Biometrika (2024), **0**, 0, *pp.* 1–8 https://doi.org/10.1093/biomet/asae038
Advance Access publication 8 August 2024

Studies in the history of probability and statistics, LI: the first conditional logistic regression

By J. A. HANLEY

Department of Epidemiology, Biostatistics and Occupational Health, McGill University, 2001 McGill College Avenue, Montréal, Québec H3A 1G1, Canada
james.hanley@mcgill.ca

Summary

Statisticians and epidemiologists generally cite the publications of Prentice & Breslow (1978) and Breslow et al. (1978) as the first description and use of conditional logistic regression, while economists cite the book chapter by Nobel laureate McFadden (McFadden, 1973). We describe the until-now-unrecognized use of, and way of fitting, this model in 1934 by Lionel Penrose and Ronald Fisher.

Some key words: Birth order; Down's syndrome; Estimating equation; Family-based selection; Maternal age; Peer review; Relative odds; Standard error.

1. Introduction

In epidemiological research, conditional logistic regression is historically thought of in the context of matched case control studies (Breslow et al., 1978; Prentice & Breslow, 1978; Keogh & Cox, 2014). It can also be seen as providing the likelihood contributions involved in the fitting of the Cox model in survival analysis, if one regards each risk set as a matched set (Liddell et al., 1977). Its first application in economics research was to 'choice-based' or 'outcome-based' sampling (McFadden, 1973). The regression approach, aided by computing advances, was a major step up from applying conditioning to a series of 2×2 tables involving just binary determinants.

Breslow (2003) traced these parallel, but relatively short, roots of conditional logistic regression and noted that those in one field were unaware of the statistical developments by professionals in the other. The aim of this present paper is to identify and revisit a much earlier application of conditional logistic regression. In addition, with the help of archival material, it describes the intense collaboration and calculations that preceded its publication, how the two publications came to be and the personae involved. To make the technical material in that application more accessible to modern-day readers, it begins with how the model is used and fitted today.

2. Conditional logistic regression today

Pardoe & Simonton's (2008) prediction of Academy Award winners is a useful orientation to the simplest version of conditional logistic regression. One has a 'set' of data for each year; for any given year, it has as many rows as there were nominees that year, with each row containing the vector (z) of predictors, and a 0/1 indicator (y) of whether the nominee in that row was the eventual winner. In a year in which there were a set of, say, five nominees, the probability that a specified nominee with a vector z will be the winner is taken to be

$$\exp[\beta z] \Big/ \sum \exp[\beta z'],$$

where β is the corresponding vector of regression coefficients, and the summation extends over all five nominees in the set. Thus, the after-the-fact likelihood contribution from that year is simply this same expression, with the specified z replaced by that of the nominee who did win ($y = 1$). The overall loglikelihood is the sum of the year-specific ones, and is easily maximized.

The idea of a 'set' of candidates for an award has a direct analogue in epidemiological research and in survival analysis, and software developers have exploited this commonality (Lumley, 2024). However, the likelihood can be more complicated. While there is only one winner per Oscar competition, there can be more than one case in a matched case control set; and if the survival time scale is coarse, 'ties' can occur: two or more candidates in a risk set may suffer the event of interest at the discrete time that defines that risk set.

3. Historical background

3.1. *The role of parental ages in Down's syndrome*

In the late 1950s, what we today call Down syndrome or Down's syndrome, or trisomy 21, was found to be a genetic disorder caused by the presence of all/part of a third copy of chromosome 21. It is typically associated with physical growth delays, characteristic facial features and a range of intellectual disability. A striking epidemiological feature is the large role of maternal age in the probability of its occurrence. One of the earliest investigators to document this role was the English human geneticist Lionel Penrose (Harris, 1973). Penrose's main work was on the genetics of intellectual deficit, but he had wide ranging interests. As a Quaker, he opposed war, and spent the World War II years working in Canada. In 1945, he returned and became the chair of genetics at University College, London, a post that had remained vacant since Ronald Fisher moved to Cambridge University in 1943.

He first studied the relative effects of the mother's and father's ages, and, after accounting for the high correlation in the ages, concluded that the father's age is 'not a significant factor', while the mother's age 'is to be regarded as very important' (Penrose, 1933). For further details, and Fisher's role in it, see the Supplementary Material. Penrose's next disentanglement project presented a much more difficult statistical problem, involving, again, two highly correlated suspects.

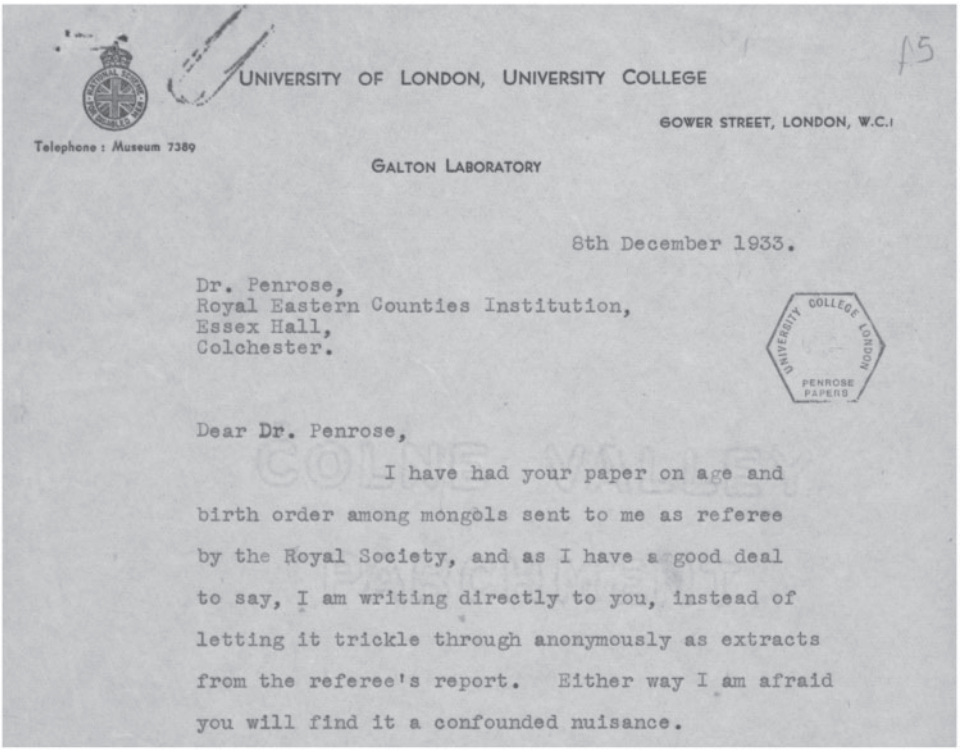
3.2. *The role of the mother's age and birth order*

Fisher had a much larger role in this. The 14 letters they exchanged between December 1933 and April 1934 show just how large it was. Also unrecognized, until now, is that the statistical analysis involved the fitting of what is known today as the conditional logistic regression model.

