Decomposition Techniques for Social Epidemiology

Advanced Social Epidemiology PhD Course

Sam Harper

University of Copenhagen

2021-10-11 to 2021-10-15

3. Decomposition

3.1 Life Table Decomposition

3.2 Concentration Index Decomposition

3.3 Kitagawa-Blinder-Oaxaca Decomposition

3. Decomposition

3.1 Life Table Decomposition

3.2 Concentration Index Decomposition

3.3 Kitagawa-Blinder-Oaxaca Decomposition

Overview of Decomposition Techniques

Today:

- Life table decomposition
- Inequality decomposition: Concentration Index
- Decomposing two-group differences: Kitagawa-Blinder-Oaxaca

Not covered here:

- Effect decomposition (i.e., mediation)
- Decomposition of population rates
- Inequality decomposition: Indexes for Nominal social groups

Moving from Description to Explanation

- Ultimately, we want to know why health inequalities are changing over time—what changed?
    - Risk factors?
    - Demographic composition?
    - Social conditions?
- Unpacking the ‘components’ of health inequality is an opportunity to better integrate the monitoring of health inequalities with the etiology of health inequalities.
- These techniques often involve various kinds of ‘counterfactual’ scenarios

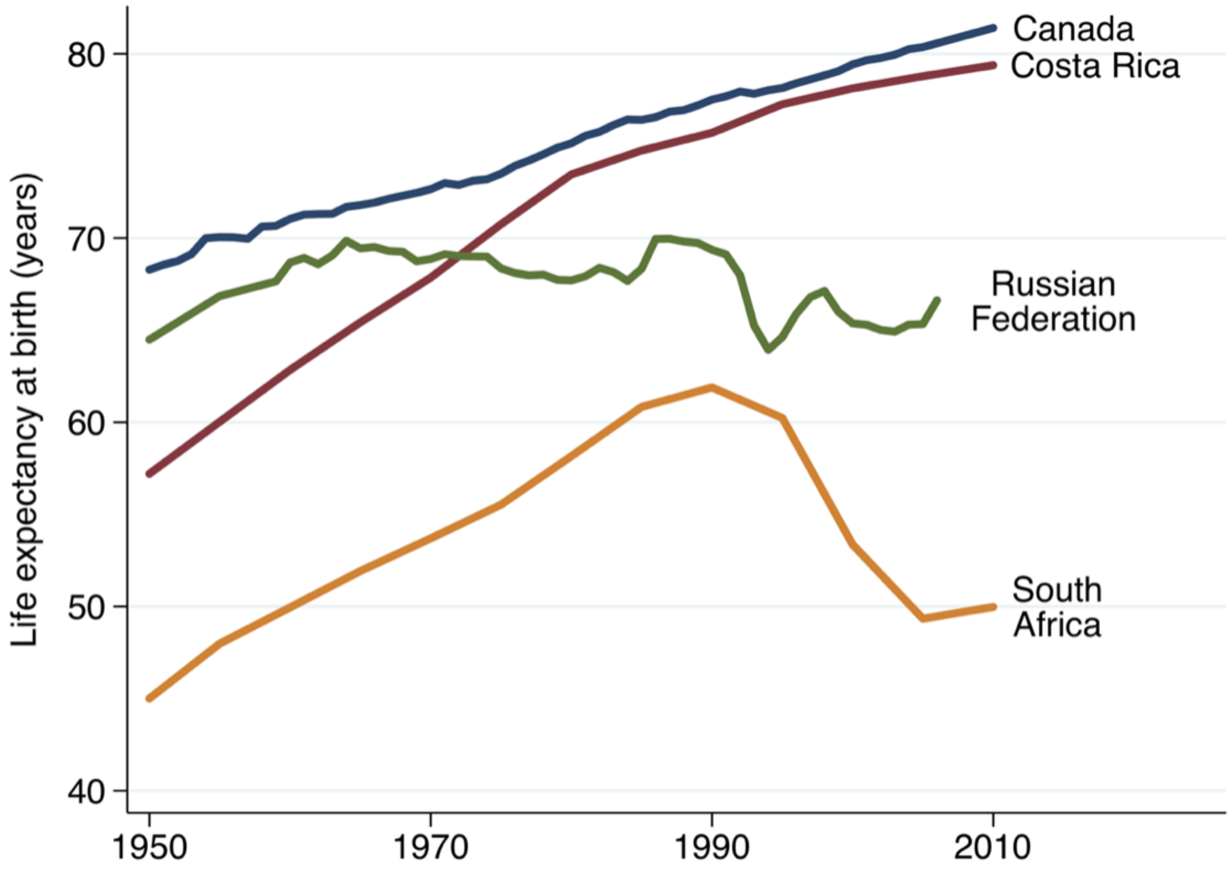
3. Decomposition

3.1 Life Table Decomposition

3.2 Concentration Index Decomposition

3.3 Kitagawa-Blinder-Oaxaca Decomposition

Why does life expectancy go up and down?



Decomposing changes in life expectancy

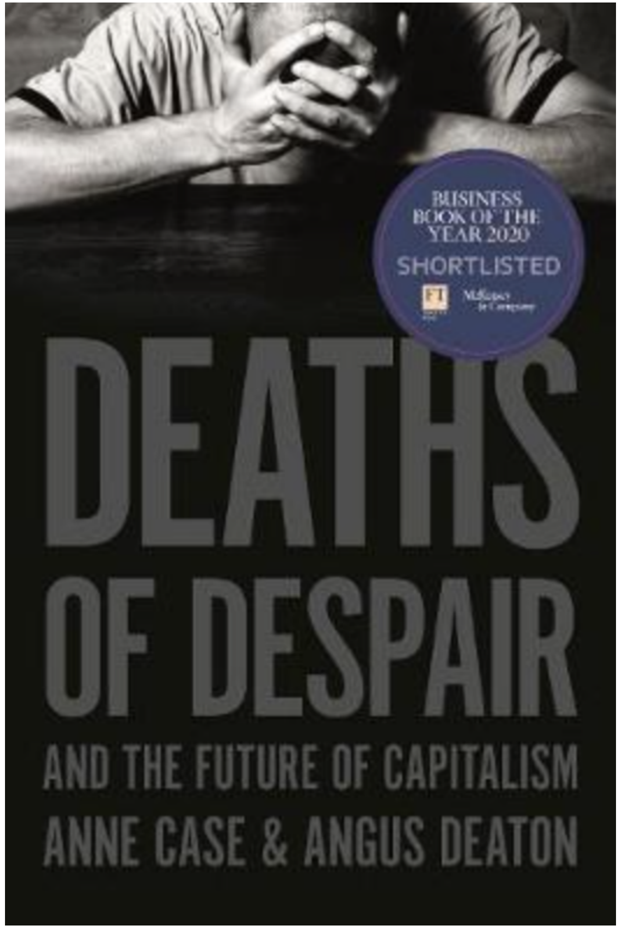
Uses age- and cause-specific mortality rate differences between two (or more) populations to estimate the contribution of specific age groups and causes of death to changes in life expectancy.

Not causal.

Can provide a means of evaluating 'explanations' for changes in mortality.

Between countries, genders, ethnic groups, social classes, etc.

Example from recent events



Over the last century, Americans' life expectancy at birth has risen from 49 to 77. Yet in recent years, that rise has faltered. Among white people age 45-54 — or a time many view as the prime of life — deaths have risen. Especially vulnerable are white men without a four-year bachelor's degree. Curiously, midlife deaths have not climbed in other rich countries, nor, for the most part, have they risen for American Hispanics or blacks.

NY Times Book Review, March 17, 2020

Specific causes are a key part of this narrative

> Although the surge in deaths in America is what we might see during the ravages of an infectious disease, like the Great Influenza Pandemic of 1918, this is an epidemic that is not carried by a virus or a bacterium, nor is it caused by an external agent, such as poisoning of the air or the fallout from a nuclear accident. Instead, people are doing this to themselves. They are drinking themselves to death, or poisoning themselves with drugs, or shooting or hanging themselves.
>
> Case and Deaton (2019, p38)

Example of using life table decomposition

ANNUAL REVIEWS

Annual Review of Public Health

Declining Life Expectancy in the United States: Missing the Trees for the Forest

Sam Harper,[1,2,3] Corinne A. Riddell,[4] and Nicholas B. King[1,2,5]

[1]Department of Epidemiology, Biostatistics and Occupational Health, McGill University, Montreal, Quebec H3A 1A2, Canada; email: sam.harper@mcgill.ca, nicholas.king@mcgill.ca
[2]Institute for Health and Social Policy, McGill University, Montreal, Quebec H3A 1A2, Canada
[3]Department of Public Health, Erasmus Medical Center, 3015 GD Rotterdam, The Netherland
[4]Division of Epidemiology and Biostatistics, School of Public Health, University of California, Berkeley, California 94720, USA; email: c.riddell@berkeley.edu
[5]Biomedical Ethics Unit, McGill University, Montreal, Quebec H3A 1X1, Canada

Annu. Rev. Public Health 2021. 42:381–403

First published as a Review in Advance on December 16, 2020

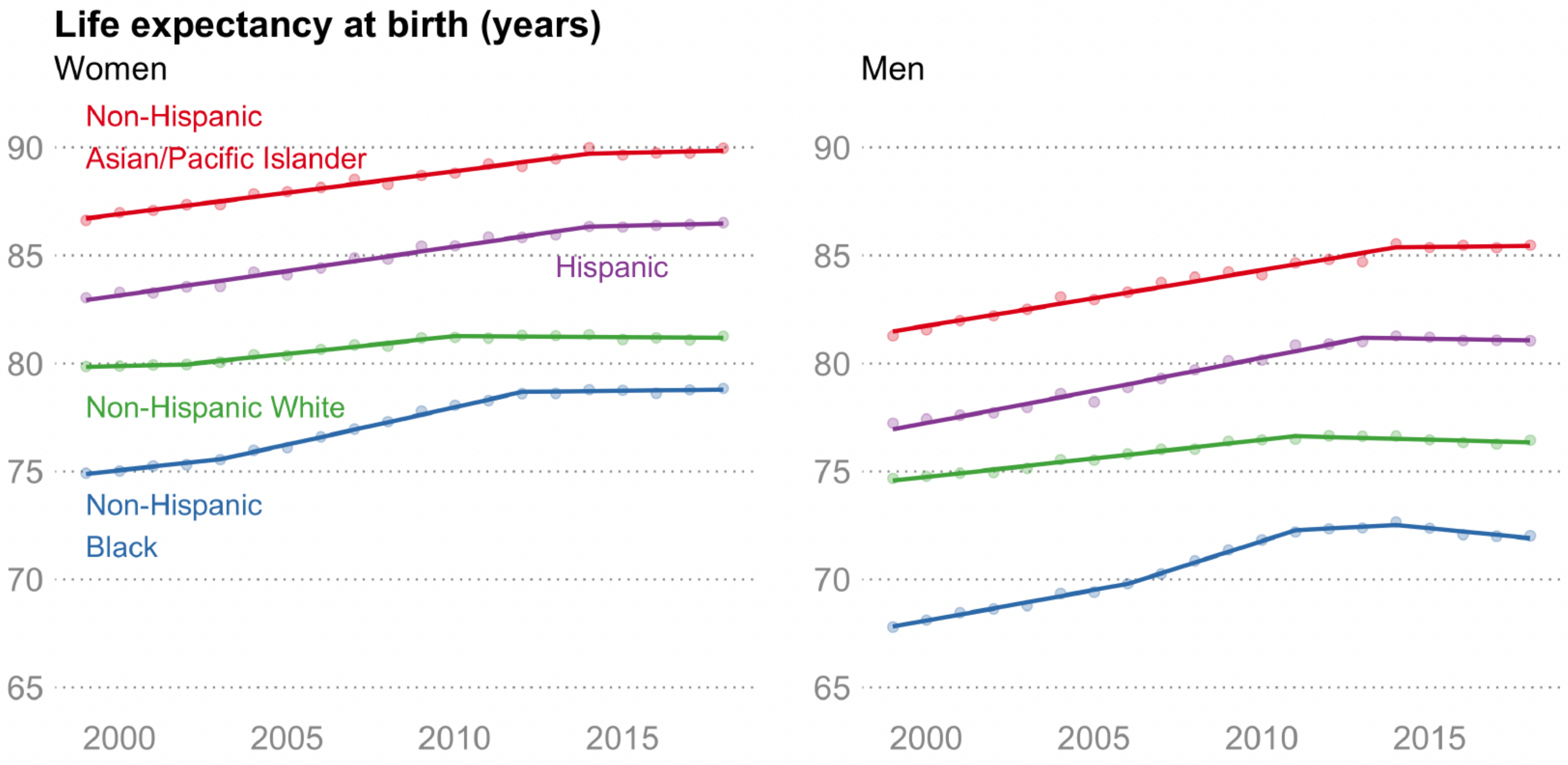
Keywords

life expectancy, opioids, cardiovascular diseases, suicide, homicide, health inequalities

Decompose the decline in life expectancy in the US between 2014 and 2017

- By age
- By cause of death
- For 8 race-ethnic groups

Trends in life expectancy



What are we explaining?

