doi:10.5281/zenodo.15209087 16 April 2025

# Anscombe dataset No. 5: Simpson's paradox

Carl McBride Ellis*
*Madrid, Spain*
ORCiD: 0000-0003-2966-7530

(Dated: April 16, 2025)

A new fifth dataset has been created, using the very same independent variable as three out of the four original datasets, and is in complete concordance with the summary statistics of the original Anscombe's quartet. On this occasion the new dataset serves as an illustration of the so-called Simpson's paradox.

## INTRODUCTION

It was little over 50 years ago that Francis J. Anscombe published a paper [1] in which four synthetic datasets were created to serve an illustration of the importance of visualizing ones data as an essential precursor of performing a good analysis. At the time the motivation for his work was to encourage the use of the emerging graphical capabilities of computers, but his message remains just as valid today. These four datasets, reproduced in Fig. 1 for completeness, have become to be popularly known as Anscombe's quartet:

- dataset $y1$: the points have a distribution such that a straight line could be considered to be a reasonable model to fit ($\hat{y} = 0.5x + 3$).
- dataset $y2$: the points form a smooth curve ($\hat{y} = -0.127x^2 + 2.78x - 6$ with effectively zero residuals) to which a straight line will underfit.
- dataset $y3$: has an outlier point. If we were to drop the single datapoint corresponding to $x = 13$ we obtain the fit $\hat{y} = 0.345x + 4$ with effectively zero residuals.
- dataset $y4$: high-leverage point; here a single outlier point dominates the fit. Dropping the datapoint corresponding to $x = 19$ results in a dataset having $\text{var}(x) = 0$.

On the other hand in 1951 Edward Simpson [2] published a paper concerning the creation of 2 × 2 contingency tables when there exists a confounding variable, or in machine learning parlance where there is a missing categorical feature. His paper has since become associated with what is known as Simpson's paradox [3, 4]; when said categorical feature is taken into account the interpretation of the inference can alter radically, a manifestation being for example a reversal in sign of the regression coefficient(s) $\beta_1$. It is worth mentioning that said regression reversal was not actually delineated in Simpson's original paper [5] but rather was described by Pearson, Lee and Bramley-Moore [6] in a paper from 1899, wherein providing an example of a spurious correlation derived from a heterogeneous mixture of male and female skulls that were sampled from the Catacombs of Paris; said correlation disappearing for the individual groups.

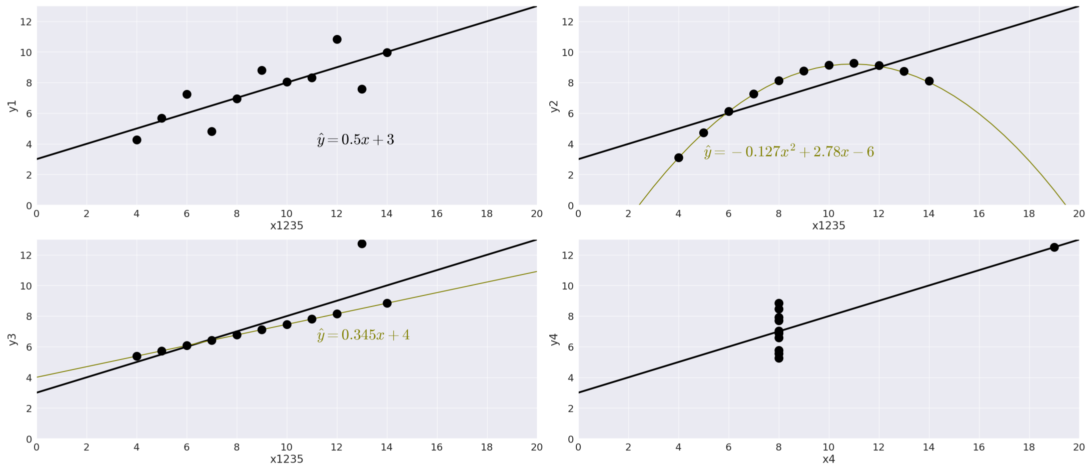

FIG. 1: The Anscombe quartet: $y1$ (upper left), $y2$ (upper right), $y3$ (lower left) and $y4$ (lower right).

TABLE I: The Anscombe quintet, organized as per the deliberate random row order of the original table. The column $y5$ (shaded) represents the contribution of this work. A `csv` file of this dataset can be downloaded from doi:10.5281/zenodo.15209087

