Methods of Applied Statistics I

STA2101H F LEC9101

Week 10

November 23 2022

Masks Cut Covid Spread in Schools, Study Finds

In a so-called natural experiment, two school districts in Boston maintained masking after mandates had been lifted in others, enabling a unique comparison.

Give this article 177

Experts wrote in an accompanying editorial that research should help dispel misinformation about the effectiveness of universal masking requirements to stem viral transmission. Christopher Capozziello for The New York Times

A natural experiment

NEJM, Nov 10 2022

The NEW ENGLAND JOURNAL of MEDICINE

ORIGINAL ARTICLE

Lifting Universal Masking in Schools — Covid-19 Incidence among Students and Staff

Tori L. Cowger, Ph.D., M.P.H., Eleanor J. Murray, Sc.D., M.P.H., Jaylen Clarke, M.Sc., Mary T. Bassett, M.D., M.P.H., Bisola O. Ojikutu, M.D., M.P.H., Sarimer M. Sánchez, M.D., M.P.H., Natalia Linos, Sc.D., and Kathryn T. Hall, Ph.D., M.P.H.

ABSTRACT

BACKGROUND
In February 2022, Massachusetts rescinded a statewide universal masking policy in public schools, and many Massachusetts school districts lifted masking requirements during the subsequent weeks. In the greater Boston area, only two school districts — the Boston and neighboring Chelsea districts — sustained masking requirements through June 2022. The staggered lifting of masking requirements provided an opportunity to examine the effect of universal masking policies on the incidence of coronavirus disease 2019 (Covid-19) in schools.

METHODS
We used a difference-in-differences analysis for staggered policy implementation to compare the incidence of Covid-19 among students and staff in school districts in the greater Boston area that lifted masking requirements with the incidence in districts that sustained masking requirements during the 2021–2022 school year. Characteristics of the school districts were also compared.

Masks Cut Covid Spread in Schools, Study Finds

In a so-called natural experiment, two school districts in Boston maintained masking after mandates had been lifted in others, enabling a unique comparison.

Give this article 177

Experts wrote in an accompanying editorial that research should help dispel misinformation about the effectiveness of universal masking requirements to stem viral transmission. Christopher Capozziello for The New York Times

A natural experiment

NEJM, Nov 10 2022

$$y_{ijk} = \mu + \underset{\substack{\uparrow \\ A}}{\alpha_i} + \underset{\substack{\uparrow \\ B}}{\beta_j} + (\alpha\beta)_{ij} + \varepsilon_{ijk}$$