	Non-Hispanic API		Non-Hispanic Black		Non-Hispanic White		Hispanic	
Year	Women	Men	Women	Men	Women	Men	Women	Men
2014	90.0	85.5	78.8	72.7	81.3	76.6	86.3	81.3
2015	89.7	85.4	78.8	72.4	81.1	76.5	86.3	81.2
2016	89.7	85.5	78.6	72.1	81.2	76.3	86.4	81.1
2017	89.7	85.3	78.8	72.0	81.1	76.3	86.4	81.1
2018	90.0	85.5	78.8	72.0	81.3	76.4	86.5	81.0
Changes								
2014-2017	-0.3	-0.2	0.0	-0.7	-0.2	-0.3	0.1	-0.2

Declines evident for all men and for most women

Largest for black men

Remember what a life table is?

Age	Length of interval	Probability of dying between ages x to x+n	Number surviving to age x	Number dying between ages x to x+n	Person-years lived between ages x to x+n	Total number of person-years lived above age x	Life exp at age x
x	n	${}_nq_x$	${}_nl_x$	${}_nd_x$	${}_nL_x$	T_x	e_x
0	1	0.0123	100,000	1,229	98,900	7,594,342	75.94
1	4	0.0016	98,771	155	394,698	7,495,442	75.89
5	5	0.0009	98,616	88	492,842	7,100,744	72.00
10	5	0.0010	98,528	98	492,389	6,607,902	67.07
15	5	0.0019	98,430	187	491,758	6,115,513	62.13
20	5	0.0035	98,243	345	490,362	5,623,755	57.24
25	5	0.0047	97,898	460	488,415	5,133,394	52.44
35	10	0.0105	96,794	1,021	481,552	4,159,267	42.97
45	10	0.0242	94,229	2,277	465,727	3,202,492	33.99
55	10	0.0483	88,782	4,287	433,781	2,284,543	25.73
65	10	0.0976	78,537	7,662	374,209	1,442,517	18.37
75	10	0.2024	60,885	12,321	274,487	738,005	12.12
85	∞	1.0000	34,617	34,617	255,202	255,202	7.37

Decomposing between 2 groups

- E.g., between 2 time periods (2014 and 2017), the general formula is:

$$
{}_n\Delta_x = \left[\underbrace{l_x^{2017}/l_0^{2017}}_{\text{fraction of survivors}} \times \overbrace{\left(\frac{{}_nL_x^{2014}}{l_x^{2014}} - \frac{{}_nL_x^{2017}}{l_x^{2017}} \right)}^{\text{direct effect}} \right] + \overbrace{\left[\frac{T_{x+n}^{2014}}{l_{x+n}^{2014}} \times \frac{\frac{l_x^{2017} l_{x+n}^{2014}}{l_x^{2014}} - l_{x+n}^{2017}}{l_0^{2017}} \right]}^{\text{indirect effect + interaction}}
$$

- Direct effect multiplies the fraction of survivors at each age by the difference between the 2 groups in 'temporary life expectancy' at a given age.

- Indirect effect happens because differences in the direct effect means more survivors at subsequent ages.

Arriaga (1984)

Partial life tables for black men

Our aim is to *decompose* the 0.7 year decline in life expectancy at birth that happened between 2014 and 2017 by age.

Black Men, 2014

Age	lx	Tx	Lx	ex
0-1	100000	98945	7266771	72.7
1-4	98828	394953	7167826	72.5
5-14	98649	985394	6772872	68.7
...				
85+	27676	204278	204278	7.4

Black Men, 2017

Age	lx	Tx	Lx	ex
0-1	100000	98919	7201581	72.0
1-4	98799	394856	7102662	71.9
5-14	98629	985064	6707806	68.0
...				
85+	27104	205713	205713	7.6

Source: Harper et al. 2020

Plug in values to estimate, e.g., contribution of 1-4 age group

Black Men, 2014

Age	lx	Tx	Lx	ex
0-1	100000	98945	7266771	72.7
1-4	98828	394953	7167826	72.5
5-14	98649	985394	6772872	68.7
...				
85+	27676	204278	204278	7.4

Black Men, 2017

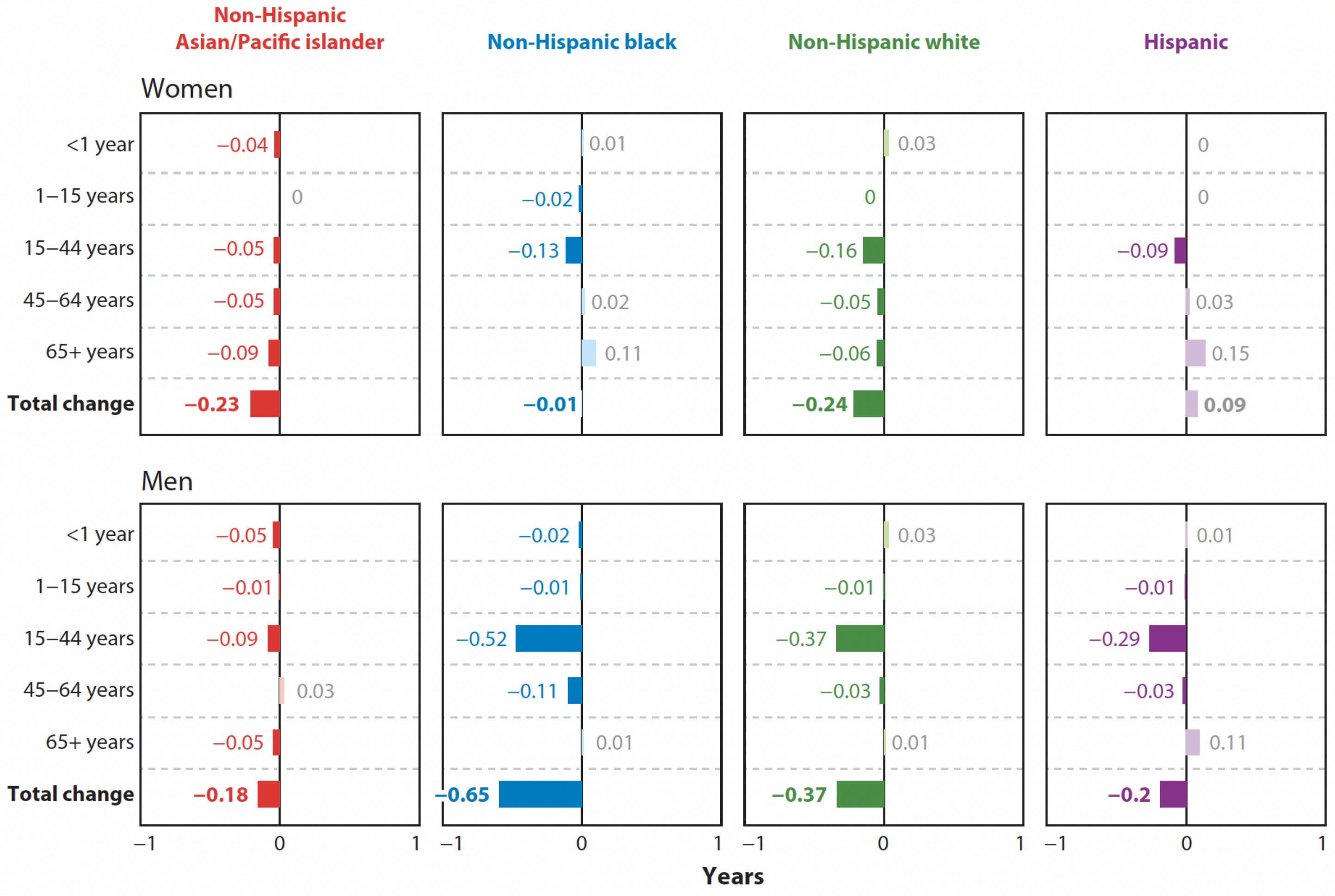
Age	lx	Tx	Lx	ex
0-1	100000	98919	7201581	72.0
1-4	98799	394856	7102662	71.9
5-14	98629	985064	6707806	68.0
...				
85+	27104	205713	205713	7.6

$$
{}_4\Delta_1 = \left[98799/100000 \times \left(\frac{7167826}{98828} - \frac{7102662}{98799}\right)\right] + \left[\frac{985394}{98649} \times \frac{\frac{98799 \times 98649}{98828} - 98629}{100000}\right]
$$

$${}_4\Delta_1 = -0.01 \text{ years}$$

Results by age

- Black men lost the most years.
- Mostly worsening mortality among the young (15-44)



Harper et al. 2020

Decomposing life expectancy differences by cause

The contribution ${}_n\Delta_x^i$ of each cause of death i within a given age group is a function of the difference between the two time periods in the proportion of deaths due to a given cause:

$$ {}_n\Delta_x^i = {}_n\Delta_x \times \frac{\overbrace{\left({}_np_x^{i,2014} \times {}_nr_x^{2014}\right) - \left({}_np_x^{i,2017} \times {}_nr_x^{2017}\right)}^{\text{difference in share of deaths for cause } i}}{\underbrace{{}_nr_x^{2014} - {}_nr_x^{2017}}_{\text{overall mortality rate difference}}} $$

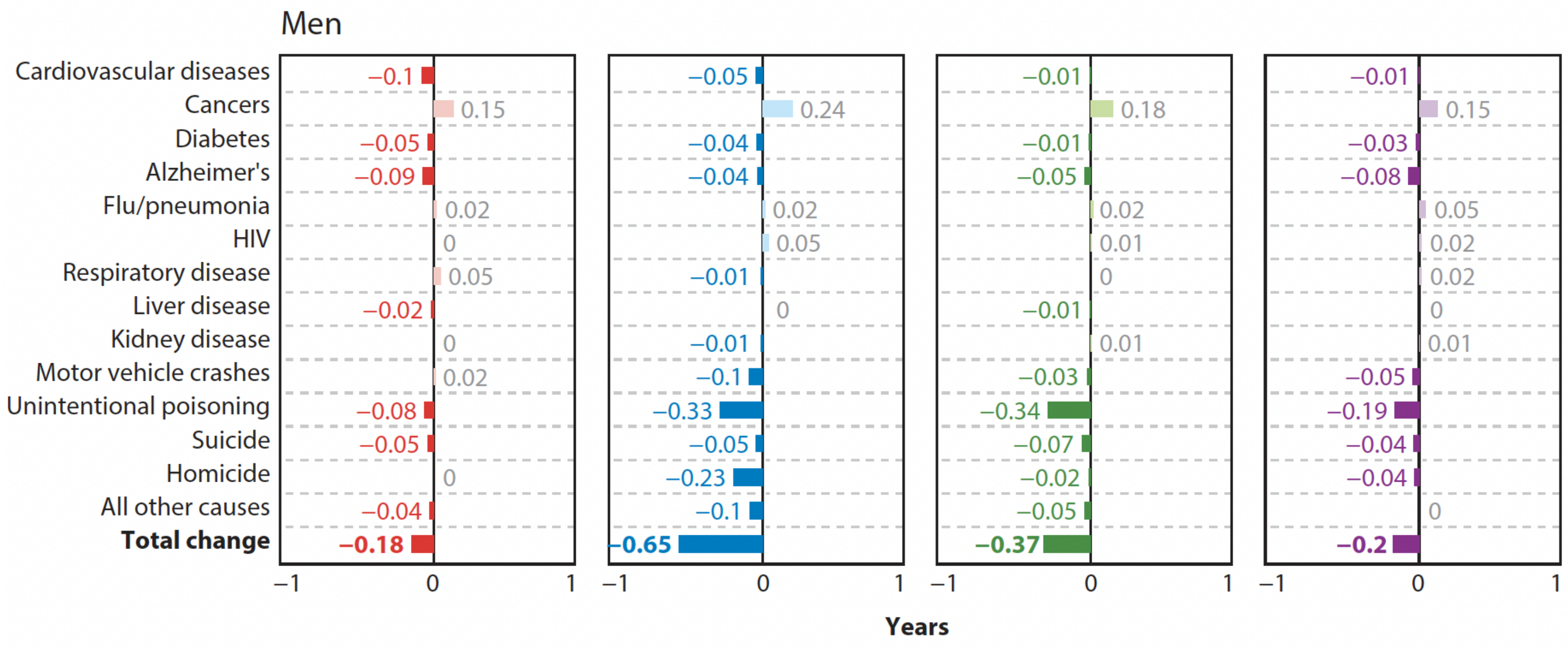
where ${}_n\Delta_x$ is the total contribution for an age group, ${}_np_x^i$ is the proportion of deaths within age group x due to cause i, and ${}_nr_x$ is the overall age-specific death rate. The total difference in life expectancy is the net sum of the age-cause components:

$$ \sum_i {}_n\Delta_x^i = {}_n\Delta_x, \text{ and } e_0^{2014} - e_0^{2017} = \sum_x {}_n\Delta_x = \sum_x \sum_i {}_n\Delta_x^i $$

