Dispersion as a Regression-Safe Segregation Measure

Online appendix, version 0.6.1, 2023-07-10

Gordon Arsenoff

Abstract

This is the online appendix for the corresponding poster presented at PolMeth 2023. This appendix is under construction. Version 0.6 adds description of the data and the two measures contrasted. Covered so far are the numerical methods used and the key findings. Theory and literature review are pending.

Executive summary

The question

Is there a measure of (e.g. residential) *segregation* ready for use on the left side of an out-of-the-box regression model?

- That is, how do its sampling bias and variance relate to *marginal* properties of the data it is build from?
- Does its variance relate to its mean as prescribed by the model, up to weighting by a function of the margins?
- Does its bias relate to the margins in ways that can be controlled for with appropriate regressors?

Such a *regression-safe* measure would admit unbiased, efficient effect estimates without a custom model or software.

The findings

The log of the data's *dispersion* is a segregation measure safe for use in the linear model using weighted least squares.

- The *dissimilarity index D* (Duncan and Duncan 1955) isn't suitable: its variance violates GLM assumptions.
- The *dispersion* ϕ (Wedderburn 1974) measures what D does, on the same data, but in a regression-safe way.
- Other indices in the literature are not treated here, but known or expected to have worse properties than D.

The *segregation-as-dispersion* idea is suggested to methodologists building measures from other types of data also.

The methods

Very simplistic numerical and visual methods are used to arrive at the findings so far reached in this project.

- Numerous simulated data sets are generated, in the format used to compute $\hat{D}$ or estimate $\hat{\phi}$.
- Data-generating parameters, including marginal aspects of the data and "true" segregation levels, are varied.
- Moments of simulated sample segregation stats are plotted against parameters and regularities are confirmed.

The informal methodology means *findings at this stage may be imprecise.* Analytical contributions would be welcome.

The literature

The immense literature on segregation measures does not yet appear to address this important question.

- Numerous proposed indices exist, but they aim at *construct validity* - capturing a theoretical idea of segregation.
- The most sophisticated, modern proposals are based on data richer than census tables as simulated here.
- Limited research exists on sampling distributions and (complicated) inference procedures for a few measures.

This project means to offer a measure practitioners can estimate *and* use with *general-purpose* tools like the GLM.

Question

Census-like data

The measures considered here are built from basic census-type macro data for a given place (e.g. a city).

- The city is assumed divided into K *fixed neighborhoods*, such as census tracts.
- For each tract k, the data report a count of residents from a minority (a_k) and a majority (b_k) group.
- In a city of total size N, a fraction $P = \sum_k a_k / N$ are from the minority group.

Such data feature the *modifiable areal unit problem* (Openshaw and Taylor 1981) of sensitivity to arbitrary boundaries.

That is, each city's data are a $K \times 2$ *contingency table*, taking the *transpose* of the form:

Group	Tract 1	$\cdots$	Tract K	**Totals**
A	a_1	$\cdots$	a_K	$\mathbf{A = NP}$
B	b_1	$\cdots$	b_K	$\mathbf{B = N - A}$
Totals	$\mathbf{M_1}$	$\cdots$	$\mathbf{M_K}$	**N**

While some alternative measures are also made from such census-like data, others take richer data sources.

- These might include maps of the physical environment to account for *space* in defining segregation.
- They may also include egocentric neighborhood concepts to mitigate the modifiable areal unit problem.
- However, use of such measures imposes on the practitioner the cost of collecting the required data.

Generalization of this work to apply to measures based on richer data is an avenue of future interest.

D, the *dissimilarity index*

The *index of dissimilarity* is defined as a sum over the neighborhoods in the city: $D = \frac{1}{2}\sum_{j=1}^{J} \left| \frac{a_j}{A} - \frac{b_j}{B} \right| \in [0,1]$

Along with the above closed form, adoption of D has been enabled by its concrete interpretation as:

- the city-wide fraction of minority-group residents…
- who would have to move between neighborhoods…
- to equalize percent minorities across all neighborhoods.

Whether this is a fact of life with any meaning for residents themselves is beyond the scope of this project.

The bounds on the dissimilarity index make *beta regression* (Smithson and Verkuilen 2006) the model of choice.

- As a limited dependent variable, D would lead to not just inefficiency but bias under the linear model.
- The beta distribution describes *ratio* variables, such as D in its interpretation as a fraction of all minority locals.
- The beta GLM is defined in terms of linear models for link functions of both the mean and the precision.

