

A course on
'Longitudinal data analysis':
what did we learn?

COMPASS
11/03/2011

Abstract

- We report back on methods used for the analysis of longitudinal data covered by the course
- We draw lessons for our own work
- This course was run by the NZ Social Statistics Network (NZSSN) Summer Programme in February 2011.
- Taught by Dr Gary Marks from the University of Melbourne.
- “This course is designed for social science researchers who wish to address research questions using appropriate statistical procedures on longitudinal data. It is not an advanced statistics course.”

Outline

1. Introduction – *Roy Lay-Yee*
2. NZSSN summer programme;
foundational topics – *Martin von Randow*
3. Fixed effects model – *Janet Pearson*
4. Random effects model – *Jessica Thomas*
5. Event history analysis – *Karl Parker*
6. Hybrid (FE/RE) model – *Roy Lay-Yee*
7. Lessons for us

Introduction

Roy Lay-Yee

What is a longitudinal study?

- Can see how, for example, persons experience change and respond to those changes
- Gathers information on the same person over time
- Repeated observations on persons, i.e. observed occasions are nested within a person
- Lack of independence between observations on same person
- Assumes data from different persons are independent

Why are longitudinal studies important?

- Capture changes over time: dynamics, sequencing, trajectories (not possible with cross-sectional studies)
- Allow greater analytical power with multiple observations per person
- Can control for time-invariant unobserved (stable) differences between people, e.g. ability
- Reduce spuriousness; can get closer to causal effects; at least can make stronger inferences

Disadvantages of longitudinal studies

- Attrition – loss of respondents over time
 - missing data (unbalanced dataset)
 - biased sample (particular sorts of people drop out - data not missing at random)
- Sample becomes less representative of population over time
- Sample members may be influenced by being part of the study

Longitudinal data analysis

- Many modelling approaches
- The course covered:
 - Fixed effects model (FE)
 - Random effects model (RE)
 - Event history analysis (EHA)
- But first things first ...



NZSSN
SOCIAL STATISTICS NETWORK

NZSSN SP11 Summary

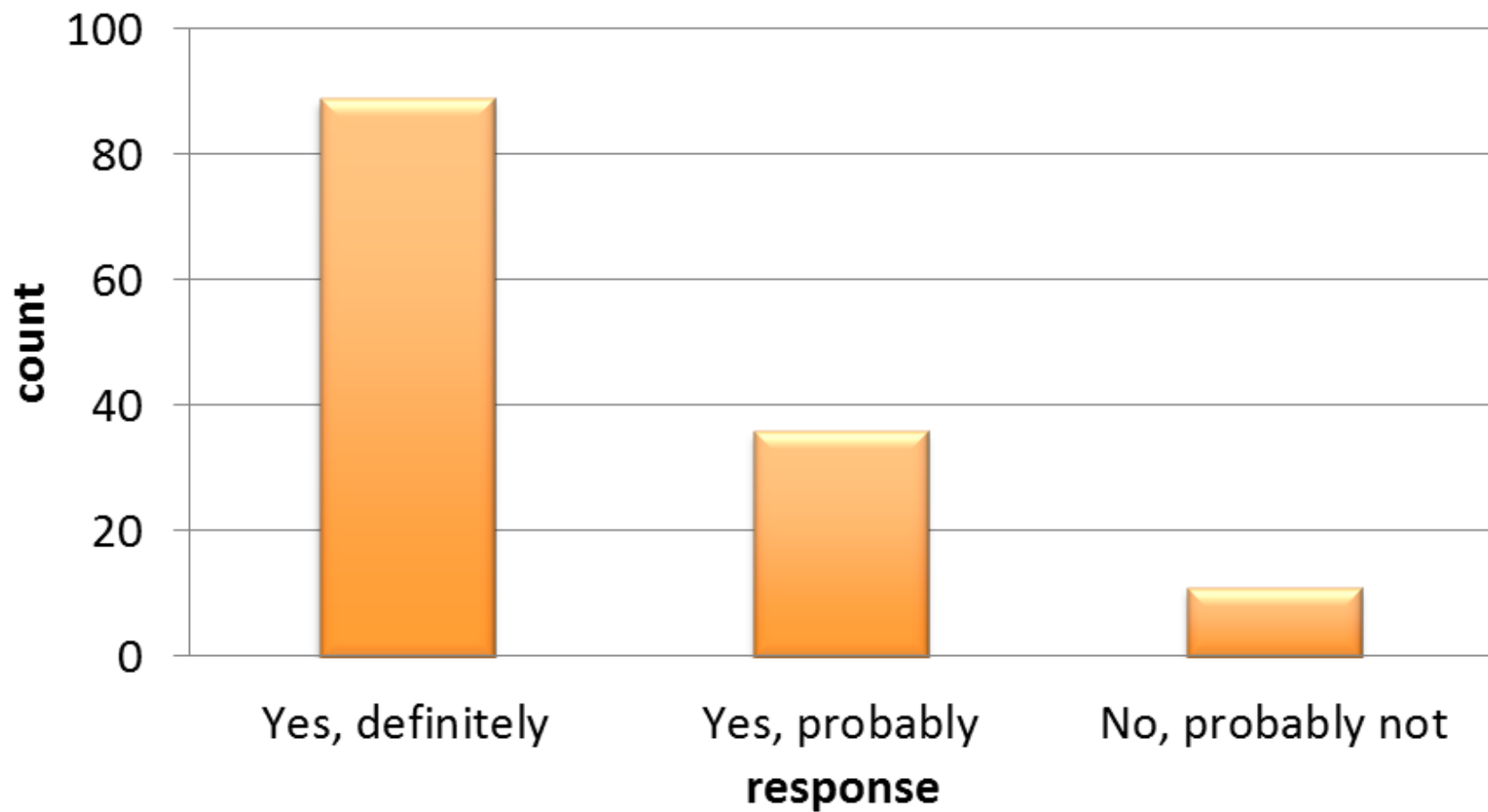
Martin von Randow



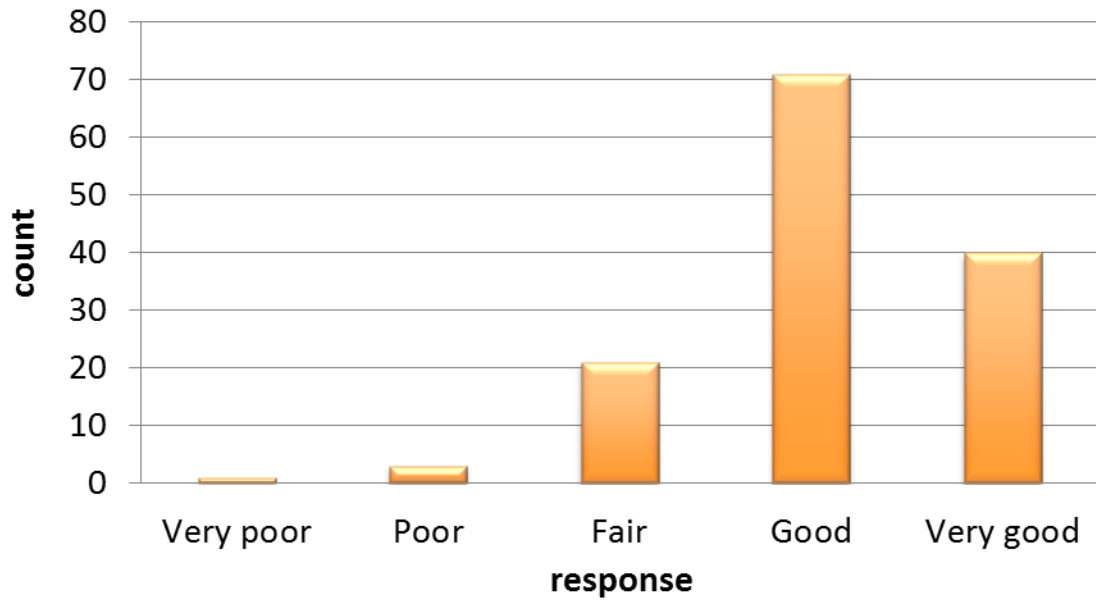
NZSSN Summer Programme

- Has been held annually since 2005, growing significantly the last 3 years
- This year:
 - 11 5-day courses in social research methods held the weeks of 7–11 & 14–18 Feb, at VUW
 - 140 attendees in total; range: 5–25 per course
- Instructors mainly from Australia, but also UoA statistics department!