Arriaga (1989)

Results by cause: Men

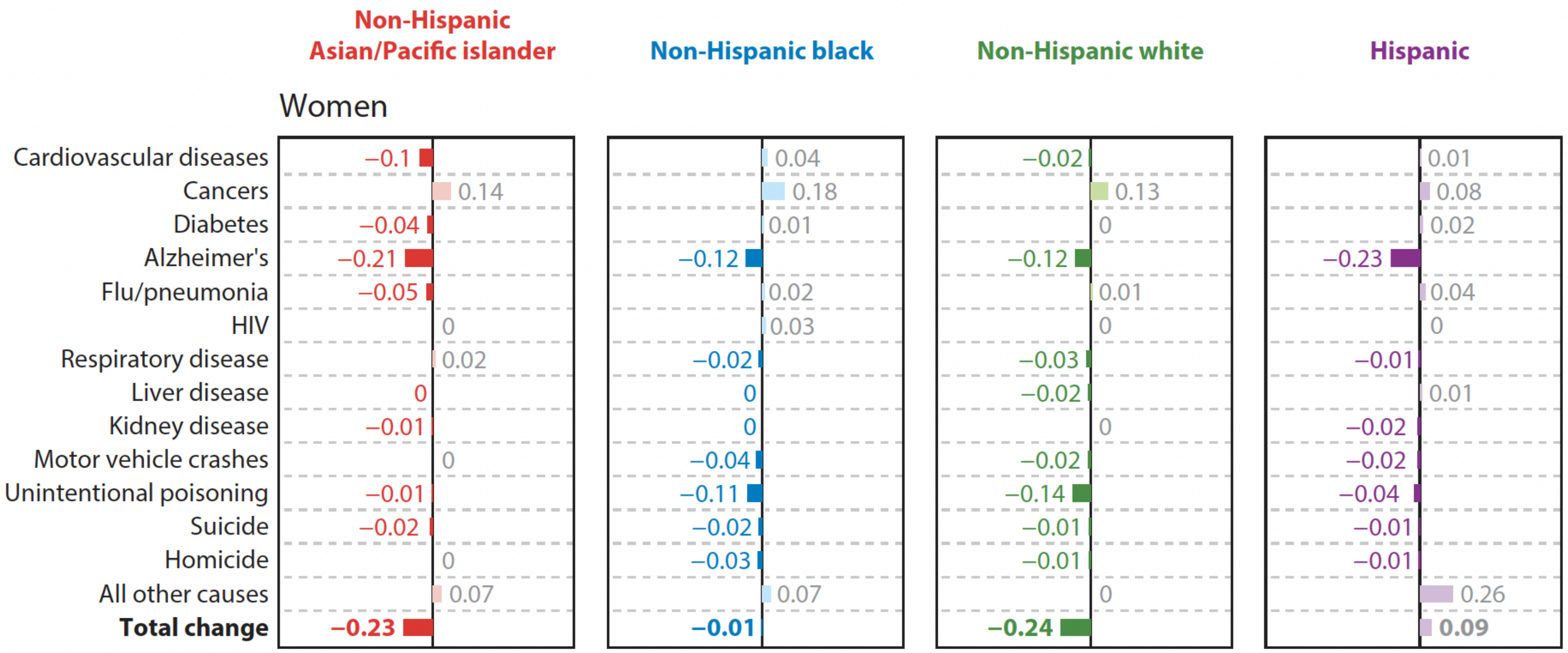
- Opioids (unintentional overdoses) played a large part.
- Homicide for black men
- Little role for suicide or alcohol.



Harper et al. 2020

Results by cause: Women

- Opioids, but also Alzheimer's.
- Variations by race-ethnicity
- Cancer mortality improved.



Harper et al. 2020

Summary

Life table decomposition useful for understanding links between proximal risks and mortality, and how they may 'explain' changing patterns of life expectancy.

Minimal assumptions, but not causal.

Example showing how the 'Deaths of Despair' narrative is hard to reconcile with diverse mortality patterns:

- Declines have affected all race-ethnic groups.
- Most of the decline due to opioid overdoses, homicide, and Alzheimer's disease.
- Deaths from suicide and alcohol-related causes have risen but explain little of America's stagnating life expectancy trends.

3. Decomposition

The 'usual' approach

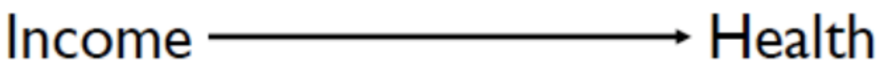
Conventional methods for “explaining” effects of social exposures

- Estimate crude or demographic-adjusted effect (logit, hazard)
- Add “conventional” risk factors (physiological, behavioural)
- Add “novel” risk factors (flavour-of-the-week)
- Interpret accordingly

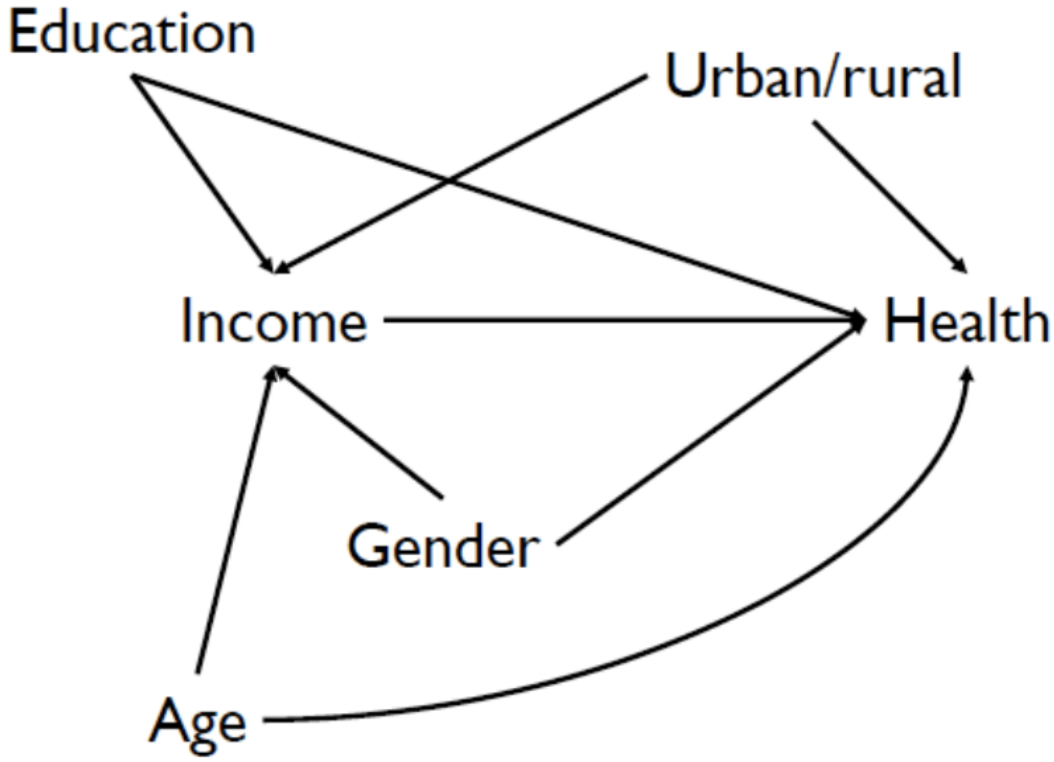
Limitations of conventional approach

- Often fail to consider entire socioeconomic distribution (typically low vs. high only) in the context of “explanation”
- Often ignore absolute risk
- Typically do not provide estimates of the specific contributions of other factors to the “explained” proportion

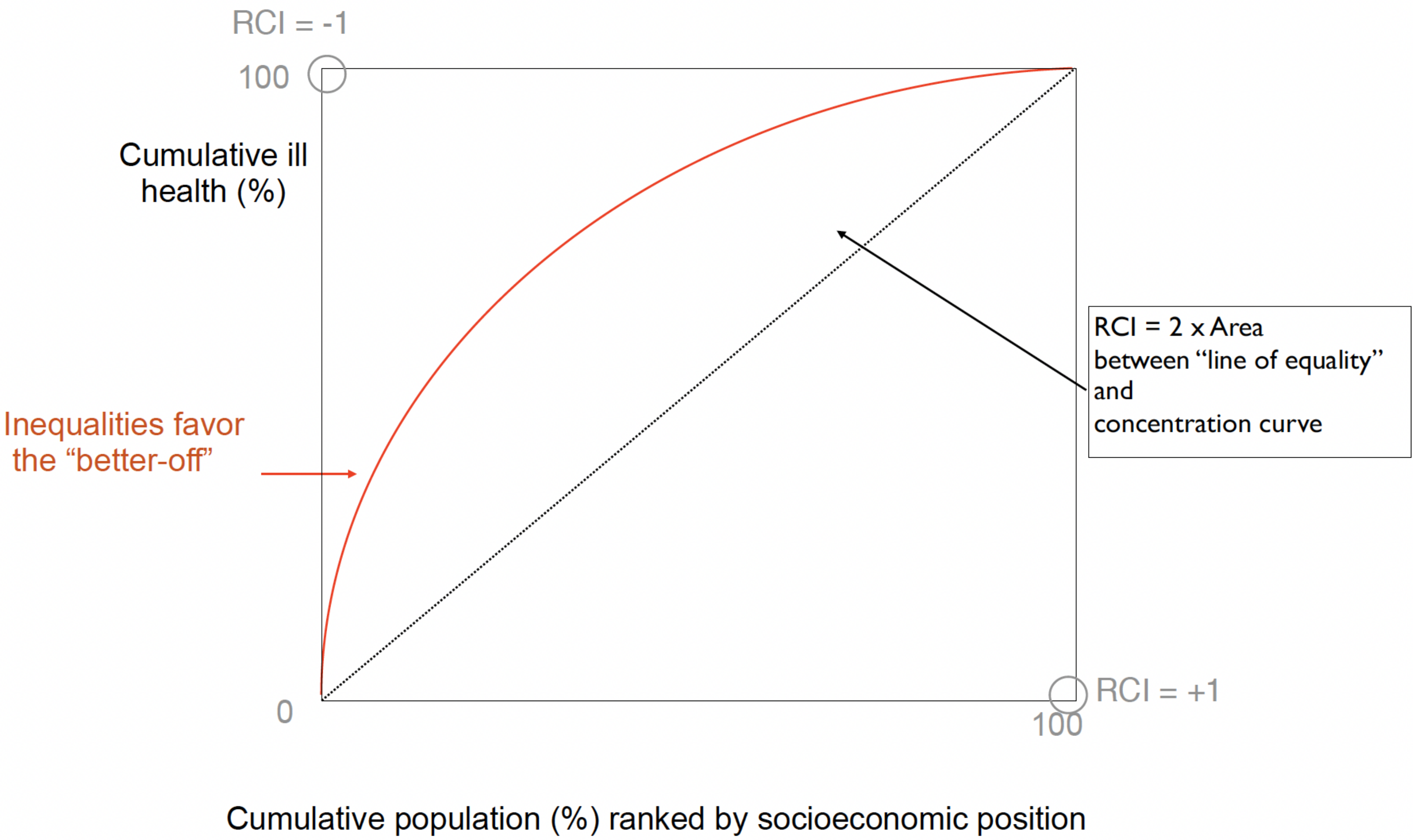
We want to understand this



By estimating something like this:



Relative Concentration Curve



Formula for writing the Concentration Index

Recall that we can write the CI as:

$$RCI = \frac{2}{n\mu}\sum_{i=1}^{n} y_i R_i - 1$$

where μ is the mean of y_i (e.g., smoking status), R_i is the fractional rank of the *i*th person in the socioeconomic (i.e., income) distribution.

The basic idea here is to develop a model for predicting y using several determinants, then plug that model back into the equation for the RCI

Kakwani et al. (1997)

Decomposition of the RCI

Since the RCI is a function of a health variable (y_i) and a socioeconomic rank variable (R_i) , i.e.

$$RCI = \frac{2}{n\mu} \sum_{i=1}^{n} y_i R_i - 1$$

Then suppose that one can write a regression equation expressing the health outcome of interest (y_i) as a function of several k_i determinants (e.g., age, gender, urban/rural status):

$$y_i = \alpha + \sum \beta_x x_{k_i} + \epsilon_i$$

Wagstaff et al. *J Econometrics* 2003

Decomposition of the RCI

Since *RCI* is a function of y_i and socioeconomic rank, one can then re-express the concentration index as:

$$RCI = \sum (\beta_k \bar{x}_k / \mu) RCI_k + gRCI_e / \mu$$

Where

- μ is the mean of *y*,
- $\bar{x}_k$ is the mean of x_k,
- β_k is the regression coefficient for x_k, and
- RCI_k is the concentration index for x_k.

