POL6: Statistics and Methods in Politics and International Relations  
HSPS, 2018-19

**Course Organiser:** Dr Pieter van Houten ([pjv24@cam.ac.uk](mailto:pjv24@cam.ac.uk))

**Teaching Staff**

Dr Claudia Abreu Lopes ([cimdal2@cam.ac.uk](mailto:cimdal2@cam.ac.uk))
Dr Pieter van Houten ([pjv24@cam.ac.uk](mailto:pjv24@cam.ac.uk))

**Aims and Objectives**

After completing this paper, students will:

- Have learned about a range of quantitative methods used in empirical research in the study of politics and international relations (and the social sciences more broadly)
- Be able to read critically, and comment on, published research using these methods
- Know how to apply these methods correctly using appropriate software packages, and how to apply statistical tests to assess the validity of results
- Appreciate the limitations of the methods taught, and common mistakes which may be made in quantitative empirical research
- Have experience of writing up the results of empirical research
- Appreciate and think about how quantitative research can feature in particular research projects
- Have a better knowledge of how quantitative methods have been applied to various topics of interest to students of politics and international relations

**Introduction to the Paper**

Quantitative information (‘data’) is ubiquitous nowadays. Statistical methods are the means to describe, analyse and present data and the patterns revealed by them. Statistical analysis is widely used in the social sciences, including the study of politics and international relations. For example, and not surprisingly, this is the case for many studies of elections and voting behaviour, public opinion, and parliamentary decisions. However, it also increasingly features in studies of issues such as inter-state wars, civil conflict, corruption, state building and regime transformations – and virtually any other issue we focus on in the study of empirical politics. (This is not to say that quantitative methods are the only, or necessarily the best, way to study these issues. As you will learn in this paper, there are both strengths and weaknesses to these research methods.) Outside the academic world, data and statistical methods are widely used in public policy-making, the financial sector, management consultancy, market research, and a whole range of other sectors and professions.
This paper introduces students to statistical methods and their applications to political and international relation issues and questions. The acquired skills will be useful for critically engaging with research that uses these methods, for conducting such research (as practised in the projects for this paper, and possibly for third-year dissertations), and, as indicated above, for a variety of possible future career paths.

The paper consists of two parts. The first part teaches students a variety of quantitative methods that can be applied to the study of political and international relations issues. This taught part takes up for the first ten weeks of the paper (all of Michaelmas term, and the first two weeks of Lent term). The second part consists of a data analysis project, on a topic that students can choose from a list provided to them. In this project, students need to formulate a research question, analyse secondary data to answer this question (applying some of the methods taught in the first part of the paper), and write a 5,000-word report on the nature and results of their project.

**Paper Format**

The first part of the paper is taught by lectures, practical sessions and supervisions. The lectures for the first part of the paper will introduce the various statistical techniques and methods, while the corresponding practical sessions will allow the students to apply these. The practical sessions are crucial to gain a better understanding of the techniques and methods (and, thus, to be able to perform well in the paper’s exam), and to obtain the skills required for the data analysis project in the second part of the paper (and, thus, to be able to do well in the other half of the assessment for the paper). For the supervisions for this part of the paper, students will be given specific tasks that will require them to go into further detail on some of the issues covered in the lectures and practical sessions. Detailed information on the supervision assignments will be provided as the course progresses. Students should expect to do two or three supervisions for this part in Michaelmas, and one in Lent.

The second part of the paper consists of an introductory lecture on how to approach a data analysis project, and three supervisions on a student’s individual project.

In Easter term, there will be a general revision session and a revision supervision in preparation for the exam.

**Assessment**

Assessment for the paper consists of two elements. First, a two-hour exam at the end of the year that tests the students’ knowledge of the material taught in the first part of the paper. Candidates will have to answer one question (which will be divided into several sub-questions). A mock exam is included at the end of this paper guide, together with the examiners’ reports from the last three years. Second, a 5,000-word report on the data analysis project, which is due early in Easter term. Each assessment element makes up 50% of the overall mark for the paper.
Course Materials

This paper differs from other papers in the HSPS Tripos in that it does not have a very extensive reading list. In fact, the core materials for this course are the lecture slides and the various other materials related to the lectures and practical sessions in the first part of the papers. These materials will be made available through the paper’s Moodle site. Students should make sure to study these materials carefully when preparing for the exam.

