Quantitative Data Analysis: A Companion for Accounting and Information Systems Research

Teaching Materials

Created by Willem Mertens, Amedeo Pugliese & Jan Recker

Teaching Notes: Quantitative Data Analysis ~ © Copyright 2017 W. Mertens, A. Pugliese & J. Recker. All Rights Reserved. ~

Copyright Notice

© Copyright 2017 W. Mertens, A. Pugliese & J. Recker. All Rights Reserved.

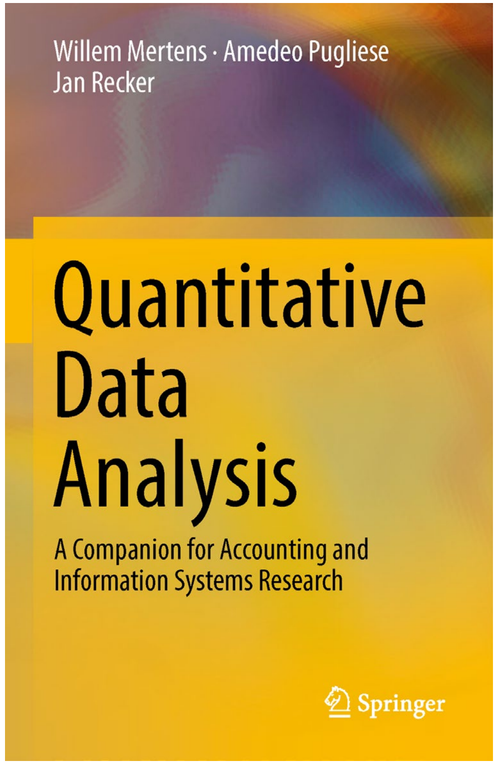
Part 6:
Time-Series and Longitudinal Analysis

 ~ © Copyright 2017 W. Mertens, A. Pugliese & J. Recker. All Rights Reserved. ~

What these materials are about

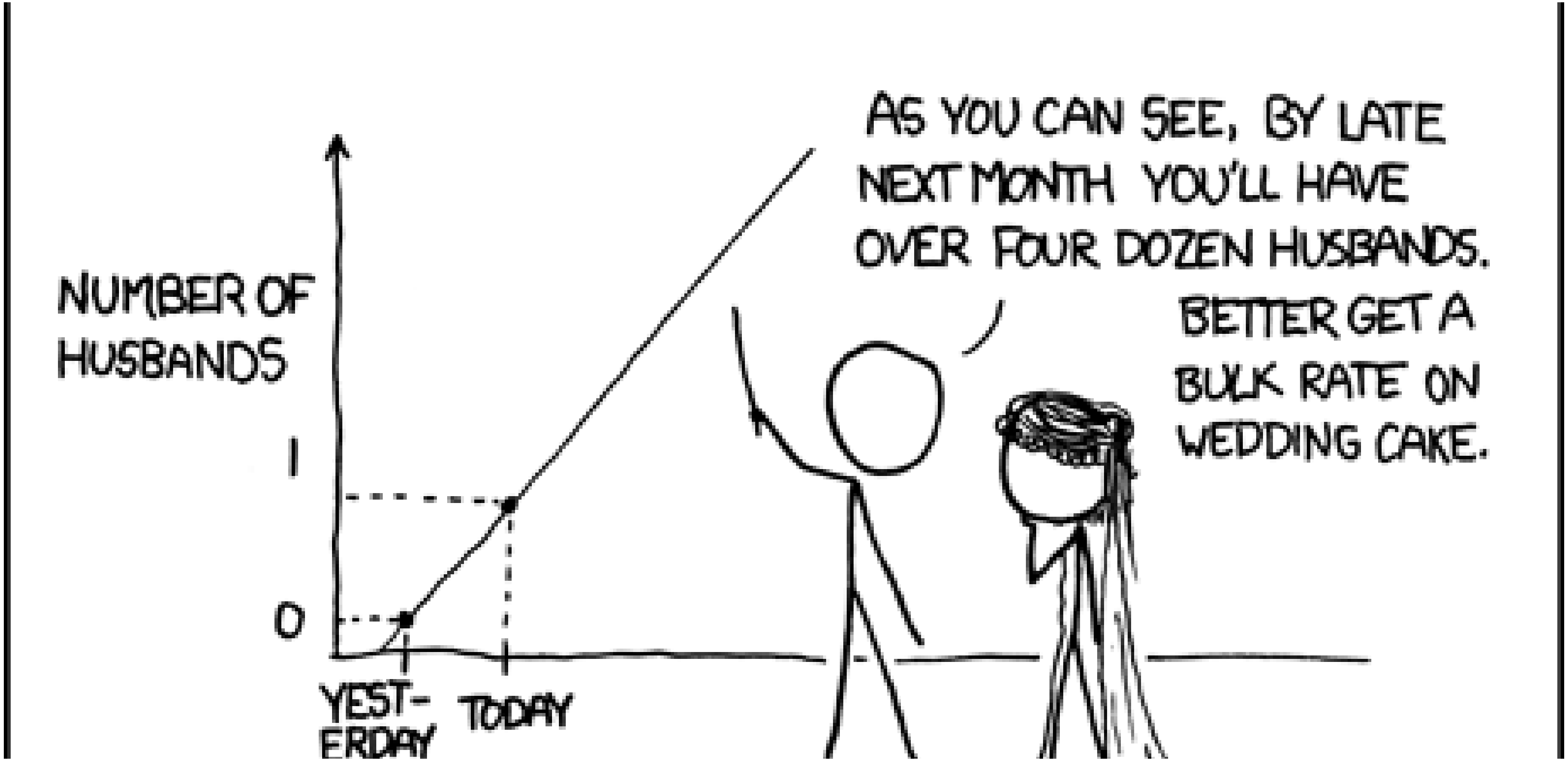
Offering a guide through the essential steps required in quantitative data analysis

