

Empirical research on economic inequality
Lecture notes (preliminary version)

Maximilian Kasy
Harvard University

January 15, 2016

Contents

1	Preface	4
1.1	Why research economic inequality?	5
1.2	Acknowledgments	6
1.3	Readings	7
2	The 1% – The Pareto distribution and maximum likelihood	10
2.1	Definition of the Pareto distribution	10
2.2	Maximum likelihood	12
2.3	Censored data	12
2.4	Maximum likelihood with censored data	13
2.5	Piketty’s $r - g$ and the Pareto parameter	14
2.6	Matlab exercises	15
3	Gender Inequality – Elasticities of Labor Supply	17
3.1	Elasticities of labor supply	17
3.2	Decomposing changes	19
3.3	Some critical remarks	20
4	Discrimination – Experiments	21
4.1	Potential outcomes and causal effects	22
4.2	(Randomized) controlled experiments	23
4.3	Estimation and t-tests	23
4.4	Matlab exercises	25
5	Unions – Distributional decompositions	26
5.1	Setup	26
5.2	Reweighting	27
5.3	Causal interpretation	28
5.4	Estimation	28
5.5	Matlab exercises	29
6	Migration, technology, education – Estimating labor demand	31
6.1	Backwards-engineering wage regressions	31
6.2	General labor demand	32

6.3	Demand based on a production function	32
6.4	The CES production function	34
6.5	Generalizations of the CES production function	35
6.6	Instruments	36
6.7	Matlab exercises	37
7	Intergenerational mobility – Measurement error	38
7.1	Classical measurement error and transitory shocks	40
7.2	Non-classical measurement error and the lifetime profile of earnings	41
7.3	Remedies	42
7.4	The causal effect of parental income; instruments	44
7.5	Matlab exercises	44
8	The distributional effect of changing prices – Equivalent variation	45
8.1	The consumer problem	46
8.2	Changing prices	47
8.3	Generalizing this result	48
8.4	Matlab exercises	49
9	Redistributive taxation – optimal tax theory	50
9.1	General principles	50
9.2	Linear income tax	52
9.3	Optimal top tax rate	54

Chapter 1

Preface

These are lecture notes for a course on “Empirical research on economic inequality.” The purpose of this class is twofold. First, to teach you about economic inequality, some of its causes, and how it is affected by policy. Second, to teach you econometric methods that have been used in the literature on economic inequality, which will help prepare you to conduct your own research on this or related topics, perhaps in an undergrad thesis.

These lecture notes are intended to accompany the reading of the original articles assigned for this class and listed in section 1.3, rather than serving as a stand-alone textbook. The purpose of these notes is to give you a compact overview of formal definitions and derivations and the econometric methods used, but often left implicit, in the papers discussed. The chapters of these lecture notes can be read in any order.

In this class, we will focus on mechanisms affecting income inequality, such as racial discrimination, (de)unionization, minimum wages, shifts in labor demand due to changes in technology and trade, shifts in labor supply due to migration, intergenerational transmission of economic status, and taxation. We will briefly talk about the historical evolution of income and wealth inequality, as well as about international inequality; mostly, however, we will focus on mechanisms affecting the distribution of incomes in the United States.

I would also like to emphasize what topics we will *not* cover in this class – particularly as these are arguably important topics:

- We will not talk about inequality along *non-economic dimensions*, such as health, education, political participation, or recognition. All of these are important, but I am not in a position to say much about them.
- We will mostly talk about inequality within the United States, rather than within *other countries* or *across countries*. Inequality within other countries matters too, of course, but the economics literature which we will discuss is focused on the United States. International inequality matters hugely – the citizenship you are assigned at birth is the single most important determinant of your life chances. International inequality is even more

complicated than domestic inequality and involves many additional considerations, so we will mostly focus on the “simpler questions” of within-country inequality.

- We will not talk about *indices* of economic inequality and their axiomatic justifications, which were central to an older literature on economic inequality, nor about issues of *measurement*.

1.1 Why research economic inequality?

A class on economic inequality raises the question: why should we care about inequality, and why should we do research on it? There are a number of distinct justifications which might be given.

First, the normative relevance of distributional questions follows from the fact that, in general, economists evaluate societal outcomes and the policies that affect them based on the **welfare of individuals**, however defined. Formally, if v_i is a measure of the welfare of individual i , social welfare evaluations are a function of all the v_i , $F(v_1, \dots, v_n)$. From this it follows that any such normative evaluation of the status quo has to start by evaluating who is doing how well. That is, how large is the welfare v_i of different individuals? Correspondingly, any normative evaluation of policy changes has to start by evaluating who wins and who loses, and by how much. That is, how much does the welfare v_i of different individuals increase or decrease? Such evaluations are the task of the kind of research considered in this class.

Any statement on whether a policy change is desirable must then take a stance on how to trade off the welfare of different individuals – e.g., how much do you care about an additional dollar for a rich person versus an additional dollar for a poor person?

Second, there is a long-standing line of ethical reasoning which suggests that we should put the bulk of normative weight on those who are worst off and aim for an equalization of welfare, which amounts to picking functions $F(v_1, \dots, v_n)$ of a particular form. Such prescriptions of **normative symmetry** or ethical equality go back as far as the biblical “golden rule;” similar prescriptions seem to appear in almost any religion. A contemporary philosophical version of such a reasoning can be found in John Rawls’s “Theory of justice” (Rawls, 1973). Rawls argues that we should evaluate societies by imagining that we are behind a “veil of ignorance,” which prevents us from knowing *who* we actually are. In such a setting of uncertainty regarding who we are, we should try to make the worst-off person as well off as possible, ensuring a minimum standard of living for ourselves should we happen to be among the worst-off. There has been much subsequent debate on this argument; a good reference is Sen (1995).

Third, we might be worried about the **consequences** of inequality, which

we might consider important in their own right. Various literatures in sociology, political science, and economics are concerned with these consequences; possible consequences that have been discussed are:

- Political consequences: An increasing concentration of income and wealth, and the rising influence of campaign donations and lobbying, might undermine democratic institutions predicated on the principle of “one person, one vote.”
- Social consequences: Increasing inequality might further social segregation (residential, educational, etc.), thus reducing knowledge of how others live and undermining social cohesion and solidarity.
- Economic consequences: Rising inequality might destabilize the economy. The increase in mortgage lending as a substitute for income growth of the bottom half of the distribution, for example, was at the origin of the financial crisis starting in 2008.

Whether rising inequality has these and other effects is a hard empirical question which we will not discuss in this class.

Fourth and finally, by studying economic inequality we learn (i) how much it has changed over time and across countries, and (ii) how much it is affected by policy decisions and other social factors. Recognizing these two facts **puts into question explanations** of inequality **that are a-historical and a-social**. These include biological explanations (such as biological racism, sexism, or justifications of inequality based on genetic differences in IQ) or explanations that reduce inequality to a matter of individual responsibility.

1.2 Acknowledgments

I thank Susanne Kimm and Ellora Derenoncourt for many helpful discussions and suggestions.

1.3 Readings

We will discuss the following articles in class. For each of these articles, the present lecture notes provide some technical and methodological background and summary.

1. Topic: The long run evolution of inequality as measured by top income shares
 Method: Pareto distribution, maximum likelihood, (interval) censored data
 Atkinson, A. B., Piketty, T., and Saez, E. (2011). Top incomes in the long run of history. *Journal of Economic Literature*, 49(1):3–71.
2. Topic: The long run evolution of gender inequality
 Method: Cohort analysis
 Goldin, C. (2006). The quiet revolution that transformed women’s employment, education, and family. *American Economic Review*, 96(2):1–21.
3. Topic: Racial discrimination
 Method: Potential outcomes, treatment effects, randomized experiments
 Bertrand, M. and Mullainathan, S. (2004). Are Emily and Greg More Employable Than Lakisha and Jamal? A Field Experiment on Labor Market Discrimination. *American Economic Review*, 94(4):991–1013.
4. Topic: The effect of de-unionization on inequality
 Method: Distributional decompositions, reweighting
 Fortin, N. M. and Lemieux, T. (1997). Institutional changes and rising wage inequality: Is there a linkage? *The Journal of Economic Perspectives*, 11(2):pp. 75–96.
 Firpo, S., Fortin, N., and Lemieux, T. (2011). Decomposition methods in economics. *Handbook of Labor Economics*, 4:1–102.
5. Topic: Labor demand and labor supply, technical change, immigration
 Method: Estimation of demand systems
 Card, D. (2009). Immigration and inequality. *The American Economic Review*, 99(2):1–21.
 Boustan, L. P. (2009). Competition in the promised land: Black migration and racial wage convergence in the north, 1940–1970. *The Journal of Economic History*, 69(03):755–782.
 Autor, D. H., Katz, L. F., and Kearney, M. S. (2008). Trends in US wage inequality: Revising the revisionists. *The Review of Economics and Statistics*, 90(2):300–323.
6. Topic: Intergenerational mobility
 Method: Measurement error

Chetty, R., Hendren, N., Kline, P., and Saez, E. (2014). Where is the land of opportunity? The geography of intergenerational mobility in the United States. *Quarterly Journal of Economics*, 129(4):1553–1623.

Black, S. and Devereux, P. (2011). Recent developments in intergenerational mobility. *Handbook of Labor Economics*, 4:1487–1541.

7. Topic: The welfare impact of changing prices and wages
Method: Equivalent variation, conditional causal effects

Deaton, A. (1989). Rice prices and income distribution in Thailand: a non-parametric analysis. *The Economic Journal*, pages 1–37.

Kasy, M. (2014). Who wins, who loses? tools for distributional policy evaluation. *working paper*.
8. Topic: Redistributive taxation
Method: Computing optimal income tax schedules

Saez, E. (2001). Using elasticities to derive optimal income tax rates. *The Review of Economic Studies*, 68(1):205–229.
9. Topic: International inequality
Method: Matching

Clemens, M. A., Montenegro, C. E., and Pritchett, L. (2009). The place premium: wage differences for identical workers across the US border.

Milanovic, B. (2015). Global Inequality of Opportunity: How Much of Our Income Is Determined by Where We Live? *The Review of Economics and Statistics*, 2(97):452–460
10. Topic: Policy options

Atkinson, A. (2014). After Piketty? *The British Journal of Sociology*, 65(4):619–638.

This class will involve some programming exercises, to be found at the end of each chapter in these lecture notes. We will be programming in Matlab. You do not need to know Matlab before this class, but it might help to look at some of the many available online resources before we get started, for instance:

1. http://www.mathworks.com/academia/student_center/tutorials/launchpad.html
2. <http://ocw.mit.edu/courses/electrical-engineering-and-computer-science/6-094-introduction-to-matlab-january-iap-2010/lecture-notes/>

If you are a student at Harvard, you can download Matlab for free from <http://downloads.fas.harvard.edu/download>. If you do not have access to Matlab, you can also use Octave which is open-source and which uses the exact same syntax as Matlab. Octave is available at <https://www.gnu.org/software/octave/>.

Recommended books

1. Normative theories of distributive justice:

- Rawls, J. (1973). *A theory of justice*. Harvard University Press, Cambridge
- Sen, A. (1995). *Inequality reexamined*. Oxford University Press, Oxford
- Roemer, J. E. (1998). *Theories of distributive justice*. Harvard University Press, Cambridge

2. Economists on the history of inequality:

- The long run evolution of wealth-inequality and its causes:
Piketty, T. (2014). *Capital in the 21st Century*. Harvard University Press, Cambridge
- Education, technology, and inequality:
Goldin, C. D. and Katz, L. F. (2009). *The race between education and technology*. Harvard University Press
- Global inequality of health and incomes:
Deaton, A. (2013). *The great escape: Health, wealth, and the origins of inequality*. Princeton University Press, Princeton
- Historical origins - the slave system
Fogel, R. W. (1994). *Without consent or contract: the rise and fall of American slavery*. WW Norton & Company
- Policy alternatives:
Atkinson, A. B. (2015). *Inequality: What Can be Done?*. Harvard University Press

3. Perspectives outside economics:

- The sociology of social classes:
Wright, E. O. (2005). *Approaches to class analysis*. Cambridge University Press, Cambridge
- Feminist perspectives:
Fraser, N. (2013). *Fortunes of Feminism: From State-Managed Capitalism to Neoliberal Crisis*. Verso Books

Chapter 2

The 1% – The Pareto distribution and maximum likelihood

Income inequality has been changing quite dramatically over time, in particular at the very top of the distribution, as illustrated by figure 2.1 on the next page, reproduced from Piketty’s “Capital in the 21st century.” How do we know this? In particular thanks to the careful historical work by various authors, reviewed in Atkinson et al. (2011), using tax data. In order to estimate top income shares, we need estimates of (i) how much income the rich received, and (ii) how large the total income generated in the economy was. In this chapter we shall be concerned with the question of how to get the first of these. Top incomes are estimated in this literature using historical tax data. The distribution of top incomes (and of top wealth holdings) is well approximated by the so-called Pareto distribution. The problem of estimating top incomes reduces to the problem of estimating the parameter α of this distribution.