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

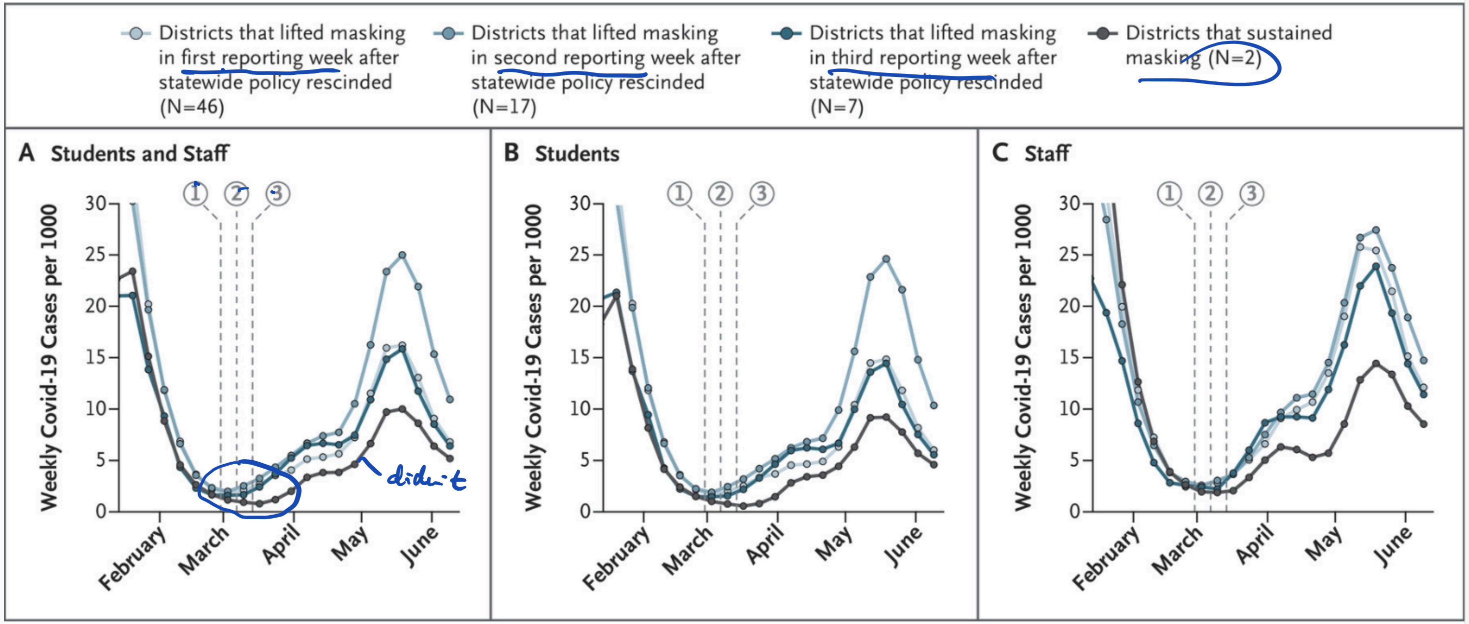


Figure 1. Incidence of Covid-19 in School Districts in the Greater Boston Area before and after the Statewide Masking Policy Was Rescinded.

The observed incidence of coronavirus disease 2019 (Covid-19) (weekly Covid-19 cases per 1000 population) among students and staff overall (Panel A), among students alone (Panel B), and among staff alone (Panel C) is shown for the 72 school districts in the greater Boston area that were included in the study. The greater Boston area was defined according to the U.S. Census Bureau as the New England city and town area of Boston–Cambridge–Newton. The Massachusetts Department of Elementary and Secondary Education rescinded the statewide masking policy on February 28, 2022. The incidence is shown according to whether the school district lifted its masking requirement in the first, second, or third reporting week after the statewide masking policy was rescinded or the district sustained its masking requirement. A school district was considered to have lifted its masking requirement in a given reporting week if the requirement had been lifted before the first day of the reporting week (reporting weeks start on Thursday). The dashed lines indicate the first (1), second (2), and third (3) school weeks (school weeks start on Monday) during which school districts lifted masking requirements. A total of 46 school districts lifted masking requirements during the first school week (starting on February 28, 2022) and in the first reporting week (starting on March 3, 2022) after the statewide masking policy was rescinded; 17 districts lifted masking requirements during the second school week (starting on March 7, 2022) and in the second reporting week (starting on March 10, 2022); 7 districts lifted masking requirements during the third school week (starting on March 14, 2022) and in the third reporting week (starting on March 17, 2022); and 2 districts sustained masking requirements. Data points are shown on the first day of the reporting week and represent 3-week trailing rolling averages to reduce statistical noise. Dates on the x axis are restricted to the period immediately before and after the statewide masking policy was rescinded.

Today

1. Upcoming events
2. Project
3. Recap
4. Finish survival data
5. Random and mixed effects models

↓ Nonpar. regr.

Project due December 19 (11.59),
no extensions
So think of it as due on December 16 :)

Project Guidelines STA 2101F: Methods of Applied Statistics I 2022

Outline

- Part I 3–5 pages, non-technical — 12 point type, 1.5 vertical spacing, thank you
  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. non-technical summary for a non-statistician of the analysis and conclusions
- Part II 3–5 pages, technical — LaTeX or R markdown; submit .Rmd and .pdf files
  1. models and analysis
  2. summary for a statistician of the analysis and conclusions
- Part III Appendix — submit .Rmd and .pdf or .html files
  R script or .Rmd file; additional plots; additional analysis; References

Project Marking

- 40 points total
- Part I:
  description of data and scientific problem 5
  suitability of plots and tables 5 — clear, non-technical, concise but thorough
  quality of the presentation 5
- Part II:
  summary of the modelling and methods 5 — justification for choices
  suitability and thoroughness of the analysis 10 — model checks, data checks
- Part III:
  relevance of additional material 5
  complete and reproducible submission 5

- November 24 3.30-4.30 Statistical Sciences Seminar
  Room 9014, Hydro Building
  and online

  $E(y|x) = x^T\beta$

  $q_{.95}(y|x) = \cdots x\beta^*$

  Keen Ming Tan, U Michigan
  "Convolution type smoothing approach for quantile regression"

- November 25 12.00-1.00 Toronto Data Workshop
  BL 520 and zoom
  Marcel Fortin and Leanne Trimble, U Toronto
  "... will talk about their newly acquired data collections, software and support"

...Upcoming

- November 28 3.30-4.30 Data Science ARES
  Room 9014, Hydro Building
  and online

  Jishnu Das, U Pittsburgh
  "Using Interpretable Machine Learning and Network Systems Approaches to Uncover Mechanisms Underlying Pathophysiology of Immune Disorders"

Recap

- generalized linear models: family, link, density
- generalized linear models: mean function, variance function, dispersion
- iteratively re-weighted least squares fitting; estimation of dispersion

- survival data: hazard function, survivor function, censoring
- parametric models: exponential, Gamma, Weibull, log-logistic
- likelihood function, log-likelihood, MLE, etc.
- nonparametric inference for survivor function   Kaplan-Meier estimator
- regression analysis: proportional hazards model; partial likelihood
- estimation of survivor function