1. Introduction
2. Comparing Differences Across Groups
3. Assessing (Innocuous) Relationships
4. Models with Latent Concepts and Multiple Relationships: Structural Equation Modeling
5. Nested Data and Multilevel Models: Hierarchical Linear Modeling
6. **Analyzing Longitudinal and Panel Data**
7. Causality: Endogeneity Biases and Possible Remedies
8. How to Start Analyzing, Test Assumptions and Deal with that Pesky p-Value
9. Keeping Track and Staying Sane



 ~ © Copyright 2017 W. Mertens, A. Pugliese & J. Recker. All Rights Reserved. ~

Longitudinal data and time series



Recap: causality

When can we conclude causal effect of A on B?

- There is a relation between A and B (covariance)
- Change in A precedes change in B in time (or is simultaneous)
- There is no other variable C that explains the change in B

Recap of previous sessions

1. Data exploration and testing
   - Structure data and delete unreliable (*not* inconvenient) cases
   - Explore descriptives and assumptions
     - Homoscedasticity
     - Independence
     - Linearity

2. Regression and ANOVA models
   - Cross-sectional analyses of
     - Linear relations between continuous variables
     - Differences between groups

3. Endogeneity and how to control for it
   - What is causality?
   - Propensity score matching: was it the treatment or the assignment to groups?
   - Instrumental variable estimation: control for ‘reverse’ causation

Most common assumptions for linear analyses

- Independence
  - Data was collected from independent sources
  - Variable measurements were independent (e.g. regression)
- Homoscedasticity/homogeneity of variance
  - Variance is equal in different (sub-)samples
- Normality
  - Sampling distribution/errors/data follow a normal distribution --> have limited skew and kurtosis

What is Longitudinal Analysis?

Longitudinal analysis is*“ research emphasizing the study of change and containing at minimum three repeated observations(although more than three is better) on at least one of the substantive constructs of interest” …*

“a study that measures the independent variable at Time 1 and the dependent variable at Time 2 […] is simply a variant of the cross-sectional design.”

(Vandenberg and Ployhart 2010)

“a rule of thumb for distinguishing between [time series and longitudinal data] is that time series have more repeated observations than subjects while longitudinal data have more subjects than repeated observations”

(Chuck Huber, 2013, http://blog.stata.com/tag/longitudinal-data/)

