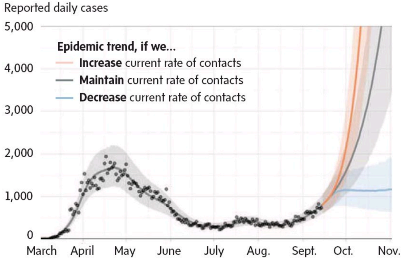
Methods of Applied Statistics I

STA2101H F LEC9101

Week 3

September 24 2020



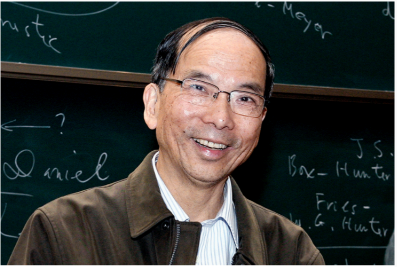
JOHN SOPINSKI/THE GLOBE AND MAIL
SOURCE: PUBLIC HEALTH AGENCY OF CANADA

Today

Start Recording

1. Office hours, Covid competition, HW1
   OH: Wednesday 9am-10.30am Monday 4pm-5.30pm, 7pm-8pm in Course Room
2. In the News: moon-shot covid testing; predictions for Canada
3. Types of studies
4. Linear Regression Part 3: model checking, model selection, $p > n$, weighted LS, mixed effects models
5. (2–3pm) Discussion, questions, etc.

- September 28 3.30 – 4.30
- September 29 3.30 – 4.30
- `http://www.fields.utoronto.ca/activities/20-21/Je`



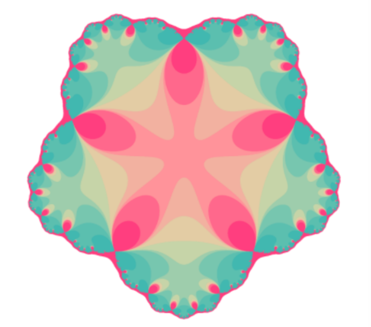
2020 Distinguished Lecture Series in Statistical Sciences: Jeff Wu

This just in

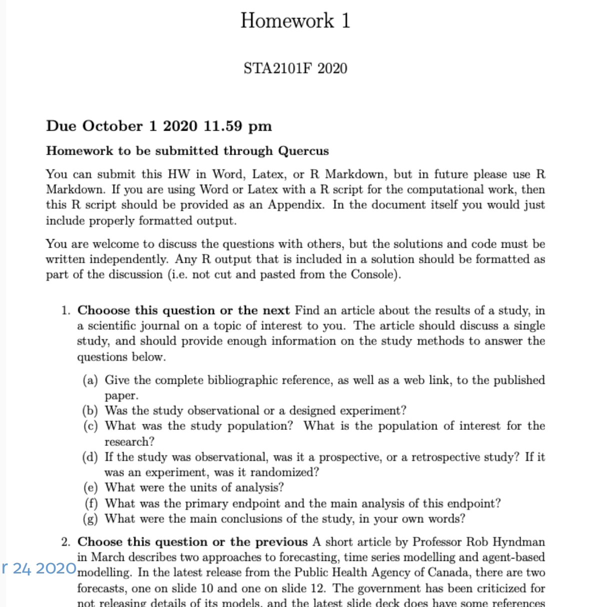
DoSS welcome mixer and aRt class Friday Sept 25 10.10 – 11.30

Zoom link

Registration Link



http://www.utstat.utoronto.ca/reid/sta2101f/Hw1Questions.pdf



Homework 1

STA2101F 2020

Due October 1 2020 11.59 pm

Homework to be submitted through Quercus

You can submit this HW in Word, Latex, or R Markdown, but in future please use R Markdown. If you are using Word or Latex with a R script for the computational work, then this R script should be provided as an Appendix. In the document itself you would just include properly formatted output.

You are welcome to discuss the questions with others, but the solutions and code must be written independently. Any R output that is included in a solution should be formatted as part of the discussion (i.e. not cut and pasted from the Console).