| $x$1,2,3,5 | $y1$ | $y2$ | $y3$ | $y5$ | $x4$ | $y4$ |
|---:|---:|---:|---:|---:|---:|---:|
| 10.0 | 8.04 | 9.14 | 7.46 | 9.82 | 8.0 | 6.58 |
| 8.0 | 6.95 | 8.14 | 6.77 | 5.56 | 8.0 | 5.76 |
| 13.0 | 7.58 | 8.74 | 12.74 | 9.51 | 8.0 | 7.71 |
| 9.0 | 8.81 | 8.77 | 7.11 | 5.45 | 8.0 | 8.84 |
| 11.0 | 8.33 | 9.26 | 7.81 | 9.72 | 8.0 | 8.47 |
| 14.0 | 9.96 | 8.10 | 8.84 | 9.40 | 8.0 | 7.04 |
| 6.0 | 7.24 | 6.13 | 6.08 | 5.80 | 8.0 | 5.25 |
| 4.0 | 4.26 | 3.10 | 5.39 | 6.03 | 19.0 | 12.50 |
| 12.0 | 10.84 | 9.13 | 8.15 | 9.61 | 8.0 | 5.56 |
| 7.0 | 4.82 | 7.26 | 6.42 | 5.68 | 8.0 | 7.91 |
| 5.0 | 5.68 | 4.74 | 5.73 | 5.92 | 8.0 | 6.89 |

In this work we deliberately set out to create a new dataset that is both illustrative of Simpson's paradox, whilst at the same time being in keeping with Anscombe's original independent variable values (*i.e.* $x \in \{4, \ldots, 14\}$) and summary statistics.

## METHODOLOGY

In 2007 Chatterjee and Firat [7] detailed a procedure based on a genetic algorithm for producing datasets that have dissimilar aspects yet similar summary statistics. In 2009 Haslett and Govindaraju [8] showed how regression fits can be 'cloned' into new datasets. In 2017 Matejka and Fitzmaurice [9] used simulated annealing to gradually morph between a starting configuration towards a desired final configuration, maintaining the statistics quasi-constant (*i.e.* to within two decimal places) along the path. They also provided a visually stunning demonstration of their technique using the Datasaurus dataset [10] composed of 182 points and, by way of 200,000 iterations, led to the creation of the 'Datasaurus dozen', the modern incarnation of the Anscombe quartet. They too, as in this work, created a Simpson's paradox dataset but did not adopt the context of the original Anscombe data. In a more recent work La Haye and Zizler [11] elegantly show how linear algebra can be used to create Anscombe like datasets, along the way demonstrating how Anscombe may have actually created his original datasets.

However, in this study we adopt a distinct approach to those in the aforementioned works. To create dataset No. 5 it was decided for aesthetic reasons that the clusters would consist of straight lines. Two clusters were created, with six points ($x \in \{4, \ldots, 9\}$) being assigned to the left hand side (LHS) cluster, and the remaining five points to the right hand side (RHS) cluster. These choices led to a three parameter search space; the regression coefficient ($\beta_1$) and the the $y$ value of the centroid of the LHS line, and $\beta_1$ of the RHS line, since the location of the $y$ value of the centroid of the RHS line is dictated by the condition that the overall $\bar{y} = 7.5$. The author then proceeded to perform a brute-force computational grid search of this parameter space, finally selecting the three parameters that led to the closest fit to the Anscombe dataset statistics. The search grid focused on values of the cluster regression coefficients that were negative so as to align with the essential message of Simpson's paradox.

## RESULTS

The data points for this new dataset are provided in the $y5$ column of Table I and are shown graphically in Fig. 2. It is worth mentioning that a viable dataset was also found by assigning five points to the LHS cluster and the remaining six to the RHS cluster. However, no suitable solution was found when assigning either four or seven of the points to the LHS cluster. The summary statistics that are calculated are as follows: the mean and the standard deviation (using Bessel's correction) of both $x$ and $y$. A linear regression is fitted to the datasets where $\hat{y} = \beta_1 x + \beta_0$ and both $\beta_1$ and $\beta_0$ are presented. The coefficient of determination, $R^2$, is given by

$$R^2 = 1 - \frac{RSS}{TSS}$$

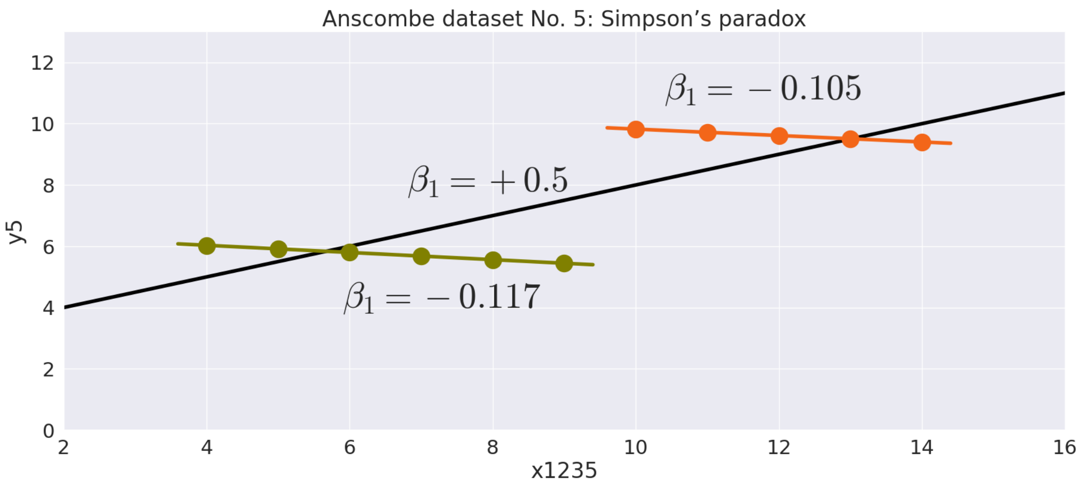

FIG. 2: Plot of the $y5$ datapoints corresponding to Simpson's paradox. We can see that the regression coefficients ($\beta_1$) for the LHS cluster (green) and the RHS cluster (orange) both have a negative slope, whereas the overall dataset regression coefficient is positive.