The model is described briefly in Penrose (1934a) and in technical detail in Penrose (1934b). To appreciate why Fisher suggested it, we begin with the first version of Penrose's manuscript, which was received by the Royal Society on November 25, 1933. Penrose's investigation

> involved the accurate determination of the maternal age at the birth of all offspring in 217 sibships, each containing one or more Down's children. The birth order was also recorded with particular care: miscarriages and stillbirths were deemed to affect the ordinal number of subsequent births, but they have been excluded from the data as presented here. It is very uncertain whether they represent offspring affected or not by Down's syndrome and I wish to include in the data only those individuals in the 217 sibships of whom it could be said with certainly that they were either Down's syndrome or not. (Penrose, 1934a)

In that version, Penrose examined differences in mean ages and mean birth ranks. After correcting for the effect brought about by the presence in the data of families of different sizes, it seemed that 'the effect of birth rank on the incidence of Down's syndrome is therefore apparently of great significance, though it is not quite as marked as the maternal age affect.' In order to assess how much of this could



UNIVERSITY OF LONDON, UNIVERSITY COLLEGE

GOWER STREET, LONDON, W.C.1

Telephone : Museum 7389

GALTON LABORATORY

8th December 1933.

Dr. Penrose,
Royal Eastern Counties Institution,
Essex Hall,
Colchester.

UNIVERSITY COLLEGE LONDON
PENROSE PAPERS

Dear Dr. Penrose,

I have had your paper on age and birth order among mongols sent to me as referee by the Royal Society, and as I have a good deal to say, I am writing directly to you, instead of letting it trickle through anonymously as extracts from the referee's report. Either way I am afraid you will find it a confounded nuisance.

Fig. 1. Beginning of Fisher's six-page letter to Penrose. The letter contains a term that is no longer used.

be an artifact of the high correlation between maternal age and birth order, he then proceeded 'to calculate the expected number of affected children occupying each birth rank on the assumption that we only know the maternal ages for these individuals and not the orders of their births'. The 'satisfactory' results of this calculation gave him 'no indication that any birth rank is more frequently occupied by a Down's syndrome child than would be expected from the mere consideration of the maternal ages at which these children are born.'

Today, thanks to the Genetics Collection in the UCL digital collections (Wellcome Library, 2024), we are able to observe a very special instance of the review process 'back then', and to assess in some detail the pioneering contributions of both the author and the reviewer.

3.3. *Fisher's pivotal role in this study*

The Penrose archives at UCL include the original of a letter sent to him less than two weeks after the manuscript was received. The carbon copy is at the University of Adelaide, Australia. The first paragraph is reproduced in Fig. 1. The second paragraph moves directly to the crux of the problem, the need to condition on the family: keeping each family as a set respects the design.

> The whole difficulty turns on the point made in section three, but that section makes it far from clear. You do not mention the essential point, that choosing families only containing [Down's syndrome cases] the proportion of [Down's syndrome cases] must be highest in the smallest families, which generally contain early, but not late children by birth rank.

The Supplementary Material may be of interest to readers who themselves have undergone peer review and are curious about the tone of Fisher's comments on Penrose's flawed approach. Still relevant today is Fisher's preamble to his suggested course of action:

> Now it seems to me that your family data are much too important for you to be satisfied with an unconvincing statistical analysis. I mean, that no one reading your paper critically will feel sure that a more exact treatment would not have yielded a different result. I may add that I entirely expect your actual conclusions to be the right ones, but that is no sufficient reason why they should not be adequately established.

The next section will describe how Penrose, with Fisher's help, followed his suggestion:

> The only convincing test for a theory, is a direct comparison between what has been observed, and what must be expected on that theory. *The appropriate theory here is, that the probability of a Down's syndrome child depends on age, in some manner unknown prior to the data, but not, given the age, on the birth rank.* As I think you already see the only relevant facts available to test this theory consist of the distribution of Down's syndrome children within families of given constitution in respect of (a) number of children recorded, (b) birth rank of these children, (c) maternal ages, and (d) number of Down's syndrome children. Families wholly Down's syndrome, like families wholly normal will give no information. [italics added]

Penrose (1934a) fitted and reported the results of what may well be the first conditional logistic regression. He shared the technical details, and the child-level data in Penrose (1934b).

4. The first conditional logistic regression: in brief

In almost the same words as in Fisher's initial letter, the essence of the model is summarized in the second paragraph of § IV of Penrose (1934a). That section first paraphrased Fisher's criticisms, referring to them as 'inaccuracies in the statistical treatment I have so far employed here'. Thus,

> to avoid these sources of ambiguity the data have been subjected to analysis by an entirely different method which was suggested by Professor R. A. Fisher. By use of this new process we are able, after a single complex reconstruction, to compare the observed number of Down's syndrome cases in any given birth rank with the number which is to be expected on the hypothesis that the probability of a Down's syndrome child depends upon maternal age (in some manner unknown prior to the data) but not, given age, upon birth rank.

In a cryptic passage that puzzled me for years, and that I explain in the next section, Penrose then presents a model, of an as-yet-to-be-specified functional form, for a specific pair of maternal ages. He adopted Fisher's symbol x to denote a relative odds. Here, I have replaced it by the Greek letter ω, and replaced his letter S for the sum by today's $\sum$.

