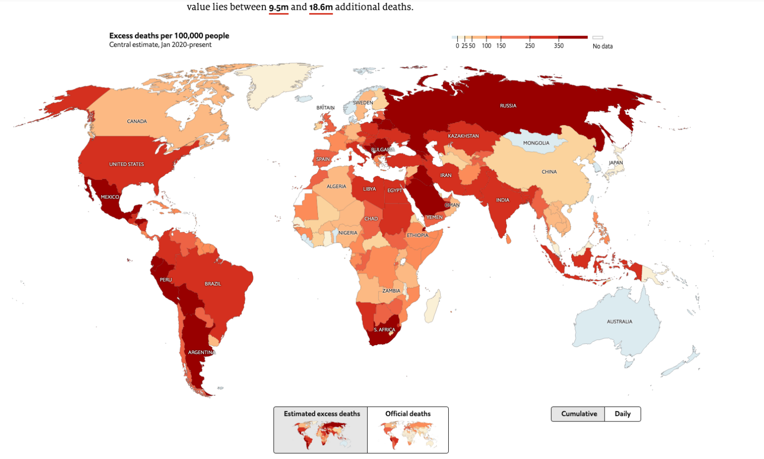
Methods of Applied Statistics I

STA2101H F LEC9101

Week 6

October 20 2021



Today

Start Recording

1. Upcoming events, Project
2. Linear Regression Part 6: randomization designs, random effects, factorial experiments
3. Logistic Regression
4. In the News

Today

1. Upcoming events, Project
2. Linear Regression Part 6: randomization designs, random effects, factorial experiments
3. Logistic Regression
4. In the News
5. Third hour – HW Comments – HW3, HW4

Syllabus – update

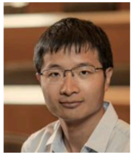
Syllabus Updated Oct 19 STA 2101F: Methods of Applied Statistics I 2021

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, steps in analysis, components of investigation, design and analysis	LM-2 Ch.1,3, Ch.14-1,2; LM-1 Ch.1,3, Ch.13; CD Ch.1
3	Sept 29	Model Comparison, diagnostics; ~~Model Selection~~, Types of Studies	LM-2 Ch.6; LM-1 Ch. 4; CD Ch.1,2

Upcoming

- Thursday Oct 21 3.30
  A top-down approach to understanding deep learning   Zoom Link

Weijie Su, University of Pennsylvania



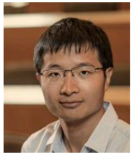
Short Bio

Weijie Su is an Assistant Professor in the Department of Statistics and Data Science of The Wharton School and the Department of Computer and Information Science, at the University of Pennsylvania. He is a co-director of Penn Research in Machine Learning. Prior to joining Penn, he received his Ph.D. in statistics from Stanford University in 2016 and his bachelor's degree in mathematics from Peking University in 2011. His research interests span privacy-preserving data analysis, optimization, high-dimensional statistics, and deep learning theory. He is a recipient of the Stanford Theodore Anderson Dissertation Award in 2016, an NSF CAREER Award in 2019, and an Alfred Sloan Research Fellowship in 2020.

Upcoming

- Thursday Oct 21 3.30
  A top-down approach to understanding deep learning   Zoom Link

Weijie Su, University of Pennsylvania



Short Bio
Weijie Su is an Assistant Professor in the Department of Statistics and Data Science of The Wharton School and the Department of Computer and Information Science, at the University of Pennsylvania. He is a co-director of Penn Research in Machine Learning. Prior to joining Penn, he received his Ph.D. in statistics from Stanford University in 2016 and his bachelor's degree in mathematics from Peking University in 2011. His research interests span privacy-preserving data analysis, optimization, high-dimensional statistics, and deep learning theory. He is a recipient of the Stanford Theodore Anderson Dissertation Award in 2016, an NSF CAREER Award in 2019, and an Alfred Sloan Research Fellowship in 2020.

- Friday Oct 22 Toronto Data Workshop   Zoom link

Toronto Data Workshop this Friday, 22 October, at noon (Toronto time) hosts Tegan Maharaj, Faculty of Information, University of Toronto.

Professor Maharaj writes:
I study AI systems and "what goes into" them, e.g. their real-world deployment context, and the effects that has on learning behaviour and generalization. I do that because I want to be able to use AI systems responsibly for problems I think are important, like impact and risk assessments for climate change, AI alignment, ecological management and other common-good problems. My website is: teganmaharaj.org.

… Upcoming

- Monday Oct 25 3.30
  Opinionated practices for teaching reproducibility: motivation, guided instruction and practice Register

Data Science ARES: Tiffany Timbers

Join us at the Data Science Applied Research and Education Seminar (ARES) with:

Dr. Tiffany Timbers
Assistant Professor of Teaching, Department of Statistics
Co-Director, Master of Data Science Program (Vancouver option)
University of British Columbia

Talk Title: Opinionated practices for teaching reproducibility: motivation, guided instruction and practice

Dataset List

1. OECD: https://stats.oecd.org/
   In addition, there is a special R package for the OECD database.

2. Ontario Government: https://data.ontario.ca/en/

3. Covid: https://www.openicpsr.org/openicpsr/search/covid19/studies
   repository for data examining the social, behavioral, public health, and economic impact of the novel coronavirus global pandemic

4. General: A great source for datasets is the Google dataset search page.

5. Climate data: NOAA Climate Data Store (CDS) contains an abundance of forecast, reanalysis, observation and climate model datasets spanning many different temporal and spatial ranges. This data can be found here.

6. Medicine: Some articles in Nature Medicine have linked datasets. A couple of such articles related to COVID19 are below:
   Immune response data
   predictors of COVID19 epidemic The latter dataset is posted on https://figshare.com/ platform that is hosting other datasets too.
7. General: You can find datasets in the UCI Machine Learning Repository: (but these are kind of tired) https://archive.ics.uci.edu/ml/datasets.php

8. Urban: Here is the link to Toronto open data portal https://open.toronto.ca/ There are many data set related to our city! For example transportation, housing, environment, etc.