The basic idea: how much of the overall inequality is due to other factors that are both differentially distributed by x (income) and also affect y (e.g., smoking)?

Explained and unexplained components

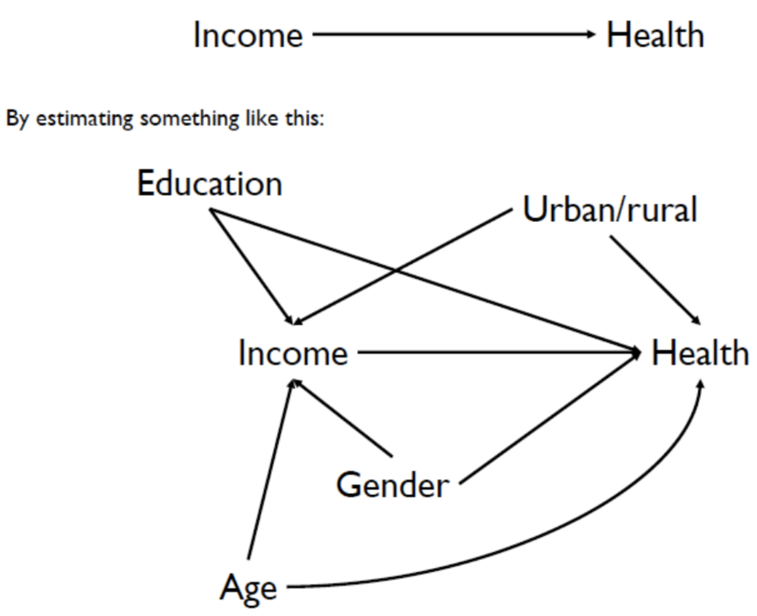
This equation results in 2 components of socioeconomic inequality:

$$RCI = \sum (\beta_k \bar{x}_k / \mu) RCI_k + gRCI_e / \mu$$

One part $(\beta_k \bar{x}_k / \mu) RCI_k$ that is due to the association between income and other factors that predict health

The other part $(gRCI_e / \mu)$ is 'unexplained', i.e., inequality that cannot be explained by systematic variation across income groups in the determinants of health.

Two types of 'explained' components



The influence of determinants depends on 2 things:

RCI_k

the strength of the relationship between each factor and income (C_k)

$\beta_k \bar{x}_k / \mu$

the strength of the relationship between each factor and health, and its prevalence in the population (elasticity).

Procedure for decomposing the Concentration Index

1 Estimate a regression equation predicting y ('health') from its determinants $(\beta_k x_k)$:

$$y_i = \alpha + \sum \beta_x x_{k_i} + \epsilon_i$$

2 Calculate the mean of y (μ) and of each of the determinants (e.g., education, age)

3 Calculate the Concentration Index for the health variable (C) *and* for each determinant in the equation predicting health (C_k).

- That is, use each determinant x_k as the "outcome" and estimate a CI for age, CI for education, etc.

Procedure for decomposing the Concentration Index

4 Calculate the absolute contribution of each determinant by multiplying its 'elasticity' by its concentration index (C_k):

$$(\beta_k \bar{x}_k/\mu)RCI_k$$

5 Calculate the percentage contribution of each determinant:

$$[(\beta_k \bar{x}_k/\mu)RCI_k]/RCI$$

A few examples...

Decomposing socioeconomic inequality in infant mortality in Iran

Ahmad Reza Hosseinpoor,[1*] Eddy Van Doorslaer,[2] Niko Speybroeck,[1] Mohsen Naghavi,[3] Kazem Mohammad,[4] Reza Majdzadeh,[4] Bahram Delavar,[3] Hamidreza Jamshidi[3] and Jeanette Vega[1]

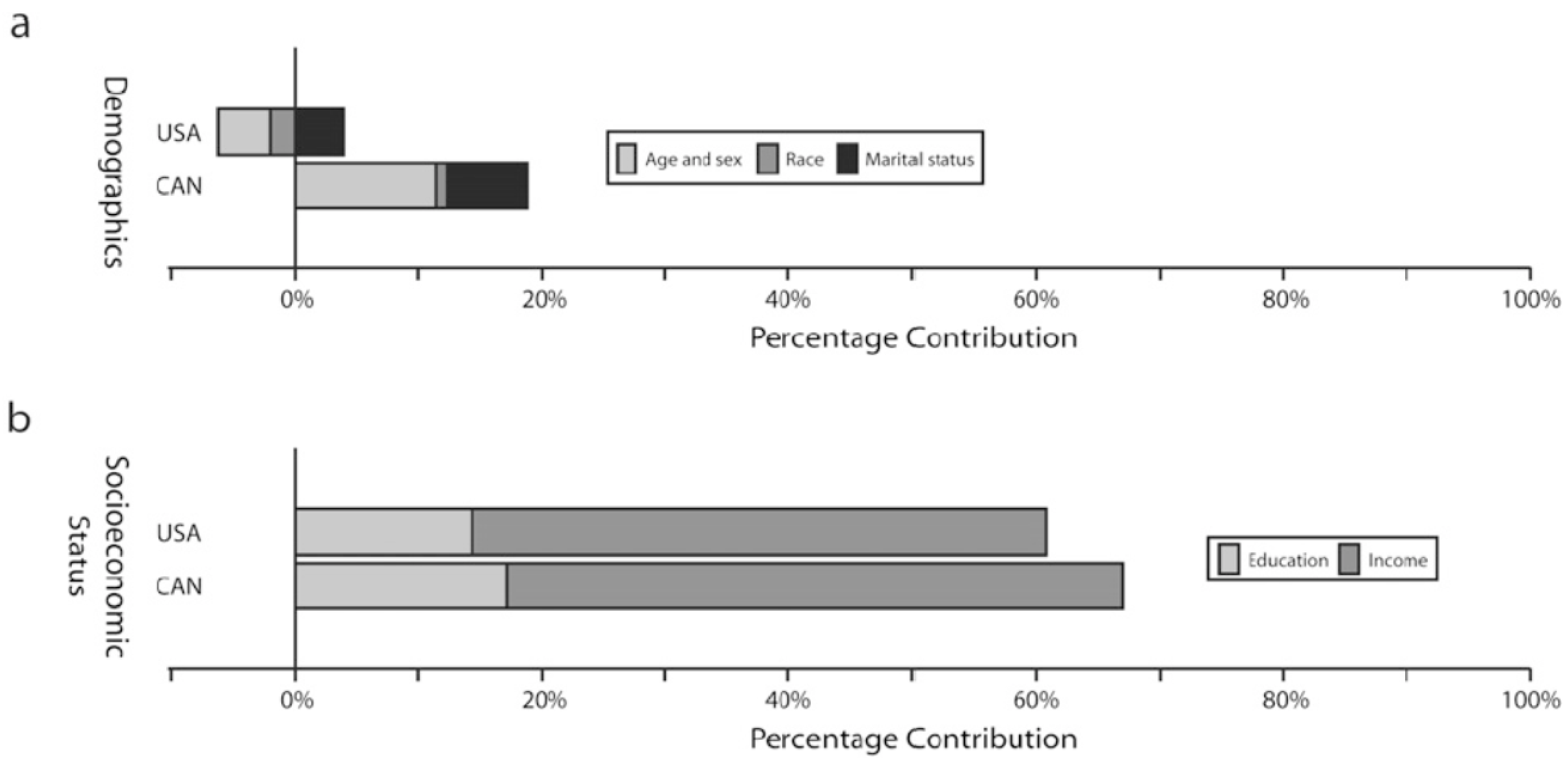
Overall Concentration index for economic status and infant mortality = 0.0413

Determinant	Beta coef.	Mean of x	Ck	Contrib to C	% of C
History of mother's stillbirth	0.5643	0.0650	-0.1001	.0010	2.5
History of mother's abortion	0.1313	0.2146	0.0396	-0.0003	-0.8
Risky birth interval	0.8028	0.1664	-0.1426	0.0054	13.0
Low economic status	0.2287	0.3634	-0.6366	0.0150	36.2
Mother's illiteracy	0.3088	0.3524	-0.2803	0.0086	20.9
Having a hygienic toilet	-0.1700	0.2916	0.3503	0.0049	11.9
Rural residency	0.1706	0.4470	-0.2663	0.0057	13.9
Total				0.0413	100.0

Hosseinpoor et al. IJE (2005)

Income-Related Health Inequalities in Canada and th United States: A Decomposition Analysis

Kimberlyn M. McGrail, PhD, Eddy van Doorslaer, PhD, Nancy A. Ross, PhD, and Claudia Sanmartin, PhD



McGrail et al. AJPH (2007)

Decomposing income-related inequality in cervical screening in 67 countries

Brittany McKinnon · Sam Harper · Spencer Moore

Contribution of education to income-related inequality in screening was highly variable across countries

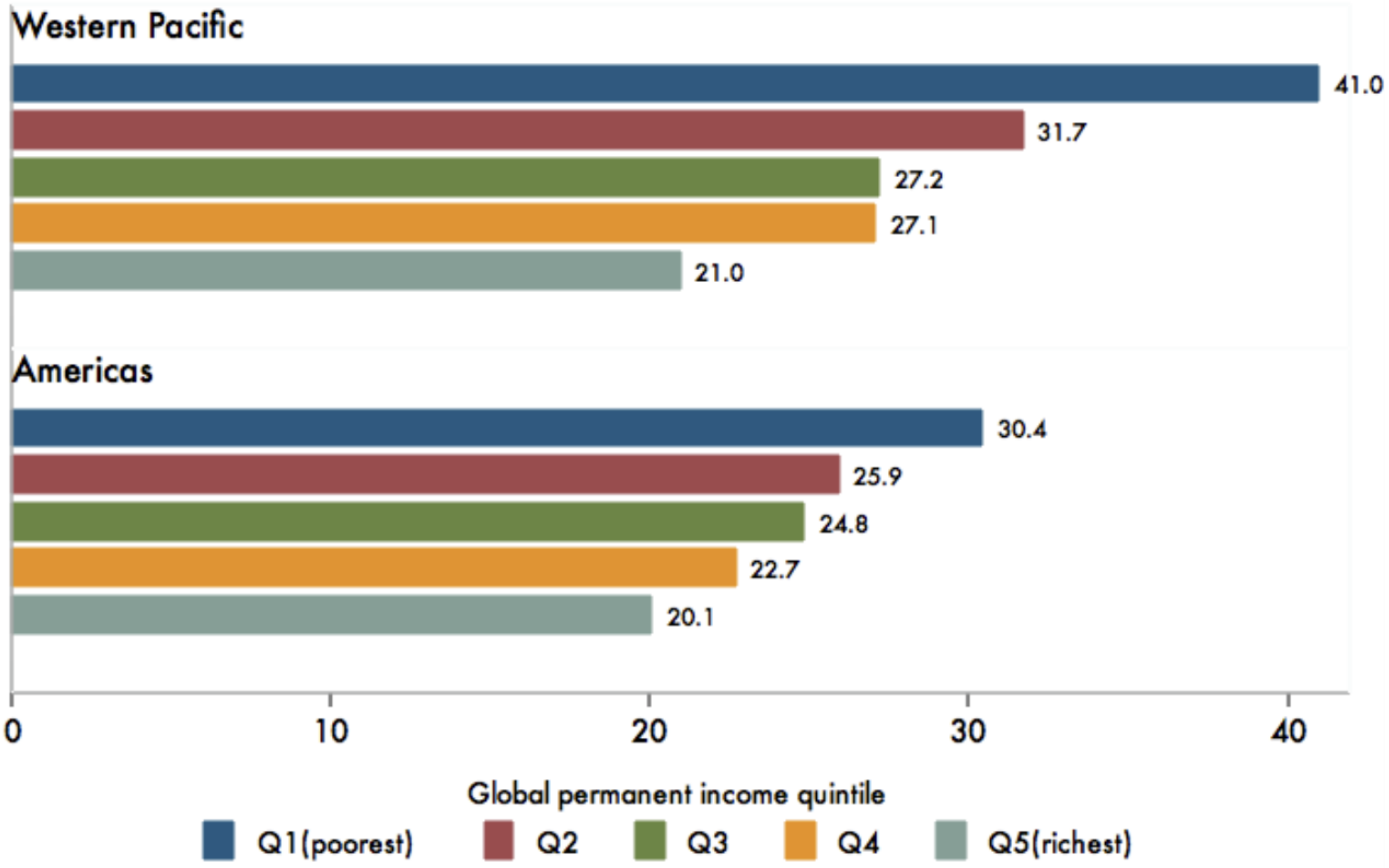
Table 4 Percentage contribution of determinants to income-related inequality in cervical screening, World Health Survey 2002–2003