> Let us suppose that there are a number of families containing only two children born at the maternal ages of 32 and 42, respectively, and that one child in each family has Down's syndrome. Call p_{32} and p_{42} the [age-specific] probabilities that a Down's syndrome child is born at these maternal ages. The frequencies of families which have the Down's syndrome child at age 32 to those which have the Down's syndrome child at 42 will be in the ratio $p_{32}/(1-p_{32}) : p_{42}/(1-p_{42})$, or, say, $\omega_{32} : \omega_{42}$ where ω is proportional to [the odds] $p/(1-p)$. In any such family the expectation that the child born at 32 is a, or in this case, the, child with Down's syndrome is $\omega_{32}/(\omega_{32}+\omega_{42})$.

He then explains that, in general, this means that

> for families containing only one Down's syndrome child, the expectation that a child whose (relative odds) was ω, is the affected one is $\omega/\sum\omega'$, where $\sum\omega'$ is the sum of the values of ω for the maternal ages of [each of the] children in the family.

With the children in the family regarded as a set, this expectation has the same structure as the conditional probability defined in § 2; thus, the likelihood contribution also has the same structure. Fisher had not yet specified a functional form for the age specific p.

The more complex expressions for the expectations involving families with more than one Down's syndrome child were left for the technical paper, and will be addressed in the next section of the

Table 1. *Trial ω values, and age-specific fitted ('calculated') frequencies of Down's syndrome (DS) children*

	Maternal age group	15–19	20–24	25–29	30–34	35–39	40–44	45–49
	Observed no. of normal children	10	114	199	228	170	67	15
	Observed no. of DS children	3	13	14	27	64	81	22
Trial no.								
1	ω values	83[4]	31[2]	19[**1**]	33[2]	104[5]	321[15]	407[20]
	Calculated no. of DS children	3.69	16.87	19.50	29.11	59.48	76.99	18.35
⋮								
7	ω values	22[4]	10[2]	6[**1**]	19[3]	88[15]	296[50]	558[90]
	Calculated no. of DS children	2.98	12.87	13.82	26.58	64.14	81.46	22.17
.....	.....							
`clogit`	Scaled ω values	[3.47]	[1.59]	[**1**]	[2.96]	[13.19]	[43.62]	[81.53]

Source: Page 440 of Penrose (1934a).
Since the ω values are relative odds, values in brackets have been scaled so that the lowest risk age group (25–29) serves as the reference category, with a scaled odds of **1**:1. See Penrose (1934b) for how he chose the ω's for each trial. The scaled ω values fitted in five iterations by the `clogit` function in the `R survival` package (R Development Core Team, 2024) yielded calculated frequencies that were, in absolute terms, within 10^{-9} of the observed ones.

present essay, along with why Fisher moved from probabilities, p, to relative odds, ω. Before he addressed the form of the ω function, Penrose stated the operational criterion of fit, which in his review, Fisher had simply stated, without justification: the best-fitting ω values will be those where the 'number of Down's syndrome children observed at any given maternal age tallies with the sum of the expectations attributed to each child at that maternal age.' As shown below, the ω's that satisfy this estimating equation are maximum likelihood estimates.

'In order to simplify the arithmetic', but also probably because of 'the considerable curvature of the regression on age', he chose to model the age function, what Fisher simply referred to as the ω '*series*', with seven parameters, one per five-year-wide age bin. Without describing how he chose his successive approximations, he reported the ω values shown in Table 1. The technical paper tells us he stopped after seven iterations, when the fitted frequencies were within 1% of the observed ones. Since the ω scale is necessarily relative rather than absolute, it is easier to appreciate their range when the fitted ω's are scaled so the lowest value is 1.

With these fitted age effects, he calculated the 17 birth-order-specific fitted numbers of Down's syndrome children and displayed them as columns of his Table IV, side by side with the 17 observed marginal frequencies. Since cases were scarce at higher birth orders, he formally examined the fit in five grouped birth ranks, and found that the agreement between the theoretical and observed numbers was 'satisfactory'.

Using the raw data in selected sibships, the calculations and other technical details are nicely laid out in Penrose (1934b). A number of them, and their 'provenance', are of modern interest.

5. The first conditional logistic regression: technical details

5.1. *Why is the model in terms of relative odds?*

Fisher wished to provide prospective probabilities for the population, and so his ultimate estimand was the probability (p) as a function of age. Why he switched from p to $\omega = p/(1-p)$, is not as 'evident' to us as it was to him. Characteristically, he did not justify this odds, or log odds, scale. As it turns out, the choice derives directly from the outcome-based study design: the data come from families and these families have at least one Down's syndrome child. And so, when one conditions on this fact, the within-family conditional distribution, within these specially

selected families, is more easily written in terms of the relative odds, the ω, than the probabilities. This can be better understood if we revisit Fisher's example of a two-child family.

The entities p_{32} and p_{42} in that example refer to the absolute age-specific probabilities of a Down's syndrome (D) child in the population that the sibships arose from. Denote by $q_{32} = 1 - p_{32}$ and $q_{42} = 1 - p_{42}$ the corresponding probabilities of a normal (N) child. Thus, among two-child families in this source population, with each family providing two independent Bernoulli trials with probabilities p_{32} and p_{42}, respectively, the relative numbers of each of the four possible family compositions are

$$N_{32}N_{42} : q_{32}q_{42}, \qquad D_{32}N_{42} : p_{32}q_{42}, \qquad N_{32}D_{42} : q_{32}p_{42}, \qquad D_{32}D_{42} : p_{32}p_{42}.$$

Only families with a discordant pattern are included in the series; thus, among these informative sibships the frequencies of the two compositions $D_{32}N_{42} : N_{32}D_{42}$ will be in the ratio

$$p_{32}q_{32} : q_{32}p_{42} \quad \text{or} \quad (p_{32}/q_{32}) : (p_{42}/q_{42}).$$

When, ultimately, Penrose referred to the fitted relative odds, the fitted ω, as the relative probabilities of a Down's syndrome child, he was most likely exploiting the fact that even at the highest maternal ages, the population-level probabilities do not exceed a few percent, so that the probability p and the odds $p/(1 - p)$ are numerically close to each other.

5.2. *Sibships containing two Down's syndrome children*

Seven of the 217 sibships contained two affected children each. In today's survival analysis terminology, these are equivalent to risk sets containing 'ties'; in case-control studies they correspond to matched sets with multiple cases per set, a duality exploited by Lumley (2024).

