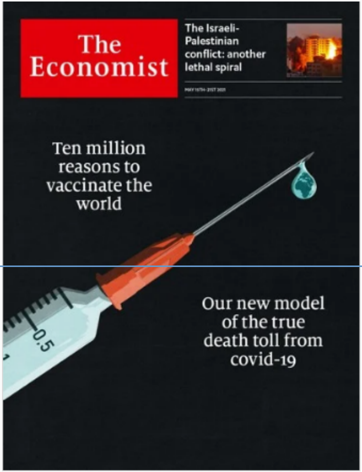
Methods of Applied Statistics I

STA2101H F LEC9101

Week 1

September 15 2021



MAY 15TH 2021

Ten million reasons to vaccinate the world

Today

Start Recording

1. Course introduction: technical issues, course details, evaluation, syllabus, people
2. Upcoming events of interest
3. Review of linear regression
4. In the news: excess deaths
5. Computing: RStudio, RMarkdown

Technical Issues

- If **you** are having technical difficulties
  - If possible, send **me** a message in chat
  - Try leaving the class and re-joining
  - Try switching to Chrome if you are using something else
  - Don't panic, the lecture is being recorded and both the recording and the slides will be posted
- If **Prof** is having technical difficulties
  - Check the chat to see if there's any information there
  - If I've disappeared completely, give me 15 minutes before closing the call
  - Look for an announcement on Quercus
  - Don't panic, Prof, you'll figure it out

STA 2101F: Methods of Applied Statistics I
Wednesday, 10am – 1 pm Eastern September 15 – December 8 2021

Updated September 14

From the calendar:

This course will focus on principles and methods of applied statistical science. It is designed for MSc and PhD students in Statistics, and is required for the Applied Paper of the PhD comprehensive exams. The topics covered include: planning of studies, review of linear models, analysis of random and mixed effects models, model building and model selection, theory and methods for generalized linear models, and an introduction to nonparametric regression. Additional topics will be introduced as needed in the context of case studies in data analysis.
Prerequisites: ECO374H1/ECO375H1/STA302H1 (regression); STA305H1 (design of studies)

Course Delivery:

On September 15 and 22, the class will be delivered online at the scheduled time

STA 2101F: Methods of Applied Statistics I
Wednesday, 10am – 1 pm Eastern September 15 – December 8 2021

Updated September 14

From the calendar:

This course will focus on principles and methods of applied statistical science. It is designed for MSc and PhD students in Statistics, and is required for the Applied Paper of the PhD comprehensive exams. The topics covered include: planning of studies, review of linear models, analysis of random and mixed effects models, model building and model selection, theory and methods for generalized linear models, and an introduction to nonparametric regression. Additional topics will be introduced as needed in the context of case studies in data analysis.
Prerequisites: ECO374H1/ECO375H1/STA302H1 (regression); STA305H1 (design of studies)

Course Delivery:

On September 15 and 22, the class will be delivered online at the scheduled time

Course Description

- Course Delivery

Piazza, Notifications

- Grading
- Academic Integrity
- Computing
- References

Modules

- Contact

Use **Piazza** for course questions; email for personal questions

Course Introductions

- about me ⟶

- TA: Ruoyong Xu

- Please turn on your camera to introduce yourself

- Name, program, current location (city)

Today

1. Course introduction: technical issues, course details, evaluation, syllabus, people
2. Upcoming events of interest
3. Review of linear regression
4. Steps in analysis
5. In the news: excess deaths

Upcoming events 1

Thursday

About Lucy Gao

Lucy is an Assistant Professor in the Department of Statistics and Actuarial Science at the University of Waterloo. She received her PhD in Biostatistics from the University of Washington. Her research interests are in statistical learning, selective inference, and experiment design.

