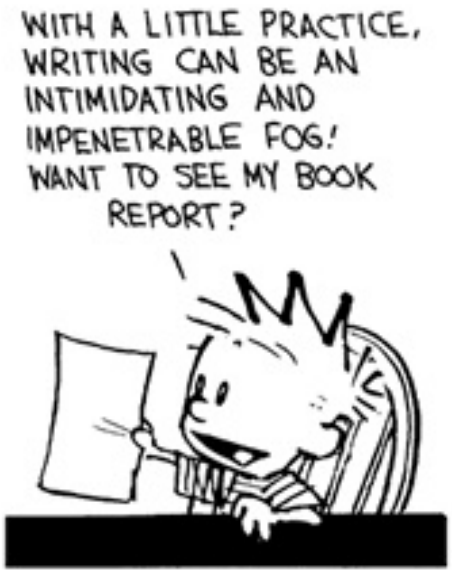
Wrapping up towards the final paper



Statistical Reasoning
and Quantitative Methods

François Briatte & Ivaylo Petev

Session 12

Outline

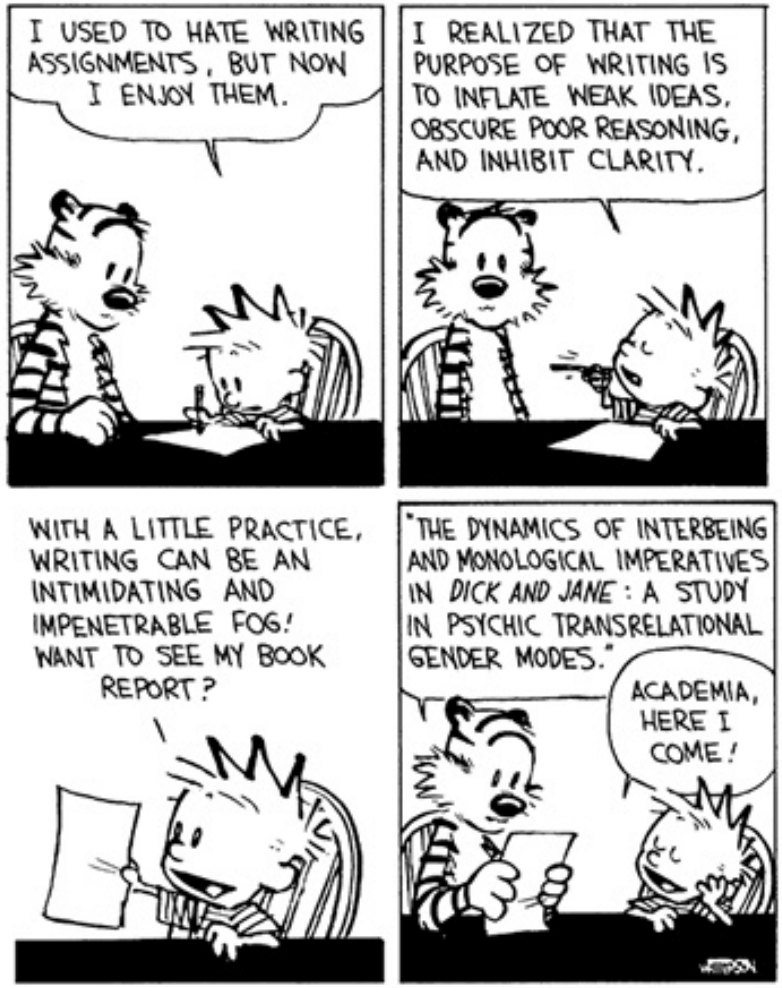
Structure

Inclusions

Improvements

Conclusions

The end



Structure

Use the IMRAD standard formatting:

1. **Introduction** to your research design (topic, theory, hypotheses).
2. **Methods** made of dataset and variable descriptions.
3. **Results** of bivariate tests and regression analysis.
4. **Discussion**, where you provide an appraisal of your original theory.

The final paper includes, approximately:

- **10–12 pages** all inclusive, with footnote references.
- **1–3 figures** of visually informative plots.
- **3–4 tables** to describe variables and cover results.

Precise instructions are all over the Stata Guide.
Examples are everywhere in the course do-files.