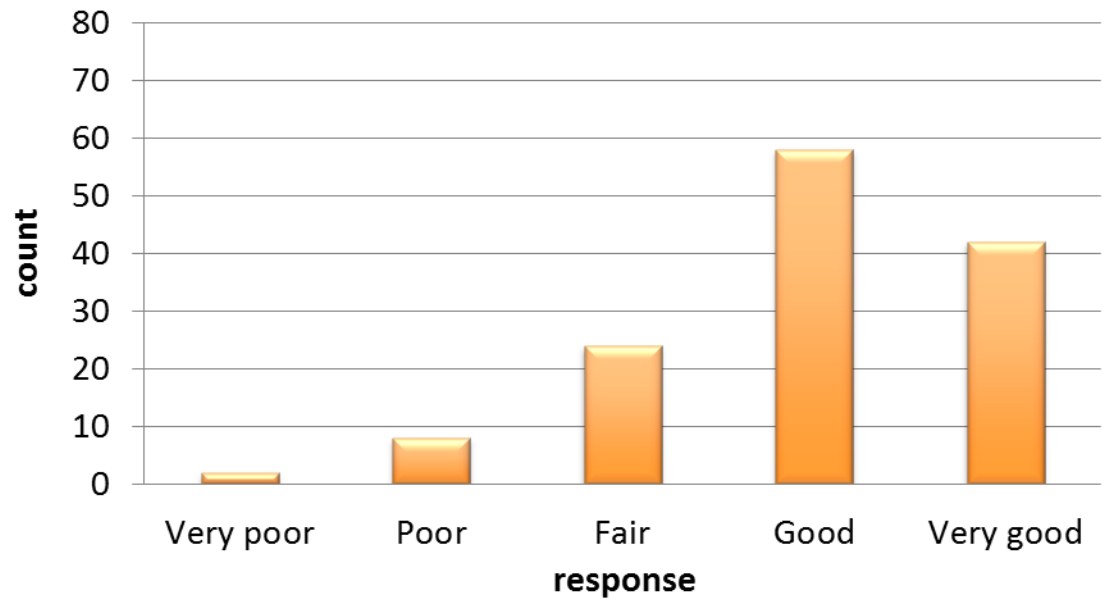
Recommending the Courses



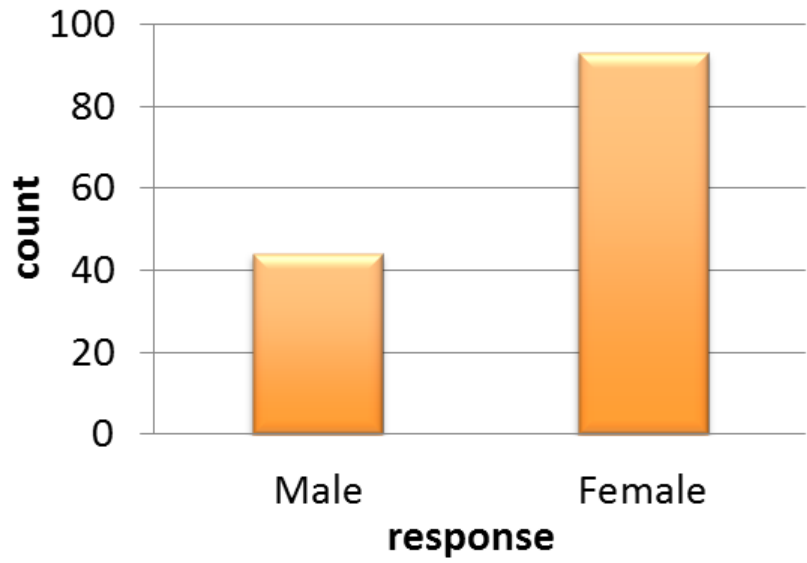
Teaching Rooms



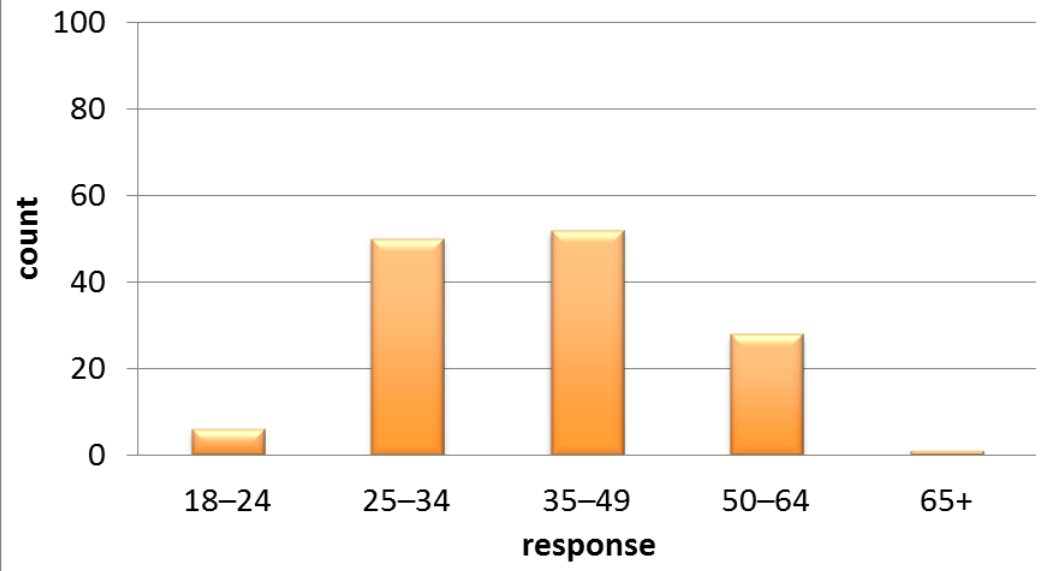
Catering



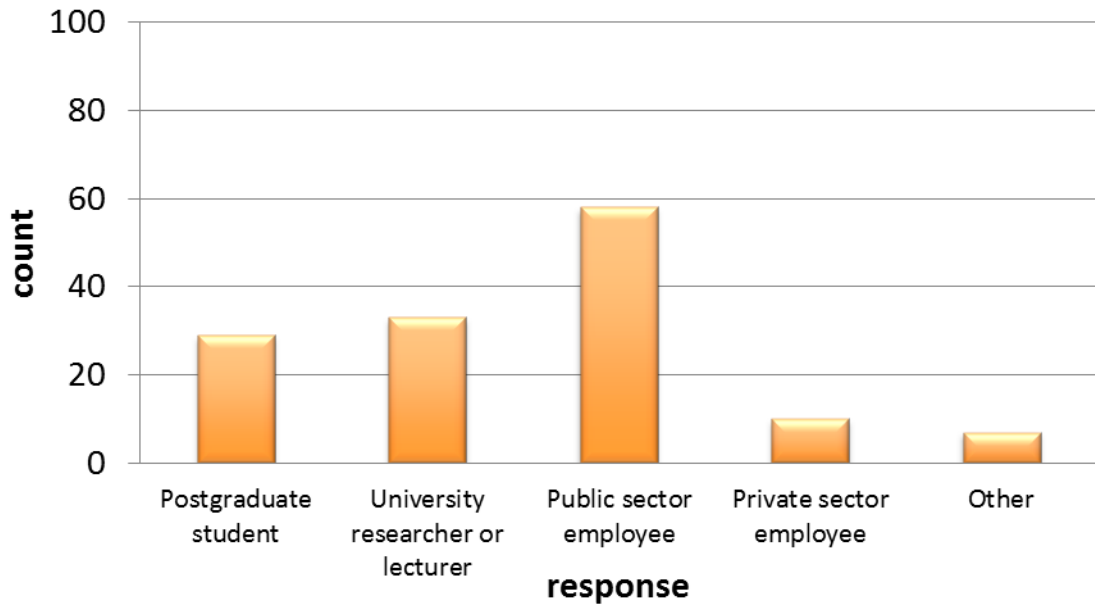
Distribution of Gender



Distribution of Age Group




Distribution of Occupation



Future Plans

- Winter Programme 2011!
 - Plans for 6 courses in the week of 11–15 July
 - Here at UoA, using Business School rooms
- Further improvements to our new website
- More targeted marketing
- Ever more new courses



Longitudinal Data Analysis – Intro –

Martin von Randow

Course details

- Intensive, Monday–Thursday, 9am–4.30pm plus Friday morning
- Instructed by Dr Gary Marks of the Australian Council for Educational Research
- 2 days of ‘revision’ covering everything from OLS regression through multinomial & ordinal
- 2.5 days specifically on the longitudinal case
- Examples from AUS surveys: HILDA & LSAY

For 'one observation per case'

- Revision
 - Normal distribution
 - Populations and samples
 - Basic univariate statistics
 - Correlations, etc.
 - Bivariate regression
- Multiple regression
- Logistic regression for dichotomous outcomes
- Multinomial regression (logit)
- Ordinal Regression (logit)



And examples covering...