Although the core materials are the lecture slides and practical session notes are the core materials for this paper, it will also be useful at times to consult a statistics textbook. There are many such textbooks. One very useful one for this paper is: Andy Field, Jeremy Miles and Zoë Field, Discovering Statistics Using R (London: Sage, 2012). Although this book is not cheap, students may want to consider buying this book. Another useful textbook is: Roger Tarling, Statistical Modelling for Social Researchers: Principles and Practice (London: Routledge, 2009). The most relevant chapters from these textbooks for this paper are indicated in the table with the specific lecture topics below.

Students may also find it useful – especially early in the course, and particularly if they have not studied statistics in any form before – to read general and non-technical accounts of statistical concepts, ideas and reasoning. Two suggestions of good books of this nature are: Charles Wheelan, Naked Statistics: Stripping the Dread from the Data (New York: Norton, 2013) and Derek Rowntree, Statistics without Tears: An Introduction for Non-Mathematicians (London: Penguin, 2018).

A list of possible further readings is provided below after the description of Part I of the paper.

Structure of the Paper

PART 1: Quantitative Methods for Politics and International Relations

This part introduces students to various statistical methods: descriptive statistics, bivariate correlation, multivariate linear regression, logistic regression, and multilevel regression. Students will be provided with an introduction to these methods in the lectures (lectures 2 to 8), illustrated with examples from the study of politics and international relations, and will start to apply them in the practical sessions (using the statistical software R).

The first lecture introduces the paper and describes the role of quantitative methods in the study of politics and international relations (providing a summary of the origins and evolution of the use of quantitative methods in this field, gives some examples of research using these methods, indicates the roles that these methods play in the research process). The last two lectures focus on so-called ‘big data’ research, which is currently receiving much attention, and – although perhaps over-hyped – has the potential to increase the scope and opportunities for quantitative research in the social sciences, but also may force us to rethink the statistical techniques and research designs that we use.
The module consists of ten weekly 2-hour lectures, ten weekly 2-hour practical sessions (the first couple of which will focus on introducing the statistical software used in the course), and three or four 1-hour supervisions.

The lecture schedule is (with a few suggested readings) is as follows:

<table>
<thead>
<tr>
<th>Topic</th>
<th>Related readings</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 Introduction: the use of statistical analyses in Politics and IR</td>
<td></td>
</tr>
<tr>
<td>2 Descriptive statistics: distributions, mean, median, etc.</td>
<td>Field et al, ch 2, Tarling, ch 3</td>
</tr>
<tr>
<td>3 Statistical significance, confidence intervals, correlation, chi-square</td>
<td>Field et al, chs 6 and 18.1-18.6</td>
</tr>
<tr>
<td>4 Simple linear regression</td>
<td>Field et al, ch 7.1-7.5</td>
</tr>
<tr>
<td>5 Multivariate linear regression</td>
<td>Field et al, ch 7, Tarling, ch 4</td>
</tr>
<tr>
<td>6 Logistic regression (I)</td>
<td>Field et al, ch 8, Tarling, chs 5 and 6</td>
</tr>
<tr>
<td>7 Logistic regression (II)</td>
<td></td>
</tr>
<tr>
<td>8 Multi-level regression</td>
<td>Field et al, ch 19, Tarling, ch 9</td>
</tr>
<tr>
<td>9 Recap of regression techniques; Big data (I)</td>
<td>Salganick, ch 2, Boyd and Crawford</td>
</tr>
<tr>
<td>10 Big data (II)</td>
<td>Salganick, ch 3, Callegaro and Yongwei</td>
</tr>
</tbody>
</table>

[Note that this lecture schedule is provisional – we may adjust the topics covered in each session depending on teaching and student needs.]

The schedule for the practical sessions is:

| 1 | Introduction to practical sessions; installing R |
| 2 | Introduction to R (I) |
| 3 | Introduction to R (II) |
| 4 | Descriptive statistics: distributions, frequencies |
| 5 | Correlations, chi-square |
| 6 | Simple linear regression |
| 7 | Multivariate linear regression |
| 8 | Logistic regression |
| 9 | Multi-level regression |
| 10 | Big data |

[Note that this practical session schedule is provisional – we may adjust the topics covered in each session depending on teaching and student needs.]
Reading list

As indicated above, the core readings and materials for this module (and the paper as a whole) are the lectures notes and additional materials provided by the lecturers. However, it will also be useful to consider some further readings on the issues and methods discussed. This list provides some suggestions in addition to the texts mentioned earlier (*: especially recommended):


Some applications of the discussed methods and models (some of these will be used in the lectures, practical sessions and supervisions):


Articles with further applications of the discussed models may be added in due course (to serve as possible further reading in preparation of the exam).

