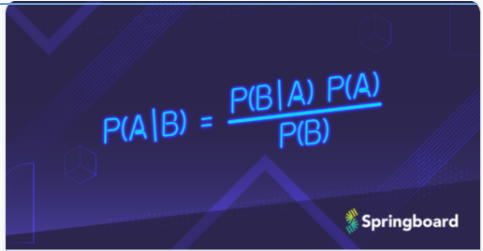
Methods of Applied Statistics I

STA2101H F LEC9101

Week 2

September 17 2020



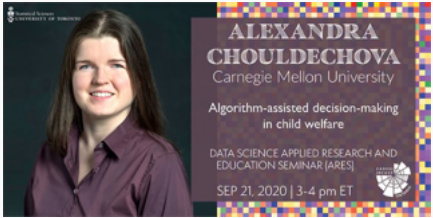
Become a Data Scientist in 6 Months
springboard.com

Today

Start Recording

1. Office hours, Covid competition, Recap last week, Zoom
   OH: Wednesday 9am-10.30am Monday 4pm-5.30pm, 7pm-8pm in Course Room
2. Linear Regression Part 2: testing groups of variables, checking model assumptions, collinearity, $p > n$
3. Types of studies
4. In the News: moon-shot covid testing; a little more on event attribution
5. RStudio and Rmd clinic

- SM – Statistical Models by Davison
- FLM – Linear Models with R by Faraway
- FELM – Extending the Linear Model with R by Faraway



Recap

- generic form of linear regression, in matrix notation $y = X\beta + \epsilon$
- least squares estimate of β is $\hat{\beta} = (X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}y$
- $\hat{\beta}$ has expected value β and variance-covariance matrix $\sigma^2(X^{\mathrm{T}}X)^{-1}$

- this is the maximum likelihood estimate if $\epsilon \sim N(0, \sigma^2 I)$
- $\hat{\beta} \sim N(\beta, \sigma^2(X^{\mathrm{T}}X)^{-1})$
- $\tilde{\sigma}^2 = (y - X\hat{\beta})^{\mathrm{T}}(y - X\hat{\beta})/(n - p)$ called s^2 in SM
- leads to t-tests for individual components β_j and confidence intervals - ntbc

- X is an $n \times p$ matrix of explanatory variables, which may be
  - measured in the sample (SM Ex 8.3),
  - fixed by design (SM Ex 8.4),
  - introduced to make the model more flexible (SM Ex 8.2) SM – Davison
  - X often called the design matrix in R, `model.matrix`

Aside: Lazy Notation

- $y_i = x_i^{\mathrm{T}}\beta + \epsilon_i, \quad , i = 1, \ldots, n$
- $\boldsymbol{y} = X\boldsymbol{\beta} + \boldsymbol{\epsilon}, \qquad \boldsymbol{y}, \boldsymbol{\epsilon}$ vectors of length n
- $y = X\beta + \epsilon, \qquad$ also vectors of length n

the lazy way

- a generic observation $y \in \mathbb{R}$ for a generic vector of covariates $x \in \mathbb{R}^1$ often written

$$y = x^{\mathrm{T}}\beta + \epsilon$$

or even $x\beta + \epsilon$

- "where we hope there is no confusion"

Comparing Models

SM 8.5, FLM 3.1,2

- residual sum of squares

$$SS(\hat{\beta}) = RSS_\Omega = (y - X\hat{\beta})^{\mathrm{T}}(y - X\hat{\beta})$$

$SS(\hat{\beta})$ SM p.366; RSS_Ω FLM-2 p.16; FLM-1, p.15

- Decomposition of variance: $y^{\mathrm{T}}y = (y - \hat{y})^{\mathrm{T}}(y - \hat{y}) + \hat{y}^{\mathrm{T}}\hat{y}$ FLM Fig 2.1