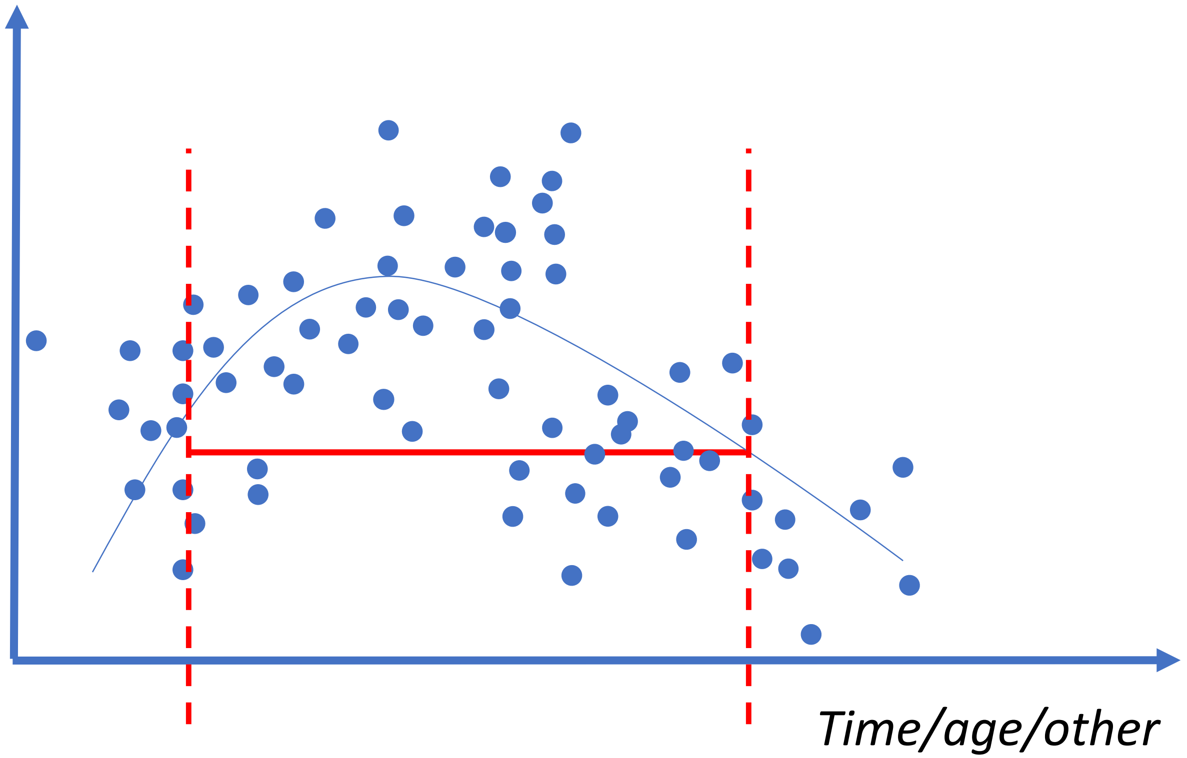
What is Longitudinal Analysis?(cont'd)

Why 3?

- Change can be non-linear
- To distinguish change from measurement error

Focus on CHANGE in IV/DV

- Time is no IV
- Simply "time" is often not the best metric
  - Age
  - Time since …
  - Quarter



Metric must be monolithic, i.e. non-reversible (e.g. body weight is no good)

Before you start: why and how?

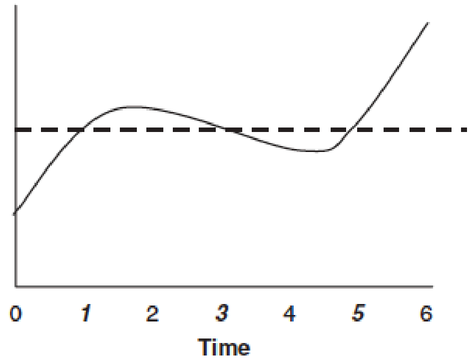
- Is there a reason to expect change over time?
- What will the change look like?
  - Linear
  - Non-linear
  - Discontinuous

Make sure measures are
- comparable across time (e.g. be careful for inflation)
- valid across time (e.g. FTE as a measure of business size)
- reliable across time (e.g. older people and dementia)

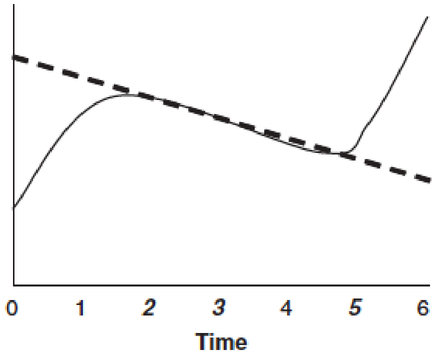
- Will all units/participants change in the same way?
  - Interested in group mean change?
    → repeated measures ANOVA
  - Interested in inter-unit differences in intra-unit change?
    → random coefficient modeling (RCM) or latent growth modeling (LGM)
  - Interested in both?
    → multilevel models for change

Before you start: why and how? (cont’d)

- Descriptive vs. explanatory design
  - Will the change be related to other variables?
  - Why? And how?
  - Usually both: how does it change and why
- Experimental vs. observational design
  - Observational (i.e. not random)
    - Mind endogeneity
    - E.g. panel data
  - Experimental
    - Balance order of conditions (if possible)
- Retrospective vs. proactive design
- Fixed vs. flexible intervals
- Enough measurements and at sensible times (e.g. seasonal differences)



a. The wrong measurement occasions miss the trend



b. Too few measurement occasions misrepresent the trend

Ployhart and Vandenberg, 2010

Before you start: your data

- Organising data
  - Multivariate/unit-level data: one variable for each metric for each measurement point

Case	Revenue Q1	Revenue Q2	…
SMB 1	$200.000	$250.000	…
SMB 2	$700.000	$600.000	…
…	…	…	…

  - four problems

1. leads naturally to non-informative summaries (e.g. wave-to-wave correlations)
2. omits an explicit “time” variable
3. is inefficient, or useless, when the number and spacing of waves varies across individuals
4. cannot easily handle the presence of time-varying predictors (e.g. year since founding)

