APPENDICES

to submitted manuscript

"Lionel Penrose's statistical consultant: and lessons from the statistical 'sudoku' they left us"

A 'Sudoku' solutions using integer linear programming

In the following toy example, from the `lp` function in the `lpSolve` package in `R`, the task is to maximize the *objective function* $\underline{1}x_1 + \underline{9}x_2 + \underline{1}x_3$ with respect to the 3 unknowns, x_1, x_2, and x_3, subject to the 2 constraints

$$\begin{array}{rcrcrcr} \underline{1}\, x_1 & + & \underline{2}\, x_2 & + & \underline{3}\, x_3 & \leq & 9, \\ \underline{3}\, x_1 & + & \underline{2}\, x_2 & + & \underline{2}\, x_3 & \leq & 15. \end{array}$$

It can be accomplished by supplying the following 5 arguments to the `lp` function:

- A character string giving direction of optimization: "min" (default) or "max."
- The numerical vector of coefficients, {1, 9, 1} that specify the objective function.
- The 2×3 matrix of coefficients that specify the constraints.
- The vector of character strings giving the direction of each constraint, here $\{\leq, \leq\}$,
- The vector of numeric values for the right-hand sides of the constraints, here {9, 15}

The maximum of the objective function (40.5) achieved at $\{x_1, x_2, x_3\} = \{0.0, 4.5, 0.0\}$.
In the Penrose data the reported 42 (rows, r) $\times$ 31 (columns, c) frequency table F is a sum of the 2 unreported sub-tables N and D :

$$F_{r,c} = N_{r,c} + D_{r,c}.$$

We wish to determine N and D subject to the constraints which are given in the reported marginals of the table, namely (with dots denoting marginal totals)

$$\texttt{OVERALL}: \quad N_{\cdot\cdot} = 573, D_{\cdot\cdot} = 154;$$

$$\texttt{ROWS}: \; N_{1\cdot} = 1, D_{1\cdot} = 0; \quad \ldots \; ; \quad N_{42\cdot} = 1, D_{42\cdot} = 0;$$

$$\texttt{COLS}: \; N_{\cdot 1} = 1, D_{\cdot 1} = 0; \quad \ldots \; ; \quad N_{\cdot 31} = 3, D_{\cdot 31} = 1.$$

Initial checks:

`ROWS`: these were consistent with `OVERALL`.
`COLS`: the $\{N_{\cdot c}\}$ summed to 57$\underline{5}$ and the $\{D_{\cdot c}\}$ to 15$\underline{2}$.
Initial efforts at balancing of the system

- Move 2 from $N_{\cdot MothersAge\ 34}$ to $D_{\cdot MothersAge\ 34}$
  Rationale : this is the minimum possible deviation from original: the text reports 154 and 573, and the values given for the row sums of 573 and 154 match the `ROWS` provided.
  Note: not all columns will allow this (especially those at the beginning) and it is also possible to involve two columns with transfer of 1 each to achieve the same consistency. These typically yield slightly different solutions, but within generally about 10-15% difference between two distinct solutions.

Ultimate edit, which balanced the system:

- Change 4:5 split at maternal age 46 to a 3:6 split.
- Change 3:1 split at maternal age 47 to a 2:2 split.

Now, we can solve the 'Sudoku' via LINEAR PROGRAMMING (using, e.g., the `lpSolve` package in `R`)

To apply it to the Penrose table...

- Limit it to to the 325 distinct '$FathersAge \times MothersAge$' cells where $F_{r,c} > 0$.
- The row index ranges from 18 to 59 (42 rows); the column index from 17 to 47 (31 columns).

Variable set: 325 unknown $d_{r\,c}$'s (# Down's cases)

Constraints, $325 + 42 + 31 = \underline{398}$ in all, 1 for each ...

$$d_{r\,.} = D_{r.} \quad 42\ r\text{'s}$$

$$d_{.\,c} = D_{.c} \quad 31\ c\text{'s}$$

$$d_{r\,c} \leq F_{r\,c} \quad 325\ d_{r\,c}\text{'s.}$$

Objective function: $\sum_1^{325} 1 \times d_{r\,d}$, [i.e., $d_{..}$]

Since the sum constraints are effectively already fulfilled, both max and min procedures yield the same solution.