- PISA test scores (15 year olds)
- University entrance performance
- Earnings
- Life satisfaction
- Poverty
- Financial stress
- Exiting unemployment
- Transition to adulthood (leaving home, marriage)

Revision – regression assumptions

- The relationship between X and Y is linear
- ε has a mean of zero
- ε is a normally distributed variable
- ε is uncorrelated with X
- ε has a constant variance across X values
- X is measured without error
- Model is properly specified; there are no other variables correlated with X that impact on Y

Revision – regression equations

$$8 \hat{Y}_i = \hat{b}_0 + \hat{b}_1 X_i \quad R^2 = \frac{\sum_{i=1}^n (\bar{Y} - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{Y})^2}$$

- $Y = b_0 + b_1 X_1 + b_2 X_2 + b_3 X_3 \dots + b_k X_k + e$

- $R_{Adj}^2 = 1 - \frac{\sum_{i=1}^n (y_i - \hat{Y}_i)^2 / (n - p - 1)}{\sum_{i=1}^n (y_i - \bar{Y})^2 / (n - 1)}$

- $Y = b_0 + b_1 X_1 + b_2 X_2 + b_3 (X_1 * X_2) + e$

- $\ln\left(\frac{p_i}{1-p_i}\right) = b_0 + b_1 X_1 + b_2 X_2 + b_3 X_3 \dots + b_k X_k + e$

More interesting – multinomial

- $Y_j = \sum_1^{j-1} b_0^{j-1,j} + \sum_1^{j-1} b_1^{j-1,j} x_1 \dots + \sum_2^k \sum_1^{j-1,j} b_k^{j-1,j} x_k$
- $\sum_1^{j-1} b_0^{j-1,j}$ represents $j-1$ intercept terms, $b_0^1, b_0^2, b_0^3 \dots b_0^{j-1}$ relative to b_0^j – the effects of the marginal distribution of Y_1^j
- $\sum_1^{j-1} b_1^{j-1,j} x_1$ represents the $j-1$ effects $b_1^1, b_1^2, b_1^3 \dots b_1^{j-1}$ relative to b_1^j of continuous variable x_1
- Interpret as odds of being in one category compared to another, e.g. $p_1/p_3, p_2/p_3$

And finally – ordinal

- Interpret odds of being in a given category or lower compared to being in a higher category, e.g. $p_1/(p_2 + p_3 + p_4)$, $(p_1 + p_2)/(p_3 + p_4)$
- $\text{logit}(F_{j \leq 1}) = \log(F_{j \leq 1}/(1 - F_{j \leq 1})) = \log \left[\frac{p_1}{(p_2 + p_3 + p_4)} \right]$
- $\text{logit}(F_{j \leq 2}) = \log(F_{j \leq 2}/(1 - F_{j \leq 2})) = \log \left[\frac{(p_1 + p_2)}{(p_3 + p_4)} \right]$
- $\text{logit}(F_{j \leq 3}) = \log((F_{j \leq 3}/(1 - F_{j \leq 3})) = \log \left[\frac{(p_1 + p_2 + p_3)}{(p_4)} \right]$

Fixed Effects Regression Model

Janet Pearson

11-3-2011

Outline

- Introduction
- What does the Fixed Effects model look like?
- Advantages of fixed effects
- Methods of implementation
- Disadvantages
- Summary/Conclusion

Introduction

- NZSSN Longitudinal Data Analysis course – Dr Gary Marks from the University of Melbourne
- Fixed effects regression
 - One way of analysing multi-level data – e.g time periods within people

Fixed Effects Model

Multiple measures for outcome on same individual – e.g. recidivism

Effect of all unobserved time invariant predictors – e.g. parent's child-rearing practices

Purely random error at each point in time

$$Y_{it} = \mu_t + \beta x_{it} + \gamma z_i + \alpha_i + \epsilon_{it}$$

$$i = 1, \dots, n \quad t = 1, \dots, T$$

Different possible intercept for each time period

Time variant predictors – e.g. marriage

Time invariant observed predictors – e.g. sex, ethnicity etc

Advantages of Fixed Effects Model 1

- Gets closer to causality of predictors – assuming unobserved time variant predictors are of no importance
- Each individual is their own control
 - E.g Recidivism ~ marriage . Look at arrest rates for same person when married and when they weren't
- Controls for the effects of unobserved stable (time invariant) variables
 - As allows unobserved variables α_i to be correlated with the observed variables X_i and Z_i (with random effects models they are orthogonal)

Advantages of Fixed Effects Model 2

- Removes ‘contaminated’ variation (remember – parental child-rearing practices when researching effect of marriage on recidivism)
- Uses only within individual differences & discards the between-individual variation (“nuisance” variation)
 - ... Sacrifices efficiency in order to reduce bias in our estimates
- Interaction effects in fixed effect models provide a very useful method for testing research hypotheses involving invariant predictors – e.g ‘do men have a different effect of marriage on recidivism than women do’?

Methods of implementing

1. Inclusion of person-specific intercepts
 - α_i added to model statement in SAS in an OLS regression
 - Does not work for large samples with only a small number of time points
2. “Difference” equation
 - For when only have two time points
 - Proc reg, modelling the differences in outcome by the differences in the predictors
 - Proc glm – with ‘absorb’ statement to give the ‘individual id’ variable
 - Data has to be balanced – same amount of missing observations for each individual
3. Time series average
 - For when have three or more time points
 - Proc means then Proc surveylogistic (with ‘cluster’ = the individual id)
 - Data has to be balanced – same amount of missing observations for each individual

Fixed Effects Methodology: Difference Equation

- For when only two time periods

- At time 1

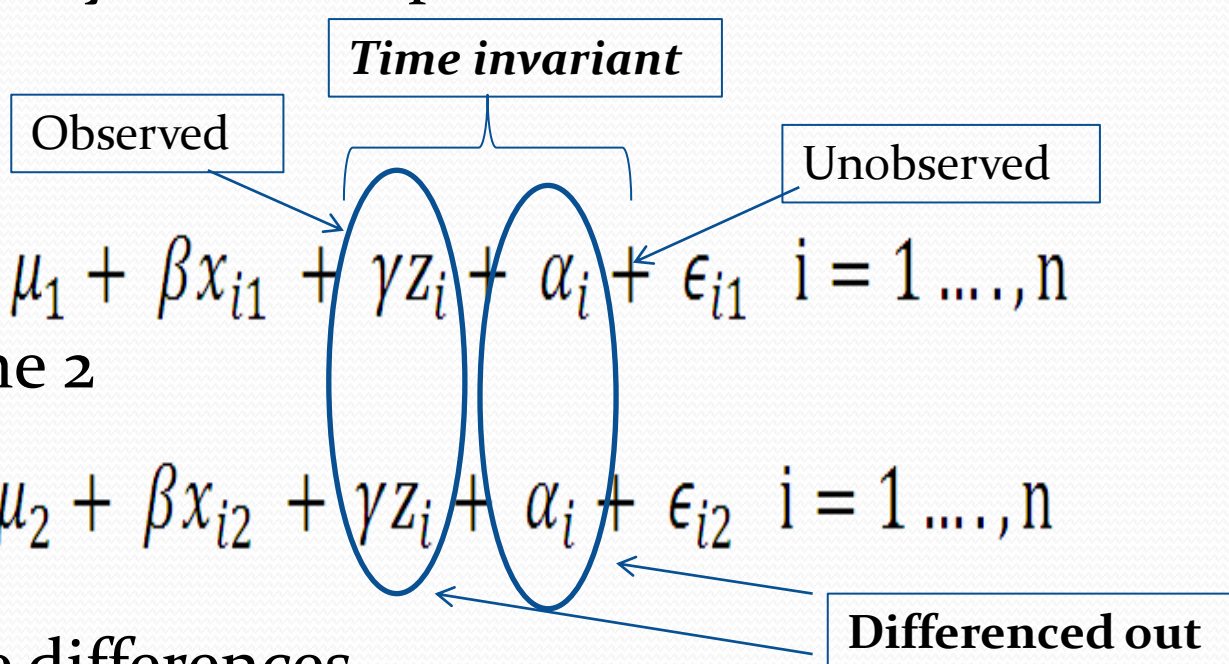
$$Y_{i1} = \mu_1 + \beta x_{i1} + \gamma Z_i + \alpha_i + \epsilon_{i1} \quad i = 1, \dots, n$$

- And at time 2

$$Y_{i2} = \mu_2 + \beta x_{i2} + \gamma Z_i + \alpha_i + \epsilon_{i2} \quad i = 1, \dots, n$$

- Taking the differences

$$Y_{i2} - Y_{i1} = (\mu_2 - \mu_1) + \beta(x_{i2} - x_{i1}) + (\epsilon_{i2} - \epsilon_{i1})$$



Fixed Effects methodology: Time series average

$$Y_{it}^* = Y_{it} - \bar{y}_i =$$

$$(\mu_t + \beta x'_{it} + \gamma Z_i + \alpha_i + \epsilon_{it})$$

$$- (\bar{u}_t + \beta \bar{x}'_i + \gamma \bar{Z}_i + \alpha_i + \bar{\epsilon}_i)$$

$$= \beta(x_{it} - \bar{x}_i) + (\mu_t - \bar{u}_t) + (\epsilon_{it} - \bar{\epsilon}_i)$$

$$= \beta x_{it}^{*'} + u_t^* + \epsilon_t^*$$

Observed and unobserved time invariant predictors are differenced out

Take difference from within subject means

If time is not included:

$$Y_{it}^* = \beta x_{it}^{*'} + \epsilon_t^{*'}$$

Fixed Effects Model disadvantages

- Inability to estimate effect of time invariant variables- e.g sex, age – it just takes account of them – differencing them out, but does not give estimates of their effect
- Causality only to the extent that time *invariant* unobserved are taken account of – unobserved variables that change over time may still be confounders
- If predictor variables vary greatly across individuals but have little variation over time for each individual, fixed effects estimates will be very imprecise
- Leads to higher p-values and wider confidence intervals than with random effects models

Summary/Conclusion

- Fixed Effects methodology:
 - Can be implemented in SAS
 - Is good for reducing bias in estimates, although reduces efficiency
 - Edges results more towards causality by having adjusted for some non-observed variables that may be confounders
 - Cannot estimate time invariant effects, although can estimate interactions with them in – e.g marriage*men
- Overall – a very interesting tool!



