Getting past ‘significance’ testing: good description & inference via modelling

Mark Haggard

mph38@cam.ac.uk

Department of Psychology, University of Cambridge

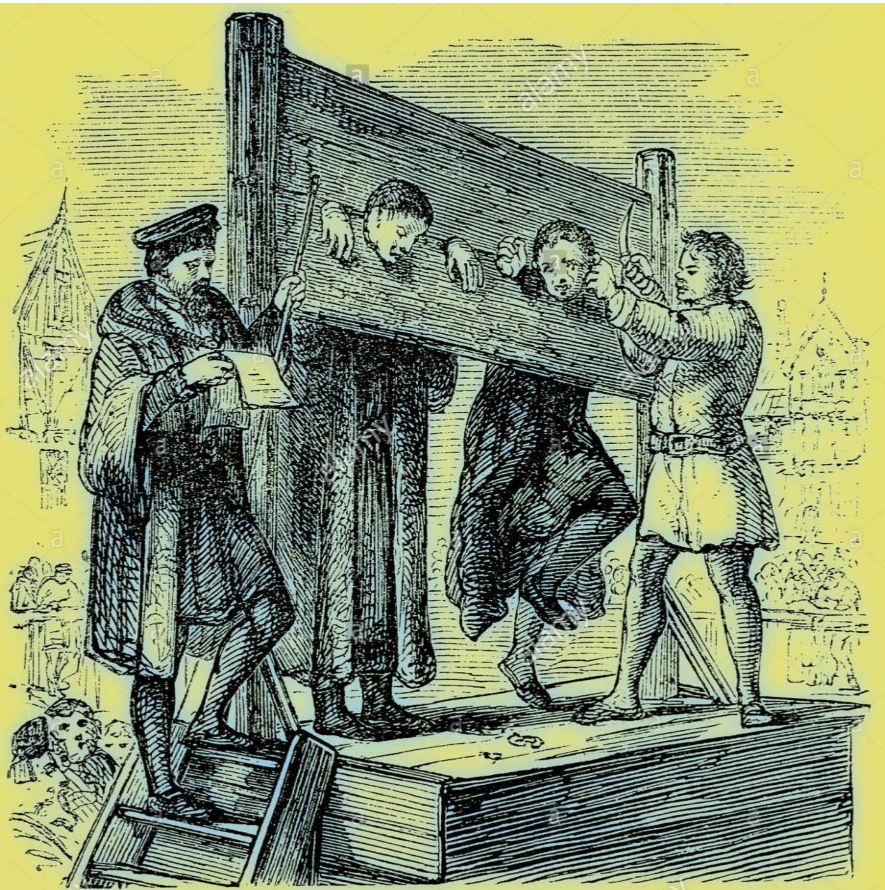
Graduates Methods Course, Lent 2022
Class 2 Jan 27

It is worth understanding enough statistics to ensure you are not shamee for mistakes

- ….even if the paper gets published through reviewers not being alert to issues (especially if ??)..
- ….even if the overall conclusion is not wrong or misleading
- ….even if ‘doing it right’ requires getting some expert help
- Checklists to preclude embarrassment can include more than citation of precedents for analysis

It is more about showing that main expectable sources of error have been identified and minimised.

Known but still frequent howlers include inflating *df* X2, by making analysis unit eyes/ears, not individuals thus (non-independence)



‘..always good to begin with a question..’

- Benjamin (2018) & 72 other authors, some with big names (who should have known better) proposed adopting p=0.005 not 0.05 as a major solution to the Replication Crisis

- One big name (Ioannides 2018, JAMA) later commented. Read carefully, it’s a recantation: harm also done. Why?

- Task: before reading Ioannides give 3 (or more) reasons why RC is not solved by adopting p=0.005 generally

- 4 chief sets of slides in this lecture: kept together but not all talked through, as some are self-supporting interspersed sets revising concepts, terminology etc:
  - 5-8 Some more comments on power for effect sizes
  - 10-14 Levels of measurement for GLM
  - 15-22 Core principles & statistical assumptions of GLM (partly)
  - 23-36 Detailed worked example of a GLM in action
  - Various extra comments (texty slides)

Idiomatic usages in scientific English for 7 overlapping words & concepts: a hierarchy ?

- ‘Data’ can be garbage, a random bit-pattern:
  - Add structure →
- Information, commodity implying multiple sources:
  - Add study design, rational data-reduction & use of control →
- Results, essentially factoids:
  - What appears in tables, a synoptic summary of the data declared relevant. Add interpretative summary of meaning →
- Findings, essentially structured :
  - Propositional or formulated; this term adds context & some wider interpretation & generalisation. Add judgement →
- Conclusions that are likely to be valid, if restricted
  - Towards likely adoption & use or application by scientific community due to context of writer’s knowledge & wisdom →
- Knowledge
  - Publicly acknowledged or accepted & value-referenced inter-relating, ed and integrated, accessing many results →
- Wisdom (an epistemic virtue)
  - implying wide, abstract reference, the strategic avoidance of error, choosing where to seek useful, trustworthy information

Emphasising magnitude, over *p*

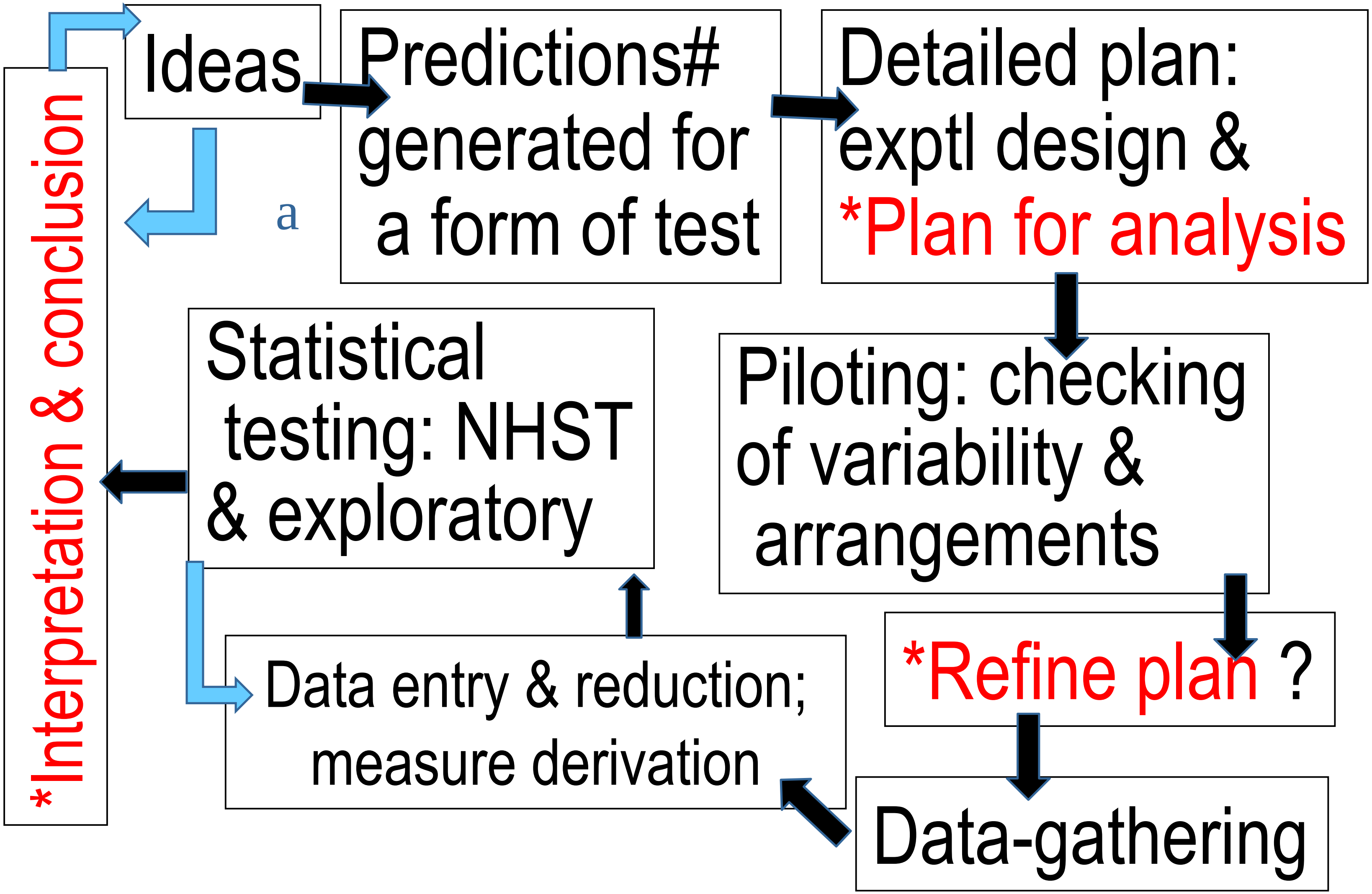
- Familiar anchors for summary discussion needed: “small”, “large” etc For other types (eg ratios, correlation, % difference, we may need to relate these to the default set for Cohen’s d
- Careful! Datasets may contain >1 form of SD to use in SDES: eg SRM for change (based on related 1-sample t-test), or change as proportion of baseline SD; value depends on sequential *r*
- Essence: standardised ES, largely* independent of N, captures magnitude not reliability for claiming non-nullity
- For wider communication with the non-numerate, it can also be expressed in terms of ranking, as “competitive leapfrogging” of salary etc eg +1 SD (d=1.0, very large effect) leap-frogs you over ~34% of the competition, if you happen to start at the median
- Also overlap: (note: other verbal anchors have also been offered)

– Small	<0.20	(>85% overlap of two distributions)
– Medium	0.50	the zone of debate (67% overlap)
– Large	>0.80	(<53% overlap) (Cohen’s ‘d’ , 1988)

- Other intraconvertible ES metrics & verbal anchor systems exist; Cohen’s (perhaps with Hedges’ correction) main one

3 stages in the scientific cycle where *power issues arise & so invoke the Effect Size

Cycle should be as unidirectional (clockwise) as possible, BUT can also include some feedback loops (turquoise) for flexibility: (a) where a re-look at literature might guide interpretation or (b) an analysis might show need for greater aggregation for power



Ideas

Predictions# generated for a form of test

Detailed plan: exptl design & *Plan for analysis

Piloting: checking of variability & arrangements

*Refine plan ?

Data-gathering

Data entry & reduction; measure derivation

Statistical testing: NHST & exploratory

*Interpretation & conclusion

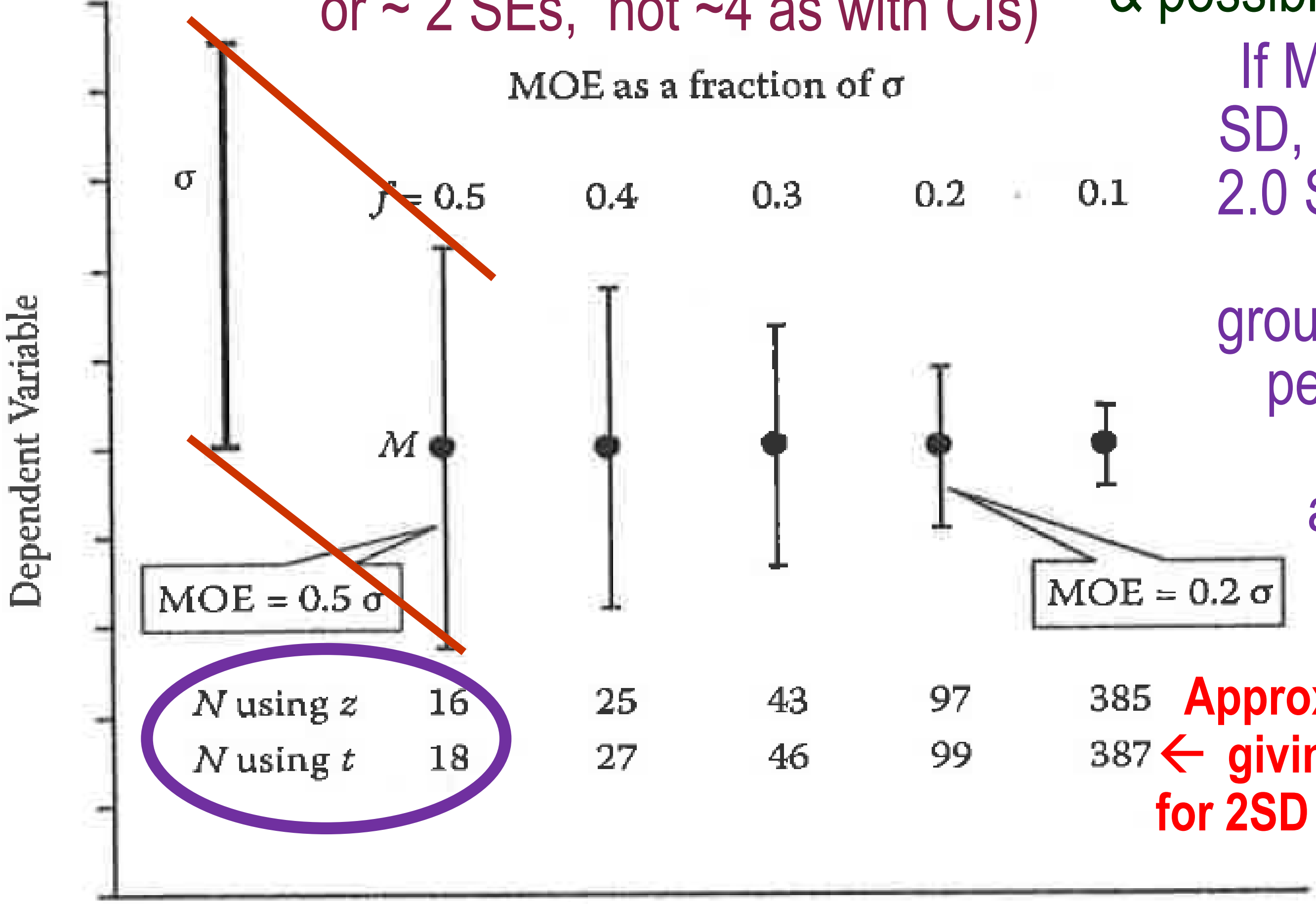
Sample-size (power) calculation can seem rigorous – but also pretentious if made over-precise, partly as the notion rests on NHST

- Obligation to postulate an explicit plausible effect size is general & important for efficiency in research
- Power level (eg 80% = 1 – False Neg rate) and acceptable False +ve rate alpha (eg 0.05) are:
  - Main convention: wish for better, when to accept ‘worse’ ?
  - Calculation is based on Neyman-Pearson logic which contains contradictions re ‘true/false’ -- not accepted by Bayesians who do not worry too much about PC
- Actual quantitative calculation mis-emphasises ‘significance’; general scenario more important than % gambles on truth-proxy. Calculation is a signpost within it
- Consideration of quantity of data to sustain a conclusion should extend beyond formal power scenario for one NHST; wider than just P for a difference, eg summarise df-ratios for stability (see later), useful narrowness of CI etc

Alternative perspective on power: sufficiently narrow CIs on estimates.

We are now back with sampling error not just magnitude; MOE,CI & SE have √N sample-Size term in, SD does not. . MOE= 0.5 one CI, or ~ 2 SEs, not ~4 as with CIs)

MOE term not strictly necessary, given CIs & possibly confusing



N using *z*	16	25	43	97	385
N using *t*	18	27	46	99	387

If MOE ~ 1.0 SD, then CI ~ 2.0 SD, so 16 is a good group size for people who can't do arithmetic

← Approx N per group giving 80% power for 2SD ES in Class 1 green table

Error reduces as sq root N of participants →

But SD differs minimally (just 'Student' t *vs* z distribution)

Ensuring the 4 beautiful virtues of GLM

(General Linear Model, ~ analysis of co-variance, multiple regression; 'general*ised*' extends term to logistic regression)

Statistical control for biased estimates and confounding, via adjustment for covariates

Examinationn of non-linearities & interactions for functional interpretatiuon eg synergism (IAs)

Informativeness of distribution of residuals for work still to be done

Comparison of effect sizes in common metric (partial eta^2) for categorical & continuous independent variables on same scale, within overall apportionment of variance

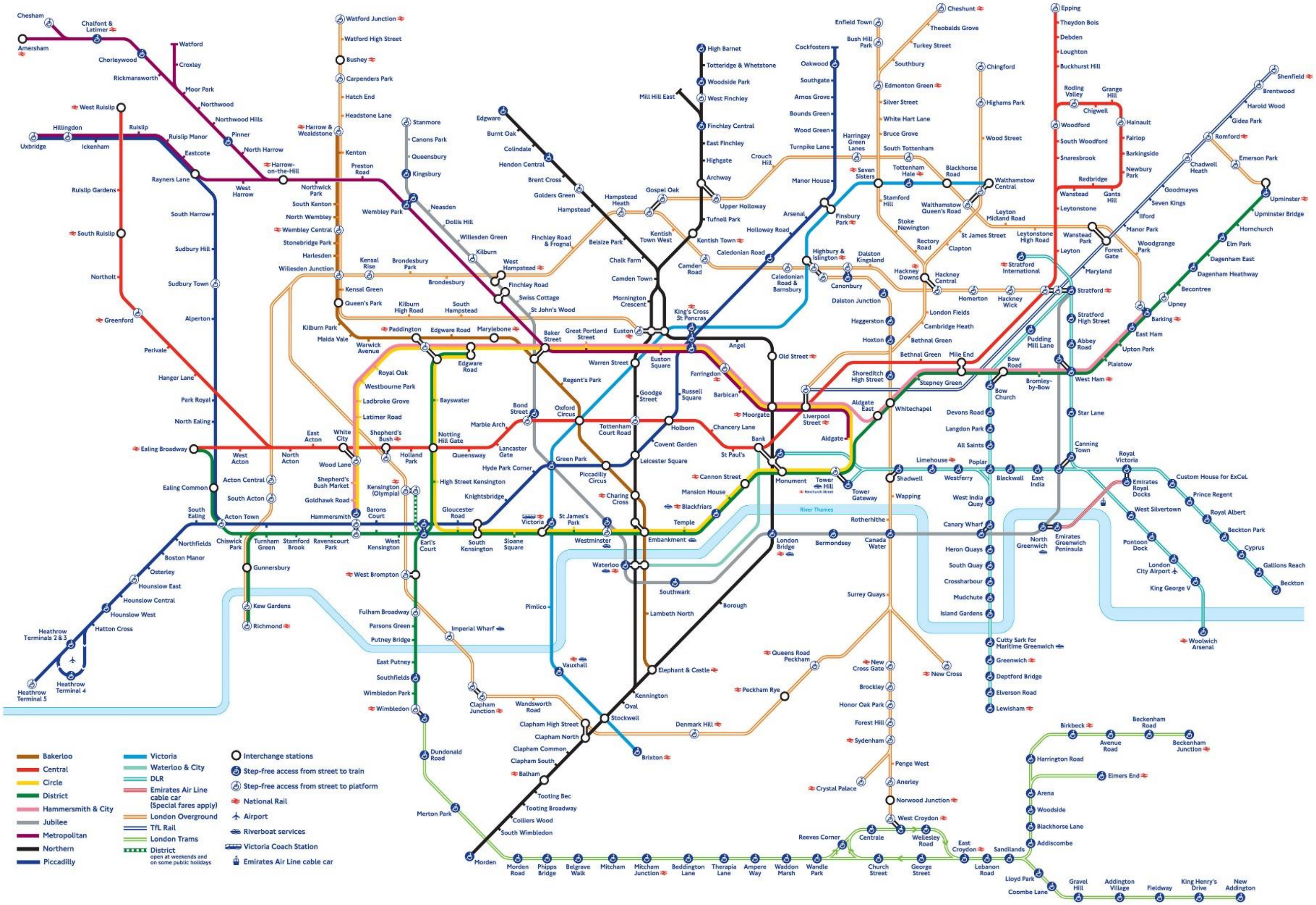
Modelling for precise adjustments to ES Issues primarily involve metrical validity

- We use them in working or daily life when need arises: seasonally adjusted employment, wind- chill factor etc
- But what do I need, to do this myself? PARAMETRIC STATISTICS
- Beware throwing away power by unnecessary degradation of level o measurement (LoM) -- not just as tests assume:
  - Interval (ratio scale even better but not used much in psych)
  - Ordered metric (cf Siegel & Castellan, re Wilcoxon Test)
  - Ordinal
  - Dichotomous but directional (See handout re dangers)
  - Nominal, unspecified categorical, not usually called ‘measurement’
- Sometimes we UP-grade the LoM (eg by aggregating data subsets)
- Generally, parametric stats require in the DV:
  - Interval-level measurement
  - Error distribution transformable to near normal (Gaussian)
  - But less important, thanks to bootstrapping (later class)

Revision: Levels of measurement

- CATEGORICAL (aka “nominal”): not really measurement as usually understood, but
  - Many category definitions stand in ordinal relationship, eg ‘often’ < ‘always’: useful so if data do bear this out, ordinal is directional so more powerful
- ORDINAL: greater/less, rankings of categories
- (EQUAL-)INTERVAL: points defined so that differences can be quantitatively compared
  - Absolute values summed and differences (Ds) can be multiplied with precise meaning
  - Integer counts (ie no decimal point) usually accepted
  - Hybrid: ‘ordered metric’ scale allows Ds to at least be ranked (eg Wilcoxon Signed-Ranks T- test)
- RATIO: meaningful zero like absolute zero for temperature at -273º C, so absolute values can be meaningfully compared (multiplied or divided)
  - Is it “better”? For what -- issue can be unanswerable
  - Exists in psychology but few applications of true 0
  - Here’s one: health status utility: death = zero

Nice example of functional *versus* metrical validity: a topological not isometric map



ARGUMENT FOR CONTINUOUS, EQUAL-INTERVAL MEASURES, ESTABLISHING SCALE PROPERTIES

If you control for a confounder by adjusting on a categorical basis (in bands) when its effect is really continuous, you will be under-adjusting; claim of having exerted statistical control will be weak.

Historical example: under-adjustment in control by categorical strata for SES effects in apparent race effect on IQ (Jensen 1969) was inadequate and this undermined his argument; degree of control was minimal so argument was false

Handling missing data is easy with categorical variables – just spend 1 more df for the variable & float the ‘missing’ level eg SES.

But this does not override the general preference for continuous. In context of missing data, which issue matters more, imputation against selection biases or precision of effect capture on those giving fuller data ? You need to state & defend your judgement.

However we have already discussed, and there are papers posted on, the loss of information with di- & tri-chotomising

Whether adjustment covariate or effect if chief interest, it is best to capture effects in their true form -- more sensitive, better adjustment for other effects

Quality of Measurement: a forgotten essential in reporting & description ?

- Largest single class of problems in my stats advice service involves sub-optimal measurement:
  - Reliability, (so power) varies with √ M items, trials etc
  - (ie as well as with √ N participants)
  - Scaling issues: resolution, range, floor & ceiling effects
  - What is sampled in measurement: ecological validity
  - Distribution is not just an issue in test legitimacy and even there is not the main issue, but it does have other forms of relevance
  - All 4 beauties of GLM require equal-interval measurement
- *How did measurement get to its present neglect* ?
- Naïve empiricism & excessive reliance on NHST
- What can I do to test, assure equal interval measurement if there is not a fundamental physical scale underlying what I measure (eg RTs, response latencies ? It may not be easy; consult me

Degrees of freedom (df) (1)

- An essential and basic concept in philosophy of science, control theory & statistical analysis – often poorly understood
- In simplest case, df = (Number of data-points - 1) because once a reference, such as the mean, is established, only (N-1) values are need to establish the relation of others to it
- In GLM, the constant expresses the grand mean, & has estimate (& CI) for being different from zero; so has 1 df, and total df→N
- Interdependences among a set of variables often entail that the true df are much lower than (N-1); this is rarely probed
- The reduction of dimensionality by Factor Analysis (SEM, PC, canonical correlation etc) exploits and delivers that lower real df
- These and related techniques summarise patterning in a correlation matrix to achieve aggregate reliability#
- Depending on the degree of investigator intervention in the data-reduction stage (DR) , it is arguable that DR may <u>conceal</u> some extra df bearing upon GFP and multiple testing adjustment (MTI)
- In certain complex modelling techniques (MLM, SEM etc) the determination of df is rather complicated and obscure

Degrees of freedom (df) (2)

- A main scientific goal is having very few degrees of freedom in the model (ie for all the effects analysed) this for parsimony….
- …..but many in the residual for reliability, stability and power
- Classical ANOVA on factorial experimental designs contained backward deletion and (re-) allocation of SOS & df:
  - In a factorial design, where higher-order IAs are not sig, you best estimate the effects and residual error by, for any NS IV, putting df & SOS into denominator
  - More generally, backwards deletion of any term from a model does this, and this is an appropriate application of NHST – this is the best thing to do, faced with the data you have
- It is widely cited as 'safe' that you only need > roughly 10 X the number of observations (or residual df) as what you have independent variables ("rule of 10"); but this is in fact marginal
- Having 25X is better & makes me feel safe. I don't even explore with <5X (avoids overfitting, instability, Type 1 & 2 errors)
- Fundamental reason is similar to growth of power with N cases in a one sample T or t-test This quantitative form of inductive inference shows no sharp cut-off at N of 10.

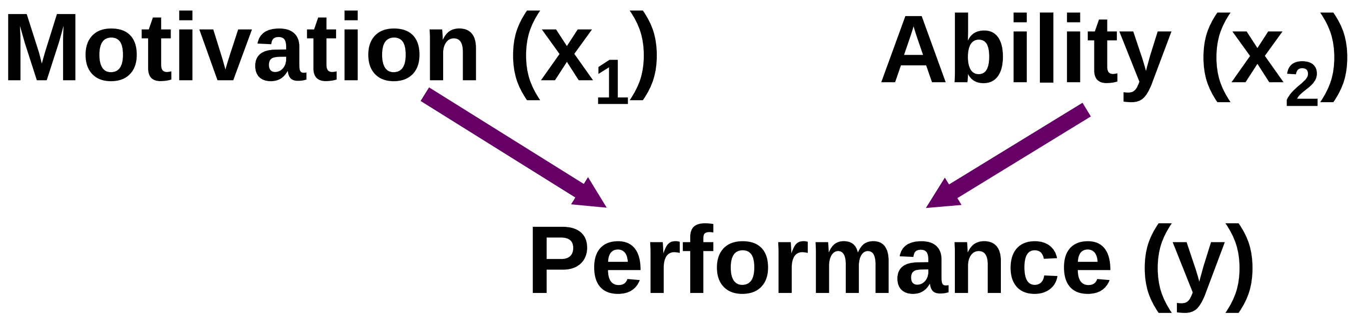
Modelling (See also Slide 40) – an alternative way of thinking to that of truth/falsity of propositions, with its own language nuances, but compatible both with conventional description & analysis of experiments

- Effect estimates in context, not raw mean difference (handles various possibilities for adjustment), enabling them to become more than a passive description of the data
- Use of coefficients to express relationships, rather than significance of some effect or raw r-value
- Significance has a place as a main but not exclusive guide to whether a variable should be in the model, not a truth proxy
- Making a more appropriate model (by error reduction and possibly thus adjusting out confounding biases) is the scientific goal and significance is only one element
- Dilemmas in model choice encourage deeper insights
- Effects of including/not-including a variable, applying a transformation can be handled quasi-experimentally
- Need for clarity on overall approach: is it largely confirmatory (hypothesis testing) or exploratory; should we use forwards stepwise or backwards deletion ?

Every measured variable is more than the measurement operation generating it: 3 types of component: representable* as a linear sum

- X1. What you should know it contains, but may forget
  - Measurement error
- X2. What you didn't realise it might contain or realised it might contain & don't want
  - Impure factor structure
  - Measurement factors: artifacts, response bias etc
  - Confounding (may overlap with two above)
- X3. What you <u>do</u> want it to contain
  - Pure operational definition of construct
- With an ingenuity similar to that used in experimental design, variance components can be separated out:
  - $Y = k + aX_1 + bX_2 + c_3$ – offers possibilities for new classes of derived variable – see late comment on 'clever' use GLM
- <u>Models</u> can either/both provide & use derived variables

Conceptualisation of many psychological models suited to GLM (ie multivariable regression)…

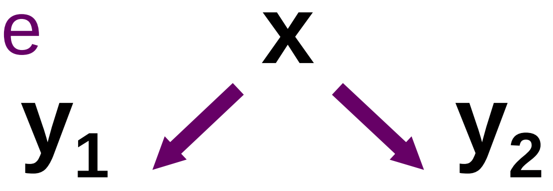


Motivation (x_1) **Ability (x_2)**

Performance (y)

$y = k + e + ax_1 + bx_2$

Bivariate GLMs (above) express many psychological theories well. Focussing on a single dependent variable: “what few things mostly cause Y ?” involves a more natural epistemological structure than a single independent variable “what list of things may X cause?” The latter, below, expresses the rarely used multivariate structure of analysis of variance (MANOVA)



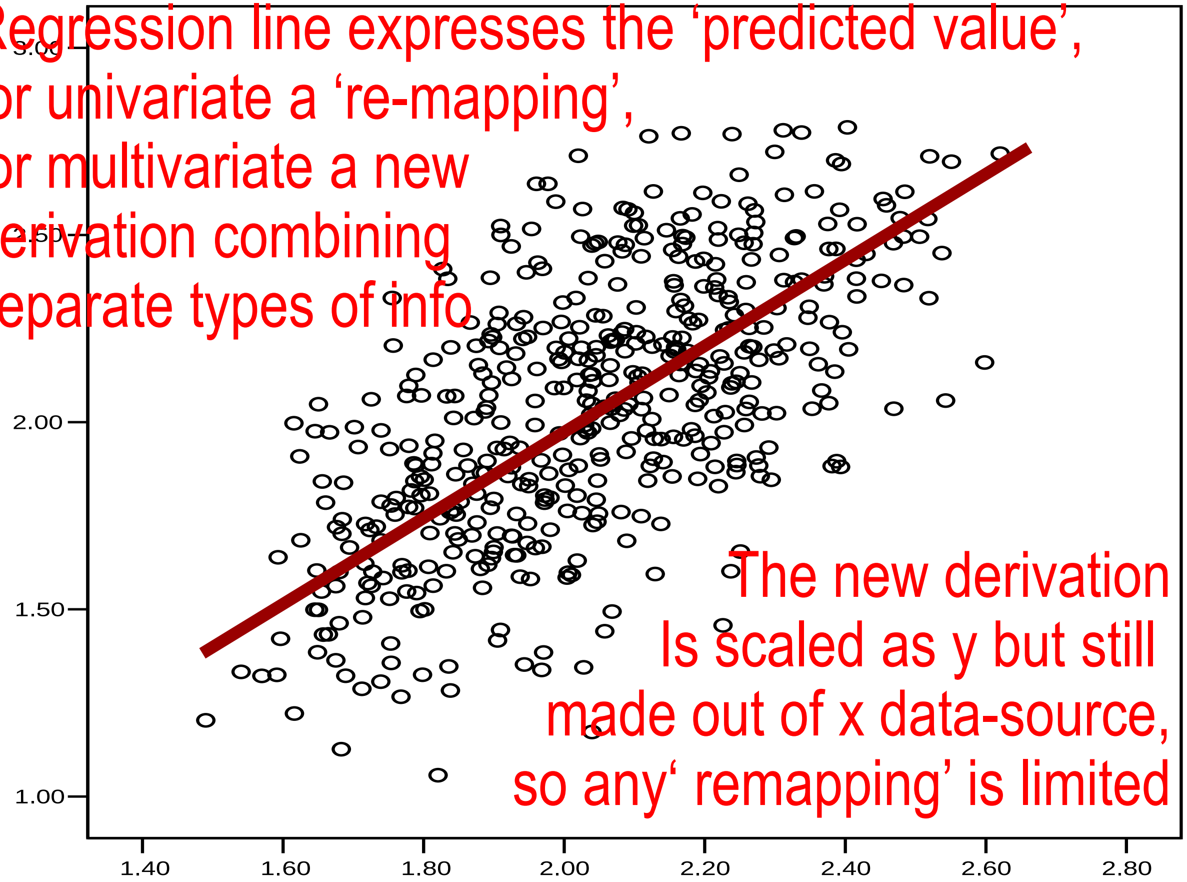
X

y_1 **y_2**

Regression: interval measurement permits transforms (for linearity), but using one may undermine = intervals

Relationship between 2 variables then has 2 aspects: linearity & strength. 100 X r^2 (Rsq) = % variance explained . Strength is b, in y=bx+c +error Predicted value = bx+c

Dependent* variable usually construed as the effect



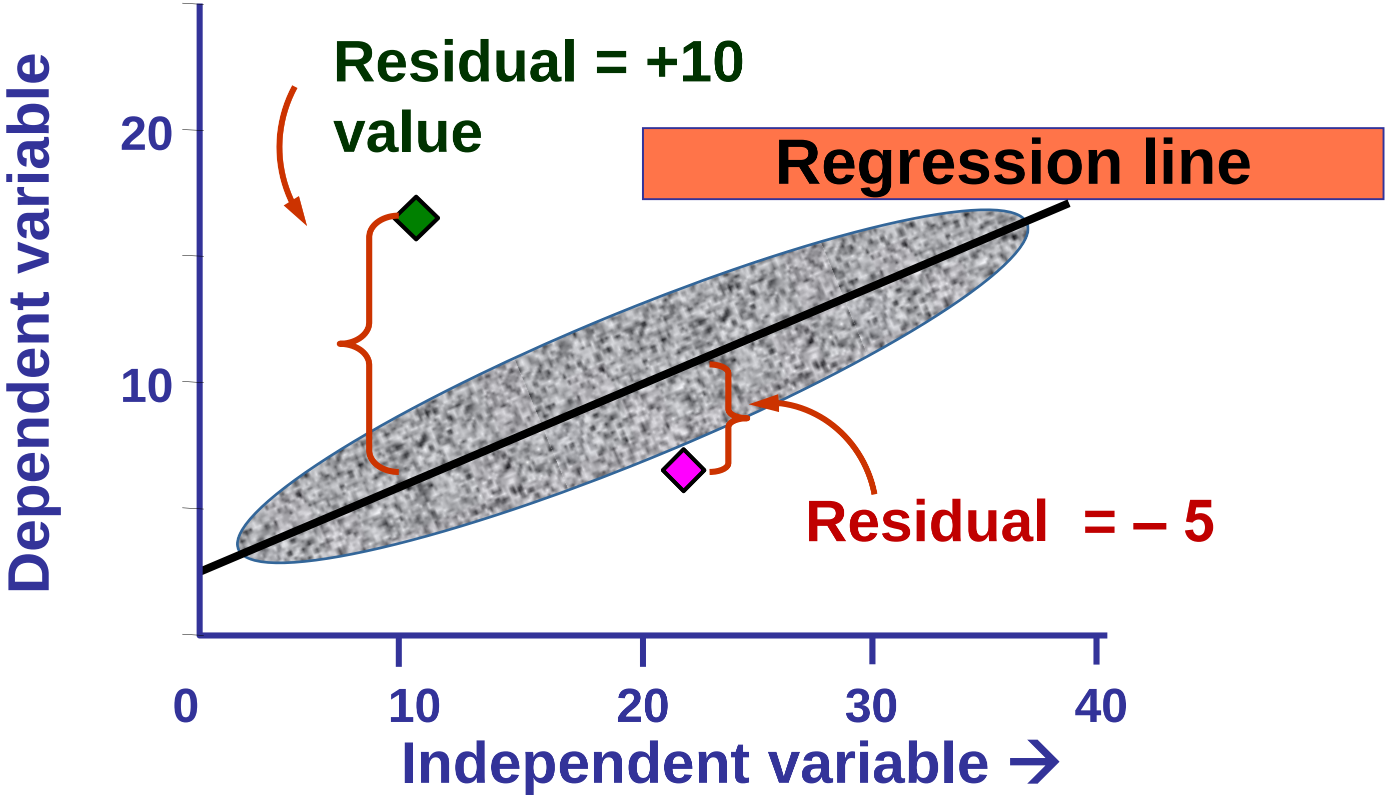
Regression line expresses the ‘predicted value’, for univariate a ‘re-mapping’, for multivariate a new derivation combining separate types of info

The new derivation Is scaled as y but still made out of x data-source, so any‘ remapping’ is limited

Independent* variable ~ what we think is probably the cause

The other crucial concept, residuals

(Discrepancies from the regression line best fitting the relation between two data variables) can be used to take out a known variance component by using it as predictor (IV)



Residuals, not necssarily raw variate, should be near-normal

Statistical control for non-experiment-able factors (categorical & continuous IVs) in GLM

- Does not easily handle within-subjects designs
- IVs entering ‘significantly’* generally also change the estimates (eg regression slope) materially & this is the main point -- not just a robustness test hoping that they do not
- We may want our estimate to come from a context summarising all the important# influences on DV
- ...*ie* in a model which we interpret in process terms, via the signs & magnitudes of each coefficient
- #Partial eta-squared ($\acute{\eta}_p^2$) allows separation of strong from weak among all IVs retained as significant, on comparable scale for both categorical & continuous
- Example is set out as forwards stepwise but, for large N & favourable df-ratio, backwards elimination is more usual#
- A single model, or modelling strategy, may combine confirmatory *and* exploratory elements; this could raise challenges in totally a priori planning and in reporting

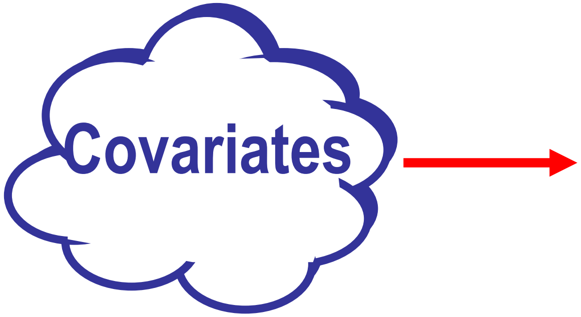
* You are forced to include a term or not, so you may use this concept here

Classes of term: passive ANCOVA approach re-expressed as equation for modelling of covariance

ANCOVA SEEN AS A FUNCTIONAL EQUATION:

Dependent variable = constant + error

+ estimates for different categorical levels of main effect(s) of interest in design



Covariates

+ continuous adjusters that there is reason to believe might influence :

(a) the dependent variable

(b) the way other main effects work (interactions)

+ interactions (usually constrained to few of a priori importance or interpretable)

Forwards stepwise or backward deletion strategy ?

- For a large sample, there is little difference in results
- With small sample or obligatory control covariates, forwards is safer.
- Forwards is naturally more *a priori* as YOU have decided the order, and it is often used 'hierarchical' for obligatory or conservative control, eg '<u>after</u> controlling for IQ, SES'
- With large sample and large M of variables, back-wards deletion (BD, elimination) is more efficient
- 'Intelligent' BD to avoid Type 2 errors: at end, bring back terms (overall or interaction) that were narrowly eliminated (eg 0.25 > p > 0.05) as they might survive in the final context once another has dropped out
- Closely related (+ve) or opposed (-ve) pairs of variables (multicollinearity) may influence one another's contribution, so in effect go in/out together, eg if it is fundamentally their difference (-ve) that is doing the work of influence

Strategy issue: what covariates to use ?

- Not a matter only of statistical significance
- In psychology we tend to think chiefly of control for otherwise uncontrolled less interesting or less specific determinants, eg ‘adjusting for general IQ’ when specific cognitive processes are of interest
- In time-structured experiments, and some other circumstances, prior commitment to **baseline scores** as a special class of covariate can be very valuable:
  - Handling within-subjects issues, & reducing error
  - Avoiding transformations, (as they have similar distribution to the DV) so offering an extra option in expressing & interpreting interactions *cf* differencing
  - Emphasising change, when variance is inhomogeneous
- Discretion & option issues always arise in choice of optimal model: we often quote more than one model, but state clear reasons for preference #

Example research Q came naturally in GLM form (kids' middle ear problems)

- We (and others) had shown that improved hearing thresholds from draining the ear and keeping fluid at bay for ~ 6-9 months when giving ventilation tubes (VTs); & that this knocked on into improved <u>development</u> (behaviour, cognition, & language)
- Adenoidectomy had been a competitor; its additive effect on hearing was smaller but longer-lasting
- So should adenoidectomy be an adjuvant (+) in children receiving surgery (hence GA) already?
  - (a) Patho-therapeutic justification in causal chain?
  - (b) If yes, then who qualifies, and what % is this?
  - (c) Does mediator variable (respiratory infections) also knock on into development? (Coherence)
- Issues (a) & (c) above were answered in related studies; analysis concentrates on issue (b)

Pseudo- (functional) regression equation summarises effects (covariance) via a set of (cumulatively) complicating model stages

Dependent Variable

Independent Variables (terms)

Post-treatment Respiratory Symptoms = Constant + Error (residual)

+ B1 x Treatment (+/-)

+ B2 X Pre-treatment baseline in same symptoms

+ *B3 X T*Baseline interaction*

+ B4 X Response bias

+ B5 X Hayfever symptoms

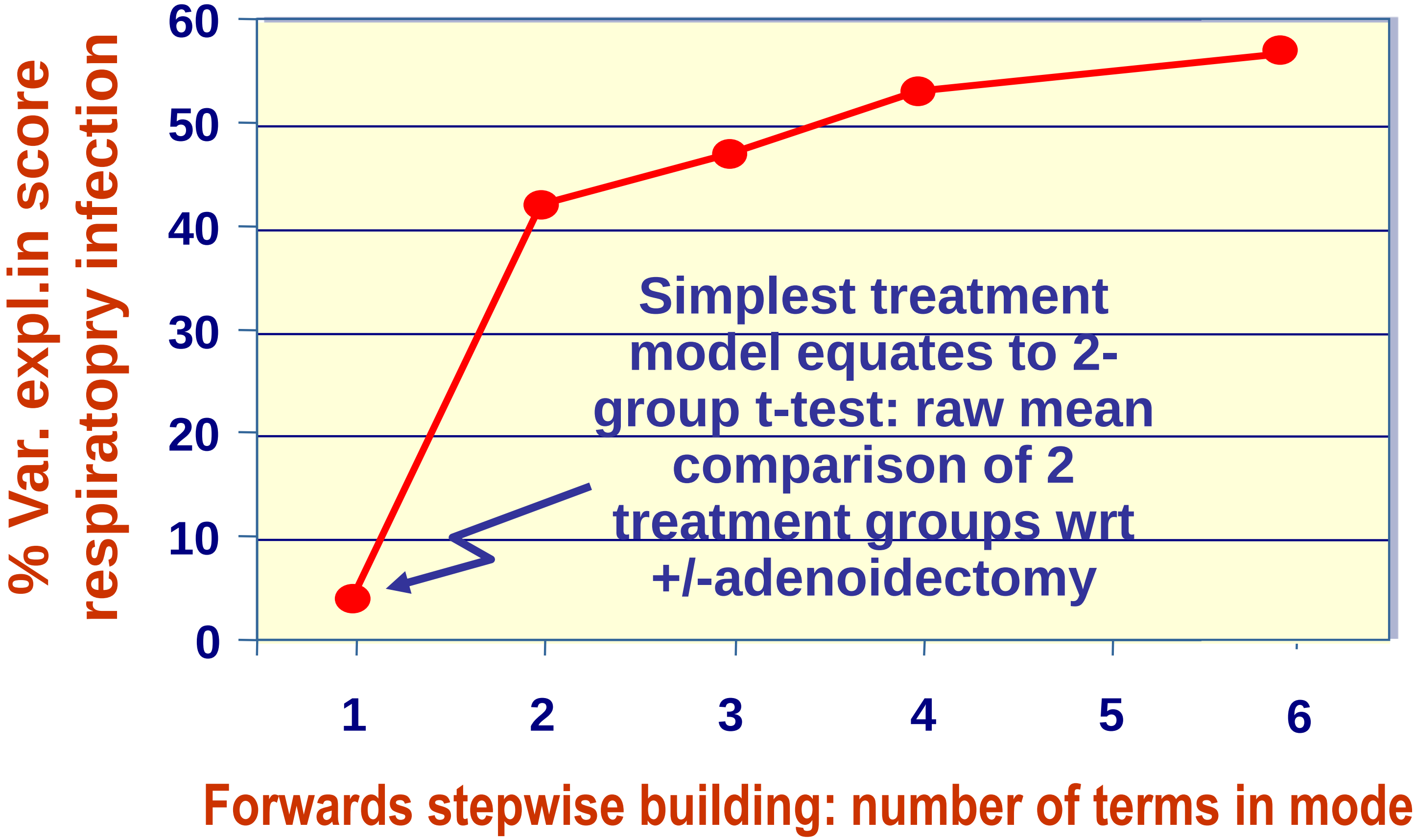
+ B6 X Socioeconomic status

2 main-effect continuous covariates

3 categorical main effects

For simple tabulation, B5 & B6 taken as pair

Motivation for programme & analysis Q is ultimately ‘psychological’ (development as DV) but has specific focus on patho-physiology to help decide who should be treated. Interaction ~ difference between differences.



Simplest treatment model equates to 2-group t-test: raw mean comparison of 2 treatment groups wrt +/-adenoidectomy

Where should we place emphasis among early *vs* late increments to cumulative Rsq ?

The foregoing 3 slides represent the approach for, planning stage, as forwards stepwise, permitting this Q to be posed

In practice after preliminaries and familiarity with results overall, it would be done as backwards deletion

The differences in Rsq depend on where you start from so are best replaced by partial eta-squared ($ή_p^2$), which de-sequences the issues

Because of the partialling, $ή_p^2$ is not strictly additive to give Rsq (for that use just $ή^2$) but it is mostly not seriously misleading to write loosely as though this were true

Generally a term significant in the model improves the model and narrows the CIs on the other terms also, except where you are overfitting. Narrowed CI = scientific goal

A priori Strategy for this analysis should state how terms would be used, what supplementary exploratory tests are allowed (+ conditionalities)

- Tests an interaction with baseline: who would benefit (most). Those with most problem in first place is the *a priori* hypothesis
- Gets a good model by statistically controlling likely sources of error and of bias:
  - in the dependent measure
  - in randomised allocation, & subsequent dropout
  - here only 2 groups of interest, both treated -- neater
- In a more complex design, more than one test may also have been *a priori*
- Exploratory analysis is not ruled out, using the same basic models; indeed that is efficient. It can be reported provided that the test for the particular effect is honestly reported as exploratory not confirmatory

Crude treatment model becomes nearer-optimal by adding appropriate further terms

(Purple entries in last 3 rows are switched from Orange for main Treatment effect to purple for the *baseline interaction, as this is of great scientific & practical interest, & units are in BL variable

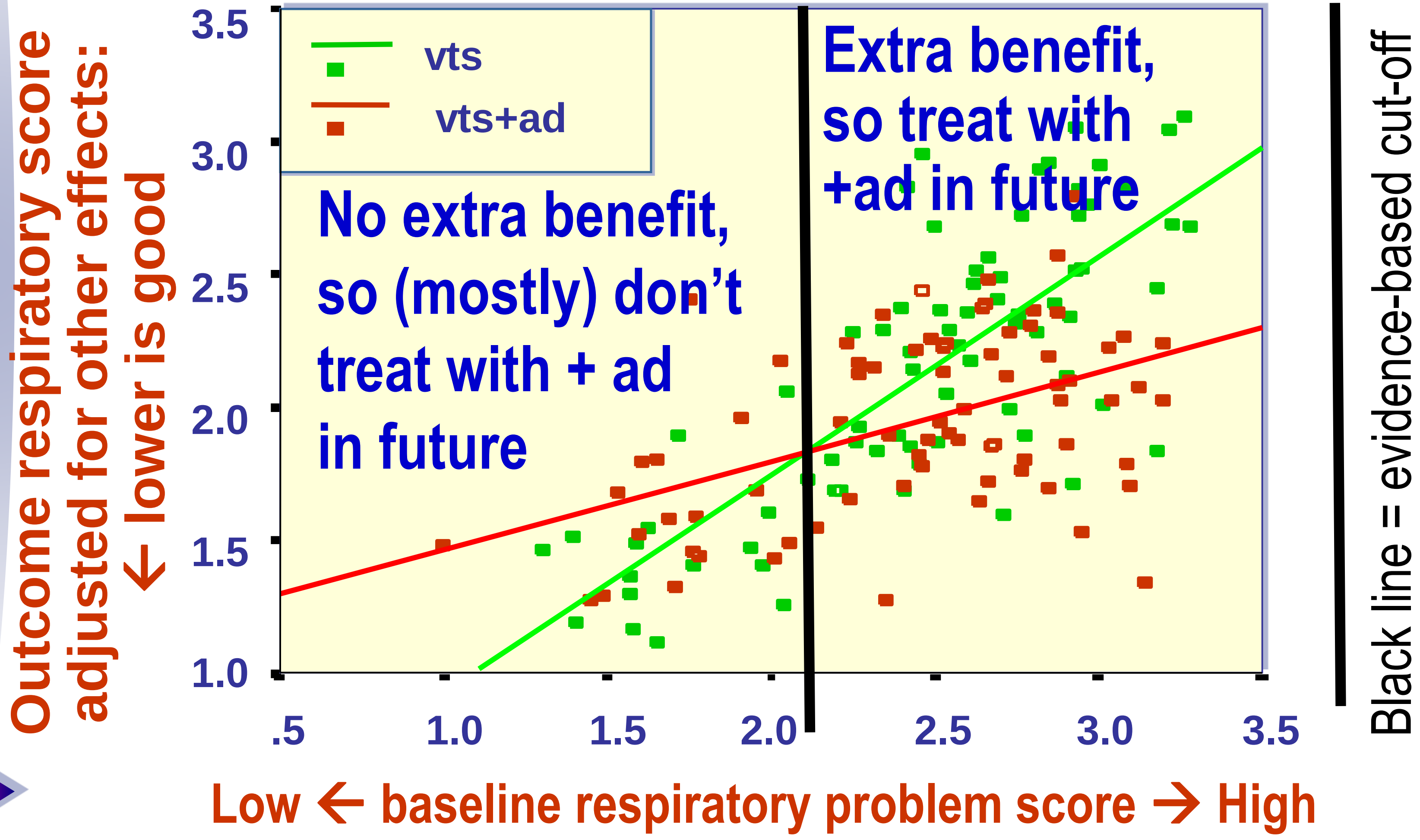
Effects included	% Variance explained	Estimates (+/-CIs) for Treatment [or IA with baseline]	P
Adenoidectomy Treatment (T)	4%	0.19 (+/-.14)	.008
T + *Baseline Respiratory score* (BL)	42%	0.17 (+/-.11)	.005
T+BL + *Interaction* (IA)	47%	[0.48/unit (+/-.212)]	[.000]
T+BL+IA+RBA *(Response bias adjustment)*	53%	[0.40/unit (+/-.208]	[.000]
T+BL+IA+RBA + two *Risk Factors* B5 & B6 (*SES, hayfever history*)	57%	[0.31/unit (+/-.202)]	[.003]

Comments on steps

- Note what happens down the last 3 rows, which presents a typical modelling set of decisions faced: where to stop?
- The interaction *baseline is so strong that attention must switch to models with it (although the overall -- also called 'main' -- effect can still be used descriptively and must be run also for interpretation in separate non-interaction models)
- Interaction is in effect a difference in regression slopes, hence meaning is always 'per unit in IV', seen in next slide
- The continuous effect of
- The CIs narrow progressively as more variables are entered (provided that they are significant, this must happen); this narrowness, not significance, should be the prime scientific goal
- Thus, the last model would be preferred unless existing good other reason was stated. It leaves an absolutely acceptable final p-value (though lately weakened)
- Narrowing of the CI accompanies reduced estimated magnitude for effect (shallower best estimate for slope)

Graphics as interpretation & explication of GLM

Interaction essential for understanding how adenoidectomy benefits child's respiratory symptoms. Also, a crossover interaction, here for *x* as a continuous covariate (but finally expressed dichotomously as a product of the analysis, not pre-stratification) helps determine good cut-off point, at which, if dichotomising IV x-variable, the Interaction would stay strong

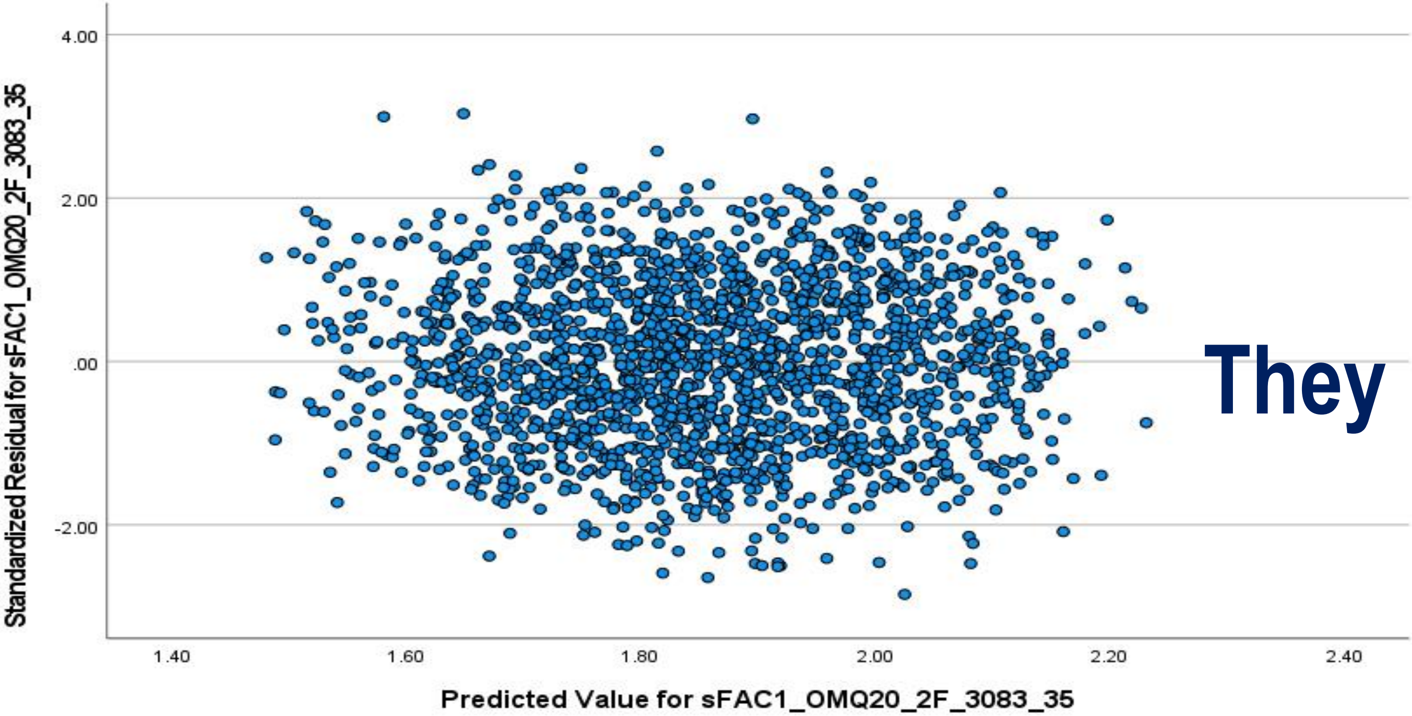


No extra benefit, so (mostly) don't treat with + ad in future

Extra benefit, so treat with +ad in future

Black line = evidence-based cut-off

Many graphics diagnostics for GLM; a basic one often used concerns homogeneity of residuals (error variance) over the predicted value, ie equal applicability of model to cases low & high in DV. Should give horizontal band from ~ –2SE to +2SE

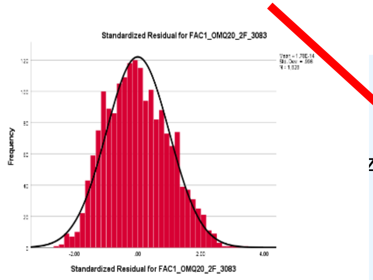


They do

In practice we would be more likely to use backwards deletion than forwards stepwise, but the next-but-1 slide consolidates the planning emphasis (on Forwards)

Graphics in diagnostics & tuning of GLM on a different set of data N=1828; Rsq 0.28→0.28 on 4 IVs

Gobbledygook name of variable means Sq root of (raw + 0.35), gentle transform for +ve skew still present after model choice



Statistics

		z for FAC1_OMQ20_2F_3083	ZRE_2 Standardized Residual for FAC1_OMQ20_2F_3083	ZRE_3 Standardized Residual for sFAC1_OMQ20_2F_3083_34	ZRE_4 Standardized Residual for sFAC1_OMQ20_2F_3083_35	ZRE_5 Standardized Residual for sFAC1_OMQ20_2F_3083_36	ZRE_6 Standardized Residual for sFAC1_OMQ20_2F_3083_37	ZRE_7 Standardized Residual for lFAC1_OMQ20_2F_3083_67	ZRE_8 Standardized Residual for lFAC1_OMQ20_2F_3083_68	ZRE_9 Standardized Residual for lFAC1_OMQ20_2F_3083_69	ZRE_10 Standardized Residual for lFAC1_OMQ20_2F_3083_70
N	Valid	1828	1828	1828	1828	1828	1828	1828	1828	1828	1828
	Missing	0	0	0	0	0	0	0	0	0	0
Mean		.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000	.0000
Median		-.0764	-.0633	-.0035	-.0055	-.0066	-.0084	-.0016	-.0017	-.0025	-.0034
Std. Deviation		.99561	.99561	.99561	.99561	.99561	.99561	.99561	.99561	.99561	.99561
Skewness		.243	.239	-.009	-.001	.007	.014	-.005	-.001	.003	.006
Std. Error of Skewness		.057	.057	.057	.057	.057	.057	.057	.057	.057	.057
Kurtosis		-.419	-.424	-.501	-.504	-.507	-.508	-.524	-.525	-.525	-.525
Std. Error of Kurtosis		.114	.114	.114	.114	.114	.114	.114	.114	.114	.114
Minimum		-2.55	-2.62	-2.86	-2.85	-2.84	-2.84	-2.85	-2.84	-2.84	-2.84
Maximum		3.41	3.46	3.03	3.03	3.04	3.04	3.02	3.03	3.03	3.03

At over 4 SEs, of course raw skew is sig, but Zeroised with transformation……

Similar with natural log function, but kurtosis worse; not much to be done about kurtosis

Recap: interpreting the improvements to the crude model as refined useful terms were added (appropriate & improving model) (Purple entries in last 3 rows are switched from main T effect to its *baseline interaction, as this is sig)

Effects included	% Variance explained	Estimates (+/-CIs) for Treatment or [IA with baseline]	P
Adenoidectomy Treatment (T)	4%	0.19 (+/-.14)	.008
T + *Baseline* (BL)	42%	0.17 (+/-.11)	.005
T+BL + *Interaction* (IA)	47%	[0.48/unit (+/-.212)]	[.000]
T+BL+IA+RBA *(Response bias adjustment)*	53%	[0.40/unit (+/-.208]	[.000]
T+BL+IA+RBA + 2 *Risk Factors* (SEG & hayfever)	57%	[0.31/unit (+/-.202)]	[.003]

CONFLICT HERE The last (circling) point is not to find 1 or 2 more ‘significant’ effects to claim, but rather to explicably improve the model. (However, ‘significant’ covariates remove error, so both things can happen)

What would I do in reporting ?

- Pre-list a priori parts; specify any freedoms, contingencies
- If new finding, check model variously: ‘nonsense’ analyses part-whole consistency, dropping one variable at a time &c
- Summarise the overall thrust of adding terms; use topic knowledge & very close inspection of data to justify & state the preferred model; interpret effect(s) of chief interest
- Other things equal, explicitly prefer model with minimum CI for effects of chief interest, not the minimum p-value
- For the IA estimate specifically, full model is also most conservative, relative to RSq modification for RBA; maxes Rsq
- State 3 considerations favouring the last model & quote it

In model choice, we address parsimony

Can various class terms add more than what they cost in df ?

1 Not surprising that baseline was larger than the treatment effect of interest (in F, t, ∆ Rsq etc); usually true and a gesture to within-subjects designs

2 Failing to consider interaction is “wrong”, medically. Treating all would expose the non-benefitting sub-group to slight risk and increases system costs

3 Randomisation of participants to treatment groups is necessary, but not sufficient control:
-- importantly, does remove allocation bias
-- largely but not fully balances group composition
-- last two extra terms in model can help with this

4 Subject to reliability & $ή_p^2$ considerations, the ‘extra’ terms could be used to project to differing populations or individuals, but this raises……….

Less clear-cut model interpretation issues: handling the limits of totally *a priori* approach

- No theoretical guarantee that social class & hay-fever which we happened to have, would be the best 2 of a bunch of epidemiologically relevant control variables (cf maternal smoking and asthma)
- So, which, & how many to fit other than the backwards elimination exercised ? Does allowing the 2 that come in rather than all the initial 4 raise a Bonferroni issue, even if we declare other 2 as 'tried but NS' ? (Next class)
- Specialised knowledge required here to make good decisions & interpretations. Declare under 'limitations' /'future research'
- Report the simplest gains truthfully & Multiverse them: do a robustness test that the result-of-interest is not largely dependent on non-*a priori* terms, and also report a model with them taken out #
- Subject to any other major considerations, choose analysis which gives most confidence (smallest CIs on the set of important estimates) not the smallest p on one!

Revisiting Modelling *versus* fixed analysis:
8 aspects of difference; they cannot all be great

- Adequacy of set of variables chosen for Rsq (% var expl) vs *a priori* framework (now these effects in this expt)
- Overall Rsq (or other GoF) important , cf individual variable Y/N significances
- Function forms: preference for continuous main-effect covariates; inspect for linearity
- Striving to capture phenomena most sensitively by expressing the form that they take: quality of fit not just some (sig) effect
- Idea of adjustment as reducer both of bias and error aspects of confounding is inherent in modelling
- Alternative models can be part of scientific output: not over-hyping or over-selecting: naturally leads to multi-versing
- It is usual to have collinearities (ie correlations among independent variables), so latent-variable approach ("factor" in FA) is implicit in how models come out and prior FA can be an alternative way to express several IVs
- Active discretion vs passive crunching by rule; benefits outweigh dangers, but you do need to know of dangers

Next class: how interactions work in GLM

more on obligations & the informativeness of the distribution of model's residuals (less important than formerly, thanks to boot-strapping)

‘Clever’ uses of modelling to solve measurement and conceptual problems, arising out of slide 18

- <u>Families</u> of models:
  - (a) alternative structures (families) -- finding them and then . testing families across members#
  - (b) interactively contrasting members of family
  - (c) accumulated explanatory power
- Destructive explaining away of a known effect (difference); can we do more than merely show that <u>some</u> effect survives adjustment (or that it does not)
- The table of adjusted means: adjusting out an effect as another manifestation of its presence (see slides 80+)
- Residual as new derived variable, with a component removed: “taking out” one of the variance components but beware the hidden extra df in doing this
- Transformations, normality assumptions and how the comfort zone of parametric statistics can be enlarged
  - When and how much does it matter?
- Knowing the limits of cleverness
  - Assumptions and applicability of covariates
  - Ratio of df required for stability; limitations of 10X rule

‘Model’ terminology: system of terms overlaps the one raised in Class 1: List not identical

- Model (1): A template for procedures at any level of analysis (also “paradigm”), often normative (ie as recommended, favoured); used widely in social science

- Model (2): A statistical model for a stochastic process enabling data with matching properties to be analysed according to its parameters, eg Poisson errors model, as used in math statistics

- Model (3): A theory (not strictly a correct usage): theory means a coherent set of more fundamental postulates than the observables, that can explain them, but not all of which may yet have been adequately verified or tested. Einstein: ‘a model for the theory’, a device enabling the theory to be tested

- Model (4): A set of formally specified relationships between variables that gives an economical (but not necessarily comprehensive) account of observables – this is most common & our usage in applied stats