As can be seen in the Supplementary Material, Penrose's description of the expression for the expected number of Down's syndrome children at a given maternal age is minimally adapted from the wording on page 4 of Fisher's review, and therefore exasperatingly cryptic. It becomes 'more evident' if we use an actual four-child sibship, $D_{31}, N_{34}, N_{38}, D_{44}$, that Penrose used to illustrate the calculation. Then, the expectation at, say, age 31 is

$$\frac{\omega_{31}(\omega_{34} + \omega_{38} + \omega_{44})}{\omega_{31}\omega_{34} + \omega_{31}\omega_{38} + \omega_{31}\omega_{44} + \omega_{34}\omega_{38} + \omega_{34}\omega_{44} + \omega_{38}\omega_{44}}.$$

The 'multiplicities' in the Fisher–Penrose context arise naturally. Some of the ties addressed in the discussion of Cox's 1972 paper (Peto, 1972), and subsequently (Breslow, 1974; Efron, 1977; Gail et al., 1981; Kalbfleisch & Prentice, 2002), are merely the result of a coarse time scale.

5.3. *The form and fitting of the 'relative odds as a function of age' model*

Fisher suggested grouping the ages in

> five or three year groups in obtaining trial values, and increasing or decreasing these for individual years, in proportion to the actual ratio of Down's syndrome children observed to expected, in each year after the first trial.

Penrose stayed with seven five-year bins, and piecewise constant ω's, throughout. As with much of his writing, Fisher did not justify his criterion for a good fit, or derive its implications. Characteristically, it was 'evident' to him what should 'tally with what' in what are effectively seven estimating equations:

> Given the series of (7 ω) values, therefore, for all ages, the expectations of each recorded child being a Down's syndrome child can be set down […]. The assigned ω values will be correct when the observed numbers at each age tally with those expected.

These balancing equations resemble a method-of-moments approach, but they also follow directly from maximizing the likelihood in §2, with the ω for a given child represented as $\exp(\sum_{j=1}^{j=7} \beta_j z_j)$,

where β_j is the log of the relative odds for a child in maternal age category j, and the binary variate z_j indicates if that child belongs in category j. To see this, suppose that the affected child is in category 3, and denote the sum of the ω's in the sibship as Ω. Thus, the likelihood contribution by this sibship is $\exp[\beta_3] / \Omega$. The loglikelihood contribution is $\beta_3 - \log(\Omega)$, and its partial derivative with respect to β_3 is $1 - d\log(\Omega)/d\beta_3$. Two applications of the chain rule for derivatives and a little calculus show this to be $1 - E_3$, where $E_3 = \omega_3/\Omega$ is the expected number of affected children at position 3. Summed over all sibships, the derivative becomes the number of affected children occupying age bin 3 minus the sum of the expected numbers of affected children in that age bin. Setting each of the partial derivatives to zero yields the estimating equations just stated by Fisher in 1933, but formally derived in equations 12–15 of Cox (1972).

Having arrived at close-to-maximum-likelihood coefficients for his maternal-age-only model, Penrose could now test his theory; see the Supplementary Material, which also shows how we might test Penrose's theory today.

6. Discussion

This 'archeological' item is of both historical and modern interest. To begin with, the logistic form of Fisher's no-name model was developed from first principles; it reflected the constraints in the family-based data collection, but its parameters reflected what occurs in the source population. This has current relevance to the ongoing, and sometimes passionate, debate about parametric models for binary data, and whether to prefer odds ratios or risk ratios, or indeed something else. In this application, modelling the odds was a practical way to reflect the design.

The investigation it was applied to is a vivid and engaging and, with multiple cases in some sibships, comprehensive example that even someone new to conditional logistic regression can easily relate to. Mapping the variable (age) into seven 'variates' made it a model that was quite large for its day.

It is not surprising that Fisher and Penrose recognized the need for a conditional approach: as geneticists, they were aware of the distortions that arise from the 'outcome-based' sampling that begins with the affected family member, or 'index' person.

Although the word 'likelihood' was not mentioned, it was central to how Fisher linked the model for the source population and the data from the studied families, and undid this distortion.

In most pioneering statistical works all we get to read is the published version, with little indication of how it came to be, or the personae involved. In this instance, apart from one possible face-to-face meeting they may have had, we have a written record of their extensive exchanges. It bears out the legendary keen insight and intuition, and sometime cryptic explanations, that characterize Fisher's written works. The correspondence with Penrose gives the impression that, contrary to the persona displayed in disputes with his critics, Fisher was a warm and collaborative colleague to those whose work he respected.

Before we celebrate how far we have come in statistical methods/computing in the last 90 years. we might wish to reflect on the price of this progress and on what we have forgotten or missed in understanding along the way. For example, by simply using the `clogit` program to fit Penrose's seven-parameter model for the ω values we miss the fact that the seven sufficient statistics are the numbers of Down's syndrome children in each of these seven maternal age categories. Equating the partial derivatives of the loglikelihood to zero results in seven balancing equations that equate these sufficient statistics to their seven expected/fitted values. Finding this balance requires an iterative search. And by simply reading off the standard errors, we miss the fact that the precision is a function of the within-sibship (co)variation of the seven indicator variables.

Both our understanding of the aetiology of Down's syndrome and its various genotypes, and its epidemiology (de Graf et al., 2015) have changed considerably in the past 90 years. Nevertheless, it can be instructive to study the first application of a statistical model, particularly when it was developed by luminaries in both statistics and in the subject matter area.

Supplementary material

The Supplementary Material contains a description of the dataset extracted from the two 1934 papers and further details.

Note regarding the references: the titles of L. S. Penrose's articles contain terms no longer used.

References

Breslow, N. E. (1974). Covariance analysis of censored survival data. *Biometrics* **30**, 89–100.

Breslow, N. E. (2003). Are statistical contributions to medicine undervalued? *Biometrics* **59**, 1–8.

Breslow, N. E., Day, N. E., Halvorsen, K. T., Prentice, R. L. & Sabai, C. (1978). Estimation of multiple relative risk functions in matched case-control studies *Am. J. Epidemiol.* **108**, 299–307.