The `R` implementation and the raw data in Penrose's Table 1 can be accessed from the link found at the bottom of the 'Statistical Sudoku' page on the section of JH's website https://jhanley.biostat.mcgill.ca/ devoted to Historical Material. [The link will be made 'public' when this article is published.]

B What summary statistics are needed to fit a logistic regression?

Take the Penrose example with a binary outcome (affected/not), two predictors, x (age of mother) and z (age of father), two separate (parametric) functions, q_1 and q_2, of x and z, respectively, with no product terms involving both.

$$\begin{aligned} P[\text{affected}] &= \frac{e^{q_1(x)+q_2(z)}}{1+e^{q_1(x)+q_2(z)}}, \\ P[\text{not affected}] &= \frac{1}{1+e^{q_1(x)+q_2(z)}}. \end{aligned}$$

Consider a cell (which we can refer by its two subscripts i and j), where the parents' ages are x_i and z_j respectively and in which there are a total of n_{ij} children. Suppose y_{ij} of these are, and the remainder $n_{ij} - y_{ij}$ are not, affected. The binomial-based log likelihood contribution from this cell is therefore

$$l_{ij} = y_{ij}[q_1(x_i) + q_2(z_j)] - n_{ij}\log(1 + e^{q_1(x)+q_2(z)}).$$

Summed over all non-empty cells, the overall log-likelihood is

$$\begin{aligned} \sum_i \sum_j l_{ij} &= \sum_i \sum_j y_{ij}[q_1(x_i) + q_2(z_j)] - \\ &\quad \sum_i \sum_j n_{ij}\log(1 + e^{q_1(x_i)+q_2(z_j)}) \\ &= \sum_i \sum_j y_{ij}\ q_1(x_i) + \sum y_{ij}\ q_2(z_j) - \\ &\quad \sum_i \sum_j n_{ij}\log(1 + e^{q_1(x_i)+q_2(z_j)}) \\ &= \sum_i Y_{i\cdot}\ q_1(x_i) + \sum_j Y_{\cdot j}\ q_2(z_j) - \\ &\quad \sum_i \sum_j n_{ij}\log(1 + e^{q_1(x_i)+q_2(z_j)}) \end{aligned}$$

where $Y_{i\cdot}$ is the marginal sum $\sum_j y_{ij}$ and $Y_{\cdot j}$ is the marginal sum $\sum_i y_{ij}$. Therefore the log likelihood remains invariant over all possible within-cell distributions as long as the marginal sums remain the same.

Thus, as long as we stay with a 'no-interaction' model, the maximum likelihood estimates for the usual logistic regression or even spline logistic regression will be unique and independent of the actual outcome distribution within each cell.

A solution eluded us and our students a decade ago. We now realize that the log likelihood as written above suggests an easy way to fit a logistic regression model without having to generate a full 3D solution: use a routine such as `optim` in `R` to directly maximize it.

Without generating a solution, we have not been able to fit a logistic regression within a GLM-type framework, but wonder if some type of EM approach might work. We welcome all suggestions.

As we said in the text, in the special/simplest case where q_1 and q_2 are just linear, and simply additive, the 3 sufficient statistics for the parameters $\{\beta_0, \beta_M, \beta_F\}$ consist of just 3 numbers: the

number of cases of Down's syndrome: 154; the sum of the ages of their 154 mothers: 5736 years; and their 154 fathers: 6065 years. Equating the partial derivatives of the log-likelihood to zero results in 3 balancing equations that equate these 3 sufficient statistics to their 3 expected/fitted values. Finding this balance requires an iterative search (the same type of search that Penrose and Fisher used in their subsequent study, [6]) but again, within any parental age "cell" one does not need to specifically know how many of the children in the cell were affected or not.