9. Economics: I found a database including quarterly economic measures for a large number of indicators, for each country separately, and for the entire EU block. We can retrieve the data at EuroStat (https://ec.europa.eu/eurostat/home). The data includes

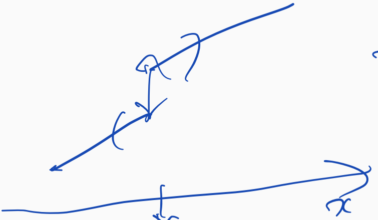
bias variance

- design of studies: systematic error, random error, estimation of uncertainty
- plan of analysis, role of individual studies
- unit of analysis; unit of interpretation

ecological bias

- interaction: between factors, between factor and continuous variables

$$\beta_0 + \beta_1 x_i + \beta_2 x_i z_i + \beta_3 z_i$$



interaction of x w group leads to different slopes

Recap: 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

418 9 · *Designed Experiments*

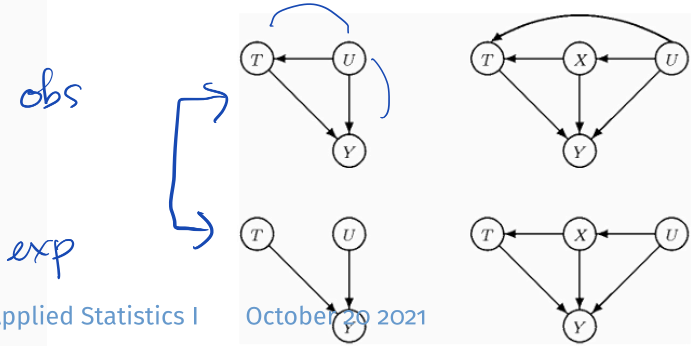


Figure 9.1 Directed acyclic graphs showing consequences of randomization. An arrow from T to Y indicates dependence of Y on T, and so forth. In general both response Y and treatment T may depend on properties U of units (upper left). Randomization (lower left) makes treatments and units independent, 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

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

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

Types of observational studies

- secondary analysis of data collected for another purpose ← weak
- 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

Stats Can (can be very good)

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

Factorial experiments: examples

- Read Ch.14 or 13 of LM – one factor variable and one continuous variable
- Example: fruitfly

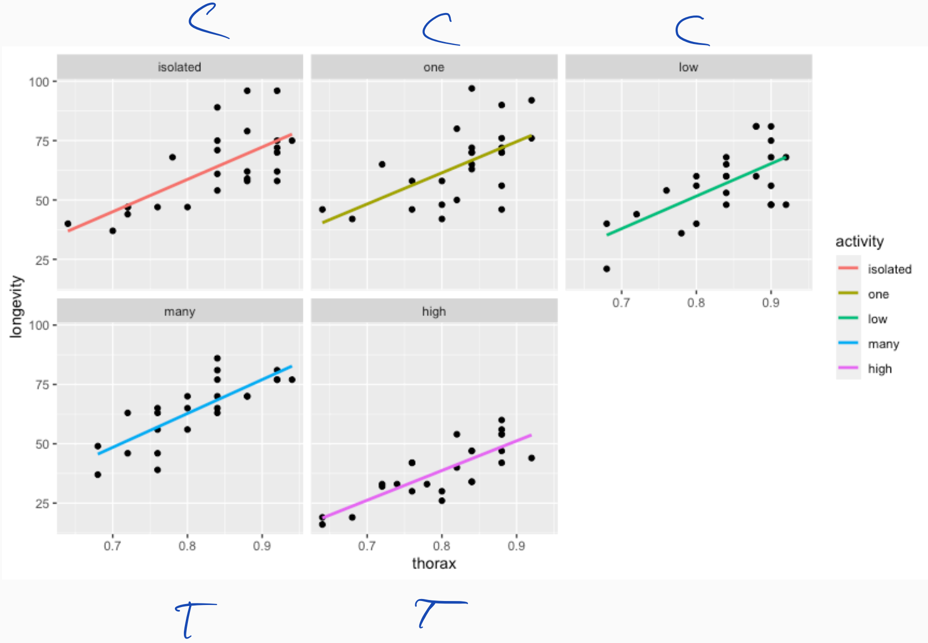


Table 8.2 Data and experimental setup for bicycle experiment (Box *et al.*, 1978, pp. 368–372). The lower part of the table shows the average times for each of the eight combinations of settings of seat height, tyre pressure, and dynamo, and the average times for the eight observations at each setting, considered separately.

Setup	Day	Run	Seat height (inches)	Dynamo	Tyre pressure (psi)	Time (secs)
1	3	2	−	−	−	51
2	4	1	−	−	−	54
3	2	2	+	−	−	41
4	2	3	+	−	−	43
5	3	3	−	+	−	54
6	2	1	−	+	−	60
7	3	1	+	+	−	44
8	4	3	+	+	−	43
9	1	1	−	−	+	50
10	4	4	−	−	+	48
11	3	5	+	−	+	39
12	4	2	+	−	+	39
13	3	4	−	+	+	53
14	1	3	−	+	+	51
15	1	2	+	+	+	41
16	2	4	+	+	+	44

2 × 2 × 2 factorial

3 factors each

… Factorial experiments: examples

Table 8.10 Poison data (Box and Cox, 1964). Survival times in 10-hour units of animals in a 3 × 4 factorial experiment with four replicates. The table underneath gives average (standard deviation) for the poison × treatment combinations.

Treatment	Poison 1	Poison 2	Poison 3
A	0.31, 0.45, 0.46, 0.43	0.36, 0.29, 0.40, 0.23	0.22, 0.21, 0.18, 0.23
B	0.82, 1.10, 0.88, 0.72	0.92, 0.61, 0.49, 1.24	0.30, 0.37, 0.38, 0.29
C	0.43, 0.45, 0.63, 0.76	0.44, 0.35, 0.31, 0.40	0.23, 0.25, 0.24, 0.22
D	0.45, 0.71, 0.66, 0.62	0.56, 1.02, 0.71, 0.38	0.30, 0.36, 0.31, 0.33

Treatment	Poison 1	Poison 2	Poison 3	Average
A	0.41 (0.07)	0.32 (0.08)	0.21 (0.02)	0.31
B	0.88 (0.16)	0.82 (0.34)	0.34 (0.05)	0.68
C	0.57 (0.16)	0.38 (0.06)	0.24 (0.01)	0.39
D	0.61 (0.11)	0.67 (0.27)	0.33 (0.03)	0.53
Average	0.62	0.55	0.28	0.48

Factor 2

3 levels

4 × 3 × 4

n = 48

4 × 3 "treatments" but in a structure.

Some classical designs

- completely randomized:
  SM Example 9.2 – one factor with 4 levels; LM-2 15.2, LM-2 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

45 students randomized to 5 groups

- completely randomized:
  SM Example 9.2 – one factor with 4 levels; LM-2 15.2, LM-2 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

- all the examples in LM-2 Ch.15, 16; LM-1 Ch. 13,14
  SM Example 9.6 (See Table 8.10) – 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

Table 9.6 Data on weight gains in pigs.