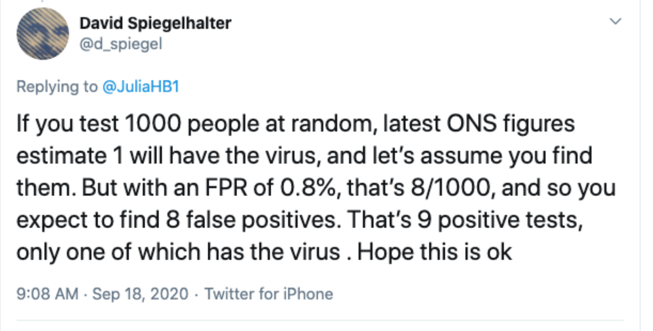
1. **Chooose this question or the next** Find an article about the results of a study, in a scientific journal on a topic of interest to you. The article should discuss a single study, and should provide enough information on the study methods to answer the questions below.

   (a) Give the complete bibliographic reference, as well as a web link, to the published paper.
   (b) Was the study observational or a designed experiment?
   (c) What was the study population? What is the population of interest for the research?
   (d) If the study was observational, was it a prospective, or a retrospective study? If it was an experiment, was it randomized?
   (e) What were the units of analysis?
   (f) What was the primary endpoint and the main analysis of this endpoint?
   (g) What were the main conclusions of the study, in your own words?

2. **Choose this question or the previous** A short article by Professor Rob Hyndman in March describes two approaches to forecasting, time series modelling and agent-based modelling. In the latest release from the Public Health Agency of Canada, there are two forecasts, one on slide 10 and one on slide 12. The government has been criticized for not releasing details of its models, and the latest slide deck does have some references

Moonshot testing in the UK

```
##   prevvals       ppv
## 1   0.0001 0.0079373
## 2   0.0010 0.0741427
## 3   0.0050 0.2867384
## 4   0.0100 0.4469274
## 5   0.0500 0.8080808
## 6   0.1000 0.8988764
```



David Spiegelhalter
@d_spiegel

Replying to @JuliaHB1

If you test 1000 people at random, latest ONS figures estimate 1 will have the virus, and let's assume you find them. But with an FPR of 0.8%, that's 8/1000, and so you expect to find 8 false positives. That's 9 positive tests, only one of which has the virus . Hope this is ok

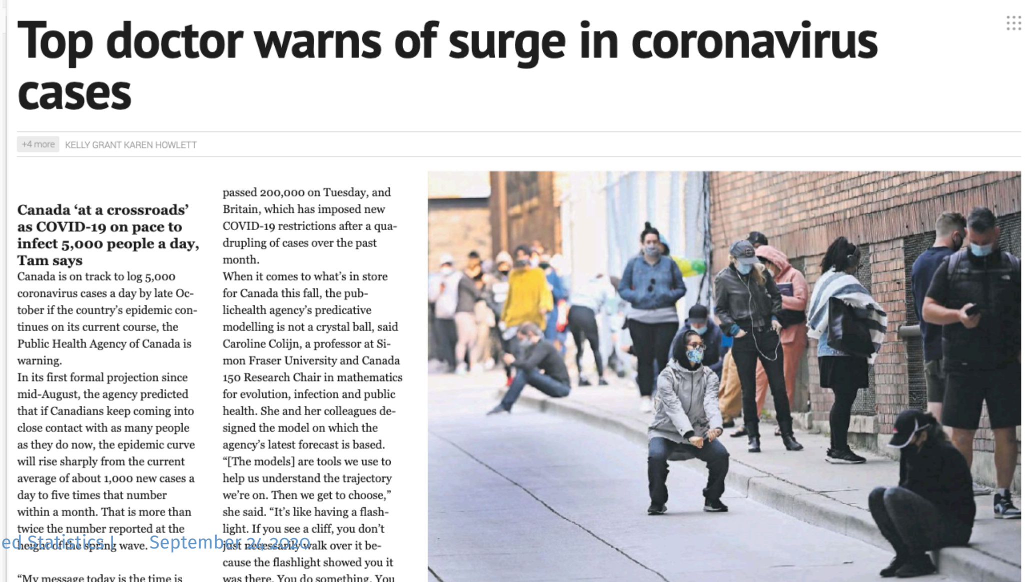
9:08 AM · Sep 18, 2020 · Twitter for iPhone

ppv: Probability $(C+ \mid T+)$

positive predictive value

I used False Positive Rate of 1%   and didn't **assume** "you find them"