C 'Model-free' statistics from 1000 possible solution-sets

Our approach is adapted from the broad principles that Fisher and Penrose used, a year later, to show that once one has accounted for the mother's age, *birth order* does not matter. The "convincing test for the theory [(hypothesis) that the once one has accounted for the mother's age, the father's age does not matter], is a direct comparison between what has been observed, and what must be expected on that theory. The appropriate theory [hypothesis] here is, that the probability of a Down's syndrome child depends on the mother's age, in some manner unknown prior to the data, but not, given the mother's age, on the [*father's age*]." We excluded children whose mothers were ages 17, 18, 25 and 30 since these maternal age-bins contain "wholly Down's syndrome, or wholly normal [children] and [thus] give no information" and limited ourselves to the 27 informative maternal age bins.

However, unlike Fisher and Penrose's 1934 paper on birth order, where they *modeled* the effect of maternal age, we accounted for maternal age by *matching* on it: Thus, within each of the 27 informative maternal age-bins, we calculated a statistic that compares the paternal ages of D and N children in that bin; we then aggregated the 27 maternal-age-specific statistics into one overall statistic, using a weighted average. We used inverse-variance weights. On the principle that in a specific maternal-age bin, a statistical comparison of n_D Downs children with n_N Normal children should have a variance proportional to $1/n_D + 1/n_N$, the weight for the sub-statistic calculated from that bin had the form $n_D \times n_N/(n_D + n_N)$. Thus, for example, the weights for maternal age bins 19 and 20 were proportional to $2 \times 5/7 = 1\frac{3}{7}$ and $1 \times 8/9 = \frac{8}{9}$ respectively, whereas the most heavily weighted age-bin was age 38, with a weight proportional to $11 \times 22/33 = 7\frac{1}{3}$.

We computed three different summary statistics, along with their null expectations and variances. The simplest and coarsest was a count: in how many of the 27 maternal-age bins did the mean age of the fathers of D children exceed the mean age of the fathers of N children. It could be thought of as a form of sign test that weighs each age-bin by the potential amount of information it provides. The second, less coarse, summary was a weighted average of the sums of the ranks of the fathers' ages of the D children in each of the 27 bins. However, because the numbers of children in the 27 bins varied considerably, this summary test statistic does not give an intuitive measure of how far apart the D and N ages are. Thus, we also calculated the (weighted average of the 27) 'placement values' [4], each of which represents the placement of the n_D fathers' ages among the n_D fathers' ages, as a way to measure how separate/overlapping the two sets of ages were. For example, consider a bin/column where the father's ages of the 2 D and 5 N children were (in increasing order) N$\underline{\text{D}}$N$\underline{\text{D}}$NN. The rank sum statistic for the 2 D children is $2 + 4 = 6$, whereas the more helpful placement (concordance, c, average of 10 pairwise u's) statistic is $3/(2 \times 5) = (1/5 + 2/5)/2 = 0.3$ or 30%. For each bin, the permutation distribution was used to compute the null expectation and variance of each statistic; where the number of possible permutations was too large, a random sample of 10,000 of them was used.

The number of unique ways in which the 154 D cases could be distributed over the 325 parental-age cells, while respecting the known marginal totals shown in Figure 4, is very large: indeed, we have not been able to determine that number. In order to assess how much the summary statistics would vary over this very large number of solution sets, we generated 1000 unique sets by using 1000 random orderings of the 398 constraints in the constraints matrix.

The main part of Figure 4 shows some details of the 1000 sampled solution sets. For example, in some 83 of the 325 parental-age cells, mostly the more sparse ones at the extreme ages, the number of D cases did not vary from solution to solution. Naturally, the cells that allowed the greatest variation were the less sparse and more central ones.