The beta model, as a GLM, prescribes a particular relationship between the mean and variance of the data.

- The model assumes the sampling variance equals the mean minus its square, all over the (modeled) precision.
- This is the same as in a quasi-likelihood binomial model, where (constant) precision is the inverse of dispersion.
- Thus, establishing regression safety will mean checking that $\mathrm{V}[\hat{D}|\cdot]$ goes as $D(1-D)$ in simulated data.

A misspecified mean-variance relationship in the GLM leads to *biased* effect estimates (Freedman 2006).

ϕ, the *dispersion*

The *dispersion* is the constant of proportion between the variance in a data set and a (given) variance function.

- The dispersion is not an *index* in the sense of having a closed-form expression for a general data set.
- The dispersion *is* a *statistic*, a sample estimate of a *parameter* in some assumed data-generating process.
- The dispersion differs in this regard from D, which does not parameterize any distribution in particular.

The estimated dispersion $\hat{\phi}$ in census-like data for a city, taken as binomial, is proposed as a segregation measure.

It is further proposed that the *linear* model is appropriate for $\log\phi$ as a left-side variable.

- ϕ should range from 0 to ∞, so $\log\phi$ may take values on the entire real line.
- This removes the issue faced by D of bias in the linear model due to bounds on the left-side variable.
- This imposes a different condition for regression safety: the variance of $\hat{\phi}$ should not depend on ϕ.

Note that even if regression safety fails, the consequence here would be inefficiency, not bias.

Methods

Basics

The methods used so far have been extremely simple:

- Draw large numbers of random $K \times 2$ contingency tables with given "true" segregation levels.
- Take *sample* segregation levels, and compute means and variances.
- Plot on appropriate scales and observe regularities.

Numerical findings reached this way await analytical derivation.

Simulated census-like data

For each combination of four marginal *parameters* (below), 2048 census-like data tables were drawn.

The following *marginal* quantities defining the tables were varied:

- N: the *grand total* (2^{14}, 2^{15}, 2^{16}, 2^{17})
- K: the *nunber of rows*[1] (2, 4, 8, 16, 32)
- P: the *population fraction in column 1* (1/2, 1/4, 1/8, 1/16, 1/32)

In the simulated data, neighborhoods are of constant size $M_k = M$ for all k.

Note that $\hat{P}$ was not held fixed in each *sample*. Methods to do this were not found.

Also varied were *population* segregation measures, of two different types:

- D, the *dissimilarity index*
- ϕ, the *dispersion*

The procedure for drawing tables varied by segregation measure.

Drawing random tables for a given ϕ

Drawing $K \times 2$ contingency tables with equal row margins M for a given ϕ is simple:

- For each k, let $\frac{a_k}{M} \overset{iid}{\sim} \text{Beta}(MP\phi, M(1-P)\phi)$.

Here, random assortment of residents *is* the mechanism of sample segregation.

- ϕ is the effective size of same-race *clumps* in which residents assort *together*.
- $\log\phi = 0$ would imply residents moving individually, $\log\phi \approx 1$ as households ($\phi \approx 2.71$), and so on.

[1] $K = 2$ output was suppressed on the poster for appearance reasons, but does not break the patterns found.

Drawing random tables for a given D

By contrast, drawing $K \times 2$ contingency tables for a given D is difficult:

- There is no known explicit distribution parameterized by a *dissimilarity* index.
- A hyper-distribution of *populations* parameterized by D, K, and P was *constructed*.
- Then sample tables were drawn from the populations for any given N.

A *population* in this context is a *scaled* Kx2 contingency table ($N = 1$).

- A population is a possible realization of a *systematic* segregation process with given D.
- *Random* assortment of city residents is a matter of taking a *sample* from that population.

This model seems to be typically assumed in literature on $\hat{D}$ as a random variable.

Hyper-distribution

Sampling $K \times 2$ scaled contingency tables with equal row margins is straightforward given *four* numbers:

- Margins K and P and dissimilarity index D
- B, the number of left-column cells below the mean cell value of the left column

From there, the procedure is as easy as:

- Find the total $\dot{A}_B$ in the left-column cells with below-mean cell counts $\dot{a}_k$ (a linear function).
- Draw B of the $\dot{a}_k$s as $\dot{A}_B$ times a B-length Dirichlet $(1, \ldots, 1)$ random variable.
- Draw the remaining $\dot{a}_k$s as $P - \dot{A}_B$ times a $(K - B)$-length Dirichlet $(1, \ldots, 1)$ random variable.
- Truncate to the feasible region, i.e., discard draws with any cell total $\dot{a}_k > 1/K$.