Recap: Analysis of data using GLMs: overview

- choose a model, often based on type of response — or on mean/variance relationship
- fit a model, using maximum likelihood estimation — convergence (almost) guaranteed
- inference for individual coefficients $\hat{\beta}_j$ from `summary`
- inference for groups of coefficients by analysis of deviance

- estimation of ϕ based on Pearson's Chi-square

typo in ELM p.121: cross out $=$ var($\hat{\mu}$)

e.g. σ^2 $1/2$

$$\hat{\phi} = \frac{1}{n-p} \underbrace{\sum_{i=1}^{n} \frac{(y_i - \hat{\mu}_i)^2}{V(\hat{\mu}_i)}}$$

analog to $\hat{\sigma}^2 = \frac{RSS}{n-p}$

- analysis of deviance: see p. 121 (near bottom) — likelihood ratio tests
- diagnostics: same as for `lm` — ELM p.124; SM p.477
  - residuals: deviance or Pearson; can be standardized — ELM likes 1/2 normal plots
  - influential observations: uses hat matrix — `SMPracticals` has very good GLM diagnostics

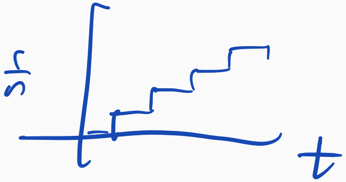
`glm.diag`, `plot.glm.diag`

- one sample $(y_1, d_1), \ldots, (y_n, d_n)$ survreg
- parametric model: Exponential, Weibull, Gamma, ...
- non-parametric: Kaplan-Meier estimator of $S(\cdot)$

$\hat{S}(t)$

generalizes empirical cdf

$\hat{F}(t)$



- covariates $(y_1, d_1, x_1), \ldots, (y_n, d_n, x_n)$
- parametric model: Exponential, Weibull, Gamma

$g(\mu_i) = x_i^T\beta$

$L(\theta; y, d)$

- semi-parametric: hazard function

$$\lambda(y; x, \beta) = \lambda_0(y)\exp(x^T\beta)$$

- semi-parametric: survivor function

$$S(y; x, \beta) = \{S_0(y)\}^{\exp(x^T\beta)}$$

$= \prod_{i=1}^{n} f(y_i; \theta)^{d_i} \times \{1 - F(y_i; \theta)\}$

$\lambda e^{-\lambda y_i}$

$\hat{\lambda}$ $\widehat{var}(\hat{\lambda})$

Inference in proportional hazards model

- model

$$\lambda(y; x, \beta) = \lambda_0(y) \exp(x^T \beta)$$

- data $(y_1, d_1, x_1), \ldots, (y_n, d_n, x_n)$ $\quad y_1 < y_2 < \ldots < y_n$

- inference about β uses partial likelihood (Cox likelihood)

$$L_{\text{part}}(\beta; t, x) = \prod_{failures} \frac{\exp(x_i^T \beta)}{\sum_{j \in \mathcal{R}_i} \exp(x_j^T \beta)}$$

- risk set $\mathcal{R}_i$ set of individuals still alive at the time the ith item fails
- inference

$$\underline{\ell'_{\text{part}}(\hat{\beta}) = 0;} \qquad \underline{-\ell''_{\text{part}}(\hat{\beta}) \doteq \{\widehat{var}(\hat{\beta})\}^{-1}}$$

$\ell = \log L$

Inference in proportional hazards model

- model

$$\lambda(y; x, \beta) = \lambda_0(y) \exp(x^T \beta)$$

- data $(y_1, d_1, x_1), \ldots, (y_n, d_n, x_n)$

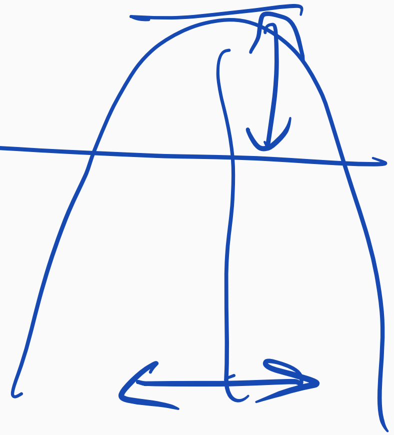
$y_1 < y_2 < \ldots < y_n$

- inference about β uses partial likelihood

$$L_{\text{part}}(\beta; t, x) = \prod_{\textit{failures}} \frac{\exp(x_i^T \beta)}{\sum_{j \in \mathcal{R}_i} \exp(x_j^T \beta)}$$

- risk set $\mathcal{R}_i$ set of individuals still alive at the time the ith item fails
- inference

$$\ell'_{\text{part}}(\hat\beta) = 0; \qquad -\ell''_{\text{part}}(\hat\beta) \doteq \{\widehat{\text{var}}(\hat\beta)\}^{-1}$$



$$\left[\begin{array}{c} \hat\beta - \beta \,\dot\sim\, N(0, \widehat{\text{var}}(\hat\beta)) \\ 2\{\ell_{\text{part}}(\hat\beta) - \ell_{\text{part}}(\beta_0)\} \,\dot\sim\, \chi^2_p \end{array}\right]$$

