable.
- Eigenvalue – A single number associated with each eigenvector. It is the amount of variance in the variables explained by that component.
- Note: each PCA has the same number of eigenvectors and eigenvalues, with the maximum equal to the number of variables

Criteria to Determine #of Components

- Eigenvalues > 1
- Scree test
- Proportion of variance accounted for
- Interpretability
  - At least 3 variables per component
  - Face validity
  - Discriminate validity
  - Simple structure

Steps in PCA

- Initial extraction
- Determine number of components
- Interpret & label components (possibly using rotation)
- Create factor/component scores

PCA Example

- Data from the MACC study examining tobacco use among youth
- Restrict the analysis to baseline, with ~4200 youth ages 12-16
- Asked respondents to report how difficult it is for youth to smoke (1=Not at all difficult, 5=Very difficult) in 10 different areas:
  - •On School Property
  - •In Restaurants
  - •In Malls
  - •In Pool Halls, Arcades, Bowling Alleys
  - •In Parks/Playgrounds
  - •In Your Home
  - •In Best Friend's Home
  - •In Coffee House
  - •In Teen Dance Club
  - •During School Hours

Descriptives


```
. corr v*, means
(obs=3724)

             | v_onsc~p v_inre~t v_insh~l v_inpo~l v_inpa~g v_inyo~e v_inbf~e v_inco~p v_inte~b v_duri~l
-------------+---------------------------------------------------------------------------------------
v_onschool~p |   1.0000
v_inrestau~t |   0.1986   1.0000
v_inshopmall |   0.0905   0.2291   1.0000
v_inpoolhall |   0.1625   0.3179   0.2051   1.0000
v_inparkpl~g |   0.1364   0.1174   0.0478   0.2798   1.0000
v_inyourhome |   0.1528   0.1321   0.0295   0.0185  -0.0315   1.0000
  v_inbfhome |   0.1783   0.1341   0.0260   0.0655   0.0393   0.4900   1.0000
v_incoffee~p |   0.1440   0.4121   0.2127   0.3010   0.1053   0.2085   0.2493   1.0000
v_inteenclub |   0.1337   0.1686   0.2035   0.2849   0.1603   0.0509   0.0873   0.2439   1.0000
v_duringsc~l |   0.4573   0.1365   0.0605   0.1830   0.1944   0.0998   0.1791   0.1515   0.1320   1.0000
```

Initial Extraction for Eigen Values

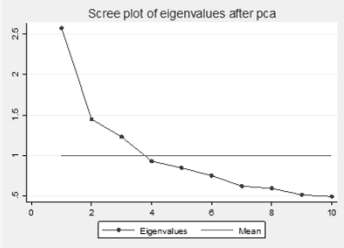

```
. pca v*

Principal components/correlation                  Number of obs    =       3724
                                                  Number of comp.  =         10
                                                  Trace            =         10
Rotation: (unrotated = principal)                 Rho              =     1.0000

--------------------------------------------------------------------------
    Component |   Eigenvalue   Difference    Proportion   Cumulative
--------------+-----------------------------------------------------------
      Comp1   |      2.57392     1.12648         0.2575       0.2575
      Comp2   |      1.44844     .219107         0.1448       0.4023
      Comp3   |      1.22933     .299677         0.1229       0.5253
      Comp4   |      .929656     .0821615        0.0930       0.6182
      Comp5   |      .847495     .0962546        0.0847       0.7030
      Comp6   |      .75124      .130856         0.0751       0.7781
      Comp7   |      .620344     .0265047        0.0620       0.8401
      Comp8   |      .59338      .0888427        0.0594       0.8995
      Comp9   |      .513037     .0214243        0.0513       0.9508
      Comp10  |      .491613     .               0.0492       1.0000
--------------------------------------------------------------------------
```

Scree Test


```
. screeplot, mean
```



Keep 3 Components


```
. pca v*, comp(3)

Principal components/correlation                  Number of obs    =       3724
                                                  Number of comp.  =          3
                                                  Trace            =         10
Rotation: (unrotated = principal)                 Rho              =     0.5253

Principal components (eigenvectors)

-------------------------------------------------------------------------------
    Variable |    Comp1     Comp2     Comp3 | Unexplained
-------------+-----------------------------+------------
 v_onschool~p |   0.3116    0.1189    0.4968 |        .393
 v_inrestau~t |   0.3792   -0.1149   -0.2581 |       .5287
 v_inshopmall |   0.2443   -0.2472   -0.3089 |       .6398
 v_inpoolhall |   0.3674   -0.1405   -0.0184 |       .4826
 v_inparkpl~g |   0.2296   -0.2781    0.2111 |       .6255
 v_inyourhome |   0.2412    0.5907   -0.1782 |       .3057
   v_inbfhome |   0.2872    0.5538   -0.0919 |       .3331
 v_incoffee~p |   0.4034   -0.0001   -0.3335 |       .4442
 v_inteenclub |   0.3031   -0.2385   -0.0995 |       .6685
 v_duringsc~l |   0.3217    0.0759    0.5698 |       .326
-------------------------------------------------------------------------------
```

Factor Rotation

- It can sometimes be easier to interpret the factors/components if they are 'rotated' – redistribute common variance across factors/components
- Pedhazur – 'While the results of a factor analysis may produce a good fitting solution it is not necessarily susceptible to a meaningful interpretation. It is in attempts to improve the interpretability of results that factors are rotated.'
- Stata – 'Rotating principal components is a disputed issue and one in which reasonable people may disagree.'
- MANY different rotation methods available, some orthogonal (i.e., right angles; components still uncorrelated) and some oblique (i.e., askewed angles; components can be correlated).