- Weekly Department Seminar Series — Sep 16 15.30 EDT
- Selective Inference on Trees
- via Zoom

Upcoming events 2

September 17 13.00 EDT

- Launch of Data Sciences Institute, U of T

Register here

- Speakers:
  - Jennifer Chayes, UC Berkeley
  - Andrew Gelman, Columbia
  - Rob Tibshirani, Stanford

- Distinguished Lecture Series in Statistical Sciences — Register here

- September 20, 2021, 4-5pm Eastern - Bayesian Modelling and Analysis of Challenging Data: Making New Sources of Data Trustworthy
- September 21, 2021, 4-5pm Eastern - Bayesian Modelling and Analysis of Challenging Data: Identifying the Intrinsic Dimension of High-Dimensional Data

Today

1. Course introduction: technical issues, people, course details, evaluation, syllabus

Timer

2. Upcoming events of interest
3. Review of linear regression
4. Steps in analysis
5. In the news: excess deaths

- Model:

$$Y_{n\times 1} = \quad X_{n\times p}\beta_{p\times 1} + \quad \epsilon_{n\times 1}$$

- Equivalently:

$$y_i =$$

- Standard Assumptions
  - y_i independent equivalently ϵ_i independent — y is often called response
  - $\mathbb{E}(\epsilon_i) = 0$ — why?
  - $\text{var}(\epsilon_i) = \sigma^2$ — constant
  - x_i known, β to be estimated — x_i often called explanatory variables
- More concisely:

$$\mathbb{E}(Y \mid X) = \qquad , \quad \text{var}(Y \mid X) =$$

I ??

Nice big equation:

$$\begin{pmatrix} y_1 \\ \vdots \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} x_{11} & \dots & x_{1p} \\ \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots \\ x_{n1} & \dots & x_{np} \end{pmatrix} \begin{pmatrix} \\ \vdots \\ \\ \end{pmatrix} + \begin{pmatrix} \vdots \\ \vdots \end{pmatrix}$$

Or, if you prefer:

$$y_i = x_{i1}\beta_1 + x_{i2}\beta_2 + \cdots + x_{ip}\beta_p + \epsilon_i, \quad \epsilon_i \qquad i = 1, \dots, n$$

Or, if you prefer:

$$\mathbb{E}(y_i \mid x_i) = x_i^{\mathrm{T}}\beta, \qquad \mathrm{var}(y_i \mid x_i) = \sigma^2, \qquad i = 1, \dots, n$$

y_i independent

- often not completely clear: X might be fixed by design, or measured on each individual

  e.g.?

- If measured, then should we consider its distribution? E.g. should our model be $(y_i, x_i^{\mathrm{T}}) \sim$??

  some $(p+1)$-dimensional distribution

- Almost always in regression settings we condition on X, as on previous slide

  ancillary statistic

- often not emphasized: interpretation of β_j
  - version 1: effect on the expected response of a unit change in jth explanatory variable, all other variables held fixed
  - version 2:

$$\beta_j = \frac{\partial \mathbb{E}(y_i \mid x_{ij})}{\partial x_{ij}} \qquad \frac{\partial \mathbb{E}(y \mid x_j)}{\partial x_j}$$

notation ambiguous, see CD §6.5.2

Least squares estimation

- Definition

$$\hat{\beta}_{LS} := \min_{\beta} \sum_{i=1}^{n} (y_i - x_i^{\mathrm{T}} \beta)^2$$

- Equivalently,

- Equivalently,

$$\hat{\beta}_{LS} :=$$

*L*2 distance

- Equivalently, $\hat{\beta}_{LS}$ is the solution of the score equation

$$X^{\mathrm{T}}(y - X\beta) = 0$$

?how?

- Solution

$$\hat{\beta}_{LS} =$$

check dimensions

… least squares estimation

- Solution

$$\hat{\beta}_{LS} = (X^{\mathrm{T}}X)^{-1}(X^{\mathrm{T}}y)$$

check dimensions

- Expected value

$$\mathbb{E}(\hat{\beta}_{LS}) =$$

why?

