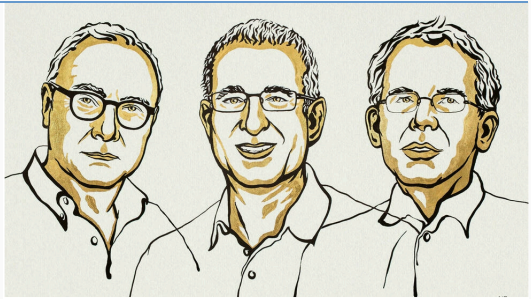
Methods of Applied Statistics I

STA2101H F LEC9101

Week 5

October 13 2021



Canadian David Card, Israeli-American Joshua Angrist and Dutch-American Guido Imbens on Monday won the Nobel Economics Prize for insights into the labour market and "natural experiments", the jury said. © Niklas Elmehed, Nobel Prize Outreach 2021 Ill.

Today

Start Recording

1. Upcoming events, HW5
2. Design of Studies
3. Linear Regression Part 5: recap, factor variables, random and mixed effects, randomization designs
4. In the News
5. Third hour – HW Comments, Refresher on factorial experiments

Syllabus – update

Syllabus Updated Sep 13 STA 2101F: Methods of Applied Statistics I 2021

Week	Date	Methods	References
1	Sept 15	Review of Linear Regression	LM-2 Ch.2-4; LM-1 Ch.2-3; CD Ch.1; SM Ch.8.2.1, 8.3
2	Sept 22	Model comparison, diagnostics, collinearity, factors, model-checking	SM Ch.8.6, 8.7; LM-2 Ch.3, Ch.10, Ch.14-1,2; LM-1 Ch.4, Ch.13; CD Ch.6; SM Ch.8.5, 8.6
3	Sept 29	Model Selection, Types of Studies	LM-2 Ch.10; LM-1 Ch. 8; CD Ch.1,2; SM 8.7.1
4	Oct 6	Factor variables; random and mixed effects; principles of	LM-2 Ch.14-17; LM-1 Ch.14-16 CD Ch.4; SM Ch.9.2.1

Upcoming

- What I Learned Applying for Faculty Positions
- Thursday 3.30 Zoom Link

Jessica Gronsbell, University of Toronto

Short Bio

Jesse is an Assistant Professor in the department. Her primary interest is in the development of statistical methods for modern digital health data sources such as electronic health records and mobile health data.

Title: What I Learned Applying for Faculty Positions

I completed my academic job search a couple of years ago. It was an overwhelming, exhausting, and amazing experience. I didn't know much about the process when I started so I hope to share what I learned. I will cover everything from start to finish - how to prepare your application materials and identify departments that are hiring, what to expect during interviews, and tips for negotiating and decision making. Hope to see you there!

- Friday Oct 15 Toronto Data Workshop Zoom link

https://utoronto.zoom.us/j/84277066292

Toronto Data Workshop this Friday, 15 October, at noon (Toronto time) hosts Drew Stommes, Yale University.

Drew will discuss his recent working paper 'On the reliability of published findings using the regression discontinuity design in political science', https://arxiv.org/abs/2109.14526. As a reminder, regression discontinuity is one of the foundational approaches in causal inference. Drew finds that 'many published using the RD design are exaggerated, if not entirely spurious'.

Due October 20 2021 11.59 pm

Homework to be submitted through Quercus

LM-2, Ex.16.6: The "High School and Beyond" data is found in `hsb`.

(a) Model the math score in terms of the five factors: `gender`, `race`, `ses`, `schtyp` and `prog`. Include all second-order interactions but no hgiher-order interactions. How many regression parameters does your model use? Explain how this can be calculated in terms of the number of levels for each factor.

(b) Determine whether some two-way interactions can be eliminated using the `anova` function.

(c) Determine whether some two-way interactions can be eliminated but now using the `drop1` function. Why do the results differ from the previous question? Which method of testing do you think is better, and why?

(d) Fit a model with only main effects and compare it to the model with all two-way interactions. Which model do you prefer and why?

(e) Check the regression diagnostics for the main effects only model and report on any anomalies.

(f) Summarize your conclusions from the analysis of the data in plain language, at most one paragraph.