Cox, D. R. (1972). Regression models and life tables (with discussion). *J. R. Statist. Soc.* B **34**, 187–220.

de Graaf, G., Buckley F. & Skotko B. G. (2015). Estimates of the live births, natural losses, and elective terminations with Down syndrome in the United States. *Am. J. Med. Genet. Part A* **167**, 756–67.

Efron, B. (1977). The efficiency of Cox's likelihood function for censored data. *J. Am. Statist. Assoc.* **72**, 557–65.

Gail, M. H., Lubin, J. & Rubenstein L. V. (1981). Likelihood calculations for matched case-control studies and survival studies with tied death times. *Biometrika* **68**, 703–7.

Harris, H. (1973). Lionel Sharples Penrose. 1898-1972. *Biogr. Mem. Fellows R. Soc.* **19**, 521–61.

Kalbfleisch, J. D. & Prentice, R. L. (2002). *The Statistical Analysis of Failure Time Data*. New York: John Wiley.

Keogh, R. H. & Cox, D. R. (2014). *Case-Control Studies*. Cambridge: Cambridge University Press.

Liddell, F. D. K., McDonald, J. C., Thomas, D. C. & Cunliffe, S. V. (1977). Methods of cohort analysis: appraisal by application to asbestos mining. *J. R. Statist. Soc.* A **140**, 469–91.

Lumley, T. (2024). `clogit` documentation in `R` `survival` package. https://CRAN.R-project.org/package=survival [last accessed 7 August 2024].

McFadden, D. L. (1973). Conditional logit analysis of qualitative choice behavior. In *Frontiers in Econometrics*, Ed. P. Zarembka, pp. 105–42. New York: Academic Press.

Pardoe, I. & Simonton, D. K. (2008). Applying discrete choice models to predict Academy Award winners. *J. R. Statist. Soc.* A **171**, 375–94.

Penrose, L. S. (1933). The relative effects of paternal and maternal age in mongolism. *J. Genet.* **27**, 219–24.

Penrose, L. S. (1934a). The relative aetiological importance of birth order and maternal age in mongolism. *Proc. R. Soc. B* **115**, 431–50.

Penrose, L. S. (1934b). A method of separating the relative aetological effects of birth order and maternal age, with specific reference to mongolian imbecility. *Ann. Eugen.* **6**, 108–32.

Peto, R. (1972). Discussion of regression models and life-tables. *J. R. Statist. Soc.* B, **34**, 187–202.

Prentice, R. L. & Breslow, N. E. (1978). Retrospective studies and failure time models. *Biometrika* **65**, 153–8.

R Development Core Team (2024). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. ISBN 3-900051-07-0, http://www.R-project.org.

Wellcome Library (2024). Birth order and Down's syndrome correspondence. http://wellcomelibrary.org/player/b20222087.

[*Received on* 23 *February* 2024. *Editorial decision on* 13 *June* 2024]

Biometrika (202x), **10x**, x, *pp*. 1–7
Printed in Great Britain Advance Access publication on xx July 20xx

Supplementary material for

"Studies in the history of probability and statistics. LI: the first conditional logistic regression"

BY J. A. HANLEY

Department of Epidemiology, Biostatistics and Occupational Health, McGill University, Montréal, Québec H3A 1G1, Canada

james.hanley@mcgill.ca

SUMMARY

This supplement expands on the above-mentioned Biometrika article. That article describes a conditional regression model used in two 1934 papers written by Lionel Penrose, but with considerable input from the referee, R A Fisher. Section 1 of this supplement describes an earlier paper (Penrose, 1933), where the topic was the role of mother's age and father's age in the probability that their child will have Down's Syndrome. In it, Penrose concluded that the father's age did not matter, and so he moved on to topic of mother's age and birth order. The two 1934 papers (one substantive (Penrose, 1934a), one methodological (Penrose, 1934b)) on mother's age and birth order are the topic of the Biometrika article. The dataset we extracted from the two 1934 papers, and are making available, is described in section 7.

1. PENROSE'S 1933 PAPER ON PARENTAL AGES

Penrose's conclusions were based on data he had collected on 150 families, each one containing at least one child with Down's syndrome. He applied two different statistical methods to the data on the 154 Down's syndrome and 573 normal children: (i) partial correlations, previously used in animal studies by the American geneticist Sewall Wright, and (ii) other-parent-adjusted age-differences between Down's and normal children, suggested to him by Ronald Fisher. [For more details, see the unpublished manuscript "Prob[Down syndrome | parental ages]: Statistical Sudoku and re-analyses of data from 1933" by Hanley and Roy.]

In his 1932 letter to Fisher thanking him again for his advice on how to disentangle the parental age effects, Penrose added

> I am also working on the relative effects of maternal age and place in family: a problem which intrigues me even more than that concerned with paternal age. Sooner or later I would like to send you the results I have got. Here, again, the maternal age seems to be the significant factor, the order of birth having no effect.

2. INFERIOR APPROACH IN FIRST SUBMISSION OF PENROSE'S 1934 PAPER

Penrose began by reporting a 6.1 year difference in the mean ages of mothers of Down's syndrome and normal children, and a 1.04 difference in their mean birth ranks. Following a correction for the effect brought about by the presence in the data of families of different sizes,

the 'displacement' in birth order increased to 1.98 ordinal places, indicating that "the effect of birth rank on the incidence of Down's syndrome is therefore apparently of great significance, though not it is not quite as marked as the maternal age affect." In order to assess how much of this could be an artifact of the high correlation between maternal age and birth order, he then proceeded "to calculate the expected number of affected children occupying each birth rank on the assumption that we only know the maternal ages for these individuals and not the orders of their births". The "satisfactory" results of this calculation gave him "no indication that any birth rank is more frequently occupied by a Down's syndrome child than would be expected from the mere consideration of the maternal ages at which these children are born."

3. FISHER'S FURTHER CRITICISM OF THIS INFERIOR APPROACH

Readers may wish to compare the tenor of Fisher's comments on Penrose's (initial) approach with the tenor of modern day comments:

> In section four, you make out a reconstruction of the expected distribution of Down's syndrome children on the assumption that neither birth rank nor age influences the incidence. From this reconstruction, you find expected numbers in each birth rank, but you do not do the same for ages, or discuss whether or not there is also a contribution to the regression on age in the reconstruction. Two other points on this part of the paper are, the absence of any discussion of the effect on the sampling error for using this reconstruction (I imagine it is greatly to reduce it) , and [...] In the interesting question of first births, you do not discuss the considerable curvature of the regression on age.