- Least squares estimates are unbiased
- Variance

really variance-covariance matrix

$$\text{var}(\hat{\beta}_{LS}) = (X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}\text{var}(y)X(X^{\mathrm{T}}X)^{-1} = (X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}\sigma^2 IX(X^{\mathrm{T}}X)^{-1} = \sigma^2(X^{\mathrm{T}}X)^{-1}$$

ASIDE: here and following all assume X is fixed

- If we further assume $\epsilon_i \sim N(0, \sigma^2)$ (and independent across i), then
- $y \mid X \sim N(X\beta, \sigma^2 I)$, and
- the likelihood function is

$$L(\beta, \sigma^2; y) = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left\{-\frac{1}{2\sigma^2}(y - X\beta)^T(y - X\beta)\right\},$$

- the log-likelihood function is

$$\ell(\beta, \sigma^2; y) = -\frac{n}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}(y - X\beta)^{\mathrm{T}}(y - X\beta),$$

constants in params don't matter

- the maximum likelihood estimate of β is

$$\hat{\beta}_{ML} = (X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}y = \hat{\beta}_{LS}$$

... what about the normal distribution?

- maximum likelihood estimate of β is

$$\hat{\beta}_{ML} = (X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}y = \hat{\beta}_{LS}$$

- distribution of $\hat{\beta}$ is normal

why?

$$\hat{\beta} \sim N_{p}(\beta, \sigma^2(X^{\mathrm{T}}X)^{-1})$$

- distribution of $\hat{\beta}_j$ is

$$N(\beta_j, \sigma^2(X^{\mathrm{T}}X)^{-1}_{jj}), \quad j = 1, \ldots, p$$

- maximum likelihood estimate of σ^2 is $\frac{1}{n}(y - X\hat{\beta})^{\mathrm{T}}(y - X\hat{\beta})$
- but we use

$$\tilde{\sigma}^2 = \frac{1}{n-p}(y - X\hat{\beta})^{\mathrm{T}}(y - X\hat{\beta})$$

Pause

(1) I'm lost

(2) I'm good

(3) I'm bored

Homework Week 1

on Quercus and home page

HW Question Week 1

STA2101F 2021

Due September 22 2021 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 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).

The dataset `wafer` concerns a study on semiconductors. You can get more information about the data with `?wafer`; you will first need `library(faraway);data(wafer)`, and possibly `install.packages`("faraway"). The questions below are adapted from LM Ch.3.

(a) Fit the linear model `resist ~ x1 + x2 + x3 + x4`. Extract the X matrix using the

Inference

- If you really like likelihood theory, the expected Fisher information is SM §8.2.3

$$\mathcal{I}(\beta, \sigma^2) = \begin{pmatrix} \sigma^{-2}X^{\mathrm{T}}X & 0 \\ 0 & \frac{1}{2}n\sigma^{-4} \end{pmatrix}$$

$\mathcal{I}^{-1}$ gives (asymptotic) variance of MLE

- but just using previous slide we have

$$\frac{\hat{\beta}_j - \beta_j}{\sigma[\{(X^{\mathrm{T}}X)^{-1}\}_{jj}]^{1/2}} \sim N(0, 1)$$

- and

$$\frac{\hat{\beta}_j - \beta_j}{\tilde{\sigma}[\{(X^{\mathrm{T}}X)^{-1}\}_{jj}]^{1/2}} \sim t_{n-p}$$

Example

See also `Sep152021.Rmd`

```
install.packages("faraway")
library(faraway)
data(prostate)
head(prostate)


model1 <- lm(lpsa ~ ., data = prostate)

summary(model1)
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.669337   1.296387   0.516  0.60693
lcavol       0.587022   0.087920   6.677 2.11e-09 ***
lweight      0.454467   0.170012   2.673  0.00896 **
```