(g) *Bonus/PhD:*

- common objectives
- to avoid systematic error, that is distortion in the conclusions arising from sources that do not cancel out in the long run
- to reduce the non-systematic (random) error to a reasonable level by replication and other techniques
- to estimate realistically the likely uncertainty in the final conclusions
- to ensure that the scale of effort is appropriate

... design of studies

- we concentrate largely on the careful analysis of individual studies
- in most situations synthesis of information from different investigations is needed

- but even there the quality of individual studies remains important
- examples include overviews (such as the Cochrane reviews)

- in some areas new investigations can be set up and completed relatively quickly; design of individual studies may then be less important

... design of studies

- formulation of a plan of analysis
- establish and document that proposed data are capable of addressing the research questions of concern

- main configurations of answers likely to be obtained should be set out
- level of detail depends on the context

- even if pre-specified methods must be used, it is crucial not to limit analysis
- planned analysis may be technically inappropriate
- more controversially, data may suggest new research questions or replacement of objectives

- latter will require confirmatory studies

Unit of study and analysis

- smallest subdivision of experimental material that may be assigned to a treatment
  context: Expt
- Example: RCT – unit may be a patient, or a patient-month (in crossover trial)
- Example: public health intervention – unit is often a community/school/...

- split plot experiments have two classes of units of study and analysis
- in investigations that are not randomized, it may be helpful to consider what the primary unit of analysis would have been, had a randomized experiment been feasible

- the unit of analysis may not be the unit of interpretation – ecological bias
  systematic difference between impact of *x* at different levels of aggregation
- on the whole, limited detail is needed in examining the variation within the unit of study

- CD: Illustration – “For country- or region-based mortality data, countries of regions respectively may reasonably constitute the units of analysis with which to assess the relationship of the data to dietary and other features
- “Yet the objective is interpretation at an individual person level
- “The situation may be eased if supplementary data on explanatory variables are available at the individual level, because this may clarify the connection of between-unit and within-unit variation”

The Economist

Menu | Weekly edition | Search

Science & technology

Sep 25th 2021 edition ›

The comforts of religion

Religious belief really does seem to draw the sting of poverty

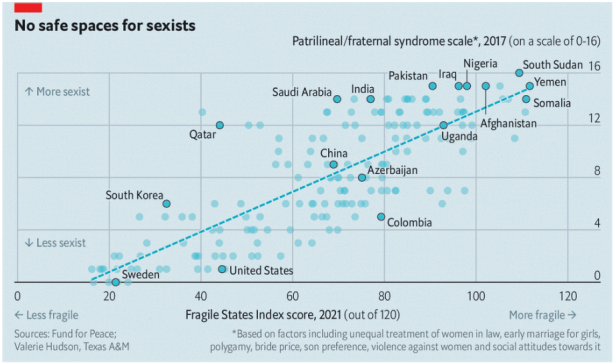
Whether the cause is spiritual or social remains to be determined

- " This set of data reproduced the finding that economic development amplifies the link between mental health and status. It also supported the idea that religiosity could attenuate that effect.

- Among rich countries, for instance, those with higher levels of self-reported religious belief had a weaker relationship between status and mental health.

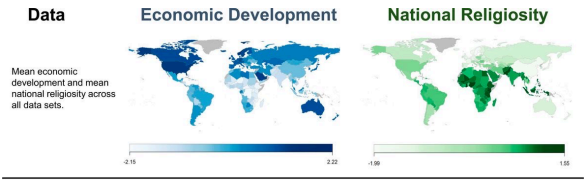
- "The upshot is that religion seems to protect people from at least some of the unpleasant effects of poverty."

Aside: ecological bias



Depicted are the moderating effects of national economic development and national religiosity on the association between SES and well-being in all three data sets.

Distribution of national economic development, national religiosity, and estimated means of the cross-level interactions (model 3).



link

ANCOVA in PNAS paper

see also LM-2, 14.1; LM-1, 13.1

- $\mathrm{E}(y_i \mid x_i, z_i) = \beta_0 + \beta_1 x_i + \beta_2 z_i + \beta_3 x_i z_i$ `WellB ~ ses + relig + ses:relig`
- y_i: Well-being, x_i: Socio-economic status, z_i: Religiosity
- z_i = (“low”, “medium”, “high”) `model.matrix`

$$
\begin{aligned}
\mathrm{E}(y_i \mid x_i, z_i = \text{"}low\text{"}) &= \beta_0 + \beta_1 x_i \\
\mathrm{E}(y_i \mid x_i, z_i = \text{"}med\text{"}) &= \beta_0 + \beta_2 + (\beta_1 + \beta_4) x_i \\
\mathrm{E}(y_i \mid x_i, z_i = \text{"}hi\text{"}) &= \beta_0 + \beta_3 + (\beta_1 + \beta_5) x_i
\end{aligned}
$$