Introduction (max. 1 page)

- **Research question**:

  "The aim of this study is to establish whether, in the United States, **obesity** *[dependent variable] is determined by the* **level of education** *[main independent variable], and how this effect varies across* **race groups** *[comparison of groups]."*

- **Context**:

  "Obesity has **increased dramatically** *since the 1980s [magnitude] and contributes to* **social distinction** *[theory]."*

- **Hypotheses**:

  H_1*: "All other things kept equal, higher levels of education are associated with lower levels of obesity.*

  H_2*: "The effect of education on obesity is highest for African Americans and weakest for Asian Americans."*

Methods (max. 2 pages)

- **Dataset**:

  "The National Health Interview Survey [cite source] contains a measure of Body Mass Index for the American population [cite the unit of analysis, sampling method and total number of observations for the dependent variable, as well as transformations or recoding]."

- **Variables**:

  "The dependent variable is a continuous variable, which is constructed from [cite the measurement method from the codebook]. Its distribution is approximately normal over the sample [describe normality]. The variable is summarised, along with independent variables, in Table 1 [include summary statistics and brief descriptions]."

Results (max. 5 pages)

- **Association**:

  "We find a statistically strong association between obesity and education [describe results through text with probability levels in brackets; reproduce the correlation matrix]. The effect holds across ethnic groups [show independent variables as controls in crosstabulations or graphs only when they are pertinent to your general argument]."

- **Regression**:

  "The model for each ethnic group is reproduced in Table 3 [include the regression output as a single table, with one column for each model]. It establishes that education is a strong predictor of obesity, along with covariates such as age, and after controlling for income [interpret R-squared and all coefficients: statistical significance, direction, magnitude]."

Discussion (max. 2 pages)

The results will indicate strong, moderate or weak effects that either confirm or reject your background assumptions.

- Interpret **confirmatory results**:

  "Our analysis provides clear evidence in support of the argument [expressed in one of your hypotheses] that... This corroborates the view that... [support, reject or amend the theoretical priors on which your research design relies]."

- Also cover **negative results**:

  "Our analysis provides no clear evidence in support of the argument [expressed in another hypothesis] that... This result challenges our intuition that..."

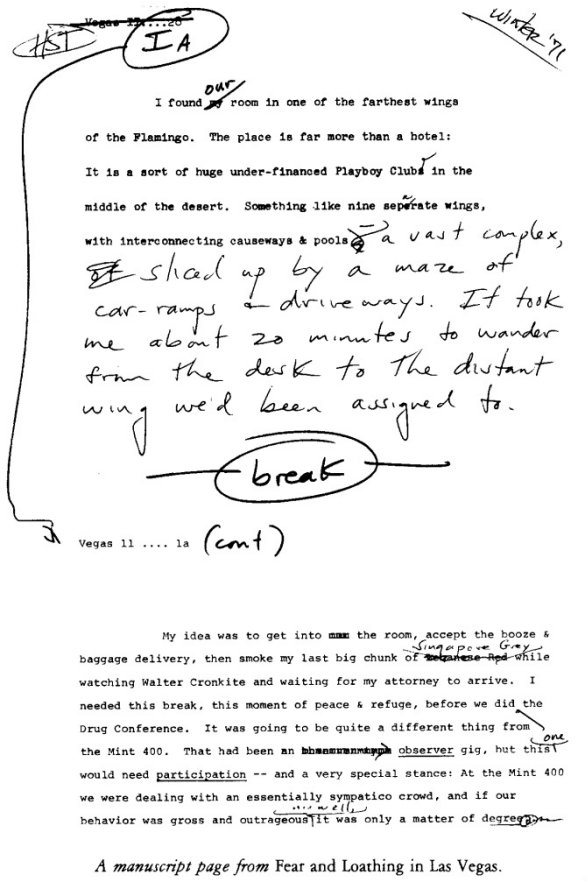
  Statistical insignificance can be attributed to sample size, measurement issues, or to a theory that might need revision.

Inclusions

Some stuff necessarily goes into your paper to make your point and support the burden of proof.

But it is just as important to realize that some stuff does not go into your paper.

The manuscript on the right is from Hunter S. Thompson's *Fear and Loathing in Las Vegas* (1971).



I A

Winter '71

I found our room in one of the farthest wings of the Flamingo. The place is far more than a hotel: It is a sort of huge under-financed Playboy Club in the middle of the desert. Something like nine separate wings, with interconnecting causeways & pools — a vast complex, sliced up by a maze of car-ramps & driveways. It took me about 20 minutes to wander from the desk to the distant wing we'd been assigned to.

break

Vegas 11 .... 1a (cont)

My idea was to get into the room, accept the booze & baggage delivery, then smoke my last big chunk of Singapore Grey while watching Walter Cronkite and waiting for my attorney to arrive. I needed this break, this moment of peace & refuge, before we did the Drug Conference. It was going to be quite a different thing from the Mint 400. That had been an observer gig, but this would need participation -- and a very special stance: At the Mint 400 we were dealing with an essentially sympatico crowd, and if our behavior was gross and outrageous ... well, it was only a matter of degree.