WHO region	Country	Age	Income	Urban	Marital Status	Education	Recent health care[a]	Unexplained
Africa	Chad	0.1	47.2	5.2	−0.7	−2.1	5.8	58.8
	Côte d'Ivoire	48.1	−0.7	15.8	−14.0	42.6	2.9	12.8
	Ethiopia	−0.6	34.2	9.8	1.4	6.0	2.6	44.4
	Ghana	−3.1	79.4	−6.4	−4.7	12.2	3.2	20.6
	Kenya	0.0	61.8	2.3	−4.3	15.3	−0.7	29.8
	Mali	−1.5	32.5	26.1	0.4	0.0	10.9	31.6
	Mauritania	2.0	11.9	18.0	−0.4	−6.4	5.8	42.9
	Mauritius	3.5	87.3	7.3	4.3	−3.0	−6.7	18.1
	Namibia	3.4	59.9	16.2	2.5	4.9	4.2	8.8
	Senegal	−8.9	83.9	2.7	−22.2	50.6	5.9	−20.3
	South Africa	2.4	46.2	14.3	7.2	33.0	−0.7	−2.7
	Swaziland	0.3	65.3	−2.5	0.0	15.7	0.9	20.2
	Zambia	19.4	15.2	26.3	1.2	9.1	0.0	31.1
Americas	Brazil	−2.4	64.5	−2.1	4.5	39.9	4.5	−8.9

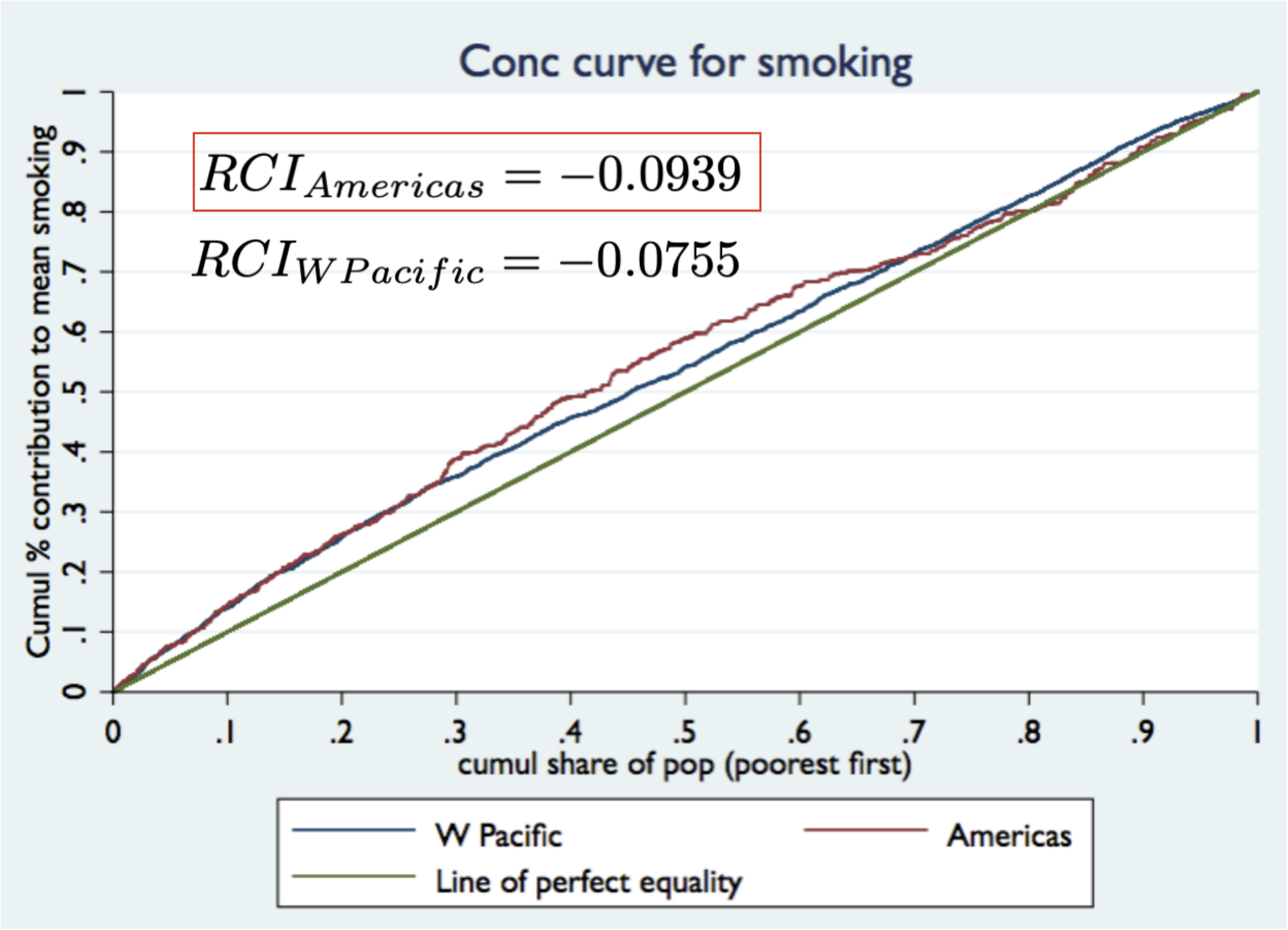
McKinnon et al. (2011)

Example: Decomposing Socioeconomic Inequality in Current Smoking

Smoking by income quintile



Concentration curve for smoking



Estimation for a specific factor: Education

Recall the decomposition formula:

$$RCI = \sum (\beta_k \bar{x}_k / \mu) RCI_k + gRCI_e / \mu$$

- Estimated β coeff on education (logit scale): -.0389 (OR = 0.96)
- Marginal effect on probability scale: -.0051 (0.5 pct points)
- Mean education: 8.9 yrs
- Mean smoking rate: 17.5%

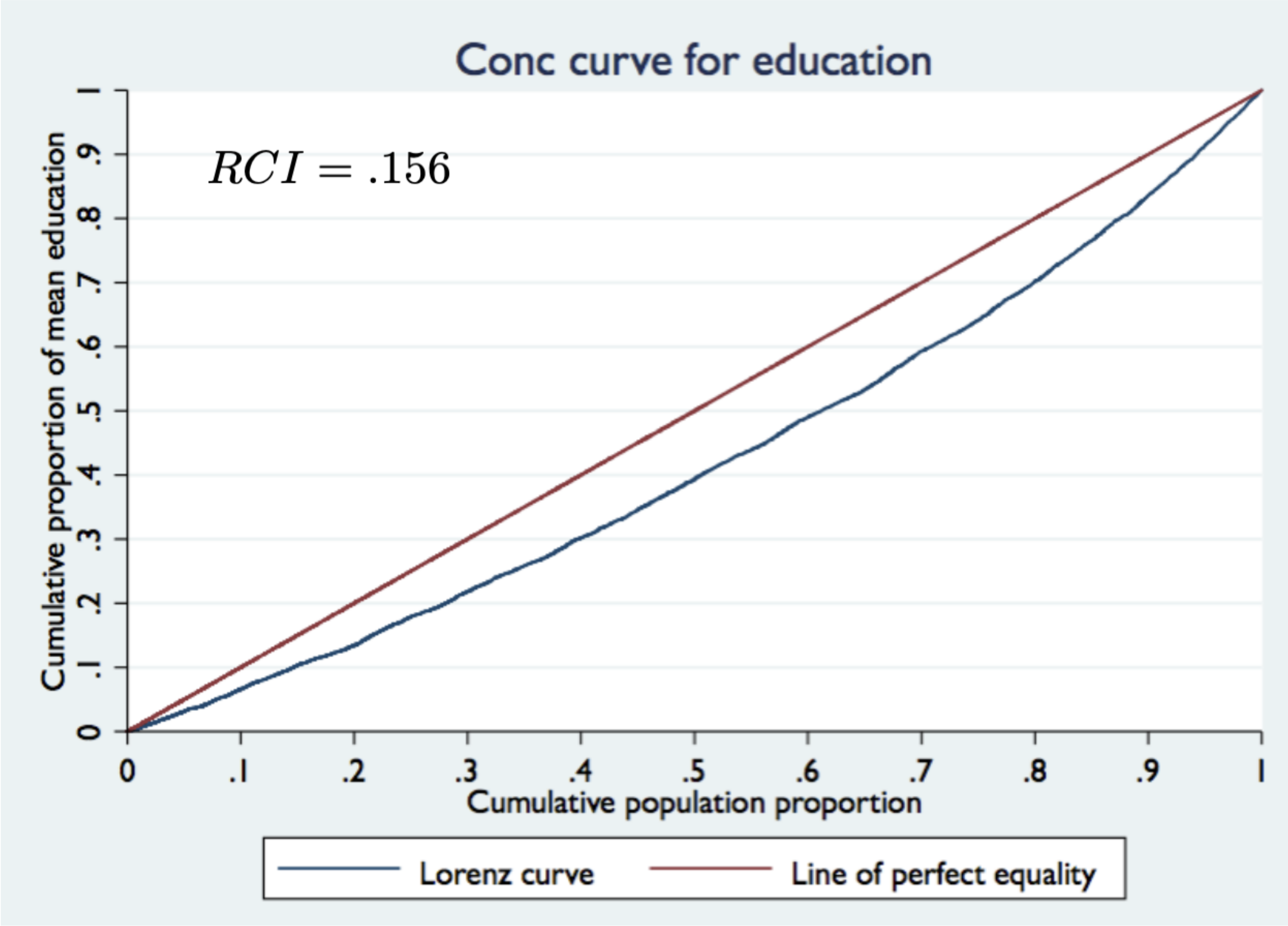
With these parameters, the elasticity of smoking with respect to education is: (-.0051 * 8.9 / .175) = -.2582

Interpretation: a 1% increase in education decreases smoking by 26% (not percentage points!).

What about the RCI for education?

Concentration curve for education

Note the y-axis is cumulative share of *education*



Estimation for a specific factor: Education

Recall the decomposition formula:

$$RCI = \sum (\beta_k \bar{x}_k / \mu) RCI_k + gRCI_e / \mu$$

So the elasticity of smoking (from the previous slide) with respect to education is (-.0051 * 8.9 / .175) = -.2582

Now we have the RCI for education = 0.156

So now we can calculate the contribution of education as:

$$\text{Elasticity} \times RCI_{ed} = -.2582 * .156 = -.04$$

Thus education accounts for -.04/ -.0939 = 41.6% of the overall RCI

Decomposition of Income-Related Inequality in Smoking: Americas region

Overall RCI = -0.094

	Elasticity	Rel Conc Index	Contribution	% Contrib
Age	3.695	0.023	0.084	-89.9%
Age^2	-1.981	0.032	-0.064	67.9%
Male	0.197	-0.055	-0.011	11.5%
BMI	-0.834	0.011	-0.009	9.6%
Urban	0.020	0.076	0.002	-1.6%
Single	0.078	-0.036	-0.003	3.0%
Divorced/Widowed	**0.161**	**-0.120**	**-0.019**	**20.7%**
Low Phys Activity	0.057	0.069	0.004	-4.2%
Mod Phys Activity	-0.023	0.025	-0.001	0.6%
Low Alcohol Consumption	**0.131**	**0.123**	**0.016**	**-17.1%**
Mod/Hi Alcohol Consumption	0.019	0.081	0.002	-1.6%
Low Fruit/Veg Consumption	0.029	-0.066	-0.002	2.0%
Self-Reported Health Good	-0.001	0.040	0.000	0.1%
Self-Reported Health Moderate	-0.043	-0.079	0.003	-3.6%
Self-Reported Health Bad/Very Bad	0.004	-0.208	-0.001	0.9%
Education	**-0.250**	**0.156**	**-0.039**	**41.6%**
Permanent Income	-0.809	0.054	-0.044	46.4%
Residual			-0.013	

Contrasting components of income-related inequality

Education:

- Elasticity stronger in W Pacific
- RCI_{ed} stronger in Americas
- Implications for intervention?

	Elasticity	RCI	Contribution	% Contribution
Western Pacific				
Income	-0.51	0.065	-0.033	43.7%
Urbanicity	0.06	0.252	0.016	-20.8%
Education	-0.43	0.096	-0.041	54.5%
Americas				
Income	-0.81	0.054	-0.044	46.4%
Urbanicity	0.02	0.076	0.002	-1.6%
Education	-0.25	0.156	-0.039	41.6%

Caveats for decomposing the RCI

Decomposition results will be sensitive to the choice of determinants included (i.e., how well-specified the model is for predicting y).

The regression equations are predictive and not causal models.

Main utility is not in estimating the potential impact on y of changing the distribution of socioeconomic position, but in indicating the potential role that other factors may play in generating socioeconomic inequalities in health.