Globe & Mail, Sep 23



Top doctor warns of surge in coronavirus cases

+4 more KELLY GRANT KAREN HOWLETT

Canada ‘at a crossroads’ as COVID-19 on pace to infect 5,000 people a day, Tam says

Canada is on track to log 5,000 coronavirus cases a day by late October if the country’s epidemic continues on its current course, the Public Health Agency of Canada is warning.

In its first formal projection since mid-August, the agency predicted that if Canadians keep coming into close contact with as many people as they do now, the epidemic curve will rise sharply from the current average of about 1,000 new cases a day to five times that number within a month. That is more than twice the number reported at the height of the spring wave.

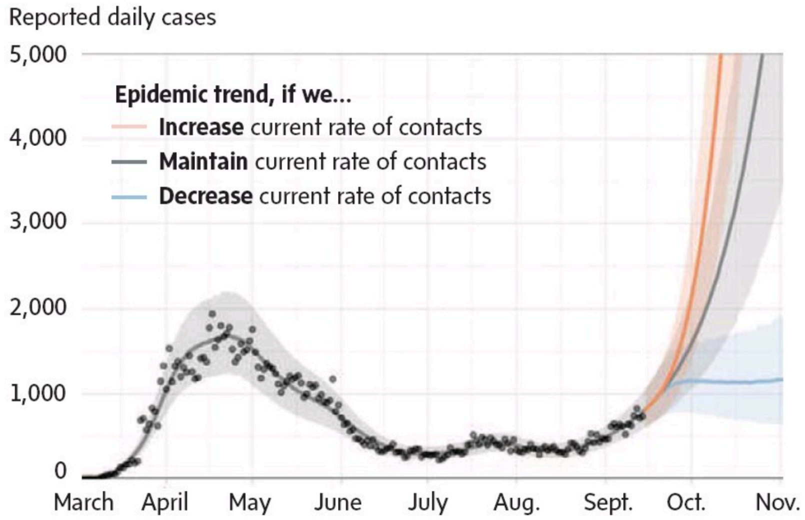
“My message today is the time is

passed 200,000 on Tuesday, and Britain, which has imposed new COVID-19 restrictions after a quadrupling of cases over the past month.

When it comes to what’s in store for Canada this fall, the publichealth agency’s predicative modelling is not a crystal ball, said Caroline Colijn, a professor at Simon Fraser University and Canada 150 Research Chair in mathematics for evolution, infection and public health. She and her colleagues designed the model on which the agency’s latest forecast is based.

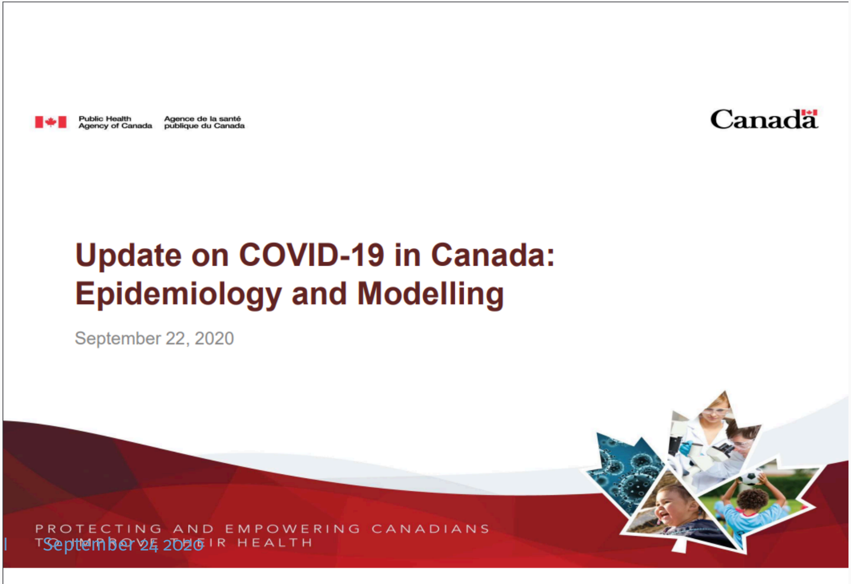
“[The models] are tools we use to help us understand the trajectory we’re on. Then we get to choose,” she said. “It’s like having a flashlight. If you see a cliff, you don’t just necessarily walk over it because the flashlight showed you it was there. You do something. You

