D. R. Cox: aspects of scientific inference

LSHTM SYMPOSIUM: 50 YEARS OF THE COX MODEL

H. S. Battey
Department of Mathematics, Imperial College London

November 10, 2022

SOME BACKGROUND TO TWO INFLUENTIAL 1972 PAPERS

Cox, D. R. (1955). Some statistical methods connected with series of events (with discussion). *JRSSB*, 17, 129–164.

Cox, D. R. (1956). Joint meeting of the IMS and Biometric Society, Princeton, N.J. – invited address later published as

Cox, D. R. (1958a). Some problems connected with statistical inference. *Ann. Math. Statist.*, 29, 357–72.

Cox, D. R. (1958b) The regression analysis of binary sequences (with discussion). *JRSSB*, 20, 215–242.

Cox, D. R. (1958c) Two further applications of a model for binary regression. *Biometrika*, 45, 562–565.

Cox, D. R. (1964). Some applications of exponential ordered scores. *JRSSB*, 26, 103–110.

Cox, D. R. (1968). Notes on some aspects of regression analysis (with discussion). *JRSSA*, 131, 265–279.

Cox, D. R. (1970). *The Analysis of Binary Data*. London: Methuen.

EARLY HISTORY: A VERY SELECTIVE OVERVIEW (1/3)

D-24 K. Pearson (1900): χ^2 test of independence in the 2×2 table: subject matter for a series of printed polemics.

D-2 Fisher (1922): likelihood, sufficiency, information, pointed out Pearson's error.

[A statistic is sufficient for a parameter ψ if no other statistic that can be calculated from the same sample provides additional information as to the value of ψ]

D+1 Fisher (1925) emphasized conditioning on an ancillary statistic for recovery of information lost by using the maximum likelihood estimate.

D+32 These ideas developed over many years culminating with Fisher (1956) and Cox (1956, 1958a).

EARLY HISTORY: A VERY SELECTIVE OVERVIEW (2/3)

Fisher (1935):

> *Some, or sometimes all of the lost information may be recovered by calculating what I call ancillary statistics, which themselves tell us nothing about the value of the parameter, but, instead, tell us how good an estimate we have made of it.*

> *Ancillary statistics are only useful when different samples of the same size can supply different amounts of information, and serve to distinguish those which supply more from those which supply less.*

> *If it be admitted that these marginal frequencies by themselves supply no information on the point at issue...*

EARLY HISTORY: A VERY SELECTIVE OVERVIEW (3/3)

Barnard (1945, 1947) put forward a test which he claimed was more powerful than Fisher's exact test, then withdrew the procedure.

Fisher (1956):

> *Professor Barnard has since then frankly avowed (1949) that further reflection has led him to the same conclusion as Yates and Fisher, as indeed Wilson with equal generosity had done earlier.*

Yates (1984):

> *That this conclusion is still not accepted in many quarters, however, is very evident from numerous recent publications.* [*]

* This is still the case. See also Brown (1990), rebutted by Fraser and Reid (1990); Buja et al. (2019), rebutted by Davison et al. (2019).

SOME BACKGROUND TO TWO INFLUENTIAL 1972 PAPERS

Cox, D. R. (1955). Some statistical methods connected with series of events (with discussion). *JRSSB*, 17, 129–164.

Cox, D. R. (1956). Joint meeting of the IMS and Biometric Society, Princeton, N.J. – invited address later published as

Cox, D. R. (1958a). Some problems connected with statistical inference. *Ann. Math. Statist.*, 29, 357–72.

Cox, D. R. (1958b) The regression analysis of binary sequences (with discussion). *JRSSB*, 20, 215–242.

Cox, D. R. (1958c) Two further applications of a model for binary regression. *Biometrika*, 45, 562–565.

Cox, D. R. (1964). Some applications of exponential ordered scores. *JRSSB*, 26, 103–110.

Cox, D. R. (1968). Notes on some aspects of regression analysis (with discussion). *JRSSA*, 131, 265–279.

Cox, D. R. (1970). *The Analysis of Binary Data*. London: Methuen.

COX (1958a). SOME PROBLEMS CONNECTED WITH STATISTICAL INFERENCE

- Brought clarity to Fisher's reasoning.
- Incisive demonstration of the need for conditioning in order to ensure scientific relevance.
- Emphasized that such relevance is sometimes incompatible with ideas of optimality that remain popular today.
- By appeal to DRC's example, Fisher's argument for conditioning on appropriate reference sets is hard to refute.
- A first attempt to define a notion of ancillarity in the presence of nuisance parameters.