2.1 Definition of the Pareto distribution

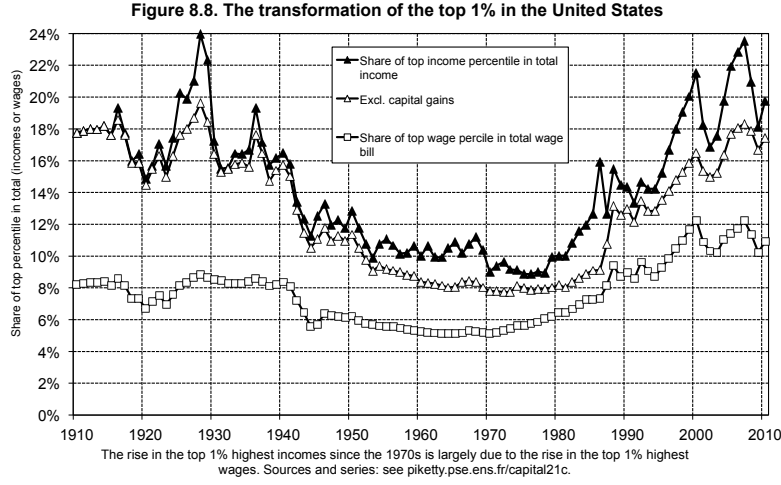
We shall suppose incomes Y above an income level of \underline{y} follow a Pareto distribution. This assumption was shown to provide a good approximation in many studies. The Pareto distribution is defined by the property that

$$P(Y > y | Y \geq \underline{y}) = (\underline{y}/y)^{\alpha_0} \quad (2.1)$$

for $y \geq \underline{y}$, where $\alpha_0 > 1$. That is, the **share of incomes above a cutoff** y declines with $y^{-\alpha_0}$. Our goal is to estimate the Pareto parameter α_0 , which we can then use to calculate top income shares.

We can calculate the **density** of the Pareto distribution, conditional on

Figure 2.1: Top 1% income share in the US. Reproduced from Piketty (2014).



$Y \geq \underline{y}$, by taking derivatives:

$$f(Y; \alpha_0) = -\frac{\partial}{\partial y} P(Y > y | Y \geq \underline{y}) = -\frac{\partial}{\partial y} (\underline{y}/y)^{\alpha_0} = \alpha_0 (\underline{y}/y)^{\alpha_0} \cdot y^{-1}. \quad (2.2)$$

The Pareto distribution has the interesting feature that, for any $y \geq \underline{y}$,

$$E[Y | Y \geq y] = \frac{\alpha_0}{\alpha_0 - 1} \cdot y. \quad (2.3)$$

Try to verify this by calculating the expectation by integration. This equation tells us that the **average income of those receiving more than y** , relative to y , equals $\alpha_0/(\alpha_0 - 1)$ – no matter what value y we pick! The smaller the parameter α_0 is, the larger are the incomes received by the very rich, and the larger will be our estimates of income inequality. Suppose we know the cutoff q^{99} such that 99% of incomes are below this cutoff. We can then calculate the **average income of the 1%** as

$$\bar{y}^{1\%} = \frac{\alpha_0}{\alpha_0 - 1} \cdot q^{99}.$$

You can find more information on the Pareto distribution at https://en.wikipedia.org/wiki/Pareto_distribution.

2.2 Maximum likelihood

Suppose you have i.i.d. observations y_1, \dots, y_n of incomes above \underline{y} from historical tax data. We want to estimate the parameter α using this data. One general method to construct estimators is to find the parameter which gives us the highest “probability” for finding the observations that we have; this idea is called maximum likelihood estimation (MLE). Formally, the **maximum likelihood estimator** is defined as

$$\hat{\alpha}^{MLE} = \operatorname{argmax}_{\alpha} \prod_{i=1}^n f(y_i; \alpha) = \operatorname{argmax}_{\alpha} \sum_{i=1}^n \log(f(y_i; \alpha)). \quad (2.4)$$

This is the value of α which maximizes the density of our observations. Note that between the second and third term we applied logs to everything – this is OK, since the logarithm is monotonically increasing and this therefore does not change the maximization problem.

The **first order condition** for the maximization problem defining the MLE (in log terms) is given by

$$\frac{\partial}{\partial \alpha} \sum_{i=1}^n \log(f(y_i; \alpha)) = 0.$$

We shall now plug in the expression for the density of the Pareto distribution which we derived before. The log likelihood of observation i , that is the log of its density given α , is equal to

$$\log(f(y_i; \alpha)) = \log(\alpha (\underline{y}/y_i)^\alpha \cdot y_i^{-1}) = \log(\alpha) + \alpha \log(\underline{y}/y_i) - \log(y_i).$$

We get

$$\begin{aligned} 0 &= \frac{\partial}{\partial \alpha} \sum_{i=1}^n \log(\alpha (\underline{y}/y_i)^\alpha \cdot y_i^{-1}) \\ &= \sum_{i=1}^n \left(\frac{1}{\alpha} + \log(\underline{y}/y_i) \right) \end{aligned}$$

Solving for α yields

$$\hat{\alpha}^{MLE} = \frac{n}{\sum_i \log(\underline{y}/y_i)}. \quad (2.5)$$

2.3 Censored data

Unfortunately we usually don’t observe the actual incomes that rich people received in the available tax data. All that is available from historical records is the number of tax filers that fall into several **tax brackets** of the form $[y_l, y_u]$. Fortunately we can still estimate the Pareto parameter from such data – which

makes the Pareto distribution a very useful model. The conditional probability that Y falls in the interval $[y_l, y_u]$ can be calculated as

$$\begin{aligned} P(Y \in [y_l, y_u] | Y \geq \underline{y}) &= P(Y > y_l | Y \geq \underline{y}) - P(Y > y_u | Y \geq \underline{y}) \\ &= (\underline{y}/y_l)^{\alpha_0} - (\underline{y}/y_u)^{\alpha_0}. \end{aligned} \quad (2.6)$$

For simplicity, suppose that we just observe two tax brackets, that is, we only have aggregate tax data which tell you the number N_1 of taxpayers falling in the bracket $[\underline{y}, y_l)$, and the number N_2 of tax payers falling in the bracket $[y_l, \infty)$. What is the distribution of N_2 conditional on $N_1 + N_2 = n$?

We can write N_2 as a sum of independent Bernoulli random variables,¹

$$N_2 = \sum_{i=1}^n \mathbf{1}(Y_i > y_l),$$

where the probability that any of these variables equals 1 is given by

$$p(\alpha_0) = P(Y > y_l | Y > \underline{y}) = (\underline{y}/y_l)^{\alpha_0}.$$

It follows that N_2 is **binomially distributed** conditional on $N_1 + N_2 = n$, that is

$$N_2 \sim \text{Ber}(n, p(\alpha_0)), \quad (2.7)$$

and

$$P(N_2 = n_2 | N_1 + N_2 = n; \alpha) = \binom{n}{n_2} \cdot p(\alpha)^{n_2} (1 - p(\alpha))^{n - n_2}.$$

You can find more information on the Binomial distribution at https://en.wikipedia.org/wiki/Binomial_distribution.

2.4 Maximum likelihood with censored data

As before, we want to construct an estimator of α_0 , but using only the interval censored data. As before, we can construct such an estimator using maximum likelihood, that is

$$\hat{\alpha}^{MLE} = \underset{\alpha}{\operatorname{argmax}} P(N_2 = n_2 | N_1 + N_2 = n; \alpha)$$

is the value which maximizes the probability of observing a number N_2 of observations in the upper tax bracket.

¹Bernoulli random variables are random variables that only take on the values 0 and 1.

The **first order condition** for the MLE is given by

$$\begin{aligned}
 0 &= \frac{\partial}{\partial \alpha} \log f(N_2 | N_1 + N_2; \alpha) \\
 &= \frac{\partial}{\partial \alpha} \log \binom{n}{N_2} + N_2 \log p + N_1 \log(1 - p) \\
 &= \left(\frac{N_2}{p} - \frac{N_1}{1 - p} \right) \cdot \frac{\partial}{\partial \alpha} p \\
 &= \left(\frac{N_2}{(\underline{y}/y_l)^\alpha} - \frac{N_1}{1 - (\underline{y}/y_l)^\alpha} \right) \cdot \frac{\partial}{\partial \alpha} p.
 \end{aligned}$$

Since $\frac{\partial}{\partial \alpha} p \neq 0$, the first term has to vanish, and after doing some algebra we see that $\hat{\alpha}^{MLE}$ is the solution to the equation

$$(\underline{y}/y_l)^{\hat{\alpha}^{MLE}} = \frac{N_2}{n},$$

so that

$$\hat{\alpha}^{MLE} = \frac{\log(N_2/n)}{\log(\underline{y}/y_l)}. \quad (2.8)$$

The larger the share of our observations is that falls in the upper tax bracket, the larger is our estimate of α .

2.5 Piketty's $r - g$ and the Pareto parameter

So far, we have discussed how to estimate α and how to use this estimate to calculate top income shares. But why do the top tails of income and wealth follow a Pareto distribution, and what determines the parameter α ? One possible story is given by the formal argument underlying Piketty's book (though hidden very deeply in the references), which relates the rate of return r on capital, relative to economic growth g , to the long run inequality of wealth.

I will give a heuristic proof of this argument; this is complicated and optional material. Suppose that the **wealth** Y of a family i **follows the process**

$$Y_{i,1} = w_i + R_i \cdot Y_{i,0} \quad (2.9)$$

over time, where $Y_{i,1}$ denotes the wealth of children, $Y_{i,0}$ the wealth of parents, w_i reflects savings from earnings, and R_i is the rate of savings from capital income, corresponding to Piketty's $r - g$. Suppose further that

$$(w_i, R_i) \perp Y_{i,0},$$

that is random shocks to earnings, rates of return, or savings are independent from past wealth. This is a strong assumption, which, however, could be relaxed.

A so-called **stationary distribution** for Y_i is one where the distribution of $Y_{i,1}$ is the same as that for $Y_{i,0}$ – inequality among children is the same

as inequality among parents. Under fairly general conditions, this process will **converge** to a stationary distribution with Pareto tail as long as $E[R] < 1$ (otherwise inequality would explode over time), so that we can assume that the distribution of Y_0 is already approximately Pareto in the tail of the distribution, that is for large values of wealth,

$$P(Y_0 > y | Y_0 \geq \underline{y}) \approx (\underline{y}/y)^{\alpha_0}.$$

What Pareto parameter yields a stationary tail? In the tail, w_i (savings from earnings) is negligibly small relative to $R_i \cdot Y_{i,0}$ (savings from capital income), so that **stationarity requires**

$$\begin{aligned} P(Y_0 > y) &= P(Y_1 > y) = P(w + R \cdot Y_0 > y) \\ &\approx P(Y_0 > y/R) = E[P(Y_0 > y/R | R)] \end{aligned}$$

for large y , where in the second line we just dropped w from the exact equality. In the last expression we first condition on R , and then average out over the distribution of R . Plugging the Pareto distribution for Y_0 into these expressions we get

$$(\underline{y}/y)^{\alpha_0} = E[(\underline{y}/(y/R))^{\alpha_0}].$$

Dividing by $(\underline{y}/y)^{\alpha_0}$ shows that this is equivalent to

$$E[R^{\alpha_0}] = 1. \tag{2.10}$$

We have derived the equation mapping the distribution of R to the Pareto parameter α_0 .

The **intuition** behind our argument and this equation is as follows: rich families move up the wealth distribution if their $R_i > 1$, they move down if their $R_i < 1$. Stationarity requires that upward- and downward movements cancel each other – as many families move down as up in any given range of the tail. If it's more likely to move down than up (R_i is mostly small), then there have to be fewer people that are very rich rather than just rich (large α , little inequality), for these movements to cancel. If it's equally likely to move up as down (R_i is centered close to 1), there have to be almost as many very rich people as rich people (small α , lots of inequality).

2.6 Matlab exercises

Write code that performs the following:

1. Generate n independent draws from the Pareto distribution with parameters \underline{y} and α

Hint: You can take $Y_i = \underline{y} \cdot U_i^{-1/\alpha}$ for U uniformly $[0, 1]$ distributed. Why?

2. Save these data to a .csv file, and exchange your file with a classmate.
3. Use your classmate's data to estimate α , using the formula in equation (2.5).
4. Now generate new data using the same procedure as before, and just tell your classmate the value \underline{y} , as well as the number of observation below / above the cutoff $2 \cdot \underline{y}$. Ask her/him to provide an estimate of α based on these numbers, using equation (2.8).
5. Now we are going to verify the argument of section 2.5 by simulations. Generate data following the process in equation (2.9), that is

$$Y_{t+1} = w_t + R_t \cdot Y_t,$$

where w_t and R_t are independent draws from uniform distributions with boundary values that you pick. Generate 10.000 observations, and only keep the last 2.000. Save them, and give them to a classmate.