Rotated Solution


```
. rotate

Rotated components

-------------------------------------------------------------------------------
    Variable |    Comp1     Comp2     Comp3 | Unexplained
-------------+-----------------------------+------------
 v_onschool~p |  -0.0152    0.6016    0.0935 |        .393
 v_inrestau~t |   0.4609   -0.0309    0.1012 |       .5287
 v_inshopmall |   0.4374   -0.1510   -0.0518 |       .6398
 v_inpoolhall |   0.4540    0.1403   -0.1632 |       .4826
 v_inparkpl~g |   0.1592    0.3799   -0.2519 |       .6255
 v_inyourhome |   0.0055   -0.0003    0.6625 |       .3057
   v_inbfhome |   0.0193    0.0955    0.6229 |       .3331
 v_incoffee~p |   0.4624   -0.0776    0.2326 |       .4442
 v_inteenclub |   0.3860    0.0599   -0.0787 |       .6685
 v_duringsc~l |  -0.0363    0.6570    0.0321 |       .326
-------------------------------------------------------------------------------
```

Rotated Solution


```
. estat rotatecompare

Component loadings
```

Variable	Rotated Comp1	Rotated Comp2	Rotated Comp3	Unrotated Comp1	Unrotated Comp2	Unrotated Comp3
v_onschool~p	-0.0152	0.6016	0.0935	0.3316	0.1189	0.4968
v_inrestau~t	0.4609	-0.0309	0.1012	0.3792	-0.1149	-0.2581
v_inshopmall	0.4374	-0.1510	-0.0518	0.2443	-0.2472	-0.3099
v_inpoolhall	0.4540	0.1403	-0.1632	0.3674	-0.3405	-0.0384
v_inparkpl~g	0.1592	0.3799	-0.2519	0.2296	-0.2781	0.3211
v_inyourhome	0.0055	-0.0003	0.6625	0.2412	0.5907	-0.1782
v_inbfhome	0.0193	0.0955	0.6229	0.2872	0.5538	-0.0919
v_incoffee~p	0.4624	-0.0776	0.2326	0.4034	-0.0001	-0.3335
v_inteenclub	0.3860	0.0599	-0.0787	0.3033	-0.2385	-0.0995
v_duringsc~l	-0.0363	0.6570	0.0321	0.3217	0.0759	0.5698

Create Component Scores


```
. predict pc1 pc2 pc3, score

Scoring coefficients
  sum of squares(column-loading) = 1
```

Variable	Comp1	Comp2	Comp3
v_onschool~p	0.3316	0.1189	0.4968
v_inrestau~t	0.3792	-0.1149	-0.2581
v_inshopmall	0.2443	-0.2472	-0.3099
v_inpoolhall	0.3674	-0.3405	-0.0384
v_inparkpl~g	0.2296	-0.2781	0.3211
v_inyourhome	0.2412	0.5907	-0.1782
v_inbfhome	0.2872	0.5538	-0.0919
v_incoffee~p	0.4034	-0.0001	-0.3335
v_inteenclub	0.3033	-0.2385	-0.0995
v_duringsc~l	0.3217	0.0759	0.5698

Component Scores


```
. list v* pc* in 1/10, compress table
```

	v_o~p	v_i~t	~mall	~ball	v_i~g	v~y~e	v~b~e	v_i~p	v_i~b	v_d~l	pc1	pc2	pc3
1.	3	3	3	2	3	5	4	4	2	3	-.4160932	.2891568	-.1752018
2.	4	5	5	5	2	5	5	5	3	3	2.271205	-.4745899	-1.485015
3.	1	3	5	5	1	4	1	5	.	1	.	.	.
4.	5	5	5	3	1	1	1	5	5	5	.9082945	-3.675824	.2631713
5.	3	2	3	3	2	4	4	3	5	4	-.2804302	-.6420401	.3612112
6.	3	3	1	1	1	5	4	5	1	1	-1.804974	1.480404	-1.281211
7.	3	2	5	1	1	1	1	5	1	5	-1.949998	-2.342704	.4730589
8.	3	5	4	2	1	5	5	4	2	4	.4198127	.899347	-1.025199
9.	2	3	5	3	1	4	4	.	1	4	.	.	.
10.	4	3	4	2	2	5	5	5	1	5	.64836	1.158134	.2148703

Component Scores


```
. corr pc*, means
(obs=3724)
```

Variable	Mean	Std. Dev.	Min	Max
pc1	-1.84e-10	1.604656	-5.390158	4.034333
pc2	5.84e-10	1.203512	-5.195009	2.496468
pc3	-5.17e-11	1.108753	-3.85657	3.356079

	pc1	pc2	pc3
pc1	1.0000		
pc2	-0.0000	1.0000	
pc3	0.0000	0.0000	1.0000

Principal Components Summary

- PCA is a statistical technique for creating weighted indexes or components from a set of observed variables
- PCA uses all of the variability in the items, not just the variability they share
- This technique is good for data reduction, showing how a smaller number of components will typically explain a fairly substantial portion of the variability in the items

Common Factor Analysis/ Exploratory Factor Analysis

- Sometimes we are only interested in using the variability that an item shares with the other items as opposed to the total variability – this is what primarily distinguishes common factors from principal components
- When the correlation between items is quite high, the difference between the total variability and the shared variability is quite small and the two approaches will yield very similar results
- However, if the overall correlation between items (as measured by the squared multiple correlation) is low, the two approaches can give fairly different answers

Common Factor Analysis

- In SEM parlance, the factor is a latent variable
- Almost all SEM techniques are confirmatory in nature; the user specifies a model, estimates this model, and then determines how well it 'fits' the data
- EFA is a precursor to this approach, developed as an exploratory technique, not requiring a user-specified model
- Many SEM users see this as a good first-pass technique, but has limitations that need to be understood when interpreting results