(Singer and Willett, 2003, p. 20)

Before you start: your data (cont'd)

- Organising data
  - Univariate/unit-period level: one case for each unit at each measurement point

Case	Quarter	Revenue	…
SMB 1	Q1	$200.000	…
SMB 1	Q2	$250.000	…
SMB 2	Q1	$700.000	
…	…	…	…

  - Typical variables: identifier (identical across time), time metric, IV, DV, controls
  - Time-invariant IV's have the same value across time

Descriptive analyses

- When *n* is small: empirical growth plots (within-unit evolution)
  - X is time, Y is outcome
  - Fit a line to explore
    - Non-parametrically: close to reality, assumption-free
    - Parametrically: assumptions, but fit statistics
- When *n* is large: scatter with regression/line fitted
  - Inter-unit variation indicated by R^2
  - Overall trend indicated by line
- To explore role of predictors/controls: stratify per level
  - E.g.: trends for men vs. women

Longitudinal analysis: options we discuss today

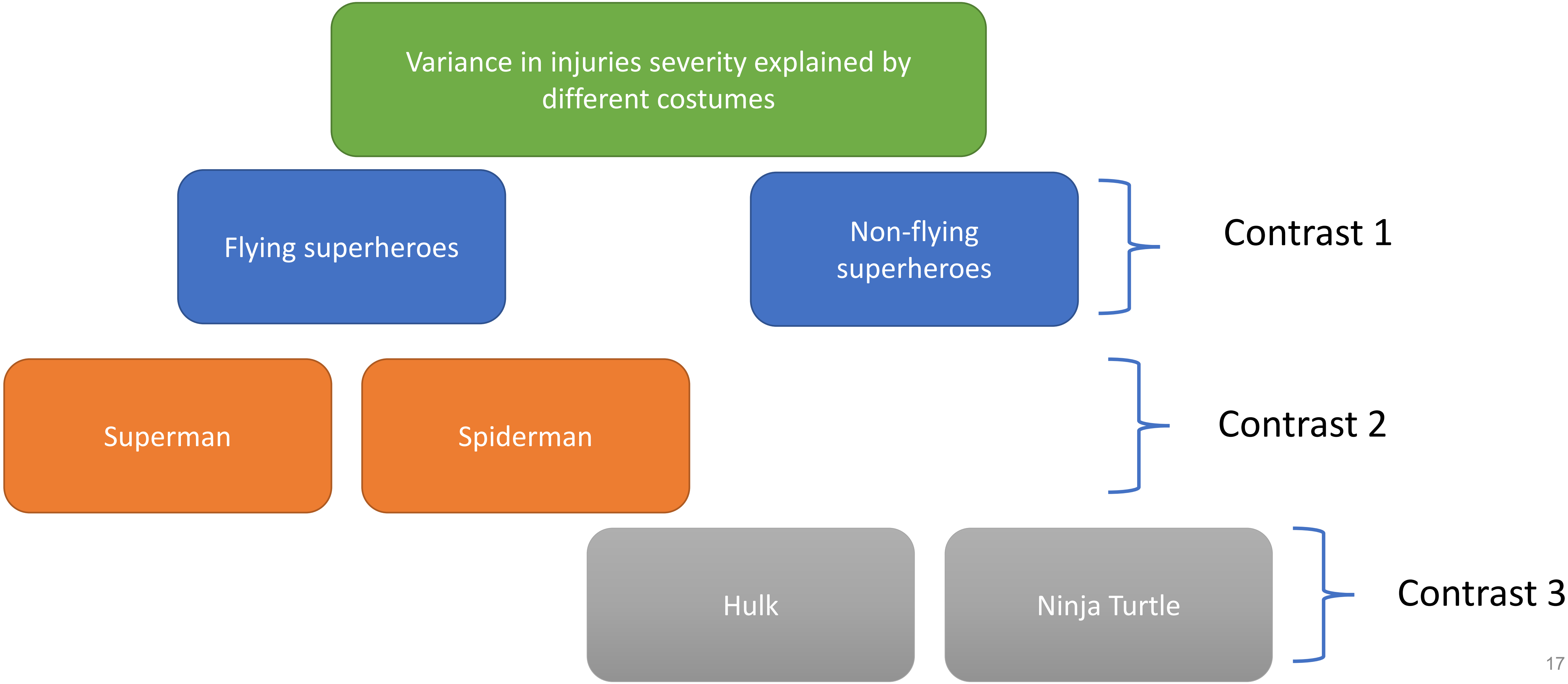
1. Differences between measurement points
   - Repeated-measures GLM/ANOVA

2. Differences between measurement points and (groups of) units
   a. Random Coefficient Modeling (commonly known as multilevel/hierarchical linear models)
   b. Latent Growth Modeling

