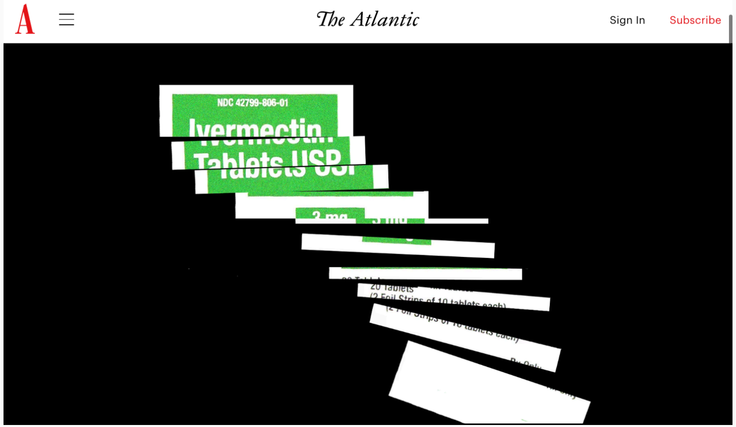
Methods of Applied Statistics I

STA2101H F LEC9101

Week 7

October 27 2021



Today

Start Recording

1. Upcoming events
2. Homework, Project
3. Linear Regression Completed: randomization designs
4. Logistic Regression
5. In the News Atlantic Oct 23 Ivermectin

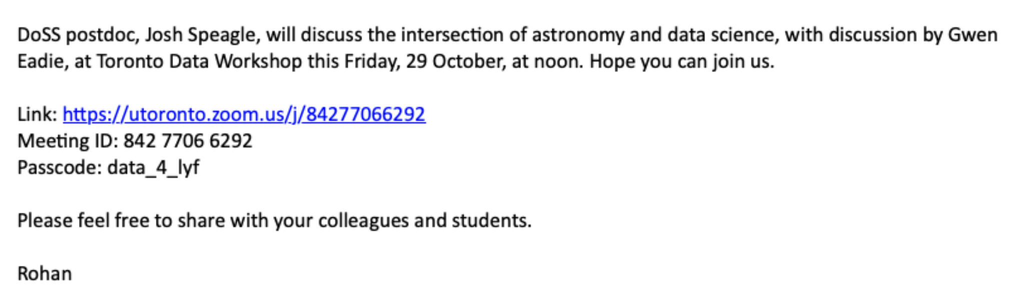
Today

Start Recording

1. Upcoming events
2. Homework, Project
3. Linear Regression Completed: randomization designs
4. Logistic Regression
5. In the News Atlantic Oct 23 Ivermectin

Upcoming

- Friday Oct 29 Toronto Data Workshop   Zoom link

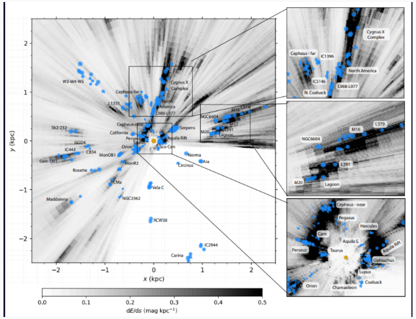


DoSS postdoc, Josh Speagle, will discuss the intersection of astronomy and data science, with discussion by Gwen Eadie, at Toronto Data Workshop this Friday, 29 October, at noon. Hope you can join us.

Link: https://utoronto.zoom.us/j/84277066292
Meeting ID: 842 7706 6292
Passcode: data_4_lyf

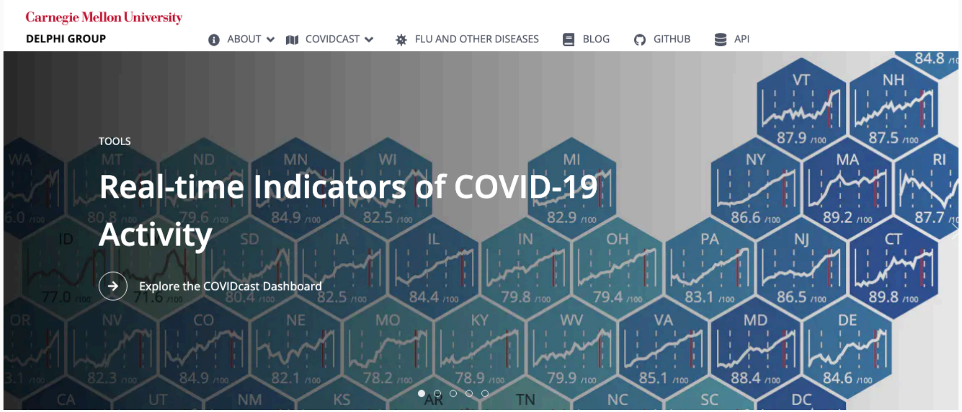
Please feel free to share with your colleagues and students.

Rohan



… Upcoming

- Monday Nov 1 15.30
  Delphi's COVIDcast Project: Lessons from Building a Digital Ecosystem for Tracking and Forecasting the Pandemic Register



Carnegie Mellon University
DELPHI GROUP
ABOUT
COVIDCAST
FLU AND OTHER DISEASES
BLOG
GITHUB
API
TOOLS
Real-time Indicators of COVID-19 Activity
Explore the COVIDcast Dashboard

Project

- Choice of dataset

unique data

Project

- Choice of dataset

unique data

- Qs for HW4/5:
  1. the data source: both bibliographic and a web link
  2. the number of observations and the number of potential explanatory variables
  3. a description of the response variable
  4. a description of the potential explanatory variables
  5. the scientific question(s) of interest
  6. unit of observation

Project

- Choice of dataset

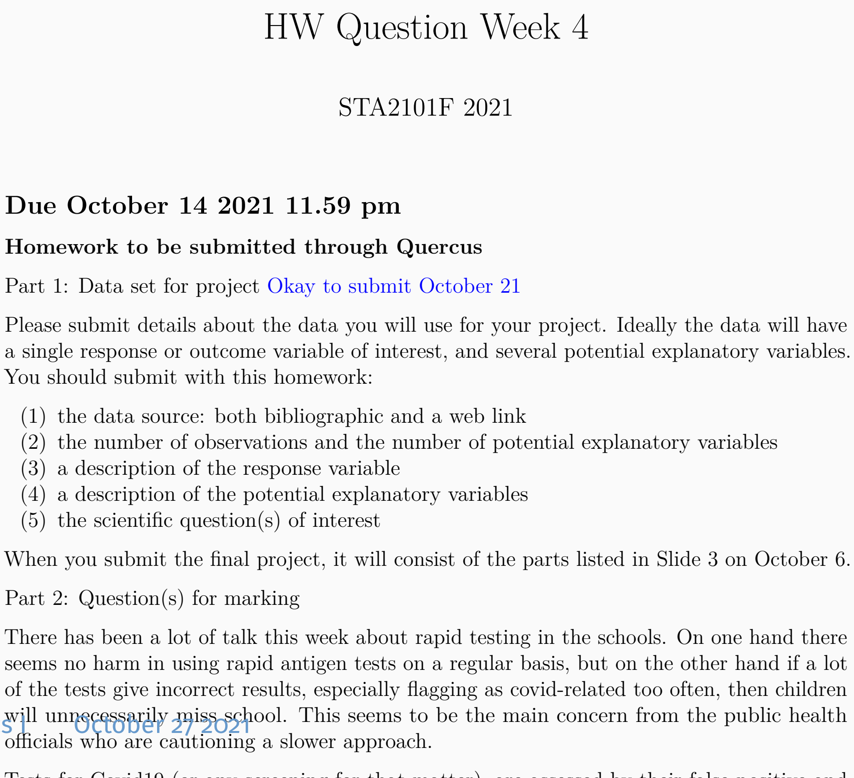
unique data

- Qs for HW4/5:
  1. the data source: both bibliographic and a web link
  2. the number of observations and the number of potential explanatory variables
  3. a description of the response variable
  4. a description of the potential explanatory variables
  5. the scientific question(s) of interest
  6. unit of observation
- Sections for Project:
  1. a description of the scientific problem of interest
  2. how (and why) the data being analyzed was collected
  3. preliminary description of the data (plots and tables)
  4. models and analysis
  5. summary for a statistician of the analysis and conclusions
  6. non-technical summary for a non-statistician of the analysis and conclusions

… Project

- Sections for Project:
  1. a description of the scientific problem of interest
  2. how (and why) the data being analyzed was collected
  3. preliminary description of the data (plots and tables)
  4. models and analysis
  5. summary for a statistician of the analysis and conclusions
  6. non-technical summary for a non-statistician of the analysis and conclusions
- Project Guidelines
  1. report: 3-5 pages: non-technical, no code – Intro, source of data, problem of interest, conclusions, a few tables, a few plots
  2. statistical appendix: main statistical methods used, summary of results, code and analysis excerpts only
  3. further plots and tables as needed
  4. R script or .Rmd file