A manuscript page from Fear and Loathing in Las Vegas.

What to include

- **Sources** for all cited items:

  "1. Author(s), Dataset Name, Year of Release <URL>."

- **Captions** for (numbered) tables and figures:

  "Fig. 2. Dot graph showing the average level of subjective happiness by geographical region."

- **Probability levels**:

  "Given the strong correlation between happiness and wealth ($r = 0.79, p < 0.01$) we were not surprised to observe our dummy for Western states return a significant increase in happiness of 3.4 percentage points at $p < 0.05$ (Table 4)."

- **Interpretation.**

Systematic interpretation

If you forget to interpret your output, you will be thrown into the gaping mouth of the sarlaac that inhabits the Great Pit of Carkoon on planet Tatooine ($p < 0.01$).

And kittens will get hurt.

Systematic references

If you forget to reference your sources, the hideous terror of Cthulhu will arise from the sunken city of R'lyeh to spread the abject curse of the Great Old Ones onto this world ($p < 0.01$).

And kittens will get hurt. Again. Different ones.

What **not** to include

- **Insignificant output**, even if you will comment on the negative findings in your Discussion.

  "Look at that cute result with a p-value of 0.697!"

- **Unexplained output**, in the form of (usually several) pasted extracts of Stata results with no solid interpretation.

  "Have a look at Tables 4–12 and have fun reading them!"

- **Virtually every possible figure** created by your do-file, regardless of the actual information it might convey to the reader.

  "Notice how good I am at producing scatterplots for all variables (see Fig. 7–77 and the 16-page Appendix H)."

- **Causality.**

What to do with causality

The basic problem concerns your language: can you establish that y is "caused" by $a + b_1 x_1 + b_2 x_2 + \epsilon$, given your regression results?

A short answer is that:

- **Statistical equations provide the conditional distribution** of y given x_1 and x_2, and its probability level. No more, no less.
- **Causal and statistical information are separate species**. Reuniting them implies crossing a wide, schismatic rift.
- **Statistical inference basically requires a background theory** to meet the requirements of causal analysis in observational studies.

Further reading:

- **Judea Pearl**, "Statistics and Causality" (2011).

The final package

- **paper**`.pdf` using family names, e.g. `Briatte_Petev.pdf`
- **do-file**`.do` using family names, e.g. `Briatte_Petev.do`
- **dataset**`.dta` using acronym and year, e.g. `wdi2010.dta`

  (only if you are using a dataset from outside the course)

The substantive qualities of your work are:

- **Accuracy** in statistical techniques.
  Involves the selection of commands and precision of their settings.
- **Appropriateness** in scientific reasoning.
  Involves the depth and terminology of your argument.
- **Readability** in all places.
  Involves the style of your writing and reproducibility of your do-file.

Improving text

> *"Anyone who cannot speak simply and clearly should say nothing and continue to work until he can do so."*
> *(Karl Popper, cited by Victoria Stodden)*

Concision applies to *all* scientific writing, regardless of its methods. Quantitative methods require as much formulation work as any other.

