Wagstaff revisited: Interpreting the Concentration Index when the variable of interest is binary, with an application to the digital divide.

Michael P. Fourman*

August 11, 2017

Abstract

Positive links between outcomes and and opportunities act to create cycles of deprivation, to increase inequality, and to reduce social mobility. Measuring and monitoring such links should be a key issue for public policy. Interventions intended to increase opportunities should be designed to reduce these effects.

For example, broadband access affords social, educational, and financial advantages. Being online can make it easier to access jobs, bargains, information, and education. Increasing digital participation is widely adopted as a policy goal. However, the digital divide can magnify existing inequality if those who are less deprived are more likely to be online.

We use this example to introduce a framework for the design of interventions that will increase opportunity without increasing inequality.

Our analysis is based on classical foundations. A generalisation of Yule's (1900) measure of association provides a natural measure of the association between an ordinal outcome and a binary opportunity — a special case of a measure introduced by Agresti (1980). We show that this corresponds precisely to Wagstaff's (2005) modification of the concentration index for the case of a binary variable. This provides a direct interpretation of Wagstaff's measure.

The major novelty of our approach is to separate the global effect into local contributions. This provides a rational basis for targetted interventions. We illustrate this by analysing postcode-level Ofcom data for availability and uptake of broadband connections at various speeds across Scotland in relation to various factors of the Scottish Index of Multiple Deprivation (SIMD).

*This is an early draft of a work in progress. It is incomplete and will surely contain uncorrected errors. Please direct any comments or suggestions to the author `michael.fourman@ed.ac.uk`

1 Methodology

We examine the effect of a binary advantage, or opportunity, on existing deprivation. Our discussion in this section is general, but we will use the language of the digital divide. Our individuals are households; the *online* households, O, enjoy the advantage; the offline households, Ø, do not. To simplify this initial discussion we assume that existing inequality is represented by a total ordering, This means that for any two households, x, y, either $x \prec y$ (x *is more deprived than* y) or $y \prec x$ — one is more deprived than the other.[1]

If we view the advantage as an opportunity, then the set of offline–online pairs represents the inequalities of opportunity that may serve to counter or reinforce existing inequality. We use a graphical representation to visualise these effects.

Figure 1 shows a grid representing the offline–online pairs from a population of one hundred individuals.[2]

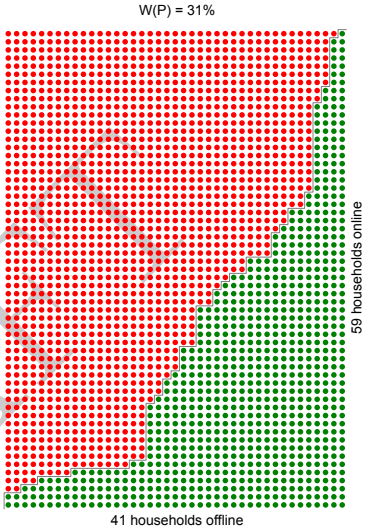


Figure 1: offline–online pairs

The dots in the diagram represent the offline–online pairs, $(d, a) \in \text{Ø} \times \text{O}$, ordered by $\prec$ in each dimension. Each column includes the pairs (d, y), for a fixed offline household d and each online household $y \in \text{O}$. Similarly, each row includes the pairs (x, a) for a fixed $a \in \text{O}$ as x ranges over the offline households, Ø.

We colour each dot to indicate whether it reinforces or counters the existing inequality. If $a \prec d$, the dot is green; the digital advantage counters the existing inequality. If $(d \prec a)$, the dot is red; the digital disadvantage compounds the existing inequality. We count the dots to determine how many times the more-deprived household has the digital advantage, A (green dots), or disadvantage, D (red dots).

$$A = |\{(d, a) \in \text{Ø} \times \text{O} \mid a \prec d\}| \qquad D = |\{(d, a)(d, a) \in \text{Ø} \times \text{O} \mid d \prec a\}| \tag{1}$$

The difference, $D - A$, normalised to a $[-1, +1]$ scale, gives a natural measure, ω, of the tendency of the digital divide to exacerbate existing inequality. It is simply related to the odds, $o_{d \prec a}$, for a randomly selected offline–online pair, that the digital disadvantage reinforces the existing inequality $(d \prec a)$.

$$\omega = \frac{D - A}{D + A} \quad o_{d \prec a} = \frac{1 + \omega}{1 - \omega} = \frac{D}{A} \tag{2}$$

[1] Adjustments to address the common case in which inequality is quantified in terms of an index of deprivation, making $\prec$ a partial linear order, will be described below.