Homework 4



HW Question Week 4

STA2101F 2021

Due October 14 2021 11.59 pm

Homework to be submitted through Quercus

Part 1: Data set for project Okay to submit October 21

Please submit details about the data you will use for your project. Ideally the data will have a single response or outcome variable of interest, and several potential explanatory variables. You should submit with this homework:

(1) the data source: both bibliographic and a web link
(2) the number of observations and the number of potential explanatory variables
(3) a description of the response variable
(4) a description of the potential explanatory variables
(5) the scientific question(s) of interest

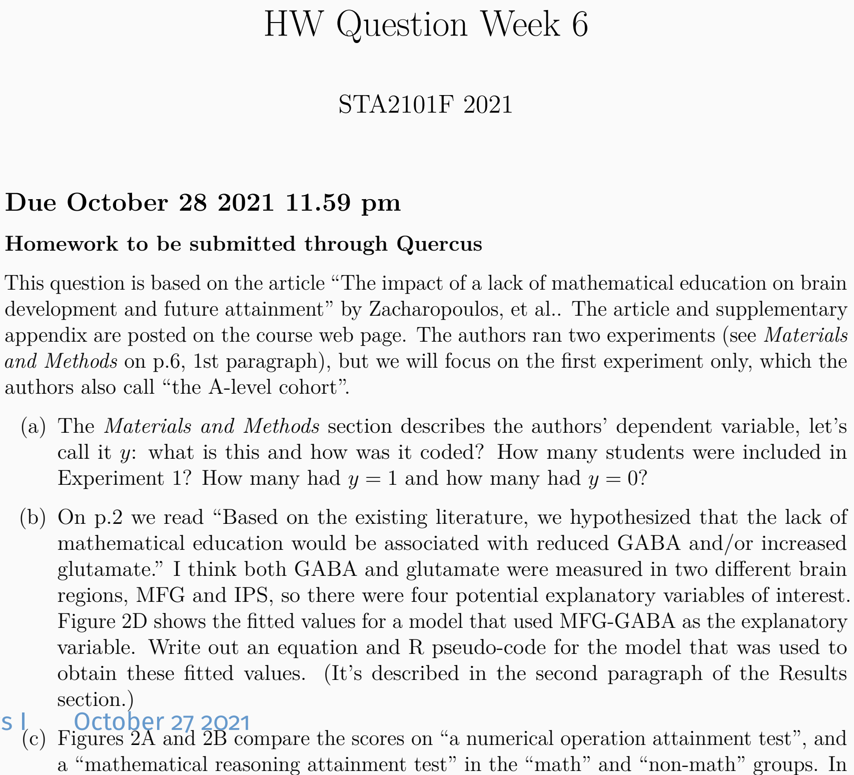
When you submit the final project, it will consist of the parts listed in Slide 3 on October 6.

Part 2: Question(s) for marking

There has been a lot of talk this week about rapid testing in the schools. On one hand there seems no harm in using rapid antigen tests on a regular basis, but on the other hand if a lot of the tests give incorrect results, especially flagging as covid-related too often, then children will unnecessarily miss school. This seems to be the main concern from the public health officials who are cautioning a slower approach.

Tests for Covid19 (or any screening for that matter), are assessed by their false positive and

Homework 6



HW Question Week 6

STA2101F 2021

Due October 28 2021 11.59 pm

Homework to be submitted through Quercus

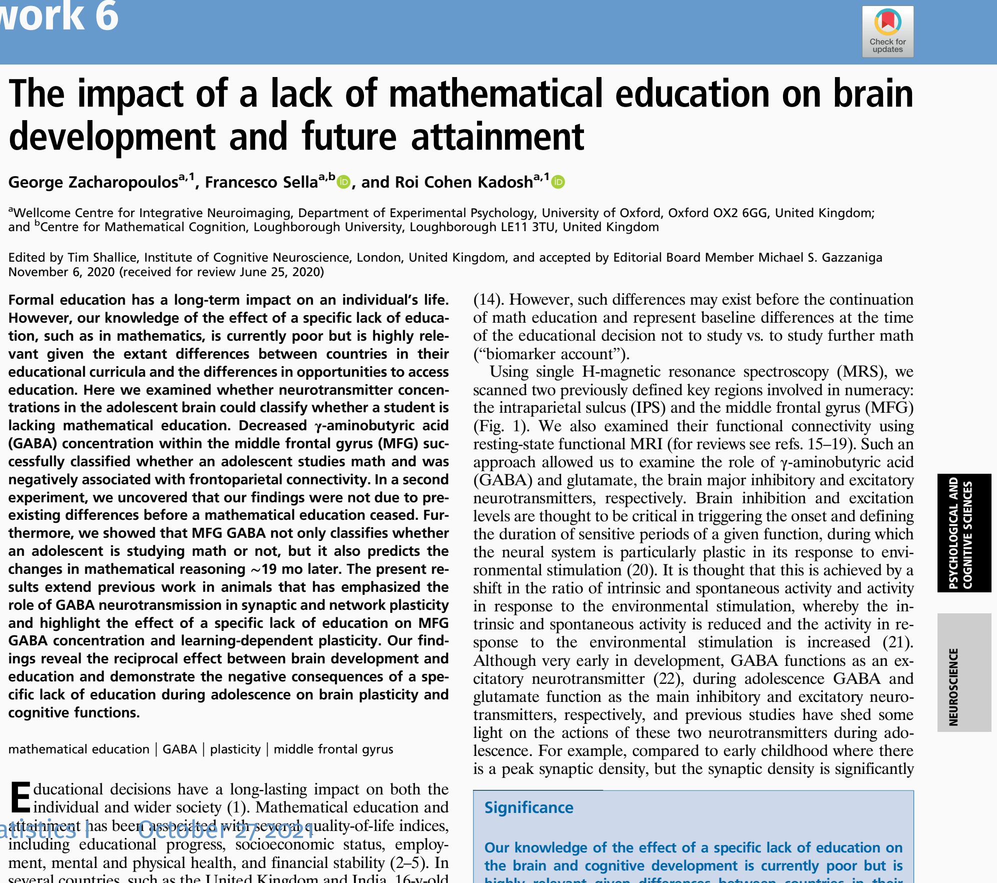
This question is based on the article “The impact of a lack of mathematical education on brain development and future attainment” by Zacharopoulos, et al.. The article and supplementary appendix are posted on the course web page. The authors ran two experiments (see *Materials and Methods* on p.6, 1st paragraph), but we will focus on the first experiment only, which the authors also call “the A-level cohort”.

(a) The *Materials and Methods* section describes the authors’ dependent variable, let’s call it y: what is this and how was it coded? How many students were included in Experiment 1? How many had $y = 1$ and how many had $y = 0$?

(b) On p.2 we read “Based on the existing literature, we hypothesized that the lack of mathematical education would be associated with reduced GABA and/or increased glutamate.” I think both GABA and glutamate were measured in two different brain regions, MFG and IPS, so there were four potential explanatory variables of interest. Figure 2D shows the fitted values for a model that used MFG-GABA as the explanatory variable. Write out an equation and R pseudo-code for the model that was used to obtain these fitted values. (It’s described in the second paragraph of the Results section.)

(c) Figures 2A and 2B compare the scores on “a numerical operation attainment test”, and a “mathematical reasoning attainment test” in the “math” and “non-math” groups. In

Homework 6


Check for updates

The impact of a lack of mathematical education on brain development and future attainment

George Zacharopoulos[a,1], Francesco Sella[a,b], and Roi Cohen Kadosh[a,1]

[a]Wellcome Centre for Integrative Neuroimaging, Department of Experimental Psychology, University of Oxford, Oxford OX2 6GG, United Kingdom; and [b]Centre for Mathematical Cognition, Loughborough University, Loughborough LE11 3TU, United Kingdom

Edited by Tim Shallice, Institute of Cognitive Neuroscience, London, United Kingdom, and accepted by Editorial Board Member Michael S. Gazzaniga November 6, 2020 (received for review June 25, 2020)

Formal education has a long-term impact on an individual's life. However, our knowledge of the effect of a specific lack of education, such as in mathematics, is currently poor but is highly relevant given the extant differences between countries in their educational curricula and the differences in opportunities to access education. Here we examined whether neurotransmitter concentrations in the adolescent brain could classify whether a student is lacking mathematical education. Decreased γ-aminobutyric acid (GABA) concentration within the middle frontal gyrus (MFG) successfully classified whether an adolescent studies math and was negatively associated with frontoparietal connectivity. In a second experiment, we uncovered that our findings were not due to preexisting differences before a mathematical education ceased. Furthermore, we showed that MFG GABA not only classifies whether an adolescent is studying math or not, but it also predicts the changes in mathematical reasoning ~19 mo later. The present results extend previous work in animals that has emphasized the role of GABA neurotransmission in synaptic and network plasticity and highlight the effect of a specific lack of education on MFG GABA concentration and learning-dependent plasticity. Our findings reveal the reciprocal effect between brain development and education and demonstrate the negative consequences of a specific lack of education during adolescence on brain plasticity and cognitive functions.

mathematical education | GABA | plasticity | middle frontal gyrus

Educational decisions have a long-lasting impact on both the individual and wider society (1). Mathematical education and attainment has been associated with several quality-of-life indices, including educational progress, socioeconomic status, employment, mental and physical health, and financial stability (2–5). In several countries, such as the United Kingdom and India, 16-y-old

(14). However, such differences may exist before the continuation of math education and represent baseline differences at the time of the educational decision not to study vs. to study further math ("biomarker account").