The ranges of the summary statistics are shown in the insert at the bottom left of Figure 4. In

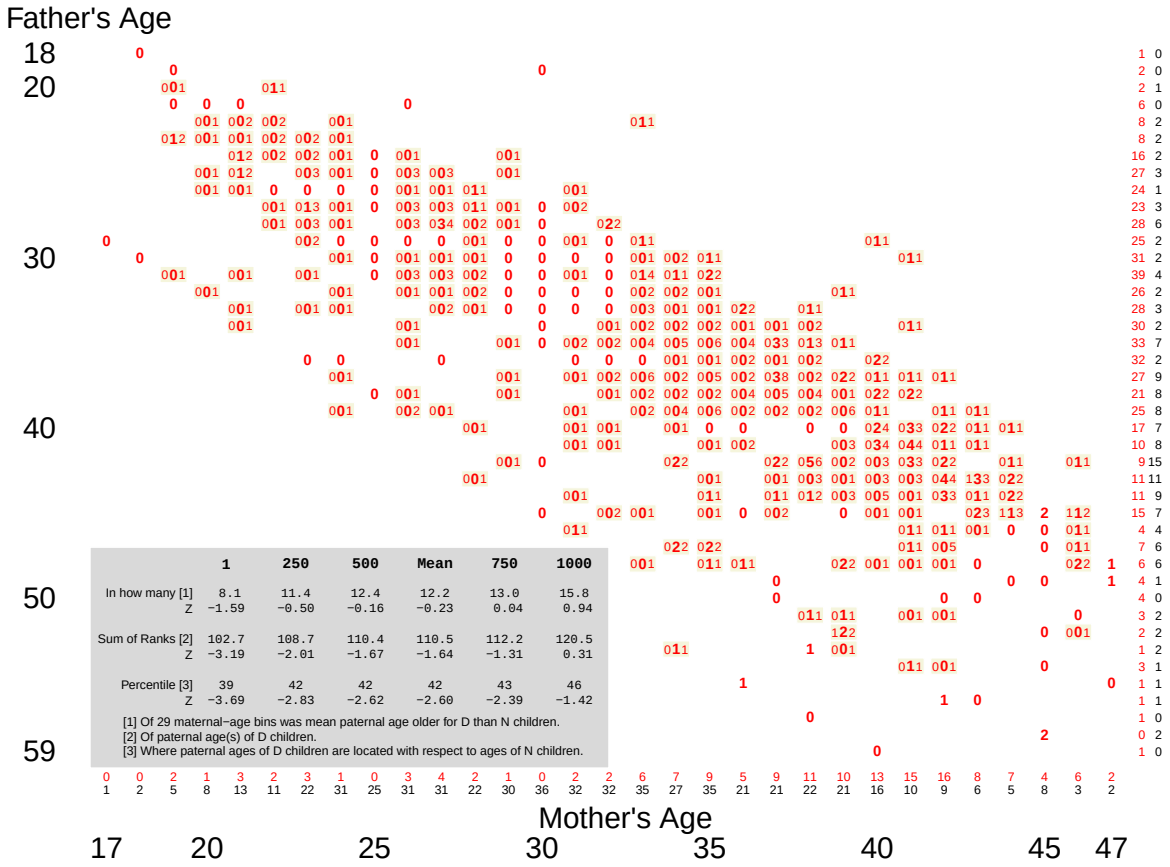


Figure 1: Cell-specific variations [across 1000 solution sets] in the numbers of Downs syndrome cases. In cells where the was no variation, only the number of cases is indicated; in cells where there was variation, the smallest, median and largest of the 1000 frequencies are shown. The insert in grey summarizes the distribution of the three summary statistics described in the text.

none of the 1000 solution sets did our variant of the 'sign test' show a remarkable difference from the null expectation. Many of the solution sets indicated that the summary rank sum of the ages of the fathers of the D children was lower than expected, whereas most of the summary statistics based on the c statistic did.