Latent Class Analysis

PUBH 8342

APRIL 14, 2016

Latent Variable

- "variables that are not directly observed but are rather inferred from other variables that are observed and directly measured"
- typically refer to observed variables as "manifest"
- more than one type of latent variable
  - can be continuously distributed (e.g., factor) or categorical (e.g., class) in nature
  - can represent theoretical constructs of interest or methodological variability

Numerous Types of Latent Variables

- Factor Analysis: "Intelligence"
- Latent Class: "Chronic Back Pain"
- Random Effect: Intercept, Slope (Growth Curve)
- Residual or error
- Variance Component: Gene vs. Environment, multi-trait multi-method
- Finite Mixtures: homogeneous subpopulations
- Missing Data: CACE models

LCA

- Latent variable that represents distinct subgroups or subtypes of individuals based on a set of observed indicators
- Unobserved (i.e., latent) groups (i.e., classes)
- Mutually exclusive and exhaustive classes
- Aka
  - Unobserved heterogeneity
  - Finite mixture model
- Siblings
  - Hierarchical cluster analysis
  - K-means clustering

Things we care about

- Class prevalences
  - Gamma
  - Proportion of sample in each class

$$\sum_{c=1}^{C} \gamma_c = 1$$

- Item-response probabilities
  - Rho
  - Proportion in each class endorsing item

$$\sum_{r_j=1}^{R_j} \rho_{j,r_j|c} = 1$$

Steps

- Estimate a number of models with varying numbers of classes (1 to ?? latent classes)
- Pick the model that best measures the latent variable
  - AIC/BIC
  - LR test/G^2
  - Entropy
  - Homogeneity
  - Separation
  - Interpretability
- Confirm lowest likelihood using numerous seeds
- Introduce structural components

Software

- Methodology Center
  - Latent GOLD
    - SAS PROC LCA
    - Stata plugin
- Mplus
- R

LCA Example

- Data come from the baseline survey of the MACC study
- N=1697 15-16 year olds
- Numerous questions about tobacco use and related behaviors
- We want a measure of 'tobacco use'
- Theory suggests this is a discrete or stage-based construct

LCA Example

- 10 dichotomous indicators (1=no, 2=yes)
  - Ever smoked a whole cigarette
  - Smoked in the last 6 months
  - Smoked in the previous month
  - Smoked today
  - Smoked first cigarette before age 14
  - Want to stop smoking
  - Have tried to quit
  - Feel you are addicted
  - Ever smoked a cigar
  - Ever chewed

LCA command

- Plugin for Stata (and add-on for SAS: proc lca)
- developed by Collins, Lanza and colleagues at the Methodology Center at Penn State
- Download from: http://methodology.psu.edu/downloads/proclcalta
  - Free, but need to register/provide email
  - 32 and 64 bit versions; install with IC a little dicey
- Collins and Lanza also wrote an excellent LCA and LTA textbook

Analysis

```
Look at the data
. list in 1/10
Look at freqencies of 'patterns'
. groups eversmok today month sixmonth wantstop tryquit chew cigar addicted firsteig, order(high) show(percent)
Create a one class Model
. dol.LCA eversmok today month sixmonth wantstop tryquit chew cigar addicted firsteig, ///
    nclass(1) ///
        seed(100001) ///
        seeddraws(100000) ///
        categories(2 2 2 2 2 2 2 2 2 2) ///
        criterion(0.000001) ///
        rhoprior(1.0)
Create a 2 class model
. dol.LCA eversmok today month sixmonth wantstop tryquit chew cigar addicted firsteig, ///
    nclass(2) ///
        seed(100001) ///                                    . return list
        seeddraws(100000) ///
        categories(2 2 2 2 2 2 2 2 2 2) ///                 . matrix list r(gamma)
        criterion(0.000001) ///
        rhoprior(1.0)                                       . matrix list r(rho)
```

Confirmatory Factor Analysis

PUBH 8342

APRIL 19, 2016

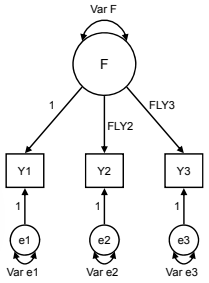
Traditionally LVs used to Measure Unobservable Constructs

- Exploratory Factor Analysis
  - Related to Principal Components Analysis
  - Determine the number of "Factors" underlying a set of variables or items
- Confirmatory Factor Analysis
  - Test whether one factor adequately describes the correlation between a set of items
  - Factor is a theoretically error-free measure of an unobservable construct
- Structural Equation Modeling
  - Path analysis with factors/latent variables

Reflective vs Formative Factors

- Formative Factor
  - A composite variable that summarizes the common variation in a collection of indicators
  - The indicator variables 'cause' the composite variable
  - Examples: SES, life stress, deviance???
  - Manipulating an indicator affects the factor – e.g., education and SES
- Reflective Factor
  - A latent or directly unobserved variable that 'causes' the responses to the indicators
  - This common cause makes the indicators correlated
  - Examples: Depression, intelligence
  - Manipulating an indicator does NOT affect the factor – e.g., sleep and depression

Confirmatory Factor Analysis



SEM "Formula"

- Unlike OLS regression, structural equation models involve solving numerous simultaneous equations (specifically, k(k-1)/n)
- Direct algebraic solution is typically not feasible, so an iterative, "trial and error" estimation technique is used, typically maximum likelihood
- Each estimated parameter in the model is given a start value (chosen either by you or the software) and an estimated variance/covariance matrix based on those values is computed
- One or more of these start values is changed in a certain direction and a second estimated variance/covariance matrix is computed
- The two estimated matrices are compared to the observed variance/covariance matrix and the model with results closer to the observed data is retained
- This process is repeated until additional changes to the parameter estimates fails to bring the estimated matrix closer to the observed matrix using a set *convergence criterion*
- This final model is said to have "converged"