4. THE 1934 PRESENTATION OF EXPECTATIONS IN FAMILIES WITH 2-AFFECTED CHILDREN

Seven of the 217 sibships contained two Down's syndrome children each. In today's survival analysis terminology, these are equivalent to risksets containing 'ties'; in case-control studies they correspond to matched sets with multiple cases per set. [The `Stata` manual explains that their `cmclogit` command fits McFadden's choice model with just 1 choice per set, which is a specific case of the more general conditional logistic regression model fit by `clogit`.] Penrose's description of the expression for the expected number of Down's syndrome children at a given maternal age is adapted from the wording in page 4 of Fisher's December 6 letter. In this online supplement, I retain the original notation, where the relative odds is denoted by x, rather than the more modern ω I use in the main text.

> For families containing two Down's syndrome children, the expectations of Down's syndrome children at each place will be $xS'(x)/SS(xx)$, adding up to two, where $S'(x)$ is the sum of the other values, and $S(xx)$ stands for the sum of all the products, two at a time.

> For families containing two Down's syndrome children, the expectations of Down's syndrome children at each place will be $xS'(x)/\sum\sum(xx')$, adding up to two, where $S'(x)$ is the sum of the other values, and $\sum(xx')$ stands for the sum of all the products, taken two at a time.

Table 1. *Trial x values, and age-specific fitted ('calculated') frequencies of Down's syndrome (DS) children*

	Maternal Age Group :	15–19	20–24	25–29	30–34	35–39	40–44	45–49
	Observed no. of normal children :	10	114	199	228	170	67	15
	Observed no. of DS children :	3	13	14	27	64	81	22
Trial No.								
1	x values :	83[4]	31[2]	19[**1**]	33[2]	104[5]	321[15]	407[20]
	Calculated no. of DS children :	3.69	16.87	19.50	29.11	59.48	76.99	18.35
2	x values :	46[4]	22[2]	12[**1**]	28[2]	106[10]	319[30]	467[40]
	Calculated no. of DS children :	3.20	15.48	16.03	28.29	63.00	78.46	19.55
3	x values :	19[4]	8[2]	5[**1**]	14[3]	82[15]	330[70]	542[110]
	Calculated no. of DS children :	3.04	12.52	13.39	23.27	53.85	86.05	21.87
...								
7	x values :	22[4]	10[2]	6[**1**]	19[3]	88[15]	296[50]	558[90]
	Calculated no. of DS children :	2.98	12.87	13.82	26.58	64.14	81.46	22.17
.....	.....							
`clogit`	scaled x values :	[3.47]	[1.59]	[**1**]	[2.96]	[13.19]	[43.62]	[81.53]

Expanded version of Table 1 in the article that this online material refers to, but using the original Fisher-Penrose notation where x, rather than ω, denotes the relative odds. Source: Page 440 of Penrose (1934a). Since the x values are *relative* odds, values in parentheses have been scaled so that the age-group with the lowest risk (25–29) serves as the reference category, with a scaled odds of **1**:1. Penrose limited himself to 1-3 digit integer values of x, and so, in the same spirit, the scaled values have also been liberally rounded. See Penrose (1934b) for how he chose the x's for each trial. The scaled x values fitted (in 5 iterations) by the `clogit` function in the `R survival` package yielded calculated frequencies that were, in absolute terms, within 10^{-9} of the observed ones. Note that there are slight discrepancies between the frequencies in Penrose (1934a) and Penrose (1934b), and between the number of unaffected children reported in both articles (807) and the number of unaffected children that appear in the Appendices (806).

5. How Penrose examined the (residual) birth-order effects

Penrose's theory was that " the probability of a Down's syndrome child depends on age [...] but not, given the age, on the birth rank." To test it, he aggregated the fitted expectations for each of the 1031 children by birth order. To Fisher, "the lack of systematic deviations" of the 17 birth-order-specific numbers of Down's syndrome children from these calculated frequencies in Table II in Penrose (1934b) was "very reassuring", but "it will be still more interesting to compare these with the random sampling deviations to be expected subject to the rather severe restrictions imposed." For formal testing purposes, Penrose collapsed the frequencies into the 5 birth-order categories shown in Table 2. In order to show that the agreement between the theoretical and observed numbers was "satisfactory", he calculated a separate standard error for each of the five (Observed − Expected) discrepancies, and noted that each discrepancy was within 1 standard error of zero. His numbers are shown in Table 2.

6. Penrose's standard errors for the birth-order-specific residuals

The complexity of the five standard errors (Table II) stemmed in part from

> the fact that the totals of the 7 columns (of the 17×7 table of fitted expectations) have been fixed by the process of fitting x values, the object of which was to make these totals correspond as closely as possible to the observed numbers.

Table 2. *Test of theory that birth order is not an aetiological factor: observed (O) and expected (E) numbers of Down's syndrome children, and standard errors*

				Standard Error	
Birth rank	O	E	Difference	1934	2024
1st	26	23.97	+2.03	2.68	2.9
2nd or 3rd	55	57.56	-2.56	3.85	3.9
4th, 5th or 6th	59	61.98	-2.98	4.07	4.1
7th to 10th	61	58.37	+ 2.63	3.41	3.3
11th to 17th	23	22.14	+0.86	1.84	1.8
Total	224	224.02			

Source: Table X page 122 of Penrose (1934a). E's computed from fitted x function in Table 1. 2024 Standard errors estimated by simulation (see text).

To correct the standard errors for "the rather severe restrictions imposed" (i.e., to "allow[...] for the fixing of the totals of the [7] columns"), Fisher showed Penrose how to calculate the several different types of variances and covariances involved, and numerically inverted a 7×7 matrix for him.