- as usual, it’s a bit more complicated
- some data collected on people, although
- “Following standard practice, we averaged person-level religiosity within each nation ”
- “Following a standard economic method, we log-transformed the GDP data”

Design of studies: randomized experiments

LM-2 5.3, LM-1 3.6

- unit of analysis – "smallest subdivision of the experimental material such that two distinct units might be randomized to different treatments"
  - example: patient in a clinical trial
  - example: plot of land in an agricultural trial
  - example: units of material in a quality control trial
- advantages of randomization?
  - balances other potential influences on responses
  - avoidance of systematic error
  - a systematic difference in response not due to treatment under study
- randomization can make causal interpretation more plausible

permutation test LM-2 §5.3

Avoidance of systematic error

CD §2.4

- “distortion in the conclusions arising from irrelevant sources that do not cancel out in the long run”
- can arise through systematic aspects of, for example, a measuring process, or the spatial or temporal arrangement of units
- this can often be avoided by design, or adjustment in analysis
- can arise by the entry of personal judgement into some aspect of the data collection process
- this can often be avoided by randomization and blinding

Observational studies

SM 9.1, LM-2 5.4, LM-1 3.7

- "treatment" is not assigned to units, only observed
- any observed effect of treatment cannot be assumed to be causal

"correlation is not causation"

- we can try to assess the effect by controlling for other variables that may also influence the response
- but we cannot control for unmeasured variables

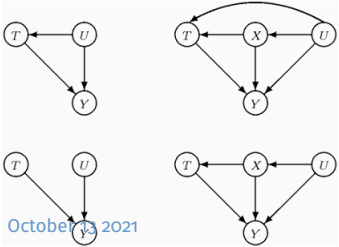


Figure 9.1 Directed acyclic graphs showing consequences of randomization. An arrow from T to Y indicates dependence of Y on T, and so forth. In general both response Y and treatment T may depend on properties U of units (upper left). Randomization (lower left) makes treatments and units independent, so any observed dependence of Y on T cannot be ascribed to joint dependence on U. The upper right graph shows the general dependence of Y, T, and covariates X on U.

Types of observational studies

- secondary analysis of data collected for another purpose

- estimation of a some feature of a defined population (could in principle be found exactly)
- tracking across time of such features

- study of a relationship between features, where individuals may be examined
  - at a single time point
  - at several time points for different individuals
  - at different time points for the same individual

- census

- meta-analysis: statistical assessment of a collection of studies on the same topic

Linear regression recap

- collinearity: difficult to separate effects of one or more covariates on the response

  or linear combinations of them
- model selection using p-values: forward, backward, stepwise
- model selection using formal criteria: AIC, BIC, C_p, R^2_a
- selecting explanatory variables through penalization (Lasso)

- model building: hierarchical principle
- model building: understanding/learning the science
- model building: empirical models, substantive models "All models are wrong"
- model building for estimation/systematic study vs model building for prediction

Factor variables: modelling

FLM-2 §14.4; FLM-1 §13.3

- a factor variable is treated as categorical
- a non-factor variable is treated as continuous
- it depends on the application which is preferred

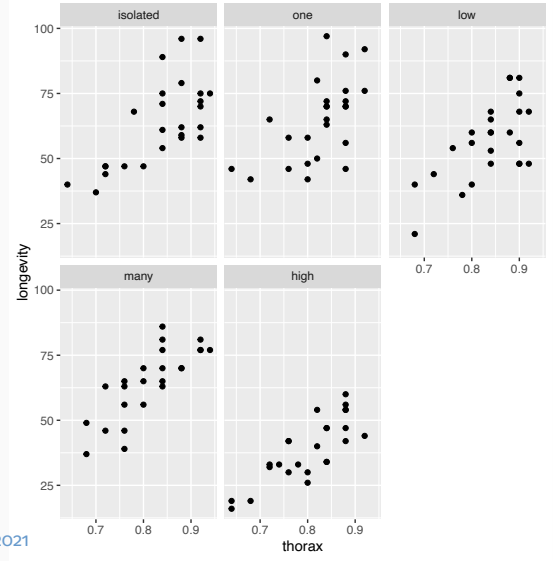
- a linear model with one factor and one continuous variable might be written as, for example:

$$y_{ij} = \mu + \alpha_j + \beta x_{ij} + \epsilon_{ij}, \quad j = 1, \ldots, J; \quad i = 1, \ldots m$$