Using single H-magnetic resonance spectroscopy (MRS), we scanned two previously defined key regions involved in numeracy: the intraparietal sulcus (IPS) and the middle frontal gyrus (MFG) (Fig. 1). We also examined their functional connectivity using resting-state functional MRI (for reviews see refs. 15–19). Such an approach allowed us to examine the role of γ-aminobutyric acid (GABA) and glutamate, the brain major inhibitory and excitatory neurotransmitters, respectively. Brain inhibition and excitation levels are thought to be critical in triggering the onset and defining the duration of sensitive periods of a given function, during which the neural system is particularly plastic in its response to environmental stimulation (20). It is thought that this is achieved by a shift in the ratio of intrinsic and spontaneous activity and activity in response to the environmental stimulation, whereby the intrinsic and spontaneous activity is reduced and the activity in response to the environmental stimulation is increased (21). Although very early in development, GABA functions as an excitatory neurotransmitter (22), during adolescence GABA and glutamate function as the main inhibitory and excitatory neurotransmitters, respectively, and previous studies have shed some light on the actions of these two neurotransmitters during adolescence. For example, compared to early childhood where there is a peak synaptic density, but the synaptic density is significantly

Significance

Our knowledge of the effect of a specific lack of education on the brain and cognitive development is currently poor but is highly relevant given differences between countries in their

PSYCHOLOGICAL AND COGNITIVE SCIENCES

NEUROSCIENCE

Recap: Design of studies

- types of observational studies: ‘found data’, survey, study, census, meta-analysis
- classical designs: completely randomized, randomized block

incomplete block, Latin square

- describes how units are assigned to treatments

Recap: Design of studies

- types of observational studies: ‘found data’, survey, study, census, meta-analysis
- classical designs: completely randomized, randomized block

  incomplete block, Latin square

- describes how units are assigned to treatments
- treatments may have a factorial structure
- regardless of the design

Recap: Design of studies

- types of observational studies: ‘found data’, survey, study, census, meta-analysis
- classical designs: completely randomized, randomized block

incomplete block, Latin square

- describes how units are assigned to treatments
- treatments may have a factorial structure
- regardless of the design
- analysis of variance partitions total sum of squares according to the treatment structure

and the blocking structure, if any

- $y_{ij} = \mu + \alpha_i + \epsilon_{ij}, \qquad j = 1, \ldots T; i = 1, \ldots, R$

α_i fixed or random

- comparison of group means $\bar{y}_{i.}$, or ←
- analysis of σ^2_α ←

α: fixed

or random

Analysis of two-factor designs

LM-2 Ch16; LM-1 Ch15; SM Ch9.2

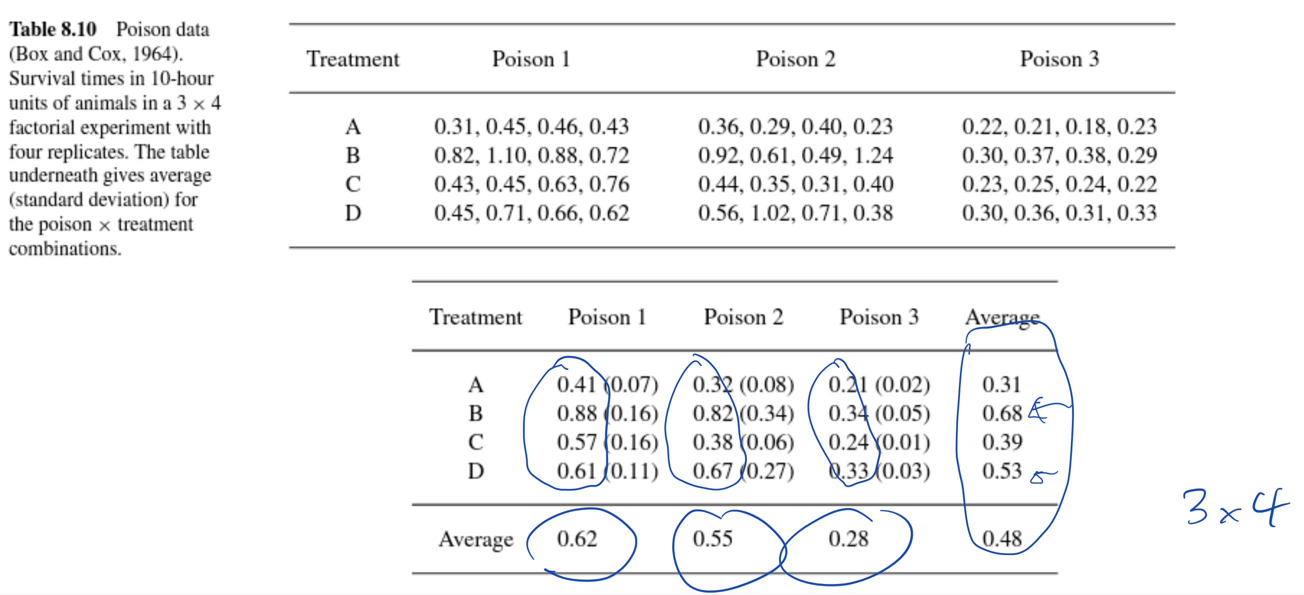


Table 8.10 Poison data (Box and Cox, 1964). Survival times in 10-hour units of animals in a 3 × 4 factorial experiment with four replicates. The table underneath gives average (standard deviation) for the poison × treatment combinations.

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

3 x 4 factorial

CR

4 obs per 'cell'

Factorial treatment structure

- model $y_{ijk} = \mu + \alpha_i + \beta_j + \boxed{(\alpha\beta)_{ij}} + \epsilon_{ijk}, \quad i = 1, \dots, I; j = 1, \dots J; k = 1, \dots, R$ (*)

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

- analysis of variance

$$\underbrace{\sum_{ijk}(y_{ijk}-\bar{y}_{\dots})^2}_{TSS} = \underbrace{\sum_{ijk}(\bar{y}_{i\cdot\cdot}-\bar{y}_{\dots})^2}_{SS_A}+\underbrace{\sum_{ijk}(\bar{y}_{\cdot j\cdot}-\bar{y}_{\dots})^2}_{SS_B}+\underbrace{\sum_{ijk}(\bar{y}_{ij\cdot}-\bar{y}_{i\cdot\cdot}-\bar{y}_{\cdot j\cdot}+\bar{y}_{\dots})^2}_{SS_{AB}}+\underbrace{\sum_{ijk}(y_{ijk}-\bar{y}_{ij\cdot})^2}_{SS_{res.}}$$

est. σ^2

under *

$$\underbrace{\bar{y}_{ij\cdot} - \bar{y}_{i\cdot}} - \underbrace{(\bar{y}_{\cdot j} - \bar{y}_{\cdot\cdot})}$$

Factorial treatment structure

LM-2 Ch16; LM-1 Ch15; SM Ch9.2

- model $\quad y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \epsilon_{ijk}, \quad i = 1, \ldots, I; j = 1, \ldots J; k = 1, \ldots, R$