Goodness of Fit

- When DF > 0, model is said to be *over-identified*
- Using *fewer* parameters to estimate *more* informations
- If you use only 8 parameters to explain 10 unique informations, it is unlikely that you will be able to perfectly recreate the data
- The difference between the estimated variance/covariance matrix and the observed variance/covariance matrix is distributed roughly as a c^2 with concordant df
- A good-fitting model would have a NON-SIGNIFICANT or lower c^2 (difference between observed and expected is small)

Goodness of Fit

- Since c^2 is often significant with large samples (and SEM is a large sample technique), have developed a number of Fit indices as alternatives
  - GFI, CFI, NFI
  - RMSEA
  - AIC, BIC, SBIC
- Also, these allow one to compare non-nested models

CFA Example

- Data come from the 2001 survey of the College Alcohol Study
- N=10,000+ college students (from ~120 colleges)
- Numerous questions about alcohol use and consequences
- We want a measure of 'overall alcohol use' or 'problem alcohol use'
- 5 indicators: frequency of drinking, quantity of drinking, binge drinking occasions, self-described drinking pattern, and frequency of hangover

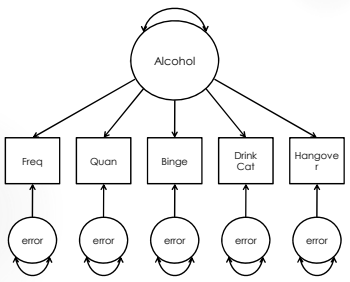
CFA

- SEM added to STATA in version 12
- Can enter the code directly like regular stata models
- Can also draw the picture and stata will generate code
  - This can be particularly helpful with large and complex SEM models
  - Also can output values back to diagram
- Includes generalized SEM (gsem)
- Although added only recently, very good implementation

CFA

```
. sem (quan freq drinkc binge hungover <- Alcohol)
Endogenous variables
Measurement:  quan freq drinkc binge hungover
Exogenous variables
Latent:       Alcohol
Standardized Solution
. sem, standardized
Goodness of fit
. estat gof, stats(all)
Output predicted factor scores
. predict predF1, latent(Alcohol)
. list in 1/15
```

RAM Diagram



STATA COMMANDS REVIEW

```
*PUBH 8342 Final

* Linking the Folder

cd "C:\Users\guillaumeo\Desktop\PUBH 8342 Final"
* Opening the data folder through stata  using the dir command

dir

*Part 1: looking at change in rates over time between races
insheet using Rate_Years_State_Race.csv, clear

describe
list in 1/10

gen logpop = log(population)
encode state, generate(stateID)
encode race, generate(race_category)
describe
tabulate race_category, nolabel // category 4 represents caucasians

tab race_category

*some descriptive statistics
* Summarizing TGCT across each year
tabstat count, by(state) stat(n mean sd min p25 p50 p75 max)
tabstat count year, by(state) stat(n mean sd min p25 p50 p75 max)
tabstat count, by(race) stat(n)
```

*Looking at our data

glm count, family(poisson) link(log) offset(logpop)

glm, eform

*Plotting change of TGCT Rate over time among races

xtset stateID //this gives you the cluster (called panel variable in stata) variable, then the timevariable so you can use them in all subsequent xt commands

xtgee count year ib4.race_category, family(poisson) link(log) corr(exchangeable) offset(logpop) vce(robust) eform

estat wcorr

meglm count year ib4.race_category, offset(logpop) || stateID:, family(poisson) link(log) eform

margins race_category, over(year)

marginsplot, xdimension(year) plotdimension (race_category)

*Modeling increase in TGCT rates by race

meglm count year if race_category==1, offset(logpop) || stateID:, family(poisson) link(log) eform

meglm count year if race_category==2, offset(logpop) || stateID:, family(poisson) link(log) eform

meglm count year if race_category==3, offset(logpop) || stateID:, family(poisson) link(log) eform

meglm count year if race_category==4, offset(logpop) || stateID:, family(poisson) link(log) eform

*Making a graph for the previous models

meglm count year ib4.race_category c.year##ib4.race_category, offset(logpop) || stateID:, family(poisson) link(log) eform

margins race_category, over(year)

marginsplot, xdimension(year) plotdimension (race_category)

*Part 2: looking at the effect of pesticide exposure on state levels of TGCT

* Importing a csv spreadsheet/dataset into stata using the "insheet" command

* Use the comma to indicate a command option: in this case, clear tells stat to remove the previous dataset because stata can only have one dataset open at the time

insheet using high_Pesticide_Exposure.csv, clear

* Using the describe command to see the number of variables, variable type and number of observations

describe

list in 1/10

gen logpop = log(population)

encode state, generate(stateID)

encode race, generate(race_category)

graph matrix mean*

if meanatrazine == . then delete

if meanacetochlor == . then delete

if meanglyphosate == . then delete

if meanmetam == . then delete

if meanmetolachlor == . then delete

g mean_atrazine= round(meanatrazine,1)

g mean_acetochlor= round(meanacetochlor,1)

g mean_glyphosate= round(meanglyphosate,1)