However, fixing B arbitrarily along with D and the margins is not on option.

- B must be integrated out to get a distribution of tables conditional *only* on D.
- Thus the probability of B given D (and the margins) is needed.
- This itself is not easy to take directly, but can be found via Bayes' Rule.

Prior A prior probability $\pi_B = \Pr(B|K,P)$ can be simulated as follows:

- Draw left-column cell fractions as P times a K-length Dirichlet$(1, \ldots, 1)$ random variable.
- Truncate to feasible draws by eliminating cell fractions larger than $1/K$.
- Count the proportion of draws with each possible B.

Likelihood A likelihood $q_D(B) = \Pr(D|B,K,P)$ can be constructed similarly:

- Fit a Beta distribution to the values of D for *each* value of B.
- Find the density of the fitted Beta distribution at the chosen value of D for each B.

Posterior Given the prior and likelihood, the distribution of B given D follows from Bayes' Rule.

Then the posterior distribution of scaled tables given D, K, and P is realizable:

- For each B, $F_B(D) = \Pr(B|D,K,P) = \pi_B F_D(B) / \sum_B (\pi_B F_D(B))$ is the posterior of B.
- For each B, draw JF_B tables using the above procedure, for a total of J populations.

Sample tables

Finally, draw a size-N *sample* from each of the J populations:

- Let $\dot{a}_k$ and $\dot{b}_k$ be the cell values in row k of each population, and $\dot{p}_k = K\dot{a}_k$.
- Then take a sample with $a_k \overset{iid}{\sim} \mathrm{B}(N\dot{p}_k, N(1-\dot{p}_k))$.

This assumes residents assort across tracts in groups of one, on top of systematic segregation D.

Findings

(log) Dispersion is regression-safe

Linear model

Bias: control by regressing on $(K-1)^{-1}$

THIS IS THE BIGGEST ERROR ON THE POSTER: the observed linearity is in $1/(K-1)$, not $1/(K-2)$.

- The bias is available for the $K=2$ simulations, but omitted from the plots for appearance reasons.
- It is clear that $\mathrm{E}[\log\hat{\phi}-\log\phi|K=2] \approx 3\times\mathrm{E}[\log\hat{\phi}-\log\phi|K=4]$.
- The $K=2$ cases thus confirm that the appropriate variable to control for is $(K-1)^{-1}$: $4-1=3\times(2-1)$.

Controlling for the inverse of $K-1$ should eliminate omitted variable bias due to *marginal* quantities.

Variance: weight by $(K-2)$

The variance of $\log\hat{\phi}$ is approximately independent of ϕ as the linear model expects.

- The sampling variance appears to go as $(K-2)^{-}1$, especially for $K>2$, and be constant in N and P.
- Thus the sampling error can be accounted for by weighting observations in proportion to $K-2$.
- This holds true in the same way at any *moderate* population value of ϕ.

Independence of mean and spread is the condition under which least-squares *may* be unbiased and efficient.

Note that weighting does not account for error due to unit-level shocks – only error due to sample size K.

- Lewis and Linzer (2005) cover performance of different responses to this two-component error situation.
- OLS with robust standard errors or feasible generalized least squares may be better than WLS.
- Of course, recommending FGLS may defeat the purpose of methodological accessibility to practitioners.

That the results will be unbiased, however, holds regardless of the least-squares method employed.

Danger zone: $\phi \ll MK$

Higher-order terms in the variance and bias dominate for ϕ not much less than the mean minority cell size.

- In this parameter region, neighborhoods are almost all minority residents or almost none.
- Plenty of these simulations would have been discarded for having some $a_k > M$.
- This is perhaps like unto *right-censoring* the a_ks in the simulated data, with attendant consequences.

Theoretical and/or analytic contributions to understanding these higher-order issues would be greatly appreciated.

This parameter region appears *realistic* for real-world data, with some important practical consequences.

- Census tracts average about 2^{12} persons; some cities or metro areas may be 100×2^{-5} percent Black.
- But it is commonly observed that an area's tracts are either almost wholly Black or almost wholly White.
- In fact, census tracts are delineated with a goal of *maximizing* their initial racial homogeneity.

Solutions are not offered here, but use of data more granular than census tracts is recommended against for now.

Effective K in neighborhoods of unequal size

In real-world data, neighborhoods do not all have equal populations M. What does K mean here?