- analysis of variance

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

- comparison of means

- interaction plots

... factorial treatment structure

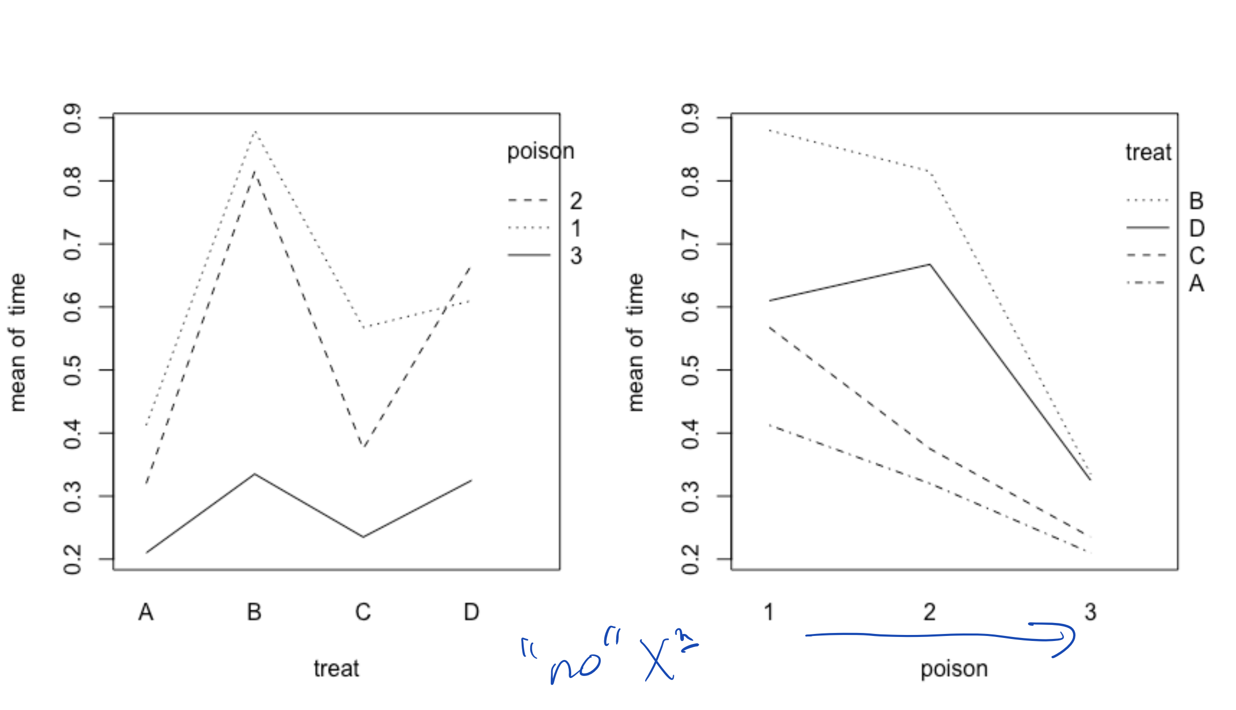
LM-2 Ch16; LM-1 Ch15; SM Ch9.2

```
> 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))
```

X: $A + B + AB$

$A + B$ no X^2

... factorial treatment structure

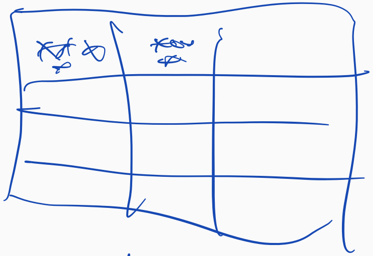


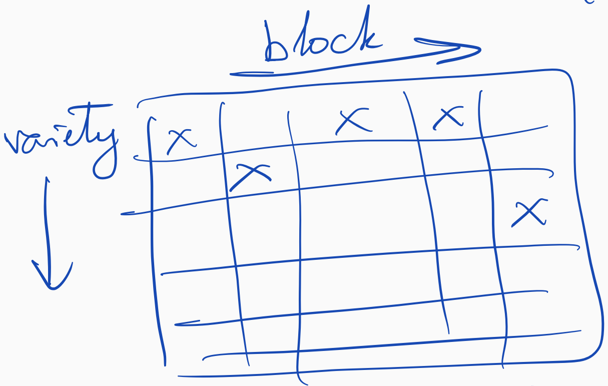
"no" X^2

One observation per cell

LM-2 16.1; LM-1 15.1

```
> data(oatvar, package = "faraway")
> xtabs(yield ~ variety + block, data = oatvar)
##          block ← not of interest
## variety    I  II III  IV   V    mean
##        1 296 357 340 331 348    334.4
##        2 402 390 431 340 320    376.6
##        3 437 334 426 320 296    362.6
##        4 303 319 310 260 242    286.8
##        5 469 405 442 487 394    439.4
##        6 345 342 358 300 308    330.6
##        7 324 339 357 352 220    318.4
##        8 488 374 401 338 320    384.2
```





⟶ Oct27.Rmd

$\bar{y}_{i\cdot} - \bar{y}_{i'\cdot} \sim N(\quad,\quad)$

Randomized block design

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

$$= \underbrace{\sum_{ij}(y_{ij}-\bar{y}_{i.}-\bar{y}_{.j}+\bar{y}_{..})^2}_{X^n \quad SS_{AB}} + \underbrace{\sum_{ij}(\bar{y}_{i.}-\bar{y}_{..})^2}_{SS_A} + \underbrace{\sum_{ij}(\bar{y}_{.j}-\bar{y}_{..})^2}_{SS_B}$$

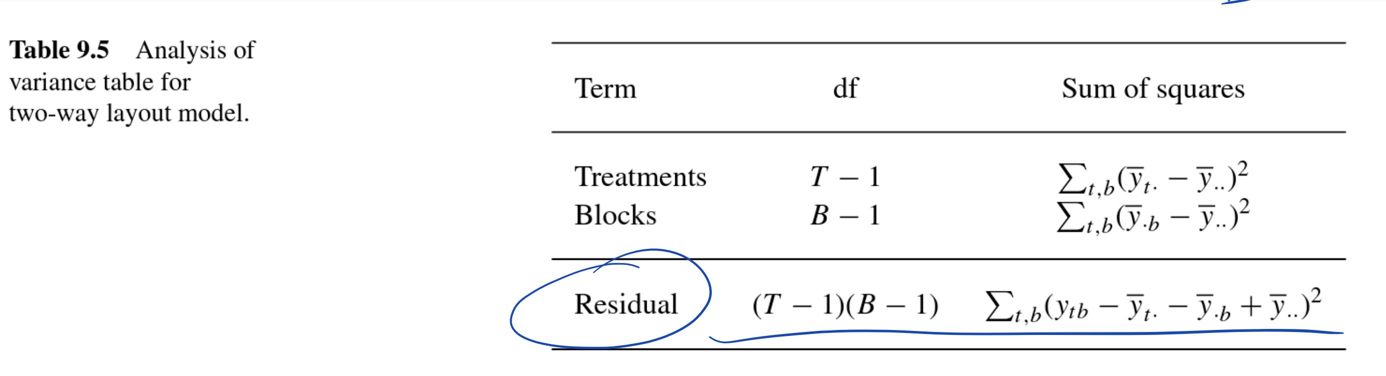


Table 9.5 Analysis of variance table for two-way layout model.

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

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
```

$$\hat{\sigma}^2 = \frac{1336.9}{28} = (36.56)^2$$

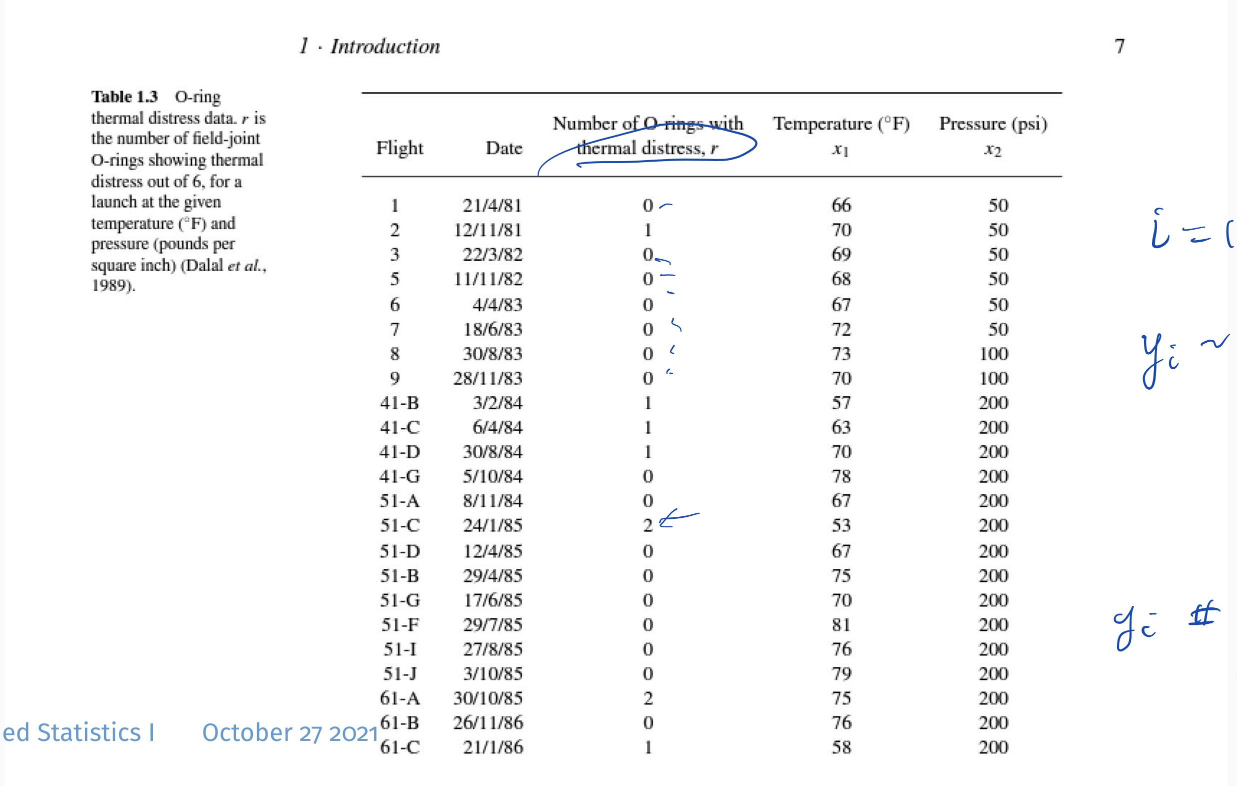
Estimation of σ^2