Example

```
summary(model1)
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.669337   1.296387   0.516  0.60693
lcavol       0.587022   0.087920   6.677 2.11e-09 ***
lweight      0.454467   0.170012   2.673  0.00896 **
age         -0.019637   0.011173  -1.758  0.08229 .
lbph         0.107054   0.058449   1.832  0.07040 .
svi          0.766157   0.244309   3.136  0.00233 **
lcp         -0.105474   0.091013  -1.159  0.24964
gleason      0.045142   0.157465   0.287  0.77503
pgg45        0.004525   0.004421   1.024  0.30886
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

It’s Just a Linear Model

Women in Statistics and Data Science

Women in Statistics and Data Science

Follow @WomenInStat 18.7K followers

Jul 23rd 2020, 14 tweets, 4 min read

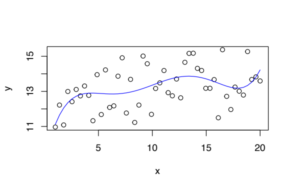
Bookmark | Save as PDF | + My Authors

Today, we’re going to play a game I’m calling “IT’S JUST A LINEAR MODEL” (IJALM).

It works like this: I name a model for a quantitative response Y, and then you guess whether or not IJALM.

- $y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \quad i = 1, \ldots, n$

1st column of X?



- $y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \beta_3 x_i^3 + \beta_4 x_i^4 + \beta_5 x_i^5 \epsilon_i$
- $y_i = \beta_0 \pm \beta_1 + \epsilon_i$
- $y_i = \beta_0 + \beta_1 \sin(x_i) + \beta_2 \cos(x_i) + \epsilon_i$
- $y_i = \gamma_0 x_{1i}^{\gamma_1} x_{2i}^{\gamma_2} \eta_i, \quad \eta_i \sim$ positive r.v.

SM Example 8.5

- $y_i = \varphi_0 + \sum_{k=1}^{K} \varphi_k s_k(x_i) + \epsilon_i$

Smoothing splines, e.g.

The linear model

- expected value $\mathbb{E}(y) =$ linear in β
- measured with additive error $y = \mathbb{E}(y) + \epsilon, \quad \epsilon \sim$
- generalizations

$$\epsilon \sim$$

$\longrightarrow$ `Sep152021.Rmd`

Today

1. Course introduction: technical issues, course details, evaluation, syllabus, people
2. Upcoming events of interest
3. Review of linear regression
4. Steps in analysis