[2] The allocation of the advantage has been drawn randomly, with the probability of being online ranging linearly from 0.4 for the most deprived individual, to 0.9 for the least deprived. In this example, 41 households remain offline, while 59 are online.

The proportion of offline–online pairs d, a for which $d \prec a$ is $(1+\omega)/2$.

In §1.2 will show that ω is precisely Wagstaff's concentration index.

1.1 Effect on Deprivation

We extend the ideas of the previous section to compare each household with the set of all those who are more fortunate.

Given a set Y, we define the sets of those less and more advantaged than an individual, x.

$$Y^{\prec x} = \{y \in Y \mid y \prec x\} \qquad\qquad Y^{\succ x} = \{y \in Y \mid y \succ x\} \tag{3}$$

We compute digital advantage and disadvantage of each household x with respect to those more fortunate, $Y^{\succ x}$. The basic equations (**??**) reduce to

$$\begin{array}{ll} \text{if } x \in \mathrm{O} & A(x, Y^{\succ x}) = \emptyset_Y^{\succ x} \quad D(x, Y^{\succ}) = 0 \\ \text{if } x \in \emptyset & D(x, Y^{\succ x}) = \mathrm{O}_Y^{\succ x} \quad A(x, Y^{\succ}) = 0 \end{array} \tag{4}$$

Each row represents an online household, a. The green dots in this row represent pairs a, d such that $a \prec d$ and $d \in \emptyset(d)$. There are $Y_{\emptyset}^{\succ a}$ such pairs. Each column represents an offline household, d. The red dots in this column correspond similarly to $Y_{\mathrm{O}}^{\succ d}$.

The *Lorenz Curve* is a line separating the red and green circles. Each dot is placed in the centre of a grid square. Since $\prec$ is total, each online household falls between two offline households, and vice-versa. For each index, $0 \le j \le 100$, the Lorenz curve plots the number of households in $P^{\prec j} = \{x_i \mid i < j\}$ that are online, $P_{\mathrm{O}}^{\prec j}$, against the number that are offline, $P_{\emptyset}^{\prec j}$.

For any set X we aggregate the effects to define $\omega_{X,Y^{\succ}}$, which measures the impact of the digital divide on X's deprivation, relative to those more fortunate.

$$\begin{array}{l} A(X, Y^{\succ}) = \displaystyle\sum_{x \in X} A(x, Y^{\succ x}) = \sum_{a \in \mathrm{O}_X} \emptyset_Y^{\succ a} \\ D(X, Y^{\succ}) = \displaystyle\sum_{x \in X} D(x, Y^{\succ x}) = \sum_{d \in \emptyset_X} \mathrm{O}_Y^{\succ d} \end{array} \qquad \omega_{X,Y^{\succ}} = \frac{D(X, Y^{\succ}) - A(X, Y^{\succ})}{D(X, Y^{\succ}) + A(X, Y^{\succ})} \tag{5}$$

In what follows, we will generally consider subsets X of some fixed population, Y. We write ω_X, for $\omega_{X,Y^{\succ}}$, and just ω for $\omega_{Y,Y^{\succ}}$. The areas above (red, disadvantage) and below (green, advantage) the Lorenz curve represent $D(Y, Y^{\succ})$ and $A(Y, Y^{\succ})$ respectively. So $\omega = \omega_{Y,Y^{\succ}}$ is the classical Gini index for our (non-classical) Lorenz curve. The index is the difference between the areas below and above the Lorenz curve, expressed as a percentage of the total area. For Figure 1 it is $\omega = 39\%$.

This index is an example of the "ordinal measure of association defined by the ratio of the proportions of concordant and discordant pairs" studied by Agresti [1]. The odds that $d \succ a$, given that a is online and d is offline, are given by ω^*.

$$\omega^* = \frac{1-\omega}{1+\omega} = \frac{A(Y, Y^{\succ})}{D(Y, Y^{\succ})} = \frac{\sum_{a \in \mathrm{O}} \emptyset^{\succ a}}{\sum_{d \in \emptyset} \mathrm{O}^{\succ d}} \tag{6}$$

Since if $a \in \mathrm{O}$ then $\text{Ø}^{\succ a} + \text{Ø}^{\prec a} = \text{Ø}$, and similarly O is split by any $d \in \text{Ø}$, we can also define $\omega_{Y,Y^\succ}$ by a sum over columns, or by a sum over rows.