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

$$= \text{Residual SS + Regression SS}$$

- Typically first column of X is $(1, \ldots, 1)^{\mathrm{T}}$, so $y = \beta_0 + X_2\beta_2 + \epsilon$, say; then decomposition becomes

$$\sum_{i=1}^{n}(y_i - \bar{y})^2 = (y - X_2\hat{\beta}_2)^{\mathrm{T}}(y - X_2\hat{\beta}_2) + \hat{\beta}_2^{\mathrm{T}}(X_2^{\mathrm{T}}X_2)\hat{\beta}$$

$$(y - \bar{y}\mathbf{1})^{\mathrm{T}}(y - \bar{y}\mathbf{1}) = \sum_{i=1}^{n}(y_i - x_{i2}^{\mathrm{T}}\hat{\beta}_2)^2 + \hat{\beta}_2^{\mathrm{T}}(X_2^{\mathrm{T}}X_2)\hat{\beta}$$

... comparing models

$$\begin{aligned}\sum_{i=1}^{n}(y_i-\bar{y})^2 &= (y-X\hat{\beta})^{\mathrm{T}}(y-X\hat{\beta})+\hat{\beta}^{\mathrm{T}}(X^{\mathrm{T}}X)\hat{\beta}\\ \text{Total SS} &= \text{Residual SS}+\text{Regression SS}\end{aligned}$$

- LHS is residual SS fitting only the 1-vector
- comparison of LHS to $SS(\hat{\beta})$ reflects importance of other βs, i.e. importance of explanatory variables
- $$F=\frac{(TSS-RSS)/(p-1)}{RSS/(n-p)} \quad \sim F_{p-1,n-p}$$
- here $\beta=(\beta_1,\beta_2,\ldots,\beta_p)$, but we don't care about β_1 $\qquad (\beta_0,\beta_1,\ldots,\beta_p)$

... comparing models

- same argument can be derived for comparing submodels
- for example, testing $(\beta_2, \beta_3, \beta_4) = (0, 0, 0)$

- fit full model $\longrightarrow$ RSS_{full}; fit reduced model $\longrightarrow$ RSS_{red}
- $$F = \frac{(RSS_{red} - RSS_{full})/(p-q)}{RSS_{full}/(n-p)}$$
- see SM §8.2 (p.367) for connection to likelihood ratio test

- when would we want to do this?

... comparing models

```
head(prostate)
#     lcavol lweight age      lbph svi      lcp gleason pgg45     lpsa
1 -0.5798185  2.7695  50 -1.386294   0 -1.38629       6     0 -0.43078
2 -0.9942523  3.3196  58 -1.386294   0 -1.38629       6     0 -0.16252
3 -0.5108256  2.6912  74 -1.386294   0 -1.38629       7    20 -0.16252
4 -1.2039728  3.2828  58 -1.386294   0 -1.38629       6     0 -0.16252
5  0.7514161  3.4324  62 -1.386294   0 -1.38629       6     0  0.37156
6 -1.0498221  3.2288  50 -1.386294   0 -1.38629       6     0  0.76547

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

... comparing models

```
> summary(model1)

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.669337   1.296387   0.516  0.60693
lcavol       0.587022   0.087920   6.677 2.11e-09 ***
lweight      0.454467   0.170012   2.673  0.00896 **
age         -0.019637   0.011173  -1.758  0.08229 .
lbph         0.107054   0.058449   1.832  0.07040 .
svi          0.766157   0.244309   3.136  0.00233 **
lcp         -0.105474   0.091013  -1.159  0.24964
gleason      0.045142   0.157465   0.287  0.77503
pgg45        0.004525   0.004421   1.024  0.30886
---
Residual standard error: 0.7084 on 88 degrees of freedom

F-statistic: 20.86 on 8 and 88 DF,  p-value: < 2.2e-16
```

... comparing models

```
model2 <- lm(lpsa ~ lcavol + lweight + svi + age + lbph, data = prostate)
anova(model2,model1)
Analysis of Variance Table