	Group								
Diet	1	2	3	4	5	6	7	8	Average
I	1.40	1.79	1.72	1.47	1.26	1.28	1.34	1.55	1.48
II	1.31	1.30	1.21	1.08	1.45	0.95	1.26	1.14	1.21
III	1.40	1.47	1.37	1.15	1.22	1.48	1.31	1.27	1.33
IV	1.96	1.77	1.62	1.76	1.88	1.50	1.60	1.49	1.70
Average	1.52	1.58	1.48	1.37	1.45	1.30	1.38	1.36	1.43

ntbc

? reduction in stochastic error ?

32 → 4 tests

- randomized blocks:
  SM Example 9.3 – one treatment factor with 4 levels, one blocking factor with 8 levels

Table 9.6 Data on weight gains in pigs.

	Group								
Diet	1	2	3	4	5	6	7	8	Average
I	1.40	1.79	1.72	1.47	1.26	1.28	1.34	1.55	1.48
II	1.31	1.30	1.21	1.08	1.45	0.95	1.26	1.14	1.21
III	1.40	1.47	1.37	1.15	1.22	1.48	1.31	1.27	1.33
IV	1.96	1.77	1.62	1.76	1.88	1.50	1.60	1.49	1.70
Average	1.52	1.58	1.48	1.37	1.45	1.30	1.38	1.36	1.43

- LM-2 17.1; LM-1 16.1

Some classical designs

- randomized blocks:
  SM Example 9.3 – one treatment factor with 4 levels, one blocking factor with 8 levels

Table 9.6 Data on weight gains in pigs.

	Group								
Diet	1	2	3	4	5	6	7	8	Average
I	1.40	1.79	1.72	1.47	1.26	1.28	1.34	1.55	1.48
II	1.31	1.30	1.21	1.08	1.45	0.95	1.26	1.14	1.21
III	1.40	1.47	1.37	1.15	1.22	1.48	1.31	1.27	1.33
IV	1.96	1.77	1.62	1.76	1.88	1.50	1.60	1.49	1.70
Average	1.52	1.58	1.48	1.37	1.45	1.30	1.38	1.36	1.43

- LM-2 17.1; LM-1 16.1
- incomplete RB — SM Example 9.4 – each block has only some treatments
- Latin square — SM Example 9.5 – two blocking factors

Analysis of variance: one-factor design

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

Analysis of variance: one-factor design

CR design

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

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

$E(y_{ij}) = \mu + \alpha_i$ $\quad$ $y_{ij} \perp$ other y's

$var(y_{ij}) = \sigma^2$

α_i change in $E(y_{ij})$ going from baseline to group i

$x_i^T \beta$

$\beta = (\mu, \alpha_1, \ldots \alpha_I, \sigma^2)$ $\quad$ $I+1$ $(+1)$ parameters

Analysis of variance: one-factor design

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

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

one constraint on $(\mu, \alpha_1, \dots, \alpha_I)$ is needed

o/w LS sol^n doesn't exist

- 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

R : $\alpha_1 = 0$ default in

another alternative $\alpha_I = 0$ $\mu = 0$ ...

$\sum_{i=1}^{I} \alpha_i = 0$ → $\hat{\mu}_{LS} = \bar{y}_{\cdot\cdot}$

$\hat{\alpha}_i = \bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot}$

$\hat{\mu}_{LS} = \bar{y}_{1\cdot}$

options(contrasts = c("contr.sum", "contr.poly"))

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

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

$E(y_{ij}) = \mu$

$\hat{\mu}$ an est. of μ $\bar{y}_{..}$ is sensible est. of μ

- 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

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

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

? $\alpha_i \equiv 0$?

do groups differ

Analysis of variance: one-factor design

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)	$IJ-1$	$\sum_{ij}(y_{ij}-\bar{y}_{..})^2$		

aov(diet ~ coag, data = ...)

$\sum_{ij}(\bar{y}_{i.}-\bar{y}_{..})^2$ ← var across grps

var within grps

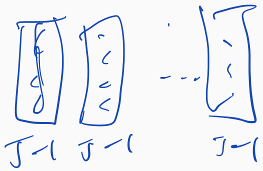
F-test $H_0: \alpha_1 = \dots = \alpha_I = 0$

Analysis of variance: one-factor design

FLM-2 Ch.15; FLM-1 Ch.14; SM 9.2.1

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}		



Analysis of variance: one-factor design

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}		

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

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

9.2 · *Some Standard Designs* 427

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

9.2 · *Some Standard Designs* 427

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

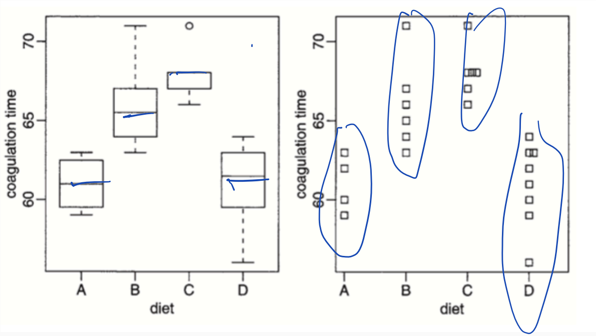
Table 9.4 Analysis of variance for data on the teaching of arithmetic.

