Statistical Perspectives on Data Mining – Summary of Modules

John Maindonald

December 23, 2006

Overview of Lectures and Laboratories

Lectures and laboratories will

- Introduce the R System
  R will be used for (almost) all computations.
- Identify statistical ideas and issues that are important for all data analysts.
- Develop data analysis skills.
- Give experience with data analysis strategies and methodologies.
  - Demonstrate classical statistical methodologies
    (Linear models and extensions are especially important).
  - Demonstrate "data mining" & related "new" methodologies

Overview of Lectures and Laboratories

Lectures and laboratories will

- Introduce the R System
  R will be used for (almost) all computations.
- Identify statistical ideas and issues that are important for all data analysts.
- Develop data analysis skills.
- Give experience with data analysis strategies and methodologies.
  - Demonstrate classical statistical methodologies
    (Linear models and extensions are especially important).
  - Demonstrate "data mining" & related "new" methodologies
- Note connections and differences between the different methodologies. Note advantages/disadvantages of each.

Key "statistical" concerns

- Ask the "right" questions, i.e. ask questions that relate to intended/likely use(s) of results.
- Different uses may require different analyses and/or different adapations of analysis results
- Some desired uses may require different data!
- These issues affect predictive generalization. Specifically:
  - What analysis, if any, will support the intended predictions?
  - There are (at least) two models – one for the data, and one for the intended predictive use of the data. They may or may not be the same model!
- Be aware of common traps, lest results are misleading or worthless.

Course Overview – Handouts

- The R System – An Introduction and Overview – 88pp. **r-dm.pdf**
- Statistical Perspectives on Data Mining (SPDM) – 49pp. **statnotes.pdf**
- 13 sets of laboratory exercises – currently 71pp. in total. **r-exercises.pdf** (We will use a selection of these.)

The R System:

Laboratories I – III

- R is currently the environment of choice for
  - specialists who are implementing new methodology
  - highly trained professional data analysts.
- It is designed for interactive data analysis: the next step may depend on the previous result

See *Rintro*; check out `http://cran.r-project.org`

The R System:

Laboratories I – III

- R is currently the environment of choice for
  - specialists who are implementing new methodology
  - highly trained professional data analysts.
- It is designed for interactive data analysis: the next step may depend on the previous result
- It can be remarkably efficient, even though:
  - data resides (mostly) in memory
  - it is an interpreted language (but one command may start a lengthy computation)

See *Rintro*; check out `http://cran.r-project.org`

The R System:

Laboratories I – III

- R is currently the environment of choice for
  - specialists who are implementing new methodology
  - highly trained professional data analysts.
- It is designed for interactive data analysis: the next step may depend on the previous result
- It can be remarkably efficient, even though:
  - data resides (mostly) in memory
  - it is an interpreted language (but one command may start a lengthy computation)
- New releases every few months bring improvements & new features.

See *Rintro*; check out `http://cran.r-project.org`

Mainly, read and work through *Rintro*.
Overheads that introduce R are in preparation. For now:

Simple R

Type following >, which is the command prompt.

```
> 2+2
[1] 4
>
```

The [1] says, perhaps a little strangely,
"first requested element will follow"

Demonstrations

```
demo(graphics)    # Gives graphics demonstrations
demo()            # List all available demonstrations
```

Data Mining versus Statistics

SPDM, Section 1.1

What is data mining

- Statistics at Scale and Speed (Daryl Pregibon)
- More cynical definitions are possible!
  - A combination of large databases and bad statistics
  - Statistics plus marketing
- The next slide gives another perspective

For statistics at Scale and Speed

- Scale and speed make automation essential.
- The skill lies in judging
  - what to automate
  - when to call on the skill of the human expert
  - in using tabular and graphical summaries to call attention to features of the data that might not otherwise be obvious

Statistical Ideas and Issues

SPDM, Section 1.1

Data Analysis – What has changed in the past two decades?