Model 1: lpsa ~ lcavol + lweight + svi + age + lbph
Model 2: lpsa ~ lcavol + lweight + age + lbph + svi + lcp + gleason +
    pgg45
  Res.Df    RSS Df Sum of Sq     F Pr(>F)
1     91 45.526
2     88 44.163  3    1.3625 0.905 0.4421
```

does this make sense?

Factor variables

- F-tests are used when the columns to be removed form a group
- if a covariate is a factor, i.e. categorical, then `lm` will construct a set of dummy variables as part of the model matrix
- these variables should either all be in, or all be out

in most cases

-
```
  prostate$gleason_factor <- factor(prostate$gleason)
  levels(prostate$gleason_factor)
[1] "6" "7" "8" "9"
  model3 <- lm(lpsa ~ .-gleason, data=prostate)
```

... factor variables

```
model3 <- lm(lpsa ~ .-gleason, data=prostate)
summary(model3)
> 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

```
> anova(model1,model3)
Analysis of Variance Table

Model 1: lpsa ~ lcavol + lweight + age + lbph + svi + lcp + gleason +
    pgg45
Model 2: lpsa ~ (lcavol + lweight + age + lbph + svi + lcp + gleason +
    pgg45 + gleason_factor) - gleason
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1     88 44.163
2     86 42.724  2    1.4392 1.4485 0.2406
```

... factor variables

- with designed experiments, covariates are often factors set at pre-determined levels
- see, e.g. Example 8.4 in SM

also Ch 14 in FLM-2; Ch 13 in FLM-1

- if the design is perfectly balanced, then X has orthogonal columns, and $X^{\mathrm{T}}X$ is diagonal
- so $\hat{\beta}_j$'s are uncorrelated, and hence independent (under normality assumption)

- more generally we might have $X^{\mathrm{T}}X$ block diagonal, e.g.

$$Y = X_1\beta_1 + X_2\beta_2 + \epsilon,$$

SM §8.5.3, FLM-2 2.11

$$X^{\mathrm{T}}X = \begin{pmatrix} X_1^{\mathrm{T}}X_1 & 0 \\ 0 & X_2^{\mathrm{T}}X_2 \end{pmatrix}$$

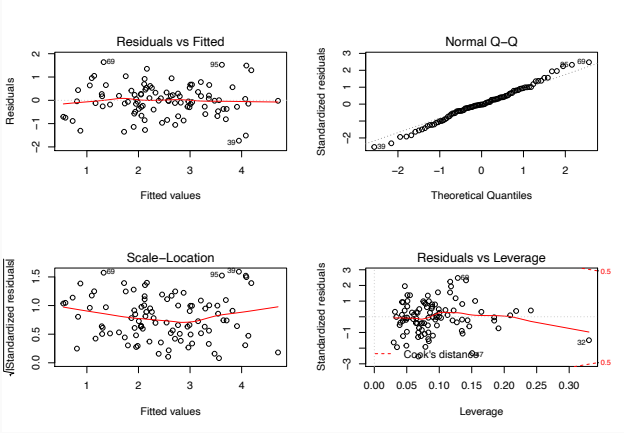
importance?

Model checking

SM 8.6, FLM-2 Ch. 6, FLM-1 Ch. 4

- assumptions on errors: $\epsilon_i \sim_{i.i.d.} N(0, \sigma^2)$ on structure $\mathbb{E}(y \mid X) = X\beta$
- normality; constant variance; independent

```
plot(model1)
```



... Model checking

- residuals: $\hat{\epsilon}_i = y_i - \hat{y}_i$
- $\mathrm{Var}(\hat{\epsilon}) = \sigma^2(I - H)$, i.e. don't all have the same variance — ntbc
- hat matrix $H = X(X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}$ $\qquad Hy = X(X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}y = X\hat{\beta} = \hat{y}$
- standardized residuals: $r_i = \dfrac{\hat{\epsilon}_i}{\tilde{\sigma}(1 - h_{ii})^{1/2}}$ — approx var 1
- Cook's distance $C_i = \dfrac{(\hat{y} - \hat{y}_{-i})^{\mathrm{T}}(\hat{y} - \hat{y}_{-i})}{p\tilde{\sigma}^2} = \dfrac{r_i^2 h_{ii}}{p(1 - h_{ii})}$ — measure of influence