The descriptions and results take up eight pages in Penrose (1934b), and are not easy to follow. Today, instead of following them, we can use our considerable computing power to derive a simulation-based approximation to the "restricted" random sampling deviations that the fitting imposes on the 17 (or 5) birth-order residuals. To do so, we began with the `clogit`-fitted x values at the foot of Table 1. We used these, together with the maternal ages in the sibship, to compute the expected number of Down's syndrome children at each position in the sibship. For each of the 217 sibships, we used these based-on-age-only expectations to randomly allocate the case(s) of Down's syndrome among the childrens' positions in the sibship. In 544 of the 1 million sets of random allocations carried out, the overall numbers of Down's syndrome in the 7 age bins very closely matched the corresponding numbers in Penrose's dataset. (We defined 'very closely' as an excess of at most 1 Down's syndrome case in at most 1 bin, counterbalanced by a deficit of at most 1 in at most 1 other bin.) The 5 estimated standard errors shown in the '2024' column in Table 2 were calculated using the standard deviations of the birth-order-specific numbers of Down's syndrome children across these 544 simulated datasets.They are similar in magnitude to those calculated theoretically by Penrose and Fisher.

Today, with computational considerations no longer an issue, we would approach this test of the theory differently (see next section). Thus, the complex and tedious calculations that went into these five standard errors, and the description of it that takes up 8 pages in the methods paper will not be discussed in detail here. It was, however, a set of calculations than Fisher not only insisted on, by also had a hand in himself.

7. FROM BRUNSVIGA (1934) TO MACBOOK AIR (2024): DATASET AND HARDWARE

The Fisher-Penrose correspondence suggests that they shared the data on 217 cards, one per sibship, probably in the same format that the data appeared in the Appendices to the two 1934 articles. In 2024, we might arrange this information electronically in an `R` data frame, with 224 + 807 = 1031 rows, one per child, and 4 columns: sibship number, Downs (1) or Normal (0), maternal age, and birth-order; such a `csv` file, is available at `https://jhanley.biostat.mcgill.ca/Penrose`.

In his study at home, Penrose had a "handle-powered desk calculating machine called the Brunsviga. This was the latest thing in computing technology at the time" and his son, then aged 5-6, remembers how "it made a very satisfying crunching noise when you turned the handle to do a big multiplication or division sum" (Penrose, 2007). Fisher would have had a computing machine at UCL; and in his March 12 letter explaining how to compute the entries in the 29×29 Table V of Penrose (1934b) that corresponded to different family configurations, he told Penrose, "I have run out these 6 families out at home with logarithm tables." The analyses in 2024 were carried out on a MacBook Air machine.

8. From Brunsviga (1934) to MacBook Air (2024): Software, and fitted conditional model involving just maternal age

Fisher's procedure has already been described, and the fits it produced are given in Table 1. The last row of Table 1 shows the x values fitted by the `clogit` function in the `survival` package in `R` version 4.3.1. The duality mentioned in the documentation is of note

> [`clogit`] [e]stimates a logistic regression model by maximising the conditional likelihood. Uses a model formula of the form `case.status` $\sim$ `exposure + strata(matched.set)`. The default is to use the exact conditional likelihood [the one Fisher used when dealing with two Down's syndrome children in the same sibship.]
>
> It turns out that the log likelihood for a conditional logistic regression model equals that from a Cox model with a particular data structure. [...] When a well tested Cox model routine is available many packages use this 'trick' rather than writing a new software routine from scratch, and this is what the `clogit` routine does. In detail, a stratified Cox model with each case/control group assigned to its own stratum, time set to a constant, status of 1=case 0=control, and using the exact partial likelihood has the same likelihood formula as a conditional logistic regression. The `clogit` routine creates the necessary dummy variable of times (all 1) and the strata, then calls `coxph` (Lumley, 2024).

Penrose stopped after 7 trials of x, when the 7 fitted frequencies were within 1% of the observed ones. [Fisher and Penrose discussed at some length whether this was close enough to not affect the conclusions.] Today, iteration in the `coxph` routine continues until the relative change in the log partial likelihood is less than some pre-specified 'epsilon', or the absolute change is less than the square root of this. In Perrose's dataset, with the default epsilon, it converges in 5 steps, and produces 7 fitted frequencies that, in absolute terns, are all within 10^{-8} of the 7 observed ones.

9. Testing Penrose's theory today

Today, most data-analysts would test Penrose's theory *directly*, by adding some representation of birth order to a conditional logistic regression model that already included maternal age.[1] Adding Penrose's 5-level factor (4 parameter) representation increases the log-likelihood in Penrose's age-only by a mere 1.6 units, whereas adding a (1 degree of freedom) linear representation

[1] It is possible, in a regular regression context where one adds a regressor x_2 to a model that already contains x_1, to fit the two β's in the joint model using a two-step procedure. In the the first step one calculates the residuals unexplained by x_1 alone; In the second step, these residuals are regressed, not on x_2 itself, but on the portion of x_2 not already explained by x_1. i.e., on the $x_2|x_1$ residuals.