… Globe & Mail, Sep 23



JOHN SOPINSKI/THE GLOBE AND MAIL
SOURCE: PUBLIC HEALTH AGENCY OF CANADA

PHAC, Sep 22



Public Health Agency of Canada
Agence de la santé publique du Canada

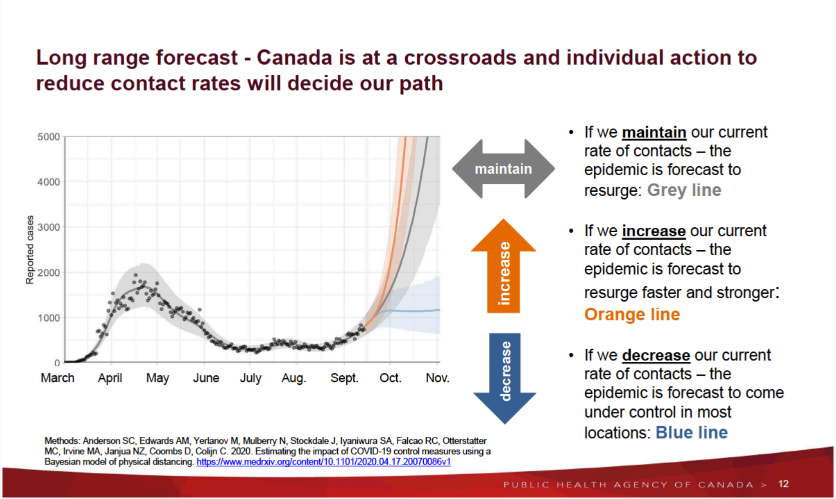
Canada

Update on COVID-19 in Canada: Epidemiology and Modelling

September 22, 2020

PROTECTING AND EMPOWERING CANADIANS TO IMPROVE THEIR HEALTH

PHAC, Sep 22

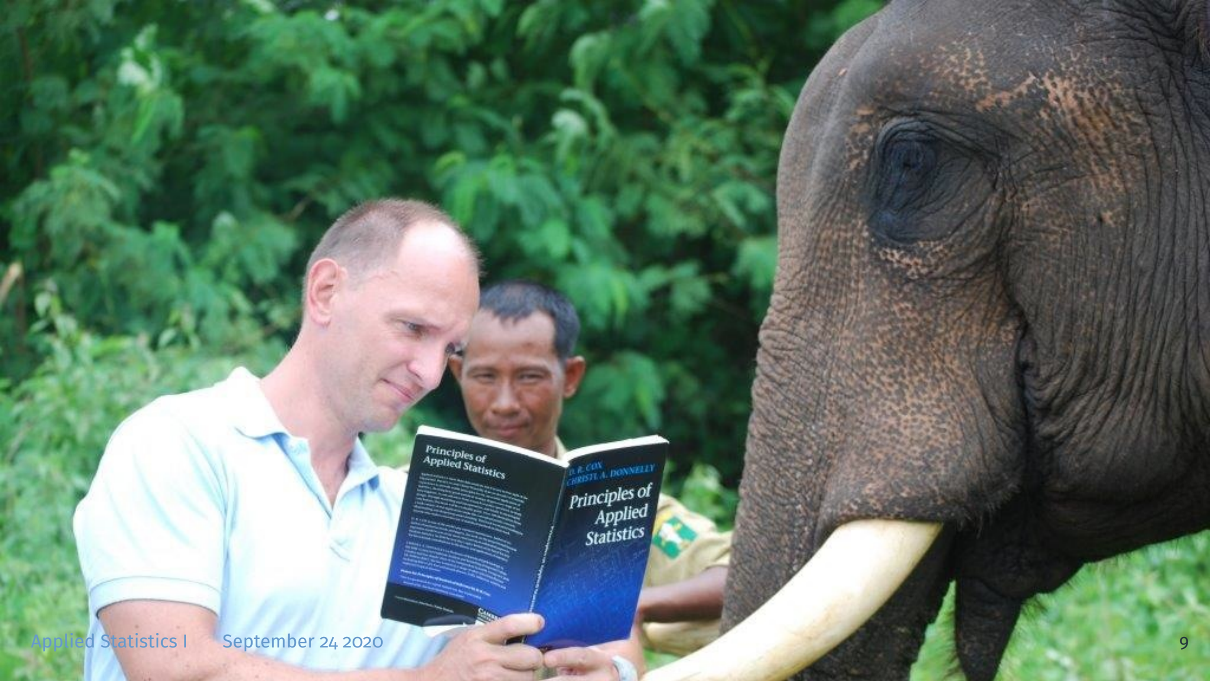


Long range forecast - Canada is at a crossroads and individual action to reduce contact rates will decide our path

- If we **maintain** our current rate of contacts – the epidemic is forecast to resurge: **Grey line**
- If we **increase** our current rate of contacts – the epidemic is forecast to resurge faster and stronger: **Orange line**
- If we **decrease** our current rate of contacts – the epidemic is forecast to come under control in most locations: **Blue line**

Methods: Anderson SC, Edwards AM, Yerlanov M, Mulberry N, Stockdale J, Iyaniwura SA, Falcao RC, Otterstatter MC, Irvine MA, Janjua NZ, Coombs D, Colijn C. 2020. Estimating the impact of COVID-19 control measures using a Bayesian model of physical distancing. https://www.medrxiv.org/content/10.1101/2020.04.17.20070086v1

PUBLIC HEALTH AGENCY OF CANADA > 12



Principles of Applied Statistics
D. R. COX
CHRISTL A. DONNELLY
Principles of Applied Statistics