**PART 2: Data analysis project**

In this part of the paper, students will undertake a data analysis project and produce a 5,000-word report on the project. This report will be formally assessed, and contributes 50% to a student’s overall mark for the paper. Students will choose a particular topic (and related dataset) from a provided list, and conduct a data analysis project on this topic. This involves formulating a research question and possible answers to the question, conducting a statistical analysis, and writing a report on the nature and results of the project.

Students will need to indicate their choice of topic by the end of Michaelmas term, and the report is to be submitted by noon on Monday 29th April 2019.

An introductory lecture (on Wednesday 6th February) will present the expectations for the project, discuss general issues of research design and how statistical analyses can contribute to this, and will provide an opportunity for students to ask questions about the projects. The main form of guidance on the individual projects will be provided by three supervisions (at least two of which will be individual ones).

A more detailed guide on these projects (including fuller descriptions of the topics/datasets, and a discussion of the expectations for the project and the report) will be provided in Michaelmas term. Here is a short description of the available topics:

1. **Voting behaviour in elections**
Elections are at the heart of the democratic process, and are consequently studied extensively. One of the key questions relates to the motivations of voters to vote for a particular party or candidate. Quantitative analyses, usually based on surveys of voters, play a crucial role in trying to answer this question. For this topic, a dataset of a British election will be made available to students to conduct a project on a specific question related to voting behaviour in the UK.

2. **Perceptions of corruption**

Corruption is seen as a major issue and problem in many parts of the world, but seems to vary considerably across countries. Although ‘corruption’ is difficult to measure, quantitative analyses into the causes and consequences of corruption have become more prominent. Most of these analyses are based on surveys of perceptions of corruption. For this project, students will be given data from surveys done by Transparency International to explore causes or consequences of corruption.

3. **Attitudes towards globalisation**

As processes of globalisation have become more pervasive and intrusive, public attitudes towards this phenomenon have become politically more important (see, for example, the apparent backlash against globalisation reflected in support for populist parties and politicians). The YouGov Centre for Public Opinion and Policy Analysis has recently done an extensive cross-country survey of attitudes towards globalisation. For this project, students will use data from this survey to explore factors influencing these attitudes.

4. **Patterns in wars**

Wars between states are a core theme in the study of international relations, while intra-state wars appear to have become more important in recent years and have consequently received increasing academic attention. Quantitative studies of these phenomena have become more prominent, pioneered by the ‘Correlates of War’ project, which has collected data on wars that have occurred since 1816. For this project, students will focus on some of these data to explore patterns in wars.

5. **Efforts to reach development goals Sustainable Development Goals (SDGs)**

States and international organisations concerned promoting development around the world have periodically formulated a set of goals that less developed countries and aid donors should aim for. These goals have most recently been phrased in terms of Sustainable Development Goals (SDGs; these are, for example, related to poverty, inequality, gender). The World Bank publishes data indicating the extent to which countries are achieving SDGs, as well as many other data on these countries. In this project, students will use these data to investigate patterns and reasons for the variation in the achievement of SDGs.

**Background reading on research design:**
Mock Exam

Candidates need to answer all questions.

Table 1 depicts the results of multilevel linear regressions on data from round 6 of the European Social Survey (2014). The population is all persons aged 15 and over resident within private households in 29 European countries, regardless of their nationality, citizenship or language. The response variable is an index of soft political protest behaviours (on a scale from 0 to 7).