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

bet → Groups

within → Residual

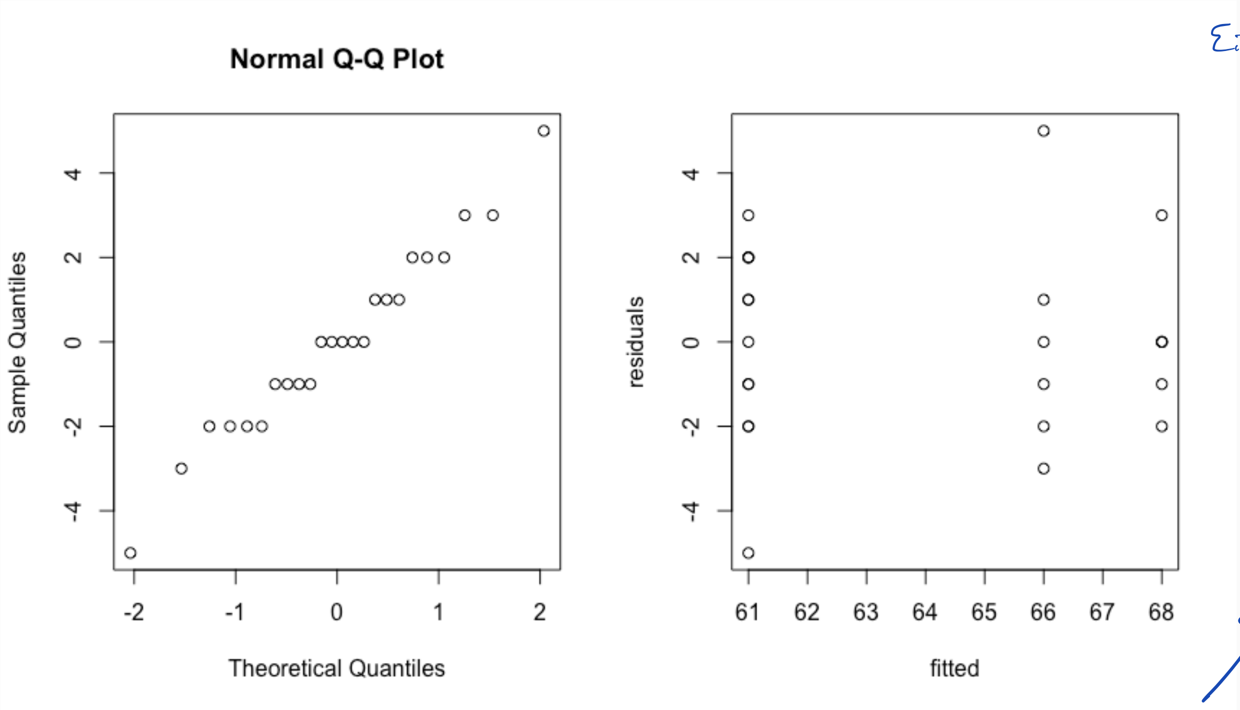
$p\{F_{4,20} \geq 15.3\}$




```
anova(lm(coag ~ diet, data = coagulation))

Response: coag
          Df Sum Sq Mean Sq F value    Pr(>F)
diet       3    228    76.0  13.571 4.658e-05 ***
Residuals 20    112     5.6
```

library(faraway)

$F_{3,20}$



$\varepsilon_{ij} \sim (0, \sigma^2)$

Can test if σ^2 same across i (Levene's test)

✗ constant variance

Reduce systematic error (bias)

- aspects of the process
  e.g. time dependence
- bias of investigators
  e.g. clinical trial

balance / randomization

Reduce random error (variance)

- compare like with like (blocking)
- use uniform material (↑)
- include background variables
- replication (↑ n)

"control what you know, randomize over the rest"

Comparison of group means

- model

  $y_{ij} = \mu + \alpha_i + \epsilon_{ij}, \quad j = 1, \ldots J_i; i = 1, \ldots I$

  group sizes unequal

- assumption $\epsilon_{ij} \sim N(0, \sigma^2)$
- $\text{var}(\bar{y}_{i.} - \bar{y}_{i'.}) =$
-

$$\frac{\bar{Y}_{i.} - \bar{Y}_{i'.}}{\tilde{\sigma}\sqrt{(1/J_i + 1/J_{i'})}} \sim$$

- 95% confidence intervals
- correction for multiple testing using HSD

```
> options(contrasts = c("contr.sum", "contr.poly"))
> summary(lm(coag~diet, data = coagulation))

Call:
lm(formula = coag ~ diet, data = coagulation)

Residuals:
   Min     1Q Median     3Q    Max
 -5.00  -1.25   0.00   1.25   5.00

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   64.000      0.498  128.54  < 2e-16 ***
diet1         -3.000      0.974   -3.08  0.00589 **
diet2          2.000      0.845    2.37  0.02819 *
diet3          4.000      0.845    4.73  0.00013 ***
```

$\bar{y}_1 \quad \bar{y}_2 \quad \bar{y}_3 \quad \bar{y}_4$

$61 \quad 66 \quad 68 \quad 61$

$$\bar{y}_{1\cdot} - \bar{y}_{2\cdot} \pm t_{20}^{\alpha/2}\, \tilde{\sigma} \cdot \left(\frac{1}{J_1} + \frac{1}{J_2}\right)^{1/2}$$

$\tilde{\sigma} = \sqrt{5.6}$

$1 - 2\alpha$

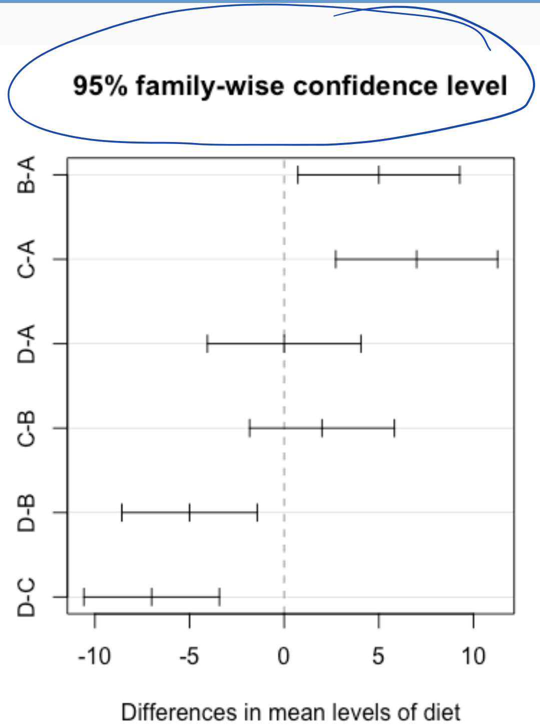
CI

for $\alpha_1 - \alpha_2$

or $(\mu + \alpha_1) - (\mu + \alpha_2)$

or

... Comparison of group means



Honest Sig diff

```
>TukeyHSD(aov(coag ~ diet, data = coagulation))
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = coag ~ diet, data = coagulation)

$diet
    diff     lwr   upr p adj
B-A    5   0.725  9.28 0.018
C-A    7   2.725 11.28 0.001
D-A    0  -4.056  4.06 1.000
C-B    2  -1.824  5.82 0.477
D-B   -5  -8.577 -1.42 0.004
D-C   -7 -10.577 -3.42 0.000

> plot(.Last.value)
```

$\bar{y}_i - \bar{y}_{i'}$

$\pm q \cdot \widehat{se}$

bigger than t dist

Multiple comparisons

- Tukey's "Honest Significant Difference" adjusts for selection

based on distribution of the largest of a set of T-statistics

Multiple comparisons

- Tukey's "Honest Significant Difference" adjusts for selection

  based on distribution of the largest of a set of T-statistics

- The Bonferroni method makes an approximate correction to the p-values:

  $p_{reported} = p_{computed} \times$ *number of comparisons*

- this controls the family-wise error rate

$P_{H_0}(\text{all } H_0 \text{ are rej.}) = 1 - P_{H_0}(\text{none } H_0 \text{ are })$

$= 1 - (1-\alpha)^k$ if they're ind't

$\approx 1 - (1 - k\alpha)$ α small $\approx \alpha k$

Multiple comparisons

- Tukey's "Honest Significant Difference" adjusts for selection

  based on distribution of the largest of a set of T-statistics

- The Bonferroni method makes an approximate correction to the p-values:
  $p_{reported} = p_{computed} \times \textit{number of comparisons}$

- this controls the family-wise error rate

- Benjamini-Hochberg controls the False Discovery Rate FDR; less conservative than Bonferroni

  genomics

- see LM-2 Ch.15.5 (posted on class web page)

STA2212S

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

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 ("trmt")					
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.