Random Effects Models

Random Effects Models

- Also called
 - Hierarchical models
 - Multilevel models
 - Variance component models
 - Mixed effects models

Random Effects Models

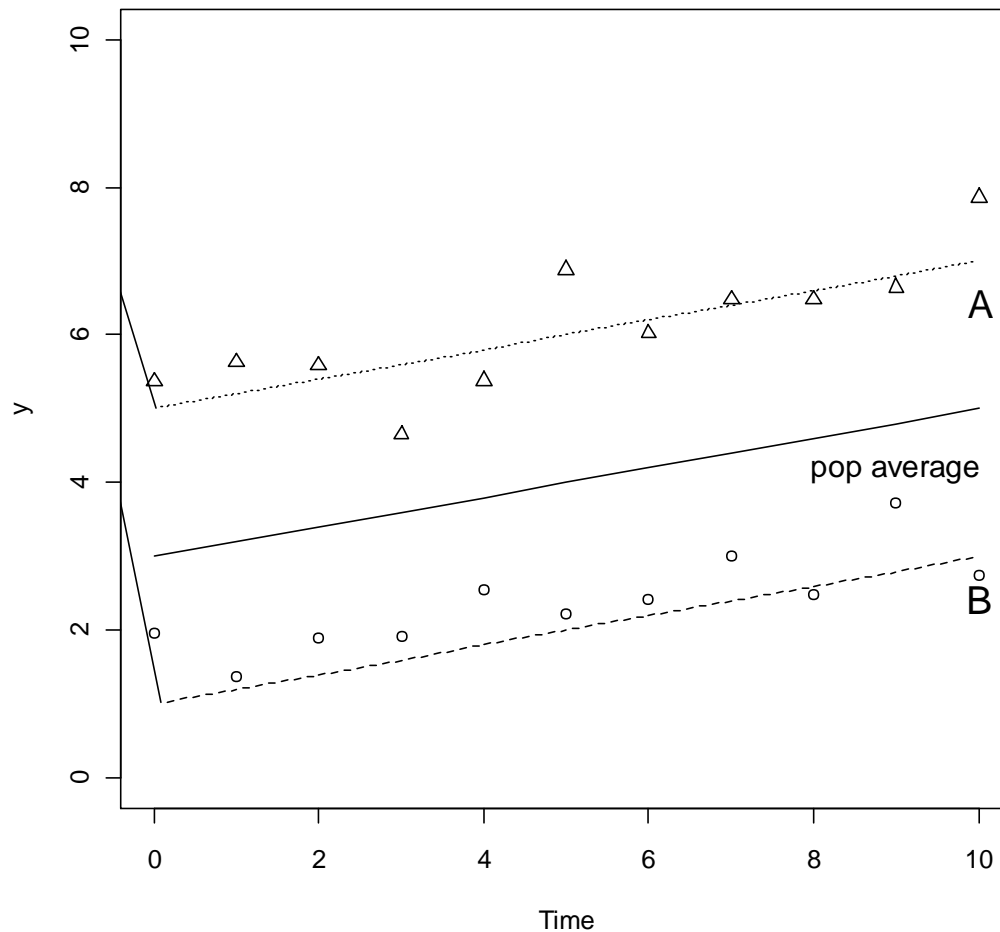
- Can use for grouped/clustered/hierarchical data
 - patients within hospitals
 - Children within schools
 - Apportion variance at each level
- Longitudinal/repeated measures data
 - Same child measured at multiple time points
 - Can look at individual response trajectories
- Data is not independent
 - A sample of children from the same school more similar to each other than a sample children from different schools
 - One individual's measurements consistently higher than average while another individual's measurements consistently lower than average.

Effect of Non-independence

- Correlation present
- Affects standard errors
 - If use ordinary OLS regression p-values and confidence intervals not correct
 - Standard errors may be too big or too small (depends on level of variable)

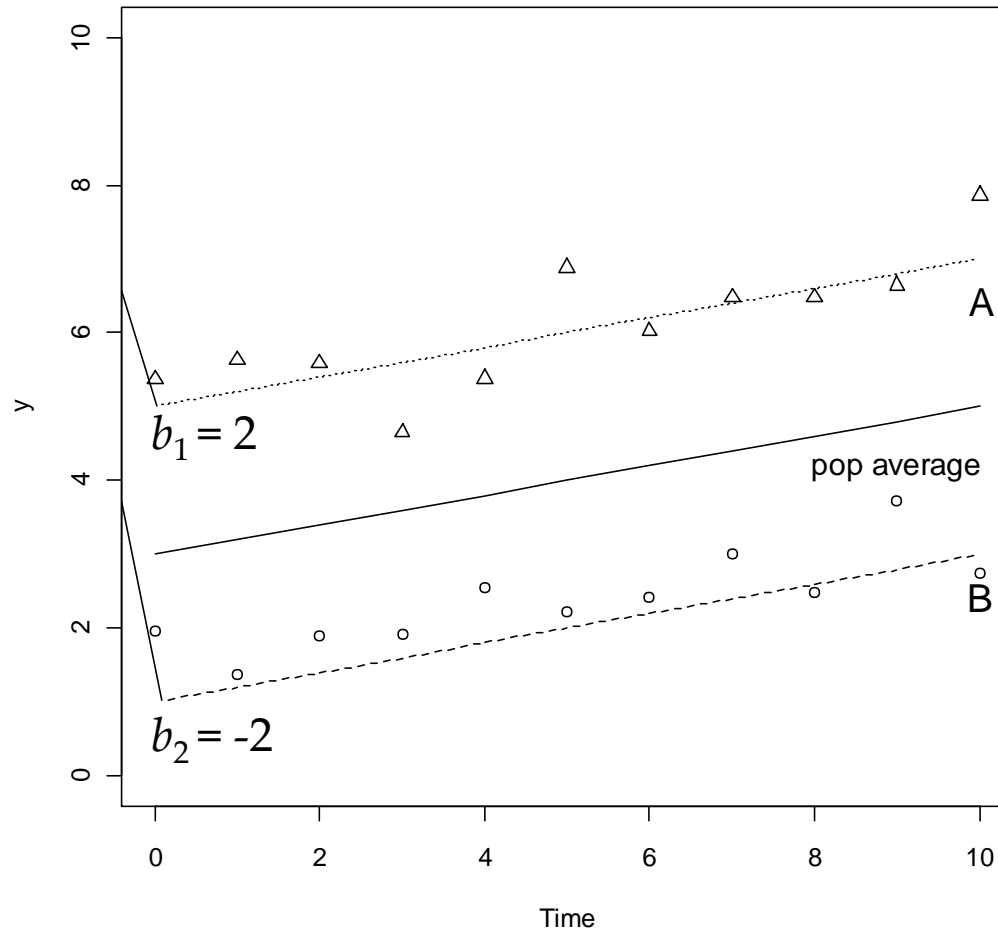
Random Intercepts Model

Each subject is assumed to have an (unobserved) underlying level of response which persists across his/her measurements



Random Intercepts Model

$$Y_{it} = \beta_0 + \beta_1 x_{it} + b_i + e_{it}$$



Random Intercepts Model

$$Y_{it} = \beta_0 + \beta_1 x_{1it} + \beta_2 x_{2it} + \dots + \beta_k x_{kit} + b_i + e_{it}$$

$$\text{Population mean} = \beta_0 + \beta_1 x_{1it} + \beta_2 x_{2it} + \dots + \beta_k x_{kit}$$

b_i = subject effect

$$\sim N(0, \sigma_b^2)$$

σ_b^2 = between-subject variance

e_{it} = within-subject error

$$\sim N(0, \sigma_e^2)$$

σ_e^2 = within-subject variance

Intra-class correlation: correlation between pairs of observations on the same individual

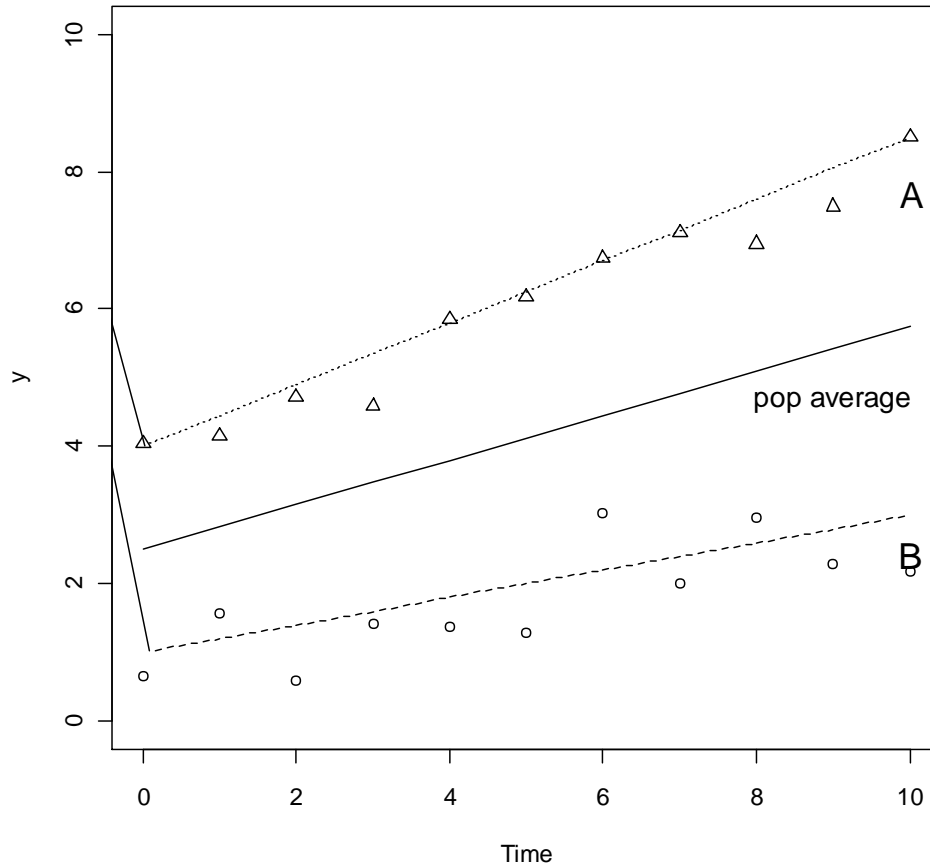
$$\sigma_b^2 / (\sigma_b^2 + \sigma_e^2)$$