Statistical science

Cox & Donnelly 2011

- start with a scientific question
- assess how data could shed light on this
- plan data collection
- consider of sources of variation and how careful planning can minimize their impact

- develop strategies for data analysis: modelling, computation, methods of analysis
- assess the properties of the methods and their impact on the question at hand

- communicate the results: accurately

  but not pessimistically
- visualization strategies, conveyance of uncertainties

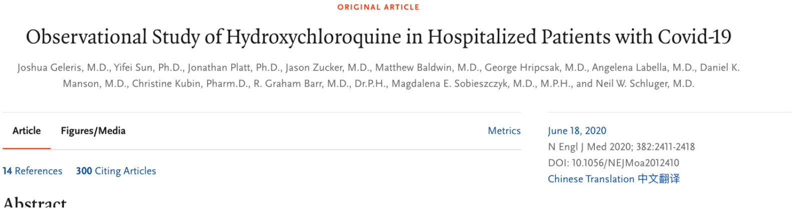
Design and Analysis

- choice of material/individuals to study – “units of analysis”
- “For studies of a new phenomenon it will usually be best to examine situations in which the phenomenon is likely to appear in the most striking form, even if this is in some sense artificial”
- statistical analysis needs to take account of the design (even if statistician enters the project at the analysis stage)
- need to be clear at the design stage about broad features of the statistical analysis – more publicly convincing **and** “reduces the possibility that the data cannot be satisfactorily analysed”
- “it is unrealistic and indeed potentially dangerous to follow an initial plan unswervingly ... it may be a crucial part of the analysis to clarify the research objectives”