g mean_metam= round(meanmetam,1)

```
g mean_metolachlor= round(meanmetolachlor,1)

egen mean_atrazine_quart = cut(mean_atrazine), group(4)

egen mean_acetochlor_quart = cut(mean_acetochlor), group(4)

egen mean_glyphosate_quart = cut(mean_glyphosate), group(4)

egen mean_metam_quart = cut(mean_metam), group(4)

egen mean_metolachlor_quart = cut(mean_metolachlor), group(4)

describe

scatter adjusted_rate2 mean_atrazine

scatter adjusted_rate2 mean_acetochlor

scatter adjusted_rate2 mean_glyphosate

scatter adjusted_rate2 mean_metam

scatter adjusted_rate2 mean_metolachlor

pwcorr adjusted_rate2 mean_atrazine mean_acetochlor mean_glyphosate mean_metam
mean_metolachlor

xtset stateID //this gives you the cluster (called panel variable in stata) variable, then the timevariable so
you can use them in all subsequent xt commands

xtgee count ib4.race_category meanatrazine meanacetochlor meanglyphosate meanmetam
meanmetolachlor, family(poisson) link(log) corr(exchangeable) offset(logpop) vce(robust) eform

estat wcorr

xtgee count ib4.race_category c.meanatrazine ib4.race_category##c.meanatrazine, family(poisson)
link(log) corr(exchangeable) offset(logpop) vce(robust) eform

xtgee count ib4.race_category c.meanacetochlor ib4.race_category##c.meanacetochlor, family(poisson)
link(log) corr(exchangeable) offset(logpop) vce(robust) eform
```

xtgee count ib4.race_category c.meanglyphosate ib4.race_category##c.meanglyphosate, family(poisson) link(log) corr(exchangeable) offset(logpop) vce(robust) eform

xtgee count ib4.race_category c.meanmetam ib4.race_category##c.meanmetam, family(poisson) link(log) corr(exchangeable) offset(logpop) vce(robust) eform

xtgee count ib4.race_category c.meanmetolachlor ib4.race_category##c.meanmetolachlor, family(poisson) link(log) corr(exchangeable) offset(logpop) vce(robust) eform

meglm count ib4.race_category meanatrazine meanacetochlor meanglyphosate meanmetam meanmetolachlor, offset(logpop) || stateID:, family(poisson) link(log) eform

margins meanatrazine meanacetochlor meanglyphosate meanmetam meanmetolachlor, over(race_category)

marginsplot, xdimension(race_category) plotdimension (meanatrazine)

marginsplot, xdimension(race_category) plotdimension (meanacetochlor)

marginsplot, xdimension(race_category) plotdimension (meanglyphosate)

marginsplot, xdimension(race_category) plotdimension (meanmetam)

marginsplot, xdimension(race_category) plotdimension (meanmetolachlor)

*Using Pesticide Exposure Quartiles

meglm count ib4.race_category i.mean_atrazine_quart i.mean_acetochlor_quart i.mean_glyphosate_quart i.mean_metam_quart i.mean_metolachlor_quart, offset(logpop) || stateID:, family(poisson) link(log) eform

meglm count ib4.race_category i.mean_atrazine_quart , offset(logpop) || stateID:, family(poisson) link(log) eform

contrast i.mean_atrazine_quart

meglm count ib4.race_category i.mean_acetochlor_quart, offset(logpop) || stateID:, family(poisson) link(log) eform

contrast i.mean_acetochlor_quart

```
meglm count ib4.race_category i.mean_glyphosate_quart, offset(logpop) || stateID:, family(poisson) link(log) eform

contrast i.mean_glyphosate_quart


meglm count ib4.race_category i.mean_metam_quart, offset(logpop) || stateID:, family(poisson) link(log) eform

contrast i.mean_metam_quart


meglm count ib4.race_category i.mean_metolachlor_quart, offset(logpop) || stateID:, family(poisson) link(log) eform

contrast i.mean_metolachlor_quart


*Crude model for pesticide exposure

meglm count i.mean_atrazine_quart i.mean_acetochlor_quart i.mean_glyphosate_quart i.mean_metam_quart i.mean_metolachlor_quart, offset(logpop) || stateID:, family(poisson) link(log) eform

contrast i.mean_atrazine_quart i.mean_acetochlor_quart i.mean_glyphosate_quart i.mean_metam_quart i.mean_metolachlor_quart


meglm count ib4.race_category i.mean_atrazine_quart i.mean_acetochlor_quart i.mean_glyphosate_quart i.mean_metam_quart i.mean_metolachlor_quart, offset(logpop) || stateID:, family(poisson) link(log) eform

contrast i.mean_atrazine_quart i.mean_acetochlor_quart i.mean_glyphosate_quart i.mean_metam_quart i.mean_metolachlor_quart

margins i.mean_atrazine_quart i.mean_acetochlor_quart i.mean_glyphosate_quart i.mean_metam_quart i.mean_metolachlor_quart, over(race_category)

marginsplot, xdimension(race_category) plotdimension (mean_atrazine_quart)

marginsplot, xdimension(race_category) plotdimension (mean_acetochlor_quart)

marginsplot, xdimension(race_category) plotdimension (mean_glyphosate_quart)

marginsplot, xdimension(race_category) plotdimension (mean_metam_quart)

marginsplot, xdimension(race_category) plotdimension (mean_metolachlor_quart)
```

*Part 3: Checking the analyses in the dataset with age groups included

* Importing a csv spreadsheet/dataset into stata using the "insheet" command

* Use the comma to indicate a command option: in this case, clear tells stat to remove the previous dataset because stata can only have one dataset open at the time

insheet using high_Pesticide_Exposure_AgeGroups2.csv, clear

* Using the describe command to see the number of variables, variable type and number of observations

describe

list in 1/10

gen logpop = log(population)

encode state, generate(stateID)

encode race, generate(race_category)

encode age_group, generate(age_group_cat)

g mean_atrazine= round(meanatrazine,1)

g mean_acetochlor= round(meanacetochlor,1)

g mean_glyphosate= round(meanglyphosate,1)

g mean_metam= round(meanmetam,1)

g mean_metolachlor= round(meanmetolachlor,1)

egen mean_atrazine_quart = cut(mean_atrazine), group(4)

egen mean_acetochlor_quart = cut(mean_acetochlor), group(4)

egen mean_glyphosate_quart = cut(mean_glyphosate), group(4)

egen mean_metam_quart = cut(mean_metam), group(4)

egen mean_metolachlor_quart = cut(mean_metolachlor), group(4)