3. Decomposition

Idea for Decomposition of Means

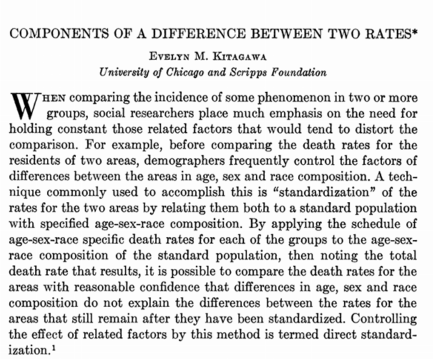
The core idea is to explain the distribution of the outcome variable in question by a set of factors that vary systematically with exposure status.

Thus, we want to know, on average, **why the mean level of health or disease differs between exposed and unexposed groups**.

Since, for most health outcomes there are multiple determinants, we may want to know which of these determinants plays more or less important roles in explaining the difference in average outcomes.

“Unpacking” or “decomposing” difference.

Origins



COMPONENTS OF A DIFFERENCE BETWEEN TWO RATES*

EVELYN M. KITAGAWA
University of Chicago and Scripps Foundation

WHEN comparing the incidence of some phenomenon in two or more groups, social researchers place much emphasis on the need for holding constant those related factors that would tend to distort the comparison. For example, before comparing the death rates for the residents of two areas, demographers frequently control the factors of differences between the areas in age, sex and race composition. A technique commonly used to accomplish this is "standardization" of the rates for the two areas by relating them both to a standard population with specified age-sex-race composition. By applying the schedule of age-sex-race specific death rates for each of the groups to the age-sex-race composition of the standard population, then noting the total death rate that results, it is possible to compare the death rates for the areas with reasonable confidence that differences in age, sex and race composition do not explain the differences between the rates for the areas that still remain after they have been standardized. Controlling the effect of related factors by this method is termed direct standardization.[1]

Evelyn Kitagawa was sociologist and demographer who devised a non-parametric method (1955) for decomposing differences between rates, refined by Prithwis das Gupta in 1978.

- Focused on understanding group contributions to rate differences.

Studies by Oaxaca (1973) and Blinder (1973) applied regression-based decomposition methods to analyze the wage gap between men and women and between whites and blacks in the USA.

- Focused on how much of wage gap was 'explained' by differences in observable characteristics

Brief note on interpretation

Decomposition methods are based on regression analyses, and thus all of the usual caveats about good specification apply

If regressions are purely descriptive, they reveal the associations that characterize the health inequality Then inequality is explained in a statistical sense but implications for policies to reduce inequality are limited

If data allow identification of causal effects, then the factors that generate the inequality are identified Then one can (potentially) draw conclusions about how policies would impact on inequality

O'Donnell 2008

Eur J Health Econ (2011) 12:17–28
DOI 10.1007/s10198-010-0220-z

ORIGINAL PAPER

Inequalities in the use of health services between immigrants and the native population in Spain: what is driving the differences?

Dolores Jiménez-Rubio · Cristina Hernández-Quevedo

Abstract In Spain, a growing body of literature has drawn attention to analysing the differences in health and health resource utilisation of immigrants relative to the autochthonous population. The results of these studies generally find substantial variations in health-related patterns between both population groups. In this study, we use the Oaxaca-Blinder decomposition technique to explore to what extent disparities in the probability of using medical care use can be attributed to differences in the determinants of use due to, e.g. a different demographic structure of the immigrant collective, rather than to a different effect of health care use determinants by nationality, holding all other factors equal. Our findings show that unexplained factors associated to immigrant status determine to a great extent disparities in the probability of using hospital, specialist and emergency services of immigrants relative to Spaniards, while individual characteristics, in particular self-reported health and chronic conditions, are much more important in explaining the differences in the probability of using general practitioner services between immigrants and Spaniards.

Kitagawa-Blinder-Oaxaca: Basic Idea

Two potential sources of mean differences in outcomes

1. Means

Differences in the prevalence of determinants of outcome

2. Effects

Differences in the effect of a given determinant on the outcome (i.e., effect measure modification)

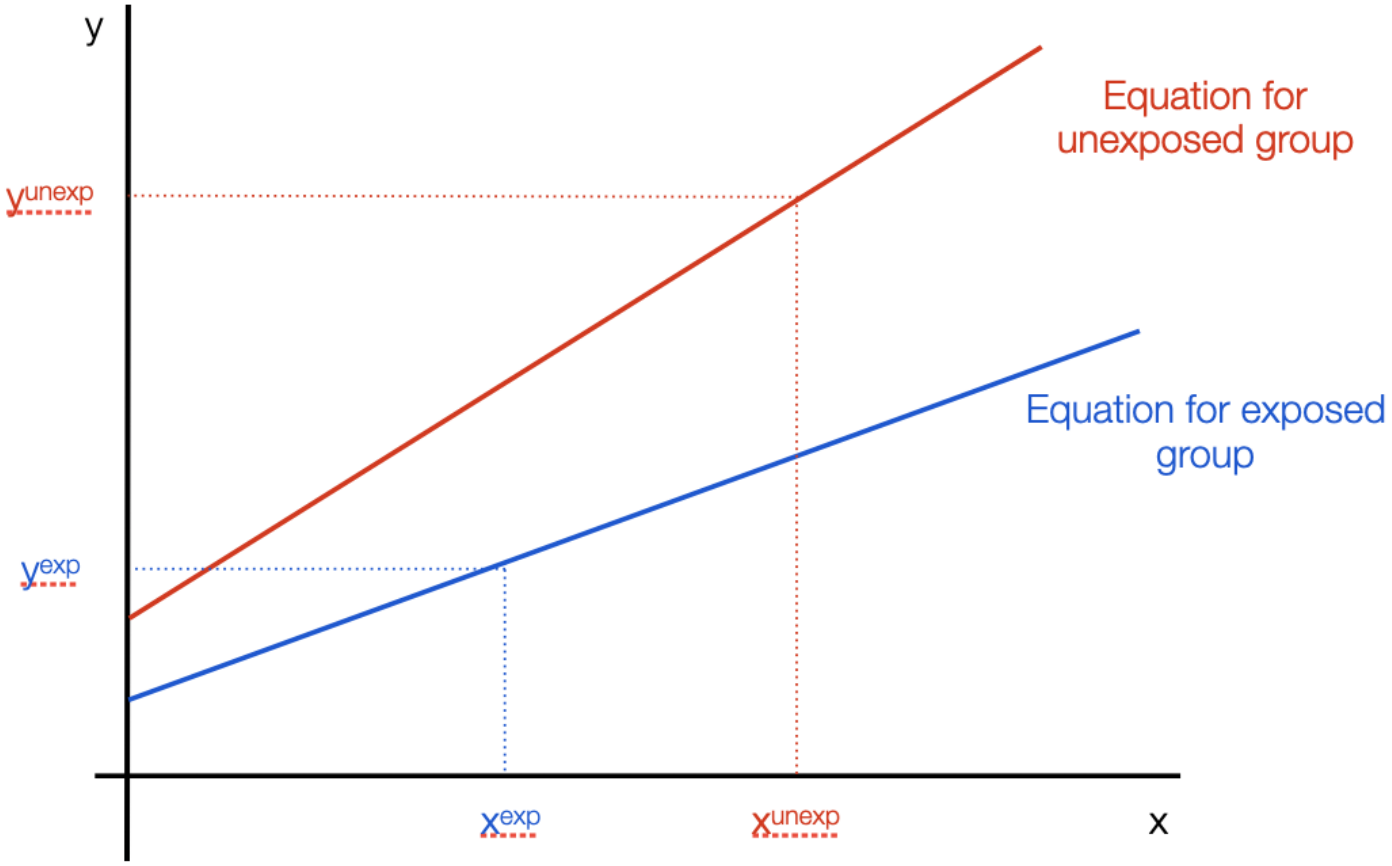
Think of 2 regressions for a given determinant X:

1. Exposed
2. Unexposed

Each generates its own coefficient and uses its own mean.

Use these to generate counterfactuals.

$$y_i = \begin{cases} \beta^{exp} x_i + \varepsilon_i^{exp} \text{ if exposed} \\ \beta^{unexp} x_i + \varepsilon_i^{unexp} \text{ if unexposed} \end{cases}$$



Two ways of expressing the mean difference in y

The overall gap between exposed and unexposed can be written as a function of differences the respective beta coefficients, evaluated at the mean for each group:

$$y^{exp} - y^{unexp} = \beta^{exp}\bar{x}^{exp} - \beta^{unexp}\bar{x}^{unexp}$$

This way:

$$y^{exp} - y^{unexp} = \Delta\bar{x}\beta^{unexp} + \Delta\beta x^{exp}$$

$$\text{where } \Delta\bar{x} = \bar{x}^{exp} - \bar{x}^{unexp} \;\; and \;\; \Delta\beta = \beta exp - \beta unexp$$

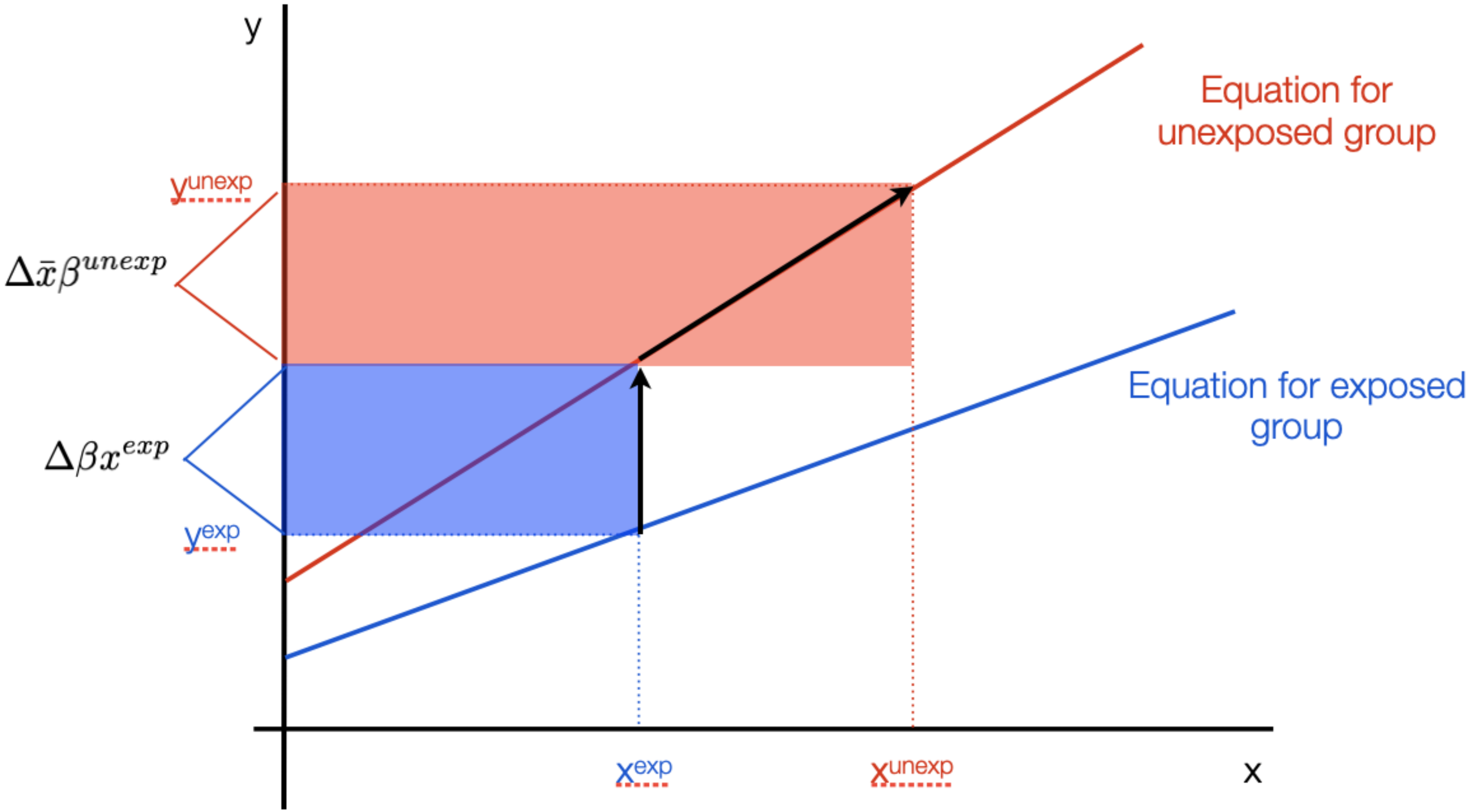
or, equivalently:

$$y^{exp} - y^{unexp} = \Delta\bar{x}\beta^{exp} + \Delta\beta x^{unexp}$$

First method

$$y^{exp} - y^{unexp} = \Delta\bar{x}\beta^{unexp} - \Delta\beta x^{exp}$$

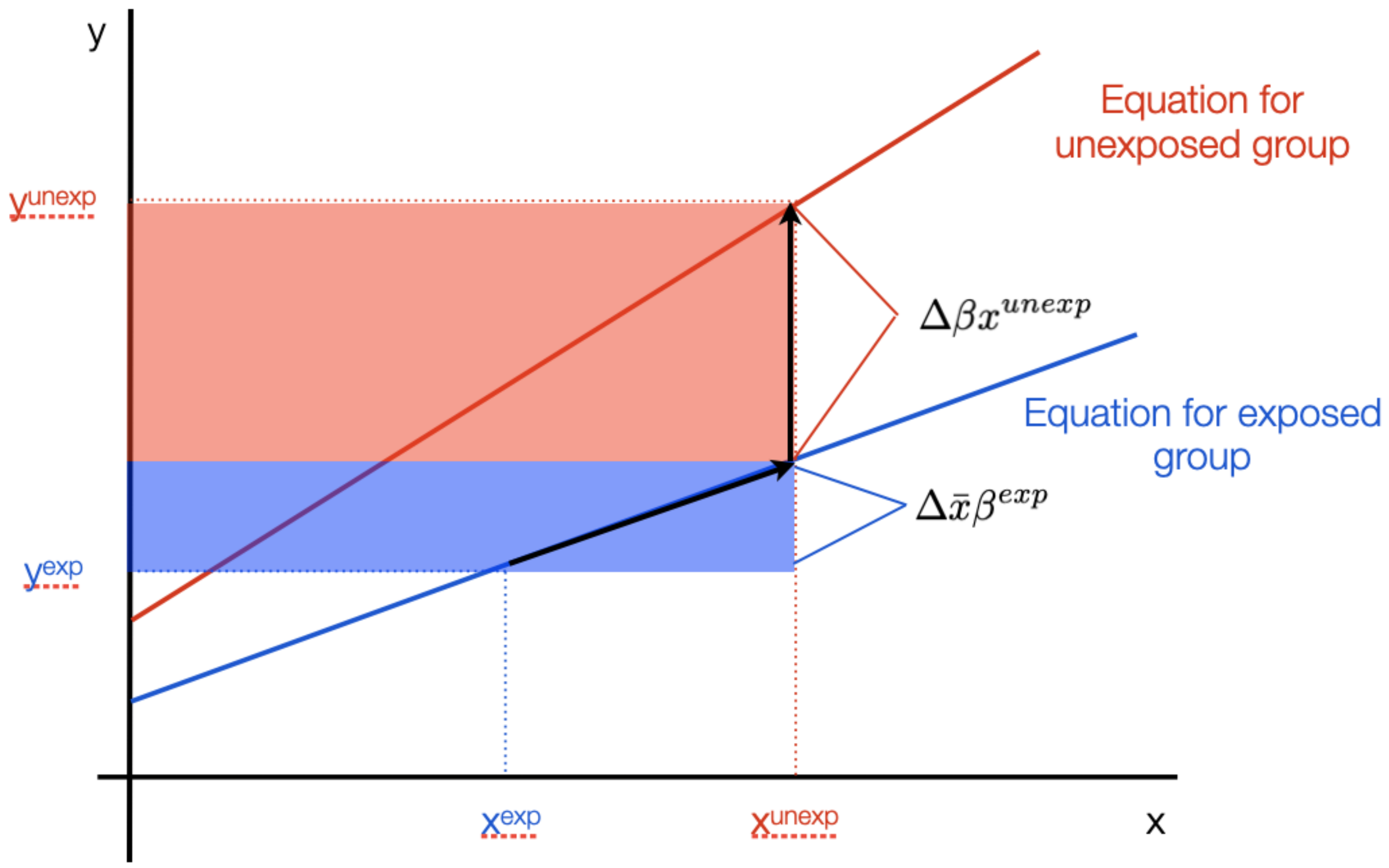
- Coefficients of unexposed
- Means of exposed



Second method

- Coefficients of exposed
- Means of unexposed

$$y^{exp} - y^{unexp} = \Delta\bar{x}\beta^{exp} - \Delta\beta x^{unexp}$$



The two methods are equally valid

In the first,the differences in the x's are weighted by the coefficients of the unexposed group and the differences in the coefficients are weighted by the x's of the exposed group:

$$y^{exp} - y^{unexp} = \Delta\bar{x}\beta^{unexp} - \Delta\beta x^{exp}$$

whereas, in the second, the differences in the x's are weighted by the coefficients of the exposed group and the differences in the coefficients are weighted by the x's of the unexposed group:

$$y^{exp} - y^{unexp} = \Delta\bar{x}\beta^{exp} - \Delta\beta x^{unexp}$$

General decomposition formula shows the mean gap as deriving from a difference in endowments (E), a gap in coefficients (C), and a gap arising from the interaction of endowments and coefficients (CE):

$$
\begin{aligned}
y^{exp} - y^{unexp} &= \Delta\bar{x}\beta^{exp} + \Delta\beta x^{exp} + \Delta\bar{x}\Delta\beta \\
&= E + C + CE
\end{aligned}
$$

- Method 1 includes interaction with “explained” part:

$$
\begin{aligned}
y^{exp} - y^{unexp} &= \Delta\bar{x}\beta^{unexp} + \Delta\beta x^{exp} \\
&= (E + CE) + C
\end{aligned}
$$

- Method 2 includes interaction with “unexplained” part:

$$
\begin{aligned}
y^{exp} - y^{unexp} &= \Delta\bar{x}\beta^{exp} + \Delta\beta x^{unexp} \\
&= E + (CE + C)
\end{aligned}
$$

Example: Decomposing Educational Differences in Blood Pressure

Basic question

What is the average difference in blood pressure between those with low vs. high education?

How much of this difference is due to the fact that determinants of blood pressure (e.g., BMI, smoking, demographics) differ between low and high educated groups?

Any residual difference is due to educational differences in the associations of risk factors for blood pressure.

Example data

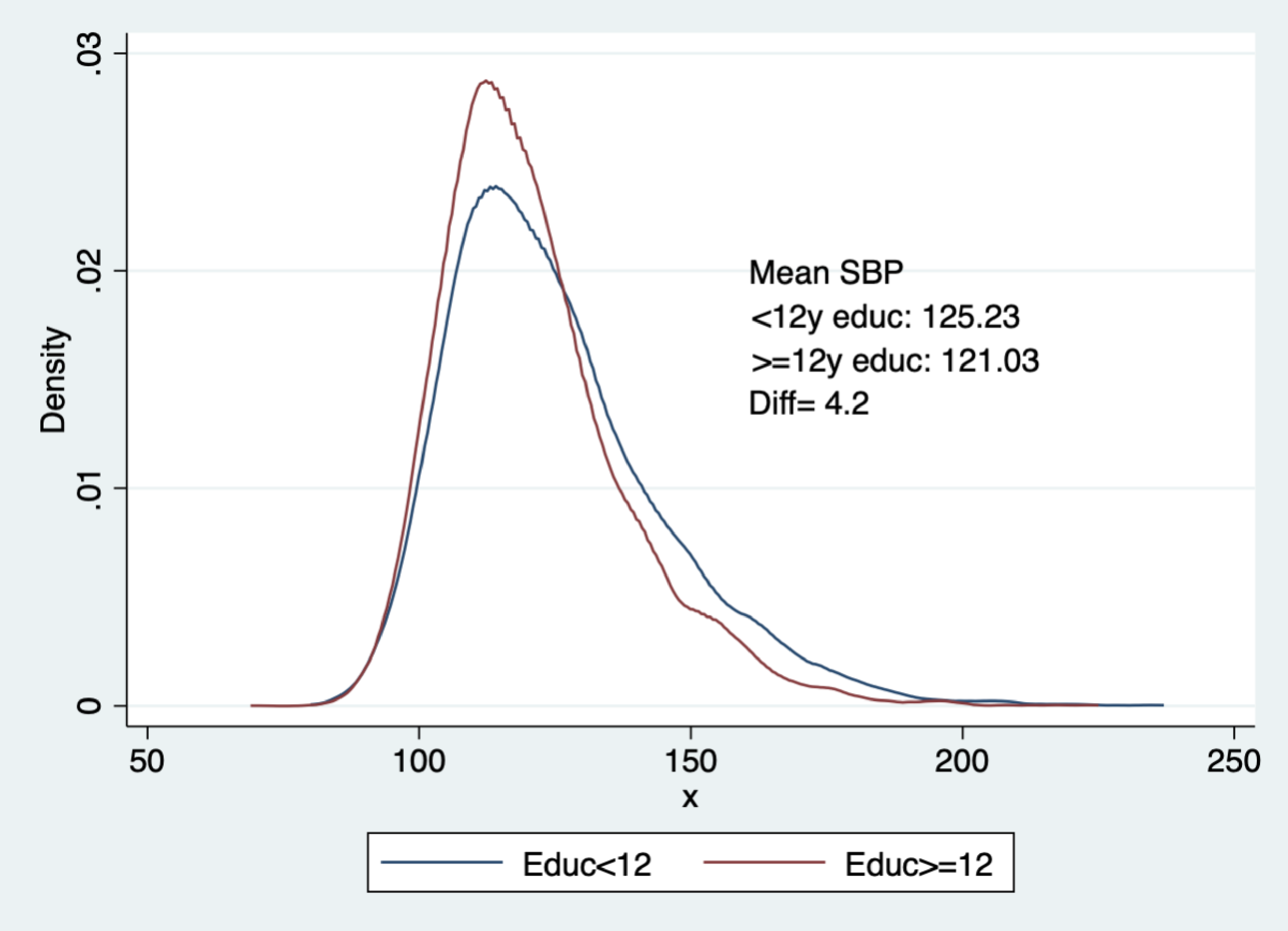
US NHANES follow up survey (1988-2006), baseline data

Systolic blood pressure as outcome (mmHg)

Overall difference by education (0: >=12y educ, 1: <12y educ)

Potential determinants (the Xs):

- age (years)
- age squared
- race (1 = non-white, 0 = other)
- marital status (1=married, 0=other)
- body mass index (kg/m^2)
- smoking (1=current smoker, 0=other)


Density
.03
.02
.01
0
50
100
150
200
250
x
Mean SBP
<12y educ: 125.23
>=12y educ: 121.03
Diff= 4.2
Educ<12
Educ>=12

Differences in determinants

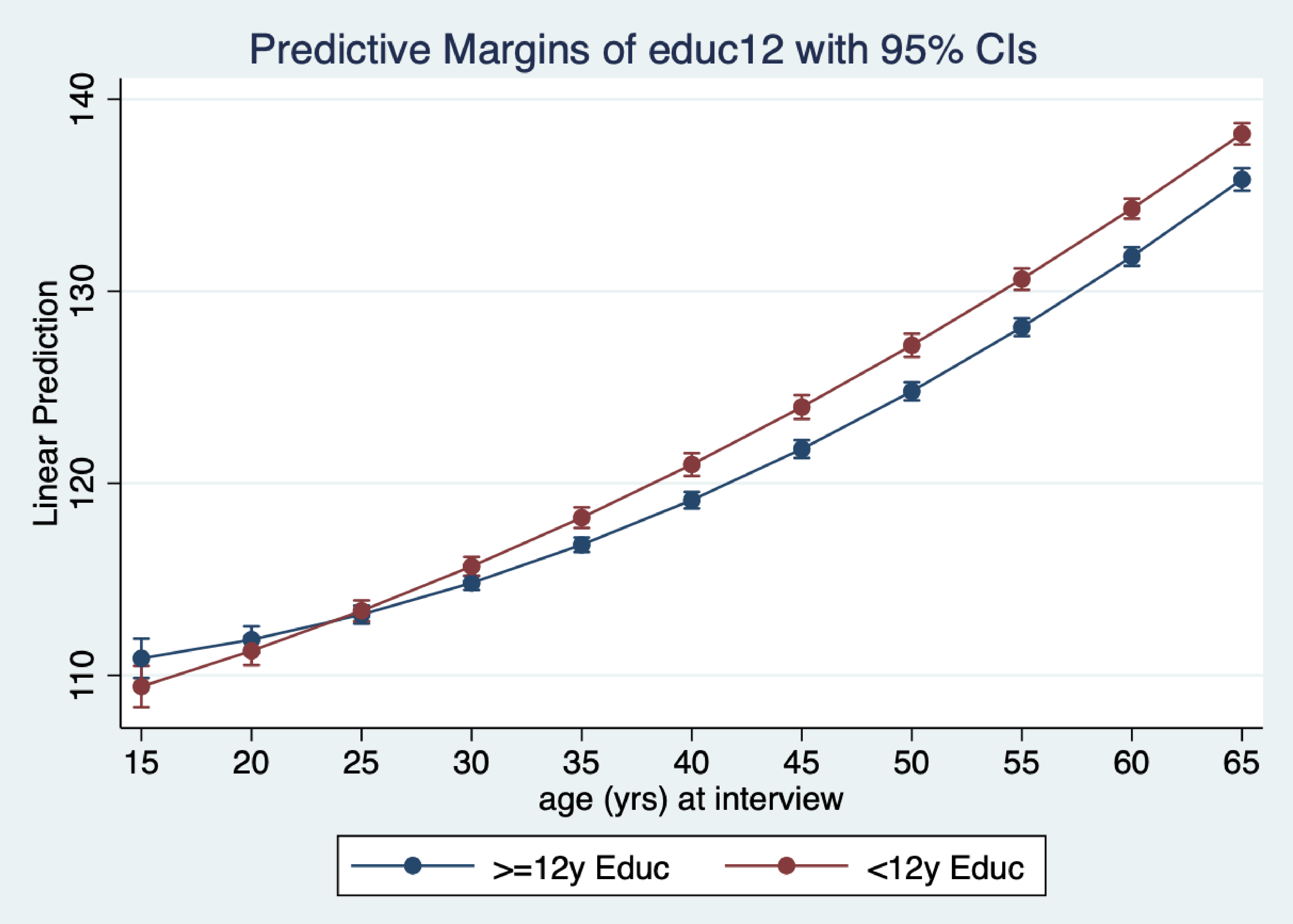
- Lower educated have higher BMI and are more likely to be smokers, as well as being older