In contrast, for any selected model-based approach, the 1000 solution sets all produced the same value of the summary statistic. For example, in a form of weighted paired t-test, across the 27 informative bins, the mean ages of the fathers of the D children were 0.23 years *younger* than those of the N children (the simulated standard error was 0.46 years). Likewise, using all 727 observations in in all 31 maternal age bins, with the indicator (0,1) variable `D` indicating a child with Downs syndrome, and with Fathers' and Mothers' ages shortened to F and M, the 1000 calls to the `R` linear regression procedure

$$\texttt{lm}(F \sim \ -1 \ + \ \texttt{as.factor}(M) + D)$$

all yielded a value of -0.23 years (SE 0.44) for the coefficient of `D`. (The point estimate remained identical, but the SE was slightly larger if this model was fitted to just the 663 observations in the 27 bins.)

These mostly negative form-free statistics bear out the Penrose statistics shown in Tables 1 and 2. We leave it to readers to explore more sophisticated logistic regression forms than those in Figure 3. It must be understood however, that the outcome-based nature of the sample precludes fitting absolute risk (probability) functions.

D SR: 'Model-based' statistics from generated solution-sets

D.1 General remarks

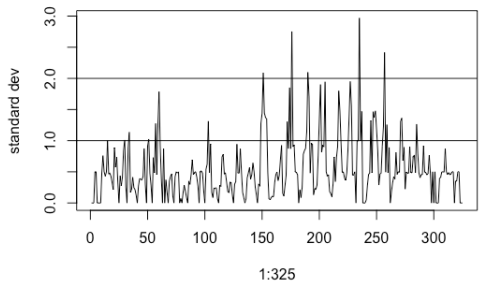
- The solution set shows certain common features shared between the different vectors of estimated Downs cases.
- The **differences** between feasible solutions are **not negligible**.
- Examples: pairwise L2 distances† between 11 solution vectors.

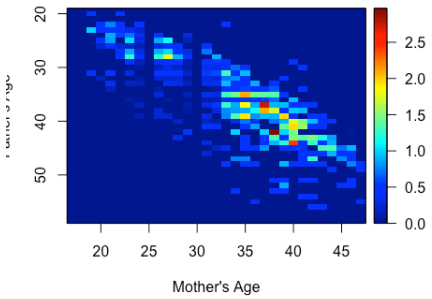
	1	2	...							...	11
1		14.8	21.5	16.1	19.5	17.2	20	17.2	20.4	15.1	20.2
2	14.8		25.5	14.1	23.6	14.1	23.4	16.4	22.6	14.9	21.9
3	21.5	25.5		24.9	12.0	23.5	13.3	22.4	14.7	22.0	14.7
4	16.1	14.1	24.9		24.0	12.6	22.3	15.4	21.0	14.8	21.5
5	19.5	23.6	12.0	24.0		24.2	12.4	22.4	14.5	21.4	13.8
6	17.2	14.1	23.5	12.6	24.2		24.3	11.7	23.2	12.9	22.3
7	20.0	23.4	13.3	22.3	12.4	24.3		23.9	11.9	22.3	14.6
8	17.2	16.4	22.4	15.4	22.4	11.7	23.9		23.9	12.9	22.6
9	20.4	22.6	14.7	21.0	14.5	23.2	11.9	23.9		23.6	11.6
10	15.1	14.9	22.0	14.8	21.4	12.9	22.3	12.9	23.6		23.2
11	20.2	21.9	14.7	21.5	13.8	22.3	14.6	22.6	11.6	23.2	

† Distance between vectors $\underline{d}^*$ and $\underline{d}^{**} = \sqrt{\sum_{i=1}^{325}(d_i^* - d_i^{**})^2}$

Remarkably, in [a sample of] 397 solutions, there were 40 locations in the solution vector of length 325 with standard deviation 0.