- linear in x, but arbitrary changes in $\mathbb{E}(y)$ by category (here indexed by j)
- `R` doesn't distinguish this at the modelling phase:
  `lm(response ~ variable1 + variable2, data = ...)`
- but uses metadata in the data frame to accommodate factors
- `is.factor(variable)` and `newvar <- as.factor(oldvar)` are helpful

Factor variables: modelling

fruitfly.Rmd



→ **fruitfly.Rmd**

- Why bother with special techniques for factor variables since we can fit them all using `lm`?
- If the experiment is designed – meaning treatment assignment under the control of the investigator, then we have stronger conclusions
- If the experiment is balanced, then the estimates of the effects of different factors are independent

  X^TX is orthogonal
- If the experiment is replicated, we can obtain reliable estimates of σ^2
- If the experiment is blocked, we can remove sources of error

- Cycling: SM Example 8.4, 8.8, 8.12, 8.22 – designed experiment with 3 factors, each at 2 levels and each of these 8 combinations used twice, for a sample size of 16
- Poison: SM Example 8.25 – 2 factors, one has 4 levels, one has 3 levels, repeated four times, for a sample size of $12 \times 4 = 48$
- Wafer: HW1 LM Ch3 – 4 factors, each at 2 levels

- Some classical designs:
  - completely randomized:
    SM Example 9.2 – one factor with 4 levels
    SM Example 9.6 (and 8.25) – two factors with 3 and 4 levels, replicated
  - randomized blocks:
    SM Example 9.3 – one treatment factor with 4 levels, one blocking factor with 8 levels
  - incomplete RB: SM Example 9.4:
  - Latin square: SM Example 9.5

- design: one factor with I levels; J responses at each level
- model

$$y_{ij} = \mu + \alpha_i + \epsilon_{ij}, \quad j = 1, \ldots J; i = 1, \ldots I; \quad \epsilon_{ij} \sim (0, \sigma^2)$$

- parameters:
  - $\mu = \mathbb{E}(y_{ij})$ if all $\alpha_i \equiv 0$;
  - α_2 is change from μ in $\mathbb{E}(y_{2j})$ in group 2, etc. using the R convention that $\alpha_1 = 0$
  - ϵ_{ij} is noise variation in response not attributed to factor variable

Analysis of variance table

Term	degrees of freedom	sum of squares	mean square	F-statistic
treatment	$(I-1)$	$\sum_{ij}(\bar{y}_{i.} - \bar{y}_{..})^2$	$\sum_{ij}(\bar{y}_{i.} - \bar{y}_{..})^2/(I-1)$	$MS_{treatment}/MS_{error}$
error	$I(J-1)$	$\sum_{ij}(y_{ij} - \bar{y}_{i.})^2$	$\sum_{ij}(y_{ij} - \bar{y}_{i.})^2/\{I(J-1)\}$	
total(corrected)	$IJ-1$	$\sum_{ij}(y_{ij} - \bar{y}_{..})^2$		

Term	degrees of freedom	sum of squares	mean square	F-statistic
treatment	$(I-1)$	$\sum_{ij}(\bar{y}_{i.}-\bar{y}_{..})^2$	$\sum_{ij}(\bar{y}_{i.}-\bar{y}_{..})^2/(I-1)$	$MS_{treatment}/MS_{error}$
error	$I(J-1)$	$\sum_{ij}(y_{ij}-\bar{y}_{i.})^2$	$\sum_{ij}(y_{ij}-\bar{y}_{i.})^2/\{I(J-1)\}$	
total(corrected)	$IJ-1$	$\sum_{ij}(y_{ij}-\bar{y}_{..})^2$		

Term	degrees of freedom	sum of squares	mean square	F-statistic
treatment	$(I-1)$	$SS_{between}$	$MS_{between}$	$MS_{between}/MS_{within}$
error	$I(J-1)$	SS_{within}	MS_{within}	
total(corrected)	$IJ-1$	SS_{total}		

$$\begin{aligned}\sum_{ij}(y_{ij}-\bar{y}_{..})^2 &= \sum_{ij}(y_{ij}-\bar{y}_{i.}+\bar{y}_{i.}-\bar{y}_{..})^2 \\ &= \sum_{ij}(\bar{y}_{i.}-\bar{y}_{..})^2+\sum_{ij}(y_{ij}-\bar{y}_{i.})^2\end{aligned}$$

See SM Table 9.3 and 9.4; FLM-2 §15.2; FlM-1 §14.2

Table 9.3 Data on the teaching of arithmetic.