```
     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
```

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

Binomial Data

ELM-1, Ch. 2, SM, Ch.1; §4.4,5;



1 · Introduction 7

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).

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

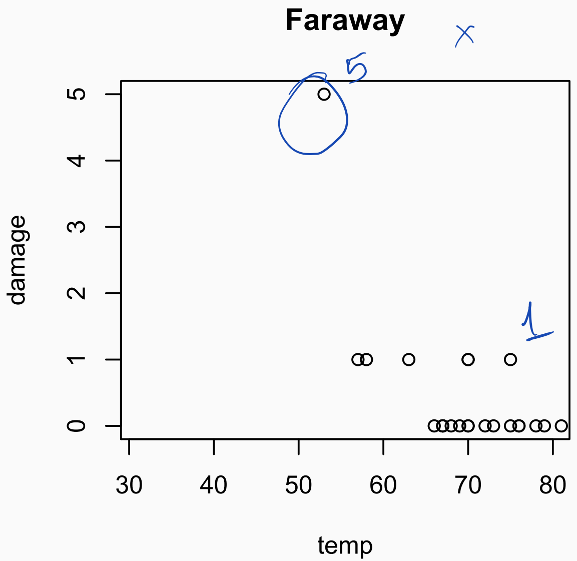
$i = 1, \dots 32$

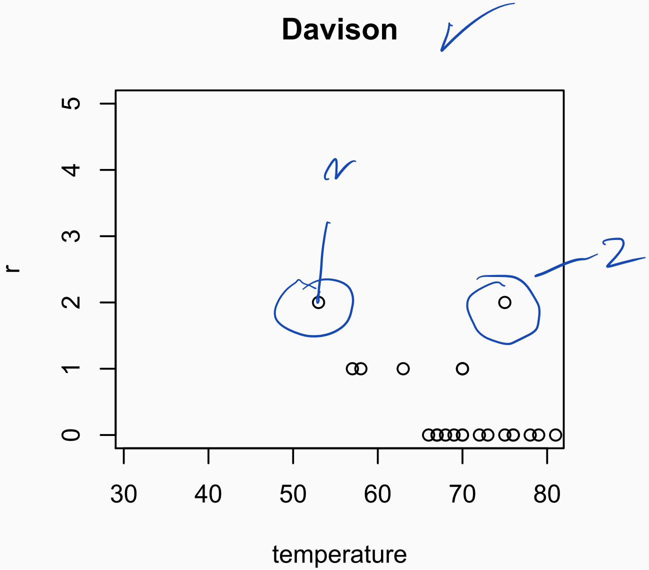
$y_i \sim Bin(6, p_i)$

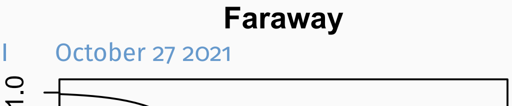
6 o-rings

y_i # O-rings damaged

… Binomial Data







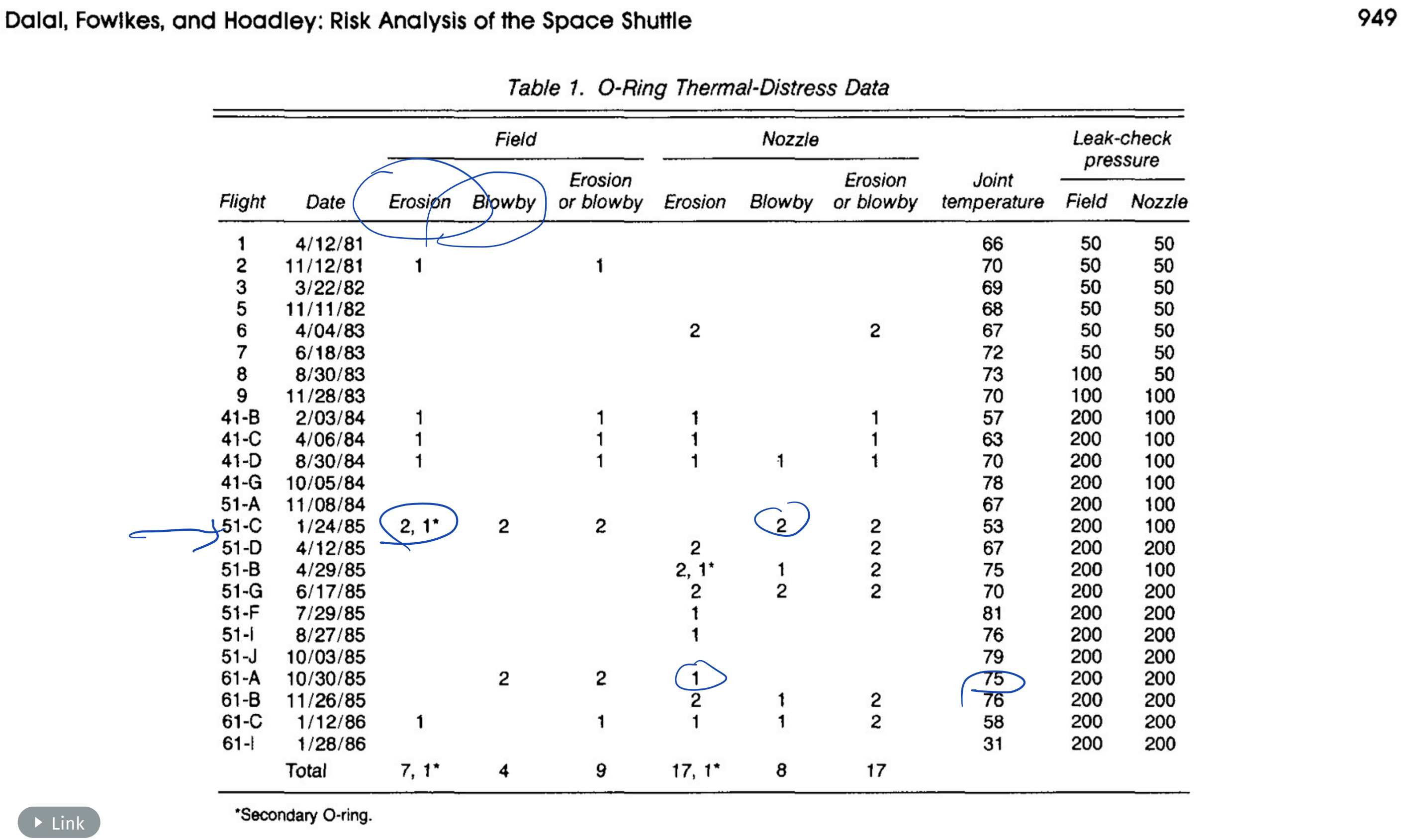


Table 1. O-Ring Thermal-Distress Data

		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.

▸ Link

Dalal et al (1989) *Journal of the American Statistical Association*

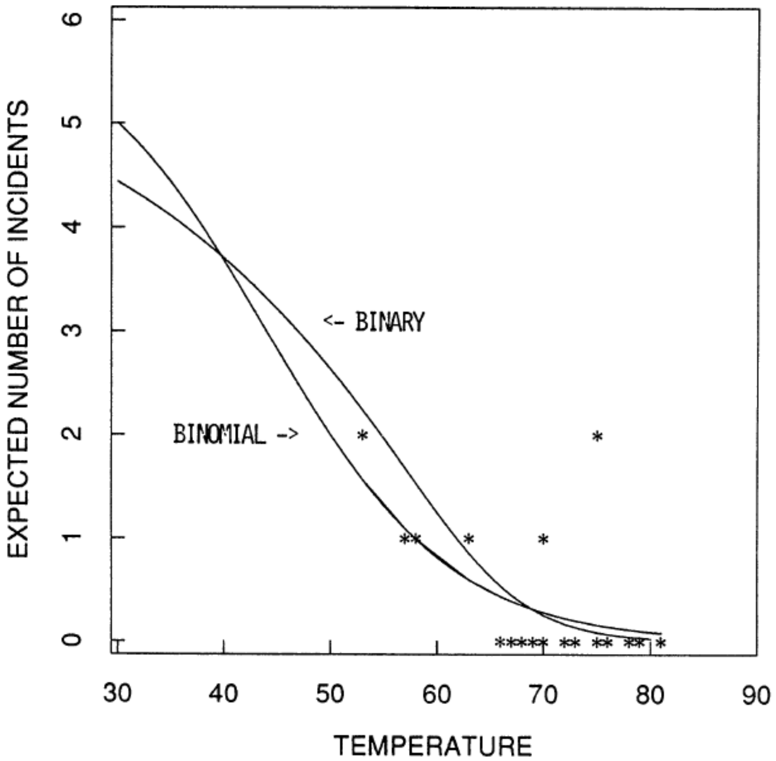


Figure 4. O-Ring Thermal-Distress Data: Field-Joint Primary O-Rings, Binomial-Logit Model, and Binary-Logit Model.

Modelling numbers/proportions of events

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

p_i dep. on temp

$p_i = \beta_0 + \underline{\underline{\beta_1}} x_i$ $\qquad x_i$ = temper.

$\beta_1 < 0$

$$\Big\downarrow$$

$$p_i(\beta) = \frac{e^{\beta_0 + \beta_1 x_i}}{1 + e^{\beta_0 + \beta_1 x_i}} \quad \in (0, 1)$$

$p_i \uparrow$ in x_i if β_1 +ve , $\downarrow$ in x_i if β_1 -ve

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

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

- in general: $\underline{\underline{n_i}}$ trials, $\underline{\underline{y_i}}$ successes, probability of success $\underline{\underline{p_i}}$

- 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

- $y_i \sim Bin(6, p_i), \quad i = 1, \dots, 23$
- 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 $\underline{\underline{y_i}}$ could in principle be the sum of n_i independent Bernoulli trials

$i = 1, \dots, 32$

$i = 1, \dots, 192$

$$\left.\begin{array}{ll} x_1 & 1/0 \\ x_1 & 1/0 \\ x_1 & \vdots \\ x_i & \\ x_1 & 1/0 \\ x_1 & \end{array}\right\} \Sigma$$

Modelling numbers/proportions of events

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

- 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

$$p_i = \frac{e^{\beta_0 + \beta_1 x_i}}{1 + e^{(\dots)}}$$

means x_i same

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

ELM-1 ‘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
```

