Random reflections on data analysis and modelling

Kelvyn Jones
School of Geographical Sciences
University of Bristol

• Informal

• Arises out of questions

• Three sections
  - some organizing principles
  - some design issues (EPI)
  - developments in modelling
    (GLM → GAM → GLLAMM)

- Resources and software
  - especially open-source R
    The Comprehensive R Archive Network
    http://www.stats.bris.ac.uk/R/

    - Not…….neural nets, kriging…. 
Some Organizing Principles I:
The GEOGRAPHICAL DATAFRAME

= DATA + STRUCTURE

<table>
<thead>
<tr>
<th>pH value</th>
<th>Geology</th>
<th>Distance from scarp</th>
<th>Agricultural Practice</th>
<th>Sample</th>
<th>Instrument</th>
<th>Field</th>
</tr>
</thead>
<tbody>
<tr>
<td>5.4</td>
<td>Non-chalk</td>
<td>101.1</td>
<td>Pasture</td>
<td>1</td>
<td>I</td>
<td>A</td>
</tr>
<tr>
<td>6.9</td>
<td>Chalk</td>
<td>150.8</td>
<td>Pasture</td>
<td>2</td>
<td>II</td>
<td>A</td>
</tr>
<tr>
<td>5.7</td>
<td>Chalk</td>
<td>160.8</td>
<td>Pasture</td>
<td>3</td>
<td>II</td>
<td>A</td>
</tr>
<tr>
<td>4.3</td>
<td>Non-chalk</td>
<td>230.3</td>
<td>Tree</td>
<td>1</td>
<td>I</td>
<td>B</td>
</tr>
<tr>
<td>4.4</td>
<td>Non-chalk</td>
<td>245.8</td>
<td>Tree</td>
<td>2</td>
<td>III</td>
<td>B</td>
</tr>
<tr>
<td>7.1</td>
<td>Chalk</td>
<td>62.1</td>
<td>Arable</td>
<td>15</td>
<td>I</td>
<td>Z</td>
</tr>
</tbody>
</table>

NOTE

- DATA is **Mixture** of types of measurement
  - discrete & continuous:

- Different **Types** of variables (EPI)
  - outcome/ response
  - exposure: primary variable of causal interest
  - covariates/cofactors: need to condition on
  - structure ~ time/space /context, measurement instrument

- **Structure** needs to be taken into account in the analysis
Some Organizing Principles II: The PURPOSE of Modelling

What is the quantitative relationship between the outcome and exposure conditional on the co-factors and co-variates?

What is the effect on precipitation of 1 metre increase in altitude taken account of distance to the coast; rain-shadow or not?

**GENERAL STRUCTURE OF A MODEL**

- DATA = SIGNAL + NOISE
- DATA = SMOOTH + ROUGH
- RESPONSE = SYSTEMATIC TREND + VARIATION
- RESPONSE = FIXED + RANDOM

<table>
<thead>
<tr>
<th>Response</th>
<th>Fixed</th>
<th>Random</th>
</tr>
</thead>
<tbody>
<tr>
<td>RAIN</td>
<td>$\sim$ Alt + Shad + Dist + (Residual)</td>
<td></td>
</tr>
</tbody>
</table>

Fixed = Averages
Residual = **Distribution** summarised by a variance

**NB**
- Both aspects are important
- Possibility of “cross-contamination”
W. S. Cleveland. (1994) *The Elements of Graphing Data*. Hobart Press,

- Aspect ratio: height divided by width:
  - Top = 1
  - Automatic by Trellis graphs in R *banking to 45 degrees*
Some Organizing Principles III: The IMPORTANCE of GRAPHS

- Allow to you see general picture & detail Simultaneously

What is relation between Y and X in the 4 graphs?

In regression terms?

\[ y = 3 + 0.5x \; ; \; \text{R-sq} = 0.67; \; \text{same t and F} \]
The IMPORTANCE of GRAPHS

• Plot
  - before you model: outliers
  - after you model: comparative size of fixed effects
  - diagnostics for inappropriate model, over-influential data, case-deletion statistics (one fit!)

Fig. 5. Influential observations for deprivation social-class interactions.
Some Design Issues I
STATISTICAL POWER

• What does finding a non-significant result mean?
  - Either no-effect
  - Or study insufficiently powerful to detect effect

• The Opposite is also a problem
  - trivial effects are found to be significant
  - waste of resources in collecting ‘too much info’

Four outcomes of a hypothesis test

<table>
<thead>
<tr>
<th>Decision</th>
<th>Null Hypothesis</th>
<th>False</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fail to reject H₀</td>
<td>Correct decision</td>
<td>Type II error</td>
</tr>
<tr>
<td></td>
<td>Prob = 1-α</td>
<td>Prob = β</td>
</tr>
<tr>
<td>Reject H₀</td>
<td>Type I error</td>
<td>correct decision</td>
</tr>
<tr>
<td></td>
<td>Prob = α</td>
<td>Prob = 1-β</td>
</tr>
</tbody>
</table>

Type 1 error: When the H₀ is true & you reject it, finding significant results where there aren’t any; controlled by α: the significance level

Type II error: not rejecting the H₀ when you should have done, controlled by β, largely ignored

Power probability of identifying significant effect when one really exists, determined by 1 – β
WHAT DETERMINES POWER?