6. Sort the data you got from your classmate, and only keep the top 200. Use these observations to estimate the Pareto parameter as in step 2.
7. Repeat the last two steps, but for a different distribution of R_t . Does the estimated Pareto parameter change in the way that you would expect?

Matlab commands which you might find useful:

```
rand
csvwrite, csvread
sort
```


Chapter 3

Gender Inequality – Elasticities of Labor Supply

The relative economic position of women and men is still quite unequal, yet has undergone great changes over the course of the last century. Economic inequality between women and men has many dimensions, including the following. There is, first, inequality of pay for the same occupation, maybe due to discrimination.² There are, second, differences in the distribution of men and women across occupations, maybe due to social norms and aspirations and due to the workings of the educational system. And there are unequal intra-household divisions of labor. Traditional divisions of roles would often require women to be primarily in charge of unpaid reproductive labor – housekeeping, taking care of children, the elderly, and the sick, etc. – while men would be in charge of paid work in the labor market.

While this description of a traditional model might be correct in some “typical” sense, there are great differences across social classes and over time, shaped by market forces, social provisions by the state, changing social norms, and other factors. These differences and this historical evolution are the subject of Goldin (2006). Claudia Goldin focuses on the changing prevalence of women’s participation in the labor market, and the changing career trajectories of women. She structures her historical description in terms of two key elasticities of women’s labor supply, the income elasticity and the substitution elasticity.

3.1 Elasticities of labor supply

Economists like to express causal effects and other relationships in terms of elasticities. Elasticities are **unit-less** magnitudes. Suppose L is a function of Y . Elasticities answer questions such as “By what **percentage** does L **increase**

²We will discuss discrimination in the next chapter.

(or decrease) if Y increases by 1%.” The elasticity ϵ is formally defined as

$$\epsilon = \frac{\partial \log L}{\partial \log Y} = \frac{\partial L}{\partial Y} \cdot \frac{Y}{L}.$$

Let L be the labor supply of a woman, that is the amount of hours or weeks of paid labor. Let w be the hourly wage that she receives or would receive in the market, and let Y be household income, which includes partners’ earnings and unearned income (from capital ownership and other sources).

Wages w have greatly increased over time, and vary across social classes / levels of education. These changes of wages are an important explanatory factor for changing patterns of women’s participation in wage labor. We can think of the effect of a change in wages on labor supply as being composed of two parts.

The first part is due to an **effect of incomes**. As households get richer, they are able to afford more. To the extent that it is considered desirable that (married) women do not work for wages, as per traditional role models (as prevalent early in the 20th century in the United States), an increase in incomes might lead to a decrease in women’s labor force participation, both across social classes and over time. Richer households can afford women’s staying home more easily. The second part is due to an **effect of relative prices**. When wages are higher, then the return to paid work relative to unpaid work is higher, creating an incentive to switch from the latter to the former. An increase in wages would suggest higher labor force participation for women, leading to increased labor force participation over time, going in the opposite direction of the income effect.

This decomposition into two parts can be made formal using two elasticities. The first is the **income elasticity**. It measures the percentage change of labor supply for a 1% change of (household) income,

$$\epsilon = \frac{\partial \log L}{\partial \log Y}.$$

Total household income Y depends not only on own-earnings, but also on partners’ earnings. If we assume that household decisions are made jointly then both these sources of income affect women’s labor supply decisions in the same manner. If that is the case then we can learn about ϵ by looking at the effect of partners’ earnings on L , since partner’s earnings do not themselves affect the incentives (relative prices) of women’s labor supply.

The second effect is measured by the **substitution elasticity** η^s . We can only measure η^s indirectly. η^s is the effect of women’s wages on their labor supply, after we subtracted the income effect. Put differently, increasing women’s wages w has a total effect η , which is the sum of substitution and income effect:

$$\eta = \frac{\partial \log L}{\partial \log w} = \frac{\partial \log L}{\partial \log Y} \frac{\partial \log Y}{\partial \log w} + \eta^s \quad (3.1)$$

To calculate the effect α of wages on household income,

$$\alpha = \frac{\partial \log Y}{\partial \log w}, \quad (3.2)$$

we just need to do some accounting, which we can do once we know household income and women's earnings. A 1% increase in w leads to a 1% increase of earnings wL , and we get

$$\alpha = \frac{wL}{Y}.$$

We can finally define η^s by

$$\eta^s = \eta - \alpha \cdot \epsilon. \quad (3.3)$$

3.2 Decomposing changes

Goldin (2006) uses this decomposition to make sense of changing patterns in women's labor force participation since the late 19th century. For any given year, we can attempt to learn about ϵ , the income elasticity, by comparing households with different partners' earnings, but similar wages for women. Formally, we might regress $\log L$ on $\log Y$, controlling for $\log w$:

$$\log L_i = \beta_0 + \epsilon \cdot \log Y_i + \beta_1 \cdot \log w_i + U_i. \quad (3.4)$$

Variation in Y_i given w_i comes from partners' earnings. We can learn about η , the total elasticity of labor force participation with respect to wages, by regressing $\log L$ on $\log W$,

$$\log L_i = \gamma_0 + \eta \cdot \log w_i + V_i. \quad (3.5)$$

There are issues in estimating these elasticities due to possible endogeneity, that is due to correlation between (Y_i, w_i) and (U_i, V_i) . We will ignore these for now, and assume that we got correct estimates of ϵ and η . We can finally learn about α by simply calculating how much women would earn when working full-time.

Plugging in the decomposition of η into equation (3.5), we get

$$\log L_i = \gamma_0 + (\eta^s + \alpha \cdot \epsilon) \cdot \log w_i + V_i. \quad (3.6)$$

This equation allows us to interpret changes of labor supply over time and differences across social classes. First, labor supply might increase over time as γ_0 increases. This is an outward shift of women's labor supply, which implies that women would work more for any given wage level. Second, labor supply might shift over time as wages w_i increase. This effect might go either way, depending on the sign of $\eta = \eta^s + \alpha\epsilon$. If negative income effects, $\epsilon < 0$, dominate, then $\eta < 0$, and increasing wages lead to a reduction of labor supply. If substitution effects, $\eta^s > 0$, dominate, then $\eta > 0$ and increasing wages lead to an increase of labor supply. Third, Labor supply might shift because of a change of these elasticities. According to Goldin (2006), the total elasticity η used to be negative but increased to become positive at some point in history, as income elasticities increased (became less negative) and substitution elasticities increased (became more positive). Possible explanations for these changing elasticities include changes in the workplace, as new technologies and occupations became available (office work as opposed to factory work), changes in the

legal environment (married women used to be barred from many occupations), changes in household technology (laundromats), public provision of care (public kindergartens; at least in European welfare states), and changes in social norms.

The net effect of these changes was an increase of labor force participation of working-age women from below 20% in the US around 1900 (and in fact close to zero for married white women) to almost 80% today. This increase was initially driven by an outward shift in labor supply (increase in γ), followed by increases in η and subsequently a positive effect of increasing wages. More recent important changes since the 1980s are not quite captured by labor force participation. In tandem with women overtaking men in educational attainment, the nature of women's occupations, in particular for well-off women, has changed from being merely a source of supplementary income with little advancement over time to involving long-term careers.

3.3 Some critical remarks

Celebratory descriptions of this changing role of women in the labor market encounter some criticism in particular in feminist discussions, see for instance Fraser (2013). We shall briefly mention two.

First, descriptions such as the one of the “quiet revolution” of women's careers since the 1980s in Goldin (2006) focus on college-educated women. Additional emphasis is put on those with professional and advanced degrees (lawyers, doctors, managers, academics...). These descriptions neglect the quite different historical changes for women at the low end of the wage distribution (in service and care occupations, in particular) in recent decades, facing stagnating low wages in a time of eroding social provisions. In addition to the focus on privileged women among all women, the consideration of labor supply differences by gender alone also neglects important heterogeneity. If one cares about inequality in general, then a focus on inequality solely along the dimension of gender might obscure other inequalities.

Second, the massive increase of women's participation in paid labor is the flip-side of an increased marketization of all spheres of life, including social spheres such as care of children and the elderly traditionally outside the reach of markets. Fraser (2013) argues that we should aim for a third alternative beyond (i) a traditional division of roles with women in charge of unpaid care-work and dependent on men's wage incomes, but also beyond (ii) a complete marketization of all spheres of life with its consequences for inequality, uncertainty, and erosion of social bonds. Such a third alternative would involve an equal role of men and women in care work organized in ways outside the anonymous market.

Chapter 4

Discrimination – Experiments

Economic chances are very unequally distributed along dimensions such as race and gender. Why is this so? There are many channels through which such inequalities might be created. These include early childhood influences, different neighborhoods of growing up, different access to and quality of primary, middle, and high school education, the creation of aspirations, different access to and treatment in higher education, different chances of being hired when applying for a job, different wages conditional on being hired, different chances of being promoted or fired in a given job, differential treatment by customers or clients, etc.

The channel of *hiring* might in turn be decomposed into several components. What is the chance of being invited to an interview, and what is the chance of being hired given an interview? How does the chance of being invited to an interview depend on the neighborhood of residence, the high school attended, or the (perceived) race and gender of an applicant? It is this very last question that the paper by Bertrand and Mullainathan (2004), which we discuss next, addresses. How does the chance of being invited to an interview depend on perceived race, for otherwise identical CVs? You should keep in mind that this is only one of many channels through which discrimination might affect labor market outcomes.

This paper gives us occasion to review potential outcomes, causality, and randomized experiments in the way they are conceptualized by empirical economists today. This framework is useful for making precise (i) what we mean by race “causing” lower/higher chances of being offered an interview, and (ii) how we can learn about this causal effect.

4.1 Potential outcomes and causal effects

Consider a **treatment** D , which can take one of two values, $D = 0$ or $D = 1$.³ In our application, D would be the implied race of the name on a given CV. Denote by Y_i the **outcome** of interest for CV i . In our application, this would be whether a CV received a call to be invited for a job interview. In order to talk about causality, we use the notion of **potential outcomes**. Potential outcomes provide the answer to “what if” questions:

- Potential outcome Y_i^0 :
Would CV i have received a callback if the race implied by the name on it were 0?
- Potential outcome Y_i^1 :
Would CV i have received a callback if the race implied by the name on it were 1?

Observed outcomes are determined by the equation

$$Y = D \cdot Y^1 + (1 - D) \cdot Y^0. \quad (4.1)$$

It is not obvious at this point that potential outcomes are an empirically meaningful idea. As we will see in section 4.2 below they are meaningful once we introduce the notion of a controlled experiment.

With the notion of potential outcomes at hand, we can define the causal effect or **treatment effect** for CV i as $Y_i^1 - Y_i^0$. Correspondingly, we can define the average causal effect or **average treatment effect**,

$$ATE = E[Y^1 - Y^0], \quad (4.2)$$

which averages the causal effect over the population of interest.

Given this formalism, we can also state **the fundamental problem of causal inference**:

We never observe both Y^0 and Y^1 at the same time!

One of the potential outcomes is always missing from the data. Which treatment D was assigned determines which of the two potential outcomes we observe (recall that $Y = D \cdot Y^1 + (1 - D) \cdot Y^0$).

Closely related is the **selection problem**: Simply comparing the average outcomes of those who got $D = 1$ and those who got $D = 0$ in general tells us nothing about causal effects. The reason is that the distribution of Y^1 among

³The language used to talk about causality by applied economists these days has its roots in biostatistics and medical trials, where $D = 0/1$ corresponds to placebo/actual treatment, hence the terminology of “treatments” and “treatment effects.”

those with $D = 1$ need not be the same as the distribution of Y^1 among everyone, similarly for Y^0 . It might, for instance, be the case that the CVs with “black” names have higher educational qualifications on average than those with “white” names, so that their chances of receiving a callback are higher no matter what the name on the CV is. Making the same point formally, we get that in general

$$\begin{aligned} E[Y|D = 1] &= E[Y^1|D = 1] \neq E[Y^1] \\ E[Y|D = 0] &= E[Y^0|D = 0] \neq E[Y^0] \\ E[Y|D = 1] - E[Y|D = 0] &\neq E[Y^1 - Y^0] = ATE. \end{aligned} \quad (4.3)$$

4.2 (Randomized) controlled experiments

The selection problem arises because potential outcomes and treatment are not statistically independent. There is one way to ensure that they actually are: by assigning treatment in a controlled way in an experiment, possibly using randomization. This guarantees that there is no selection, i.e.

$$(Y^0, Y^1) \perp D.$$

In this case, the selection problem is solved and

$$\begin{aligned} E[Y|D = 1] &= E[Y^1|D = 1] = E[Y^1] \\ E[Y|D = 0] &= E[Y^0|D = 0] = E[Y^0] \\ E[Y|D = 1] - E[Y|D = 0] &= E[Y^1 - Y^0] = ATE. \end{aligned} \quad (4.4)$$

The statistical independence ensures that, when comparing averages for the treatment and control group, we actually compare “apples with apples.” Note how this idea of controlled experiments gives **empirical content** to the “meta-physical” notion of **potential outcomes**!

