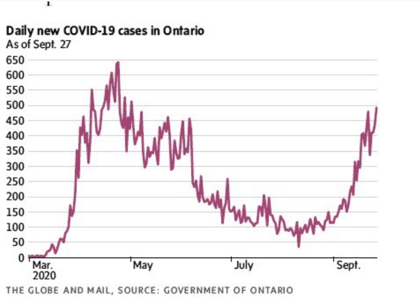
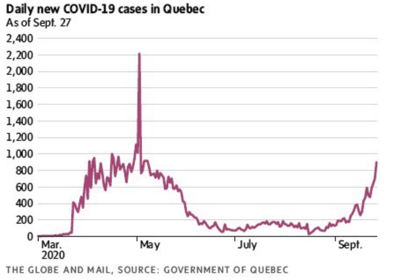
Methods of Applied Statistics I

STA2101H F LEC9101

Week 4

October 1 2020



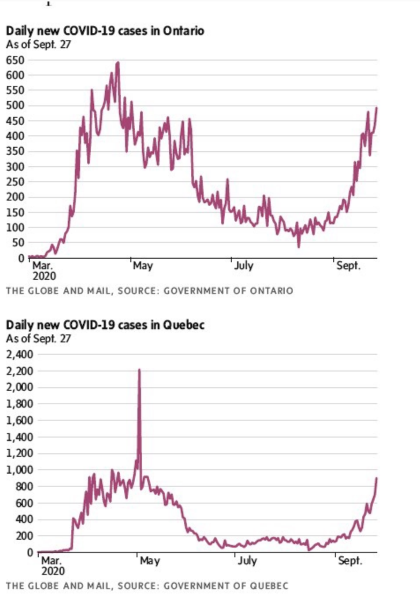


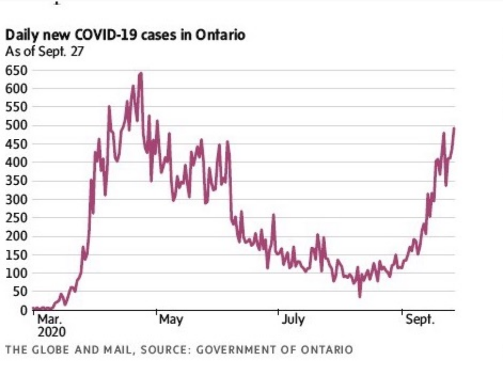
On a quick look, which province

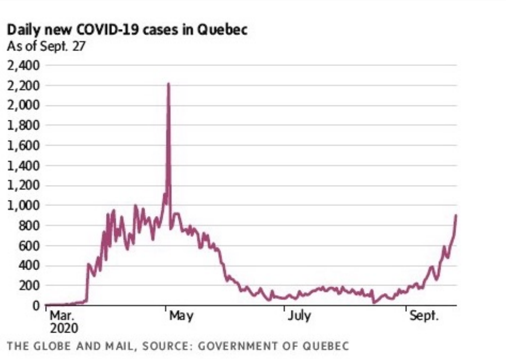
is in worse shape now?

(1) Ontario — top graph

(2) Quebec — bottom graph







On 3 May 2020, the large increase in cases includes the increase from the previous day (892 new cases) and 1,317 cases from April that hadn't yet been tabulated due to a technical problem. Source: CBC News via Wikipedia

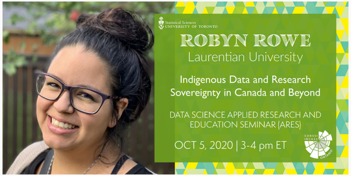
Institut nationale de santé publique

Public Health Ontario

Today

Start Recording

1. Syllabus updates; Two editions of Faraway; Next week; student life
2. In the News: the story that won't die
3. Linear Regression Part 4: Factors, random and mixed effects
4. Principles of Measurement – CD Ch. 4
5. (2–3pm) Discussion, questions, etc.