See Appendix to *An R Companion to Applied Regression*

```
> library(car)
> data(Rossi)
> Rossi[1:5, 1:10]
  week arrest fin age  race wexp         mar paro prio educ
1   20      1  no  27 black   no not married  yes    3    3
2   17      1  no  18 black   no not married  yes    8    4
3   25      1  no  19 other  yes not married  yes   13    3
4   52      0 yes  23 black  yes     married  yes    1    5
5   52      0  no  19 other  yes not married  yes    3    3

> mod.allison <- coxph(Surv(week,arrest) ~ fin + age + race + wexp + mar + paro
    + prio, data = Rossi)
```

… Example

```
> summary(mod.allison)
Call:
coxph(formula = Surv(week, arrest) ~ fin + age + race + wexp +
    mar + paro + prio, data = Rossi)

  n= 432, number of events= 114

                    coef exp(coef) se(coef)      z Pr(>|z|)
finyes          -0.37942   0.68426  0.19138 -1.983  0.04742 *
age             -0.05744   0.94418  0.02200 -2.611  0.00903 **
raceother       -0.31390   0.73059  0.30799 -1.019  0.30812
wexpyes         -0.14980   0.86088  0.21222 -0.706  0.48029
marnot married   0.43370   1.54296  0.38187  1.136  0.25606
paroyes         -0.08487   0.91863  0.19576 -0.434  0.66461
prio             0.09150   1.09581  0.02865  3.194  0.00140 **
```

$\hat\beta$ $e^{\hat\beta}$

$$\frac{e^{50\cdot\hat\beta}}{e^{49\cdot\hat\beta}} = .944$$

"holding the other covariates constant, an additional year of age reduces the weekly hazard of re-arrest by 0.944, that is, by 5.6%

- $$S(y;x) = \mathrm{pr}(Y \geq y \mid x) = \{S_0(y)\}^{\exp(x^T\beta)}$$

  $\prod \left(1 - \frac{d_i}{n-i+1}\right)$

- use partial likelihood to estimate β by $\hat\beta$
- estimate baseline survivor function as $\widehat{S}_0(y) = \prod_{i:y_i \leq y}\left(1 - \frac{d_i}{\sum_{j\in\mathcal{R}_{\rangle}} \exp(x_j^T\hat\beta)}\right)$

  K-M type

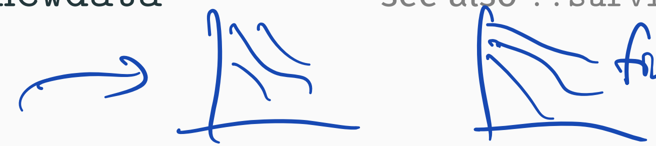
- estimate survivor function for individual with covariates x_+:

  $x_j(t)$

  $$\widehat{S}(y;x_+) = \{\widehat{S}_0(y)\}^{exp(x_+^T\hat\beta)}$$

- "the `survfit` function estimates $S(\cdot)$ by default at the mean value of the covariates"
- " we may wish to display how estimated survival depends on the value of a covariate"
- " this is passed to `survfit` through the argument `newdata`"

no



see also `??survfit`

Random effects models

- single source of variation: $y_1, \dots, y_n$, independent, $f(y_i \mid x_i; \beta, \sigma^2) = \dots$
- if observations arise in groups, or repeated measurements on the same individual, then sets of observations may be correlated
- or it may be natural to model more than one source of randomness

1-way layout

9 · Designed Experiments

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

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

$j = 1, \dots, 7$

$i = 1, \dots, 6$

subj. effect

One-way layout

- blood data: seven measurements on six subjects
- possible model

$$y_{ij} = \mu + \alpha_i + \epsilon_{ij}, \quad j = 1, \dots, 7; i = 1, \dots, 6$$

- using linear model formulation, rather than glm
- if parameters α_i viewed as constants, then interpretation is

$$\alpha_i - \alpha_{i'} = \mathrm{E}(y_{ij}) - \mathrm{E}(y_{i'j})$$

- e.g. expected difference in response between subject i and subject i'

One-way layout

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

$0 < = \sigma_\alpha^2 / (\sigma_\alpha^2 + \sigma_\varepsilon^2) < 1$

- blood data: seven measurements on six subjects
- possible model

$\downarrow$ 1 r.e.   error term

$$y_{ij} = \mu + \alpha_i + \epsilon_{ij}, \quad j = 1, \dots, 7; i = 1, \dots, 6$$

$\mathrm{var}(y_{ij}) = \sigma_\alpha^2 + \sigma_\varepsilon^2$

- using linear model formulation, rather than glm
- if parameters α_i viewed as constants, then interpretation is