Experimental and observational studies

- experiment is a study in which all key elements are under the control of the investigator
- in an observational study key elements cannot be manipulated by the investigator.
- “It often, however, aids the interpretation of an observational study to consider the question: what would have been done in a comparable experiment?”
- Example: early studies of hydroxychloroquine for Covid19 were observational
- Randomized controlled trials are underway

Hydroxychloroquine



ORIGINAL ARTICLE

Observational Study of Hydroxychloroquine in Hospitalized Patients with Covid-19

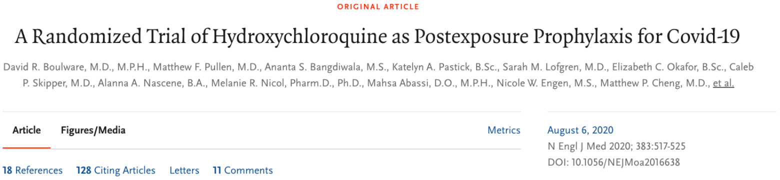
Joshua Geleris, M.D., Yifei Sun, Ph.D., Jonathan Platt, Ph.D., Jason Zucker, M.D., Matthew Baldwin, M.D., George Hripcsak, M.D., Angelena Labella, M.D., Daniel K. Manson, M.D., Christine Kubin, Pharm.D., R. Graham Barr, M.D., Dr.P.H., Magdalena E. Sobieszczyk, M.D., M.P.H., and Neil W. Schluger, M.D.

June 18, 2020
N Engl J Med 2020; 382:2411-2418
DOI: 10.1056/NEJMoa2012410
Chinese Translation 中文翻译

14 References   300 Citing Articles

Abstract

June



ORIGINAL ARTICLE

A Randomized Trial of Hydroxychloroquine as Postexposure Prophylaxis for Covid-19

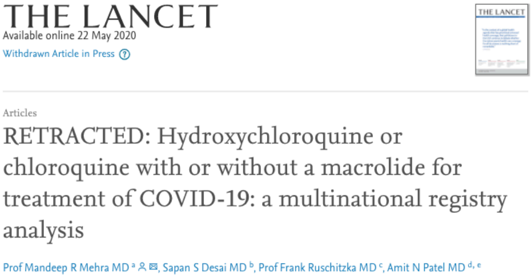
David R. Boulware, M.D., M.P.H., Matthew F. Pullen, M.D., Ananta S. Bangdiwala, M.S., Katelyn A. Pastick, B.Sc., Sarah M. Lofgren, M.D., Elizabeth C. Okafor, B.Sc., Caleb P. Skipper, M.D., Alanna A. Nascene, B.A., Melanie R. Nicol, Pharm.D., Ph.D., Mahsa Abassi, D.O., M.P.H., Nicole W. Engen, M.S., Matthew P. Cheng, M.D., et al.

August 6, 2020
N Engl J Med 2020; 383:517-525
DOI: 10.1056/NEJMoa2016638

18 References   128 Citing Articles   Letters   11 Comments

August

Hydroxychloroquine



THE LANCET
Available online 22 May 2020
Withdrawn Article in Press

Articles

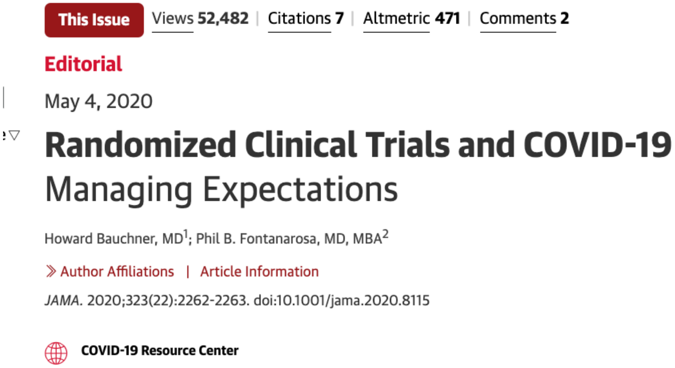
RETRACTED: Hydroxychloroquine or chloroquine with or without a macrolide for treatment of COVID-19: a multinational registry analysis