In Bertrand and Mullainathan (2004), for instance, statistically “white” or “black” names were randomly assigned to given resumes which were sent out as job applications. This allows one to estimate the causal effect of race on the likelihood of getting invited to a job interview by simply comparing means. They actually used a design that was slightly more complicated than simple randomization: for each job-opening they submitted two (or four) randomly chosen CVs, and out of those one (or two) were randomly assigned a “black” / “white” name.

4.3 Estimation and t-tests

So far we have talked about expectations, that is population averages, for the treatment and control groups. We can easily construct estimators by replacing

expectations with sample averages in equation (4.4). Consider a randomized trial with N individuals. Suppose that the estimand of interest is ATE:

$$ATE = E[Y_1 - Y_0] = E[Y|D = 1] - E[Y|D = 0].$$

Replacing the conditional expectation $E[Y|D = 1]$ by its sample analog, the conditional mean \bar{Y}_1 , and similarly for $E[Y|D = 0]$, we construct an estimator:

$$\hat{\alpha} = \bar{Y}_1 - \bar{Y}_0,$$

where

$$\begin{aligned}\bar{Y}_1 &= \frac{\sum Y_i \cdot D_i}{\sum D_i} = \frac{1}{N_1} \sum_{D_i=1} Y_i \\ \bar{Y}_0 &= \frac{\sum Y_i \cdot (1 - D_i)}{\sum (1 - D_i)} = \frac{1}{N_0} \sum_{D_i=0} Y_i\end{aligned}\tag{4.5}$$

with $N_1 = \sum_i D_i$ and $N_0 = N - N_1$. As you can easily show, $\hat{\alpha}$ is an unbiased estimator of the ATE,

$$E[\hat{\alpha}] = ATE.$$

We not only want to get a point-estimate of the average treatment effect, we also want to calculate a range of likely values, to assess whether our estimates are just the result of chance or reflect some true causal effects. This can be done using the t -statistic, which is defined as

$$t = \frac{\hat{\alpha} - \alpha_{ATE}}{\hat{\sigma}_\alpha},\tag{4.6}$$

where

$$\hat{\sigma}_\alpha = \sqrt{\frac{\hat{\sigma}_1^2}{N_1} + \frac{\hat{\sigma}_0^2}{N_0}}$$

and

$$\hat{\sigma}_1^2 = \frac{1}{N_1 - 1} \sum_{D_i=1} (Y_i - \bar{Y}_1)^2.$$

$\hat{\sigma}_0^2$ is analogously defined. The t -statistic is approximately standard normal distributed (for samples of a reasonable size),

$$t \sim^{approx} N(0, 1).$$

We get a range of plausible values – a 95% confidence interval – by calculating the interval

$$CI = [\hat{\alpha} - 1.96 \cdot \hat{\sigma}_\alpha, \hat{\alpha} + 1.96 \cdot \hat{\sigma}_\alpha].\tag{4.7}$$

As an exercise, try to show that

$$P(\alpha \in CI) \approx 0.95.$$

Note that in this expression α is fixed, while CI is random!

4.4 Matlab exercises

Write code which performs the following:

1. Generate n pairs of potential outcomes Y_i^1, Y_i^0 which are just independent draws from the standard normal distribution. What is the ATE for this data generating process?
2. Generate $D = \mathbf{1}(Y_i^1 > Y_i^0)$, and the corresponding Y based on equation (6.4). Calculate $\bar{Y}_1 - \bar{Y}_0$
3. Repeat 2, but with $D = \mathbf{1}(Y_i^1 \leq Y_i^0)$.
4. Repeat 2, but with D independent Bernoulli 0.5 draws.
5. Using the data from 4, calculate a 0.95 confidence interval for the ATE .

Matlab commands which you might find useful:

`randn`

Chapter 5

Unions – Distributional decompositions

Suppose we are interested in the effect of declining unionization on the distribution of wages. One might be tempted to simply compare the distribution of wages of union- and non-union members in order to learn about this effect. This is problematic, however, since these two groups might be quite different in terms of their distribution of age, education, gender, ethnicity, sector of the economy, state of residence, etc. A better approach would thus compare people who look similar along all these dimensions and differ only in terms of their union membership. This is the basic idea behind distributional decompositions, as pioneered by DiNardo et al. (1996), underlying the discussion in Fortin and Lemieux (1997), and reviewed in Firpo et al. (2011). Distributional decompositions provide the answer to hypothetical questions such as the following: what if (i) the distribution of demographic covariates (age, gender,...) had stayed the same, (ii) the distribution of wages given demographics *and* union membership status had stayed the same, but (iii) we consider actual historical changes of union membership for different demographic groups – how, in this hypothetical scenario, would the distribution of wages have changed? Intuitively, such distributional decompositions provide an answer to the question: To what extent is de-unionization responsible for the rise in inequality?

5.1 Setup

Suppose we observe **repeated cross-sections** with i.i.d. draws from the time t distributions P^t of the variables (Y, D, X) . Here X denotes **covariates** such as age, education, and location. As in chapter 4, D is a binary “**treatment**” variable such as union membership. The variable Y denotes an **outcome** such as real income.

We are interested in isolating the effect of historical changes in the prevalence of union membership D on the distribution $P(Y)$ of incomes Y , and in the effect

of these historical changes on **statistics of the income distribution**, $\nu(P(Y))$. Possible choices for the statistics ν include the mean, the variance, the share below the poverty line, quantiles or the Gini coefficient.

Let $P^1(Y, D, X)$ in particular denote the joint distribution of (Y, D, X) in **period 1** (the year at the end of the historical period which we are considering), and $P^0(Y, D, X)$ the corresponding distribution in **period 0** (the year at the beginning of the historical period). Our goal is to identify the **counterfactual** distribution P^* of Y in which the **effect of changing D is “undone,”** while holding constant the current (period 1) distribution of covariates X as well as the distribution of income Y given X and D . The change from P^* to P^1 will be interpreted as the causal effect of changing D on the income distribution. Formally, define P^* as

$$P^*(Y \leq y) := \int_{X,D} P^1(Y \leq y|X, D) P^0(D|X) P^1(X) dD dX. \quad (5.1)$$

This expression asks us to consider the following scenario: take the share $P^1(X)$ in the population of period 1 of 40 year old women with a high school degree living in the Midwest as given, and similarly for all other demographic groups. For each of these groups, however, suppose that their prevalence of union membership was the same as in period 0, $P^0(D|X)$. Consider finally the distribution of incomes for each demographic group and each union-membership status of period 1, again, $P^1(Y \leq y|X, D)$. Putting all groups together (formally: integrating out the distributions of X and D), we get the counterfactual income distribution $P^*(Y \leq y)$.

5.2 Reweighting

We can rewrite the distribution P^* as defined in equation (5.1) in a useful way as follows. First, multiply and divide the integrand by $P^1(D|X)$. Second, rewrite the probability $P^1(Y \leq y|X, D)$ as an expectation $E^1[\mathbf{1}(Y \leq y)|X, D]$. Third, give the fraction $P^0(D|X)/P^1(D|X)$ a new name, $\theta(D, X)$. Finally, pull θ into the conditional expectation, and use the “law of iterated expectations” to get an unconditional expectation. Executing these steps yields

$$\begin{aligned} P^*(Y \leq y) &= \int_{X,D} P^1(Y \leq y|X, D) \frac{P^0(D|X)}{P^1(D|X)} P^1(D|X) P^1(X) dD dX \\ &= \int_{X,D} E^1[\mathbf{1}(Y \leq y)|X, D] \theta(D, X) P^1(D|X) P^1(X) dD dX \\ &= E^1[E^1[\mathbf{1}(Y \leq y) \cdot \theta(D, X)|X, D]] \\ &= E^1[\mathbf{1}(Y \leq y) \cdot \theta(D, X)], \end{aligned} \quad (5.2)$$

where

$$\theta(D, X) := \frac{P^0(D|X)}{P^1(D|X)}. \quad (5.3)$$

In the exercises, you will be asked to do these calculations in some simple examples, to see that nothing very complicated is going on. Equation 5.2 states that P^* is a **reweighted version of the current distribution**, P^1 . Any counterfactual distributional characteristic ν of P^* can be estimated based on estimates of P^* . Estimating P^* requires estimation of the ratio in equation (5.3).

Consider an individual who is a union member, $D = 1$, and has covariate values $X = x$. Suppose for this value of X it was more likely to be a union member in period 1 than in period 0. This implies that $\theta(X, D) > 1$ for this person – we should *upweight* that person’s income to get the counterfactual income distribution P^* , where union membership probabilities are assumed not to have changed over time. Consider another individual, who is also a union member, of a demographic group $X = x$ for which union membership did increase over time. For this individual, equation 5.2 tells us to *downweight* her income to get the counterfactual distribution.

5.3 Causal interpretation

Our definition and discussion of counterfactual distributions was purely statistical, and did not make any reference to potential outcomes or causality. The counterfactual distribution P^* can however be interpreted causally under an assumption of conditional independence. Denote Y^d the potential outcome (e.g. wage, or income) of a person with treatment status $D = d$; this is just the same kind of potential outcome which we encountered in chapter 4. In chapter 4 we showed that differences in means can be interpreted as average treatment effects if treatment is independent of potential outcomes. Something slightly more complicated works in the present context. Assume that treatment and potential outcomes are independent *conditional* on X , that is

$$Y^d \perp D | X \quad \forall d. \quad (5.4)$$

Under this assumption

$$P(Y | D = d, X = x) = P(Y^d | X = x), \quad (5.5)$$

and the conditional distribution $P(Y | D = d, X = x)$ does not change as the probability of treatment assignment $P(D = d | X)$ is changed, for instance through de-unionization. Under the **conditional independence assumption**, we can therefore **interpret P^* as the counterfactual** that would actually prevail if union-membership probabilities $P(D = d | X)$ had not changed over time.

5.4 Estimation

Implementing an estimator of $\nu(P^*)$ based on equation (5.2) involves two steps. First, we need to estimate the weight function θ . Second, we need to calculate ν for the distribution P^1 reweighted by the estimated θ .

In order to estimate θ , we need to estimate the ratio between $P^1(D|X)$ and $P^0(D|X)$, corresponding to the change in the prevalence of D within demographic groups defined by X . Suppose first that X takes on only finitely many values x . Then we can directly estimate these conditional probabilities by the corresponding population shares in our sample,

$$\hat{\theta} = \frac{P_n^0(D|X)}{P_n^1(D|X)} = \frac{P_n^0(D, X)}{P_n^1(D, X)} \cdot \frac{P_n^1(X)}{P_n^0(X)}, \quad (5.6)$$

where we use the subscript n to denote sample shares.

If X takes on many values, or has continuous components, we can not do this anymore. In that case, however, we can use a logit model for the distribution of D given X , with parameters β^t changing over time:

$$P^t(D = 1|X; \beta^t) = \frac{\exp(X \cdot \beta^t)}{1 + \exp(X \cdot \beta^t)}. \quad (5.7)$$

Based on estimates of the parameters β^t , we can estimate the weights θ by

$$\hat{\theta} = \frac{\hat{P}^0(D|X)}{\hat{P}^1(D|X)} = \frac{\exp(X \cdot \hat{\beta}^0)}{1 + \exp(X \cdot \hat{\beta}^0)} \cdot \frac{1 + \exp(X \cdot \hat{\beta}^1)}{\exp(X \cdot \hat{\beta}^1)}. \quad (5.8)$$

The parameters β can be estimated using maximum likelihood, similarly to the estimators for the Pareto parameter that we came up with in chapter 2:

$$\hat{\beta}^t = \operatorname{argmax}_{\beta} \prod_{i=1}^{n_t} P^t(D_i|X_i; \beta) = \operatorname{argmax}_{\beta} \prod_{i=1}^{n_t} \frac{\exp(D_i \cdot X_i \cdot \beta)}{1 + \exp(X_i \cdot \beta)}, \quad (5.9)$$

where the product is taken over all time t observations. Implementations of this logit estimator are readily available in most statistical software packages.