$y \sim$

x

not t

approx. lik. th.

y_i

$n_i - y_i$

cbind(d, 6-d)

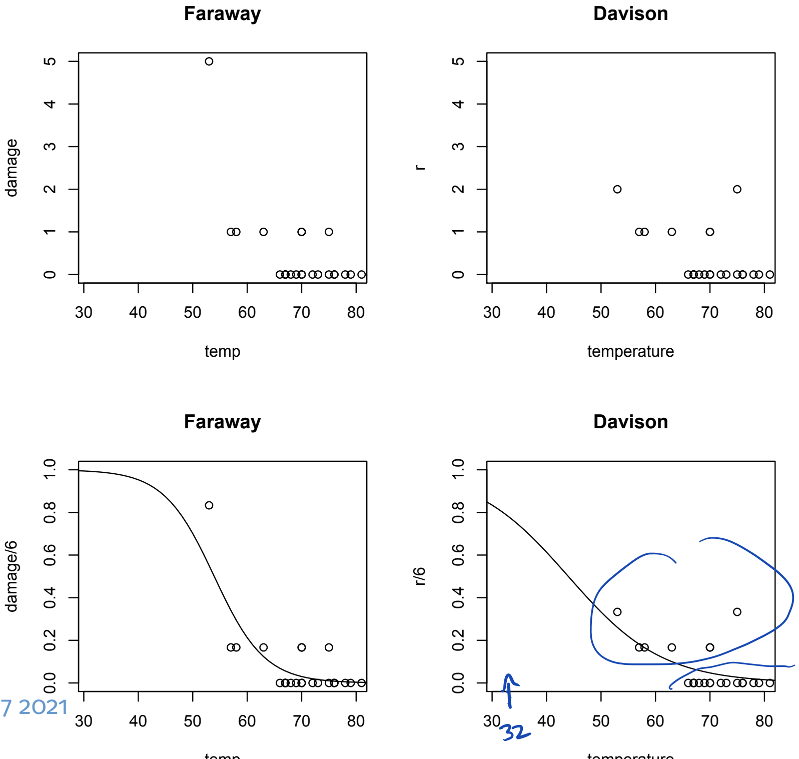
"success"

"failure"

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



Regression modelling with binomial

- model:

$$y_i \sim Bin(n_i, p_i)$$

$$n_i = 6, i = 1, \dots, n$$

Regression modelling with binomial

- model:

$$y_i \sim Bin(n_i, p_i)$$

$n_i = 6, i = 1, \ldots, n$

- regression: link the p_i's through x_i

Regression modelling with binomial

- model:

$$y_i \sim Bin(n_i, p_i)$$

$n_i = 6, i = 1, \ldots, n$

- regression: link the p_i's through x_i
- for example,

$$p_i = \frac{\exp(\beta_0 + x_{i1}\beta_1 + \cdots + x_{iq}\beta_q)}{1 + \exp(\beta_0 + x_{i1}\beta_1 + \cdots + x_{iq}\beta_q))}$$

Regression modelling with binomial

- model:

$$y_i \sim Bin(n_i, p_i)$$

$n_i = 6, i = 1, \dots, n$

- regression: link the p_i's through x_i
- for example,

$$p_i = \frac{\exp(\beta_0 + x_{i1}\beta_1 + \cdots + x_{iq}\beta_q)}{1 + \exp(\beta_0 + x_{i1}\beta_1 + \cdots + x_{iq}\beta_q))}$$

- more concisely

$$p_i = \frac{\exp(x_i^{\mathrm{T}}\beta)}{1 + \exp(x_i^{\mathrm{T}}\beta)}$$

Regression modelling with binomial

$$\boxed{y_i = x_i^T \beta + \varepsilon_i}$$

$$y_i = \mu_i + \varepsilon_i$$

- model:

$$y_i \sim Bin(n_i, p_i)$$

$n_i = 6, i = 1, \ldots, n$

- regression: link the p_i's through x_i
- for example,

$$p_i = \frac{\exp(\beta_0 + x_{i1}\beta_1 + \cdots + x_{iq}\beta_q)}{1 + \exp(\beta_0 + x_{i1}\beta_1 + \cdots + x_{iq}\beta_q)}$$

- more concisely

$$p_i = \frac{\exp(x_i^{\mathrm{T}}\beta)}{1 + \exp(x_i^{\mathrm{T}}\beta)}$$

$$p_i = \underset{\substack{\uparrow \\ \text{cdf}}}{F}(x_i^T \beta)$$

- $x_i^{\mathrm{T}} = (1, x_{i1}, \ldots, x_{iq})$; $\quad \beta = (\beta_0, \beta_1, \ldots, \beta_q)^{\mathrm{T}}$

all vectors are column vectors

… regression modelling with binomial

- Probability of event:

$$p_i = \frac{\exp(x_i^{\mathrm{T}}\beta)}{1 + \exp(x_i^{\mathrm{T}}\beta)}$$

… regression modelling with binomial

- Probability of event:

$$p_i = \frac{\exp(x_i^{\mathrm{T}}\beta)}{1 + \exp(x_i^{\mathrm{T}}\beta)}$$

- Linear on the logit scale:

$$\log \frac{p_i}{1 - p_i} = x_i^{\mathrm{T}}\beta$$

… regression modelling with binomial

- Probability of event:

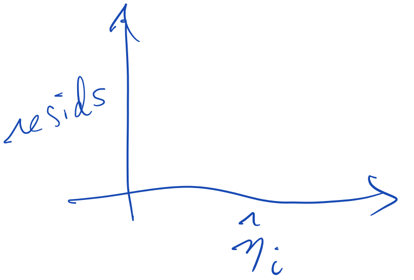
$$p_i = \frac{\exp(x_i^{\mathrm{T}}\beta)}{1+\exp(x_i^{\mathrm{T}}\beta)}$$

- Linear on the logit scale:

$$\log\frac{p_i}{1-p_i} = x_i^{\mathrm{T}}\beta$$

- linear predictor:

$$x_i^{\mathrm{T}}\beta = \eta_i$$



this scale for diagnostics

… regression modelling with binomial

- Probability of event:

$$p_i = \frac{\exp(x_i^{\mathrm{T}}\beta)}{1+\exp(x_i^{\mathrm{T}}\beta)}$$

- Linear on the logit scale:

$$\log \frac{p_i}{1-p_i} = x_i^{\mathrm{T}}\beta$$

- linear predictor:

$$x_i^{\mathrm{T}}\beta = \eta_i$$

- p_i is always between 0 and 1

… regression modelling with binomial

- Probability of event:

$$p_i = \frac{\exp(x_i^{\mathrm{T}}\beta)}{1 + \exp(x_i^{\mathrm{T}}\beta)}$$

- Linear on the logit scale:

$$\log \frac{p_i}{1 - p_i} = x_i^{\mathrm{T}}\beta$$

- linear predictor:

$$x_i^{\mathrm{T}}\beta = \eta_i$$

- p_i is always between 0 and 1
- see ELM-1 §2.1 for a linear fit

… regression modelling with binomial

- Probability of event:

$$p_i = \frac{\exp(x_i^{\mathrm{T}}\beta)}{1+\exp(x_i^{\mathrm{T}}\beta)}$$

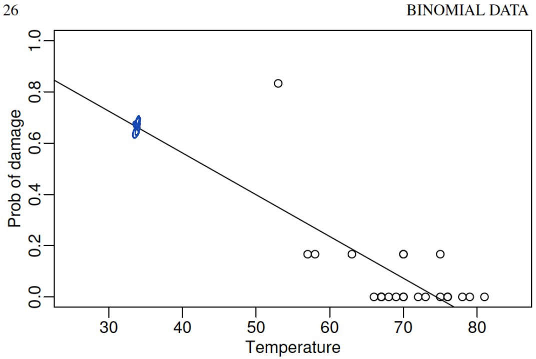
- Linear on the logit scale:

$$\log \frac{p_i}{1-p_i} = x_i^{\mathrm{T}}\beta$$

- linear predictor:

$$x_i^{\mathrm{T}}\beta = \eta_i$$

- p_i is always between 0 and 1
- see ELM-1 §2.1 for a linear fit



… regression modelling with binomial

```
> summary(logitmodcorrect)

Call:
glm(formula = cbind(r, m - r) ~ temperature, family = binomial,  data = shuttle2)


Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  5.08498    3.05247   1.666   0.0957 .
temperature -0.11560    0.04702  -2.458   0.0140 *
```