- September 28 3.30 – 4.30
- October 5 3.30 – 4.30
- https://canssiontario.utoronto.ca/?mec-events=data-science-ares-robyn_rowe

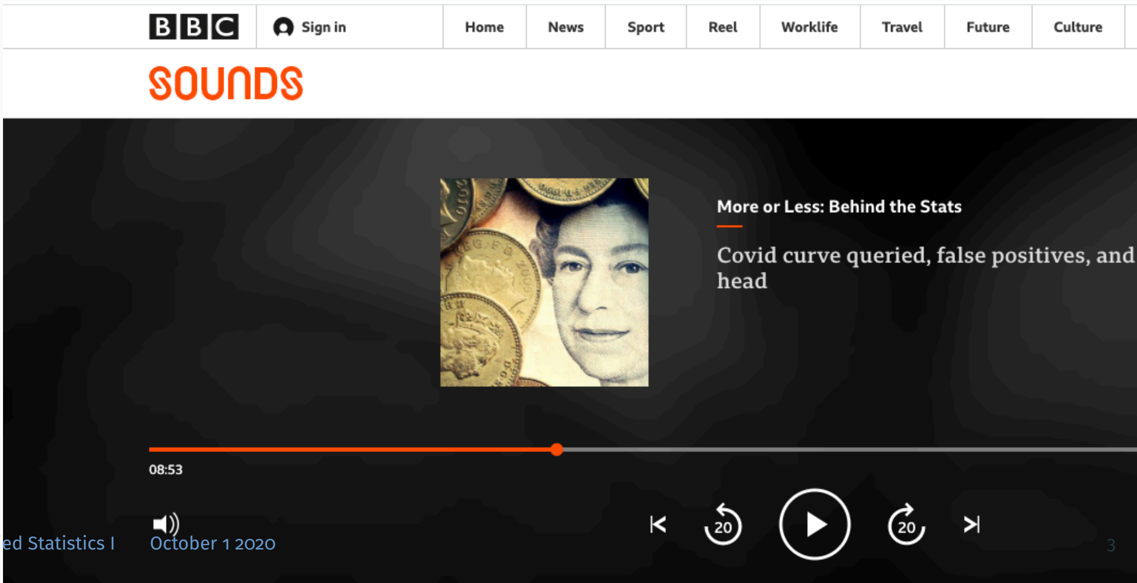
http://www.utstat.utoronto.ca/reid/sta2101f/syllabus20Update-1.pdf

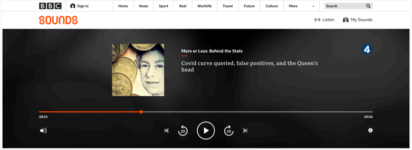
Syllabus updated Sep 28 STA 2101F: Methods of Applied Statistics I

Week	Date	Methods	References	Computing
1	Sept 10	Review of Linear Regression	SM Ch.8.2.1, 8.3; FLM-2 Ch.2-4; FLM-1 Ch.2-3; CD Ch.1	RStudio and RMarkdown
2	Sept 17	~~Model Selection~~ Comparing models; factors; model checking; diagnostics; collinearity	SM Ch.8.5,6; FLM Ch.3; FLM-2 14,2.11,6; FLM-1 4,13; CD Ch.6	~~tidyverse~~
3→HW1	Sept 24	~~Random and Mixed Effects Models~~Model selection; Types of studies	SM 8.7; FLM-2 Ch. 10; FLM-1 Ch.8; CD Ch.1,2	~~ggplot~~ HW 1 Qs
4←HW1	Oct 1	~~Designed Experiments~~ Factor variables; Random and Mixed Effects; Principles of Measurement	SM Ch. 9.1,9.2; FLM-2 Ch.14-17;FLM-1 Ch.13-16; CD Ch.4	
5	Oct 8	~~Binary Responses~~Designed Experiments; Preliminary Analysis	SM Ch.9.1,2; FLM-2 Ch.14-17 FLM-1 Ch.13-16; Ch.2; CD Ch.5	
6	Oct 15	Logistic Regression	SM 10.6.1; FELM Ch.3	
7→HW2	Oct 22	Generalized Linear Models	FELM Ch.6,7; SM 10.3	

False Positives, Again!