The explanatory variables are:

- Trust in country’s parliament (ranging from ‘no trust’ – coded 0 – to ‘completely trust’ – coded 10);
- Years of full-time education completed (in years);
- Gender of respondent (0=man; 1=woman);
- Age of respondent (in years);
- Economist Intelligence Unit Democracy Index (ranging from ‘the least democratic’ – coded 0 – to ‘the most democratic’ – coded 10)
- Gross Domestic Product per capita (in purchasing power standards)

<table>
<thead>
<tr>
<th>Model 1</th>
<th>Model 2</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Coef</strong></td>
<td><strong>Std error</strong></td>
</tr>
<tr>
<td>Intercept</td>
<td>-0.506</td>
</tr>
<tr>
<td>Trust in parliament</td>
<td>0.063**</td>
</tr>
<tr>
<td>Years of full-time education completed</td>
<td>0.073**</td>
</tr>
<tr>
<td>Age in years</td>
<td>-0.002**</td>
</tr>
<tr>
<td>Gender</td>
<td>-0.100**</td>
</tr>
<tr>
<td>Democracy Index</td>
<td></td>
</tr>
<tr>
<td>Gross Domestic Product</td>
<td></td>
</tr>
<tr>
<td>Intra-class correlation coefficient</td>
<td>0.133</td>
</tr>
</tbody>
</table>

*p<.01; **p<.001

1. Is it justified to use linear regression models in this analysis?

2. Write the regression equation for Model 1 for a male, 30 years old, 12 years of education and with trust in parliament=5.
3. Which factor do you think has the largest effect on the soft political protest behaviours of individuals in Model 1? Justify your answer by giving some examples of the substantive effects of the coefficients in Table 1.

4. What is the rationale for using a multilevel regression this specific example?

5. Interpret the regression coefficients and associated p-values for the individual and country level variables for Model 2.

6. Write a hypothesis that considers a cross-level interaction and justify your choice of explanatory variables.

7. Considering the response variable, can you think of alternative ways of measuring 'soft protest behaviour'?

Examiners Reports

[Note that the exam papers in previous years differed from this year’s. They contained 3 main questions; only the first – and largest – of these is relevant for this year’s exam.]

2015-16

This paper was taught for the first time this year, and 21 candidates took the exam. The exam consisted of a mandatory Question 1 (which had several parts), and two optional questions of which candidates had to answer one. Of these optional questions, Question 2 (on how to design a quantitative project on voting behaviour in the 2014 Scottish referendum) was clearly the more popular with 19 answers. The performance of the candidates was mostly good: 6 candidates received a First class mark, 11 candidates a 2.1 mark (of which two had a mark of 69), and 2 candidates received 2.2 marks. Unfortunately, there were also one Third class mark and one Fail mark.

The average quality of the scripts was very good, and the examiners were impressed with the quality of quite a few of the answers. In particular, many of the answers to Question 2 were very strong. Most of the candidates showed a good understanding of the main elements of research design, and the best answers formulated precise research questions, indicated possible hypotheses, discussed the data required to test these hypotheses, and provided some reflections on the methodological challenges that would be involved in such a project. The few weak answers to this question covered only some of these elements, or included incoherent or inconsistent statements in parts of the answer. More generally, most answers could have discussed possible methodological challenges more (potential issues to discuss here would be the quality of the data, the difficulty to operationalize certain concepts, and the representativeness of samples). The two answers to Question 3 were not very strong, and failed to discuss in any detail the advantages and drawbacks of quantitative data to explore ‘reality’ and integrate a useful example into this discussion.
The quality of the answers to the various parts of Question 1 was somewhat mixed. The best answers showed a good understanding of statistical concepts and the models that were used in the question, and provided good substantive interpretations of the presented statistical results. Quite a few answers showed solid understanding of the issues involved, but did not sufficiently explain the answers or discuss the substantive interpretations to get a higher mark. For example, several candidates gave very concise answers to Question 1.5, focusing on only one possible rationale for using the indicated model, where there were, in fact, several reinforcing rationales (and answers were rewarded for focusing on several or all of them). Unfortunately, there were also several weak answers which showed limited understanding of the statistical issues and concepts, or made claims that were clearly wrong.

It was disappointing to see that some of the answers to Questions 1.1 and 1.5 were missing or very limited, as the basic information to address these questions was provided on some of the lecture slides. Overall, the examiners had the impression that some candidates would have benefited from attending more lectures and practical sessions in the course of the year. However, it should be emphasized again that the performance of many of the candidates was very good.

2016-17

This year, 33 candidates took the exam: 23 second-year students and 10 third-year students. As in the previous year, the exam consisted of a mandatory Question 1 (which had several parts), and two optional questions of which candidates had to answer one. Of these optional questions, Question 2 (on how to design a quantitative project on the effects of globalisation on the economic situation of countries and/or individual well-being) was clearly the more popular with 29 answers. The performance of the candidates was as follows: 7 candidates received a First class mark, 16 candidates a 2.1 mark (of which three had a mark of 69), and 8 candidates received 2.2 marks. Unfortunately, there were also one Third class mark and one Fail mark.