$i = 1, \ldots, 6.$

$j = 1, \ldots, 7$

$\alpha_i \sim (0, \sigma_\alpha^2)$

random effects models

...components of variance

Now

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

...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,

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

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

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

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

SS_{Betw}

MS_{Betw}

(fixed)

$$E(MS) = \sigma^2 + \frac{1}{I}\sum \alpha_i^2$$

$\bar{y}_{i\cdot} - \bar{y}_{i'\cdot}$

$E(\;) = 0$

$$var(\bar{y}_{i\cdot} - \bar{y}_{i'\cdot}) = f(\sigma_A^2, \sigma^2)$$

$$H_0 : \sigma_A^2 = 0$$

$\sigma_A^2 \neq 0$ if $p <$ small

$$\frac{MS_{bet}}{MS_{err}} \sim F$$

$$y_{ij} = \mu + \alpha_i + \varepsilon_{ij}$$

$\alpha_i \sim (0, \sigma_a^2)$

$\varepsilon_{ij} \sim (0, \sigma^2)$

Analysis of two-factor designs

Table 8.10 Poison data (Box and Cox, 1964). Survival times in 10-hour units of animals in a 3×4 factorial experiment with four replicates. The table underneath gives average (standard deviation) for the poison $\times$ treatment combinations.

Treatment	Poison 1	Poison 2	Poison 3
A	0.31, 0.45, 0.46, 0.43	0.36, 0.29, 0.40, 0.23	0.22, 0.21, 0.18, 0.23
B	0.82, 1.10, 0.88, 0.72	0.92, 0.61, 0.49, 1.24	0.30, 0.37, 0.38, 0.29
C	0.43, 0.45, 0.63, 0.76	0.44, 0.35, 0.31, 0.40	0.23, 0.25, 0.24, 0.22
D	0.45, 0.71, 0.66, 0.62	0.56, 1.02, 0.71, 0.38	0.30, 0.36, 0.31, 0.33

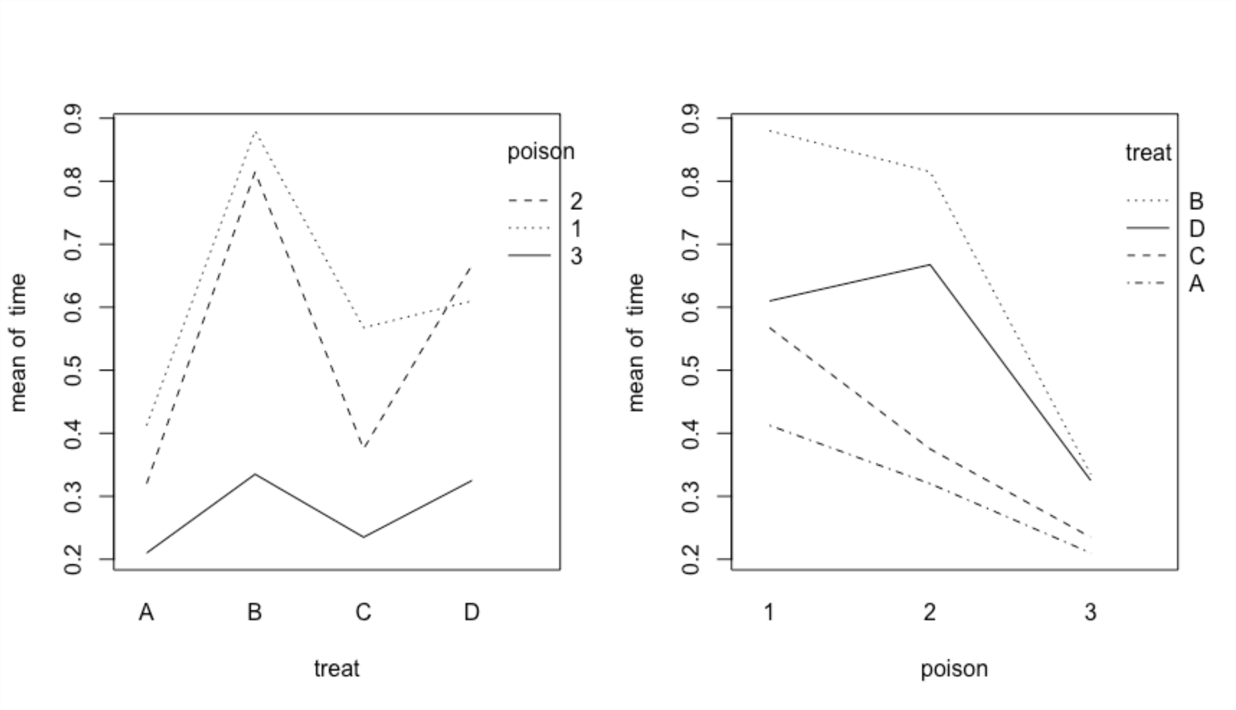
Treatment	Poison 1	Poison 2	Poison 3	Average
A	0.41 (0.07)	0.32 (0.08)	0.21 (0.02)	0.31
B	0.88 (0.16)	0.82 (0.34)	0.34 (0.05)	0.68
C	0.57 (0.16)	0.38 (0.06)	0.24 (0.01)	0.39
D	0.61 (0.11)	0.67 (0.27)	0.33 (0.03)	0.53
Average	0.62	0.55	0.28	0.48

Analysis of two-factor designs

- model: $y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \epsilon_{ijk}, \quad i = 1, \ldots, I; j = 1, \ldots J; k = 1, \ldots, R$
- analysis of variance
- comparison of means
- interaction plots

... analysis of two-factor designs

```
> library(SMPracticals}
> data(poisons)
> pmod <- lm(time ~ poison + treat, data = poisons)
> anova(pmod)
Analysis of Variance Table

Response: time
             Df Sum Sq Mean Sq F value   Pr(>F)
poison        2  1.033   0.517   23.22 3.3e-07 ***
treat         3  0.921   0.307   13.81 3.8e-06 ***
poison:treat  6  0.250   0.042    1.87    0.11
Residuals    36  0.801   0.022

> with(poisons, interaction.plot(treat,poison,time))
> with(poisons, interaction.plot(poison,treat,time))
```


mean of time
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
poison
2
1
3
A
B
C
D
treat
mean of time
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
treat
B
D
C
A
1
2
3
poison

Two-way layout without replication

Randomized block design

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

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

Table 9.5 Analysis of variance table for two-way layout model.