SM data

$\hat{\beta}_0 = 5$ $\hat{\beta}_1 = -0.116$

Estimation

- $\ell(\beta; y) = \sum_{i=1}^{n} [y_i(\beta_0 + \beta_1 x_i) - n_i \log\{1 + \exp(\beta_0 + \beta_1 x_i)\}]$ +

$$\ell(\beta; y) = \log \prod_{i=1}^{n} f(y_i \mid x_i ; \beta) \leftarrow \text{log Likelihood}$$

jt density

$$\hat{\beta} = \arg\max_{\beta} \prod f(\ )$$

vector $\beta_1, \dots, \beta_2$

$$= \sum \log f(y_i \mid x_i, \beta)$$

$$y_i \sim \text{Bin}\binom{n_i}{p_i} \qquad f(y_i) = p_i(\beta)^{y_i}(1 - p_i(\beta))^{n_i - y_i}\binom{n_i}{y_i}$$

$$\log f = y_i \log p_i + (n_i - y_i)\log(1 - p_i)$$

Estimation

- $\ell(\beta; y) = \sum_{i=1}^{n} [y_i(\beta_0 + \beta_1 x_i) - n_i \log\{1 + \exp(\beta_0 + \beta_1 x_i)\}]$

$\log f(y_i)$

- maximum likelihood estimate $\hat{\beta}_0$, $\hat{\beta}_1$

$\partial \ell(\beta; y)/\partial \beta = 0$

$$p_i = \frac{e^{x_i^T \beta}}{1 + e^{x_i^T \beta}} \qquad 1 - p_i = \frac{1}{(\quad)}$$

$$= y_i \log\left(\frac{p_i}{1-p_i}\right) + n_i \log(1 - p_i)$$

$$y_i \sim Bin(\overset{6}{n_i}, p_i)$$

$$\log\left(\frac{p_i}{1-p_i}\right) = x_i^T \beta$$

$$= \sum y_i(x_i^T \beta) + n_i \log\{1 - p_i(\beta)\}$$

$$i = 1, \dots, \underset{32}{n}$$

general case

Estimation

- $\ell(\beta; y) = \sum_{i=1}^{n} [y_i(\beta_0 + \beta_1 x_i) - n_i \log\{1 + \exp(\beta_0 + \beta_1 x_i)\}]$
- maximum likelihood estimate $\hat\beta_0, \hat\beta_1$
-

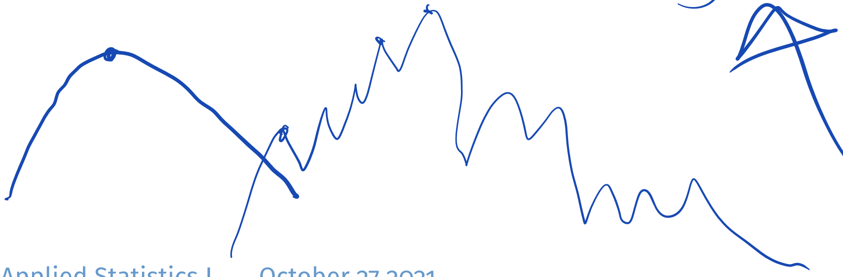
$\left.\frac{\partial \ell}{\partial \beta_0}\right|_{\hat\beta_0 \hat\beta_1} = 0 \qquad \left.\frac{\partial \ell}{\partial \beta_1}\right|_{\hat\beta_0, \hat\beta_1} = 0$

$\underline{\underline{\partial \ell(\beta; y)/\partial \beta = 0}}$

$$\underline{\hat\beta_0 = 5.08498}, \quad \hat\beta_1 = \underline{-0.11560} \qquad j(\beta) \equiv -\frac{\partial^2 \ell(\beta)}{\partial \beta \partial \beta^{\mathrm{T}}}$$

IRWLS iteratively re-weighted LS ← max lik.

algorithm in glm's can be solved using



Estimation

$$\hat\beta \xrightarrow{d} N(\beta, i^{-1}(\beta))$$

- $\ell(\beta; y) = \sum_{i=1}^{n} [y_i(\beta_0 + \beta_1 x_i) - n_i \log\{1 + \exp(\beta_0 + \beta_1 x_i)\}]$
- maximum likelihood estimate $\hat\beta_0$, $\hat\beta_1$
- $\partial \ell(\beta; y)/\partial\beta = 0$

$$i(\beta) = E\{-\ell''(\beta)\}$$

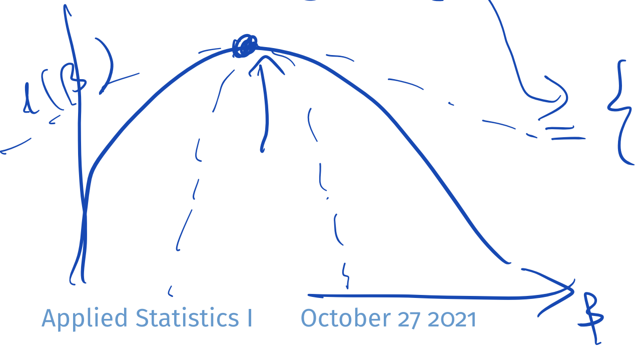
$$\underline{\hat\beta_0 = 5.08498, \quad \hat\beta_1 = -0.11560} \qquad j(\beta) \equiv -\frac{\partial^2 \ell(\beta)}{\partial\beta\partial\beta^{\mathrm{T}}}$$

- $\widehat{\operatorname{var}}(\hat\beta) \doteq j^{-1}(\hat\beta)$

$$\ell'(\hat\beta) = 0$$

$$\left.\frac{\partial \ell}{\partial \beta_0}\right|_{\hat\beta_0, \hat\beta_1} = 0$$

$$\left.\frac{\partial \ell}{\partial \beta_1}\right|_{\hat\beta_0, \hat\beta_1} = 0$$



$$\doteq \{-\ell''(\hat\beta)\}^{-1}$$

$$\approx \operatorname{var}(\hat\beta)$$

Estimation

- $\ell(\beta; y) = \sum_{i=1}^{n} [y_i(\beta_0 + \beta_1 x_i) - n_i \log\{1 + \exp(\beta_0 + \beta_1 x_i)\}]$

- maximum likelihood estimate $\hat{\beta}_0$, $\hat{\beta}_1$

$\partial\ell(\beta; y)/\partial\beta = 0$

-

$$\hat{\beta}_0 = 5.08498, \quad \hat{\beta}_1 = -0.11560 \qquad j(\beta) \equiv -\frac{\partial^2 \ell(\beta)}{\partial\beta\partial\beta^{\mathrm{T}}}$$

- $\mathrm{var}(\hat{\beta}) \doteq j^{-1}(\hat{\beta})$

SM dataset

```
> vcov(logitmodcorrect)
            (Intercept)  temperature
(Intercept)   9.3175983 -0.142564339
temperature  -0.1425643  0.002211221
```

$$\begin{bmatrix} \mathrm{var}\,\hat{\beta}_0 & \mathrm{cov}(\hat{\beta}_0, \hat{\beta}_1) \\ & \mathrm{var}(\hat{\beta}_1) \end{bmatrix}$$

$\frac{\hat{\beta}}{\hat{se}} \dot{\sim} N(0,1) \leftarrow$ p value

$\hat{se}(\hat{\beta}_1) = \sqrt{.0022}$

Interpretation of estimated coefficients

```
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  5.08498    3.05247   1.666   0.0957 .
temperature -0.11560    0.04702  -2.458   0.0140 *
```

LS

$x_1 \uparrow$ by 1 unit

$y \uparrow$ by $\hat{\beta}_1$ units

"all other..."