3. Prediction of events
   - Survival analysis

4. Prediction of evolution
   - Time series analysis

1. Differences between measurement points



Repeated-measures ANOVA

- Share of variance that is explained by manipulation/time of measurement
- Assumption of independence no longer holds
- Extra assumption: sphericity
  - Equal variance of the differences between conditions
  - Test: Mauchly's test (H_0 = equal variances of differences between conditions)
  - If violated: adjust degrees of freedom through
    - Greenhouse-Geisser correction
    - Huynh–Feldt correction (less conservative)
    - Lower-bound: avoid use

Example

Descriptive Statistics

	Mean	Std. Deviation	N
Knowledge_2011	5.78	1.886	98
Knowledge_2012	4.85	2.688	98
Knowledge_2013	5.18	2.391	98

Mauchly's Test of Sphericity[a]

Measure: MEASURE_1

Within Subjects Effect	Mauchly's W	Approx. Chi-Square	df	Sig.	Epsilon[b] Greenhouse-Geisser	Epsilon[b] Huynh-Feldt	Epsilon[b] Lower-bound
Year	.956	4.354	2	.113	.958	.976	.500

Tests the null hypothesis that the error covariance matrix of the orthonormalized transformed dependent variables is proportional to an identity matrix.

Tests of Within-Subjects Effects

Measure: MEASURE_1

Source		Type III Sum of Squares	df	Mean Square	F	Sig.
Year	Sphericity Assumed	43.313	2	21.656	4.340	.014
	Greenhouse-Geisser	43.313	1.915	22.617	4.340	.016
	Huynh-Feldt	43.313	1.953	22.180	4.340	.015
	Lower-bound	43.313	1.000	43.313	4.340	.040
Error(Year)	Sphericity Assumed	968.020	194	4.990		
	Greenhouse-Geisser	968.020	185.763	5.211		
	Huynh-Feldt	968.020	189.418	5.111		
	Lower-bound	968.020	97.000	9.980		

Tests of Within-Subjects Contrasts

Measure: MEASURE_1

Source	Year	Type III Sum of Squares	df	Mean Square	F	Sig.
Year	Level 1 vs. Level 2	84.500	1	84.500	7.063	.009
	Level 2 vs. Level 3	11.112	1	11.112	1.159	.284
Error(Year)	Level 1 vs. Level 2	1160.500	97	11.964		
	Level 2 vs. Level 3	929.888	97	9.586		

Longitudinal analysis: Options

1. Differences between measurement points
   - Repeated-measures GLM/ANOVA

2. Differences between measurement points and (groups of) units
   a. Random Coefficient Modeling (commonly known as multilevel/hierarchical linear models)
   b. Latent Growth Modeling

3. Prediction of events
   - Survival analysis

4. Prediction of evolution
   - Time series analysis

2.a Random Coefficient Modeling

- Multilevel model
  - Level 1: within-unit change
  - Level 2: between-unit differences