Power is increased in the following circumstances.

• little noise in the system; clear signal
• the effect is substantial
• \( \alpha \) is set leniently (0.05 and not 0.01)
• large sample size
  Can usually only do something about size

Power analysis

Prospectively: how large should my study be?
Retrospectively: when non-significant’ effect found; is there sufficient power?

Problem!

• Need to know size of effect
• Cohen’s solution for very large range of procedures
  - defines small, medium and large effects
  - as ratio of effect to variation

Cohen, J (1988) *Statistical power analysis for the behavioural sciences* Lawrence Erlbaum, New Jersey

Software

PS, freely available from:
http://www.mc.vanderbilt.edu/prevmed/ps/

Power Calculator online
http://ebook.stat.ucla.edu/calculators/powercalc/

G*Power freely available from
http://www.psycho.uni-duesseldorf.de/aap/projects/gpower/index.html;
‘compromise power’ between Type I and II errors.
### Logistic regression, test that $\zeta = 0$ for one covariate, $x$, after adjustment for prior covariates

- **Power**
  - 50
  - 60
  - 70
  - 80
  - 90
  - 100

- **Total Sample Size**
  - 0
  - 100
  - 200
  - 300
  - 400
  - 500
  - 600
  - 700

---

- **$nQuery Advisor®$ industry standard**

- **Binary outcome**: Dead or alive
- **Binary exposure**: Polluted or not
- **Effect size**: odds of 2; exposure doubles risk of outcome
- **$\alpha$**: two-sided 0.05
- **Confounding**: correlation between exposure and covariates linked to both exposure and outcome
  - $R^2$ between exposure & covariates
  - (0, 0.2, 0.5, 0.7)

- **Results**: No confounding: 111 cases for power of 80%
  - $R^2$ is 0.75, need 443 cases

- **NB**: diminishing returns of increasing size
Some Design Issues II

EFFICIENCY

• For a given sample size can I make stronger inferences?

• Choose design: rare outcomes: case comparison design
  rare exposure: prospective design

Case-comparison design

Comparison of exposure rates between cases and controls provides estimate of effect (usually odds ratio [OR])


• Sample on outcomes: Case with ‘disease’
  : Comparison without

• How differed in exposure?

• Example thalidomide: mothers of limb-defect babies had been exposed to drug in comparison to ‘normal’ births

• Modelling: what is increased risk of being a case given exposure taking account of covariates?
**CASE-COMPARISONS & EFFICIENCY**

**Question:** is there any protective effect for BCG vaccination scar on leprosy (Leprosy & TB: similar bacillus)

**Study design 1: prevalence study in Central Africa (Huge!)**

<table>
<thead>
<tr>
<th>Data</th>
<th>BCG Scar</th>
<th>Cases</th>
<th>Population</th>
</tr>
</thead>
<tbody>
<tr>
<td>Present</td>
<td>101</td>
<td>46,028</td>
<td></td>
</tr>
<tr>
<td>Absent</td>
<td>159</td>
<td>34,594</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>260</td>
<td>80,622</td>
<td></td>
</tr>
</tbody>
</table>

**Results**

<table>
<thead>
<tr>
<th>Size</th>
<th>OR</th>
<th>Low 95% OR</th>
<th>High 95% OR</th>
</tr>
</thead>
<tbody>
<tr>
<td>80,622</td>
<td>0.48</td>
<td>0.37</td>
<td>0.62</td>
</tr>
</tbody>
</table>

odds of leprosy halved (0.48) in presence of BCG scar

**Study design 2:** What if case comparison design?

Drawing comparisons at random from non-cases

<table>
<thead>
<tr>
<th>BCG Scar</th>
<th>Cases</th>
<th>a) 1:1</th>
<th>b) 2:1</th>
<th>c) 5:1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Present</td>
<td>101</td>
<td>148</td>
<td>296</td>
<td>740</td>
</tr>
<tr>
<td>Absent</td>
<td>159</td>
<td>112</td>
<td>224</td>
<td>560</td>
</tr>
<tr>
<td>Total</td>
<td>260</td>
<td>260</td>
<td>520</td>
<td>1300</td>
</tr>
</tbody>
</table>