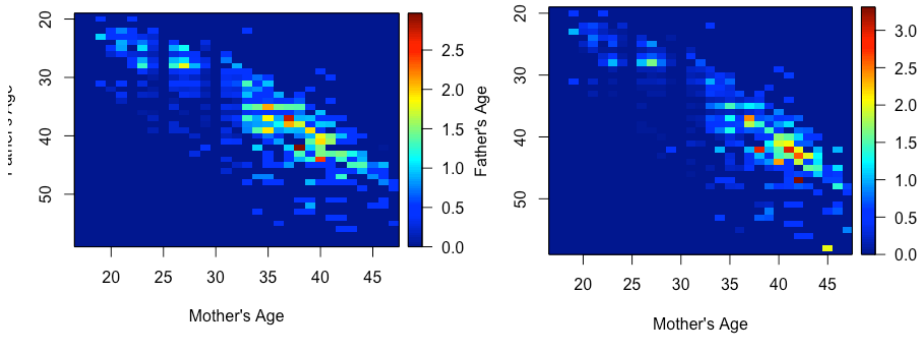
D.2 Standard deviation in solutions



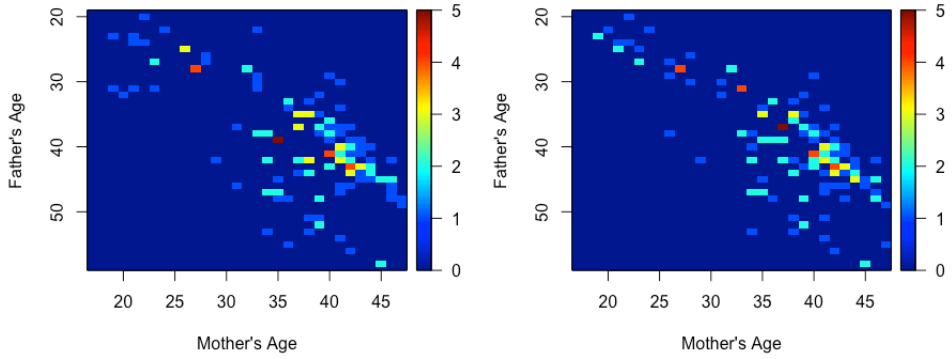


sd	# locations
0	40
< 1	282
$= 1$	1
< 1.3	296
< 2	320
$= 2$	0

D.3 Standard deviation and average solution by age

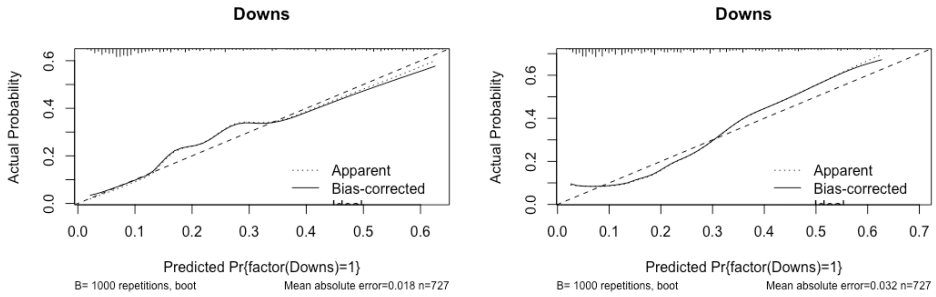


D.4 Two distinct solutions

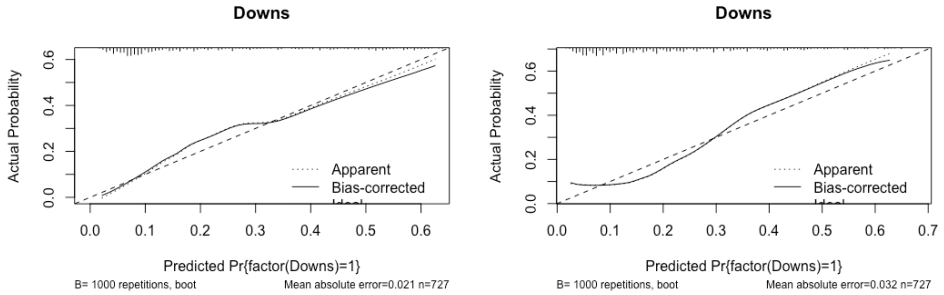


D.5 LEFT: 'spline(age)' and RIGHT: 'linear(age)' logistic calibration for 1st solution