$$\text{Level 1: } Y_{ti} = \pi_{0i} + \pi_{1i} T_{ti} + e_{ti}$$

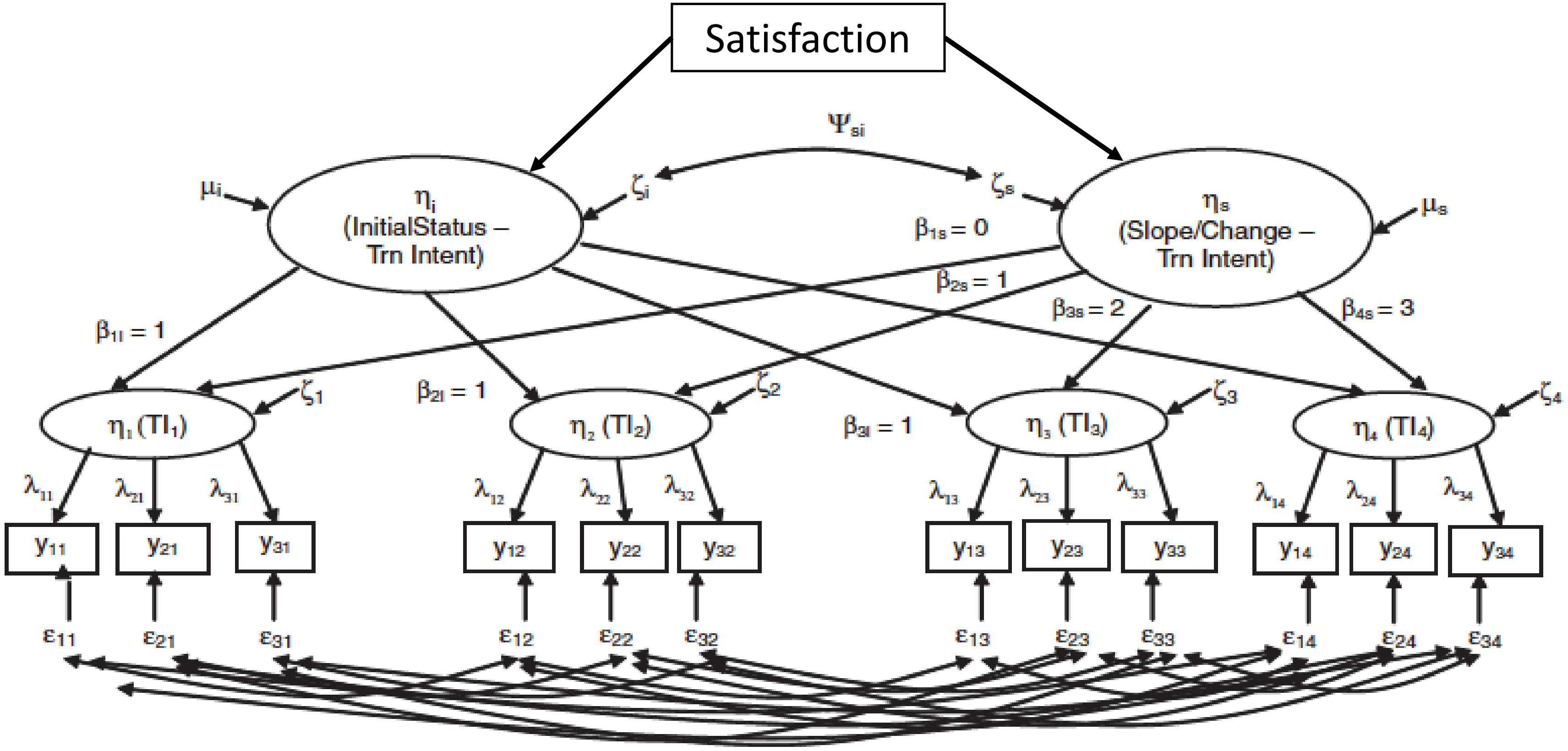
$$\text{Level 2: } \pi_{0i} = \beta_{00} + \beta_{01} X_i + r_{0i}$$

$$\pi_{1i} = \beta_{10} + \beta_{11} X_i + r_{1i}$$

Ployhart and Vandenberg, 2010

2.b Latent Growth Modeling

- More or less a combination of SEM and RCM
  - Structural equation model where latent variables are intercept and slope coefficients



Latent Growth Modeling

- Advantages
  - accounts for measurement error in the estimation process: errors are modelled to be related across time
  - more flexible (e.g. mediation, moderation)
  - allows testing for longitudinal validity: invariance testing
    - making sure that each construct is measured equally at each point in time
    - only possible if multiple indicators per construct are available

“factorial invariance in longitudinal models concerns whether relations between latent variables and their manifest indicators are invariant across occasions. Stated differently, the expected value of a person’s score on manifest variable j at time t should be a function of her score on the latent variable and the associated unique factor at time t, and should not additionally depend on time of measurement”

(Widaman, K. F., Ferrer, E., & Conger, R. D., 2010)

Longitudinal analysis: Options

1. Differences between measurement points
   - Repeated-measures GLM/ANOVA

2. Differences between measurement points and (groups of) units
   a. Random Coefficient Modeling (commonly known as multilevel/hierarchical linear models)
   b. Latent Growth Modeling

3. Prediction of events
   - Survival analysis

4. Prediction of evolution
   - Time series analysis

3. Survival analysis

- Purpose: to predict *whether* and *when* an event will take place

- Typical elements
  - Target event, e.g. bankruptcy, turnover, unemployment
  - Starting point, e.g. data breach, major restructuring, starting on first job
  - Metric for time, e.g. quarters, fortnights, years
  - Life table: probability of event for each given time interval

- One issue: censoring
    - Units that don't experience event during data collection period
    - Units that drop out because of other reason than event
  - Non-informative vs. informative censoring: why did they drop out?
  - Right- vs. left-censoring: end of data collection/drop-out vs. start of data collection/inclusion unknown or repeatable events

Survival analysis with discrete-time data

- Discrete-time means there are set measurement points, with no data in between
- Three functions can be estimated:
  - The hazard function:
    - The probability that a unit will experience the event in a certain time interval
    - Similar to proportion that drops out in a certain time interval (in its simplest form)
    - Odds ratio: probability that it will happen/1-probability that it will happen
  - The survivor function
    - The probability of ‘survival’: probability that individual will not experience event past a certain point in time
    - Similar to proportion of units that experienced the event before the end of the time interval out of the total number of units in the data set
  - The median lifetime
    - The value of time for which the value of the estimated survivor function is .5