5.5 Matlab exercises

1. Suppose X can take on 4 values, depending on an individual's gender and on whether he/she graduated from college. Suppose you have matrices A^1 and A^0 completely describing $P^1(X, D)$ and $P^0(X, D)$, as well as matrices B^1 and B^0 describing $E^t[Y|X, D]$. What is the dimension of these matrices? Pick some numbers for these matrices A^t and B^t (be careful to make sure that probabilities add up to 1).
2. Write a Matlab script that calculates θ and $E^*[Y]$ based on the matrices A^t and B^t . Where possible, do calculations by matrix multiplication rather than using loops.
3. Next, generate random samples from the distributions $P^t(Y, D, X)$ under the same assumptions, that is with distributions of X and D determined

by some matrices A and B that you picked, and for values Y drawn from the distributions

$$Y|X, D \sim N(E^t[Y|X, D], 1).$$

Export a dataset containing draws of the variables t, Y, X, D , and give it to a classmate.

4. Use the data you got from your classmate to estimate $E^*[Y]$, building on the code you wrote for 1. Compare your estimate to the number they got in step one based on the true matrices A and B .

Chapter 6

Migration, technology, education – Estimating labor demand

This section discusses the approach taken by the empirical literature on labor demand. Applications which we will discuss in class include Card (2009) on the impact of international migration, Boustan (2009) on domestic migration and racial inequality, and Autor et al. (2008) on the impact of technical change.

6.1 Backwards-engineering wage regressions

This literature aims to explore the impact of the relative labor supply of different groups on relative wages. Papers in this literature estimate regressions of the form

$$\log \left(\frac{w_j}{w_{j'}} \right) = \text{controls} + \beta \cdot \log \left(\frac{N_j}{N_{j'}} \right) + \epsilon_{j,j'}, \quad (6.1)$$

where j and j' are different “types” of labor, w denotes wages, N denotes labor supply, the controls account for some factors other than labor supply (including trends in technology), β is interpreted as the inverse of the elasticity of substitution, and ϵ captures all other (unobserved) factors affecting relative wages. The models invoked in this literature are justifications of this regression.

We will start with a very general model, and get increasingly specific, until we end up with a model rationalizing such regressions. This approach allows us to discuss the assumptions invoked along the way, as well as their implications. We begin with a **general demand function**, mapping labor supply of various groups, in conjunction with unobserved factors, into the wages of these groups. We then consider the **neoclassical theory of wage determination**, which assumes that wages correspond to marginal productivities with respect to some aggregate production function. We finally consider a set of specific parametric

production functions, including the “**CES-production function**” and some of its variants.

6.2 General labor demand

Suppose there are $j = 1, \dots, J$ types of labor. As for “types,” think in particular of the level of education; however, types might also depend on age, gender and country of origin. For each type j , denote by N_j the number of people of that type which are employed in a given labor market. We might think of a labor market as a city, as a state, or as a nation. Denote by

$$N = (N_1, \dots, N_J)$$

the labor supply of each type in this labor market, and by

$$w = (w_1, \dots, w_J)$$

the (average) wage of each type.

As N changes, whether through immigration, demographic shifts, or education, this has consequences for wages. We can denote the **counterfactual wages** that would prevail if labor supply were equal to N (holding all else equal) by

$$w = w(N, \epsilon). \quad (6.2)$$

Here “all else” is captured by ϵ , denoting all other factors influencing wages besides labor supply.

We would like to learn about the function w , so that we can tell to what extent historical changes in N are responsible for changes in wage inequality. If changes in **labor supply** are **random**, that is independent of ϵ , this is in principle possible by regressing w on N . The problem is that we usually have only a few observations, but potentially many variables (N_1, \dots, N_J) on which to regress, which makes it hard to get precise estimates. The literature therefore imposes restrictions on the function $w(N, \epsilon)$, which are justified by theoretical models.

If changes in N are not random, which is likely the case if the labor markets considered are cities, then we additionally need to find valid **instruments** for N . We will get back to this point below.

6.3 Demand based on a production function

How are wages determined? Neoclassical theory assumes that there is an **aggregate production function**

$$y = f(N_1, \dots, N_J), \quad (6.3)$$

which determines the amount of output y (in US\$ terms) that can be achieved for a given level of inputs of different types of labor. Implicit in this formulation is that the supply of capital and demand for products have already been “concentrated out.”⁴

Neoclassical theory additionally assumes that wages are determined by the **marginal productivity** of different types of labor, that is by the amount output would be increased by increasing inputs by one unit:

$$w_j = \frac{\partial f(N_1, \dots, N_J)}{\partial N_j} \quad (6.4)$$

This theory is justified by assuming that employers are profit maximizing, and that labor markets clear.

There are many **reasons to be skeptical** about this theory:

- What if effort or the qualification of applicants depend on offered wages? Then employers would be ill-advised to pay just the marginal productivity.
- Who even knows what the marginal productivity of a given type of labor is?
- What about social norms for remuneration, and what about collective bargaining?
- What if employers face upward sloping labor supply, maybe because of search frictions? Then they would depress wages below marginal productivity, acting as a “monopsony.”
- What if labor markets don’t clear, for whatever reason?

That said, as far as the function $w(N, \epsilon)$ is concerned, the assumption that wages are determined by the marginal productivity of types of labor with respect to some aggregate production function does not impose much of a restriction. Its only implication for the behavior of the demand function is that it implies the **symmetry** condition

$$\frac{\partial w_j}{\partial N_{j'}} = \frac{\partial^2 f}{\partial N_{j'} \partial N_j} = \frac{\partial^2 f}{\partial N_j \partial N_{j'}} = \frac{\partial w_{j'}}{\partial N_j}. \quad (6.5)$$

This symmetry is sufficient and necessary for the existence of a function f such that equation (6.4) holds.

Usually it is also assumed that the aggregate production function exhibits **constant returns to scale**, so that

$$f(\alpha N_1, \dots, \alpha N_J) = \alpha \cdot f(N_1, \dots, N_J) \quad (6.6)$$

⁴Concentrating out a variable means to plug in its maximizing value given the other variables. For instance, concentrating out K from $F(N, K)$ gives the function $f(N) := \max_K F(N, K)$.

for all $\alpha > 0$. This says that if all inputs are increased by a factor α , then so is aggregate output. Note that constant returns to scale in terms of labor inputs implicitly requires an infinitely elastic supply of capital and other factors, as well as an infinitely elastic demand for final products.

Constant returns to scale implies that wages only depend on relative supplies of labor. To see this, differentiate both sides of equation (6.6) with respect to N_j , which yields

$$\alpha \cdot \partial_j f(\alpha N_1, \dots, \alpha N_J) = \alpha \cdot \partial_j f(N_1, \dots, N_J),$$

where I use $\partial_j f$ to denote the partial derivative of f with respect to its j th argument. This in turn implies

$$w_j(\alpha N_1, \dots, \alpha N_J, \epsilon) = w_j(N_1, \dots, N_J, \epsilon). \quad (6.7)$$

6.4 The CES production function

Empirical work often assumes a specific functional form for the aggregate production function. The most common form is the “constant elasticity of substitution” production function. This production function takes the form

$$y = f(N_1, \dots, N_J) = \left(\sum_{j'=1}^J \alpha_{j'} N_{j'}^\rho \right)^{1/\rho}, \quad (6.8)$$

where the α_j and ρ are unknown parameters which we might try to estimate from data. Assuming the marginal productivity theory of wages, equation (6.4), we get

$$\begin{aligned} w_j &= \frac{\partial f(N_1, \dots, N_J)}{\partial N_j} = \left(\sum_{j'=1}^J \alpha_{j'} N_{j'}^\rho \right)^{1/\rho-1} \cdot \alpha_j \cdot N_j^{\rho-1} \\ &= y^{1-\rho} \cdot \alpha_j \cdot N_j^{\rho-1}. \end{aligned} \quad (6.9)$$

This implies that the relative wage between groups j and j' is given by

$$\frac{w_j}{w_{j'}} = \frac{\alpha_j}{\alpha_{j'}} \cdot \left(\frac{N_j}{N_{j'}} \right)^{\rho-1}$$

Denote

$$\sigma = \frac{1}{\rho} - 1,$$

so that $\rho - 1 = -1/\sigma$. σ is called the **elasticity of substitution**. It describes the slope of how relative wages depend on relative supply. This slope is constant for the given production function, lending it the name “constant elasticity of substitution.”

In log terms, the relative wage can be written as

$$\log \left(\frac{w_j}{w_{j'}} \right) = \log \left(\frac{\alpha_j}{\alpha_{j'}} \right) - \frac{1}{\sigma} \log \left(\frac{N_j}{N_{j'}} \right). \quad (6.10)$$

Substituting controls (including trends), as well as an unobserved residual ϵ , for the term $\log \left(\frac{\alpha_j}{\alpha_{j'}} \right)$, we get our initial regression specification.

6.5 Generalizations of the CES production function

The basic CES production function is fairly restrictive. The literature uses various generalizations, where each type of labor is considered to be an aggregate of sub-types.

Generalization 1: Aggregate types

Assume that the J types of labor can be grouped into K aggregate types, where type j belongs to aggregate type k_j . Many papers assume that the production function f is CES with respect to the supply L_k of these aggregate types. Within them, different types are perfectly substitutable, but might have different marginal productivities (different θ_j). Formally,

$$y = f(N_1, \dots, N_J) = \left(\sum_{k=1}^K \alpha_k L_k^\rho \right)^{1/\rho}, \quad (6.11)$$

$$L_k = \sum_{j:k_j=k} \theta_j N_j.$$

We get wages, relative wages, and their logarithm to equal

$$\begin{aligned} w_j &= \frac{\partial f(N_1, \dots, N_J)}{\partial N_j} = \left(\sum_{k=1}^K \alpha_k L_k^\rho \right)^{1/\rho-1} \cdot \alpha_{k_j} \cdot L_{k_j}^{\rho-1} \cdot \theta_j \\ &= y^{1-\rho} \cdot \alpha_{k_j} \cdot L_{k_j}^{\rho-1} \cdot \theta_j \\ \frac{w_j}{w_{j'}} &= \frac{\theta_j}{\theta_{j'}} \cdot \frac{\alpha_{k_j}}{\alpha_{k_{j'}}} \cdot \left(\frac{L_{k_j}}{L_{k_{j'}}} \right)^{\rho-1} \\ \log \left(\frac{w_j}{w_{j'}} \right) &= \log \left(\frac{\alpha_{k_j}}{\alpha_{k_{j'}}} \cdot \frac{\theta_j}{\theta_{j'}} \right) - \frac{1}{\sigma} \log \left(\frac{L_{k_j}}{L_{k_{j'}}} \right) \end{aligned}$$

The relative wages of types j and j' thus depend only on the relative supply of aggregate types k_j and $k_{j'}$.

Generalization 2: Nested CES

As in generalization 1, assume that there are K aggregate types which are substitutable with elasticity σ_1 . Within these aggregate types, however, the sub-types

are not perfectly substitutable, but instead have an elasticity of substitution of σ_2 . Formally,

$$y = f(N_1, \dots, N_J) = \left(\sum_{k=1}^K \alpha_k L_k^{\rho_1} \right)^{1/\rho_1}$$

$$L_k = \left(\sum_{j:k_j=k} \theta_j N_j^{\rho_2} \right)^{1/\rho_2}$$

We get wages, relative wages, and their logarithm to equal

$$w_j = \frac{\partial f(N_1, \dots, N_J)}{\partial N_j} = y^{1-\rho_1} \cdot \alpha_{k_j} \cdot L_{k_j}^{\rho_1-\rho_2} \cdot \theta_j \cdot N_j^{\rho_2-1}$$

$$\frac{w_j}{w_{j'}} = \frac{\theta_j}{\theta_{j'}} \cdot \frac{\alpha_{k_j}}{\alpha_{k_{j'}}} \cdot \left(\frac{L_{k_j}}{L_{k_{j'}}} \right)^{\rho_1-\rho_2} \cdot \left(\frac{N_j}{N_{j'}} \right)^{\rho_2-1}$$

$$\log \left(\frac{w_j}{w_{j'}} \right) = \log \left(\frac{\alpha_{k_j}}{\alpha_{k_{j'}}} \cdot \frac{\theta_j}{\theta_{j'}} \right) + \left(\frac{1}{\sigma_2} - \frac{1}{\sigma_1} \right) \log \left(\frac{L_{k_j}}{L_{k_{j'}}} \right) - \frac{1}{\sigma_2} \log \left(\frac{N_j}{N_{j'}} \right)$$

The relative wages of j and j' thus depend on both $L_{k_j}/L_{k_{j'}}$ and $N_j/N_{j'}$.

6.6 Instruments

Contrary to what we have assumed so far, variation in labor supply might not be random relative to other factors determining wages. We should especially worry about this when comparing wages across cities. It might for instance be the case that workers with a college degree migrate to cities where the returns to a college degree are highest.