Group	Test result y									Average	Variance
A (Usual)	17	14	24	20	24	23	16	15	24	19.67	17.75
B (Usual)	21	23	13	19	13	19	20	21	16	18.33	12.75
C (Praised)	28	30	29	24	27	30	28	28	23	27.44	6.03
D (Reproved)	19	28	26	26	19	24	24	23	22	23.44	9.53
E (Ignored)	21	14	13	19	15	15	10	18	20	16.11	13.11

Table 9.4 Analysis of variance for data on the teaching of arithmetic.

Term	df	Sum of squares	Mean square	F
Groups	4	722.67	180.67	15.3
Residual	40	473.33	11.83	

Components of variance

- in some settings, the one-way layout refers to sampled groups
- not an assigned treatment
- e.g. a sample of people, with several measurements taken on each person
- $y_{ij} = \mu + \alpha_i + \epsilon_{ij}$ as before, but with different assumptions

Subject					
1	2	3	4	5	6
68	49	41	33	40	30
42	52	40	27	45	42
69	41	26	48	50	35
64	56	33	54	41	44
39	40	42	42	37	49
66	43	27	56	34	25
29	20	35	19	42	45

Table 9.22 Blood data: seven measurements from each of six subjects on a property related to the stickiness of their blood.

...components of variance

- $y_{ij} = \mu + \alpha_i + \epsilon_{ij}, \quad \epsilon_{ij} \sim (0, \sigma^2), \quad \alpha_i \sim (0, \sigma_a^2) \qquad i = 1, \ldots, T; j = 1 \ldots R$
- variance of response within subjects
- variance of response between subjects

- as before,

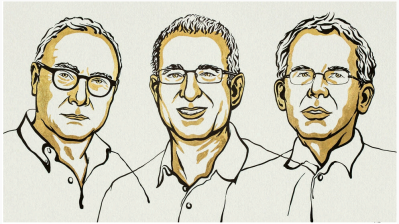
$$\sum_{ij}(y_{ij} - \bar{y}_{..})^2 = \sum_{ij}(\bar{y}_{i.} - \bar{y}_{..})^2 + \sum_{ij}(y_{ij} - \bar{y}_{i.})^2$$

- random effects induce dependence among measurements on the same subject: ntbc

$$\text{cov}(y_{ij}, y_{ij'}) = \sigma_A^2$$

- $SS_{within} \sim \sigma^2 \chi^2_{T(R-1)}$ $\qquad SS_{between} \sim (R\sigma_A^2 + \sigma^2)\chi^2_{T-1}$ leads to F-test for $H_0 : \sigma_A^2 = 0$

In the News



Canadian David Card, Israeli-American Joshua Angrist and Dutch-American Guido Imbens on Monday won the Nobel Economics Prize for insights into the labour market and "natural experiments", the jury said. © Niklas Elmehed, Nobel Prize Outreach 2021 III.

The Nobel in economics goes to three who find experiments in real life.

The 2021 Nobel Memorial Prize in Economic Sciences honored the work of David Card, Joshua Angrist and Guido Imbens, which changed the way that labor markets in particular are studied.

Claudio Bresciani/TT News Agency, via Agence France-Presse — Getty Images

Causal inference from observational studies

- qualitative support for causation LM-2 5.7
- propensity score matching LM-2 5.5
- instrumental variables
- regression discontinuity designs
- $y_i = \beta_0 + \beta_1 x_i + \beta_2 z_i + \beta_3 x_i z_i + \epsilon_i, \qquad i = 1, \ldots, n; \quad \epsilon \sim (0, \sigma^2)$
- x is continuous, used as a ranking variable; z is a two-level factor "treatment"
- Owen & Varian Tuning the tie-breaker design arxiv