SAS code – random intercepts

```
PROC MIXED CL;  
  CLASS id group;  
  MODEL y = group time group*time / SOLUTION;  
  RANDOM INTERCEPT / SUBJECT=id G;  
RUN;
```

- G option
 - gives the “G matrix” from which you can see the estimate of σ_b^2
 - If it is small intercepts do not vary much, if it is large, they do
- CL option
 - gives confidence limits for covariance parameter estimates (σ_b^2)
 - If the confidence interval does not contain zero, then σ_b^2 is significantly different from zero.
 - If zero is in the confidence interval you could consider using OLS regression

Random Intercepts and Slopes Model



$$\beta_0 = 2.5$$

$$b_{01} = 1.5$$

$$b_{02} = -1.5$$

$\beta_1 = 0.29$ (y increases by 2.9 units over the time period)

$$b_{11} = .41$$

$$b_{12} = .18$$

$$Y_{it} = \underbrace{\beta_0 + \beta_1 x_{it}}_{\text{Population line}} + \underbrace{b_{0i}}_{\substack{\uparrow \\ \text{Random} \\ \text{intercepts}}} + \underbrace{b_{1i} x_{it}}_{\substack{\uparrow \\ \text{Random} \\ \text{slopes}}} + \underbrace{e_{it}}_{\substack{\uparrow \\ \text{Within-subject} \\ \text{error}}}$$

SAS code – random intercepts and slopes

```
PROC MIXED CL;  
  CLASS id group;  
  MODEL y = group time group*time / SOLUTION;  
  RANDOM INTERCEPT time / SUBJECT=id TYPE=UN G;  
RUN;
```

TYPE=UN

Allows correlated intercepts and slopes

UN = unstructured covariance structure of G

$$\begin{bmatrix} \mathfrak{g}_{11} & \mathfrak{g}_{12} \\ \mathfrak{g}_{12} & \mathfrak{g}_{22} \end{bmatrix} = \begin{bmatrix} \text{var}(b_0) & \text{cov}(b_0, b_1) \\ \text{cov}(b_0, b_1) & \text{var}(b_1) \end{bmatrix}$$

Software estimates the correlation between the subject-specific intercepts and the subject-specific slopes

E.g. higher intercept may mean steeper slope

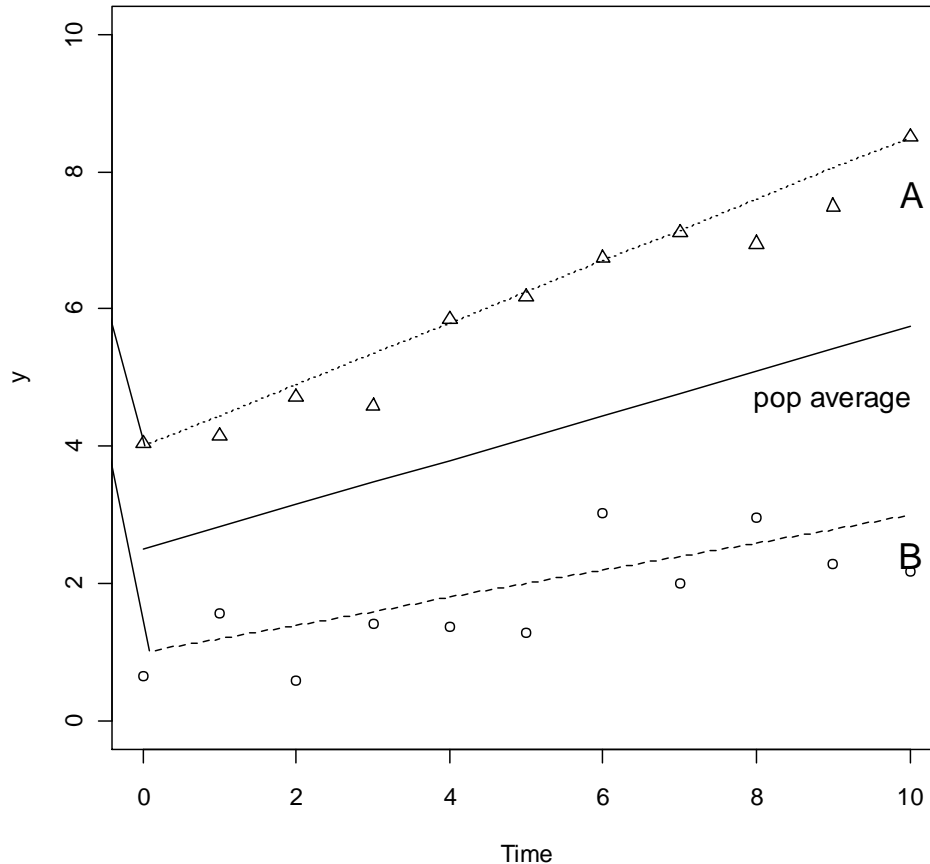
Can allow any subset of the regression parameters to vary randomly

Intercepts and slopes assumed to have a multivariate normal distribution

Prediction of Random Effects

- Usually most interested in population parameters (the β 's)
- But can also “estimate” or predict the subject-specific effects, b_i
- known as
 - best linear unbiased predictors (BLUPs) or
 - Empirical Bayes (EB) estimates
- These are “shrunk” toward the population mean for each individual
- Less shrinkage when n_i is large and when σ_b^2 is large relative to σ_e^2

Random Intercepts and Slopes Model



$$\beta_0 = 2.5$$

$$b_{01} = 1.5$$

$$b_{02} = -1.5$$

$\beta_1 = 0.29$ (y increases by 2.9 units over the time period)

$$b_{11} = .41$$

$$b_{12} = .18$$

$$Y_{it} = \underbrace{\beta_0 + \beta_1 x_{it}}_{\text{Population line}} + \underbrace{b_{0i}}_{\substack{\uparrow \\ \text{Random} \\ \text{intercepts}}} + \underbrace{b_{1i} x_{it}}_{\substack{\uparrow \\ \text{Random} \\ \text{slopes}}} + \underbrace{e_{it}}_{\substack{\uparrow \\ \text{Within-subject} \\ \text{error}}}$$

SAS code – predicting random effects

```
PROC MIXED;  
  CLASS id group;  
  MODEL y = group time group*time / SOLUTION OUTPRED=yhat;  
  RANDOM INTERCEPT time / SUBJECT=id TYPE=UN SOLUTION G;  
RUN;
```

- **OUTPRED=yhat**
 - saves a subject's predicted response profile
 - yhat is the SAS dataset name
- **SOLUTION** option in the **RANDOM** statement
 - Gives the empirical bayes estimates, \hat{b}_i

Limitations

- Need a reasonable number of subjects (>30)
- Flexible in accommodating any degree of imbalance in the data
 - e.g. due to missing data or measurements being taken at different times)
- But validity of results depends on the assumption about missingness
- More..

To Conclude

- Linear mixed effects models are increasingly used for the analysis of longitudinal data
- The introduction of random effects accounts for the correlation among repeated measures
- Appealing because
 - Flexible in accommodating a variety of study designs, data models and hypotheses
 - Of ability to predict individual trajectories over time as well as a population level response



Thank You

Questions?

Event History Analysis

aka Survival Analysis

What?

- A set of statistical procedures for the analysis of time to event data
- Longitudinal data is suited to event history analysis
- Time to event.
 - Can be a bad event: Death, Heart attack, disease, accidents or unemployment.
 - But also can be good event: leaving home, partnering and marriage, gaining a full-time job, exiting unemployment.
- Uses terms such as ‘hazard’, ‘risk’, ‘survival’, ‘failure’ etc.

Why Survival Analysis?

- Can we just use OLS regression?
- Censoring!!!!
- Non-normality, most often 'events' are rare and do not follow a normal distribution.
- What can you do?
 - Estimate time-to event for a group of individuals.
 - Compare different groups (treatment vs. placebo)
 - Study the relationship between the survival time and covariates in the model.

What is Censoring?

- We only have partial information on a person.
 - May have exited the study before study was completed.
- Types of censoring.
 - Right Censoring.
 - Lost to follow up.
 - Changes address but does not inform researchers
 - Death from cause unrelated to cause of interest.
 - Death from a car accident rather than heart disease.
 - Still not “failed at end of the study.
 - Left Censoring
 - Event of interest happens before recording begins.
 - Interval Censoring
 - Event of interest happens between 2 inspection times.