5. In the news: excess deaths

Steps in Analysis

- understand the physical background
- understand the objective
- make sure you know what the client wants
- put the problem into statistical terms

- How were the data collected:
    - are the data observational or experimental? etc.
    - is there nonresponse
    - are there missing values
    - how are the data coded
    - what are the units of measurement
    - beware of data entry errors

Components of investigation

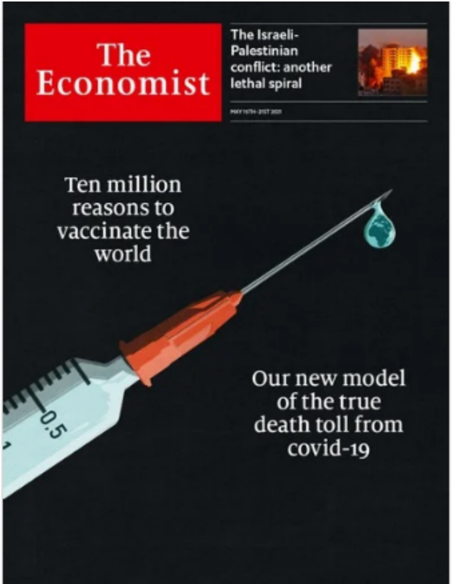
- start with a scientific question
- assess how data could shed light on this
- plan data collection
- consider of sources of variation and how careful planning can minimize their impact

- 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

In the news

Economist, May 15

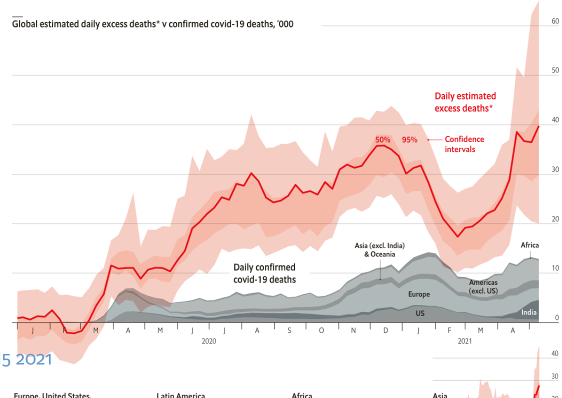


Modelling covid-19's death toll

There have been 7m-13m excess deaths worldwide during the pandemic

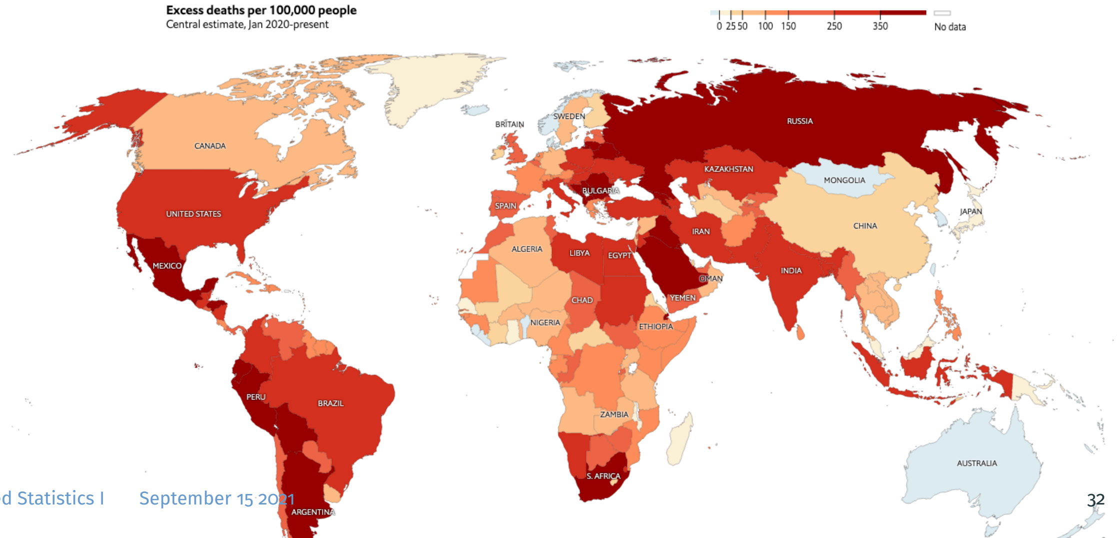
The rich world suffered relatively badly, but most of the dying has been elsewhere

MAY 15TH 2021



Economist updates

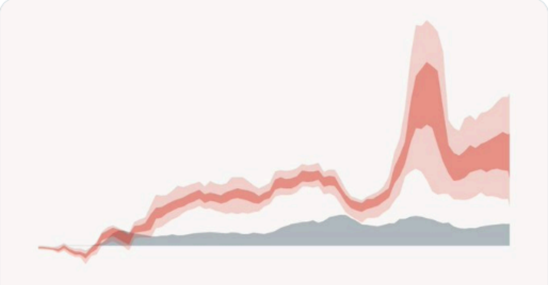
value lies between **9.5m** and **18.6m** additional deaths.



Shotwell, 2021, Sep 7

Adrian Blomfield
@adrianblomfield

Global statistical modelling done by The Economist estimates that the true number of those who died in Kenya as a result of the covid-19 pandemic is between 19,000 and 110,000, versus an official death toll of 4,746.



The pandemic's true death toll
Our daily estimate of excess deaths around the world | Graphic detail
economist.com

Why the Economist's excess death model is misleading

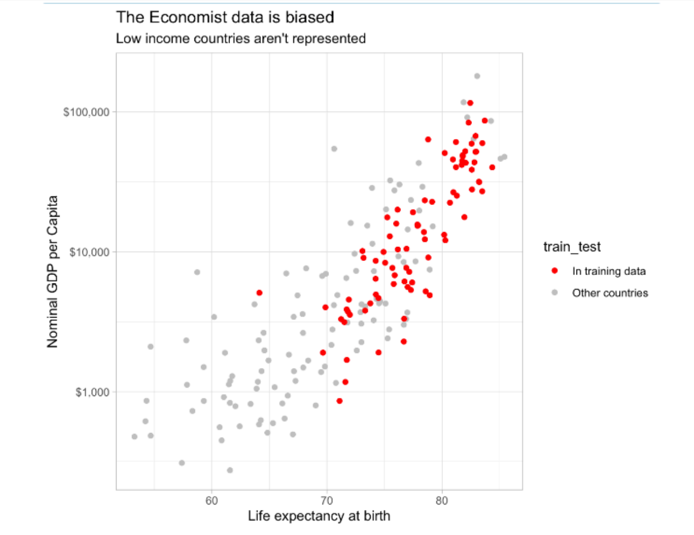
Sep 7, 2021
10 min read

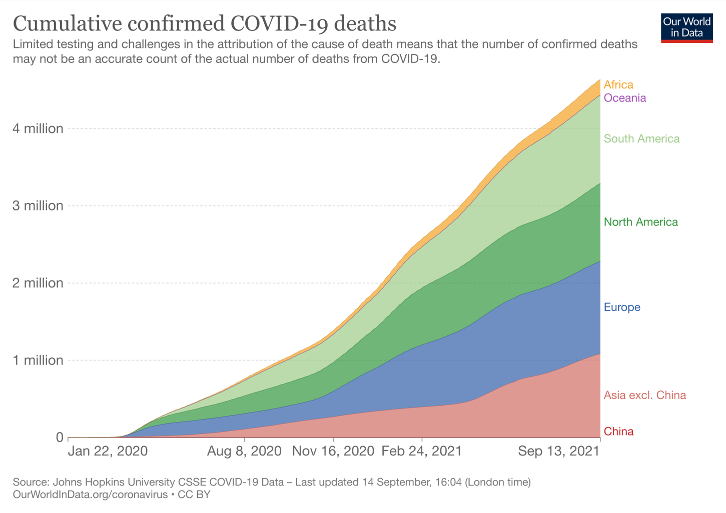