Many of the scripts were good, and there were some truly impressive answers to each of the questions. On average, the answers to Question 2 were not quite as strong as last year. Most of these answers had a good discussion of some elements of a quantitative research project on one of the themes indicated in the question, but only relatively few were able to do all the things required for an excellent answer: formulate a precise research question, indicate possible hypotheses, discuss the indicators and data required to test these hypotheses, and provide some reflections on the methodological challenges that would be involved in such a project. In particular, several answers did not provide any discussion of what measures might be used to evaluate the effect of globalisation (the main explanatory variable). Also, and as last year, more discussion of possible methodological challenges (potential issues to discuss here would be the quality of the data, the difficulty to operationalise certain concepts, and the representativeness of samples of individuals) would have improved most answers. The few answers to Question 3 were of mixed quality. The best essay provided an insightful critique and defence of the use of statistics, linking it to a discussion of the nature of the social sciences. The weaker answers made various
strong but unsubstantiated claims about the use of statistics, or failed to use examples in
the answer.

The quality of the answers to the various parts of Question 1 varied. The best answers
showed an excellent understanding of statistical concepts and the models that were used in
the question, and provided good and elaborate substantive interpretations of the presented
statistical results. A very small number of scripts suggested that the candidates
misunderstood the nature of the response variable (although this was explained several
times in the question, including in plain language). Particularly noteworthy was the variation
in the length of answers between scripts. While some questions (for example, Questions 1.1
and 1.3) could be answered satisfactorily that way, the other questions required longer
answers to be fully satisfactory and receive high marks. For example, as explained
throughout the paper, ‘interpreting regression coefficients’ (Question 1.4) involves not only
indicating whether variables are statistically significant, but also discussing what the
substantive effect is, what conclusions can be derived from this (in relation to earlier
formulated hypotheses), and what the reasons for this result might be. Similarly,
‘formulating hypotheses’ (Question 1.2) involves both indicating hypotheses and discussing
possible rationales for them.

Finally, and similarly to last year, the examiners had the impression that some candidates
would have benefited from attending more lectures and practical sessions in the course of
the year. Overall, however, the examiners were impressed with how much knowledge and
understanding of quantitative methods the candidates had been able to develop in this
paper.

2017-18

This year, 22 candidates took the exam: 13 second-year students and 9 third-year students.
As in the previous year, the exam consisted of a mandatory Question 1 consisting of several
parts; and two optional questions of which candidates had to answer one. Of these optional
questions, Question 2 on how to design a quantitative research project was much more
popular with 19 answers. Question 3 pertained to the complications of employing “big
data” in research.

4 candidates received a First class mark, 14 candidates a 2.1 mark (of which two had a mark
of 69), and 2 candidates received 2.2 marks. Unfortunately, there were also one Third class
mark and one Fail mark. Students who received First marks often received close to perfect
scores on Question 1. Though this is unusual in POLIS examinations, the nature of POL6 is
exceptional in that—specifically for Question 1—answers either are or are not objectively
correct. The best answers to Question 1 showed an excellent understanding of statistical
concepts and the models that were used in the question, and provided good and elaborate
substantive interpretations of the presented statistical results. Weaker answers would
provide incomplete interpretations of the quantitative results presented; for instance,
discussing a coefficient’s statistical significance but not its effect size on the outcome
variable. Students also sometimes did not provide a rationale for hypotheses about
relationships between different variables.
Good answers for Question 2 were able to formulate a precise research question, derive testable hypotheses, discuss the data required to test these hypotheses, and provide an overview of the methodological challenges that would be involved in their proposed project. Weaker answers often failed to define key terms or esoteric statistical language, and/or did not provide a clear rationale for the proposed estimation technique they put forward. Students sometimes forgot to discuss methodological challenges that could arise, or gave this part of their answer only superficial attention. Though only a handful of students opted to answer Question 3, responses to this latter question were of about the same quality as a typical answer to Question 2. To be of top quality, essays on Question 3 had to address the general philosophy behind applying statistical analysis to social questions, as well as developments in computing power and data management techniques that made the use of big data feasible.