More or Less, Sep 23 8:53–15:30





prevvals <dbl>	ppv <dbl>	ppv2 <dbl>
0.0001	0.0079373	0.0259766
0.0010	0.0741427	0.2106927
0.0050	0.2867384	0.5726557
0.0100	0.4469274	0.7292616
0.0500	0.8080808	0.9334889
0.1000	0.8988764	0.9673519

Chance of a false positive case: 1 % 0.03%

This just in

RStudio Blog

Home About Categories Tags Archives

RStudio v1.4 Preview: Visual Markdown Editing

J.J. Allaire

2020-09-30

Categories: RStudio IDE R Markdown Tags: preview rstudio rmarkdown

Today we're excited to announce availability of our first Preview Release for RStudio 1.4, a major new release which includes the following new features:

- A visual markdown editor that provides improved productivity for composing longer-form articles and analyses with R Markdown.
- New Python capabilities, including display of Python objects in the Environment pane, viewing of Python

Recap of Linear Regression Part 3

- Estimation of β, σ^2 — t-statistic for testing $\beta_j = 0$
- Estimation of $\mathbb{E}(y \mid x_+)$ — estimated error of $x_+\hat{\beta}$
- Prediction of response at covariate values x_+ — prediction error

- Model selection: hierarchical models
- Model selection: testing procedures – forward, backward, stepwise
- Model selection: information criteria AIC, BIC, adjusted R^2, C_p

- Model selection via Lasso

- Question: if you remove a factor variable with k levels, does the AIC penalty decrease by $k - 1$ or by 1?
- Answer: it decreases by $k - 1$, as it should — see `prostate.R` on web page

Factor variables

- in `prostate` data, variable `gleason` takes on just 4 values
- if introduced as it is in the data frame, it will be treated as a continuous variable
- we can make it into a factor variable using `gleason-f <- factor(gleason)`
- what's the difference?

```
                 Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)      0.913282   0.840838   1.086  0.28044
lcavol           0.569988   0.090100   6.326 1.09e-08 ***
lweight          0.468791   0.169610   2.764  0.00699 **
age             -0.021749   0.011361  -1.914  0.05890 .
lbph             0.099685   0.058984   1.690  0.09464 .
svi              0.745879   0.247398   3.015  0.00338 **
lcp             -0.125112   0.095591  -1.309  0.19408
pgg45            0.004990   0.004672   1.068  0.28848
gleason_factor7  0.267607   0.219419   1.220  0.22595
gleason_factor8  0.496820   0.769267   0.646  0.52011
gleason_factor9 -0.056215   0.500196  -0.112  0.91078
```

Factor variables: coefficient estimates

svi	0.745879	0.247398	3.015	0.00338	**
lcp	-0.125112	0.095591	-1.309	0.19408	
pgg45	0.004990	0.004672	1.068	0.28848	
gleason_factor7	0.267607	0.219419	1.220	0.22595	
gleason_factor8	0.496820	0.769267	0.646	0.52011	
gleason_factor9	-0.056215	0.500196	-0.112	0.91078	

expected response at level 7, relative to level 6 is .267 units higher;

level 8 relative to level 6
level 9 relative to level 6

svi	0.766157	0.244309	3.136	0.00233	**
lcp	-0.105474	0.091013	-1.159	0.24964	
gleason	0.045142	0.157465	0.287	0.77503	
pgg45	0.004525	0.004421	1.024	0.30886	

for every unit increase in `gleason`, expected response increases by 0.045

all other variables held fixed

top estimates are difficult to interpret, as they are all referenced to level 6

bottom estimates assume the score is quantitative

Factor variables: analysis of variance

```
Analysis of Variance Table

Response: lpsa
               Df Sum Sq Mean Sq
lcavol          1 69.003  69.003
lweight         1  5.949   5.949
age             1  0.420   0.420
lbph            1  1.069   1.069
svi             1  5.952   5.952
lcp             1  0.129   0.129
pgg45           1  1.192   1.192
gleason_factor  3  1.480   0.493
Residuals      86 42.724   0.497
```

```
Analysis of Variance Table

Response: lpsa
          Df Sum Sq Mean Sq
lcavol     1 69.003  69.003
lweight    1  5.949   5.949
age        1  0.420   0.420
lbph       1  1.069   1.069
svi        1  5.952   5.952
lcp        1  0.129   0.129
gleason    1  0.708   0.708
pgg45      1  0.526   0.526
Residuals 88 44.163   0.502
```

Factor variables: modelling

- a factor variable is treated as categorical
- a non-factor variable is treated as continuous
- it depends on the application which is preferred