	Covariate means			
	<12y Educ		>=12y Educ	
Variable	$\overline{x}$	SD($\overline{x}$)	$\overline{x}$	SD($\overline{x}$)
Age	44.6	18.7	40.9	15.8
Age*Age	2338	1705	1920	1436
Non-white	0.33	0.47	0.36	0.48
Married	0.42	0.49	0.40	0.49
BMI	27.4	5.6	26.9	5.6
Smoker	0.31	0.46	0.25	0.43

Differences in coefficients

- BMI and smoking both have larger coefficients for the better educated group.
- Age has a slightly stronger association for the less educated.

	Regression coefficients			
	<12y Educ		>=12y Educ	
Variable	β	SE(β)	β	SE(β)
Age	0.60	0.01	0.53	0.01
Age*Age	0.00	0.00	0.01	0.00
Non-white	2.17	0.44	2.43	0.31
Married	0.92	0.44	0.89	0.32
BMI	0.38	0.04	0.61	0.02
Smoker	0.73	0.44	1.10	0.33
Intercept	110.86	1.11	102.20	0.74


Predictive Margins of educ12 with 95% CIs
Linear Prediction
140
130
120
110
15
20
25
30
35
40
45
50
55
60
65
age (yrs) at interview
>=12y Educ
<12y Educ

Contribution of covariate differences → Covariate Means

Contribution of coefficient differences → Coefficients

Interaction between coefficients and covariates → Interaction

	Coefficients used in decomposition:					
	<12y Educ		>=12y Educ		Pooled	
SBP (mmHg)	Est.	SE	Est.	SE	Est.	SE
>=12y Educ	125.23	0.25	125.23	0.25	121.03	0.17
<12y Educ	125.23	0.25	125.23	0.25	125.23	0.25
Difference	-4.20	0.30	-4.20	0.30	-4.20	0.30
Δ due to:						
Covariate Means	**-2.77**	**0.20**	**-2.88**	**0.19**	**-2.85**	**0.19**
Age	-2.14	0.17	-1.89	0.16	-2.00	0.16
Age*Age	-0.46	0.08	-0.69	0.07	-0.59	0.06
Non-white	0.07	0.02	0.07	0.02	0.07	0.02
Married	-0.02	0.01	-0.02	0.01	-0.02	0.01
BMI	-0.18	0.04	-0.29	0.06	-0.25	0.05
Smoker	-0.04	0.03	-0.06	0.02	-0.06	0.02
Coefficients	**-1.29**	**0.25**	**-1.40**	**0.26**	**-1.32**	**0.25**
Age	-0.13	0.03	0.11	0.03	-0.02	0.01
Age*Age	0.79	0.35	0.56	0.25	0.69	0.32
Non-white	0.08	0.18	0.09	0.19	0.08	0.19
Married	-0.01	0.23	-0.01	0.21	-0.01	0.23
BMI	0.06	0.02	-0.05	0.02	0.02	0.01
Smoker	0.11	0.17	0.09	0.14	0.11	0.16
Intercept	-2.20	0.48	-2.20	0.48	-2.20	0.47
Interaction	**-0.11**	**0.11**	**0.11**	**0.11**		

	Coeffi	
	<12y Educ	
SBP (mmHg)	Est.	SE
>=12y Educ	125.23	0.25
<12y Educ	125.23	0.25
Difference	-4.20	0.30
Δ due to:		
Covariate Means	**-2.77**	**0.20**
Age	-2.14	0.17
Age*Age	-0.46	0.08
Non-white	0.07	0.02
Married	-0.02	0.01
BMI	-0.18	0.04
Smoker	-0.04	0.03
Coefficients	**-1.29**	**0.25**
Age	-0.13	0.03
Age*Age	0.79	0.35
Non-white	0.08	0.18
Married	-0.01	0.23
BMI	0.06	0.02
Smoker	0.11	0.17
Intercept	-2.20	0.48
Interaction	**-0.11**	**0.11**

Contribution of covariate differences

SBP among the low educated group would be 2.8 mmHg lower if they had the same covariate characteristics as the higher educated.

Most of this difference comes from differences in the distribution of age.

Why positive? This means that the SBP difference would be even larger if the low educated had the same percentage non-white as the higher educated.

SBP (mmHg)	Coeffi <12y Educ Est.	SE
>=12y Educ	125.23	0.25
<12y Educ	125.23	0.25
Difference	-4.20	0.30
Δ due to:		
Covariate Means	**-2.77**	**0.20**
Age	-2.14	0.17
Age*Age	-0.46	0.08
Non-white	0.07	0.02
Married	-0.02	0.01
BMI	-0.18	0.04
Smoker	-0.04	0.03
Coefficients	**-1.29**	**0.25**
Age	-0.13	0.03
Age*Age	0.79	0.35
Non-white	0.08	0.18
Married	-0.01	0.23
BMI	0.06	0.02
Smoker	0.11	0.17
Intercept	-2.20	0.48
Interaction	**-0.11**	**0.11**

Contribution of coefficient differences

SBP among the low educated group would be 1.3 mmHg lower if they had the same regression coefficients as the higher educated.

Most of this difference is captured by the intercept (i.e., unmeasured factors).

Why positive? This means that the SBP difference would be even larger if smoking had the same effect in low educated as it does in the higher educated.

Similar results if we use the coefficients of the higher educated to weight the covariate differences

	Coefficients used in decomposition:					
	<12y Educ		>=12y Educ		Pooled	
SBP (mmHg)	Est.	SE	Est.	SE	Est.	SE
>=12y Educ	125.23	0.25	125.23	0.25	121.03	0.17
<12y Educ	125.23	0.25	125.23	0.25	125.23	0.25
Difference	-4.20	0.30	-4.20	0.30	-4.20	0.30
Δ due to:						
Covariate Means	**-2.77**	**0.20**	**-2.88**	**0.19**	**-2.85**	**0.19**
Age	-2.14	0.17	-1.89	0.16	-2.00	0.16
Age*Age	-0.46	0.08	-0.69	0.07	-0.59	0.06
Non-white	0.07	0.02	0.07	0.02	0.07	0.02
Married	-0.02	0.01	-0.02	0.01	-0.02	0.01
BMI	-0.18	0.04	-0.29	0.06	-0.25	0.05
Smoker	-0.04	0.03	-0.06	0.02	-0.06	0.02
Coefficients	**-1.29**	**0.25**	**-1.40**	**0.26**	**-1.32**	**0.25**
Age	-0.13	0.03	0.11	0.03	-0.02	0.01
Age*Age	0.79	0.35	0.56	0.25	0.69	0.32
Non-white	0.08	0.18	0.09	0.19	0.08	0.19
Married	-0.01	0.23	-0.01	0.21	-0.01	0.23
BMI	0.06	0.02	-0.05	0.02	0.02	0.01
Smoker	0.11	0.17	0.09	0.14	0.11	0.16
Intercept	-2.20	0.48	-2.20	0.48	-2.20	0.47
Interaction	**0.11**	**0.11**	**0.11**	**0.11**		

	Coefficients used in decomposition:					
	<12y Educ		>=12y Educ		Pooled	
SBP (mmHg)	Est.	SE	Est.	SE	Est.	SE
>=12y Educ	125.23	0.25	125.23	0.25	121.03	0.17
<12y Educ	125.23	0.25	125.23	0.25	125.23	0.25
Difference	-4.20	0.30	-4.20	0.30	-4.20	0.30
Δ due to:						
Covariate Means	**-2.77**	**0.20**	**-2.88**	**0.19**	**-2.85**	**0.19**
Age	-2.14	0.17	-1.89	0.16	-2.00	0.16
Age*Age	-0.46	0.08	-0.69	0.07	-0.59	0.06
Non-white	0.07	0.02	0.07	0.02	0.07	0.02
Married	-0.02	0.01	-0.02	0.01	-0.02	0.01
BMI	-0.18	0.04	-0.29	0.06	-0.25	0.05
Smoker	-0.04	0.03	-0.06	0.02	-0.06	0.02
Coefficients	**-1.29**	**0.25**	**-1.40**	**0.26**	**-1.32**	**0.25**
Age	-0.13	0.03	0.11	0.03	-0.02	0.01
Age*Age	0.79	0.35	0.56	0.25	0.69	0.32
Non-white	0.08	0.18	0.09	0.19	0.08	0.19
Married	-0.01	0.23	-0.01	0.21	-0.01	0.23
BMI	0.06	0.02	-0.05	0.02	0.02	0.01
Smoker	0.11	0.17	0.09	0.14	0.11	0.16
Intercept	-2.20	0.48	-2.20	0.48	-2.20	0.47
Interaction	**0.11**	**0.11**	**0.11**	**0.11**		

Using coefficients from a model pooling both groups together also gives similar results.

No interaction term because only one set of coefficients is used for both group predictions.

Caveat: results depend on specification

Adding gender increases the "explained" component (i.e., "endowments") from -2.77 to -2.95, so important consequences for how much of the gap is "unexplained"

```
. oaxaca systolic agec agec2 nonwhite married bmic current male, by(educ12) nodetail

Blinder-Oaxaca decomposition                      Number of obs     =     15,859
                                                    Model           =     linear
Group 1: educ12 = 0                                 N of obs 1      =       9532
Group 2: educ12 = 1                                 N of obs 2      =       6327

------------------------------------------------------------------------------
    systolic |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
overall      |
     group_1 |   121.0268   .1744272   693.85   0.000     120.6849    121.3686
     group_2 |   125.1985   .2500719   500.65   0.000     124.7084    125.6886
  difference |  -4.171762   .3048947   -13.68   0.000    -4.769345    -3.57418
  endowments |  -2.949963   .2080375   -14.18   0.000     -3.35771   -2.542217
coefficients |  -1.023872   .2494773    -4.10   0.000    -1.512839   -.5349059
 interaction |  -.1979264   .1126793    -1.76   0.079    -.4187737    .0229209
------------------------------------------------------------------------------
```

Methods frontier

- Attempting to reconcile the non-causal framework of KBO with mediation methods, new estimators.

Meaningful Causal Decompositions in Health Equity Research

Definition, Identification, and Estimation Through a Weighting Framework

John W. Jackson[a,b,c,d,e]

Abstract: Causal decomposition analyses can help build the evidence base for interventions that address health disparities (inequities). They ask how disparities in outcomes may change under hypothetical intervention. Through study design and assumptions, they can rule out alternate explanations such as confounding, selection bias, and measurement error, thereby identifying potential targets for intervention. Unfortunately, the literature on causal decomposition analysis and related methods have largely ignored equity concerns that actual interventionists would respect, limiting their relevance and practical value. This article addresses these concerns by explicitly considering what covariates the outcome disparity and hypothetical intervention adjust for (so-called allowable covariates) and the equity value judgments

(*Epidemiology* 2021;32: 282–290)

Health disparities represent differences across
ileged versus socially marginalized groups
considers inequitable, avoidable, and unjust.[1]
that address disparities[2] usually affect risk fac
overrepresented among marginalized groups.
evidence base draws from studies that compare
disparities before and after adjustment for a ris
difference method[3]). But the changes seen after

Jackson (2021)

Summary

Various decomposition techniques exist that may be useful for analyzing social determinants of health Life table decomposition—over time or between groups, or both Regression-based decomposition of Concentration Index Oaxaca decomposition of mean health between groups

All of these techniques make assumptions that need to be evaluated in the course of analysis

When used properly, decomposition techniques can help to provide key evidence on why health inequalities exist and change over time.