THE LONDON SCHOOL OF ECONOMICS AND POLITICAL SCIENCE

Department of Psychological and Behavioural Science

PB4A7: Quantitative Applications in Behavioural Science

Department of Psychological and Behavioural Science
Autumn Semester

Course Information

Instructor:	Dr. Thomas Curran	**Time:**	Wednesdays 11am–1pm
E-mail:	t.curran@lse.ac.uk	**Room:**	See timetable
Office:	CON 3.16	**Seminars:**	Wednesday 2pm–3.30pm & 3.30pm–5pm
Office Hours	Mondays 12pm-1pm	**Room:**	See timetable

Teacher Information

GTA:	Mr. Mohamed Karim Merabtine	**Office Hours:**	On Student Hub
E-mail:	m.k.merabtine@hss23.qmul.ac.uk	**Help Sessions**	Tuesday 1pm-2pm

Course Description

The primary objective of this course is to familiarise students with the comprehensive statistical toolkit necessary to comprehend the multifaceted and individual-level causes of human behaviour and to equip them to conduct their own research. The course will cover leading methods used by psychologists and economists to test behavioural science hypotheses and examine relationships in data. Beginning with essential data cleaning and screening techniques for identifying and handling missing data, outliers, and ensuring data quality, students will master the core statistical foundations of the General Linear Model (GLM), which serves as the unifying framework for t-tests, Analysis of Variance (ANOVA), and regression analysis. Building on these foundations, the course covers sophisticated multivariate analyses, including factor analysis for identifying underlying constructs, structural equation modelling (SEM) for testing theoretical models and examining latent variables, and multilevel modelling for analysing hierarchical data structures common in psychological and economic research. This course complements 'Experimental Design and

Methods for Behavioural Science' (PB413), which covers experimental design and research methods for MSc Behavioural Science students, providing a comprehensive methodological foundation across both experimental and observational research paradigms.

Requirements

Download and install R and R Studio. Throughout the course, we will rely on R for workshops.

Student Hours

You can book office hours for Dr Thomas Curran via Student Hub.

Class Attendance and Participation

The learning process of this class is based on in-class discussion and participation. Attendance and careful preparation of the course material are therefore highly recommended. This includes coming to class on time.

Materials

There are numerous great websites, books, articles, and videos that provide interesting information beyond the lectures. Some of them we will use as required or optional readings for the course. An indicative (but not exhaustive) list:

- Applied Data Skills: Processing & Presenting Data by Emily Nordmann and Lisa DeBruine
- Yarrr! The pirate's guide to R by Nathaniel D. Phillips
- Learning Statistics with R: A Tutorial for Psychology Students and Other Beginners by Danielle Navarro
- Statistical Thinking for the 21st Century by Russell A. Poldrack
- Hayes, A. F. (2015), *Introduction to mediation, moderation and conditional process analysis: a regression based approach* New York: The Guilford Press.
- Keith, T.Z. (2019). *Multiple regression and beyond: An introduction to multiple regression and structural equation modelling* (3rd ed). London: Routledge.
- Urdan, T. C. (2011). *Statistics in plain English*. London: Routledge.
- R coding Cheat sheets

Readings are available to download from the course Github: https://github.com/thomcurran/PB4A7.

Assessment

The summative assessment comprises two parts:

1. **Secondary data analysis** on real datasets (worth **70%**)
2. **In-class exam** with multiple choice and short text responses (worth **30%**)

Secondary data analysis details

You will be given four data analysis tasks on secondary data:

- Task 1 – ANOVA
- Task 2 – Regression
- Task 3 – Path Analysis
- Task 4 – Structural Equation Modelling
- Task 5 – Multilevel Modelling

You need to write the Methods and Results sections for EACH OF Task 1, 2, and 3 AND ONE OF EITHER Task 4 or Task 5. The research questions are provided on the task sheets and will require you to employ analyses covered in PB4A7 (including assumption testing).

Deadline: TBC (around mid-January)

In-class exam details

The in-class exam will be a 90-minute multiple-choice and short-answer test. Students will complete the test under invigilated conditions in the PBS lab.

Date: TBC (in AT11)

Course Outline

Week 0: Preparation

Before starting the course, students are asked to work through at least the first two chapters of Emily Nordmann and Lisa DeBruine's Advanced Data Skills online textbook, but ideally the whole thing (which is sensational). At a minimum, before you begin, you should have R and R Studio installed on your laptop and be familiar with the R Studio panes.

📄 **Install R and R Studio** (www.posit.co)

📖 **Required Readings**

- Applied Data Skills: Processing & Presenting Data. Available here.

📖 **Further readings**

- Phillips (2017), Chapters 4-10.
- Navarro (2015), Chapters 1-4.