**Results**

<table>
<thead>
<tr>
<th>Size</th>
<th>OR</th>
<th>Low 95% OR</th>
<th>High 95% OR</th>
</tr>
</thead>
<tbody>
<tr>
<td>80,622</td>
<td>0.48</td>
<td>0.37</td>
<td>0.62</td>
</tr>
<tr>
<td>260</td>
<td>0.48</td>
<td>0.33</td>
<td>0.69</td>
</tr>
<tr>
<td>520</td>
<td>0.48</td>
<td>0.35</td>
<td>0.66</td>
</tr>
<tr>
<td>1300</td>
<td>0.48</td>
<td>0.36</td>
<td>0.64</td>
</tr>
</tbody>
</table>

- same result is found as when all 80,000 used;
- little gain in precision (either 1,300 or 80,000)
- C-C highly efficient for rare outcomes
Developments in modelling I
GENERAL LINEAR MODEL
- Developments on the right-hand side of the equation

- Linear model (continuous predictors, additive and linear)
  Eg  Rain ~ Alt + Dist + (Residual)
  Results <- lm(Rain ~ Alt + Dist)  R-code

  • General linear model
  1. Factors as predictors (qualitative states)
     Rain ~ Alt + Dist + SHAD + (Residual)
     Rain ~ Alt + Dist + FullShad + PartShad
     + (Residual)

  2. Interactions between variables
     Rain ~ Alt + Dist * SHAD + (Residual)

  3. Non-linearity as polynomial function
     Rain ~ Alt + Poly(Dist, 3) + (Residual)
     Rain ~ Alt + Dist + Dist^2 + Dist^3 + (Residual)

  4. In combination
     Rain ~ Alt + Poly(Dist, 3)* SHAD + (Residual)
General linear model (cont)

**Dummy variable trick:** drop one of the qualitative states (here: None) and 1/0 code the others

**Results**

Precip $= 42.8 + 20.3\text{Alt} - 25.1\text{Full} - 15.8\text{Partial} - 7.20\text{F*Alt}$

$- 2.97\text{ P*Alt} - 1.91\text{Alt-sq}; \quad \text{R-Sq} = 87.9\%$
Developments in modelling II

**GENERALIZED LINEAR MODEL**

- Developments on the **left-hand side** of the equation (incorporates all that on the right)
- So far $Y$ is **continuous** but it need not be….some examples

<table>
<thead>
<tr>
<th>Type</th>
<th>Example</th>
<th>Property</th>
<th>Model</th>
</tr>
</thead>
<tbody>
<tr>
<td>Count</td>
<td>No of rain days in a year</td>
<td>Positive values; discrete count</td>
<td>Log link and Poisson or NBD distribution</td>
</tr>
<tr>
<td>Binary</td>
<td>Eroding/ Not eroding</td>
<td>Discrete qualitative states</td>
<td>Logit/Probit/clog Link &amp; binomial distribution</td>
</tr>
<tr>
<td>Ordered &amp; unordered categorical</td>
<td>Eroding, stable, deposition</td>
<td>Multiple discrete states</td>
<td>Logit and multinomial distribution</td>
</tr>
<tr>
<td>Time to event</td>
<td>Time to death</td>
<td>Positive &amp; censored</td>
<td>Cox model</td>
</tr>
<tr>
<td>Multiple states &amp; episodes</td>
<td>Single, married, divorced</td>
<td>Recurrent discrete events</td>
<td>Event–history models</td>
</tr>
</tbody>
</table>


EXAMPLE OF LOGIT MODEL: BINARY OUTCOME

- **Dataframe** (Atkinson and Massari, 1998 Susceptibility to landsliding in Central Apennines Computers and Geosciences, 373-385) KJ’s re-analysis

<table>
<thead>
<tr>
<th>Pixel(20*20m)</th>
<th>Lslide</th>
<th>Geology (7cat)</th>
<th>Dip (5 cat)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Yes</td>
<td>LAML(base)</td>
<td>5-20</td>
</tr>
<tr>
<td>2</td>
<td>No</td>
<td>CSSL</td>
<td>45-80</td>
</tr>
<tr>
<td>3</td>
<td>No</td>
<td>CSSL</td>
<td>20-45 (base)</td>
</tr>
<tr>
<td>1900</td>
<td>No</td>
<td>SCSD</td>
<td>20-45</td>
</tr>
</tbody>
</table>

- **Model**
  \[ \text{Res} \leftarrow \text{glm(Lslide } \sim \text{Geology + Dip, binomial)} \]