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

$\alpha_i \perp \varepsilon_{ij}$

$$\alpha_i - \alpha_{i'} = \mathrm{E}(y_{ij}) - \mathrm{E}(y_{i'j})$$

$= \mathrm{cor}(\mu + \alpha_i + \varepsilon_{ij}, \mu + \alpha_i + \varepsilon_{ij'})$

$= \sigma_\alpha^2$

- e.g. expected difference in response between subject i and subject i'
- depending on the context, this may not be of interest
- e.g. if the subjects are a random sample, meant to represent a population
- if we view α_i as random, e.g. $\alpha_i \sim N(0, \sigma_\alpha^2)$, then σ_α^2 is the between-subject variance

$\mathrm{logit}\{\Pr(Y_{ij} = 1 \mid \alpha_i)\} = \mu + \alpha_i$ Y_{ij} binary

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

$L(\mu, \sigma_\alpha^2) = \int \prod_{j=1}^{n} f(y_{ij} \mid \alpha_i) \cdot f(\alpha_i) d\alpha_i$

One-way layout

- blood data: seven measurements on six subjects
- possible model

$$y_{ij} = \mu + \alpha_i + \epsilon_{ij}, \quad j = 1, \ldots, 7; i = 1, \ldots, 6$$

- using linear model formulation, rather than glm
- if parameters α_i viewed as constants, then interpretation is

$$\alpha_i - \alpha_{i'} = \mathrm{E}(y_{ij}) - \mathrm{E}(y_{i'j})$$

- e.g. expected difference in response between subject i and subject i'
- depending on the context, this may not be of interest
- e.g. if the subjects are a random sample, meant to represent a population
- if we view α_i as random, e.g. $\alpha_i \sim N(0, \sigma^2_\alpha)$, then σ^2_α is the between-subject variance
- if ϵ_{ij} is modelled as $N(0, \sigma^2_\epsilon)$, then σ^2_ϵ is within-subject variance
- interest may well focus on estimation of these two components of variance, and possibly estimation of μ, the population mean

In contrast

SM Example 9.2

Table 9.3 Data on the teaching of arithmetic.

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

Table 9.3 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

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

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

$y \sim$ group

anova (lm.fit)

$H_0: \alpha_i = 0$

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

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

Analysis of variance table

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

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

bet

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

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

Analysis of variance table

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

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

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

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

Analysis of variance table

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

Term	degrees of freedom	sum of squares	mean square	F-statistic	Expected MS
treatment	$(I-1)$	$SS_{between}$	$MS_{between}$	$MS_{between}/MS_{within}$	$J\sigma_\alpha^2 + \sigma_\varepsilon^2$
error	$I(J-1)$	SS_{within}	MS_{within}		σ_ε^2
total(corrected)	$IJ-1$	SS_{total}			

$\tilde{\sigma}_\varepsilon^2 = MS_w \cancel{\varepsilon} = SS_w/df_w$

Components of variance

$\frac{MS_b}{\cancel{SS}} \; \cancel{=} \; J\tilde{\sigma}_\alpha^2 + \tilde{\sigma}_\varepsilon^2$

- 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

$$\frac{MS_b - \tilde{\sigma}_\varepsilon^2}{J} = \tilde{\sigma}_\alpha^2$$

$(y_1 \quad d_1)$

5 0

10 1

Components of variance

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

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

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

...components of variance

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

- $y_{ij} = \mu + \alpha_i + \epsilon_{ij}, \quad \epsilon_{ij} \sim (0, \sigma^2_\epsilon), \quad \alpha_i \sim (0, \sigma^2_\alpha) \qquad i = 1, \dots, I; j = 1 \dots J$
- variance of response within subjects
- variance of response between subjects

$E(SS_{bet}) = (I-1)(J\sigma^2_\alpha + \sigma^2_\epsilon)$

$E(SS_w) = I(J-1)\sigma^2_\epsilon$

- as above,

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

$$SST = SS_{between} + SS_{within}$$

- $\mathrm{E}(SS_{within}) = I(J-1)\sigma^2_\epsilon \quad \mathrm{E}(SS_{between}) = (I-1)(J\sigma^2_\alpha + \sigma^2_\epsilon)$
- $SS_{within} \sim \sigma^2_\epsilon \chi^2_{I(J-1)} \qquad SS_{between} \sim (J\sigma^2_\alpha + \sigma^2_\epsilon)\chi^2_{I-1}$ leads to F-test for $H_0 : \sigma^2_\alpha = 0$
- and estimates $\tilde{\sigma}^2_\epsilon = SS_{within}/I(J-1)$, $\tilde{\sigma}^2_\alpha = (MS_{between} - MS_{within})/J$

(fixed) $\alpha_1 = \dots = \alpha_I = 0$

$$SS_w \sim \sigma_\varepsilon^2 \chi^2_{I(J-1)}$$

under N model

$\varepsilon_{ij} \perp \alpha_i$

$$SS_b \sim (J\sigma_\alpha^2 + \sigma_\varepsilon^2)\chi^2_{I-1}$$

independent.