RELATIONSHIP BETWEEN COX (1958a) AND SUBSEQUENT WORK

The question of where to limit the conditioning is a challenging one. In the simplest setting, an arbitrarily granular choice renders each individual uninformative about others, while too coarse a conditioning typically yields conclusions irrelevant to the question at hand. When there are many nuisance parameters the appropriate conditional formulation becomes particularly elusive, although the conceptual argument for distinguishing samples of varying degrees of information remains compelling.

SOME BACKGROUND TO TWO INFLUENTIAL 1972 PAPERS

Cox, D. R. (1955). Some statistical methods connected with series of events (with discussion). *JRSSB*, 17, 129–164.

Cox, D. R. (1956). Joint meeting of the IMS and Biometric Society, Princeton, N.J. – invited address later published as

Cox, D. R. (1958a). Some problems connected with statistical inference. *Ann. Math. Statist.*, 29, 357–72.

Cox, D. R. (1958b) The regression analysis of binary sequences (with discussion). *JRSSB*, 20, 215–242.

Cox, D. R. (1958c) Two further applications of a model for binary regression. *Biometrika*, 45, 562–565.

Cox, D. R. (1964). Some applications of exponential ordered scores. *JRSSB*, 26, 103–110.

Cox, D. R. (1968). Notes on some aspects of regression analysis (with discussion). *JRSSA*, 131, 265–279.

Cox, D. R. (1970). *The Analysis of Binary Data*. London: Methuen.

Journal of the Royal Statistical Society

SERIES B (METHODOLOGICAL)

Vol. XX, No. 2, 1958

THE REGRESSION ANALYSIS OF BINARY SEQUENCES

By D. R. COX

Birkbeck College, University of London

[Read before the RESEARCH SECTION of the ROYAL STATISTICAL SOCIETY, March 5th, 1958, Professor G. A. BARNARD in the Chair]

⋮

a narrow range, because of the restriction that θ_t must lie in [0, 1] and, in the absence of special considerations for a particular problem, the best form seems to be the logistic law

$$\text{logit } \theta_t \equiv \log \{\theta_t/(1 - \theta_t)\} = \alpha + \beta x_t, \tag{1}$$

i.e.

$$\theta_t \equiv \text{pr } (Y_t = 1) = e^{\alpha+\beta x_t}/(1 + e^{\alpha+\beta x_t}), \tag{2}$$

and

$$1 - \theta_t \equiv \text{pr } (Y_t = 0) = 1/(1 + e^{\alpha+\beta x_t}). \tag{3}$$

This form has been extensively used in work on bioassays, notably by Berkson.

COX (1958b): AN ALTERNATIVE EXPOSITION

- From the Bernoulli likelihood function, match up minimal sufficient statistics with those of a normal-theory linear model.
- Recover the logistic form as the unique model that produces such unification and satisfies the boundary conditions.

Q. Why match up the sufficient statistics?

A.1 Simple sufficient statistics allow an exact conditional analysis.

- Ensures relevance by attaching to conclusions the precisions actually achieved.
- Eliminates nuisance parameters.

A.2 A unified theory has some aesthetic appeal.

Fisher's exact test is recovered as a special case.

A.3 Stabilizes interpretation to some extent.

LOGISTIC REGRESSION VIA SUFFICIENCY (1/4)

The normal theory linear model has associated with it simple sufficient statistics for the regression coefficient vector $\beta = (\beta_1, \ldots, \beta_p)^{\mathrm{T}}$ and unknown error variance. These are $S = X^{\mathrm{T}} Y$, i.e. $(S_j)_{j=1}^p = (\sum_{i=1}^n x_{ij} Y_i)_{j=1}^p$ and the residual sum of squares.

Suppose now that $(Y_i)_{i=1}^n$ are binary, with outcomes encoded arbitrarily as $\{0, 1\}$. There is no floating dispersion parameter.

LOGISTIC REGRESSION VIA SUFFICIENCY (2/4)

Parameterize the likelihood function in terms of the binomial success probabilities $(\theta_i)_{i=1}^n$, where "success" corresponds to $y_i = 1$:

$$\ell(\theta_1, \ldots, \theta_n; y_1, \ldots, y_n) = \prod_{i=1}^n \theta_i^{y_i}(1-\theta_i)^{(1-y_i)}, \quad (y_i)_{i=1}^n \in \{0,1\}^n$$

and consider which function $\theta_i(x_i) : \mathbb{R}^p \to [0,1]$ produces sufficient statistics for the unknown parameter of the form $(\sum_{i=1}^n x_{ij} y_i)_{j=1}^p$.

By analogy with linear regression we might consider a smaller class of functions of the form $\theta_i(x_i^{\mathrm{T}} \beta) : \mathbb{R} \to [0,1]$.