TABLE II: A selection of statistics of each of the five datasets, rounded to the same precision as the original publication. The standard deviations are calculated using Bessel's correction.

| statistic | $y1$ | $y2$ | $y3$ | $y5$ | $y4$ |
|---|---|---|---|---|---|
| $\bar{x}$ | 9.0 | 9.0 | 9.0 | 9.0 | 9.0 |
| $\bar{y}$ | 7.5 | 7.5 | 7.5 | 7.5 | 7.5 |
| $\sigma(x)$ | 3.317 | 3.317 | 3.317 | 3.317 | 3.317 |
| $\sigma(y)$ | 2.032 | 2.032 | 2.030 | 2.031 | 2.031 |
| $\beta_1$ | 0.5 | 0.5 | 0.5 | 0.5 | 0.5 |
| $\beta_0$ | 3.0 | 3.0 | 3.0 | 3.0 | 3.0 |
| $\sum(x-\bar{x})^2$ | 110.0 | 110.0 | 110.0 | 110.0 | 110.0 |
| RSS | 13.76 | 13.78 | 13.76 | 13.76 | 13.74 |
| $R^2$ | 0.667 | 0.666 | 0.666 | 0.666 | 0.667 |
| regression sum of squares | 27.51 | 27.50 | 27.47 | 27.48 | 27.49 |
| standard error of $\beta_1$ | 0.118 | 0.118 | 0.118 | 0.118 | 0.118 |
| Pearson $r_{xy}$ | 0.816 | 0.816 | 0.816 | 0.816 | 0.817 |

where $RSS$ is the sum of squared residuals given by $\sum \varepsilon_i^2$ where $\varepsilon_i = y_i - \hat{y}_i$, and $TSS$ is the total sum of squares given by $\sum(y-\bar{y})^2$. The regression sum of squares is given by $\sum(\hat{y}_i - \bar{y})^2$. The estimated standard error of $\beta_1$ is given by

$$SE\beta_1 = \sqrt{\frac{RSS/(n-2)}{\sum(x-\bar{x})^2}}$$

where $n$ is the number of datapoints, in this case 11. The final statistic we present is the Pearson correlation coefficient, $r_{xy}$, which measures linear correlation between $x$ and $y$ and is given by

$$r_{xy} = \frac{\sum(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum(x_i-\bar{x})^2}\sqrt{\sum(y_i-\bar{y})^2}}$$

In Table II we can see that the results for the new $y5$ dataset are within the bounds of the results of the other four datasets.

This work is dedicated to Maureen J. Ellis.

* carl.mcbride@protonmail.ch

[1] F. J. Anscombe, Graphs in statistical analysis, The American Statistician **27**, 17 (1973).

[2] E. H. Simpson, The interpretation of interaction in contingency tables, Journal of the Royal Statistical Society: Series B (Methodological) **13**, 238 (1951).

[3] C. R. Blyth, On simpson's paradox and the sure-thing principle, Journal of the American Statistical Association **67**, 364 (1972).

[4] D. V. Lindley and M. R. Novick, The role of exchangeability in inference, The Annals of Statistics **9**, 45 (1981).

[5] M. A. Hernán, D. Clayton, and N. Keiding, The simpson's paradox unraveled, International Journal of Epidemiology **40**, 780 (2011).

[6] K. Pearson, A. Lee, and L. Bramley-Moore, Mathematical contributions to the theory of evolution VI, Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character **192**, 257 (1899).

[7] S. Chatterjee and A. Firat, Generating data with identical statistics but dissimilar graphics, The American Statistician **61**, 248 (2007).

[8] S. J. Haslett and K. Govindaraju, Cloning data: Generating datasets with exactly the same multiple linear regression fit, Australian & New Zealand Journal of Statistics **51**, 499 (2009).

[9] J. Matejka and G. W. Fitzmaurice, Same stats, different graphs: Generating datasets with varied appearance and identical statistics through simulated annealing, CHI '17: Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems , 1290 (2017).

[10] C. Gillespie, S. Locke, A. Cairo, R. Davies, J. Matejka, G. Fitzmaurice, L. D'Agostino McGowan, R. Cotton, T. Book, and J. Rivers, datasaurus: Datasets from the datasaurus dozen, CRAN package (2025).

[11] R. La Haye and P. Zizler, Using linear algebra to construct anscombe's quartet, The College Mathematics Journal **56**, 124 (2025).