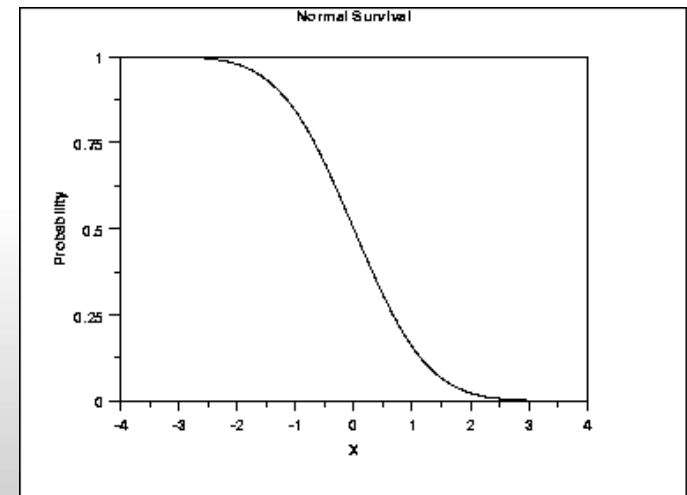
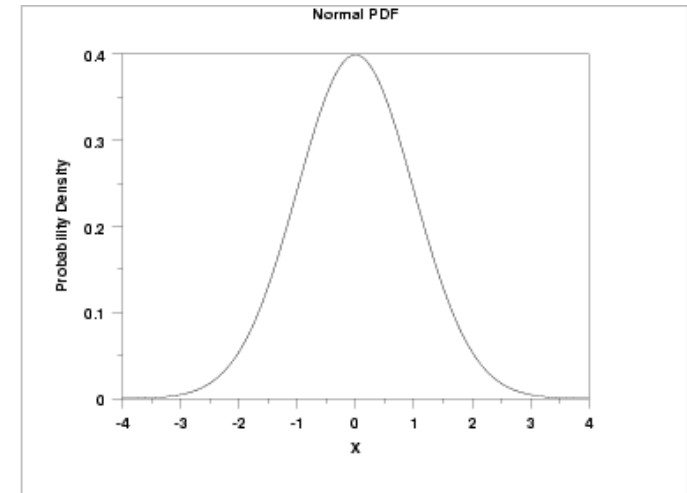
Basic Concepts 1

- Probability density function
 - The probability of the failure time occurring at exactly time t (out of the whole range of possible t 's)

- $$f(t) = \lim_{\Delta t \rightarrow 0} \frac{P(t \leq T < t + \Delta t)}{\Delta t}$$

- Survival time function

- Cumulative Survival
- $S(t) = 1 - F(t)$ where $F(t)$ is the CDF of $f(t)$
- $S(t) = P(T > t)$

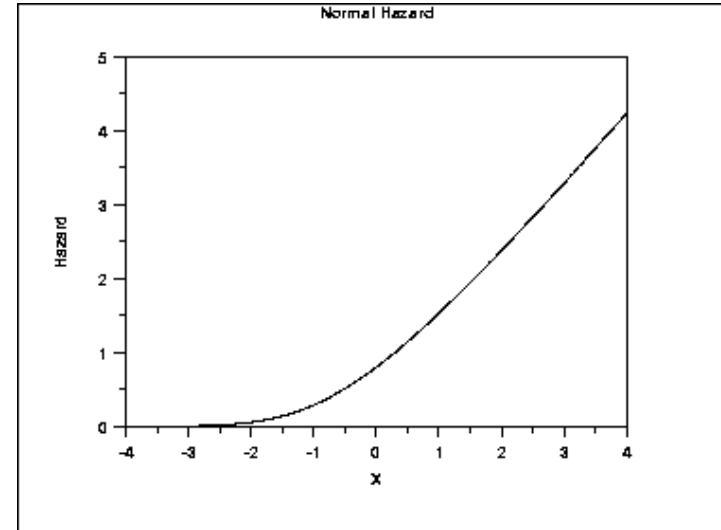


Basic Concepts 2

- Hazard Function

- The probability that **if you survive to t** , you will succumb to the event in the next instant.

- $$h(t) = \lim_{\Delta t \rightarrow 0} \frac{P(t \leq T < t + \Delta t | T \geq t)}{\Delta t}$$



- Relationships

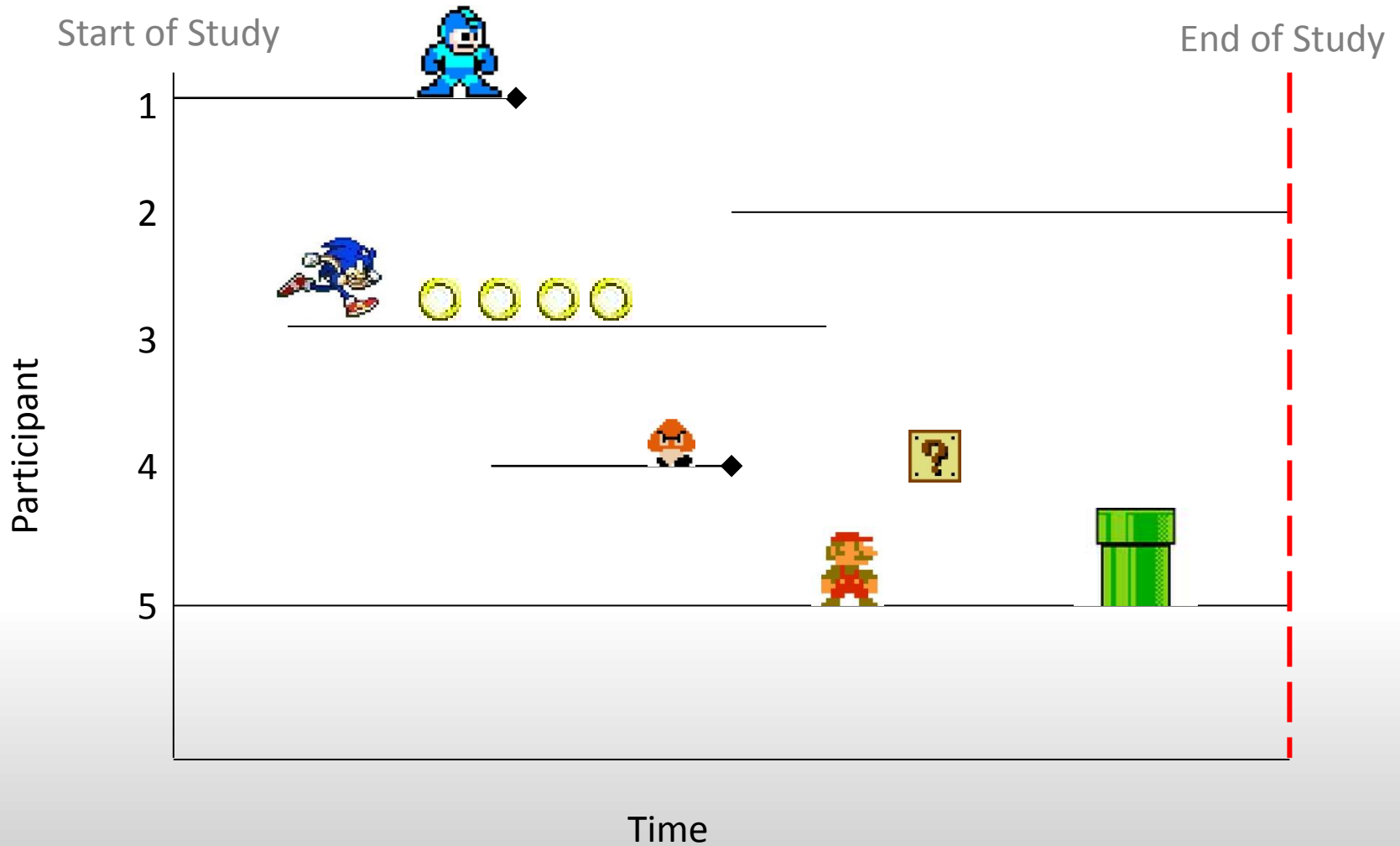
$$h(t) = \frac{f(t)}{S(t)}$$

$$S(t) = \exp\left(-\int_0^t h(u) du\right)$$

$$h(t) = \frac{-d \log S(t)}{dt}$$

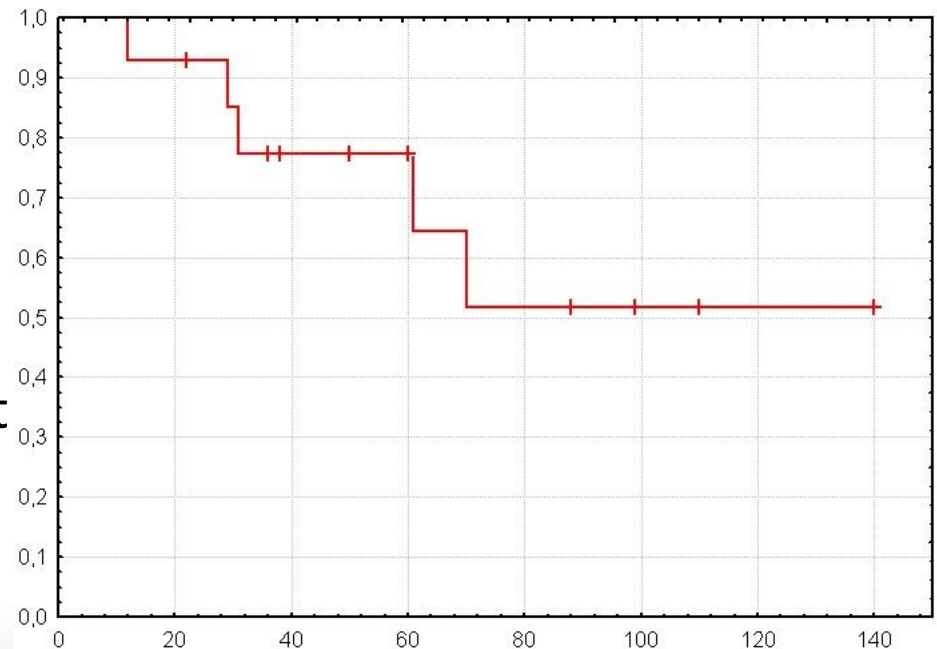
$$f(t) = h(t) \exp\left(-\int_0^t h(u) du\right)$$