- Recommended is the Laakso and Taagepera (1979) effective number of units, or inverse of the *concentration*.
- That is, with varying neighborhood sizes M_k, try $K_{\mathrm{eff}} = 1/\sum_k (M_k/N)^2$ in place of K.
- Old simulations and plots for which output and code is now lost seemed to show this adaptation worked well.

Numerical confirmation of the K_{eff} proposal's appropriateness will be pursued down the road.

Dissimilarity is not regression-safe

Bias: dependent on margins in low limit only

Absent unrealistic *unsegregated* cases, sampling bias of D might well just be *ignored* under the logit link.

- Absent *systematic* segregation, $\Lambda^{-1}\hat{D}$ looks linear in $\log K - 1$, $\log N - 1$, and $\log P(1-P)$.
- Margin bias has a more complicated form away from the low-D limit, of a similar shape at any level.
- However, its magnitude ranges from barely noticeable at $D = 0.2$ to negligible at $D = 4$ and up.

Note, assuming substantial segregation everywhere and ignoring margin bias would have to be a *theoretical* call.

Variance: independent of mean contra GLM

The variance of $\hat{D}$, however, is not related to the mean as required to estimate the beta GLM.

- The log of the variance is linear in $\log P(1-P)$ and $\log N - 1$, but nearly independent of D (and K as well).
- Weighting by $(D - D^2)^{-1}$ would be required to even approximately unbias beta regression effect estimates.
- Magnitude of the bias induced in the GLM by the misspecified mean-variance relationship is not covered here.

The (unique) canned model for the range of D is simply unusable for its spread as a function of its mean.

How to model D if you must

Where D is available but ϕ cannot be estimated, perhaps in historic data sets, practitioners could perhaps:

- Use OLS with robust standard errors and hope $0 \ll D \ll 1$ to avoid multiple sorts of bias.
- Use the linear model on the inverse logit of $\hat{D}$ and accept inefficiency likely to be severe.
- Build a custom model with user-defined mean-variance function – i.e., give up on accessible methods.

None allows making substantive statements from the model like "how much X would achieve *total* desegregation?".

Literature

Suggestions of *comprehensive* review articles on segregation measures are welcome. A few example papers include:

- Massey and Denton (1988) introduce separate *dimensions* of segregation to be measured separately.
- Roberto (2016) (*inter alia*) proposes a census-data measure and contrasts with the most-used ones.
- Yao et al. (2018) review measures going beyond census data to account for *spatial* features.

A larger picture of the literature will be build up here and/or above as this appendix expands.

References

Duncan, Otis Dudley, and Beverly Duncan. 1955. "A Methodological Analysis of Segregation Indices." *American Sociological Review* 20(2): 210–17.

Freedman, David A. 2006. "On the so-Called "Huber Sandwich Estimator" and "Robust Standard Errors"." *The American Statistician* 60(4): 299–302. https://www.tandfonline.com/doi/abs/10.1198/000313006X152207.

Laakso, Marku, and Rein Taagepera. 1979. "'Effective' Number of Parties: A Measure with Application to West Europe." *Comparative Political Studies* 12: 3–27.

Lewis, Jeffrey B., and Drew A. Linzer. 2005. "Estimating Regression Models in Which the Dependent Variable Is Based on Estimates." *Political Analysis* 13(4): 345–64.

Massey, Douglas S., and Nancy A. Denton. 1988. "The Dimensions of Residential Segregation." *Social Forces* 67: 281–315.

Openshaw, Stan, and P. J. Taylor. 1981. "Quantitative Geography." In eds. N. Wrigley and R. J. Bennett. Routledge; Kegan Paul.

Roberto, Elizabeth. 2016. "The Divergence Index: A Decomposable Measure of Segregation and Inequality." https://arxiv.org/abs/1508.01167.

Smithson, Michael, and Jay Verkuilen. 2006. "A Better Lemon Squeezer? Maximum-Likelihood Regression with Beta-Distributed Dependent Variables." *Psychological Methods* 11(1): 54–71.

Wedderburn, R. W. M. 1974. "Quasi-Likelihood Functions, Generalized Linear Models, and the Gauss—Newton Method." *Biometrika* 61: 439–47.

Yao, Jing, David W. S. Wong, Nick Bailey, and Jonathan Minton. 2018. "Spatial Segregation Measures: A Methodological Review." *Tijdschrift voor Economische en Sociale Geografie* 110(3): 235–50.