$$A(Y, Y^\succ) = \sum_{a \in \mathrm{O}} \text{Ø}^{\succ a} = \sum_{d \in \text{Ø}} \mathrm{O}^{\prec d} \qquad \omega = \frac{1}{\text{Ø}} \sum_{d \in \text{Ø}} \frac{\mathrm{O}^{\succ d} - \mathrm{O}^{\prec d}}{\mathrm{O}} \tag{7}$$

$$D(Y, Y^\succ) = \sum_{d \in \text{Ø}} \mathrm{O}^{\succ d} = \sum_{a \in \mathrm{O}} \text{Ø}^{\prec a} \qquad = \frac{1}{\mathrm{O}} \sum_{a \in \mathrm{O}} \frac{\text{Ø}^{\prec a} - \text{Ø}^{\succ a}}{\text{Ø}} \tag{8}$$

If we define the advantage of an individual x relative to a set X by

$$x_Y = \frac{X^{\prec x} - X^{\succ x}}{X^{\prec x} + X^{\succ x}}, \qquad \text{then} \qquad \omega = \frac{1}{\mathrm{O}} \sum_{a \in \mathrm{O}} a_{\text{Ø}} = -\frac{1}{\text{Ø}} \sum_{d \in \text{Ø}} d_{\mathrm{O}}\,. \tag{9}$$

Thus the index ω may be interpreted as the average advantage of an online houshold relative to the set of offline households, or the average disadvantage of an offline household relative to the set of online households.

If we add an extra online household, x, to our population, we add an extra row. The new index ω^+ is given by

$$\mathrm{O}\,\omega = \sum_{a \in \mathrm{O}} a_{\text{Ø}} \qquad (\mathrm{O}+1)\,\omega^+ = \mathrm{O}\,\omega + x_{\text{Ø}} \qquad \omega^+ - \omega = \frac{x_{\text{Ø}} - \omega}{\mathrm{O}+1} \tag{10}$$

Similarly, if we remove an offline household, x, we remove a column. In this case, the new index ω^- is given by

$$\text{Ø}\,\omega = -\sum_{d \in \text{Ø}} d_{\mathrm{O}} \qquad (\text{Ø}-1)\,\omega^- = \text{Ø}\,\omega + x_{\mathrm{O}} \qquad \omega^- - \omega = \frac{x_{\mathrm{O}} + \omega}{\text{Ø}-1} \tag{11}$$

Suppose we have a population with index ω. We apply (11) followed by (10) to calculate the effect of taking an offline household, x, online.

$$\omega^- = \omega + \frac{x_{\mathrm{O}} + \omega}{\text{Ø}-1} \qquad \omega^{-+} = \omega + \frac{x_{\mathrm{O}} + \omega}{\text{Ø}-1} + \frac{x_{\text{Ø}} - \omega}{\mathrm{O}+1} - \frac{x_{\mathrm{O}} + \omega}{(\text{Ø}-1)(\mathrm{O}+1)} \tag{12}$$

Rearrange the second equation:

$$(\text{Ø}-1)(\mathrm{O}+1)(\omega^{-+} - \omega) = (x_{\mathrm{O}} + \omega)(\mathrm{O}+1) + (x_{\text{Ø}} - \omega)(\text{Ø}-1) + x_{\mathrm{O}} + \omega \tag{13}$$

$$= (\mathrm{O} - \text{Ø})\omega + \mathrm{O}x_{\mathrm{O}} + \text{Ø}x_{\text{Ø}} + 3\omega + 2x_{\mathrm{O}} - x_{\text{Ø}} \tag{14}$$

Divide the rhs by the number of individuals, N to conclude that

$$\omega^{-+} > \omega \;\text{ iff }\; (p-q)\omega + x_\prec + \frac{3\omega + 2x_{\mathrm{O}} - x_{\text{Ø}}}{N} > 0 \tag{15}$$

Thus, the more households there are online, and the greater the digital divide, ω, the more likely it is that getting more households online will increase inequality. For large N we ignore the final term and apply the following criterion to determine the households, x, which must go online if we are to reduce ω.