Time to heart attack

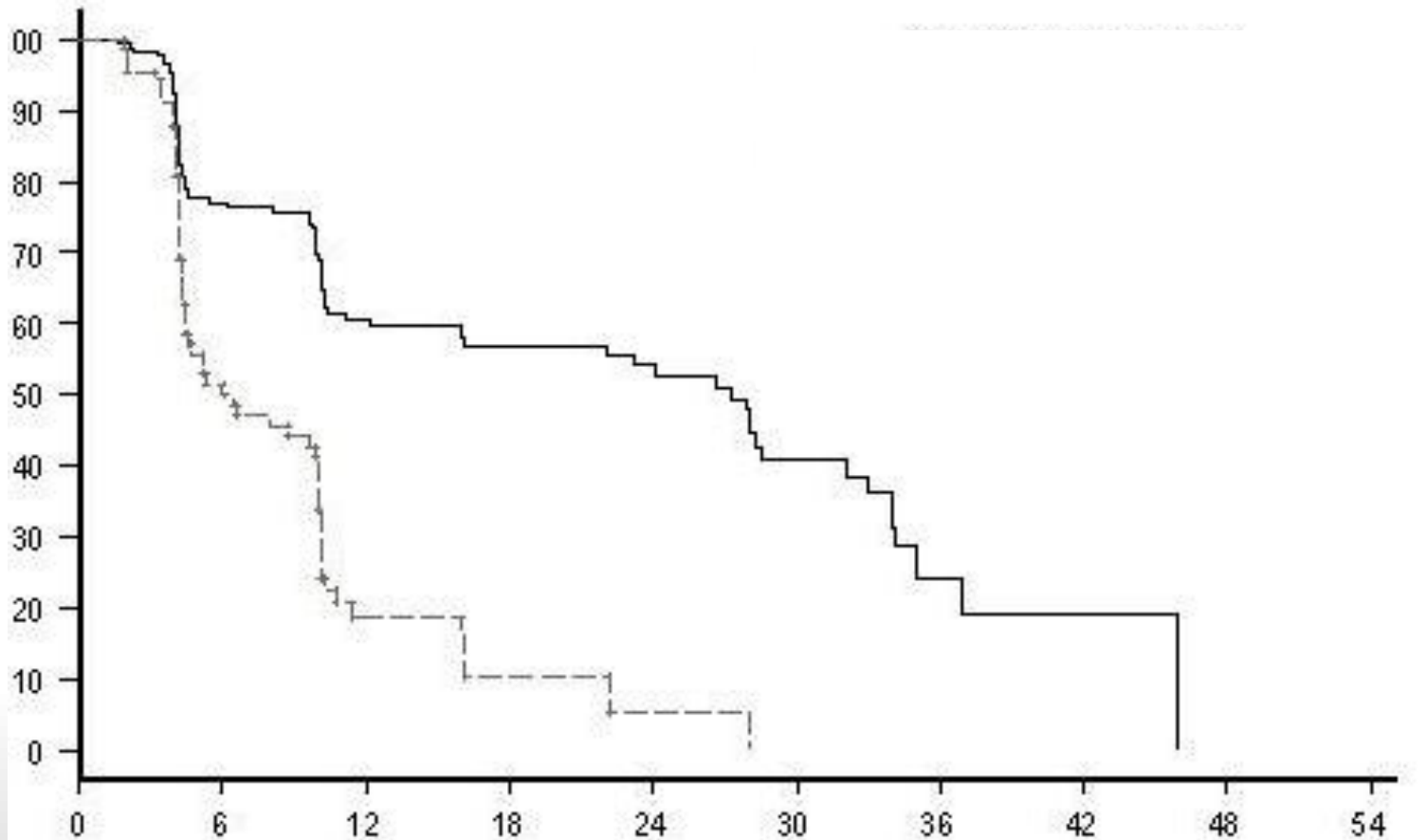


Kaplan-Meier curves

- Non parametric estimation of the survival function.
- No math assumptions (about either the underlying hazard function or about proportional hazards).
- The empirical probability of surviving past certain times in the sample (taking into account censoring).
- Used to describe survivorship of study population/s.
- Used to compare two study populations.



Kaplan-Meier curves



Accelerated Time Failure Models (ATF)

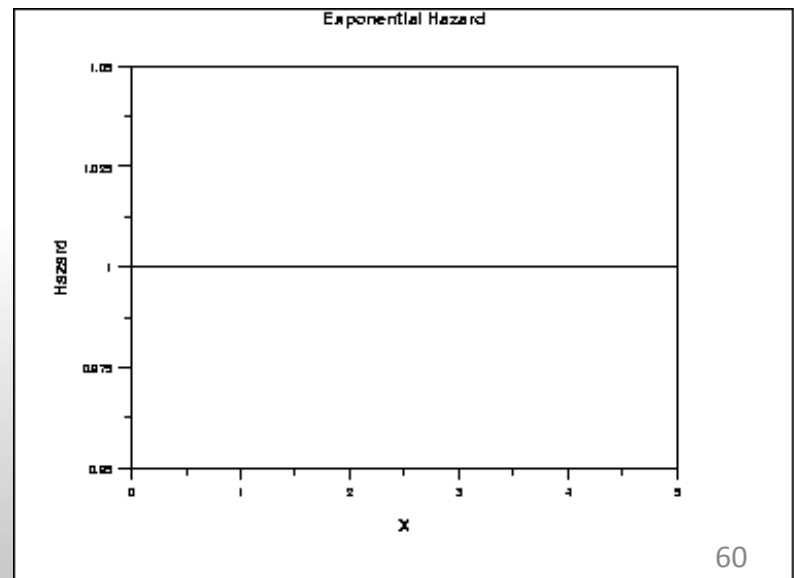
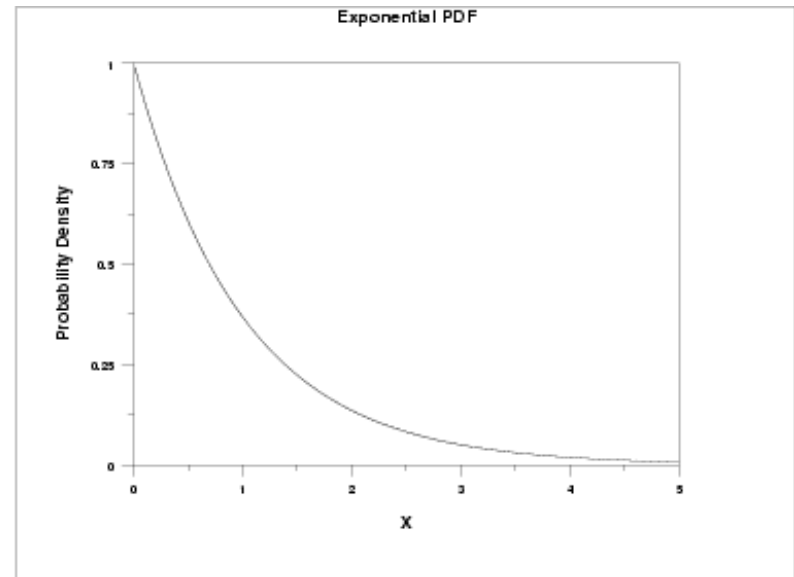
$$\text{Log}T_i = \beta_0 + \beta_1 x_{i1} + \dots + \beta_k x_{ik} + \sigma \varepsilon_i$$

Where:

- T_i is the time for the event of interest by individual i .
 - $x_{i1} \dots x_{ik}$ are the values of k covariates for individual i .
 - ε_i error term.
 - $\beta_0 \dots \beta_k$ and σ are the parameters to be estimated.
-
- Models survival time.