Term	df	Sum of squares
Treatments	$T-1$	$\sum_{t,b}(\bar{y}_{t\cdot} - \bar{y}_{\cdot\cdot})^2$
Blocks	$B-1$	$\sum_{t,b}(\bar{y}_{\cdot b} - \bar{y}_{\cdot\cdot})^2$
Residual	$(T-1)(B-1)$	$\sum_{t,b}(y_{tb} - \bar{y}_{t\cdot} - \bar{y}_{\cdot b} + \bar{y}_{\cdot\cdot})^2$

Estimation of σ^2

```
    Analysis of Variance Table

Response: yield
          Df Sum Sq Mean Sq F value    Pr(>F)    
variety    7  77524 11074.8  8.2839 1.804e-05 ***
block      4  33396  8348.9  6.2449  0.001008 ** 
Residuals 28  37433  1336.9                      
---

Residual standard error: 36.56 on 28 degrees of freedom
```

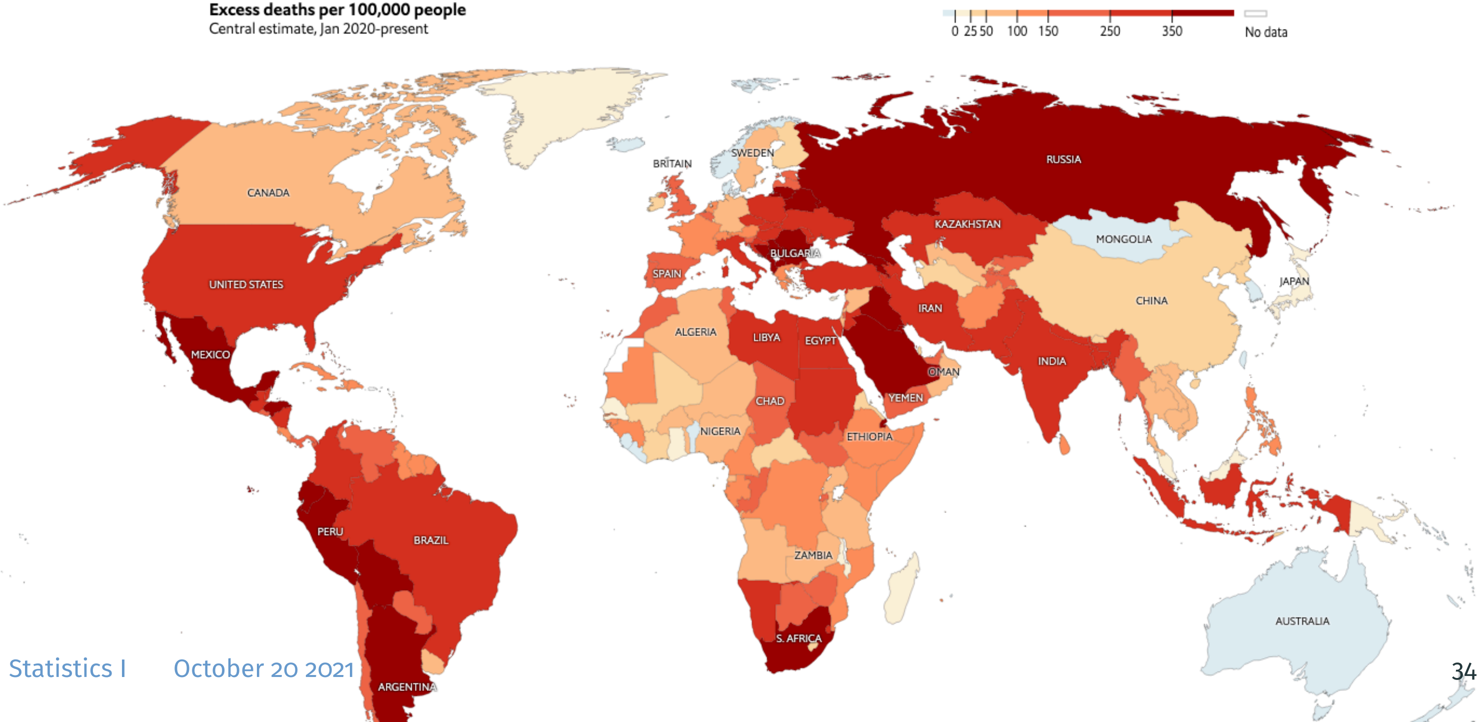
Estimation of σ^2

```
     Analysis of Variance Table

Response: yield
          Df Sum Sq Mean Sq F value    Pr(>F)
variety    7  77524 11074.8  8.2839 1.804e-05 ***
block      4  33396  8348.9  6.2449  0.001008 **
Residuals 28  37433  1336.9
---

Residual standard error: 36.56 on 28 degrees of freedom
```

The interaction between blocks and treatments is used to estimate error. This is sometimes justified by assuming the block effects β_j are random.

In the News

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

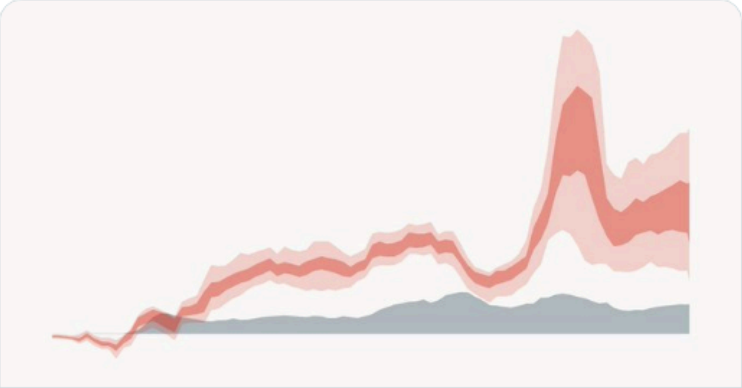


… in the news

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

9/9/2021 Why the Economist's excess death model is misleading • Gordon Shotwell

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

Africa's COVID-19 cases are seven times higher than official count, WHO says

GEOFFREY YORK › AFRICA BUREAU CHIEF
JOHANNESBURG
PUBLISHED OCTOBER 14, 2021

4 COMMENTS SHARE — A A+ TEXT SIZE BOOKMARK

TRENDING

1 EXPLAINER
Canada's COVID-19 benefits are set to expire on Oct. 23. Here's what you need to know

2 Rob Carrick: So you think you'll teach your online broker a lesson by moving your account

3 Councillor Jyoti Gondek wins mayoral race in Calgary; former Liberal cabinet minister Amarjeet Sohi wins in Edmonton

4 Rogers family, independent directors to meet Tuesday to discuss boardroom rift

WHO Africa

link

WHO Africa

link

> As WHO in Africa, we are using a model to estimate the degree of underestimation. Our analysis indicates that as few as one in seven cases is being detected, meaning that the true COVID-19 burden in Africa could be around 59 million cases.
>
> The proportion of underreporting on deaths is lower, our estimates suggest around one in three deaths are being reported. Deaths appear to be lower on the continent in part because of the predominantly younger and more active population.