- **Results**

<table>
<thead>
<tr>
<th>Variable</th>
<th>Odds</th>
<th>Prob</th>
<th>No of sites</th>
<th>Chi$^2$</th>
<th>Prob</th>
</tr>
</thead>
<tbody>
<tr>
<td>Geology</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>LAML</em></td>
<td>1.0</td>
<td></td>
<td>249</td>
<td>46.42</td>
<td>0.00</td>
</tr>
<tr>
<td>CSSL</td>
<td>3.52</td>
<td>0.00</td>
<td>967</td>
<td></td>
<td></td>
</tr>
<tr>
<td>LIMB</td>
<td>4.18</td>
<td>0.04</td>
<td>18</td>
<td></td>
<td></td>
</tr>
<tr>
<td>LMTL</td>
<td>1.00</td>
<td>0.99</td>
<td>183</td>
<td></td>
<td></td>
</tr>
<tr>
<td>MASC</td>
<td>0.42</td>
<td>0.11</td>
<td>242</td>
<td></td>
<td></td>
</tr>
<tr>
<td>MLTL</td>
<td>2.11</td>
<td>0.12</td>
<td>78</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SCSD</td>
<td>1.21</td>
<td>0.67</td>
<td>163</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Dip</td>
<td></td>
<td></td>
<td></td>
<td>4.12</td>
<td>0.39</td>
</tr>
<tr>
<td>20-45</td>
<td>1.0</td>
<td></td>
<td>1127</td>
<td></td>
<td></td>
</tr>
<tr>
<td>45-80</td>
<td>1.30</td>
<td>0.23</td>
<td>262</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5-20</td>
<td>1.43</td>
<td>0.26</td>
<td>177</td>
<td></td>
<td></td>
</tr>
<tr>
<td>O’turned</td>
<td>0.90</td>
<td>0.65</td>
<td>285</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Vertical</td>
<td>1.52</td>
<td>0.26</td>
<td>49</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Geology is important Clay and sandstone 3.5, Limestone with marls below 4.2 in Cf Layered & Massive Limestone 1.
Developments in modelling III

GENERALIZED ADDITIVE MODEL

- **Relaxes** the non-linearity assumption
- Much more **flexible** form than a polynomial
- **Data-driven** via tuning parameter
  - do different degrees of roughness improve the fit?
  - Smoothest possible = straight line


**Model:**
Res <-gam(Wood ~s(band1,5) +s(band2,5)+.., binomial)
Plot.gam(Res)

**Results**

![Graphs and tables showing results from a Generalized Additive Model analysis.]
Developments in modelling IV
TREE REGRESSION

Origins
- AID
- Data mining especially marketing
- Non-additive and non-linear behaviour captured

Method
- recursively splitting on the predictors to maximize being able to predict the outcome

Example 1: classification tree
Res <- tree(risk ~ age + pressure)
Example 2: regression tree

model <- tree(ozone ~ temp + wind + rad)
plot(model, type = "u")
text(model)

EG very low ozone mean = 12.22; relatively low temperatures (<82.5); high wind (> 7.15), low levels of radiation (< 79.5)
high ozone mean of 102.4: high temp (>82.5), relatively still days (<10.6), high radiation (> 205).

Problem: the Texas sharp shooter
TREE REGRESSION (continued)
Over-fitting: modelling noise not signal

Overfitting results in the model fitting the training set too well, and in undermining the model’s performance on unseen data set (with its own, yet different peculiarities)

Cross-validation (Beirman et al)
- gauges size of tree warranted by data
- random selections are omitted and predicted from tree of the remainder

4-6splits
Developments in modelling V
MULTILEVEL MODELS

- AKA as random co-efficient modelling; hierarchical modelling, mixed modelling; highly structured stochastic systems; generalised linear latent and mixed models

- ‘statistical models as a formal framework of analysis with a complexity of structure that matches the system being studied;

- GLM models averages (fixed); multilevel models averages AND variances (random)

- Complexity = ‘dependencies’ + ‘missingness’

- Dependencies due to context (space & time), design (eg clustered sample); measurement

- Missingness: not completely balanced data

- Some problems cast as multilevel models
  - unit diagrams
  - classification diagrams
CONCLUSIONS

Realistically complex modelling:

In COMBINATION
Complex dependencies due to structures + different types of responses + different types of predictors + relaxing parametric assumptions + diagnostics

Increasingly: Bayesian-inspired engine for calibration
MCMC
General implementation in R
Specialist software
(MLwiN, BUGS, GeoBugs, Bayes-X)

BUT KISS & caveat emptor;
NB experience with techniques in its infancy (Gam onwards!)

NB All based on assumptions

“in my experience it is scientists themselves who are the keenest purveyors of statistical snake oil
Some Design Issues II
SENSITIVITY & SPECIFICITY

Synopsis
The aim of this talk is to cover in an informal way some recent and not so recent developments in data analysis. It arises out of questions from physical geography staff and postgraduates; usually after they have collected their data. Consequently I will first talk about research designs especially statistical power, sensitivity, specificity and efficiency. The second half of the talk is about modelling and covers the journey from general linear model to generalised linear model to generalised linear latent and mixed models, taking in generalized additive models and tree regression on the way! I point to appropriate software tools; and given our financial exigencies these will be open-source wherever possible.