LOGISTIC REGRESSION VIA SUFFICIENCY (3/4)

$$\ell(\theta_1, \ldots, \theta_n; y_1, \ldots, y_n) = \prod_{i=1}^n \theta_i^{y_i}(1-\theta_i)^{(1-y_i)}, \quad (y_i)_{i=1}^n \in \{0,1\}^n$$

For the sufficient statistic to be a sum, either θ_i or $(1-\theta_i)$ must contain an exponential and these probabilities need to be equal up to the exponential term in order that the exponential of the sum be factorable in the likelihood. More explicitly, sufficiency of $(\sum_{i=1}^n x_{ij} y_i)_{j=1}^p$ for β requires

$$\begin{aligned} \theta_i &= f(x_i^{\mathrm{T}}\beta) e^{x_i^{\mathrm{T}}\beta} \in [0,1] \\ 1-\theta_i &= f(x_i^{\mathrm{T}}\beta) = \frac{\theta_i}{e^{x_i^{\mathrm{T}}\beta}} \in [0,1], \end{aligned}$$

LOGISTIC REGRESSION VIA SUFFICIENCY (4/4)

Enforcing the boundary conditions leads to the logistic law:

$$\theta_i = \frac{e^{x_i^{\mathrm{T}}\beta}}{1 + e^{x_i^{\mathrm{T}}\beta}}, \quad 1 - \theta_i = \frac{1}{1 + e^{x_i^{\mathrm{T}}\beta}}.$$

The argument is symmetric and these could be defined the other way round, as would be hoped since the coding of the binary variables is arbitrary.

EXACT CONDITIONAL ANALYSIS (COX 1958b, 1970)

Suppose β_p is the parameter of interest. The conditional distribution of $S_p = \sum_{i=1}^{n} x_{ip} Y_i$ given $S_1 = s_1, \ldots, S_{p-1} = s_{p-1}$ is

$$\text{pr}(S_p = s_p \mid S_1 = s_1, \ldots, S_{p-1} = s_{p-1}) = \frac{c(s_1, \ldots, s_p) e^{\beta_p s_p}}{\sum_u c(s_1, \ldots, s_{p-1}, u) e^{\beta_p u}},$$

where $c(s_1, \ldots, s_p)$ is the number of possible realizations of $Y_1, \ldots, Y_n$ such that the values of $S_1, \ldots, S_p$ are equal to those actually observed.

Cox (1958b) used this to construct exact confidence sets for β_p.

REMARKS (1/3)

- $S_1, \ldots, S_{p-1}$ are ancillary for β_p and sufficient for $\beta_1, \ldots, \beta_{p-1}$.
- Ancillarity: if we were given only the values $s_1, \ldots, s_{p-1}$ no conclusions could be drawn about β_p. Achieve relevance by conditioning on them.
- By sufficiency, conditioning eliminates $p - 1$ nuisance parameter from the analysis.
- Both justifications of conditioning lead to the same conclusion in this case.

REMARKS (2/3)

It is clear that absence of Newton-Raphson and maximum likelihood fitting was deliberate. Fisher proposed Fisher scoring (an application of Newton-Raphson) for solving ML estimating equations in 1925. DRC was certainly aware of it in 1958.

WHY IS THE EXACT CONDITIONAL ANALYSIS BARELY USED?

A.1 Calculation of the combinatorial quantity $c(s_1, \ldots, s_p)$ would have been difficult with 1950s computation.

A.2 Rigid application of Neyman-Pearson theory became widespread.

REMARKS (3/3)

Suppose as before that the outcome Y is encoded as $\{0, 1\}$. Let $\theta(x) = \mathbb{E}(Y(x)) = \mathrm{pr}(Y(x) = 1)$. The logistic model is

$$\underbrace{\log \frac{\theta(x)}{1 - \theta(x)}}_{g(\theta)} = x^{\mathrm{T}}\beta.$$

$$\eta_{jk} := \frac{\beta_j}{\beta_k} = \frac{(d/d\theta)g(\theta) \cdot (\partial\theta(x)/\partial x_j)}{(d/d\theta)g(\theta) \cdot (\partial\theta(x)/\partial x_k)} = \frac{\partial\theta(x)/\partial x_j}{\partial\theta(x)/\partial x_k}.$$

Thus, η_{jk} specifies by how much we need to change x_k in order to have the same effect on θ as a unit change in x_j, all other things equal, and this interpretation is the same as in linear regression.

More on this presently...

SOME BACKGROUND TO TWO INFLUENTIAL 1972 PAPERS

Cox, D. R. (1955). Some statistical methods connected with series of events (with discussion). *JRSSB*, 17, 129–164.