ATF Exponential Distribution

- Exponential is the “original” distribution in survival theory.
- Constant hazard function (flat).
 - Memoryless property
 - Prob. of failure in next time interval given age does not depend on age.
- $h(t) = \lambda$
- Constant hazards not realistic

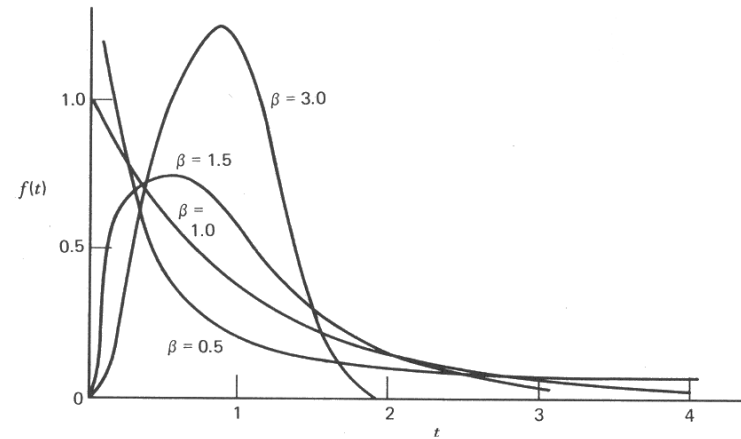


ATF Weibull Distribution

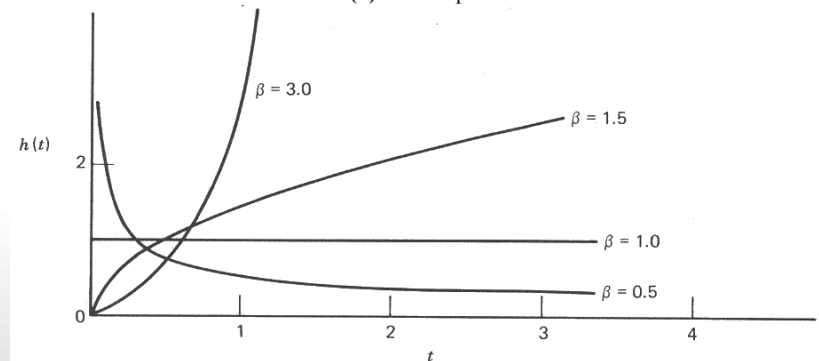
- Weibull Distribution
 - Most common used.
 - Includes the exponential distribution as a special case.

$$h(t) = \lambda \gamma t^{\gamma-1}$$

- Gamma Distribution is sometimes used also.



(b) Weibull pdf



(a) Weibull hrf (power function of time)

Cox Regression – Proportional Hazards Models

$$h_i(t) = h_0(t) e^{\beta_1 X_{i1} + \dots + \beta_k X_{ik}}$$

- Covariates act multiplicatively on a baseline hazard $h_0(t)$.
- Relative risks between individuals are constant over time
- So if RR comparing a male to a female is 1.5, male has a 1.5 times the risk **at all times**.
- Does not require choice of a particular distribution.
- Can include time variant predictors, the accelerated time failure models do not allow this.
- However, it is based on the assumption that the hazards for a different subgroup are proportional to each other.

Comparing PH and AFT models

- Accelerated time failure models :
 - Covariates act multiplicatively on “average” survival time.
- Proportional hazards:
 - Covariates act multiplicatively on risk at any time.
- Big advantage of PH
 - Can estimate β 's without making any assumptions about the form of $h_0(t)$ i.e. no distributional assumptions.
- SAS code:
 - PROC LIFETEST : Kaplan-Meier
 - PROC LIFEREG : Accelerated time failure model
 - PROC PHREG : Proportional hazards

Hybrid model & lessons learnt

Roy Lay-Yee

Outline

1. Thanks to the team for doing the hard parts ... helping us to understand the various methods
2. My part is trying to integrate them to meet our needs (apologies to other colleagues)
3. Preface - longitudinal data analysis
4. Comparing fixed effects (FE) & random effects (RE) models (apologies to EHA)
5. Hybrid (FE & RE) model - a promising lead
6. Lessons learnt

Our needs: driving dynamic micro-simulation

What is dynamic micro-simulation modelling?

1. Start with a base sample of individual units, e.g. children
2. Each person has a set of initial attributes, e.g. gender, ethnicity
3. Then probabilistic rules are applied in a each year to these persons to mimic changes in state & behaviour, e.g. how is children's GP visiting influenced by family circumstances (adjusting for other factors)
4. This produces simulated estimates of outcomes, both aggregate and distributional, e.g. average number of GP visits broken down by gender & ethnicity

Longitudinal data analysis informs the rules!

1. Use person-year data (years nested in person)
2. Want measures of effect and variation

Terminology

$$Y \sim X + E$$

Y: **outcome**, dependent variable, response

X: **predictor**, independent variable, explanatory variable, covariate

X: **observed**, measured

vs E: **un-observed**, unmeasured, omitted, e.g. ability

X: **time-variant** (changing over time), e.g. student achievement

vs **time-invariant** (stable) variables, e.g. gender, ethnicity

E: **error**, disturbance, residual, unobserved / unmeasured / omitted variables

Longitudinal data: reminder

- Repeated observations over time on same persons, e.g. person-year data
- Observations on same person are not independent
- Assume data from different persons are independent

Longitudinal data analysis: Considerations, esp. FE and RE models

- *Number of people, number of time points?*
- *Are data binary, categorical or continuous?*
- *Are data evenly spaced in time?*
- *Are data balanced vs unbalanced (missing data by time)*
- *What hierarchies (levels) are present in data?*
- What kinds of **correlation** expected?
- What **variation** present and have to be accounted for?
- What **effects** do we want to estimate?

FE and RE models: similarities

- Closely related mathematically
- Both models account for within-person variation
- Both can estimate effects of observed time-variant predictors

FE and RE models: differences

Considerations		Model	
		Random Effects	Fixed Effects
Correlation	Intercept	uncorrelated with X	correlated with X
	Error	must be uncorrelated with X	can be correlated with X
Variation	Within-person	yes	yes
	Between-person	yes	no
	Precision	smaller standard errors	larger standard errors
Effects (estimation)	Bias	more	<ul style="list-style-type: none"> • less • controls for all time-invariant, both observed & un-observed (differenced out)
	Observed	time-variant & time-invariant variables	<ul style="list-style-type: none"> • estimates only time-variant • controls for time-invariant • but can include interactions between time-invariant and time-variant
	Un-observed (controlled)	no	controls for time-invariant variables

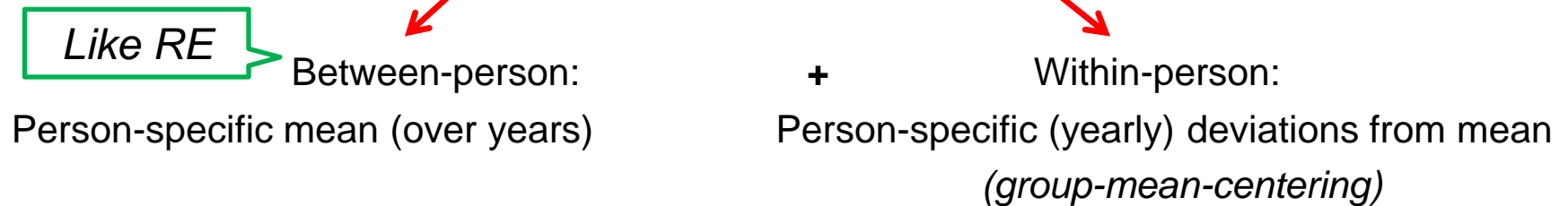
A 'Hybrid' model

- We had heard of this mythical beast before
- I asked Gary
- He said:
 - FE ~ OLS on person-year data
 - RE ~ FE with random intercept
 - Hybrid ~ FE + RE
- He said read all about it here:
 - Paul Allison - Fixed effects regression methods for longitudinal data using SAS, 2005.
- I read his copy ...

(FE& RE) Hybrid model

- Hybrid ~ FE + RE
- $Y \sim X$
- On left-hand-side (Y): can handle various types of outcome variable?
- On right-hand-side (X): *observed time-variant* + *observed time-invariant*

decomposes components



- Controls for un-observed time-invariant variables *Like FE*
- Reference: Paul Allison - Fixed effects regression methods for longitudinal data using SAS, 2005.

What can models do for us?

Considerations	Models		
	Random effects	Fixed effects	Hybrid
Variation			
<i>Within-person</i>	yes	yes	yes
<i>Between-person</i>	yes	no (differenced out)	yes
Effects			
Observed (estimated)			
<i>Time-variant</i>	yes	yes	yes
<i>Time-invariant</i>	yes	no (but yes interaction w time-variant)	yes
Unobserved (controlled)			
<i>Time-variant</i>	no	no	no
<i>Time-invariant</i>	no	yes	yes
State dependence (lagged variables)	no	no	?

Hybrid model pros & cons

- ✓ Accounts for both within- and between-person variation (like RE)
- ✓ Can estimate effects of observed time-variant & observed time-invariant predictors (like RE)
- ✓ Controls for un-observed time-invariant variables (like FE)
- Can handle all types of outcome variable?
- Observed time-variant predictor is in 2-component form of mean & mean-deviation – difficult to implement?
- Relies on change over time
- Cannot include lagged variable as predictor?
- Other assumptions and limitations?

Lessons learnt from the course

- Thanks to NZSSN and especially Gary Marks
- Taught us about various methods for longitudinal analysis:
FE, RE, EHA
- Particularly applied focus using real data and SAS
- Referred us to the Hybrid model (based on FE and RE) –
we will investigate
- May or may not be appropriate solution to our situation
- But an excellent point of departure