- simple linear regression $E(y_i \mid x_i) = \beta_0 + \beta_1 x_i$, $\text{var}(y_i \mid x_i) = \sigma^2$
- suppose $y \in \{0, 1\}$
- examples
- $E(y_i \mid x_i) =$

$$\Pr(Y_i = 1 \mid X_i) = \frac{e^{\beta_0 + \beta_1 x_i}}{1 + e^{\beta_0 + \beta_1 x_i}}$$

$\uparrow$

$\in (0, 1)$

Stochastic $\rightarrow$ Normal $\rightarrow$ Bernoulli

Systematic parts $\beta_0 + \beta_1 x_i$

$$\rightarrow \frac{e^{\beta_0 + \beta_1 x_i}}{1 + e^{\beta_0 + \beta_1 x_i}}$$

use any CDF $: \mathbb{R} \rightarrow [0, 1]$

$p_i = \Pr(Y_i = 1 \mid x_i)$

$=$

$$\log\left(\frac{p_i}{1 - p_i}\right) = \beta_0 + \beta_1 x_i$$

log odds depends on $\uparrow$

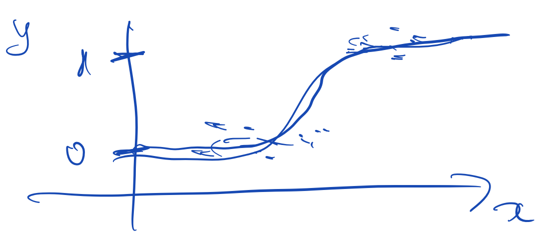
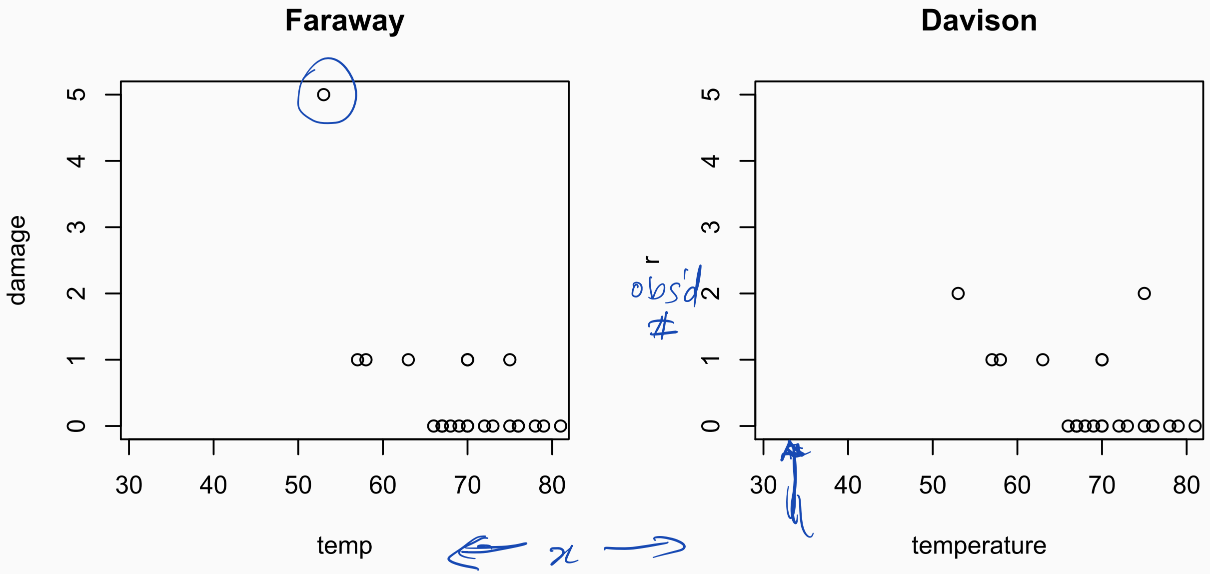


Table 1.3 O-ring thermal distress data. r is the number of field-joint O-rings showing thermal distress out of 6, for a launch at the given temperature (°F) and pressure (pounds per square inch) (Dalal *et al.*, 1989).

m = 6

Flight	Date	Number of O-rings with thermal distress, r	Temperature (°F) x_1	Pressure (psi) x_2
1	21/4/81	0	66	50
2	12/11/81	1	70	50
3	22/3/82	0	69	50
5	11/11/82	0	68	50
6	4/4/83	0	67	50
7	18/6/83	0	72	50
8	30/8/83	0	73	100
9	28/11/83	0	70	100
41-B	3/2/84	1	57	200
41-C	6/4/84	1	63	200
41-D	30/8/84	1	70	200
41-G	5/10/84	0	78	200
51-A	8/11/84	0	67	200
51-C	24/1/85	2	53	200
51-D	12/4/85	0	67	200
51-B	29/4/85	0	75	200
51-G	17/6/85	0	70	200
51-F	29/7/85	0	81	200
51-I	27/8/85	0	76	200
51-J	3/10/85	0	79	200
61-A	30/10/85	2	75	200
61-B	26/11/86	0	76	200
61-C	21/1/86	1	58	200

m = 6


Faraway
Davison
damage
r
temp
temperature
0
1
2
3
4
5
30
40
50
60
70
80


Faraway
Davison
1.0

Table 1. O-Ring Thermal-Distress Data

		Field			Nozzle				Leak-check pressure	
Flight	Date	Erosion	Blowby	Erosion or blowby	Erosion	Blowby	Erosion or blowby	Joint temperature	Field	Nozzle
1	4/12/81							66	50	50
2	11/12/81	1		1				70	50	50
3	3/22/82							69	50	50
5	11/11/82							68	50	50
6	4/04/83				2		2	67	50	50
7	6/18/83							72	50	50
8	8/30/83							73	100	50
9	11/28/83							70	100	100
41-B	2/03/84	1		1	1		1	57	200	100
41-C	4/06/84	1		1	1		1	63	200	100
41-D	8/30/84	1		1	1	1	1	70	200	100
41-G	10/05/84							78	200	100
51-A	11/08/84							67	200	100
51-C	1/24/85	2, 1*	2	2		2	2	53	200	100
51-D	4/12/85				2		2	67	200	200
51-B	4/29/85				2, 1*	1	2	75	200	100
51-G	6/17/85				2	2	2	70	200	200
51-F	7/29/85				1			81	200	200
51-I	8/27/85				1			76	200	200
51-J	10/03/85							79	200	200
61-A	10/30/85		2	2	1			75	200	200
61-B	11/26/85				2	1	2	76	200	200
61-C	1/12/86	1		1	1	1	2	58	200	200
61-I	1/28/86							31	200	200
	Total	7, 1*	4	9	17, 1*	8	17			

*Secondary O-ring.