Week 1: Introduction to the linear model

This week, we will introduce the linear model as an overarching framework for understanding common statistical tools in the psychological sciences. You will learn about the most basic form of a statistical model, that is: DATA = MODEL + ERROR. The model part of this formula being our "best guess" given available information, and the error part being the imprecision or variance of that best guess from the observed data. The concepts introduced here are the fundamental building blocks for the more complex statistical models that we will construct in subsequent weeks, which allow us to add variables that might reduce (or explain) the variance around our "best guess".

📖 **Required readings**

- Urdan (2017), Chapters 4 and 5*
- Poldrack (2019), Chapter 5.

Week 2: Explaining variation I: Comparing groups

This week, we will move on from data cleaning and assumptions to data analysis itself. Here, we'll add categorical explanatory variables to our empty models that might help to reduce the error or imprecision of statistical estimation. Let's say we have an empty model that contains 100 student scores of narcissism. Without any other information, our best guess of what any given student might score on narcissism is the mean (i.e., the empty model). However, let's say we speculate that we might be able to improve our best guess, or reduce our estimated variance, if we knew the gender of the student because males typically score higher on narcissism than females. This is, in essence, what we are doing by adding a categorical explanatory variable to our empty model -- attempting to improve the accuracy of our best guess by adding variables that explain the error variation in the empty model.

📖 **Required readings**

- Navarro (2015), Chapter 13 and 14.
- Urdan (2017), Chapters 8 and 9 *

Week 3: Explaining variation II: Relationships

Having taken a look at explaining error variance in the empty model with a categorical explanatory variable to compare groups, this week it's the turn of continuous or quantitative explanatory variables to examine relationships. Although comparing groups and examining relationships might sound like different analyses, they are, in fact, just special cases of the linear model, with both seeking to reduce the error in our empty model's "best guess". The only difference is that we can't use the mean as a model for continuous variables because we don't have discrete groups (e.g., males or females). Instead, we use a line as a model, allowing us to make estimations using any data point along a whole spectrum of possible scores (e.g., height or weight). You may remember this line from school; it's called the line of best fit or the regression line.

📖 **Required readings**

- Hayes (2015), Chapters 2 and 3.*
- Urdan (2017), Chapters 12 and 13.*

- Navarro (2015), Chapter 15.

Week 4: Evaluating linear models: Hypothesis testing and causal inference

At this stage, students should now have sufficient knowledge to begin understanding the basic form of linear models and the purpose of explanatory variables in reducing the error variance around their "best guess". Rather than the amount of error within our models, this week we'll be asking a different question: how much error is there in our estimates? In other words, what is the spread or variance of our linear model estimates across many possible samples? We will introduce confidence intervals as a way of describing the error variance in the estimate and discuss the various methods for constructing them.

📖 **Required readings**

- Poldrack (2019), Chapters 7-10.
- Navarro (2015), Chapters 9-11.
- Phillips (2017), Chapter 13.
- Meuleman et al (2013). Chapter 5.

Week 5: Moderation and mediation

To date, we have examined relationships as if the predictors were independent of each other. But what if they interact? Perhaps the relationship between perfectionism and burnout is stronger in highly controlling contexts, such as work? This is classic moderation -- the effect of the predictor (perfectionism) on the outcome variable (burnout) is conditional on the level of a third variable, the moderator (control). The second part of this session will be dedicated to tests of causal processes. Rather than examining the interaction of third variables, we will focus on the relationships as they operate through third variables. Perfectionism is a vulnerability to anxiety disorders, for example. But this relationship is not direct -- perfectionists suffer anxiety because they have a ruminative response style. This is classic mediation -- the relationship between perfectionism and anxiety goes through rumination. In this session, we will extend our knowledge of multiple regression using the linear model to conduct tests of moderation and mediation.

📖 **Required readings**

- Hayes (2015), Chapters 4, 7, and 8.
- Hayes, A. F. (2009). Beyond Baron and Kenny: Statistical mediation analysis in the new millennium. *Communication monographs, 76*(4), 408-420.

Week 6: Reading Week

No lectures or seminars this week

Week 7: The general linear model: A recap and things I didn't tell you 🙊

The second half of this semester sees us taking a recap of the linear model and the class of analyses that we covered in the first half: t-test, ANOVA, and regression. We'll go back over these univariate tests and quickly recap their core principles and applications. Then, we're *going to cover some auxiliary issues that I didn't tell you about, but are nevertheless important to know. These are the four core assumptions of the linear model and how we diagnose for them in R.*

📖 **Required readings**

- Carey, G. (2013). The general linear model: A gentle introduction. In G. Carey (Ed.) *Quantitative Methods in Neuroscience*, Chapter 9.
- Nimon, K. F. (2012). Statistical assumptions of substantive analyses across the general linear model: a mini-review. *Frontiers in psychology*, *3*, 322.
- Meuleman, B., Loosveldt, G., & Emonds, V. (2013). Regression Analysis: Assumptions and Diagnostics. In H. Best and C. Wolf (Eds.) *The SAGE Handbook of Regression Analysis and Causal Inference*, pp. 83-110. London: Sage.