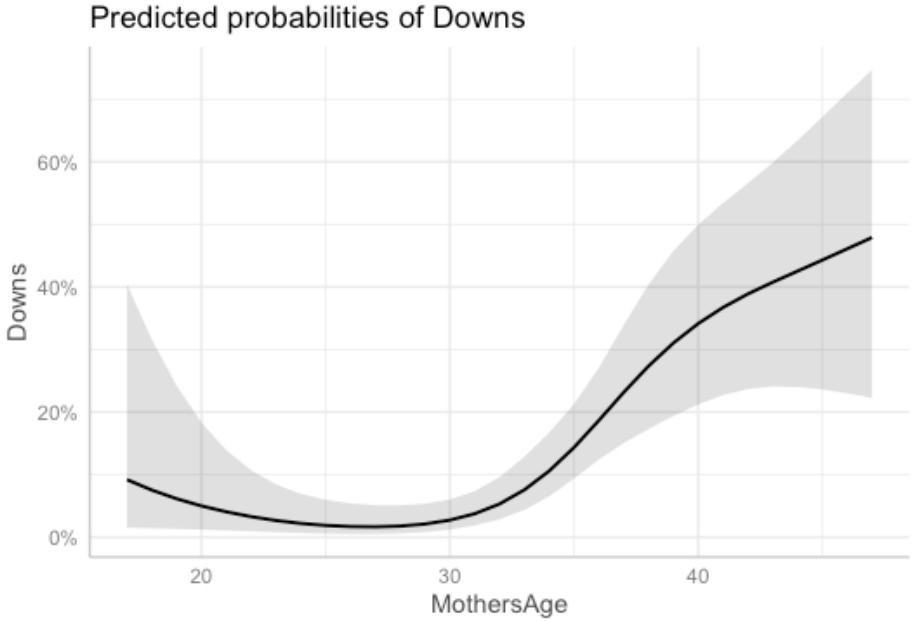
Here is a check with generalised additive model



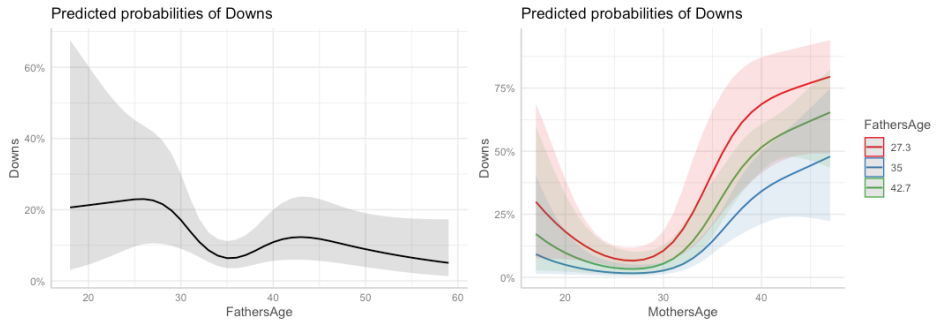
D.6 'spline(age)' and 'linear(age)' logistic calibration for second solution



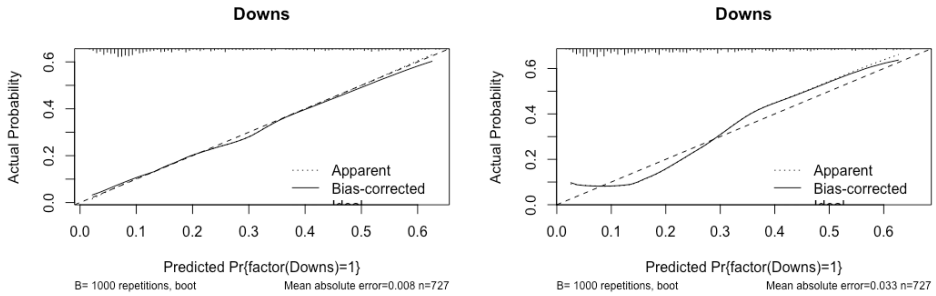
D.7 spline fits for solutions by Mother's age