Prof Mandeep R Mehra MD [a], Sapan S Desai MD [b], Prof Frank Ruschitzka MD [c], Amit N Patel MD [d, e]

retracted

Clinical Trials registry

ClinicalTrials.gov



This Issue Views **52,482** | Citations **7** | Altmetric **471** | Comments **2**

Editorial

May 4, 2020

Randomized Clinical Trials and COVID-19

Managing Expectations

Howard Bauchner, MD[1]; Phil B. Fontanarosa, MD, MBA[2]

» Author Affiliations | Article Information

JAMA. 2020;323(22):2262-2263. doi:10.1001/jama.2020.8115

COVID-19 Resource Center

Design of Studies

- 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

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
- experiment: investigator has complete control over treatment assignment
- census
- meta-analysis: statistical assessment of a collection of studies on the same topic

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

Principles of measurement

- “construct validity – measurements do actually record the features of concern”
- “record a number of different features sufficient to capture concisely the important aspects”
- reliable – i.e. reasonably reproducible
- “cost of the measurements is commensurate with their importance”
- “measurement process does not appreciably distort the system under study”

Types and phases of analysis

- “A general principle, sounding superficial but difficult to implement, is that analyses should be as simple as possible, but no simpler.”
- the method of analysis should be transparent
- main phases of analysis
    - data auditing and screening;
    - preliminary analysis;
    - formal analysis;
    - presentation of conclusions

Recap of Linear Regression Part 2

- Residual Sum of Squares $SS(\hat{\beta}) = (y - X\hat{\beta})^{\mathrm{T}}(y - X\hat{\beta})$
  - can be compared to Total Sum of Squares to see if fitting the model has reduced variation
  - estimates σ^2 when divided by $n - p$
  - can be compared to residual sum of squares from a smaller model to test sub-hypotheses, such as $\beta_2 = \beta_3 = 0$
  - this is useful, and necessary, when some variables are factors
- Checking model assumptions
  - plots: residuals vs. fitted values, QQplots of residuals, plots of Cook's distance
  - looking for evidence of: nonlinearity in residuals (omitted variable(s)?); non-normality of residuals (extreme); influential cases (fit with and without)
  - residuals are often standardized so they have approximately the same variance
- Collinearity
  - if X is ill-conditioned, then estimates of β are unstable
  - interpretation of individual coefficients also unstable
- dependence among responses may be more important, depending on the context

auto-correlation in residuals

Three tasks related to linear regression

-2 Ch.3,4; FLM-1 Ch.3; SM 8.3

- Estimation of β, and estimation of its standard error – for inference about $\mathbb{E}(y \mid x)$

alternatively comparing sub-models using F-tests

- Prediction of y_+, say, given a new vector of explanatory variables x_+
- Model Selection: which explanatory variables do we need for prediction or inference?

These same questions arise in other models such as logistic regression, analysis of survival data, and so on, but the generic linear model is often a good starting point

- Prediction: $y_+ = x_+^{\mathrm{T}}\beta + \epsilon;$ $\quad \hat{y}_+ = x_+^{\mathrm{T}}\hat{\beta};$ $\quad \mathrm{var}(\hat{y}_+) = \sigma^2 x_+ (X^{\mathrm{T}}X)^{-1} x_+$

assuming ...

- error in expected response different from

prediction error $\mathbb{E}(y_+ - \hat{y}_+)^2 = \sigma^2 + \mathrm{var}(\hat{y}_+)$