"a unit increase in temperature is associated with an increase in log-odds of O-ring damage of -0.116"

"an increase in the odds of $\exp(-0.116) = 0.89$"

so actually a decrease

" an increase in the probability of ??

depends on the baseline probability

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

$$\log(\text{odds}) = \frac{P(\text{succ})}{P(\text{fail})}$$

Nested models

- Comparing two models:

Nested models

- Comparing two models:
- likelihood ratio test

$$2\{\ell_A(\hat{\beta}_A) - \ell_B(\hat{\beta}_B)\}$$

compares the maximized log-likelihood function under model A and model B

Nested models

- Comparing two models:
- likelihood ratio test

  $$2\{\ell_A(\hat{\beta}_A) - \ell_B(\hat{\beta}_B)\}$$

  compares the maximized log-likelihood function under model A and model B
- example

  model A: $\mathrm{logit}(p_i) = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i}, \quad \beta_A = (\beta_0, \beta_1, \beta_2)$

  model B: $\mathrm{logit}(p_i) = \beta_0 + \beta_1 x_{1i}, \quad \beta_B = (\beta_0, \beta_1)$

Nested models

- Comparing two models:
- likelihood ratio test

$$2\{\ell_A(\hat\beta_A) - \ell_B(\hat\beta_B)\}$$

  compares the maximized log-likelihood function under model A and model B
- example
  model A: $\mathrm{logit}(p_i) = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i}, \quad \beta_A = (\beta_0, \beta_1, \beta_2)$
  model B: $\mathrm{logit}(p_i) = \beta_0 + \beta_1 x_{1i}, \quad \beta_B = (\beta_0, \beta_1)$

- when model B is nested in model A, LRT is approximately χ^2_ν distributed, under model B

Nested models

- Comparing two models:
- likelihood ratio test

  $$2\{\ell_A(\hat{\beta}_A) - \ell_B(\hat{\beta}_B)\}$$

  compares the maximized log-likelihood function under model A and model B
- example
  model A: $\mathrm{logit}(p_i) = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i}, \quad \beta_A = (\beta_0, \beta_1, \beta_2)$
  model B: $\mathrm{logit}(p_i) = \beta_0 + \beta_1 x_{1i}, \quad \beta_B = (\beta_0, \beta_1)$

- when model B is nested in model A, LRT is approximately χ^2_ν distributed, under model B
- $\nu = dim(A) - dim(B)$

… nested models

```
> logitmodcorrect2 <- glm(cbind(r,m-r) ˜ temperature + pressure, family = binomial, data = shuttle2)

> summary(logitmodcorrect2)

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  2.520195   3.486784   0.723   0.4698
temperature -0.098297   0.044890  -2.190   0.0285 *
pressure     0.008484   0.007677   1.105   0.2691
---
    Null deviance: 24.230  on 22  degrees of freedom
Residual deviance: 16.546  on 20  degrees of freedom
AIC: 36.106
Number of Fisher Scoring iterations: 5
```

… nested models

```
> logitmodcorrect2 <- glm(cbind(r,m-r) ~ temperature + pressure, family = binomial, data = shuttle2)

> anova(logitmodcorrect,logitmodcorrect2)
Analysis of Deviance Table

Model 1: cbind(r, m - r) ~ temperature
Model 2: cbind(r, m - r) ~ temperature + pressure
  Resid. Df Resid. Dev Df Deviance
1        21     18.086
2        20     16.546  1   1.5407
```

...nested models

- Model A: $\mathrm{logit}(p_i) = \beta_0 + \beta_1 \texttt{temp}_i + \beta_2 \texttt{pressure}_i$

…nested models

- Model A: $\operatorname{logit}(p_i) = \beta_0 + \beta_1 \texttt{temp}_i + \beta_2 \texttt{pressure}_i$
- Model B: $\operatorname{logit}(p_i) = \beta_0 + \beta_1 \texttt{temp}_i$

...nested models

- Model A: $\mathrm{logit}(p_i) = \beta_0 + \beta_1 \texttt{temp}_i + \beta_2 \texttt{pressure}_i$
- Model B: $\mathrm{logit}(p_i) = \beta_0 + \beta_1 \texttt{temp}_i$
- nested: Model B is obtained by setting $\beta_2 = 0$

…nested models

- Model A: $\mathrm{logit}(p_i) = \beta_0 + \beta_1 \mathtt{temp}_i + \beta_2 \mathtt{pressure}_i$

- Model B: $\mathrm{logit}(p_i) = \beta_0 + \beta_1 \mathtt{temp}_i$

- nested: Model B is obtained by setting $\beta_2 = 0$

- Under Model B, the change in deviance is (approximately) an observation from a χ^2_1

...nested models

- Model A: $\operatorname{logit}(p_i) = \beta_0 + \beta_1 \texttt{temp}_i + \beta_2 \texttt{pressure}_i$

- Model B: $\operatorname{logit}(p_i) = \beta_0 + \beta_1 \texttt{temp}_i$

- nested: Model B is obtained by setting $\beta_2 = 0$

- Under Model B, the change in deviance is (approximately) an observation from a χ^2_1
- $\Pr(\chi^2_1 \geq 1.5407) = 0.22$
  this is a p-value for testing $H_0 : \beta_2 = 0$

...nested models

- Model A: $\operatorname{logit}(p_i) = \beta_0 + \beta_1 \texttt{temp}_i + \beta_2 \texttt{pressure}_i$

- Model B: $\operatorname{logit}(p_i) = \beta_0 + \beta_1 \texttt{temp}_i$

- nested: Model B is obtained by setting $\beta_2 = 0$

- Under Model B, the change in deviance is (approximately) an observation from a χ^2_1
- $\Pr(\chi^2_1 \geq 1.5407) = 0.22$
  this is a p-value for testing $H_0 : \beta_2 = 0$

- so is $1 - \Phi\{\frac{\hat{\beta}_2}{\widehat{s.e.}(\hat{\beta}_2)}\} = 1 - \Phi(1.105) = 0.27$

ELM-1 p.30

Inference

- confidence intervals for β_1

Inference

- confidence intervals for β_1

- based on normal approximation: $\hat{\beta}_1 \pm \widehat{\text{s.e.}}(\hat{\beta}_1) * 1.96$

Inference

- confidence intervals for β_1

- based on normal approximation: $\hat{\beta}_1 \pm \widehat{\text{s.e.}}(\hat{\beta}_1) * 1.96$
- `(-0.208, -0.023)`

Inference

- confidence intervals for β_1

- based on normal approximation: $\hat{\beta}_1 \pm \widehat{\text{s.e.}}(\hat{\beta}_1) * 1.96$
- `(-0.208, -0.023)`

- based on profile log-likelihood

$\ell_{\text{p}}(\beta_1)$, details to follow

Inference

- confidence intervals for β_1

- based on normal approximation: $\hat{\beta}_1 \pm \widehat{\text{s.e.}}(\hat{\beta}_1) * 1.96$
- (-0.208, -0.023)

- based on profile log-likelihood $\ell_p(\beta_1)$, details to follow
- confint(logitmodcorrect):
  (-0.212~~2262~~, -0.0244~~701~~)

ELM-1 p. 31

Bernoulli data

ELM-2, Ch.2

- each response is $y_i = 0, 1$ instead of $0, 1, \dots, m_i$
- explanatory variables x_i^T as usual
- same model

$$\mathrm{pr}(y_i = 1 \mid x_i) = p_i(\beta) = \frac{\exp(x_i^T\beta)}{1 + \exp(x_i^T\beta)}$$

- example `wcgs` data, ELM-2, Ch.2
- example HW6: “The math group, the single dependent variable of this work, was coded as a dichotomous variable (1: math group vs. 0: nonmath group).”
- “To classify the math vs. nonmath groups, we also executed a binary logistic regression.”

$\longrightarrow$ Oct27-2.Rmd

In the News

$$y_{ij} = x_i^T \underline{\beta} + z_i \gamma + \epsilon_{ij} \qquad j = 1, \dots, n_i$$

$$\uparrow \qquad\qquad \uparrow \text{rand} \qquad\qquad i = 1, \dots, k$$

$$\gamma \sim (0, \sigma_r^2)$$

- The Real Scandal About Ivermectin

  Atlantic, Oct 23

$$\text{cor}(y_{ij}, y_{ij'})$$

- Nonreplicable publications are cited more than replicable ones

  Science Advances, May 21

- Post COVID-19 in children, adolescents and adults: results of a matched cohort study including more than 150,000 individuals with COVID-19

  MedRXiv, Oct 21 — not yet peer-reviewed

$$\bar{y}_1 - \bar{y}_2 \sim N\left(\mu_1 - \mu_2, \frac{\sigma^2}{m}\right)$$

$$\alpha_1 - \alpha_2$$

$$\mu_1 = \mu + \alpha_1$$

$$\mu_2 = \mu + \alpha_2$$

$$H_0: \alpha_1 - \alpha_2 = 0$$

CI for