To take care of this endogeneity issue, Card (2009) proposes to instrument changes of labor supply making clever use of prior migration patterns.⁵ The idea is that new migrants tend to settle in the same cities as previous migrants from the same source countries, while the amount of new migrants at the national level is arguably not affected by city-specific economic conditions. Suppose for instance that prior migrants from country m happened to settle mostly in Chicago and Los Angeles. If then some political or economic crisis in country m compels a new set of people from m to leave their home country, Chicago and Los Angeles will likely experience an increase in their labor force which is unrelated to local economic conditions in these two cities.

Let us make this more formal, cf. (Card, 2009, p8). Let m index different source countries for migrants, and let M_m denote the number of new migrants

⁵If you need to review instrumental variables, have a look at chapter 4 of Angrist and Pischke (2010) or my lecture slides at <http://scholar.harvard.edu/files/kasy/files/appliedemxslides.zip>.

from country m arriving in a given time period. Let λ_m denote the share of prior migrants from country m living in a given city i , and $\delta_{m,j}$ the share of migrants from country m that are of type j . If new migrants from m have the exact same settlement patterns as prior migrants, then we should see

$$\sum_m \lambda_m M_m$$

new migrants arriving in a given city. Given their distribution of types, this would imply a growth of the population of workers of type j in the given city by a factor

$$Z_j = \frac{1}{N_j} \cdot \sum_m \lambda_m M_m \delta_{m,j}. \quad (6.12)$$

This is the instrument for changes in $\log N_j$ that Card proposes. For the relative change of N_j and $N_{j'}$ we can use the instrument

$$Z_{j,j'} = \frac{Z_j}{Z_{j'}}.$$

This is a valid instrument if it satisfies the condition

$$E[Z_{j,j'} \cdot \epsilon_{j,j'}] = 0.$$

Under this condition, we can estimate β in equation (6.1) by

$$\hat{\beta} = \frac{E_n[Z_{j,j'} \cdot \log(w_j/w_{j'})]}{E_n[Z_{j,j'} \cdot \log(N_j/N_{j'})]}, \quad (6.13)$$

where E_n denotes sample averages across cities.

6.7 Matlab exercises

Chapter 7

Intergenerational mobility – Measurement error

In this chapter we are discussing the literature on intergenerational mobility as reviewed in Black and Devereux (2011), and the important recent contribution of Chetty et al. (2014).

To what extent does our parents' economic situation determine our own economic chances? And to what extent is “equality of opportunity” a reality? This seemingly well-posed question is conceptually quite a bit trickier than it might seem at first. To make this point, let us consider a number of different objects. Each of these objects might be considered a measure of intergenerational mobility.

1. **Predictability** of (log) child income in a given year (or a few years) **using** (log) parent **income in a given year** (or a few years):

$$E[Y_{c,s}|Y_{p,t}]$$

This conditional expectation describes the average (log) of child income in a given year, or at a given age, among all those children whose parents received an income of $Y_{p,t}$ in a given (earlier) year.

$$\frac{\text{Cov}(Y_{p,t}, Y_{c,s})}{\text{Var}(Y_{p,t})}$$

This fraction gives the slope of a linear approximation to the conditional expectation in the previous display. This approximation is called the best linear predictor. If income is measured in log terms, this slope describes the percentage increase in average child income for a 1% increase in parent income. This slope is the most common measure of intergenerational mobility. Another variant of this measure uses ranks in the national income distribution, instead of levels or logs.

2. **Predictability of (log) child lifetime income using (log) parent lifetime income:**

$$E[\bar{Y}_c | \bar{Y}_p]$$

This conditional expectation describes the average (log) of child lifetime income,⁶ among all those children whose parents received a lifetime income of \bar{Y}_p .

$$\frac{\text{Cov}(\bar{Y}_p, \bar{Y}_c)}{\text{Var}(\bar{Y}_p)}$$

As before, this fraction gives the slope of a linear approximation to the conditional expectation in the previous display.

Measured income varies significantly over time, because of the life cycle of earnings, because of transitory shocks, and because of measurement error. Arguably we might be more interested in the relationship between the lifetime incomes of parents and children, that is, in the relationship between long-run average incomes. Lifetime income is in general more strongly related between parents and children than short-run income. We will discuss why in sections 7.1 and 7.2 below. The relationship between lifetime incomes is often considered the actual object of interest of mobility studies. Lifetime incomes are hard to observe, however, which is why short-term incomes are studied more often.

3. **Predictability using additional variables:** But why stop there? Is it not equally relevant how other factors such as parent education, location of residence, etc. predict child outcomes? Philosophers such as Rawls argue that features such as these, determined at birth and out of our control, are “morally arbitrary” – they should not determine our chances in life. More generally, we might be interested in knowing to what extent life outcomes are predictable at birth. The more predictive factors we consider, the better we will be able to predict child outcomes. This motivates consideration of objects such as the following:

$$E[\bar{Y}_c | \bar{Y}_p, X_p, W_p]$$

This conditional expectation describes the average (log) of child lifetime income, among all those children whose parents received a lifetime income of \bar{Y}_p , who had education level X_p , location of residence W_p , etc.

$$\text{Var}((\bar{Y}_p, X_p, W_p))^{-1} \cdot \text{Cov}((\bar{Y}_p, X_p, W_p), \bar{Y}_c).$$

As before, this vector gives the slopes of a linear approximation to the conditional expectation shown above.

4. **The causal effect of parent lifetime income:**

$$\bar{Y}_c = g(\bar{Y}_p, \epsilon).$$

⁶This is also called “permanent income.”

The structural function g describes the causal effect of parent income on child income.⁷ If parent income is changed, it is assumed that the function g and the set of unobserved factors ϵ do *not* change. Not all correlations are causal. Child and parent income might be statistically related because education is transmitted across generations, for instance, without there being a causal effect of parent income. The causal effect of parent income on their children is the subject of a more recent literature, usually using instrumental variables; we will discuss this in section 7.4. We might, for instance, care about this causal effect if we are interested in the effect of redistributive taxation on the next generation.

5. The **causal effect of additional variables**:

$$\bar{Y}_c = h(\bar{Y}_p, X_p, W_p, \epsilon')$$

The structural function h describes the causal effect of parent income, education, location of residence, etc., on child income. As before, other factors X_p, W_p might have a causal effect, in addition to the effect of parental income \bar{Y}_p . We might for instance be interested in the effect of current educational policy on future generations.

7.1 Classical measurement error and transitory shocks

As mentioned, much of the literature on intergenerational income mobility is interested in objects of the form

$$\beta := \frac{\text{Cov}(\bar{Y}_p, \bar{Y}_c)}{\text{Var}(\bar{Y}_p)}, \quad (7.1)$$

describing the **predicted percentage increase** in child **lifetime income** for a 1% increase in parent lifetime income. A key concern is that this slope is different than the slope we would get from a regression on short-run parental income. To see why, suppose that

$$\begin{aligned} Y_{p,t} &= \bar{Y}_p + \epsilon_{p,t} \\ Y_{c,s} &= \bar{Y}_c + \epsilon_{c,s}, \end{aligned} \quad (7.2)$$

where

$$\begin{aligned} \text{Cov}(\bar{Y}_p, \epsilon_{p,t}) &= \text{Cov}(\bar{Y}_p, \epsilon_{c,s}) = 0 \\ \text{Cov}(\bar{Y}_c, \epsilon_{c,s}) &= \text{Cov}(\bar{Y}_c, \epsilon_{p,t}) = 0 \\ \text{Cov}(\epsilon_{p,t}, \epsilon_{c,s}) &= 0. \end{aligned} \quad (7.3)$$

⁷Structural functions are an alternative, equivalent notation for the potential outcomes which we learned about in chapter 4.

These equations say that, for both parents and children, income in a given year is equal to permanent income plus a shock ϵ of mean 0. The shocks might either be due to transitory fluctuations in actual income, or due to measurement error. These shocks are assumed to be uncorrelated with permanent incomes, and with each other. This assumption holds true for what is called “**classical measurement error**.”

Suppose we estimate the slope of a **regression of short run incomes**,

$$\gamma := \frac{\text{Cov}(Y_{p,t}, Y_{c,s})}{\text{Var}(Y_{p,t})}. \quad (7.4)$$

How are the parameters β and γ related to each other? Let us start by looking at the covariance in the numerator of γ :

$$\begin{aligned} & \text{Cov}(Y_{p,t}, Y_{c,s}) \\ &= \text{Cov}(\bar{Y}_p, \bar{Y}_c) + \text{Cov}(\bar{Y}_p, \epsilon_{c,s}) + \text{Cov}(\epsilon_{p,t}, \bar{Y}_c) + \text{Cov}(\epsilon_{p,t}, \epsilon_{c,s}) \\ &= \text{Cov}(\bar{Y}_p, \bar{Y}_c). \end{aligned}$$

All except the first of the covariance terms in the second line of this display are zero by the assumptions we imposed on the measurement errors, that is by equation (7.3). So we get the same covariance in the numerator for both β and γ . What about the denominator?

$$\text{Var}(Y_{p,t}) = \text{Var}(\bar{Y}_p) + 2 \cdot \text{Cov}(\epsilon_{p,t}, \bar{Y}_p) + \text{Var}(\epsilon_{p,t}) = \text{Var}(\bar{Y}_p) + \text{Var}(\epsilon_{p,t}),$$

and therefore

$$\gamma = \frac{\text{Var}(\bar{Y}_p)}{\text{Var}(\bar{Y}_p) + \text{Var}(\epsilon_{p,t})} \cdot \beta. \quad (7.5)$$

The short run coefficient γ is smaller than the long run coefficient β by a factor which depend on the relative variance of measurement error (transitory shocks), and lifetime income. This phenomenon is known as **attenuation bias**.

Note that, for this type of measurement error, only the error on the parent side matters, while the variance of $\epsilon_{c,s}$ does not show up in the formula for the attenuation bias!

7.2 Non-classical measurement error and the lifetime profile of earnings

In the last section we assumed that measurement error for both parents and children is “classical,” that is, it has mean 0 and is independent of actual lifetime earnings. That assumption might not be correct, especially when earnings are measured early in life (say, before age 35). This is often the case for children, who are not old enough in the available data-sets for us to observe their income at a later point in life.

The reason is that the **profile of earnings over a lifetime** is quite different depending on the qualifications required for a particular occupation. The annual earnings of those with higher lifetime earnings tend to rise more steeply with experience relative to those with lower lifetime earnings.

For illustration, consider the following model: maintain the same assumption as in the previous section, except that

$$Y_{c,s} = \bar{Y}_c \cdot (1 + \alpha \cdot (s - \bar{s})) + \epsilon_{c,s}. \quad (7.6)$$

This equation says that the earnings of children rise over time in a way that is positively related to lifetime earnings, by a factor which is determined by the parameter $\alpha > 0$. The earnings of children are on average equal to lifetime earnings at age \bar{s} .

Under equation (7.6), we get

$$\begin{aligned} & \text{Cov}(Y_{p,t}, Y_{c,s}) \\ &= \text{Cov}(\bar{Y}_p, \bar{Y}_c \cdot (1 + \alpha \cdot (s - \bar{s}))) + \text{Cov}(\bar{Y}_p, \epsilon_{c,s}) \\ & \quad + \text{Cov}(\epsilon_{p,t}, \bar{Y}_c \cdot (1 + \alpha \cdot (s - \bar{s}))) + \text{Cov}(\epsilon_{p,t}, \epsilon_{c,s}) \\ &= (1 + \alpha \cdot (s - \bar{s})) \cdot \text{Cov}(\bar{Y}_p, \bar{Y}_c), \end{aligned}$$

so that

$$\gamma = (1 + \alpha \cdot (s - \bar{s})) \cdot \frac{\text{Var}(\bar{Y}_p)}{\text{Var}(\bar{Y}_p) + \text{Var}(\epsilon_{p,t})} \cdot \beta.$$

Suppose, for simplicity, that measurement error for parents' income was not an issue, i.e. $\text{Var}(\epsilon_{p,t}) = 0$. In that case

$$\gamma = (1 + \alpha \cdot (s - \bar{s})) \cdot \beta. \quad (7.7)$$

If we observe children at age \bar{s} , there is no problem. If we observe them at an age s younger than \bar{s} , however, there is a downward bias in γ relative to β , since $(1 + \alpha \cdot (s - \bar{s})) < 1$ in this case.

7.3 Remedies

It is important to recognize that the reasons for the downward bias are different between sections 7.1 and 7.2, so that the remedies in either case are different, as well. In the case of classical measurement error, as in section 7.1, the problem is that we **over-estimate the inequality** of lifetime incomes on the **parent** side. Since part of the estimated inequality is simply due to measurement errors and transitory shocks, this part of inequality is not inherited by children, and we conclude erroneously that inequality of incomes is transmitted to a lesser extent than it actually is. In section 7.2, the problem is related to the fact that we **under-estimate the inequality** of the lifetime incomes of **children**. This again implies that regressions of short-term incomes suggest that parental inequality of incomes is transmitted to a lesser extent than it actually is.