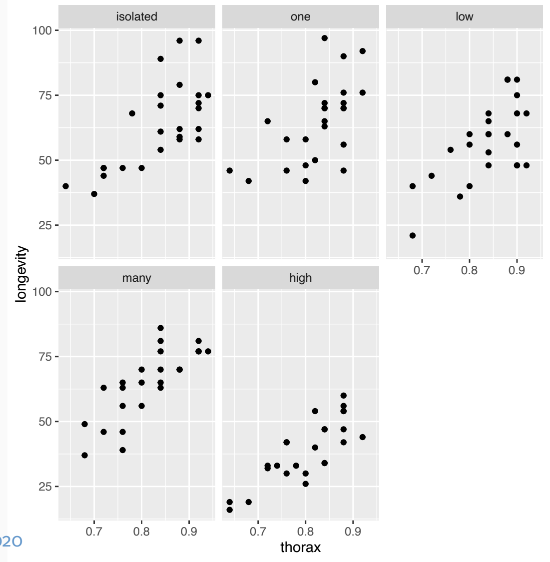
- a linear model with one factor and one continuous variable might be written as, for example:

$$y_{ij} = \mu + \alpha_j + \beta x_{ij} + \epsilon_{ij}, \quad j = 1, \ldots, J; \quad i = 1, \ldots m$$

- linear in x, but arbitrary changes in $\mathbb{E}(y)$ by category (here indexed by j)
- `R` doesn't distinguish this at the modelling phase:
  `lm(response ~ variable1 + variable2, data = ...)`
- but uses metadata in the data frame to accommodate factors
- `is.factor(variable)` and `newvar <- as.factor(oldvar)` are helpful

Factor variables: modelling

fruitfly.Rmd



⟶ **fruitfly.Rmd**

Factor variables: examples

- Cycling: SM Example 8.4, 8.8, 8.12, 8.22 – designed experiment with 3 factors, each at 2 levels and each of these 8 combinations used twice, for a sample size of 16
- Poison: SM Example 8.25 – 2 factors, one has 4 levels, one has 3 levels, repeated four times, for a sample size of $12 \times 4 = 48$
- Some classical designs: SM §9.2 – Example 9.2, 9.3, 9.49.5, 9.6(8.25), 9.13
- FLM-2 – Chapters 14 through 17; FLM-1 – Chapters 13 through 16
- Why bother with special techniques for factor variables since we can fit them all using `lm`?

… factor variables: examples

- Why bother with special techniques for factor variables since we can fit them all using `lm`?
- If the experiment is designed – meaning treatment assignment under the control of the investigator, then we have stronger conclusions
- If the experiment is balanced, then the estimates of the effects of different factors are independent

  $X^{\mathrm{T}}X$ is orthogonal
- If the experiment is replicated, we can obtain reliable estimates of σ^2
- If the experiment is blocked, we can remove sources of error

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

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

- 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

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

Analysis of variance: one-factor design

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

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}		

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

See SM Table 9.3 and 9.4; FLM-2 §15.2; FlM-1 §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

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	

Pause

(1) New to me

(2) Not sure

(3) Seen it before

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.

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

- as before,

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

- induces dependence among measurements on the same subject: ntbc

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

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

Principles of measurement

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

- $\longrightarrow$ **CD Ch.4, p54,55**

54 *Principles of measurement*

of sampling may give higher quality than the study of a complete population of individuals.

Types and phases of analysis

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

Avoidance of systematic error

- "distortion in the conclusions arising from irrelevant sources that do not cancel out in the long run"
- can arise through systematic aspects of, for example, a measuring process, or the spatial or temporal arrangement of units
- this can often be avoided by design, or adjustment in analysis
- can arise by the entry of personal judgement into some aspect of the data collection process
- this can often be avoided by randomization and blinding

Randomized block design

- two factor variables, treatment and block
- design: treatments assigned at random within blocks
- model:

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

- parameters:
  - $\mu = \mathbb{E}(y_{ij}$ if all $\alpha_i \equiv 0; \beta_j \equiv 0;$
  - α_i is change in $\mathbb{E}(y)$ from μ due to treatment i
  - β_j is change in $\mathbb{E}(y)$ due to effect of block j
  - ϵ_{ij} unexplained variation
- analysis:

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