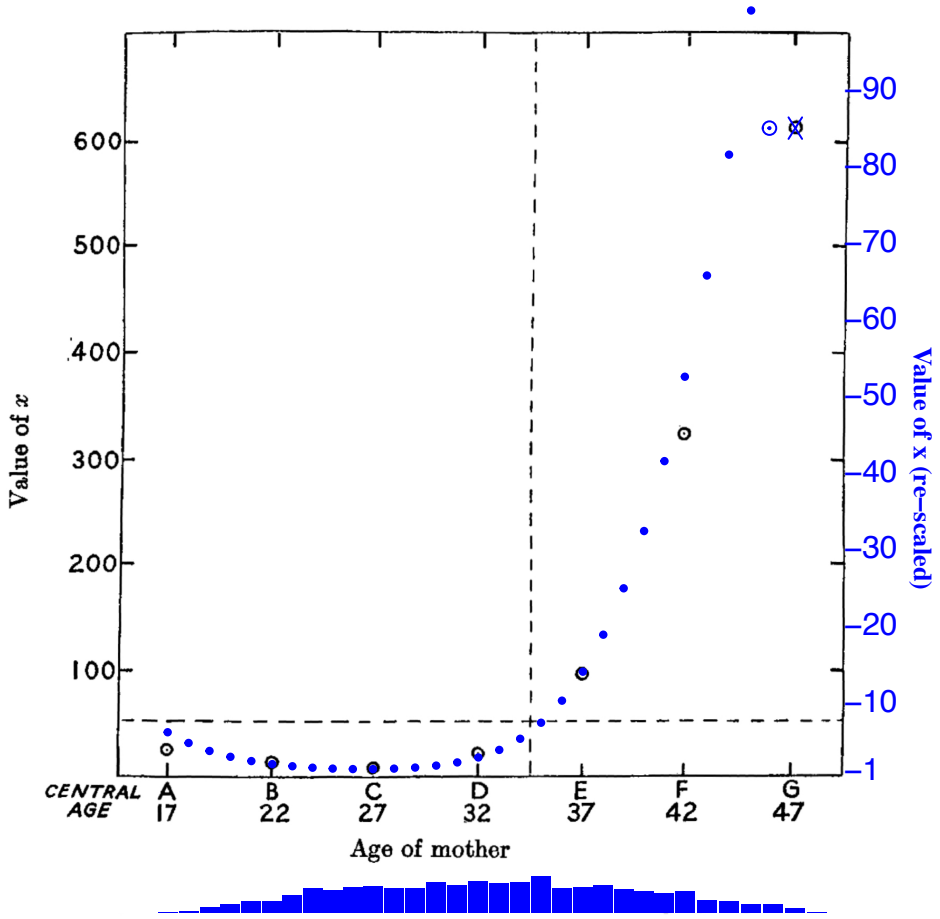


Fig. 1. Reproduction of Figure in Penrose 1934b, entitled "final estimate of values of x for maternal age groups," and, superimposed on it, in blue, (i) the distribution of the maternal ages (ii) a scale in which the relative probabilities begin at 1 (iii) a smooth x curve fitted with a 3 degrees of freedom spline by the `clogit` function, and (iv) a horizontal relocation of the fitted x value that Penrose had plotted above a 'central' age of 47 so that its stands above age 45.8: of the 37 children in that '45–49' age bin, some 15, 14, 7 and 1 of their mothers were aged 45, 46, 47 and 48 respectively. Since he regarded the fitted x values as population based estimates, Penrose compared them with ones from a large series of cases, where the rescaled values were 0.8, 0.9, 1, 2.2, 5, 19, and 58.

increases it by some 2.3 units. However, this increasing-probability-with birth-order may be an artifact of the wide age bins, especially at the upper ages, that Penrose used: when 4- or 3-years-wide bins are used, the addition of birth order no longer improves the log-likelihood substantially.

After Penrose's "first shots" at x, it became very clear to Fisher that "x increases very rapidly from 35 years upwards, so that the evidence [the birth-order residuals] must be somewhat distorted by a change in x value within your [age-]group." Thus, he suggested "unless either you take smaller age groups for the older ages , or if you prefer, take a smooth curve with two or three constants to give the x values for the individual years in this region." Penrose stayed with seven 5-year bins, and, as is seen in Fig. 2, he plotted the final x values over the midpoints of these age groups, rather than their statistical centres of gravity. Fisher's suggested smooth curve (see Fig.

2), easily computed today, shows considerable statistical parsimony over Penrose's 6-parameter model.

Given what we have seen so far, it is no surprise that the addition to this parsimomious age-only model of a spline representation of birth order does not improve it. Nor was this a surprise to Fisher

> in fact one might say *a priori* that birth rank is in fact so closely associated with maternal age that even a biggish lot of data gives not much scope for one cause to manifest Itself when the other is eliminated.

10. Less-appreciated issues in conditional logistic regression

With the ease with which this model can be fitted today, a number of subtleties specific to the *conditional* version of logistic regression are not widely appreciated. The first of these is the issue of correlated predictors raised by Fisher. Penrose thought that the correlation between birth order and maternal age of "only" 0.66 left some scope to separate the effect of maternal age from the effect of birth order. Although Fisher does not specify his measure of the association between birth rank and maternal age, his "so closely' phrase suggests he may well have appreciated that it is not the 'crude' or marginal correlation of 0.66 that matters, but the *within-sibship* correlation. In Penrose's dataset, the 'local' or within-sibship correlation is 0.94! [Thus, in the (albeit insufficient) model with age and birth order each represented as a 1 degree of freedom (linear) term, the correlation between the 2 fitted coefficients is -0.92!]

The second, and related, issue is the precision with which the coefficients in a conditional logistic regression are estimated. The two determinants are best appreciated by examining the form of the information matrix, conveniently given in equations 16 and 17 in Cox (1972). Applied to Penrose's example, it is the sum of (a special version of) the 217 within-sibship variance-covariance matrices of the variates in the linear predictor. When calculated at the null (as in a score test), all members of the sibship are weighted equally in computing the within-sibship covariance matrix; at a non-null value of the parameter(s), the members are weighted according to the exponentiated values of their linear predictors. In what Cox (1972) called "an 'exponentially weighted' form of sampling", an member with a weight w is the statistical equivalent of w persons in the reference category (Hanley, 2008). Thus the larger the number the sibships, the larger the variances, and the smaller the correlations of the predictors within each sibship, the greater will be the amount of information, and the better the precision or power.

References

Cox, D. R. (1972). Regression models and life tables (with Discussion). *J. R. Statist. Soc.* B **34**, 187–220.

Hanley, J. A., (2008) The Breslow Estimator of the Nonparametric Baseline Survivor Function in Cox's Regression Model: Some Heuristics. *Epidemiology*, **19** 101–102.

Hanley, J. A. & Roy, S., (2024) Prob[Down syndrome | parental ages]: Statistical Sudoku and re-analyses of data from 1933. *under review at JRSS A*

Lumley, T. (2024) `clogit` documentation in `R survival` package.

Penrose, L. S. (1933). The relative effects of paternal and maternal age in mongolism. *Journal of Genetics* **27** 219–224.

Penrose, L. S. (1934a). The relative aetiological importance of birth order and maternal age in mongolism. *Proceedings of the Royal Society B* **115** 431–450.

Penrose, L. S. (1934b). A method of separating the relative aetological effects of birth order and maternal age, with specific reference to mongolian imbecility. *Annals of Eugenics* **6** 108–132.

Penrose, O. (2007). A beautiful method of analysis. In *Fifty years of human genetics: a Festschrift and liber amicorum to celebrate the life and work of George Robert Fraser*. Mayo, O & Leach, C. eds. ISBN 9781862547537. Adelaide: Wakefield Press, pp. 434–451.

[*Received on* 24 $February$ 2024. *Editorial decision on* dd mmm 2024]