Cox, D. R. (1956). Joint meeting of the IMS and Biometric Society, Princeton, N.J. – invited address later published as

Cox, D. R. (1958a). Some problems connected with statistical inference. *Ann. Math. Statist.*, 29, 357–72.

Cox, D. R. (1958b) The regression analysis of binary sequences (with discussion). *JRSSB*, 20, 215–242.

Cox, D. R. (1958c) Two further applications of a model for binary regression. *Biometrika*, 45, 562–565.

Cox, D. R. (1964). Some applications of exponential ordered scores. *JRSSB*, 26, 103–110.

Cox, D. R. (1968). Notes on some aspects of regression analysis (with discussion). *JRSSA*, 131, 265–279.

Cox, D. R. (1970). *The Analysis of Binary Data*. London: Methuen.

Notes on Some Aspects of Regression Analysis

By D. R. Cox

Imperial College

[Read before the Royal Statistical Society on Wednesday, March 20th, 1968, the President, Dr F. Yates, C.B.E., F.R.S., in the Chair]

⋮

The justification of maximum likelihood methods is asymptotic but sometimes analogues of at least a few of the "exact" properties of normal-theory linear models can be obtained. The simplest case is when the ith observation on the dependent variable has a distribution in the exponential family (Lehmann, 1959, p. 50)

$$\exp\{A_i(y)\,B(\theta_i)+C_i(y)+D(\theta_i)\},$$

where θ_i is a single parameter and there is a linear model

$$B(\theta_i)=\sum x_{ir}\beta_r,$$

where the β's are unknown parameters and the x's known constants. Special cases are the binomial, Poisson and gamma distributions when the "linear" model applies to the logit transform, to the log of the Poisson mean and to the reciprocal of the mean of the gamma distribution. Sufficient statistics are obtained and in very fortunate cases useful "exact" significance tests for single regression coefficients emerge.

1970

14. The independent random variables $Y_1, \ldots, Y_n$ are such that the p.d.f. of Y_i is

$$\exp\{A_i(\theta_i)\, B_i(y) + C_i(\theta_i) + D_i(y)\},$$

where θ_i is a scalar parameter. A 'linear' model for the θ_i asserts that

$$A_i(\theta_i) = \sum a_{is}\beta_s.$$

Show that there are simple sufficient statistics for the unknown parameters $\boldsymbol{\beta}$, and that much of the discussion of Chapter 4 can be paralleled. Show that for binomial and Poisson distributions the 'linear' models are consistent with natural constraints on the signs of the parameters but that for gamma distributions with

p.20: discussion of alternative link functions (but does not use that terminology)

FROM EARLIER

- $S_1, \ldots, S_{p-1}$ are ancillary for β_p and sufficient for $\beta_1, \ldots, \beta_{p-1}$.
- Ancillarity: if we were given only the values $s_1, \ldots, s_{p-1}$ no conclusions could be drawn about β_p. Achieve relevance by conditioning on them.
- By sufficiency, conditioning eliminates $p-1$ nuisance parameter from the analysis.
- Both justifications of conditioning lead to the same conclusion in this case.

The above conclusions extend to exponential families when the canonical parameter $g(\theta)$ is modelled linearly as in Cox (1968, 1970).

FROM EARLIER

Let $\theta(x) = \mathbb{E}(Y(x)) = \mathrm{pr}(Y(x) = 1)$. The logistic model is

$$\underbrace{\log \frac{\theta(x)}{1-\theta(x)}}_{g(\theta)} = x^{\mathrm{T}}\beta.$$

$$\eta_{jk} := \frac{\beta_j}{\beta_k} = \frac{(d/d\theta)g(\theta) \cdot (\partial\theta(x)/\partial x_j)}{(d/d\theta)g(\theta) \cdot (\partial\theta(x)/\partial x_k)} = \frac{\partial\theta(x)/\partial x_j}{\partial\theta(x)/\partial x_k}.$$

Thus, η_{jk} specifies by how much we need to change x_k in order to have the same effect on θ as a unit change in x_j, all other things equal, and this interpretation is the same as in linear regression.

The above conclusion extends to exponential families when the canonical parameter $g(\theta)$ is modelled linearly as in Cox (1968, 1970).

SOME BACKGROUND TO TWO INFLUENTIAL 1972 PAPERS

Cox, D. R. (1955). Some statistical methods connected with series of events (with discussion). *JRSSB*, 17, 129–164.

Cox, D. R. (1956). Joint meeting of the IMS and Biometric Society, Princeton, N.J. – invited address later published as

Cox, D. R. (1958a). Some problems connected with statistical inference. *Ann. Math. Statist.*, 29, 357–72.