Survival analysis with continuous-time data

- Problem:
  - ‘hazard’ is infinitely small because time intervals are infinitely small
  - more useful measure: hazard rate
    - the rate with which hazard increases over time
    - e.g. bankruptcies occur at a rate of 78 per quarter
- Allows estimating effects of IV’s on evolution of hazard rate and survival
  - E.g. people with low lecture satisfaction have higher unemployment rate and lower median survival

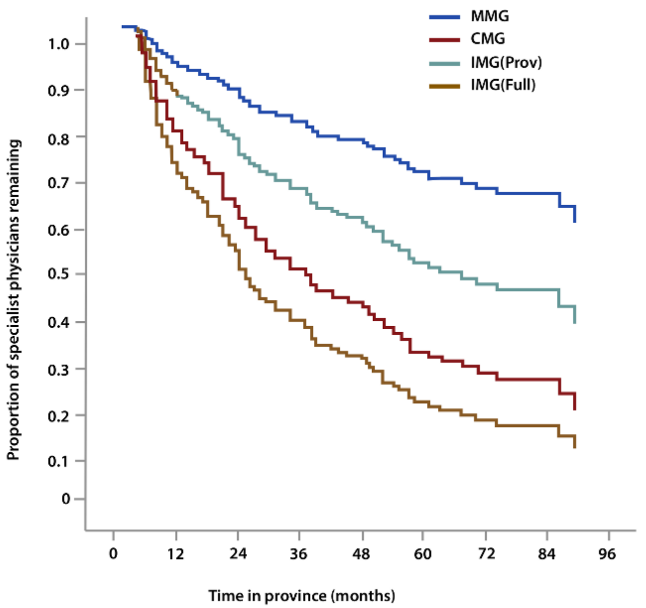


Figure 3
Survival curve obtained from Cox regression analysis of the length of practice in NL of specialist physicians in the 2000–04 cohort (n = 180), differentiated by physician group

Longitudinal analysis: options we discuss today

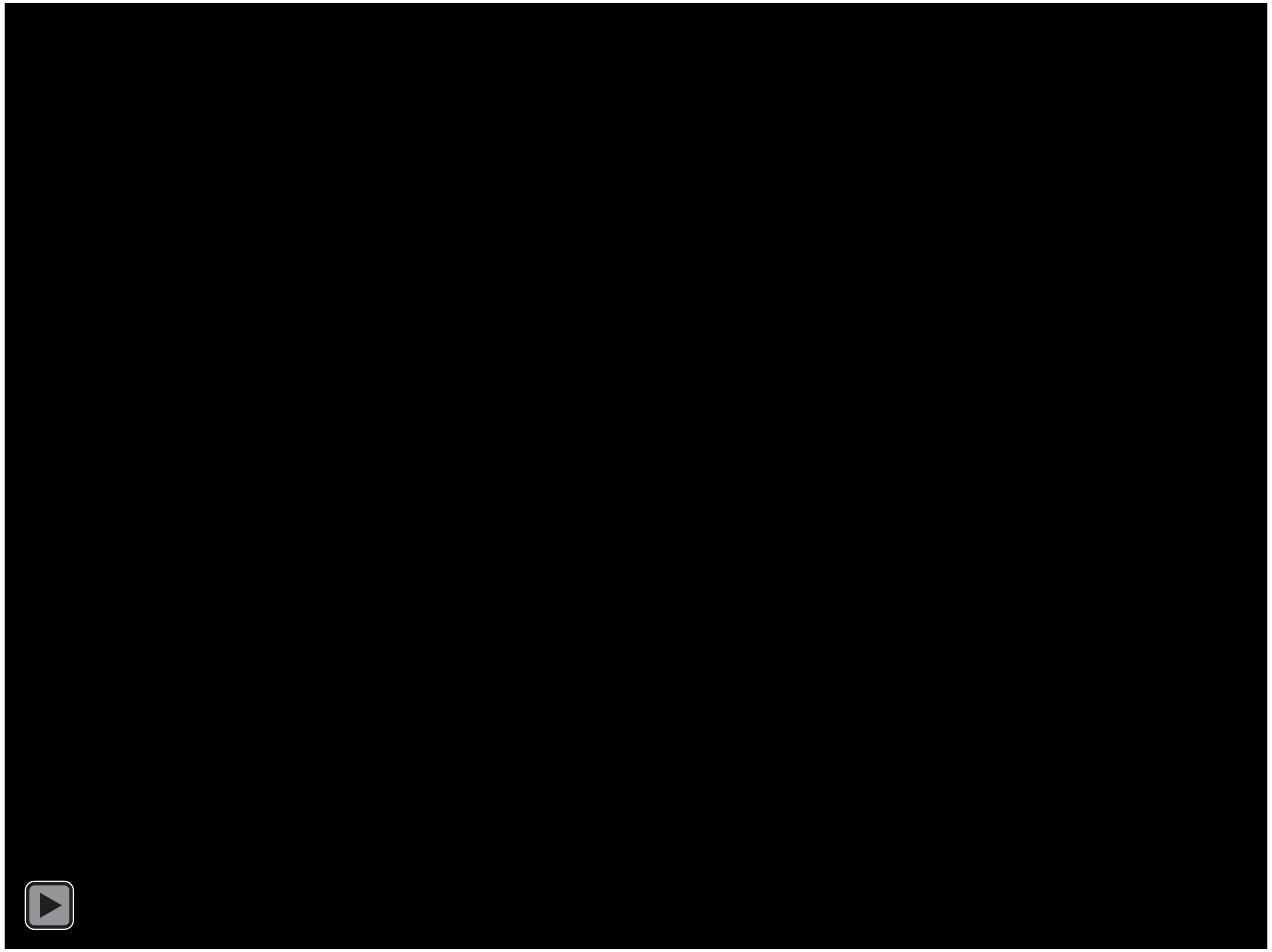
1. Differences between measurement points
   - Repeated-measures GLM/ANOVA