https://data.library.virginia.edu/diagnostic-plots/

Collinearity

- simple model $y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \epsilon_i, \quad i = 1, \ldots n$
- if $x_1 \perp x_2$, then interpretation of β_1 and β_2 clear
- if $x_1 = x_2$ then β_1 and β_2 not separately identifiable
- usually we're somewhere in between, at least in observational studies
- may be very difficult to dis-entangle effects of correlated covariates
- example: health effects of air pollution
- measurable increase in mortality on high-pollution days
- measurable increase in mortality on high-temperature days
- high temperatures and high levels of pollutants tend to co-occur
- mathematically, $X^{\mathrm{T}}X$ is nearly singular, or at least ill-conditioned, so calculation of its inverse is subject to numerical errors
- if $p > n$ then $X^{\mathrm{T}}X$ not invertible, no LS solution

ridge, Lasso

In the News

Coronavirus outbreak

What is No 10's 'moonshot' Covid testing plan and is it feasible?

Plan to provide rapid tests for 10 million people a day would be hugely costly – and the technology does not yet exist

- Boris Johnson pinning hopes on Covid testing 'moonshot'
- Coronavirus – latest updates
- See all our coronavirus coverage

Sarah Boseley *and* Robert Booth

… In the News

ROYAL STATISTICAL SOCIETY

DATA | EVIDENCE | DECISIONS

Sir,

We are concerned that the government's leaked "moonshot" plan ("Doubts cast on Boris Johnson's 'moonshot'",10 September), to test millions of people daily for Covid-19 does not appear to take account of fundamental statistical issues. This plan goes well beyond test-and-trace, for which the statistical basis is well established, and – judging on the basis of the leaked plan – its success may require new tests to be more accurate than diagnostic tests for any other disease.

This is not to say that the approach of mass testing is not right but to build a consensus among the scientific community for this, we must first understand precisely what the government's objective is and then assess whether mass testing is the best way to achieve it.

… In the News

for any other disease.

This is not to say that the approach of mass testing is not right but to build a consensus among the scientific community for this, we must first understand precisely what the government's objective is and then assess whether mass testing is the best way to achieve it.

There are harms associated with testing – as there are with not-testing – and before the UK decides to move towards mass-testing, the balance of these harms needs to be assessed.

Tests cause harm when they miss or wrongly diagnose cases. Our current tests have 1 and 2% false positive rates – which, when millions are being tested every day, risks causing personal and economic harm to tens of thousands of people. This problem is exacerbated if the new tests, as is likely, are less accurate than the ones used currently.

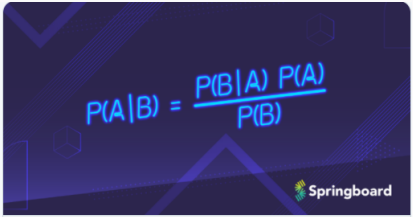
If mass-testing can give people confidence that they are disease-free, tests need to detect nearly all cases. Our current tests miss around a fifth of those with the disease – if the new tests are even less sensitive, they may not be accurate enough for the safe running of events but could be useful for complementing social distancing measures.

We urge the government to make information about the new tests and their planned use available to enable broad discussion with experts and reach consensus and understanding on the balance of risks. The Royal Statistical Society is here to provide support with the essential statistical issues.

Professor Sylvia Richardson and Professor Jon Deeks on behalf of the Royal Statistical Society Covid-19 Task Force

A shorter version of this letter appeared in the Times on 11 September 2020:
https://www.thetimes.co.uk/article/times-letters-loneliness-and-the-tender-care-of-the-elderly-

Testing for disease



Become a Data Scientist in 6 Months
springboard.com

$\longrightarrow$ R Markdown