There are a number of **remedies** for the problem of **classical measurement error** on the parent side:

1. **Use better data:** When we are interested in earnings, administrative data (for instance from the IRS, or from social security administrations in other countries) tend to be more reliable than self-reported earnings in surveys, that is, they have measurement error with smaller variance.

This strategy of course only takes care of measurement error, but not of transitory shocks to actual earnings

2. **Average earnings over several years:** Suppose for illustration that shocks are uncorrelated across years and have constant variance. Then

$$\frac{1}{k} \sum_{t=t_0}^{t_0+k} Y_{p,t} = \bar{Y}_p + \frac{1}{k} \sum_{t=t_0}^{t_0+k} \epsilon_{p,t},$$

and

$$\text{Var} \left(\frac{1}{k} \sum_{t=t_0}^{t_0+k} \epsilon_{p,t} \right) = \frac{1}{k^2} \sum_{t=t_0}^{t_0+k} \text{Var}(\epsilon_{p,t}) = \frac{1}{k} \text{Var}(\epsilon_{p,t_0}).$$

Averaging earnings over k years thus reduces the variance of measurement error by a factor $1/k$, and correspondingly reduces the attenuation bias from a factor of $1 / (1 + \text{Var}(\epsilon_{p,t}) / \text{Var}(\bar{Y}_p))$ to a factor of

$$\frac{1}{1 + \frac{1}{k} \frac{\text{Var}(\epsilon_{p,t})}{\text{Var}(\bar{Y}_p)}}. \quad (7.8)$$

3. **Assessing the reliability of the data:** Suppose we have two measurements Y_{p,t_1} and Y_{p,t_2} of parental income with independent measurement error. Then the correlation of these two variables is equal to

$$\text{corr}(Y_{p,t_1}, Y_{p,t_2}) = \frac{\text{Cov}(Y_{p,t_1}, Y_{p,t_2})}{\sqrt{\text{Var}(Y_{p,t_1}) \cdot \text{Var}(Y_{p,t_2})}} = \frac{\text{Var}(\bar{Y}_p)}{\text{Var}(\bar{Y}_p) + \text{Var}(\epsilon_{p,t})}. \quad (7.9)$$

But this factor is exactly the same as the one describing the attenuation bias from β to γ .

The situation is more complicated for **non-classical** measurement error, for instance of the form discussed in section 7.2. The main remedy for measurement error of this kind is to **use child income measured at a later point in life**, when the dispersion of annual earnings more closely resembles the dispersion of lifetime earnings. Another remedy would be to “move the goalpost,” and to focus on **other outcomes** that are determined earlier in life. The leading example would be educational attainment, which is usually well determined by the time children have reached their late 20s.

7.4 The causal effect of parental income; instruments

So far, we considered regressions of short-term incomes (the first of the objects introduced at the beginning of this chapter), regressions of lifetime incomes (the second of the objects), and their relationship. We shall now turn to the causal effect of parental income (the fourth object we introduced), and how it relates to regressions of (lifetime) income.

7.5 Matlab exercises

Chapter 8

The distributional effect of changing prices – Equivalent variation

So far, these lecture notes have considered (potentially) **observable** outcomes – income and wealth, earnings and wages. These outcomes all measure different things. Depending on what we are ultimately interested in, consideration of one or the other of these outcomes might be the most relevant object to study.

Suppose now that we are ultimately interested in some notion of **welfare**.⁸ How could we possibly measure welfare? Do any of the observable outcomes considered so far correspond to welfare? Economists like to think of welfare in terms of realized **utility**. Utility, however, is **not observable**. It is therefore not obvious whether “utility” provides a meaningful concept of welfare at all.

Rather than attempting to directly measure the level of utility, we might **change the question**. Instead of the level we might try to measure changes in utility, induced by a given change in prices, wages, taxes, or some other policy. And instead of measuring changes in utility itself, we might try to measure how these changes in utility compare to changes that would be induced by a simple transfer of money. We have thus shifted the question in two ways:

1. **Changes** in utility, rather than **levels** of utility.
2. Transfers of **money** that would induce similar changes of utility, rather than changes in **utility** itself.

It turns out that, when we modify the question in this way, then it has a well-defined and surprisingly simple answer – at least if we assume that individuals

⁸Great discussions on various notions of welfare and how it might be measured can be found in Sen (1995) and Roemer (1998).

are utility maximizing. In this chapter, we will derive this answer for the case of changing prices of consumer goods. We will show that a change dp_j of the price p_j of good j has the same effect on utility of individual i as a reduction of income by $dp_j \cdot x_{j,i}$, where $x_{j,i}$ is the individual's current consumption of good j . In the next chapter, we will use the same idea to talk about redistributive taxation.

8.1 The consumer problem

Standard economic theory assumes that individuals choose their **consumption** to **maximize** their **utility**. They do so subject to the constraint that their expenses do not exceed their income. Denote individuals by i . Assume there are two consumption goods, good 1 and good 2, with prices p_1 and p_2 . Individual i has an income y_i and chooses her consumption $x_i = (x_{1,i}, x_{2,i})$ to maximize her utility u_i . Formally,

$$x_i(p, y_i) = \operatorname{argmax}_x u_i(x) \quad (8.1)$$

subject to the **budget constraint**

$$x_{1,i} \cdot p_1 + x_{2,i} \cdot p_2 \leq y_i. \quad (8.2)$$

The utility v_i that a household can achieve for given prices and income is equal to the utility of the chosen consumption bundle,

$$v_i(p, y_i) = u_i(x_i(p, y_i)). \quad (8.3)$$

We can rewrite the individual's budget constraint (assuming that it holds with equality) to express consumption of good 1 in terms of the other variables,

$$x_{1,i} = \frac{1}{p_1}(y_i - x_{2,i} \cdot p_2)$$

We can next substitute the budget constraint, rewritten in this form, into the optimization problem, to get an unconstrained problem:

$$x_{2,i} = \operatorname{argmax}_{x_2} u_i \left(\frac{1}{p_1}(y_i - x_2 \cdot p_2), x_2 \right).$$

The solution to this unconstrained problem has to satisfy the **first order condition**

$$\frac{\partial}{\partial x_2} \left[u_i \left(\frac{1}{p_1}(y_i - x_2 \cdot p_2), x_2 \right) \right] = 0,$$

which we can rewrite as

$$\frac{\partial_{x_1} u_i(x_i)}{p_1} = \frac{\partial_{x_2} u_i(x_i)}{p_2}. \quad (8.4)$$

This first order condition is sometimes interpreted as saying that the ratio of marginal benefits to marginal costs has to be the same for both goods.

8.2 Changing prices

When the price of good 2 changes, how does that affect the welfare of different individuals? Formally, what is $\partial_{p_2} v_i(p, y_i)$?

To make the notation less cluttered, we drop the subscript i from the following derivation. You should not forget that everything is different for different individuals, though. We can calculate the **welfare effect of changing** p_2 using (i) the chain rule, (ii) substituting for $\partial_{p_2} x_1$ using the rewritten budget constraint, (iii) rearranging, and (iv) using the first order condition of utility maximization:

$$\begin{aligned}\partial_{p_2} v(p, y) &= \partial_{x_1} u(x) \cdot \partial_{p_2} x_1 + \partial_{x_2} u(x) \cdot \partial_{p_2} x_2 \\ &= \partial_{x_1} u(x) \cdot \left(-\frac{x_2}{p_1} - \partial_{p_2} x_2 \frac{p_2}{p_1} \right) + \partial_{x_2} u(x) \cdot \partial_{p_2} x_2 \\ &= -\partial_{x_1} u(x) \cdot \frac{x_2}{p_1} + \left(-\frac{\partial_{x_1} u(x)}{p_1} + \frac{\partial_{x_2} u(x)}{p_2} \right) \cdot p_2 \cdot \partial_{p_2} x_2 \\ &= -x_2 \cdot \frac{\partial_{x_1} u(x)}{p_1}.\end{aligned}$$

Make sure you understand each step of this proof! The most important step in this derivation is the last one: because we assume that **individuals** maximize utility, the **first order condition** holds. And because the first order condition holds, we can **drop the term involving** $\partial_{p_2} x_2$. As far as their welfare is concerned, it does not really matter how individuals react to price changes!⁹

We can do a completely similar calculation to get the effect of increasing income y :

$$\begin{aligned}\partial_y v(p, y) &= \partial_{x_1} u(x) \cdot \partial_y x_1 + \partial_{x_2} u(x) \cdot \partial_y x_2 \\ &= \partial_{x_1} u(x) \cdot \left(-\frac{1}{p_1} - \partial_y x_2 \frac{p_2}{p_1} \right) + \partial_{x_2} u(x) \cdot \partial_y x_2 \\ &= -\frac{\partial_{x_1} u(x)}{p_1} + \left(-\frac{\partial_{x_1} u(x)}{p_1} + \frac{\partial_{x_2} u(x)}{p_2} \right) \cdot p_2 \cdot \partial_{p_2} x_2 \\ &= -\frac{\partial_{x_1} u(x)}{p_1}.\end{aligned}$$

As before: as far as their welfare is concerned, it does not really matter how individuals react to income changes. Now we are almost done. We can calculate how the welfare effect of a price change dp_2 compares to the welfare effect of a change in income. This is called **equivalent variation**, we abbreviate it by

⁹In fact, there is a deep sense in which this is the *only* implication of “welfarism,” where welfarism is the idea of evaluating household welfare based on their realized utility. This is sometimes also called utilitarianism.

EV:

$$EV = \frac{\partial_{p_2} v(p, y) \cdot dp_2}{\partial_y v(p, y)} = \frac{-x_2 \cdot \frac{\partial_{x_1} u(x)}{p_1} \cdot dp_2}{-\frac{\partial_{x_1} u(x)}{p_1}} = -x_2 \cdot dp_2. \quad (8.5)$$

Increasing the price of good 2 by one dollar has the same effect on individual i as decreasing her income by $-x_{2,i}$ dollars, where $x_{2,i}$ is the amount she consumes of good 2.

8.3 Generalizing this result

So far, we have considered a fairly special case. Only the price of good 2 changes, there are only two goods, and individuals are only consumers. All of these are easily generalized:

- There is nothing special about good 2, so we get the same result for good 1:

$$EV = \frac{\partial_{p_1} v_i(p, y_i) \cdot dp_1}{\partial_y v_i(p, y_i)} = -x_{1,i} \cdot dp_1.$$

- There is nothing special about the case of 2 goods. We might just as well assume that there are J goods, and do the exact same proof. (This would be a good exercise!). When the price of good j changes by dp_j for $j = 1 \dots J$, this implies a welfare change of

$$EV = \frac{dv_i(p, y_i)}{\partial_y v_i(p, y_i)} = -\sum_j x_{j,i} \cdot dp_j.$$

- We also don't need to assume that individuals are only consumers. Suppose they start out with an endowment

$$\omega_i = (\omega_{1,i}, \dots, \omega_{J,i}),$$

which they can either consume or sell on the market. Their *net* consumption of good j is equal to $x_{j,i} - \omega_{j,i}$. In such a setting, we get

$$EV = \frac{dv_i(p, y_i)}{\partial_y v_i(p, y_i)} = \sum_j (\omega_{j,i} - x_{j,i}) \cdot dp_j.$$

These formulas are the basis of so-called Paasche price indices. They say that we should evaluate price changes by weighting them with an appropriate consumption basket, corresponding to the amounts consumed of each good. This formula gives a different price index for every individual. To evaluate the distributional impact of price changes, all we have to do is to collect information on every individual's net consumption of various goods. Once we have estimates of the welfare changes (in dollar terms) for each individual, we can plot how these welfare changes relate to income or various demographic factors, for instance

location of residence.

In fact, the same logic carries us much further. We won't prove it here, but we get similar evaluations of the welfare effect of price changes if:

- Individuals also face discrete choices, rather than just continuous ones, as we have assumed so far.
- Individuals make intertemporal choices, that is choices over time.
- Individuals also face constraints other than their budget constraint, such as informational constraints, credit constraints, etc.

In all of these cases we can still compare the welfare effect of changing prices (or wages, or interest rates, for that matter), to the welfare effect that a lump-sum transfer of money would have. And in all of these cases we derive this welfare effect by essentially ignoring any behavioral responses to a change in prices.

8.4 Matlab exercises

Chapter 9

Redistributive taxation – optimal tax theory

One of the primary policy tools to address economic inequality is redistributive taxation. Redistributive taxation is, obviously, a very contested field of policy. There is a field of economics that aims to derive “optimal taxes,” including optimal redistributive income taxes, inheritance taxes, etc. We will discuss some of the basic ideas of this field in the present chapter. A key reference for our discussion is Saez (2001).