The Economist has published a model which estimates that Kenyans are only detecting 4-25% of the true deaths which can be attributed to Covid. I think this is a good opportunity to learn about why many machine learning models are problematic. I'm going to talk about this particular model, but I should note that I've only spent about ten hours looking at this problem and I'm sure the authors of this model are smart thoughtful people who don't mean to mislead. That said, I think it's an excellent example of how machine learning models can lend a sheen of credibility to things that are basically unsupported assertions. When someone says that their model says something, most people assume that means that it's supporting that thing with hard data when it's often just making unsupported assertions. It's possible that the authors of this model have sound reasons about why they can make global excess death predictions based on a small unrepresentative sample of countries, but even so I think these observations are helpful for figuring out which models you should trust.

What got me started thinking about this subject was this tweet by one of the writers at The Economist suggesting that Kenya was radically undercounting deaths which have resulted from the Covid-19 pandemic.

Adrian Blomfield
@adrianblomfield

Global statistical modelling done by The Economist estimates that the true number of those who died in Kenya


The Economist data is biased
Low income countries aren't represented
Nominal GDP per Capita
$100,000
$10,000
$1,000
60
70
80
Life expectancy at birth
train_test
In training data
Other countries


Cumulative confirmed COVID-19 deaths
Limited testing and challenges in the attribution of the cause of death means that the number of confirmed deaths may not be an accurate count of the actual number of deaths from COVID-19.
Our World in Data
4 million
3 million
2 million
1 million
0
Jan 22, 2020
Aug 8, 2020
Nov 16, 2020
Feb 24, 2021
Sep 13, 2021
Africa
Oceania
South America
North America
Europe
Asia excl. China
China
Source: Johns Hopkins University CSSE COVID-19 Data – Last updated 14 September, 16:04 (London time)
OurWorldInData.org/coronavirus • CC BY