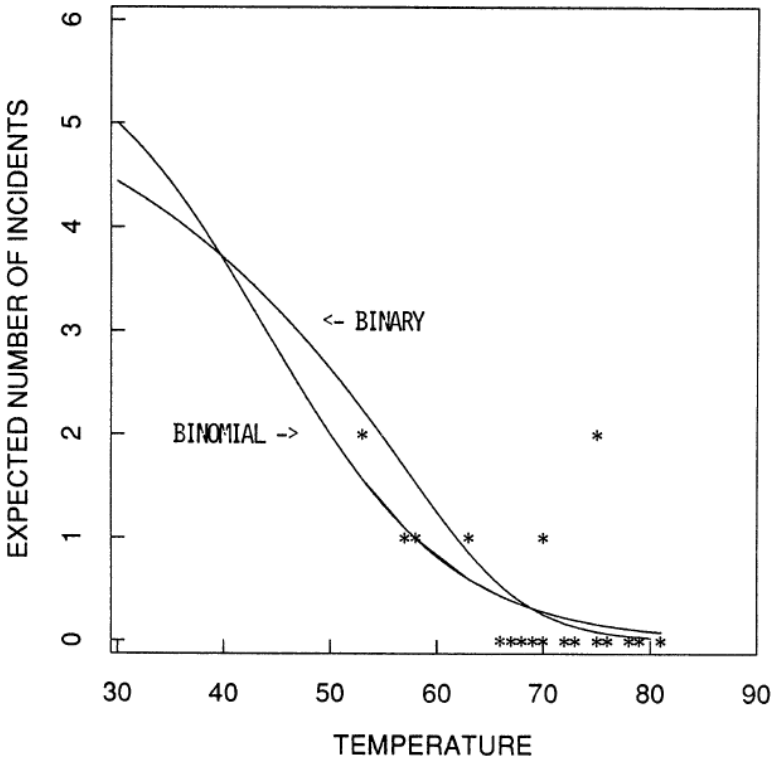


Figure 4. O-Ring Thermal-Distress Data: Field-Joint Primary O-Rings, Binomial-Logit Model, and Binary-Logit Model.

- $y_i \sim Bin(6, p_i), \quad i = 1, \ldots, 23$

$$y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \varepsilon_{ijk}$$

$$H_0: (\alpha\beta)_{ij} = 0 \qquad \text{F-test}$$

$$\bar{y}_{ij\cdot} - \bar{y}_{i\cdot\cdot}$$

$$\bar{y}_{\cdot j\cdot}$$

Modelling numbers/proportions of events

- $y_i \sim Bin(6, p_i), \quad i = 1, \ldots, 23$

- in general: n_i trials, y_i successes, probability of success p_i

Modelling numbers/proportions of events

- $y_i \sim Bin(6, p_i), \quad i = 1, \ldots, 23$
- in general: n_i trials, y_i successes, probability of success p_i
- for regression: associated covariate vector x_i, e.g. temperature

Modelling numbers/proportions of events

- $y_i \sim Bin(6, p_i), \quad i = 1, \ldots, 23$
- in general: n_i trials, y_i successes, probability of success p_i
- for regression: associated covariate vector x_i, e.g. temperature
- SM uses m_i and r_i instead of n_i and y_i

Modelling numbers/proportions of events

- $y_i \sim Bin(6, p_i), \quad i = 1, \ldots, 23$
- in general: n_i trials, y_i successes, probability of success p_i
- for regression: associated covariate vector x_i, e.g. temperature
- SM uses m_i and r_i instead of n_i and y_i
- each y_i could in principle be the sum of n_i independent Bernoulli trials

Modelling numbers/proportions of events

- $y_i \sim Bin(6, p_i), \quad i = 1, \ldots, 23$
- in general: n_i trials, y_i successes, probability of success p_i
- for regression: associated covariate vector x_i, e.g. temperature
- SM uses m_i and r_i instead of n_i and y_i
- each y_i could in principle be the sum of n_i independent Bernoulli trials
- each of the n_i trials having the same probability p_i

Modelling numbers/proportions of events

- $y_i \sim Bin(6, p_i), \quad i = 1, \ldots, 23$
- in general: n_i trials, y_i successes, probability of success p_i
- for regression: associated covariate vector x_i, e.g. temperature
- SM uses m_i and r_i instead of n_i and y_i
- each y_i could in principle be the sum of n_i independent Bernoulli trials
- each of the n_i trials having the same probability p_i
- with the same covariate vector x_i

FELM 'covariate classes', p.26

Challenger data: Faraway

```
> library(faraway); data(orings)
> logitmod <- glm(cbind(damage,6-damage) ~ temp, family = binomial, data = orings)
> summary(logitmod)
Call:
glm(formula = cbind(damage, 6 - damage) ~ temp, family = binomial,
    data = orings)
...
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) 11.66299    3.29626   3.538 0.000403 ***
temp        -0.21623    0.05318  -4.066 4.78e-05 ***
---

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 38.898  on 22  degrees of freedom
Residual deviance: 16.912  on 21  degrees of freedom
```

Challenger data: Davison

```
> library(SMPracticals) # this is for datasets in
                        #Statistical Models by Davison
> data(shuttle) # same example, different name
> shuttle2 <- data.frame(as.matrix(shuttle)) # this is a kludge to avoid
                                     #an error with head(shuttle)
> head(shuttle2)
  m r temperature pressure
1 6 0          66       50
2 6 1          70       50
3 6 0          69       50
4 6 0          68       50
5 6 0          67       50
6 6 0          72       50
> par(mfrow=c(2,2)) # puts 4 plots on a page


> with(orings,plot(temp,damage,main="Faraway",xlim=c(31,80)))
> with(shuttle,plot(temperature,r,main="Davison",xlim=c(31,80),
+ ylim=c(0,5)))
```

Challenger data fits