$$\frac{\frac{1}{(J\sigma_\alpha^2+\sigma_\varepsilon^2)} SS_b / (I-1)}{\frac{1}{\sigma_\varepsilon^2} SS_w / I(J-1)} \sim F_{I-1,\, I(J-1)}$$

$$\frac{\chi^2_{\nu_1}}{\chi^2_{\nu_2}}$$

$$L < \frac{\cancel{\sigma_\varepsilon^2}\ 1}{\frac{J\sigma_\alpha^2}{\sigma_\varepsilon^2} + \frac{\cancel{\sigma_\varepsilon^2}}{1}} \leq u$$

$\left(\sigma_\alpha^2 / \sigma_\varepsilon^2 \right.$ or $\sigma_\varepsilon^2 / \sigma_\alpha^2$

if can be done

Example

```
> anova(lm(y ~ subject, data = sticky))
Analysis of Variance Table

Response: y
          Df Sum Sq Mean Sq F value  Pr(>F)
subject    5 1465.9  293.18  2.3198 0.06327 .
Residuals 36 4549.7  126.38
---
```

Subj

1 2 . . . 6

j = 1 … 7

MS_b MS_w

```
> (293.18-126.38)/7
[1] 23.82857
```

$= J$

$\tilde{\sigma}_\varepsilon^2 = 126.38$

$= (11.2)^2$

$\tilde{\sigma}_\alpha^2 = (4.8)^2$ (?)

$= 23.8$

random, nested within intercept

```
> library(lme4)
> mmod <- lmer(y ~ 1 +(1|subject), data = sticky)
> summary(mmod)
Linear mixed model fit by REML ['lmerMod']
Formula: y ~ 1 + (1 | subject)
   Data: sticky
...
Random effects:
 Groups   Name        Variance Std.Dev.
 subject  (Intercept)  23.83    4.881
 Residual             126.38   11.242
Number of obs: 42, groups:  subject, 6

Fixed effects:
            Estimate Std. Error t value
(Intercept)   41.905      2.642   15.86
```

σ_α^2 σ_α σ_ε σ_ε^2 $\hat{\mu}$

$H_0: \sigma_\alpha^2 = 0$ using F-st.

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

$\mu = E(y_{ij})$ in pop^n

α_i random ~~offset~~ for ith subject

intercept α_i for ith subject

$(\mu + \alpha_i)$

fixed random

CI for $\frac{\sigma_\alpha^2}{2}$

- we might have more than one level of variation
- SM Example: H hospitals; S surgeons at each hospital; P patients treated by each surgeon
- response is a measure of success of surgery   assume continuous
- linear model:

$$y_{hsp} = \mu + b_h + \underline{a_{hs}} + \epsilon_{hsp}, \quad h = 1, \ldots, H; s = 1, \ldots, S; p = 1, \ldots P$$

- patient 1 treated by surgeon 1 in hospital 1 has no relation to patient 1 treated by surgeon 1 in hospital 2, etc.
- interpretation? $\underline{b_h}$ departure from average success (μ) in hospital h
- a_{hs}

- depending on the context, we may treat factors as fixed, or random

… Nested variation

Term	degrees of freedom	sum of squares	Expected mean square
between hospitals	$(H-1)$	$\Sigma_{h,s,p}(\bar{y}_{h..} - \bar{y}_{...})^2$	$PS\sigma_b^2 + P\sigma_a^2 + \sigma^2$
between surgeons, within hospitals	$H(S-1)$	$\Sigma_{h,s,p}(\bar{y}_{hs.} - \bar{y}_{h..})^2$	$P\sigma_a^2 + \sigma^2$
between patients within surgeons	$HS(P-1)$	$\Sigma_{h,s,p}(y_{hsp} - \bar{y}_{hs.})^2$	σ^2

$\tilde{\sigma}_a^2 =$

$\tilde{\sigma}^2 = MS_{bpws}$

linear model:

$$y_{hsp} = \mu + b_h + a_{hs} + \epsilon_{hsp}, \quad h = 1, \ldots, H; s = 1, \ldots, S; p = 1, \ldots P$$

$$b_h \sim N(0, \sigma_b^2), \quad a_{hs} \sim N(0, \sigma_a^2), \quad \epsilon_{hsp} \sim N(0, \sigma^2)$$

ELM-2 10.8

see also ELM-2 §10.8 (ELM-1 §8.6) for another example

- more usual to have a model with some fixed effects: treatments, explanatory variables (age, income, ...)
- and some random effects: cluster, family, school, hospital, ...
- the general form of a linear mixed effect model is

$$y = X\beta + Z\gamma + \epsilon$$

$$\underset{n\times 1}{y} = \underset{n\times p}{X}\underset{p\times 1}{\beta} + \underset{n\times 1}{\varepsilon}$$