*Using Pesticide Exposure Quartiles

```
meglm count ib4.race_category i.age_group_cat, offset(logpop) || stateID:, family(poisson) link(log) eform

contrast i.age_group_cat


meglm count ib4.race_category i.age_group_cat ib4.race_category##i.age_group_cat, offset(logpop) || stateID:, family(poisson) link(log) eform

margins i.race_category, over(age_group_cat)

marginsplot, xdimension(age_group_cat) plotdimension (race_category)


meglm count ib4.race_category i.age_group_cat i.mean_atrazine_quart i.mean_acetochlor_quart i.mean_glyphosate_quart i.mean_metam_quart i.mean_metolachlor_quart, offset(logpop) || stateID:, family(poisson) link(log) eform

contrast i.mean_atrazine_quart i.mean_acetochlor_quart i.mean_glyphosate_quart i.mean_metam_quart i.mean_metolachlor_quart
```

TRENDS AND SPLINES STATA CODE

```
*Bringing in the data
use  "/Users/richardmaclehose/Dropbox/PubH 8342/Splines/sample.dta",

*creating many small categories for our predictor of interest, birthweight
g catbw=.
replace catbw=0 if bweight<1000
replace catbw=1 if bweight>=1000 &bweight<1500
replace catbw=2 if bweight>=1500 &bweight<2000
replace catbw=3 if bweight>=2000 &bweight<2500
replace catbw=4 if bweight>=2500 &bweight<3000
replace catbw=5 if bweight>=3000 &bweight<3500
replace catbw=6 if bweight>=3500 &bweight<4000
replace catbw=7 if bweight>=4000 &bweight<4500
replace catbw=8 if bweight>=4500

*label the new variables for spline knots
label define lcatbw 2"1.5-2kg" 3"2-2.5kg" 4"2.5-3kg" 5"3-3.5kg" 6"3.5-4kg"
7"4-4.5kg" 8">4.5kg"
label val catbw lcatbw

*making a dichotomized value for birthweight
generate dbw=bweight>=3000
label define ld 0"<3kg" 1">=3kg"
label val dbw ld

svyset [pw=wt]
svy: tab catbw death, row
recode catbw 0=1

*Running logistic regression on birthweight, then using the "two way
command" to make plots looking at linear trend in the log odds
logistic death bweight [pw=wt] ,cformat(%9.3f)
predict xb_linear,xb
predict p_linear
twoway (line p_linear bweight if bweight<5000, sort lwidth(medthick)),
xtitle(Birth Weight (g))
twoway (line xb_linear bweight if bweight<5000, sort lwidth(medthick)),
xtitle(Birth Weight (g))

*Running logistic regression with a dichotomized birthweight (note how to
use i.variable syntax in the regression model
logistic death i.dbw [pw=wt]
predict p_dcat
twoway (line p_linear bweight if bweight<5000, sort lwidth(medthick))
(scatter p_dcat bweight if bweight<5000,sort col(green)), xtitle(Birth
Weight (g))

*Running the logistic regression with birthweight categorized in multiple
smaller categories
logistic death i.catbw [pw=wt]
predict p_cat
```

```
twoway (line p_linear bweight if bweight<5000, sort lwidth(medthick))
(scatter p_cat bweight if bweight<5000,sort col(red))(scatter p_dcat
bweight if bweight<5000,sort col(green)), xtitle(Birth Weight (g))

*making a spline using the mkspline command: the last variable before the
= sign tells stata which predictor to base the splines off
*Using <variable_a>-<variable_b> will tell stata to use all the variables
between the 2 indicated values
mkspline spline0 2000 spline1 2500 spline2 3000 spline3 3500 spline4 4000
spline5 4500 spline6 =bweight, di marginal
logistic death bweight spline1-spline6 [pw=wt]
predict p_linspline
twoway (line p_linear bweight if bweight<5000, sort lwidth(medthick))
(line p_cat bweight if bweight<5000,sort col(red)) (line p_linspline
bweight if bweight<5000,sort col(green)), xtitle(Birth Weight (g))
twoway  (line p_linspline bweight if bweight>3000&bweight<5000,sort
col(green)), xtitle(Birth Weight (g))

* Making a quadratic transformation of the splines
foreach var of varlist bweight spline1-spline6{g `var'_sq=`var'*`var'}
logistic death bweight bweight_sq spline1_sq spline2_sq spline3_sq
spline4_sq spline5_sq spline6_sq [pw=wt]
predict p_qspline
twoway (line p_cat bweight if bweight<5000, sort lwidth(medthick)) (line
p_qspline bweight if bweight<5000,sort col(red)) (line p_linspline bweight
if bweight<5000,sort col(green)), xtitle(Birth Weight (g))

foreach var of varlist spline1-spline6{g rq`var'=`var'*`var'-
spline6*spline6}
logistic death bweight rqspline1 rqspline2 rqspline3 rqspline4 rqspline5
[pw=wt]
predict p_rqspline
twoway (line p_linear bweight if bweight<5000, sort lwidth(medthick))
(line p_rqspline bweight if bweight<5000,sort col(red)) (line p_linspline
bweight if bweight<5000,sort col(green)), xtitle(Birth Weight (g))
twoway (line p_rqspline bweight if bweight>3000&bweight<5000,sort
col(red)) (line p_linspline bweight if bweight>3000&bweight<5000,sort
col(green)), xtitle(Birth Weight (g))
```