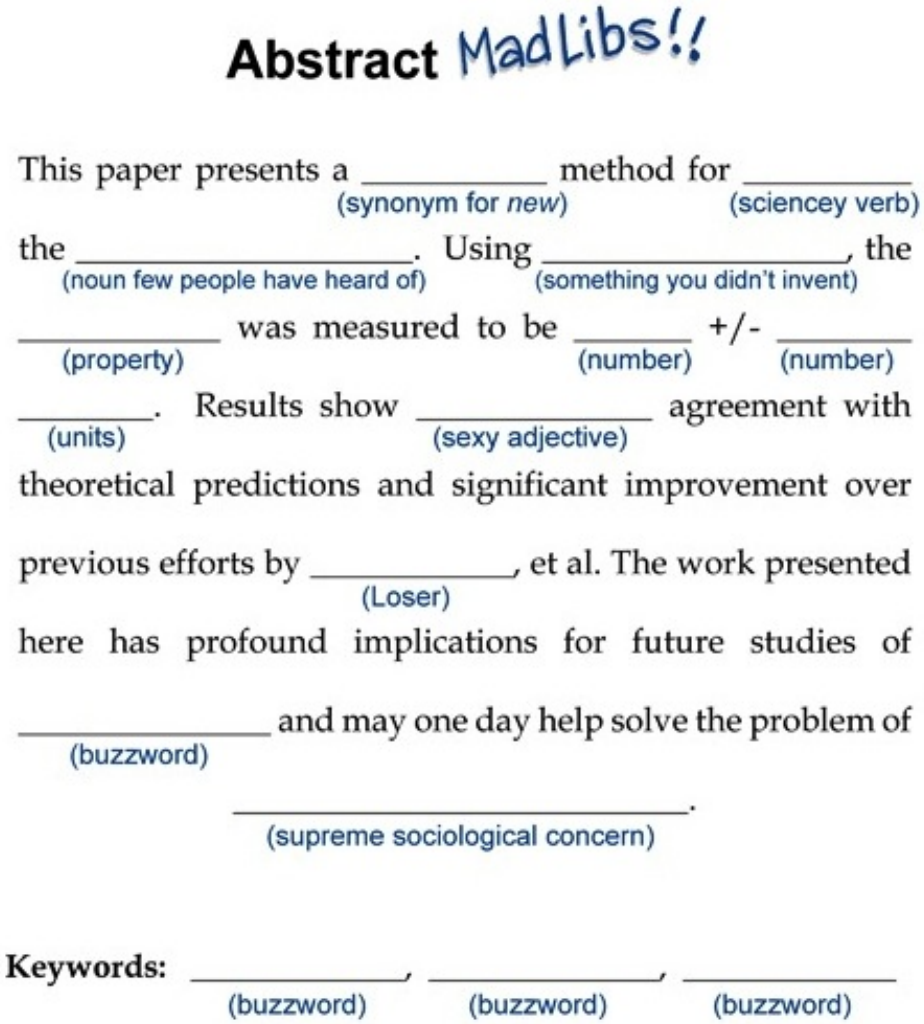
> *"Do not worry. You have always written before and you will write now. All you have to do is write one true sentence. Write the truest sentence that you know."*
> *(Ernest Hemingway, cited by Thomas Basbøll)*

Sentences are the fundamental component of any text; paragraphs and sections only come on top as structural markup.

Scientific writing uses standard terms

What really counts is the 'flesh' that you add between the 'bones' (not minding the gruesome analogy).

There will be grades for structure and statistics, but overall, substantive content prevails in the overall assessment.



Abstract MadLibs!!

This paper presents a _____ (synonym for *new*) method for _____ (sciencey verb) the _____ (noun few people have heard of). Using _____ (something you didn't invent), the _____ (property) was measured to be _____ (number) +/- _____ (number) _____ (units). Results show _____ (sexy adjective) agreement with theoretical predictions and significant improvement over previous efforts by _____ (Loser), et al. The work presented here has profound implications for future studies of _____ (buzzword) and may one day help solve the problem of _____ (supreme sociological concern).

Keywords: _____ (buzzword), _____ (buzzword), _____ (buzzword)

Improving code

Literate progamming:

> *"Let us change our traditional attitude to the construction of programs: Instead of imagining that our main task is to instruct a computer what to do, let us concentrate rather on explaining to human beings what we want a computer to do."*
> *(Donald Knuth)*

- **Variables** should have short and readable names.
- **Graphs** should include the `name()` option to stay in memory.
- **Comments** and sections should help understand the code.

Use the course do-files to find some inspiration and make up your own style of literate programming.

Improving graphs

- **Inter-ocular trauma test**: if the graph does not hit you between the eyes, disregard it.
  - `graph box` spots outliers, while `histogram` qualifies a distribution.
  - `graph dot` applies to a categorical and a continuous variable.
  - `scatter` applies only when the variables are sufficiently continuous.

  Save graphs in PNG or PDF format for inclusion into your text.
- **Graph options** will make your graphs much more readable and useful: read their documentation, do-files, and ask for help.
  - `ylabel(1(10)100)` creates a vertical 100-point labelled scale.
  - `yscale(reverse)` reverses the y-axis for a reverse-coded variable.
  - `ytitle("GDP growth (%)")` provides a concise title to the y-axis.
  - `mlabel(cname)` adds a label to data points on a scatterplot.

  All graph options are fully documented in Stata. Some options will apply only to the y-axis or x-axis depending on the plot.