Week 8: Structural equation modelling I: Path analysis

This week we are going to move from univariate tests to multivariate tests and introduce something called structural equation modelling. The underlying principles are the same: the general linear model. However, we now move to applying these principles in a multivariate framework (i.e., multiple outcome variables). The most basic structural equation model is the path model, which tests for relationships between multiple criterion and multiple outcome variables in some causal chain. This is sometimes called path analysis. We will see how path models build on the limitations of multiple regression and look at some of their practical applications to real world psychological phenomena. We will then apply this knowledge to real word data in the workshop.

📖 **Required readings**

- Keith, T. Z. (2014). Chapters 11, 12, and 13.

Week 9: Structural equation modelling II: Factor analysis

Last week, we introduced structural equation modelling as a multivariate tool of analysis using the linear model. We examined the simplest structural equation model -- the path model. Here we introduced issues of parsimony and model fit, and we saw how mediation models might be tested simultaneously (i.e. not using separate regressions). This week, we build on this understanding and introduce factor analysis as a way of assessing the validity of measurement. Along the way, we will examine the differences between exploratory and confirmatory factor analysis. We will see how capturing variance with latent factors is preferable in psychology because latent factors are constructed in the absence of error. And we will also see how factor analysis is conducted in R.

📖 **Required readings**

- Keith, T. Z. (2014). Chapters 14 and 15.
- Field, A. (2012). Chapter 17.

📖 **Further readings**

- Flora, D. B., & Flake, J. K. (2017). The purpose and practice of exploratory and confirmatory factor analysis in psychological research: Decisions for scale development and validation. Canadian Journal of Behavioural Science, 49, 78.
- Anderson, J. C., & Gerbing, D. W. (1988). Structural equation modeling (2nd Ed) in practice: A review and recommended two-step approach. Psychological bulletin, 103, 411.

Week 10: Structural equation modelling III: Latent variable SEM

To this point, we have looked at path models and confirmatory factor analysis. One can test relationships and the other the adequacy of measurement. However, neither of these models can test both relationships and measurement at the same time. This week we'll look at combining path analysis and confirmatory factor analysis in something called latent variable structural equation modelling. The advantage of this approach is that it can test relationships and causal models in the absence of error. It is indeed an elegant approach and it is not one of the most common approaches to hypothesis testing in the psychological and behavioural sciences.

Before this lecture and workshop, please read the following JEPS blog post: https://blog.efpsa.org/2015/12/14/structural-equation-modeling-what-is-it-what-does-it-have-in-common-with-hippie-music-and-why-does-it-eat-cake-to-get-rid-of-measurement-error/

📖 **Required readings**

- Keith, T. Z. (2014). Chapter 16.
- Loehlin, J. C. & Beaujean, A. A. (2017). Chapter 1.

Week 11: Multilevel modelling: When the data are clustered

To this point, we have looked at various univariate and multivariate statistical models that are part of the general linear model. Each time we have run these analyses we have assumed that the data are sampled independently. This assumption, however, is almost always breached in real-world settings. When doing research on children's academic achievement, for example, kids are clustered within schools, which are clustered within geographical regions. In repeated measures designs, time points are nested within days, which are nested within people. At each of these levels, there is within-cluster variance and between-cluster variance that must be accounted for in analyses but are not in the techniques we have covered to date. Multilevel models (also known as linear mixed models, hierarchical linear models, or mixed-effect models) have become increasingly popular in psychology for analysing clustered data. Helpfully for us, they are also underpinned by the linear model and so you are already familiar with the basic concepts! We will build on this knowledge to demonstrate how multilevel modelling accounts for data clustering and how to apply them to cross-sectional and repeated measures data using R.

Before this lecture and workshop, please read the following BPS blog post: Cartwrigth, M., Traviss, G., & Blance, A. (2012). Murder, muddled thinking and multilevel modelling. Available here: https://thepsychologist.bps.org.uk/volume-25/edition-9/murder-muddled-thinking-and-multilevel-modelling

📖 **Required readings**

- Peugh, J. L. (2010). A practical guide to multilevel modeling. *Journal of school psychology, 48*, 85-112.
- Nezlek, J. B. (2008). An introduction to multilevel modeling for social and personality psychology. *Social and Personality Psychology Compass, 2*, 842-860.
- McNeish, D., & Matta, T. (2018). Differentiating between mixed-effects and latent-curve approaches to growth modeling. *Behavior Research Methods, 50*, 1398-1414.