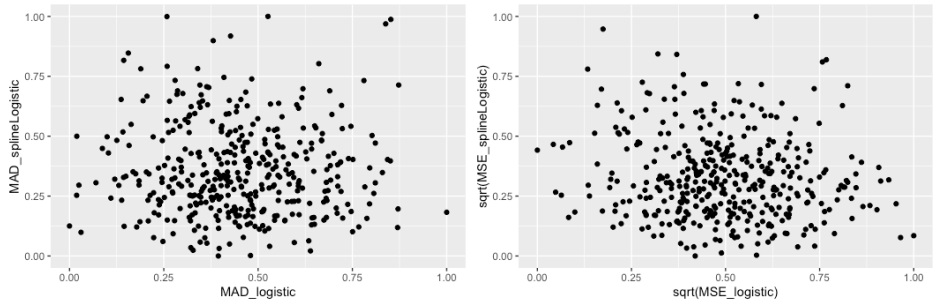
D.8 spline fits for solutions by Fathers age and by levels



D.9 spline and logistic calibration for the "best" or 24th solution



D.10 Comparison of errors for each solution



JH & SR 2025.03.31

References

[1] Cornfield, J. (1962) Joint dependence of risk of coronary heart disease on serum cholesterol and systolic blood pressure: a discriminant function analysis. *Federation Proceedings* **21(4)Pt 2** 58–61.

[2] 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* **167A**:756–767.

[3] Fisher, R. A. (1922). On the mathematical foundations of theoretical statistics. *Philosophical Transactions of the Royal Society A.* **222** 309–368.

[4] Hanley J. A. and Hajian-Tilaki K. O. (1997). Sampling Variability of Nonparametric Estimates of the Areas under Receiver Operating Characteristic Curves: An Update. *Academic Radiology* **4** 49–58.

[5] Hanley J. A. (2024). Statistical Sudoku. `https://jhanley.biostat.mcgill.ca/ StatisticalSudoku/`.

[6] Hanley J. A. (2024). Studies in the history of probability and statistics. LI: the first conditional logistic regression *Biometrika* https://doi.org/10.1093/biomet/asae038

[7] Harris H. (1973). Lionel Sharples Penrose. *Biographical memoirs of Fellows of the Royal Society* **19** 521–561.

[8] Hodgson S. (2025). Shirley Hodgson *Wikipedia* `https://en.wikipedia.org/wiki/Shirley_Hodgson`

[9] Horton N. J. (2013). I Hear, I Forget. I Do, I Understand: A Modified Moore-Method Mathematical Statistics Course. *The American Statistician* **67** 219-228.

[10] Penrose L. S. (1933). The relative effects of paternal and maternal age in Down's syndrome. *Journal of Genetics* **27** 219–224.

[11] Penrose L.S. and 18 other signatories (1961). Mongolism. [letter] *The Lancet,* **1** 775 (April 8).

[12] Penrose O. (2007). A beautiful method of analysis. *Fifty years of human genetics: a Festschrift and liber amicorum to celebrate the life and work of George Robert Fraser* edited by Oliver Mayo and Carolyn Leach. ISBN 9781862547537, Wakefield Press, Adelaide. 434-451.

[13] Rodríguez-Hernández ML, Montoya E. (2011). Fifty years of evolution of the term Down's syndrome. *Lancet* **378**: 402.

[14] Stigler S. (1973). Studies in the History of Probability and Statistics. XXXII: Laplace, Fisher and the Discovery of the Concept of Sufficiency. *Biometrika* **60**: 439–445.

[15] Wellcome Library (2024). Birth Order and Down's syndrome Correspondence. `http://wellcomelibrary.org/player/b20222087`.

[16] Wright S. (1926). Effects of age of parents on characteristics of the guinea-pig. *The American Naturalist* **60** 552–559.