Improving tables

Your do-file produces tons of tables; your final paper shows only two:

- **Descriptive statistics**:
  - `sum` or `tabstat` describe continuous variables.
  - `tab` or `fre` describe categorical variables.
  - `tabout` export descriptions to text.
- **Regression results**:
  - `reg` and its options produces all regression analysis.
  - `mkcorr` exports correlation matrixes to text.
  - `estout` exports regression results to text.

Use a spreadsheet editor to import and format your tables, following the options documented in the Stata Guide.

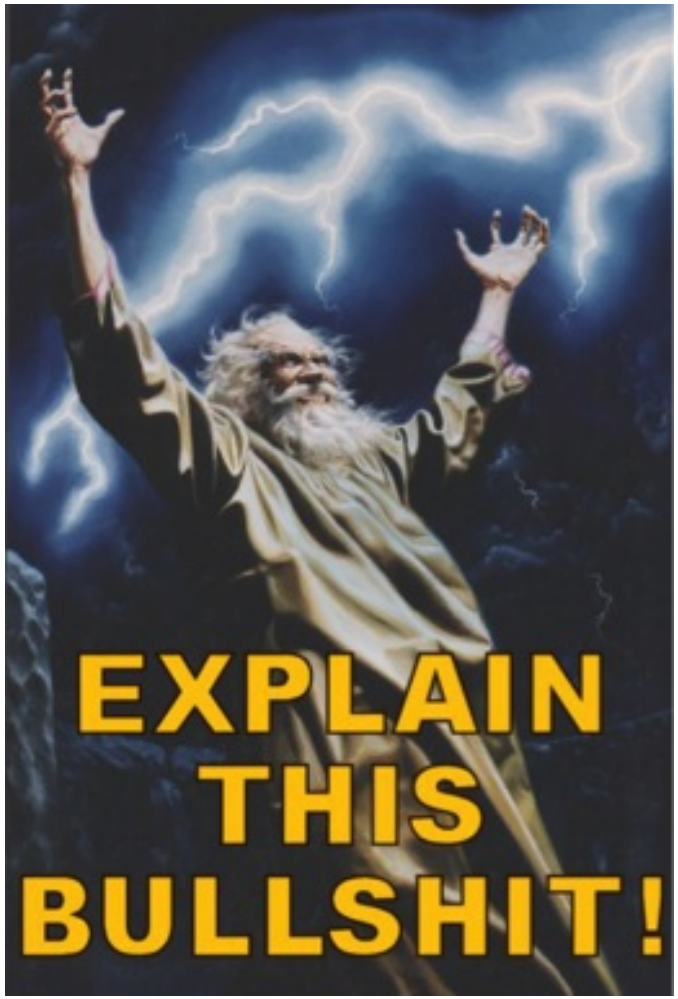
Note: presentational aspects are part and parcel of quantitative methods, just like grammar, syntax and punctuation are to writing.

Systematic proofreading

If you forget to proofread your work, a gigantic hole might open in the earth under your feet, and you might burn in the flames of the monstrous Moloc'h for eternity ($p < 0.01$).

And your graders will get angry at their laptops.

All remaining kittens will be decimated without any sign of human mercifulness.



Further help

- Course-specific help:
  - Stata Guide
  - Session do-files
  - Course slides
- General help:
  - Handbook chapters
  - Stata documentation (`help` *`command`*)
  - Online tutorials

Handbook chapters and course emails are available from the ENTG. Everything else is systematically archived on the course website:

`http://f.briatte.org/teaching/quanti/`

Happy coding!

Concluding thoughts on your data

Your Discussion ends the paper by answering some fundamental questions on your research.

Start with an appraisal of your data. In the end:

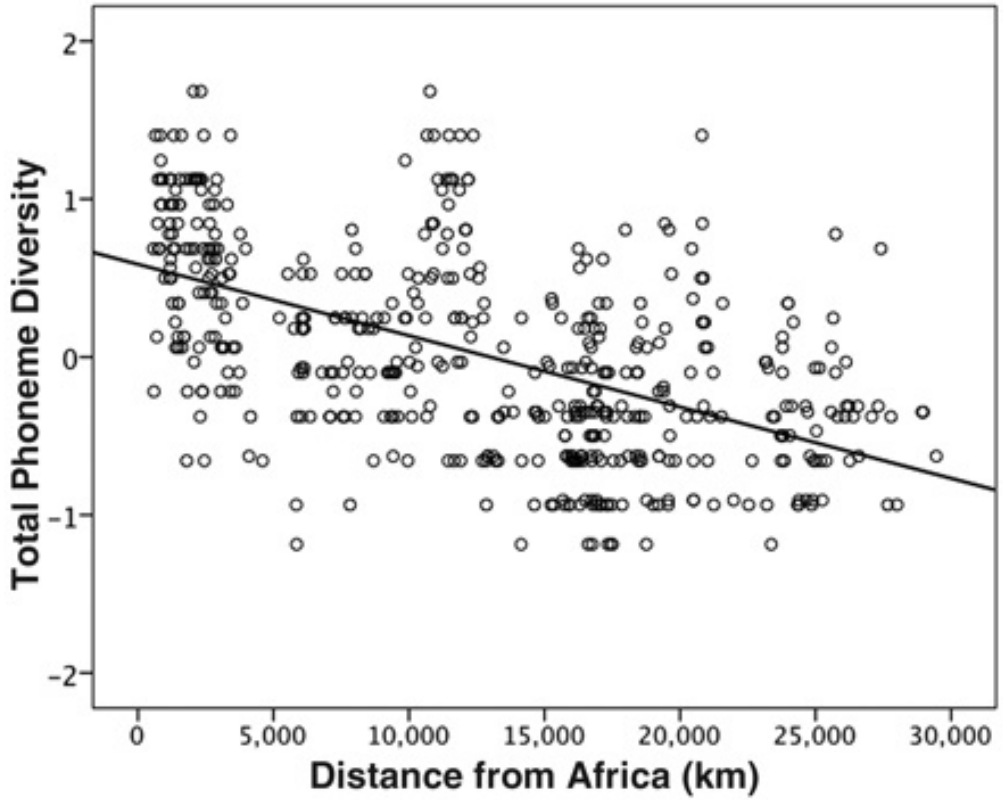
- **How precise** is your measurement of the issue at stake? Are your variables reliable proxies for the phenomena you are interested in? What limitations did you run in?
- **How representative** is your final sample, on which you ran your model? Would you be able to generalize your results, to what population and with what confidence?

Data limitations are always expectable in the social sciences. The current data revolution is pushing for open data of higher quality and clarity, but there are serious obstacles and pitfalls.

Organized skepticism by example

"Human genetic and phenotypic diversity declines with distance from Africa"
(Quentin Atkinson)

Published in Science, 15 April 2011; see also The Economist, 16 April 2011.



Organized skepticism by example

"I have some concerns about what lies in between the assumptions and the results, especially concerning the way 'Total Phoneme Diversity' is estimated. This measure gives (what seems to me to be) excessive weight to certain features, ignores syllable structure, and (as a result) is heavily influenced by a few areal characteristics [by which he means continent-level distributions] that are as likely to be innovations as survivals."
(Mark Liberman)

- **Blog entry**: "Phonemic diversity decays 'out of Africa'?"
- **URL**: `http://languagelog.ldc.upenn.edu/nll/?p=3090`

Concluding thoughts on your model

Continue with an appraisal of your model. In the end:

- **How predictive** [R-squared] is your model? Did you manage to formulate a reasonable interpretation of the relationships that emerged between your variables?
- **How informative** was your overall research? Did the interpretation of your model [coefficients] provide an interesting way to think about your general topic?

By definition, no model perfectly embraces reality. Linear models are extremely sensitive to how you set them up (model specification), and violating their background assumptions is usually devastating.

Bayesian statistics are gradually getting us out of this mess.

Concluding thoughts on your methods

Finish with a mental assessment of your methods:

- **Our course is introductory**. Much more advanced techniques exist to refine our models.
- **Our technique is limited**. Frequentist methods like linear regression have known intrinsic flaws.
- **Our knowledge is imperfect**. "The order of human thought will never reflect the order of things," as Simon Shapin puts it.

Further reading:

- **Michel Foucault**, "The Order of Things" (1966).
- **Paul Schrodt**, "Seven Deadly Sins..." (2010).
- **Simon Shapin**, "Never Pure..." (2010).

Course outline

Univariate statistics

- Introduction
- Datasets
- Distributions
- Estimation

Assignment No. 1

Bivariate statistics

- Significance
- Crosstabulations
- Correlation
- Linear regression

Assignment No. 2

Statistical modelling

- Basics
- Extensions
- Diagnostics
- Conclusion

Final paper

corrected
revised
appended


THIS
IS
STATA
Warner Bros

Congratulations, and thank you.

`exit, clear`