2. Differences between measurement points and (groups of) units
   a. Random Coefficient Modeling (commonly known as multilevel/hierarchical linear models)
   b. Latent Growth Modeling

3. Prediction of events
   - Survival analysis

4. Prediction of evolution
   - Time series analysis

4. Time series analysis



Time series analysis: the basics

- Nature of data: many measurement points
- Goal
  - To describe trends
  - To predict trends
- How?
  - Descriptive: as above
  - Interpolation: Estimate values between measurement points
  - Extrapolation: Estimate values after measurement points

    e.g. professor Osborne

    → by fitting linear and non-linear models to the data

Time Series Analyses: the basics (cont'd)

- Basic assumption:

$$Y_t = T_t + Z_t + S_t + R_t, \qquad t = 1, \ldots, n.$$

  - Y_t = variable of interest — e.g. revenue
  - T_t = monotone trend — e.g. growth
  - Z_t = long term cyclic trend — e.g. recession, recovery, growth, decline
  - S_t = short term cyclic trend — e.g. seasonality
  - R_t = random variance (error) — e.g. emotional CEO
  - $G_t = T_t + Z_t$ --> long-term behaviour of time series

Chair of Statistics, University of Würzburg (2012)

Time Series Analyses: the basics (cont'd)

- Autocorrelation:
  - Similarity between observations as a function of the time lag between them
  - Function:
    - *s* and *t* are certain points in time
    - When stationary
      - *t* is time interval
      - Stationary means average and variance do not vary over time

$$R(s,t) = \frac{\mathrm{E}[(X_t - \mu_t)(X_s - \mu_s)]}{\sigma_t \sigma_s}$$

$$R(\tau) = \frac{\mathrm{E}[(X_t - \mu)(X_{t+\tau} - \mu)]}{\sigma^2},$$

Importantly derived from https://en.wikipedia.org/

Time Series Analyses: the basics (cont'd)

- Autocorrelation: *similarity between observations as a function of the time lag between them*
  - Shows seasonality
  - Similar to regression:
    - Value X_7 is a function of values X_{1-6}
    - Linear least square estimates (i.e. minimizing variance)

$$X_t = \mu + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \cdots + \theta_q \varepsilon_{t-q}$$

- Moving average
  - 'Filters out' fluctuations
  - Iterative non-linear fitting procedures (Error is not observed)

- Box-Jenkins
  - Autoregressive moving average
  - Autoregressive integrated moving average

- More complex models can be fitted
  - Logistic
  - Bayesian

Importantly derived from https://en.wikipedia.org/

Resources and references

- https://statistics.laerd.com/statistical-guides/sphericity-statistical-guide-2.php
- Andy Field – Discovering Statistics Using SPPS/R/?
- Vandenberg, Robert J. and Ployhart Robert E. (2010). *Longitudinal Research: The Theory, Design, and Analysis of Change.* Journal of Management 36(1)
- Singer, Judith D. and Willett, John B. (2003). Applied Longitudinal Data Analysis: Modeling Change and Event Occurrence. Published to Oxford Scholarship Online: September 2009. DOI: 10.1093/acprof:oso/9780195152968.001.0001
- Widaman, K. F., Ferrer, E., & Conger, R. D. (2010). Factorial Invariance within Longitudinal Structural Equation Models: Measuring the Same Construct across Time. *Child Development Perspectives*, *4*(1), 10–18. http://doi.org/10.1111/j.1750-8606.2009.00110.x
- Chair of Statistics, University of Würzburg (2012). A First Course on Time Series Analysis: Examples with SAS. http://www.statistik-mathematik.uni-wuerzburg.de/fileadmin/10040800/user_upload/time_series/the_book/2012-August-01-times.pdf

End of Part 6

© Copyright 2017 W. Mertens, A. Pugliese & J. Recker. All Rights Reserved.