- Large datasets (from automation; merging of databases)
- New types of data (e.g., images, web pages, sound tracks)
- There are new vastly better data analysis tools
  (Computing and statistical theory have advanced hand in hand)
- The new methods and software often allow:
  - Analyses that better reflect the scientific questions
  - Analyses that extract more of the information
  - Meta-analysis across multiple studies, yielding information not available from individual studies.
- Traps for the unwary are as serious as ever.
  Watch this space!

Data mining encompasses, broadly, the first three items above.

Statistical Ideas and Issues

SPDM, Sections 1.2 and 1.3

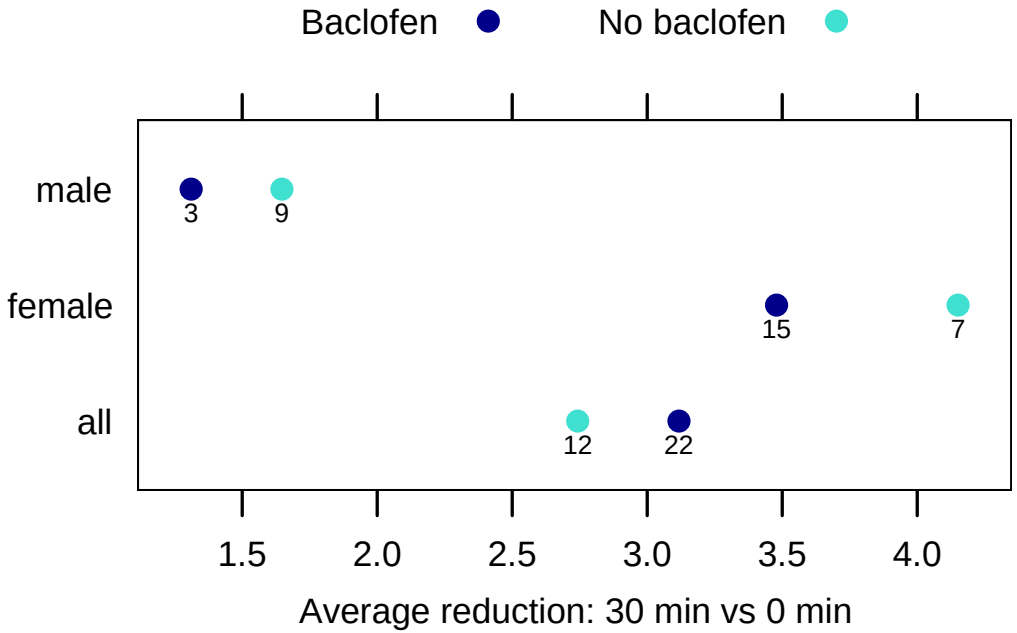
Questions

- Why am I undertaking this investigation?
- What is the intended use of results?
  What is the target population? Will results be used under the same conditions that generated the data used to fit the model? If not, are the differences fatal?
- What limitations in the data, from manner of collection or from incompleteness, may constrain the intended use?
  Issue: How were the data obtained (sampled)?

Here, nothing has changed in recent decades!
Such questions are as important as ever, even more for observational than for experimental data.

Unequal subgroup weights – traps

Laboratory VIII; *SPDM*, Section 3



Does baclofen, following operation (additional to earlier painkiller), reduce pain?

Unequal subgroup numbers spell trouble!

Subgroup numbers, shown below each point in the graph, weight the overall averages when sex is ignored.

Baclofen: 15f to 3m, i.e. $\frac{15}{18}$ to $\frac{3}{18}$ (a little less than f average)

No baclofen: 7f to 9m, i.e. $\frac{7}{16}$ to $\frac{9}{16}$ ($\approx \frac{1}{2}$-way between m & f)

Unequal subgroup weights – a strategy

Laboratory VIII; *SPDM*, Section 3

Simple approach

Calculate means for each subgroup separately.
Overall treatment effect is average of subgroup differences.
Effect of baclofen (reduction in pain score from time 0) is:

Females: 3.479 - 4.151 = -0.672 (-ve, *Rightarrow* an increase)

Males: 1.311 - 1.647 = -0.336

Average of m and f = -0.5 $\times$ (0.672+0.336) = -0.504

Fit a model that accounts for sex and baclofen effects

y = overall mean + sex effect + baclofen effect + interaction
(At this point, we are not including an error term).