- model matrix $X_{n\times p}$, fixed effects β
- model matrix Z, random effects γ

$n \times q$ $\quad q \times 1$

$$\gamma \sim N(0, \sigma^2 D)$$

$$y_{ij} = \beta_0 + \beta_1 age_{ij} + \gamma_1 school_{ij} + \varepsilon_{ij} \qquad j = 1, \ldots, J$$

$$+ \gamma_2 \overline{age_{ij}}$$

repeat obs =

Mixed effects models

- more usual to have a model with some fixed effects: treatments, explanatory variables (age, income, ...)
- and some random effects: cluster, family, school, hospital, ...
- the general form of a linear mixed effect model is

$$y = X\beta + Z\gamma + \epsilon$$

- model matrix $X_{n\times p}$, fixed effects β
- model matrix Z, random effects γ
- if we assume $\epsilon \sim N(0, \sigma^2 I)$, then model is

$$Y \mid \gamma \sim N(X\beta + Z\gamma, \sigma^2 I)$$

$$\to f_Y(y) = \int \underset{\text{N. density}}{\downarrow} \cdot \overset{f(\gamma_i;\sigma^2 D)}{\cancel{N(0,\sigma^2 D)}}\, d\gamma_i$$

- $$Y \mid \gamma \sim N(X\beta + Z\gamma, \sigma^2 I)$$
- If in addition $\gamma \sim N(0, \sigma^2 D)$,

- $$Y \mid \gamma \sim N(X\beta + Z\gamma, \sigma^2 I)$$

$\gamma_{q\times 1}$

- If in addition $\gamma \sim N(0, \sigma^2 D)$,

$D_{q\times q}$

- $$Y \sim N\{X\beta, \sigma^2(I + ZDZ^T)\}$$

marginal distribution

- but still conditional on X and Z

explanatory variables

- unknown parameters: β, D, and σ^2
- could estimate by maximum likelihood $Y \sim N(X\beta, \sigma^2 V)$, $V = I + ZDZ^T$
- $$L(\beta, \sigma^2, D; Y) = \frac{1}{(2\pi)^{n/2}|\sigma^2 V|^{1/2}} \exp -\frac{1}{2\sigma^2}(y - X\beta)^T V^{-1}(y - X\beta)$$
- default in `lme4` is to use "REML"

restricted maximum likelihood

$\hat{\beta}$ $\hat{\sigma}^2$ $\hat{D}$

- inference for fixed effects:

$$\hat{\beta} \sim N(\beta, \sigma^2\{X^T(I + ZDZ^T)^{-1}X\}^{-1})$$

- need estimates of D and σ^2 $\sigma^2 D = var(\gamma)$
- the normal distribution is only approximate, when D is estimated
- and can be a poor approximation, if true var(γ) is very small
- we might also want to test whether some components of variance are 0
- standard likelihood theory does not apply boundary
- extensive discussion in ELM 10.2 rather confusing

- conceptually simpler to think of $N(0, D)$ as a prior distribution for γ, and compute (or sample from) the posterior distribution AS II

Rat growth data SM Ex 9.18

- repeated measurements on the 30 individuals, at 5 time points
- might expect that regression relationship against time is similar for each individual, subject to random variation
- model $y_{jt} = \beta_0 + b_{j0} + (\beta_1 + b_{j1})x_{jt} + \epsilon_{jt}, \quad t = 1, \ldots, 5$
- x_{jt} takes values $0, 1, 2, 3, 4$ for $t = 1, 2, 3, 4, 5$
- same for each j
- `data(rat.growth, library="SMPracticals")`
- $(b_{j0}, b_{j1}) \dot\sim N_2(0, \Omega_b)$, $\quad \epsilon_{jt} \dot\sim N(0, \sigma^2)$ independent
- two fixed parameters β_0, β_1
- four variance/covariance parameters: $\sigma^2_{b0}, \sigma^2_{b1}, \text{cov}(b_0, b_1), \sigma^2$

$j = 1, \ldots 30$

$y_{jt} = \beta_0 + \beta_1 x_{jt} + \epsilon_{jt}$

$y_{jt} = \beta_0 + \beta_1 d_j + \beta_2 x_{jt} + \beta_3 x_{jt} d_j$

$\beta_0 + \beta_1 x_{jt} + b_0 + b_{j1} x_{jt}$

… Example 9.18

time → covariate

animals

	Week 1	2	3	4	5		Week 1	2	3	4	5
1	151	199	246	283	320	16	160	207	248	288	324
2	145	199	249	293	354	17	142	187	234	280	316
3	147	214	263	312	328	18	156	203	243	283	317
4	155	200	237	272	297	19	157	212	259	307	336
5	135	188	230	280	323	20	152	203	246	286	321
6	159	210	252	298	331	21	154	205	253	298	334
7	141	189	231	275	305	22	139	190	225	267	302
8	159	201	248	297	338	23	146	191	229	272	302
9	177	236	285	340	376	24	157	211	250	285	323
10	134	182	220	260	296	25	132	185	237	286	331
11	160	208	261	313	352	26	160	207	257	303	345
12	143	188	220	273	314	27	169	216	261	295	333
13	154	200	244	289	325	28	157	205	248	289	316
14	171	221	270	326	358	29	137	180	219	258	291
15	163	216	242	281	312	30	153	200	244	286	324

Table 9.27 Weights (units unknown) of 30 young rats over a five-week period (Gelfand *et al.*, 1990).