SPLINES SAMPLE STATA CODE

```
* PUBH 8342 Assignment 1

*Question 1
*exp(B0)
di exp(log(0.1))
*exp(B1)
di exp(log(0.3)-log(0.1))
*exp(B2)
di exp(log(0.2)-log(0.1))
*B0
di log(0.1)
*B1
di log(0.3)-log(0.1)
*B2
di log(0.2)-log(0.1)
*B3
di log(0.4)
di (-.91629073 - .69314718 - 1.0986123 - (-2.3025851))
*exp(B3)
di exp(-.40546511)

*Question 2
*bringing in the framingham data
use  "C:\Users\onyea005\Desktop\Homework 1\frmgham2.dta"

*examine the data. 'd' is short for describe
describe

*descriptive statistics on bmi
summarize bmi, detail

*Getting the percentiles for bmi
sort bmi
centile bmi, centile(20 40 60 80)
return list

*making a spline using the mkspline command
mkspline spline0 22.688 spline1 24.568 spline2 26.382 spline3 28.792
spline4 =bmi, di marginal

*running a logistic regression relating bmi to the risk of death using the
linear splines
logistic death bmi spline1-spline4
predict p_linspline
twoway (line p_linspline bmi,sort col(green))

*Question 3
*running a logistic regression relating bmi to the risk of death using
quadratic splines
foreach var of varlist bmi spline0-spline4 {
g `var'_sq=`var'*`var'
}
```

```
logistic death bmi bmi_sq spline0_sq spline1_sq spline2_sq spline3_sq
spline4_sq
predict p_qspline
twoway (line p_qspline bmi,sort col(red))

*Question 4
*running a logistic regression relating bmi to the risk of death using
restricted quadratic splines
foreach var of varlist spline0-spline4 {
g rq`var'=`var'*`var'-spline4*spline4
}
logistic death bmi rqspline0 rqspline1 rqspline2 rqspline3
predict p_rqspline
twoway (line p_rqspline bmi,sort col(blue))

*Question 5
*getting descriptive statistics for sex
tabulate sex
recode sex 1=0 2=1
tabulate sex

*repeating the logistic regression using restricted quadratic splines
adjusting for sex
logistic death bmi rqspline0 rqspline1 rqspline2 rqspline3 sex
predict p_rqsplinesex
twoway (line p_rqsplinesex bmi,sort col(blue))

*Adjusting the sex variable
logistic death bmi rqspline0 rqspline1 rqspline2 rqspline3 i.sex
predict p_rqsplinesex2
twoway (line p_rqsplinesex2 bmi,sort col(blue))

logistic death bmi rqspline0 rqspline1 rqspline2 rqspline3 if sex ==0
predict p_rqsplinev2
twoway (line p_rqsplinev2 bmi,sort col(blue))

logistic death bmi rqspline0 rqspline1 rqspline2 rqspline3 if sex ==1
predict p_rqsplinev3
twoway (line p_rqsplinev3 bmi,sort col(blue))
twoway (line p_rqsplinev3 bmi,sort col(black)) (line p_rqsplinev2 bmi,sort
col(green))

* Ignore this section
*I think the graph looks funky due to the splines constraining the values
that the slope can take, and having to switch between 0 and 1 for each
individual observation's sex value
*I will make a spline for sex based on the proportion of gender in the
dataset to correct this

tabulate sex
mkspline splinesex0 0.425 splinesex1=sex, di marginal
foreach var of varlist sex splinesex0-splinesex1 {
g `var'_sq=`var'*`var'
}
```

```
logistic death bmi rqspline0 rqspline1 rqspline2 rqspline3 sex splinesex0
predict p_rqsplinesexmk1
twoway mspline(line p_rqsplinesexmk1 bmi,sort col(blue))

* Resuming code proper

*Question 6
*Dichotomizing the BMI variable
generate dbmi=bmi>=25.92239
label define lbl 0"0" 1"1"
label val dbmi lbl
tab dbmi
tab dbmi sex

*running a logistic regression with an interaction term for bmi and sex
logistic death i.dbmi##i.sex
binreg death i.dbmi##i.sex

* Making a table for the relative risks
tab death dbmi if sex==0
tab death dbmi if sex==1

*ICR Calculation
di exp(_b[1.dbmi])*exp(_b[1.sex])*exp(_b[1.dbmi#1.sex])-exp(_b[1.dbmi])-
exp(_b[1.sex])+1
```

SURVIVAL ANALYSIS STATA CODE

```
use "/Users/maclehose/Google Drive/pubh8342_2016/data/gehan.dta", clear

list
*life table with 2 week intervals
ltable weeks relapse if group==2, intervals(2) noadjust


*set up survival data
stset week, f(relapse)

list

*Generate a KM survival estimate
sts list if group==2

sts list if group==1

sts list, by(group)

*graph the KM curves

sts graph, by(group) risktable ci



*Test equality of KM curves
sts test group


*Lecture 2
*Recode treatment so new therapy is baseline
recode group 2=0
*Run Cox Model


stcox i.group

*Check proportionality
* Graphical test
sts graph, by(group)
stphplot, by(group)

* Residual test
stcox i.group
estat phtest

*Lecture 3
* Interaction with ln(time)
stcox i.group, tvc(group) texp(ln(_t))
*Or interact with time
stcox i.group, tvc(group) texp(_t)
*note that neither of these is equivalent to the following
* THIS IS WRONG!!!!
```

```
g grouptime=group*week
stcox i.group grouptime


*Estimating effects
stcox group, tvc(group) texp(_t)
*HR at time 0
lincom _b[main:group]+_b[tvc:group]*0,eform
*HR at time 1
lincom _b[main:group]+_b[tvc:group]*1,eform

*HR at time 10
lincom _b[main:group]+_b[tvc:group]*10,eform

*or for the log-t interaction
stcox group, tvc(group) texp(ln(_t))
*HR at time 0
lincom _b[main:group]+_b[tvc:group]*(-4.9),eform
*HR at time 1
lincom _b[main:group]+_b[tvc:group]*0,eform

*HR at time 10
lincom _b[main:group]+_b[tvc:group]*2.302,eform


use "/Users/maclehose/Google Drive/pubh8342_2016/data/gehan2.dta", clear
stset week, f(relapse)

recode group 2=0
stcox group sex, tvc(group sex)
stcox group sex, tvc(sex)

*Stratify by sex
stcox group, strata(sex)



use "/Users/maclehose/Google Drive/pubh8342_2016/data/leukemia.dta", clear
stset week, f(relapse)




*time varying covariate
use "/Users/maclehose/Google Drive/pubh8342_2016/data/stanford.dta", clear
stset stime, f(died) id(id)

stsplit post,at(0) after(wait)
recode post -1=0 0=1
replace post=0 if wait==0

stcox post
```

COX REGRESSION SAMPLE STATA CODE

```
* PUBH 8342 Assignment 5

*bringing in the dataset
use  "C:\Users\guillaumeo\Desktop\Graduate School\PhD\Spring 2016\PUBH
8342\Homeworks\Homework 5\smokesurvival.dta"

*examine the data. 'd' is short for describe
describe
list in 1/10

*eversmoke: Whether a person ever smoked prior to the beginning of the
study (1= smoked)
*disease: Whether the person developed lung disease (disease=1) or
censored (disease=0)
*years: The number of years until the person developed disease or was
censored
*ow: An indicator for whether the person was overweight (ow=1) or not
(ow=0) at the beginning of the study
*Sex: 1=male; 0=female

*Question 1

*Setting up the time and censoring variables for the survival analysis
*In this case, our follow up time variable is years, and our failure
indicator variable is disease
stset years, failure(disease)

* Getting the KM survival estimate

sts list
sts graph, by(ow) risktable ci

*Question 2
*Running a cox model for the effect of being overweight on occurence of
COPD

*Crude model
stcox i.ow i.eversmoke i.sex

*Method 1: Checking for proportionality by plotting the log (negative log)
curves
stphplot, by(ow)

*Method 2: Checking for proportionality using the goodness of fit test
estat phtest

*Method 3: Checking for time varying coefficients
stcox i.ow i.eversmoke i.sex, tvc(i.ow i.eversmoke i.sex) texp(_t)

*stcox i.ow i.eversmoke i.sex, tvc(i.ow i.eversmoke i.sex) texp(ln(_t))
*Looks like you can just keep eversome as the tvc
```

```
*Question 2a
stcox i.ow i.eversmoke i.sex, tvc(i.ow i.eversmoke) texp(_t)
*stcox i.ow i.eversmoke, tvc(i.ow) texp(ln(_t)) strata(sex)
*stcox i.ow i.sex, tvc(i.ow) texp(ln(_t)) strata(eversmoke)
stcox i.ow, tvc(i.ow) texp(_t) strata(eversmoke sex)

*Question 2b
*HR at time 2
lincom _b[main:1.ow]+_b[tvc:1.ow]*2,eform

*HR at time 10
lincom _b[main:1.ow]+_b[tvc:1.ow]*10,eform

*Question 3
 stcox i.eversmoke i.sex i.ow , tvc(i.ow i.eversmoke i.sex) texp(_t)
 *Most appropriate cox model,, since sex and ow were not statistically
significant tvcs
 stcox i.eversmoke i.sex i.ow , tvc(i.eversmoke) texp(_t)

 *Quesiton 3b
 *Regular time
 stcox i.eversmoke i.sex i.ow , tvc(i.eversmoke) texp(_t)

*HR at time 5
lincom _b[main:1.eversmoke]+_b[tvc:1.eversmoke]*5,eform

*HR at time 10
lincom _b[main:1.eversmoke]+_b[tvc:1.eversmoke]*10,eform

*HR at time 20
lincom _b[main:1.eversmoke]+_b[tvc:1.eversmoke]*20,eform

*HR at time 30
lincom _b[main:1.eversmoke]+_b[tvc:1.eversmoke]*30,eform

*Log Time
stcox i.eversmoke i.sex i.ow , tvc(i.eversmoke) texp(ln(_t))

*HR at time 5
lincom _b[main:1.eversmoke]+_b[tvc:1.eversmoke]*5,eform

*HR at time 10
lincom _b[main:1.eversmoke]+_b[tvc:1.eversmoke]*10,eform

*HR at time 20
lincom _b[main:1.eversmoke]+_b[tvc:1.eversmoke]*20,eform

*HR at time 30
lincom _b[main:1.eversmoke]+_b[tvc:1.eversmoke]*30,eform

*Question 4
stcox i.eversmoke i.sex i.ow , tvc(i.eversmoke) texp(_t)
*Plotting the log (negative log) curves
stphplot, by(eversmoke)
```

```
di exp(2.3)

*The cutpoint seems to be at approximately 10 years
stcox i.eversmoke i.sex i.ow , tvc(i.eversmoke) texp(10)
stcox i.eversmoke i.sex i.ow , tvc(i.eversmoke) texp(9)
stcox i.eversmoke i.sex i.ow , tvc(i.eversmoke) texp(11)

*HR at time 5
lincom _b[main:1.eversmoke]+_b[tvc:1.eversmoke]*5,eform

*HR at time 10
lincom _b[main:1.eversmoke]+_b[tvc:1.eversmoke]*10,eform

*HR at time 20
lincom _b[main:1.eversmoke]+_b[tvc:1.eversmoke]*20,eform

*HR at time 30
lincom _b[main:1.eversmoke]+_b[tvc:1.eversmoke]*30,eform

*Waiiiiit thats a trick question!!! it should be interaction = 0 at 5
years, and interaction = 1 at 10, 20 and 30 years
*HR at time 5 and 10
lincom _b[main:1.eversmoke]+_b[tvc:1.eversmoke]*0,eform

*HR at time 20 and 30
lincom _b[main:1.eversmoke]+_b[tvc:1.eversmoke]*1,eform
```