Why specify a model?

It makes assumptions explicit. More anon!

When/how can models untangle subgroup effects?

Laboratory VIII; *SPDM*, Section 3

An Example – Is moderate wine-drinking good for the heart?

Many factors affect the risk of a heart attack – genetic, lifestyle, diet, etc. Modeling will be effective only if:

- Data include all factors that might inflence risk
- The response is correctly modeled (for important variables, moderation may be better than either extreme)
- The model correctly accounts for interactions between factors.

Here, it may be impossible to disentangle subgroup effects (but several papers claim to demonstrate an effect.)

See Jackson et al. (2005). (reference in *SPDM*)

Conditions for effective modeling

Laboratory VIII; *SPDM*, Section 3

Modeling is likely to be effective if

- One or two factors have a large and dominant effect
  e.g., health effects of smoking or of exposure to asbestos
- OR, a theoretical equation explains most variation, e.g.
  book weight = constant $\times$ height $\times$ breadth $\times$ thickness
- OR (best of all), data are from a carefully designed experiment.

Example – hormone replacement therapy (HRT)

- Initial data were observational (many factors, no good guidance from theory)
- More recently, there've been randomized controlled trials.

The experimental data gives definitive (& different) results.

Subgroup Effects – tables of counts

Laboratory VIII; *SPDM*, Section 3

Example with contrived data

	Engineering		Sociology		Total	
	Female	Male	Female	Male	Female	Male
Admit	10	30	30	15	*40*	*45*
Deny	10	30	10	5	*20*	*35*

Admission rate calculations – the weights

Females: $\frac{10}{20} \times \frac{20}{60} + \frac{30}{40} \times \frac{40}{60}$ [0.33 (Eng) : 0.66 (Soc)]

Males: $\frac{30}{60} \times \frac{60}{80} + \frac{15}{20} \times \frac{20}{80}$ [0.75 (Eng) : 0.25 (Soc)]

The Overall Rates

- females ($\frac{2}{3}$): bias is towards the Sociology rate (0.75)
- males ($\frac{45}{80}$): bias is towards the Engineering rate (0.5).

Models – the two parts of a model

SPDM, Section 2, esp. Subsection 2.2, and Sections 14 and 15

Models have fixed and random parts. Common aims are:

- Estimate the fixed part(s)
- Assess implications, for accuracy, of the random part(s).

Three levels of sophistication in accounting for random error

1. Ignore the random part (i.e., deterministic modeling)
2. Independently & identically distributed (i.i.d.) errors
3. Assume a "complex" error structure
   (may be essential if there is > 1 level of random variation)

How much sophistication is required?

Use a simplified model if it serves the purpose.
(but careful modeling of the random part may be needed to show that a simplified model is OK!)

Models – their importance for data analysts

SPDM, Section 2

There is not one model, but (at least) three!

- Data were sampled from a source population?
  Was the sampling, effectively, random? Is bias an issue?
- A model must be assumed, for purposes of analysis,
  e.g. Depression in lawn is proportional to roller weight.
  Results perhaps apply to the lawn(s) in the sample.
- How will results be used? Assuming use for prediction:
  Are target and source populations essentially the same?
  Are predictions for the sampled lawn(s), or for new lawns?

Worst Case Scenario

Data source and target population are totally different, e.g.
Data on 2006 oil prices will give hopeless predictions for 2008!

Models – their importance for data analysts

SPDM, Section 2

Alternatively, the model might describe the total process

- For analysis a model is used that purports to describe the processes (sampling & other) that generated the data.
  e.g. Depression in lawn is proportional to roller weight.
  It is now harder to say what is meant by bias.
- For prediction, there must be a model that describes the processes that will apply when predictions are made.
  Will the model that was used for analysis do the job?

Worst Case Scenario

Is it possible to create a model for 2006 oil prices that will still be usable in 2008?

Choosing and Comparing Models

Possible strategies

- Ensure that the model fits well (diagnostic checks, etc.)
  Predictions should then be accurate and robust.
  Accurate and robust for what range of target populations?
- Aim directly for accurate prediction (a data mining emphasis?) What population is in mind?

Commentary

- Predictive accuracy will vary with the target population.
- Often, good stats properties $\Rightarrow$ acceptable accuracy, for different possible different target populations.
- Tuning predictions to a specific target can be risky.
  Times change, populations also change!

Assessment of Predictive Accuracy

Laboratories X and XI will demonstrate cross-validation

Two Empirical Methods (both assume one model does all!)

- The training/test approach can be used when source and target populations differ.
  Requires a sample from target population.
  But what if, e.g., the target is a future population.
  e.g. train a model to detect computer intrusions.
- Cross-validation is popular and effective.
  It assumes equivalence of source and target populations.

Cross-validation uses multiple repeats of a training/test split, with a new training/test split at each repeat (see next slide).

Theoretical Assessments

- Classical statistical theory may give accuracy estimates.
  Such estimates may be highly sensitive to assumptions.

Test/training Set and Cross-Validation

Steps

- Split data into k parts (below,k=4)
- At the ith repeat or *fold* ($i = 1, \ldots k$) use:
  the ith part for *testing*, the other k-1 parts for *training*.
- Combine the performance estimates from the k folds.

Random Error in Models – Examples

SPDM, Section 2

Example – Distance fallen under gravity $d = 0.5gt^2$

- There is random error
  - g is not quite constant as the object falls
  - there are fixed & random effects from air resistance
  - there is measurement error
- Often, the small random error can be ignored.

Corn yields on Antigua (in the Caribbean)

8 sites; 4 blocks per site. Sources of variation are:

- Within sites – the relevant error for prediction to a new block on one of the 8 sites
- Between sites – need to a/c for this, for prediction to a new block on a new site.

As complex error structures go, the error here is simple!

Distributions for the random part

Note in particular:

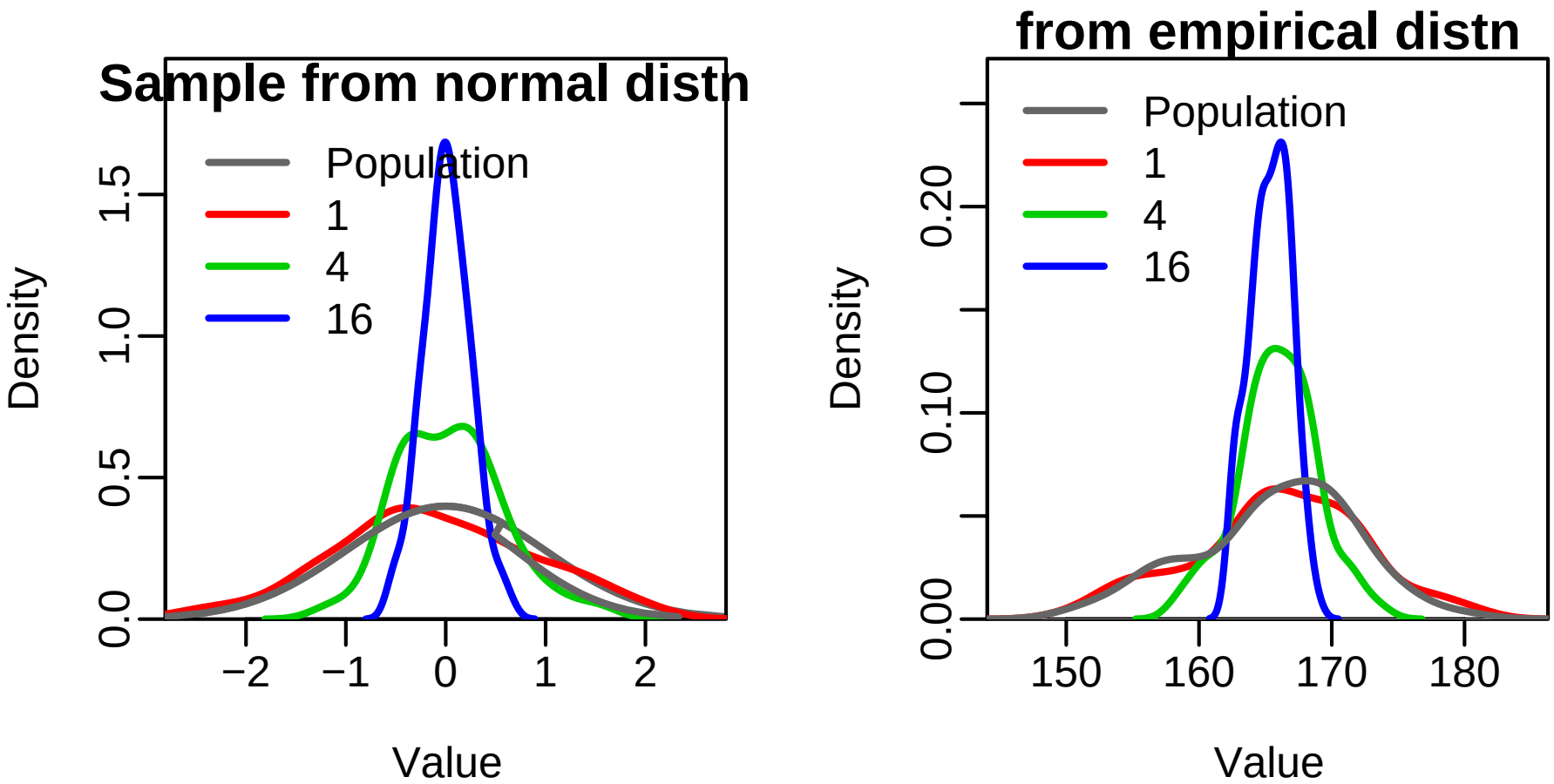
- the normal distribution
- the empirical data distribution
  (treat the sample data as a microcosm of the population)
- the distributions covered in laboratories III and IV are often used as building blocks for "complex" distributions.

When/where are "complex" distributions important?

- There may be > 1 relevant source of random variation.
- Time series.
- Other examples abound, but this will do for now.

The Sampling Distribution of the Mean

Laboratory V; *SPDM*, Part II



Numbers (1, 4, 16) in the legend are sample sizes.
The right panel used heights of female Adelaide U students.

Sampling Distributions – their Use for Inference

Laboratory V; *SPDM*, Part II

The sampling distribution of the mean

There is (usually) a single sample of data, & a single mean. The sampling distribution estimates the distribution of the mean under repeated sampling.

Estimating the sampling distribution

Alternatives are:

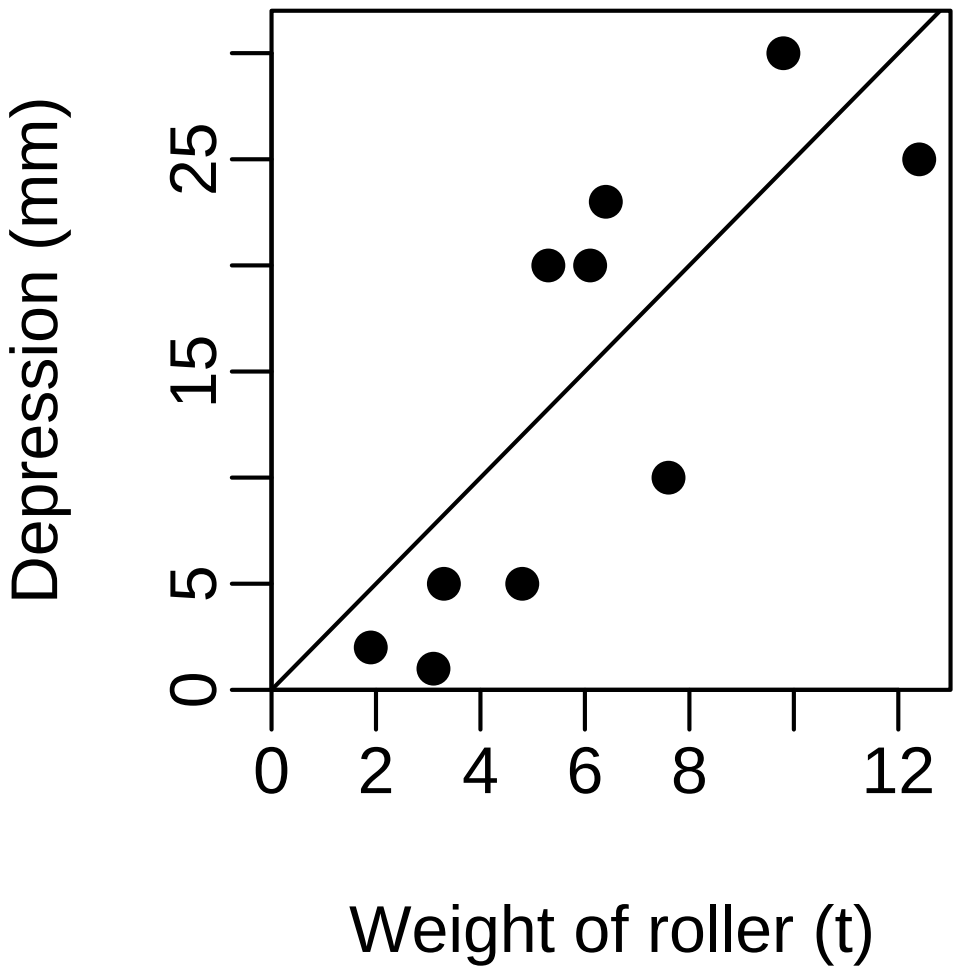
- Use theory, usually making i.i.d. assumptions
  Inference is then based on the t-distribution;
- Base inference on the empirically estimated distribution under repeated resampling (the bootstrap method).
  [Most useful for statistics other than the mean.]

Further discussion is deferred till later

Linear Models - a simple example

Laboratory VI; *SPDM*, Section 9

Straightline model: $y_i = bx_i + e_i \quad (i = 1, 2, \ldots n)$



Model in the graph:

depression = $b \times$ weight + error

The line has been forced through the origin. Is this desirable?

Fitting the model

Typically, use least squares, i.e., choose b to minimize $\sum e_i^2$.

Linear Models - Why?

Laboratories VI and VII; *SPDM*, Sections 9–13

Start discussion of models with linear models because:

- the statistical theory is well understood;
- linear models are widely used;
- linear models are simple in concept, and easy to fit;
- ideas that will be explained in this context are important also for other models, including "data mining" models;
- non-linear models are often built together, in part, from linear models.

Multiple Linear Regression

Laboratory VII; *SPDM*, Sections 10–13

Multiple explanatory variables

- Adding more explanatory variables is easy, e.g.
  $\log(\text{bookweight}) =$
  $a + b_1 \times \log(\text{area of cover}) + b_2 \times \log(\text{thickness}) + \text{error}$
- Knowing what results mean can be hard or impossible:
  - One variable can mask or modify the effect of another.
  - It is often impossible to be sure that all relevant effects have been accounted for.
  - There may be biases in data collection.
  - Some variables may be so inaccurately measured that their effects are undetectable.

Practical use of multiple regression, to give results that carry conviction, can be (in many contexts, is) challenging!

The Wide Reach of Linear Models

Laboratory VII; *SPDM*, Sections 10–13

Linear models can do much, much more!

- Qualitative effects can be modeled
- Effects can be smooth curves (use regression splines)

These extensions use the same linear modeling theory.

Linear components in non-linear models

Non-linear models may have "linear" component(s).
Model-fitting algorithms can take advantage of this.

Models for Classification & Discrimination

Laboratories X and XI – insert a graph that shows a 2-dim discriminant line

There are many different types of discriminant model

- "Classical" linear discriminants (NB: "non-linear" uses)
- "Algorithmic' (e.g., tree-based, neural nets)
  (Data miners have often preferred algorithmic models.)

Comparison

- Classical models have a well-developed statistical theory.
  (accuracy estimates may be approximate, but adequate?).
- Classical models may allow mechanistic interpretations.
- Algorithmic models may give more directly usable results.
  (e.g., a simple diagnostic rule)
- Non-classical may outperform classical models when there is a complex pattern of effects, and data are extensive.