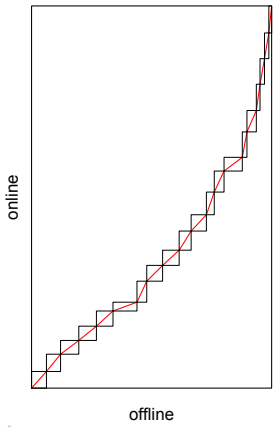
$$\omega^{-+} < \omega \;\text{ iff }\; x_\prec < (q-p)\omega \tag{16}$$

1.2 Wagstaff

We extend our treatment to a deprivation ordering $\prec$ defined in terms of an index, $I : Y \to \mathbb{R}$, and show that ω is precisely Wagstaff's [2] concentration index for a binary variable. We define:

$$\begin{aligned} x \prec y &\text{ iff } I(x) < I(y) \\ x \preceq y &\text{ iff } I(x) \leq I(y) \\ x \approx y &\text{ iff } I(x) = I(y) \end{aligned} \tag{17}$$

The relation $\prec$ is no longer a total order, but it is linear in the sense that, for any z, if $x \prec y$ then $x \prec z$ or $z \prec y$.



Again, for each index, we plot the cumulative number of households online against the cumulative number offline. This figure shows data for a population of 20,000 individuals, with 20 levels of deprivation. To generate this example, the uptake for each level of deprivation has been sampled from a beta-binomial whose mean varies linearly with deprivation in the same way as in our previous example.

In this case, for each value of the index, the offline–online pairs of households with the same index lie in a rectangle not covered by the rules given above (4). We draw the Lorenz curve as a diagonal line across each of the rectangles. This corresponds to a linear interpolation of the Lorenz curve, between points given by the cumulative offline/online data for each level of deprivation.

Each advantage within the community of households sharing a common level of deprivation, is balanced by a corresponding disadvantage. The equations (18) adjust the definitions (4) of A and D to share the advantages and disadvantages equally, over the online and offline households (respectively) at this level of deprivation.

$$\begin{aligned} A(a) &= \#\{y \in \emptyset_Y \mid a \prec y\} + \frac{1}{2}\#\{y \in \emptyset_Y \mid y \approx a\}, \quad \text{for each } a \in \mathrm{O}_Y\,, \\ D(d) &= \#\{y \in \mathrm{O}_Y \mid d \prec y\} + \frac{1}{2}\#\{y \in \mathrm{O}_Y \mid y \approx d\}, \quad \text{for each } d \in \emptyset_Y\,. \end{aligned} \tag{18}$$

If we scale both online and offline populations to $[0, 1]$, then ω is given as the Gini index –the difference between the areas below and above the curve– for a Lorenz curve plotting cumulative proportion of the online population against cumulative proportion of the offline population.

We now show that ω is exactly Wagstaff's index. Consider a population U with V households online Wagstaff plots the cumulative online population, v, against cumulative population, u, ordered by index of deprivation, to give a concentration curve.

The concentration index is the difference between the areas below and above the curve, expressed as a proportion of the total area, $U \times V$. Wagstaff corrects the concentration index for this curve, dividing it by $q = \frac{U-V}{V}$, the proportion

of the population offline. This is equivalent to expressing the Gini difference as a proportion of the area of the parallelogram shown in Figure **??**.

For each point (u, v), on Wagstaff's Lorenz curve, the corresponding offline population is $u - v$. so, transforming the parallelogram linearly to the unit square

$$(u, v) \mapsto \left(\frac{u - v}{U - V}, \frac{v}{V} \right)$$

transforms Wagstaff's Lorenz curve to ours, which shows that our index does indeed correspond to his.

1.3 Interpretation

$$\{d, a \mid d \prec a\} \qquad pq\frac{1+\omega}{2} \tag{19}$$

$$\{d, a \mid a \prec d\} \qquad pq\frac{1-\omega}{2} \tag{20}$$

$$\{d, d' \mid d \prec d'\} \tag{21}$$

$$\{a, a' \mid a \prec a'\} \tag{22}$$

We now interpret ω_X as a generalised odds ratio statistic We analyse inequalities in the distribution of fixed broadband connections across Scotland, and their effects on existing inequalities, as measured by the Scottish Index of Multiple Deprivation (SIMD).

References

[1] Alan Agresti. Generalized odds ratios for ordinal data. *Biometrics*, 36(1):59–67, 1980.

[2] Adam Wagstaff. The bounds of the concentration index when the variable of interest is binary, with an application to immunization inequality. *Health Economics*, pages 429–432, 2005.

[3] G. Udny Yule. On the association of attributes in statistics: With illustrations from the material of the childhood society, &c. *Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, 194(252-261):257–319, 1900.