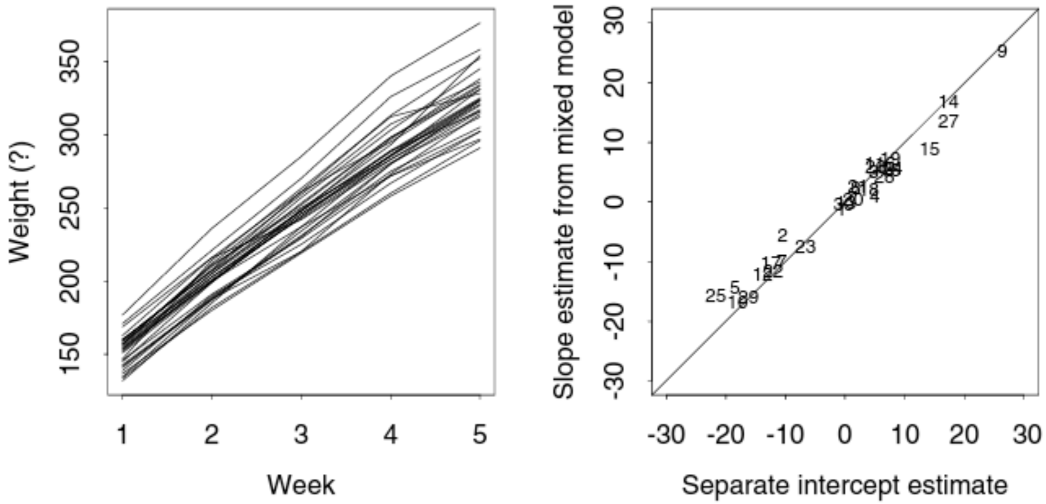


Figure 9.9 Rat growth data. Left: weekly weights of 30 young rats. Right: shrinkage of individual slope estimates towards overall slope estimate; the solid line has unit slope, and the estimates from the mixed model lie slightly closer to zero than the individual estimates.

We treat the rats as a sample from a population of similar creatures, with different initial weights and growing at different rates. To model this we express the data from the jth rat as

$$y_{jt} = \beta_0 + b_{j0} + (\beta_1 + b_{j1})x_{jt} + \varepsilon_{jt}, \quad t = 1, \ldots 5,$$

… Example 9.18

- maximum likelihood estimates of fixed effects: $\hat{\beta}_0 = 156.05(2.16)$, $\hat{\beta}_1 = 43.27(0.73)$
- weight in week 1 is estimated to be about 156 units, and average increase per week estimated to be 43.27
- there is large variability between rats: estimated standard deviation of 10.93 for intercept, 3.53 for slope
- there is little correlation between the intercepts and slopes

```
 with(rat.growth, plot( y ~ week , type="l"))
> separate.lm = lm(y ~ week + factor(rat)+ week:factor(rat), data = rat.growth)

> rat.mixed = lmer(y ~ week + (week|rat), data = rat.growth) # REML is the defaul
> summary(rat.mixed)  # compare Table 9.28
```

$\beta_0 + \beta_1$ time

$b_{0j} + b_{1j}$ time

```
> summary(rat.mixed)
Linear mixed model fit by REML ['lmerMod']
Formula: y ~ week + (week | rat)
   Data: rat.growth
...
Random effects:
 Groups   Name        Variance Std.Dev. Corr
 rat      (Intercept) 119.54   10.933
          week         12.49    3.535   0.18
 Residual              33.84    5.817
Number of obs: 150, groups:  rat, 30

Fixed effects:
            Estimate Std. Error t value
(Intercept) 156.0533     2.1590   72.28
week         43.2667     0.7275   59.47
```

b_0

b_1

σ^2

$\frac{\mathrm{cov}(b_0, b_1)}{\sqrt{\sigma^2_{b_0}\sigma^2_{b_1}}}$

"the estimated mean weight in week 1 is 156, but the variability from rat to rat has standard deviation of about 11 about this.

The slopes show similarly large variation.

The measurement error variance $\hat{\sigma}^2 = 5.82^2$ is smaller than the inter-rat variation in intercepts but exceed that for slopes"

Fox & Weisberg

Mixed Models in R

Generalized linear mixed models

- linear model: random effect induces correlation

- binary regression: