Common uses and abuses of regression models: A proposal for reform of teaching and practice

John Carlin

Murdoch Children's Research Institute & The University of Melbourne

Stata Biostatistics & Epidemiology Virtual Symposium 23-Feb-23

[Based on ISCB President's Invited talk, ISCB43, Newcastle, U.K. 22-Aug-22]

A talk about *practice* and *teaching*…

… not about new methods or methodological research (or Stata tricks ☹)

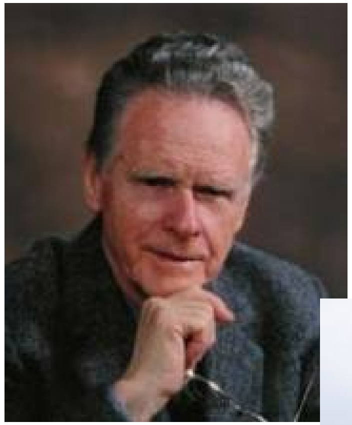
- … and ideas will be familiar to some, just not sufficiently connected to the mainstream of practice

- ***Practice*** is important: without active engagement in improving statistical standards, biostatisticians put their discipline at risk
- Bad science is a problem, often fed by bad statistics…
  and misuse of regression methods in particular!
- ***Teaching and training*** are important: academic statisticians should contribute at all levels to improve practice

Arthur P. Dempster

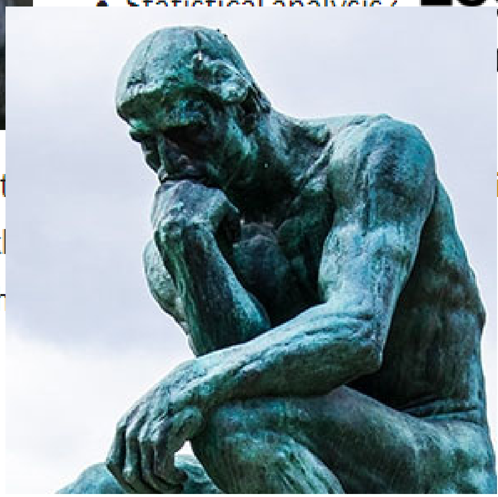
Emeritus Professor of Theoretical Statistics

Advice to PhD students:
"Knowledge of techniques is not the important thing in statistics"



Research Interests

- Methodology and logic of applied statistics.
- Computational aspe
- Modelling and analy
- Statistical analysis



- 1956: Ph.D. in Mat ... iversity
- 1953: M.A. in Mat
- 1952: B.A. in Math ... Toronto

Statistical Science
1998, Vol. 13, No. 3, 248–276

Logicist Statistics I. Models and Modeling

Dempster

2. WHAT "IS" A MODEL?

"Model" is used here interchangeably with the awkwardly long "mathematical model." The long form draws attention to abstract or purely mathematical content, while the short form suggests a type of replica, here a formal representation of objective reality through a corresponding mathematical structure. The term model implies, in addition to the abstract structure, a defined set of connections of the structure to the objective world,

Journey of an applied statistician (me!)

- Training: BSc, Masters & PhD in Statistics, limited practical experience
- Motivation/interest: using statistical methods to 'make sense of data', i.e. answer questions in health & medical research
- On the job: a big gap between training and confidence in practice
- How to cope? Looked around and noticed that a lot of statistical analysis published in medical journals uses **regression models**… so that was it!
- How to succeed? Become skilled at fitting regression models and back-engineering stories that sound meaningful to collaborators!
  - From ~1995, this was greatly facilitated by becoming a keen Stata user!

Example 1

- Question:

How much is a child's kidney enlarged after acute infection?

- Use of regression?

Sonographic Measurement of Renal Enlargement in Children with Acute Pyelonephritis and Time Needed for Resolution:
Implications for Renal Growth Assessment

Frederick E. Pickworth[1]
John B. Carlin[2]
Michael R. Ditchfield[1]
Margaret P. de Campo[1]
John F. de Campo[1]
David J. Cook[1]
Terry Nolan[3]
Harley R. Powell[4]
Robert Sloane[3]
Keith Grimwood[3]

OBJECTIVE. Failure of a kidney to grow satisfactorily in childhood is evidence of renal disease. Because kidneys may enlarge during an episode of acute pyelonephritis, concomitant renal length measurements cannot be used as baselines for growth assessment. This study was designed to determine the degree of renal enlargement in children with acute pyelonephritis and the time the enlargement takes to resolve after treatment is started to find the optimum time for obtaining baseline measurements.

SUBJECTS AND METHODS. In a cohort study, 180 children younger than 5 years old with their first proven acute urinary tract infection, with or without pyelonephritis, had renal scintigraphy and sonography within 15 days of starting treatment. The presence of cortical defects on scintigrams indicated pyelonephritis. The lengths of kidneys with and without scintigraphic defects (i.e., with and without pyelonephritis) were compared, adjusting for age and sex, and the length of kidneys with defects was related to time elapsed between the start of treatment and sonography.

RESULTS. Ninety-nine kidneys (28%) in 77 children (43%) had scintigraphic defects. Kidneys with defects were an average of 3.2 mm longer than kidneys without defects. Length and time interval between treatment and sonography in kidneys with defects correlated negatively, with mean length approaching that of kidneys without defects by 10–11 days.

CONCLUSION. Kidneys with acute pyelonephritis initially increase in length but return to normal on average by the 11th day of treatment. If poor renal growth is used as an indication of renal disease, sonography should be delayed or repeated at least 2 weeks after the start of treatment to determine the length of the uninflamed kidney.

AJR 1995;165:405–408

Example 1

Use of regression?

- *Kidney length increases with age*
- *Regression allows estimation of mean difference controlling for age… assuming difference ~constant!*

Nice descriptive summary of data

- *Aside: cubic curve used for age dependence*

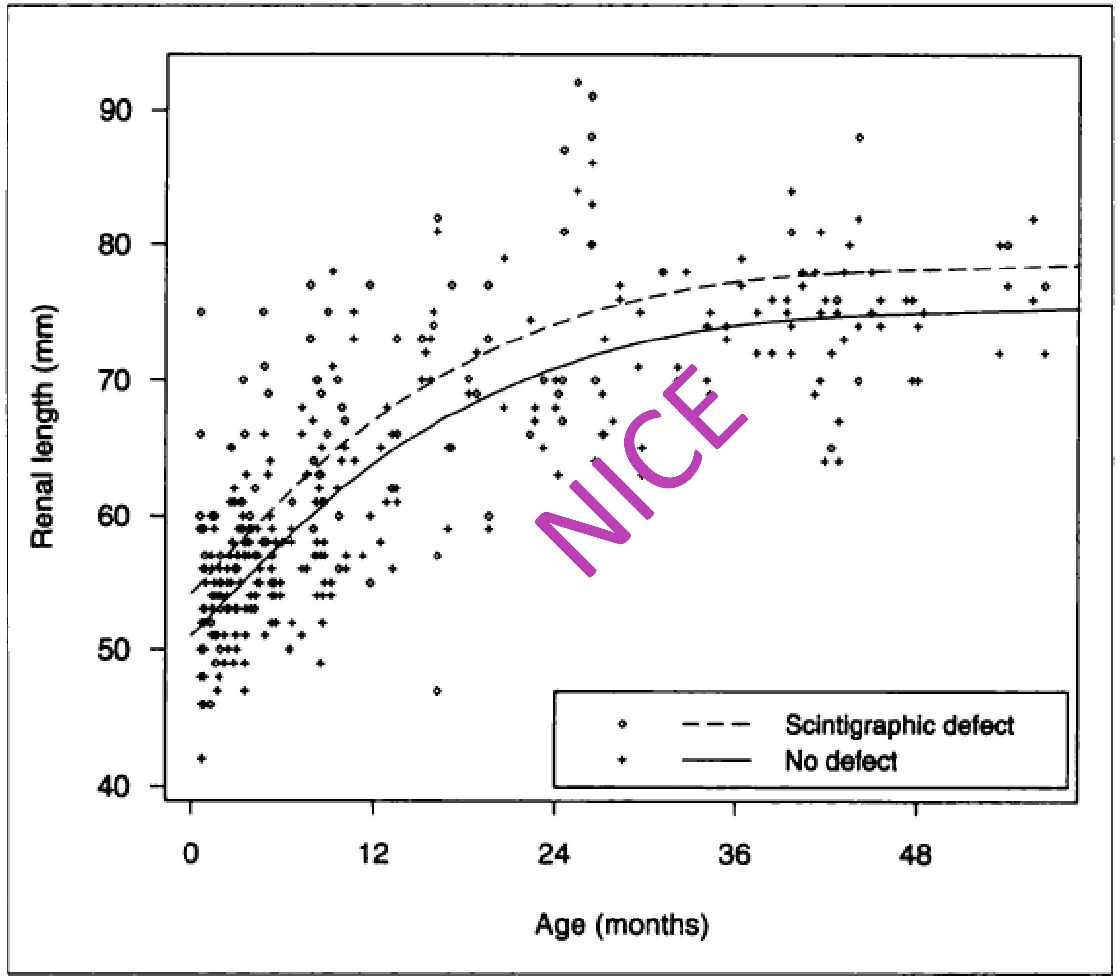


Fig. 1.—Scatter plot of renal length measured on sonograms versus age for kidneys with and without defects shown on scintigrams. Curved lines represent cubic model used in analysis of covariance calculations. Because curves are parallel, there is a similar absolute increase in renal length at all ages.

Example 2

- Question:

Can clinical factors predict successful gas enema for intussusception?

- Use of regression?

Gas Enema for the Reduction of Intussusception: Relationship Between Clinical Signs and Symptoms and Outcome

M. Katz[1]
E. Phelan[1]
J. B. Carlin[2]
S. W. Beasley[3]

OBJECTIVE. The aim of this study was to establish the extent to which the clinical features of intussusception can be used to predict successful outcome of gas enema and to determine whether the nonsurgical management of intussusception in children can be improved by refining the criteria used to select patients for gas enema.

SUBJECTS AND METHODS. Clinical data on 282 consecutive episodes of intussusception (255 patients) were collected prospectively from January 1987 to July 1991. Gas enema was performed in 273 episodes, in which the clinical signs and symptoms were studied by using logistic regression. Nine patients had primary surgery.

RESULTS. Gas enema was successful in 216 (79%) of 273 enemas attempted. Fifty-seven patients had surgery after unsuccessful enema. Univariate analysis showed significant associations between successful enema and duration of signs and symptoms less than 12 hr, no rectal bleeding, absence of small-bowel obstruction, presence of a palpable mass, and normal hydration. Multivariate analysis showed that dehydration, small-bowel obstruction, and duration of signs and symptoms longer than 12 hr were significant predictors of unsuccessful enema; yet, in these groups the rate of success still justified attempted enema. Even in severe dehydration, the successful enema reduction rate was 31%.

CONCLUSION. Our data suggest that although the factors identified had some predictive value in determining the outcome of attempted enema reduction, they could not be used to indicate patients in whom enema reduction should not be attempted. All patients with intussusception should have a gas enema if the absolute contraindications to enema (i.e., peritonitis or perforation) are absent.

AJR 1993;160:363–366

Example 2

- Question:

Can clinical factors predict successful gas enema for intussusception?

- Use of regression?

Logistic regression "to determine which variables were predictive… forward selection procedure was used…"

TABLE 4: Results of Logistic Regression Analysis with Successful Gas Enema as Outcome for Children with Intussusception

Predictor Variable[a]	p Value[b]	Odds Ratio[c]	95% Confidence Interval
Dehydration level	<.001		
1–4%		0.32	(0.13, 0.80)
5%		0.13	(0.05, 0.33)
6–10%		0.10	(0.02, 0.42)
Duration of symptoms >12 hr	.03	0.42	(0.02, 0.90)
Small-bowel obstruction	.005		
1–2 fluid levels		0.78	(0.32, 1.90)
>3 fluid levels		0.24	(0.10, 0.57)
Palpable mass present	.03	2.43	(1.07, 5.50)

MISGUIDED

Note.—Baseline odds of successful gas enema for well-hydrated patients who had signs and symptoms for less than 12 hr, no obstruction, and no palpable mass were 10.1.

[a]Variables in Table 1 that are omitted from this table showed no significant contribution to the multivariate model. No significant interactions were found between the independent variables.

[b]Likelihood ratio test for variable when entered last.

[c]Odds ratio comparing given level of each variable with baseline: for example, at dehydration level 1, odds of success are .32 times the odds for normal hydration, assuming all other variables remain constant.

Example 3

- Question:

Estimate "strength of association" of numerous factors with risk of childhood asthma

- Use of regression?

Paediatric and Perinatal Epidemiology 1993, 7, 67–76

The associations between childhood asthma and atopy, and parental asthma, hay fever and smoking

Mark A. Jenkins*, John L. Hopper*, Louisa B. Flander*, John B. Carlin† and Graham G. Giles‡

**Faculty of Medicine, Epidemiology Unit, The University of Melbourne, Parkville, Victoria, †Clinical Epidemiology and Biostatistics Unit, Royal Children's Hospital, Parkville, Victoria, and ‡Anti Cancer Council of Victoria, 1 Rathdowne St, Carlton, Victoria, Australia*

Example 3

- Question:

Estimate "strength of association" of numerous factors with risk of childhood asthma

- Use of regression?

"Relationships between [Y] and explanatory variables [$X_1, X_2, \ldots$] were analysed by multiple logistic regression"

Table 2. Odds ratios and 99% confidence intervals for child's asthma after adjustment for all other factors in the model

Risk factor	n=7368
Maleness	1.56 (1.30–1.86)
Hay fever	3.86 (3.12–4.78)
Eczema	2.04 (1.63–2.55)
Hives	1.34 (1.09–1.65)
Allergy to foods or medicines	1.70 (1.26–2.30)
Maternal asthma	2.63 (2.08–3.31)
Paternal asthma	2.52 (1.99–3.19)
Maternal smoking	1.26 (1.05–1.51)

WRONG

Table 2 fallacy!!!

A snapshot of current practice in clinical research


PEDIATRICS®

OFFICIAL JOURNAL OF THE AMERICAN ACADEMY OF PEDIATRICS

www.pediatrics.org

American Academy of Pediatrics

DEDICATED TO THE HEALTH OF ALL CHILDREN®

Pediatrics Zsp258=0$Mwwyi$: $Nyri$6466-

- 18 research articles
- 11/18 report results based on regression analysis
  - Others: 6/7 descriptive aims (2 purely qualitative), 1/7 RCT
- Of the 11:
  - 2: essentially descriptive (trends over time)
  - 1: time trends compared between “groups”
  - 4: regression to estimate a causal effect controlling for confounders
  - 4: “investigate associations”, “identify risk factors”…
- CLAIM: purpose (and therefore value) of 4-5 (of 11) uses of regression analysis are questionable

Regression abuse

Claim: regression models are poorly understood by most (non-statistician) users…

- "Exploring risk factors…": data-driven regression modelling is seen as a valid approach for illuminating cause and effect
  - E.g. from an anonymous reviewer (*J Cystic Fibrosis*)

  "It might also be interesting to include some multiple regression models with various health [markers] predicting QoL scores in the same model to understand the relative contribution of each marker on QoL."

- "Adjustment" = statistical magic to ensure quality of conclusions?

Regression models: what are they?

- Represent the variation in a "response" or "outcome" as "systematic + random" or "smooth + error"

$$Y = \beta_0 + \beta_1 X + \epsilon$$

- Simple, univariate

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \epsilon$$

- Multiple, multivariable

- Why are these models so compelling, but poorly used and understood?
- Can we find clues in the way we train people to use them…?

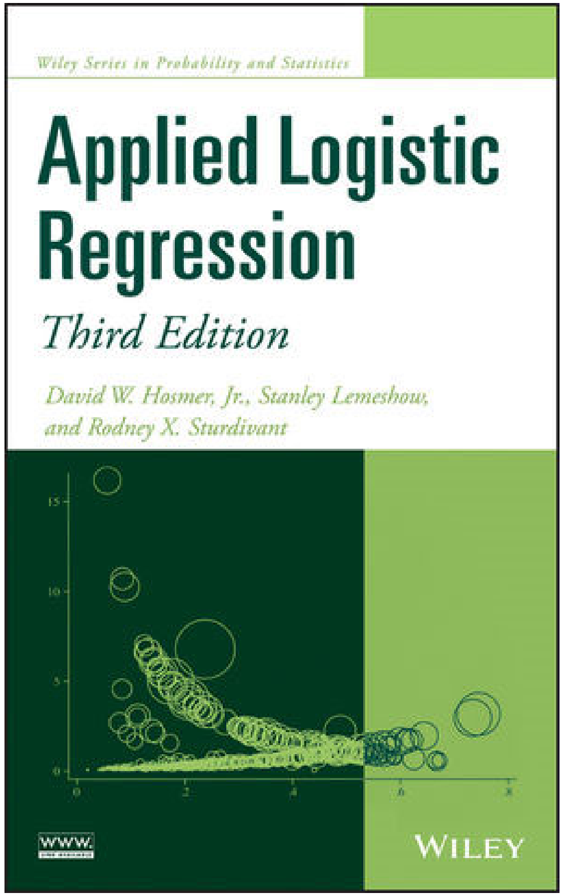
How do we (currently) teach regression?

- Classic texts

e.g. Hosmer & Lemeshow (1989, 2000, 2013)

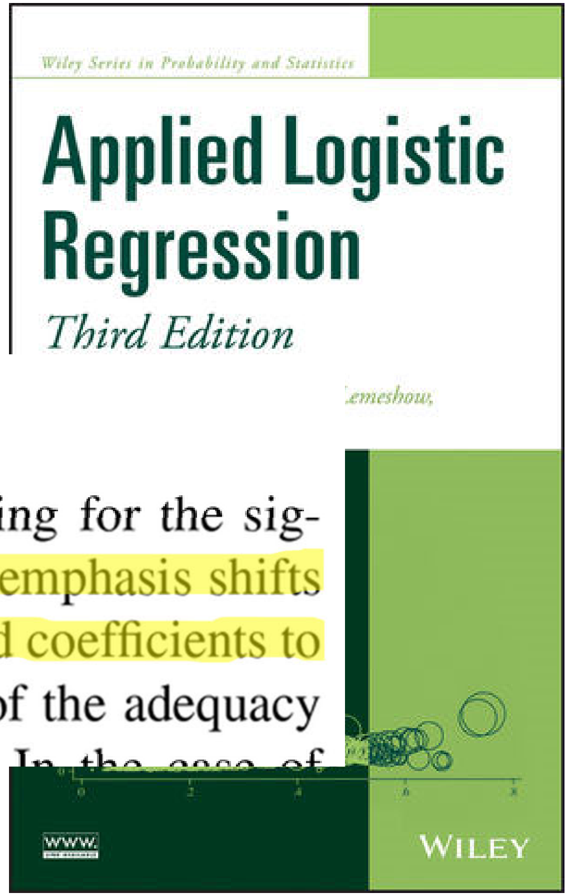
From the Introduction:

"Before beginning a thorough study of the logistic regression model it is important to understand that the goal of an analysis using this model is the same as that of any other regression model used in statistics, that is, to find the best fitting and most parsimonious, clinically interpretable model to describe the relationship between an outcome (dependent or response) variable and a set of independent (predictor or explanatory) variables."



How do we (currently) teach regression?

Contents



3.1 INTRODUCTION

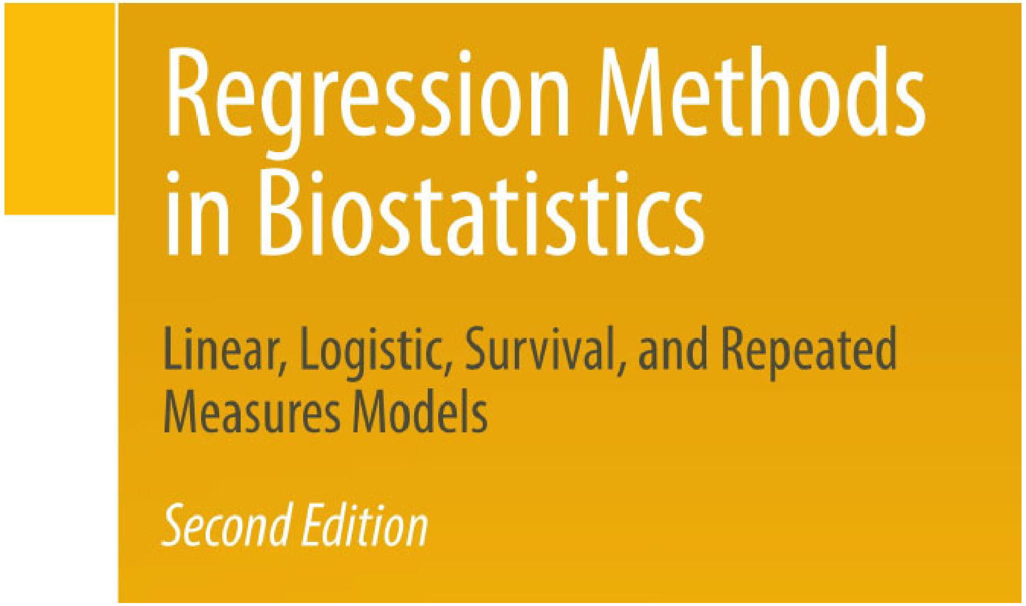
In Chapters 1 and 2 we discussed the methods for fitting and testing for the significance of the logistic regression model. After fitting a model the emphasis shifts from the computation and assessment of significance of the estimated coefficients to the interpretation of their values. Strictly speaking, an assessment of the adequacy of the fitted model should precede any attempt at interpreting it. In the case of

NO!!
We should interpret the coefficients *before* fitting the model

How do we (currently) teach regression?

- Contemporary texts

e.g. Vittinghoff et al (Springer, 2nd ed. 2012)

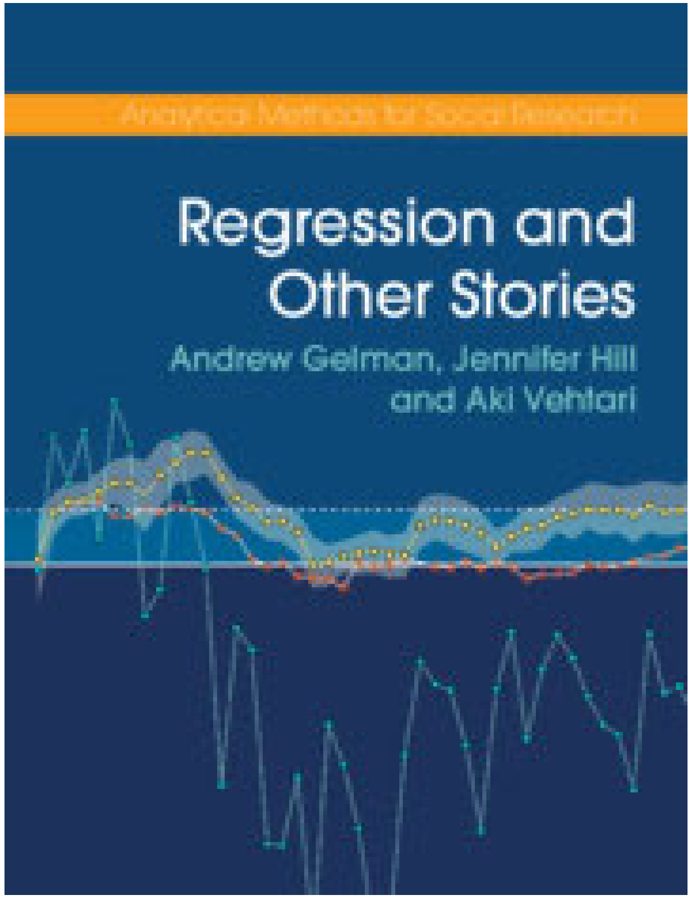


"The book describes a family of statistical techniques that we call *multipredictor* regression modeling. This family is useful in situations where there are multiple measured factors (also called predictors, covariates, or independent variables) to be related to a single outcome (also called the response or dependent variable). The applications of these techniques are diverse, including those where we are interested in prediction, isolating the effect of a single predictor, or understanding multiple predictors."

How do we (currently) teach regression?

Gelman, Hill & Vehtari
Regression and Other Stories (2020)

- Emphases the importance of purpose and tentativeness of models, but still ambiguous about whether the model or the purpose comes first



How do we (currently) teach regression?

- First define the general model,

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \epsilon$$

  - Establishes mathematical framework with clear notation
- Then discuss applications…
- This feeds the **true model myth**, that *lurking within every dataset is an underlying "true model" that we should find…*
  - Seductive – mathematical facts can be claimed AND scientific conclusions may be possible, IF the model is true
  - Students internalise that once you have the "correct" model everything else follows…

Regression models: why so important?

Two possibilities?

1) Natural phenomena follow "laws" that can be captured by regression models…
   - N.B. Assumes all variables measured and the model identifiable in whatever finite sample you might have!
   - Could then determine the "independent effect" of each variable on *Y*?
   - But surely this is never true of health and disease in populations, which are complex, with a lot of variability…
2) Regression models provide useful tools to study aspects of phenomena…
   - What exactly do we mean? How?
   - To answer, we need to revisit the purpose of statistical analysis!

Three types of research question

- Descriptive
  - Summarising and describing phenomena
- Prediction
  - Turning inputs into output: if we measure *a*, *b*, *c*, *d*, what value of *Y* should we expect?
- Causal
  - What value of *Y* should we expect if we change input *X*? (counterfactual prediction)

Hernán et al (*CHANCE,* 2019)
”three tasks in data science: description, prediction, counterfactual prediction”

Three purposes for regression methods

- Descriptive
  - Simple (“univariate”) regression *describes* average rate of change in *Y* with one-unit change in *X*?; multiple regression…?
- Predictive
  - Multiple regression seems a good place to start, with inputs/predictors the “independent variables”
- Causal
  - Role for regression not so immediately obvious…? (what do we mean by the “effect of *X* on *Y*”?)

Three examples revisited: three purposes?

Example 1

- Question:

How much is child's kidney enlarged after acute infection?

- Use of regression:

Estimate (mean) difference between infected and not, adjusting for age and sex

DESCRIPTIVE 👍

Example 2

- Question:

Can clinical factors predict successful gas enema for intussusception?

- Use of regression:

Logistic regression "to determine which variables were predictive… forward selection procedure was used…"

PREDICTIVE (but…)

Example 3

- Question:

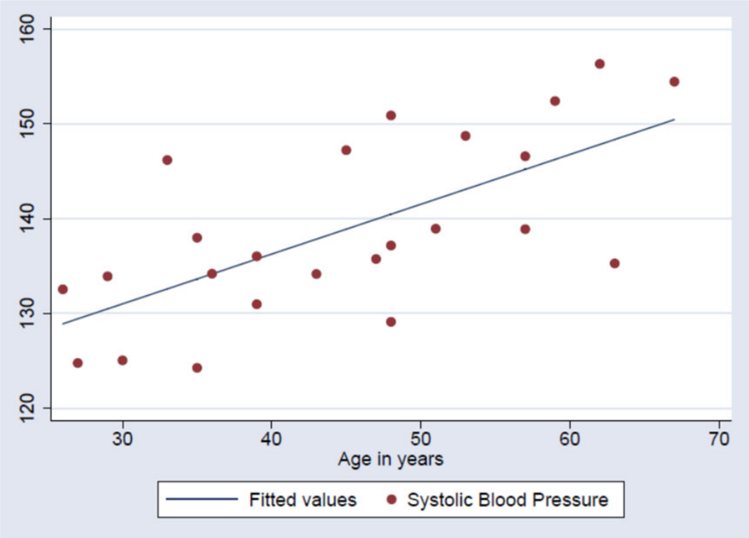
Estimate "strength of association" of numerous factors with risk of childhood asthma

- Use of regression:

"Relationships between [Y] and explanatory variables [$X_1, X_2, \ldots$] were analysed by multiple logistic regression"

CAUSAL??

How should we teach regression?



Simple linear regression

- Scatterplot depicts (co-)variation of *Y* and *X*
  - Regression line *describes* average rate of change in *Y* with one-unit change in *X*
- Why not *X* vs. *Y*? Correlation or regression?
- Regression describes variation of *Y* (*on average*) as function of *X*
  - What form of function? *Straight line?* (may need more flexible functions…)
- For what other purposes could this be useful?
  - Simple *prediction*: if we only know *X*, what do we expect for *Y*?
  - *Causal inference*: not clear how??
- Finally, need statistical inference for rate of change, simple prediction (mean, individual value), etc

How should we teach regression?

Multiple regression

Begin with purpose/questions, build theory as needed around these…

Introduce for each of the 3 purposes…

Descriptive purpose/question:

- Scatterplot *Y* vs. age for two groups (e.g. infected, not infected)
  - E.g. Example 1 (renal lengths)
- Regression as curve fitting/ smoothing
- May be useful to describe difference between groups using a simple model
- Inference for average difference between groups, adjusted? (for age, in example)
- All as exemplifying a general approach…

Fox MP et al, "On the Need to Revitalize Descriptive Epidemiology" *Am J Epi* 2022

How should we teach regression?

Multiple regression: for **prediction**

- Multiple *X*'s available for *prediction* of *Y*
- Standard independent-predictors linear model ($Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \epsilon$) a useful starting point
- Strategies for building models
  - Emphasise dependence on sample size!
  - Selection of variables, considering interactions, non-linearities etc
- Validation (internal/external)
- Coefficients not useful!! Traditional inference (tests) not useful!
  - Instead need to consider measures of prediction accuracy etc.

How should we teach regression?

Regression for causal inference

- First requires clarity of purpose: what are we seeking to estimate?
- Potential outcomes & target trial: define the causal effect as the true difference we would obtain in the ideal study (a perfect infinite RCT)
  - Difference in means, risk difference, RR, OR, …
  - Key *statistical* issue: the target parameter!
- Then map from the ideal to the actual study, invoking causal assumptions…
- A regression model for the outcome *may* enable us to estimate the target effect… how?

Regression for a causal purpose: how?

- Begin by defining the target parameter…

  i.e. causal effect (estimand) of interest…

  = (say) Mean difference in a continuous outcome Y:

$$\delta = \mu^{(1)} - \mu^{(0)}$$

  where $\mu^{(x)} = E\left(Y^{(x)}\right)$ = mean value in population under treatment/exposure condition x $(= 0{,}1)$

Regression for a causal purpose: how? (RCT)

- **Suppose target trial is feasible to conduct as a real RCT**
- If a **perfect** RCT, then the target parameter is (trivially) *identified* as

$$\delta = E\left(Y^{(1)}\right) - E\left(Y^{(0)}\right) = E(Y|X = 1) - E(Y|X = 0)$$

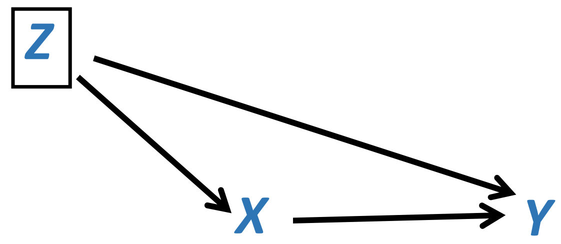
- ... which in turn is unbiasedly *estimated* by

$$\hat{\delta} = \hat{E}(Y|X = 1) - \hat{E}(Y|X = 0) = \bar{Y}_1 - \bar{Y}_0$$

- Inference for $\hat{\delta}$?
  - Can get from standard "*t*-test", or *equivalently* using regression estimation (in Stata-speak: `regress y i.trt`)

Regression for causal purpose (beyond RCT)

- **If target trial not feasible**: observational study (or an ‘imperfect’ RCT), we seek to *emulate* it…
- Target parameter (estimand) remains the same but is no longer identifiable without *causal assumptions*
  - Encode (some of) these in a causal diagram (DAG), which guides analysis planning to minimise biases
  - In particular, must try to control confounding…
  - …by standardisation (g-computation, weighting etc), or conditioning (blocking back-door paths)
- Conditioning = holding confounders (***Z***) constant: regression is one approach…



Regression for causal purpose (beyond RCT)

- How does regression provide control of a confounder?
- Need assumptions! Causal identifiability (consistency, exchangeability, positivity) $\Rightarrow$

$$\delta = E_Z(E(Y \mid X = 1, Z = z) - E(Y \mid X = 0, Z = z))$$

Now, *IF* (we further assume) $E(Y \mid X = x, Z = z) = \beta_0 + \beta_1 x + \beta_2 z$

THEN $E(Y \mid X = 1, Z = z) - E(Y \mid X = 0, Z = z) =$

$$= (\beta_0 + \beta_1 + \beta_2 * z) - (\beta_0 + \beta_2 * z) = \beta_1$$

So, under these assumptions,

$\delta = \beta_1$ = difference in means between exposure groups *at every value of* Z...

= desired target effect (*IF the effect is indeed constant for all* z)

In Stata-speak: `regress y i.trt z`

Regression for causal purpose (beyond RCT)

Extends in two ways:

1. Control for multiple confounders

$$E(Y \mid X = x, Z_1 = z_1, Z_2 = z_2, \ldots) = \beta_0 + \beta_1 x + \eta_1 z_1 + \eta_2 z_2 + \cdots$$

- N.B. the equation now encodes many more assumptions:
  - Effect is constant across all strata of the confounders!
  - Default linear specification for non-categorical confounder effects

2. Apply to different target estimands

- Risk ratio: log-link regression ("GLM") for binary ("binomial") outcome
- Odds ratio (if you must!): logistic regression

The causal revolution

- Upswell of interest in development of methods among biostatisticians
    - Newer “g-methods” also use regression models within them
- Standard practice in epidemiology & biostatistics lags behind
    - Yet to recognise that majority of research addresses causal questions
    - Statistical tradition largely responsible: “correlation does not equal causation”…

  Hernán, M. A. (2018). "The C-Word: Scientific Euphemisms Do Not Improve Causal Inference From Observational Data." *American Journal of Public Health* **108(5): 616-619.**
- In summary, most published data analysis both:
  (a) addresses causal questions, and
  (b) uses regression analysis…
  BUT with insufficient clarity of purpose (and so unclear method)!

Regression for causal purpose: done wrong…

An unfortunate but still common approach:

- Fit multiple regression model using all “risk factors of interest”, including “adjustment for covariates”
- Present estimates of coefficients (after variable selection) as “effects mutually adjusted for each other”

THIS HAS NO LOGICAL BASIS!

- The “Table 2 fallacy”
    - Greenland, S. and D. Westreich (2013). "The Table 2 Fallacy: Presenting and Interpreting Confounder and Modifier Coefficients." *American Journal of Epidemiology* **177(4): 292-298.**

Where does Stata fit in all of this?

- Provider of tools *par excellence*!
- Most of these agnostic to purpose…

Good…

- Flexibility for sophisticated users to adapt
- Ready-made tools for many purposes
  - e.g. for causal inference (`margins`, `teffects`, `pweights` option, as well as many regression commands, when used appropriately)

… and Bad…

- Estimation commands produce way too much output and provide freedom for regression abuse on massive scale!

Concluding messages

- Regression is not a method for “fitting models” but a tool for answering questions
  - “All models are wrong, but some models are useful…” (George Box, 1970s)
  - Useful for what? ***Until the purpose is defined, the construction of models should wait!***
- Extensive reform is needed in the practice and teaching of regression methods in biostatistics, epidemiology and health science
  - Could software developers play a role?

Acknowledgment: Margarita Moreno-Betancur

- Responsible for many of the ideas presented, e.g. the terms “true model myth”, “regression abuse”