Cox, D. R. (1958b) The regression analysis of binary sequences (with discussion). *JRSSB*, 20, 215–242.

Cox, D. R. (1958c) Two further applications of a model for binary regression. *Biometrika*, 45, 562–565.

Cox, D. R. (1964). Some applications of exponential ordered scores. *JRSSB*, 26, 103–110.

Cox, D. R. (1968). Notes on some aspects of regression analysis (with discussion). *JRSSA*, 131, 265–279.

Cox, D. R. (1970). *The Analysis of Binary Data*. London: Methuen.

COX (1972): MAIN CONNECTION TO THE FOREGOING

Inferential separations enable elimination of the baseline hazard function, an infinite-dimensional nuisance parameter (Nancy Reid's lecture).

ARE THE LOGISTIC AND PH MODELS COMPATIBLE?

Suppose that lifetimes are generated from a distribution with proportional hazards but that we only observe the $\{\text{alive}, \text{dead}\} = \{0, 1\}$ indicator at the end of the study. Let β_{L} and β_{H} be the coefficient vectors (without intercept) in the logistic regression and PH models respectively.

Q. Is the logistic model compatible with the PH model in the sense that $\beta_{\mathsf{L}} = \beta_{\mathsf{H}}$?

ARE THE LOGISTIC AND PH MODELS COMPATIBLE?

A. No. Logistic and proportional odds (McCullagh, 1980) are compatible. PH is compatible with a double logarithmic transform of the survival probabilities:

$$\underbrace{\log[-\log\{1-\theta(x)\}]}_{g(\theta)} = \text{constant} + x^{\mathrm{T}}\beta_{\mathrm{H}}$$

$g(\theta)$ does not lead to simple sufficient statistics for β_{H} in the dichotomized model.

A more flexible extension of Cox's (1968, 1970) canonical exponential family regression models sacrifices the simple sufficient statistics and the possibility for exact conditional analysis.

This extension is the class of Generalized Linear Models (Nelder and Wedderburn, 1972).

CLOSELY RELATED DEVELOPMENTS (1/4)

Generalized linear models (Nelder and Wedderburn, 1972) followed quite directly. Emphasised maximum likelihood fitting by the Newton-Raphson algorithm. Based on the full likelihood function.

CLOSELY RELATED DEVELOPMENTS (2/4)

A body of work sought to achieve the appropriate conditioning approximately, beyond exponential family canonical form, e.g.

- Fraser (1964). Local conditional sufficiency. *JRSSB*.
- Barndorff-Nielsen and Cox (1979). Edgeworth and saddle-point approximations with statistical applications (with discussion). *JRSSB*
- Cox (1980). Local ancillarity. *Biometrika*
- Barndorf-Nielsen (1983). On a formula for the distribution of the maximum likelihood estimator. *Biometrika*
- McCullagh (1984). Local sufficiency. *Biometrika*
- Cox and Reid (1987). Parameter orthogonality and approximate conditional inference (with discussion). *JRSSB*
- Fraser and Reid (1988). On conditional inference for a real parameter: a differential approach on the sample space. *Biometrika*
- Barndorff-Nielsen (1990). Approximate interval probabilities. *JRSSB*
- Fraser (1990). Tail probabilities from observed likelihoods. *Biometrika*

CLOSELY RELATED DEVELOPMENTS (3/4)

Sufficiency provides a separation of the information in the data into that relevant for inference on the parameters of a given model, and that relevant for assessing model adequacy.

On the basis of this, Cox (1968) emphasised the construction of confidence sets of models: all low-dimensional subsets of variables that are statistically indistinguishable at an arbitrary threshold.

This was emphasised repeatedly, e.g. Cox and Snell (1974, 1989), Cox (1995), Cox and Battey (2017).

CLOSELY RELATED DEVELOPMENTS (4/4)

Parameter-based factorisation of the likelihood function:

$$L(\psi, \lambda; y) = L_{\mathsf{pa}}(\psi; y) L_{\mathsf{r}}(\psi, \lambda; y).$$

- Base inference for ψ on $L_{\mathsf{pa}}(\psi; y)$, thereby eliminating λ.
- Ideally, little or no information for inference on ψ is lost through relinquishment of the remainder likelihood $L_{\mathsf{r}}(\psi, \lambda; y)$.
- Constructive procedure for finding factorisable transformations or approximately factorisable transformations.

(Photo credit: Christiana Kartsonaki). Clockwise from anterior:
Christiana Kartsonaki, Nanny Wermuth, Ruth Keogh, David Cox, Heather Battey

THE END