Model Selection

FLM-2 Ch.10; FLM-1 Ch.8; SM, Ch.8.7

- “analyses should be as simple as possible, but no simpler”
- What variables should we keep in the model ?
- Hierarchical models: some models have a natural hierarchy: polynomials, factorial structure, auto-regressive, sinusoidal, ...
- in these models the ‘highest’ level of the hierarchy is removed first
- e.g. $y = \beta_0 + \beta_1 x + \beta_2 x^2 + \epsilon$ should *not* be simplified to $y = \beta_0 + \beta_2 x^2 + \epsilon$
- e.g. if interaction terms are included, then main effects and other 2nd-order terms also need to be included: $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{12} x_1 x_2 + \beta_{11} x_1^2 + \beta_{22} x_2^2 + \epsilon$
- *not* $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{12} x_1 x_2 + \epsilon$ unless $x = 0/1$
- $y = \beta_0 + \beta_1 \sin(2\pi x) + \beta_2 \cos(2\pi x) + \beta_3 \sin(4\pi x) + \beta_4 \cos(4\pi x) + \epsilon$
- $y_t = \beta_0 + \alpha y_{t-1} + \epsilon$ $y_t = \beta_0 + \alpha_1 y_{t-1} + \alpha_2 y_{t-2} \epsilon$ *not* $y_t = \beta_0 + \alpha_2 y_{t-2} + \epsilon$

… Model Selection

FLM-2 Ch.10; FLM-1 Ch.8; SM, Ch.8.7

- testing procedures: forward selection, backward selection, stepwise selection
- it is quite common to fit all explanatory variables, and then drop if $p > 0.05$
- if estimates and estimated standard errors don't change very much, may be okay
- if estimates and estimated standard errors change a lot, cause for concern
- if estimates change sign, points to possibly extreme confounding

```
library(SMPracticals)
data(nuclear)
head(nuclear)
lm1 <- lm(log(cost) ~ date + log(t1) + log(t2) +
            log(cap) + pr + ne + ct + bw +
            log(cum.n) + pt, data = nuclear)
step(lm1)
```

SM Example 8.29; see esp. Table 8.14

... Model Selection

FLM-2 Ch.10; FLM-1 Ch.8; SM, Ch.8.7

```
step(lm1)
...
Call:
lm(formula = log(cost) ~ date + log(t2) + log(cap) + pr + ne +
    ct + log(cum.n) + pt, data = nuclear)

Coefficients:
(Intercept)         date      log(t2)     log(cap)
  -15.22561      0.22722      0.30186      0.68246
         pr           ne           ct   log(cum.n)
   -0.09336      0.25895      0.11462     -0.07873
         pt
   -0.21572
```

as implemented in `MASS`, `step` uses the AIC criterion to choose the model, not the *p*-value

... Model Selection

FLM-2 Ch.10; FLM-1 Ch.8; SM, Ch.8.7

- Criterion-based procedures

most widely used

- AIC, BIC, Mallows C_p, R_a^2

RSS: residual sum of squares

- $$AIC = n\log(RSS/n) + 2p$$
- $$BIC = n\log(RSS/n) + \log(n)p$$
- $$C_p = RSS_p/\tilde{\sigma}^2 + 2p - n$$
- $$R_a^2 = 1 - \frac{\tilde{\sigma}^2_{model}}{TSS/(n-1)}$$
- SM has yet another version AIC_c which may be better than AIC for linear models
- C_p and R_a^2 are only useful for linear models; AIC and BIC more general

see Appendix A of Kirchmeier-Young (wildfire), for example

Biased estimation

FLM-2 Ch.11.3, 11.4; FLM-1 Ch.9.3

- if p is large relative to n, or even $p > n$, then LS estimates don't exist
- $X^{\mathrm{T}}X$ is not invertible

X has rank $\min(n, p)$

- ridge regression:

$$\min_{\beta} \{(y - X\beta)^{\mathrm{T}}(y - X\beta) + \lambda \sum_{j=1}^{p} \beta_j^2\}$$

-

$$\hat{\beta}_{ridge} = (X^{\mathrm{T}}X + \lambda I)^{-1}X^{\mathrm{T}}y$$

$(X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}y$

- Lasso regression

$$\min_{\beta} \{(y - X\beta)^{\mathrm{T}}(y - X\beta) + \lambda \sum_{j=1}^{p} |\beta_j|\}$$

- minimizing value $\hat{\beta}_{Lasso}$ has many elements = 0 (depending on λ)

model selection