Recall that we discussed the distributional impact of changes in prices on individuals’ welfare in the previous chapter. What we will do next is very similar, with changing taxes taking the place of changing prices. Additional complications arise because we need to talk about government revenues, and about how to compare the welfare of different people.

There are many different kinds of taxes in practice, including value-added taxes, income taxes, wealth taxes, inheritance taxes, etc. The framework we discuss applies, in principle, to the analysis of all of these.

9.1 General principles

There are some general principles in common to the analysis of “optimal taxes” for different kinds of taxes:

1. **Marginal policy changes:**

The theory of optimal taxation is concerned with finding policies that maximize some notion of social welfare. As usual, we can characterize maximizers by first order conditions. At the optimum, any (feasible) marginal policy change has no effect on social welfare. We thus need to understand the effect of marginal policy changes on welfare.

2. **Envelope theorem:**

The first key ingredient to understand such marginal changes is the result

we proved in the previous chapter: If we (i) measure individual welfare by realized utility and (ii) assume that individuals are maximizing utility subject to the constraints they face, then we can ignore the effect of behavioral responses to policy changes. This result is called the envelope theorem.

3. Welfare weights:

The envelope theorem allows us to evaluate the effect of a policy change on any *individual*, in terms of the amount of dollars that we could equivalently have given or taken from them. But how do we get from there to social welfare? We have to somehow decide how much we care about an additional dollar for a rich person versus an additional dollar for a poor person, or an additional dollar for a disabled person versus for an able-bodied person, etc. It is important to recognize that there is no “scientific” way to make this decision! In particular, it is meaningless from the point of view of economic theory to sum up dollars across people. The decision how to make these trade-offs depends on our moral judgments, and in practice, on the outcome of distributional struggles between different groups.

If we have settled on how to make these trade-offs between different people, we can express them in terms of welfare weights ω_i that measure the value we attach to an additional dollar for person i .

4. Government budget constraint:

When we think about changing taxes, we also have to think about the impact of these changes on government revenues. One way to do this is to only consider tax changes that do not change total revenues. Another way, which is mathematically equivalent, assumes that there is a marginal value λ of additional government revenues, where λ is on the same scale as the welfare weights ω_i . This is the approach we will take.

When we are considering the effect of tax changes on government revenues, we can *not* ignore behavioral responses to these changes. Usually, the tax base, and thereby government revenues, are affected by such behavioral responses. Rich individuals might for instance respond to a tax increase by exploiting additional loopholes in the tax code or by tax evasion.

5. Effects on prices:

When thinking about the effect of changing some tax, we also have to think about how prices and wages are affected by this change. This can be complicated, and is an empirical matter. To simplify our exposition, we will assume in this chapter that prices and wages do not change in response to policy changes.

Let us now state these principles in a more **formal** way. Suppose we are changing a **tax parameter** α , **individual welfare** for person i is given by v_i , and **government revenues** are given by g . A choice of α is optimal if

$$\partial_{\alpha} SWF = 0.$$

Adding up all components of **social welfare**, and using the appropriate welfare weights, we get

$$\partial_\alpha SWF = \sum_i \omega_i \cdot \partial_\alpha v_i + \lambda \cdot \partial_\alpha g. \quad (9.1)$$

The **envelope theorem** tells us that $\partial_\alpha v_i$ can be calculated as the effect on the individual's budget constraint, holding behavior constant.

The effect on **government revenue** $\partial_\alpha g$ has two components, the **direct effect** (holding behavior fixed), and the **behavioral effect** of individuals reacting to the policy change.

9.2 Linear income tax

Let us go through these terms in a more specific context, where individuals choose their labor supply l and consumption x subject to a linear income tax $t = \alpha + \beta \cdot l \cdot w$, where l denotes labor supply and w denotes the wage. Real income taxes are rarely linear, but this assumption allows us to considerably simplify our discussion. Different individuals have different utility functions and different wages. In generalization of the setup we considered in section 8.1, assume **individuals solve**

$$(x_i, l_i) = \operatorname{argmax}_x u_i(x, l) \quad (9.2)$$

subject to the **budget constraint**

$$x_i \cdot p \leq -\alpha + w_i \cdot l_i \cdot (1 - \beta). \quad (9.3)$$

Note that the choice variables x_i and l_i are functions of prices p , wages w_i , and the tax parameters α and β . Realized utility, as before is given by

$$v_i = u_i(x_i, l_i).$$

By exactly the same arguments as in chapter 8, we get that the envelope theorem in this setting implies that the **equivalent variation** of marginally increasing α , and of marginally increasing β , is given by

$$\begin{aligned} EV_\alpha &= -1 \\ EV_\beta &= -w_i \cdot l_i. \end{aligned}$$

As an exercise, try to prove this, going step by step through the arguments of chapter 8.

What about **government revenues**? Effects on these are given by the sum of a mechanical and a behavioral component,

$$\begin{aligned} \partial_\alpha g &= N + \beta \cdot \sum_i w_i \cdot \partial_\alpha l_i \\ \partial_\beta g &= \sum_i w_i \cdot l_i + \beta \cdot \sum_i w_i \cdot \partial_\beta l_i, \end{aligned}$$

where N is the number of people in the population. To simplify exposition, we shall assume that there are no effects of changing α on labor supply, so that $\partial_\alpha l_i = 0$ and thus $\partial_\alpha g = N$.

Now we have all terms that we need to calculate the **marginal effect on social welfare** of changing α and β :

$$\begin{aligned}\partial_\alpha SWF &= \sum_i (\lambda - \omega_i) \\ \partial_\beta SWF &= \sum_i (\lambda - \omega_i) \cdot w_i \cdot l_i + \lambda \cdot \beta \cdot \sum_i w_i \cdot \partial_\beta l_i.\end{aligned}$$

These expressions are obtained by simply adding up everyone's equivalent variation, weighted by ω_i , and the impact on government revenues, weighted by λ .

At the **optimal linear income tax**, both of these expressions have to equal zero. This implies

$$\begin{aligned}\lambda &= E[\omega_i] \\ \lambda \cdot \beta \cdot E[w \cdot \partial_\beta l] &= \text{Cov}(\omega, w \cdot l),\end{aligned}$$

where E denotes the average across individuals, and Cov the covariance across individuals.

The first equation says that the value of an additional dollar for the government is the same as the average value of an additional dollar across the population. The second equation can be rewritten as

$$\beta = \frac{\text{Cov}(\omega/\lambda, w \cdot l)}{E[w \cdot \partial_\beta l]}.$$

This equation says that the marginal tax rate β , that is the **degree of redistribution**,

1. **is decreasing in the covariance of welfare weights and earnings.**

This covariance is negative if we assign larger welfare weights to people with lower earnings, and it is more negative the more welfare weights reflect a desire for redistribution.¹⁰

2. **is decreasing in the behavioral response of the tax base** to an increase in tax rates, $-E[w \cdot \partial_\beta l]$.

If higher tax rates lead to increased tax evasion, for instance, than this behavioral response is negative, as well. This second item reflects constraints on feasible redistribution. To the extent that there are behavioral responses to taxation, it is not possible to take 1\$ from a rich person and give 1\$ to a poor person. If behavioral responses are small (as seems to be the case, with the exception of tax evasion), we might get close, though.

¹⁰It would be a good exercise, involving some calculus, to verify this claim assuming that ω is a decreasing function of $w \cdot l$.

9.3 Optimal top tax rate

Let us now turn to nonlinear income taxes, where we go through a simplified exposition of the arguments in Saez (2001). We will only consider how to set the top tax rate. In standard models, welfare weights (“the marginal welfare value of additional income”) go to zero as income goes to infinity, *relative* to the welfare weights of people with average income. Put differently, an additional dollar for a billionaire is considered to be of much smaller value than an additional dollar for a poor person. If that is so, we want to set the top tax rate to maximize revenues, since the assumption implies

$$\partial_\tau SWF = \lambda \cdot \partial_\tau g,$$

where τ is the top tax rate. This top tax rate applies to everyone above the income threshold \underline{y} .

Assume, returning to chapter 2, that top incomes follow a **Pareto distribution** with parameter α :

$$P(Y > y | Y \geq \underline{y}) = (\underline{y}/y)^\alpha.$$

Assume further that the **elasticity of taxable income** with respect to the “net of tax” rate $1 - \tau$ is equal to η for those above the income threshold \underline{y} , that is

$$\eta = -\partial_\tau y_i \cdot \frac{1 - \tau}{y_i}. \quad (9.4)$$

Government revenues from taxes on top income receivers are equal to

$$g(\tau) = \tau \cdot N \cdot (E[Y | Y \geq \underline{y}] - \underline{y}),$$

where N is the number of individuals above the threshold. We have all terms that we need to calculate the effect of a change of τ on government revenues, which is given by a sum of mechanical and behavioral effects:

$$\begin{aligned} \frac{1}{N} \cdot \partial_\tau g &= (E[Y | Y \geq \underline{y}] - \underline{y}) - \frac{\tau}{1 - \tau} \cdot \eta \cdot E[Y | Y \geq \underline{y}] \\ &= \underline{y} \cdot \left(\frac{\alpha}{\alpha - 1} \cdot \left(1 - \frac{\tau}{1 - \tau} \cdot \eta \right) - 1 \right) \end{aligned} \quad (9.5)$$

Solving the **first order condition** $\partial_\tau g = 0$ yields

$$\left(\frac{\alpha}{\alpha - 1} \cdot \left(1 - \frac{\tau}{1 - \tau} \cdot \eta \right) - 1 \right) = 0,$$

or, after some algebra

$$\tau = \frac{1}{1 + \alpha \cdot \eta}.$$

If we plug in the realistic parameter values $\alpha = 2$ and $\eta = .25$, this formula implies an **optimal top tax rate** of $1/(1 + 0.5) = 67\%$. More generally, optimal top tax rates are larger (i) the more unequal the distribution of incomes is (small α), and (ii) the less responsive taxable incomes are to changes in tax rates (less tax loopholes, better tax enforcement).

Bibliography

- Angrist, J. D. and Pischke, J.-S. (2010). The Credibility Revolution in Empirical Economics: How Better Research Design is Taking the Con out of Econometrics. *Journal of Economic Perspectives*, 24(2):3–30.
- Atkinson, A. (2014). After Piketty? *The British Journal of Sociology*, 65(4):619–638.
- Atkinson, A. B. (2015). *Inequality: What Can be Done?*. Harvard University Press.
- Atkinson, A. B., Piketty, T., and Saez, E. (2011). Top incomes in the long run of history. *Journal of Economic Literature*, 49(1):3–71.
- Autor, D. H., Katz, L. F., and Kearney, M. S. (2008). Trends in US wage inequality: Revising the revisionists. *The Review of Economics and Statistics*, 90(2):300–323.
- Bertrand, M. and Mullainathan, S. (2004). Are Emily and Greg More Employable Than Lakisha and Jamal? A Field Experiment on Labor Market Discrimination. *American Economic Review*, 94(4):991–1013.
- Black, S. and Devereux, P. (2011). Recent developments in intergenerational mobility. *Handbook of Labor Economics*, 4:1487–1541.
- Boustan, L. P. (2009). Competition in the promised land: Black migration and racial wage convergence in the north, 1940–1970. *The Journal of Economic History*, 69(03):755–782.
- Card, D. (2009). Immigration and inequality. *The American Economic Review*, 99(2):1–21.
- Chetty, R., Hendren, N., Kline, P., and Saez, E. (2014). Where is the land of opportunity? The geography of intergenerational mobility in the United States. *Quarterly Journal of Economics*, 129(4):1553–1623.
- Clemens, M. A., Montenegro, C. E., and Pritchett, L. (2009). The place premium: wage differences for identical workers across the US border.

- Deaton, A. (1989). Rice prices and income distribution in Thailand: a non-parametric analysis. *The Economic Journal*, pages 1–37.
- Deaton, A. (2013). *The great escape: Health, wealth, and the origins of inequality*. Princeton University Press, Princeton.
- DiNardo, J., Fortin, N., and Lemieux, T. (1996). Labor market institutions and the distribution of wages, 1973-1992: A semiparametric approach. *Econometrica*, 64:1001–1044.
- Firpo, S., Fortin, N., and Lemieux, T. (2011). Decomposition methods in economics. *Handbook of Labor Economics*, 4:1–102.
- Fogel, R. W. (1994). *Without consent or contract: the rise and fall of American slavery*. WW Norton & Company.
- Fortin, N. M. and Lemieux, T. (1997). Institutional changes and rising wage inequality: Is there a linkage? *The Journal of Economic Perspectives*, 11(2):pp. 75–96.
- Fraser, N. (2013). *Fortunes of Feminism: From State-Managed Capitalism to Neoliberal Crisis*. Verso Books.
- Goldin, C. (2006). The quiet revolution that transformed women’s employment, education, and family. *American Economic Review*, 96(2):1–21.
- Goldin, C. D. and Katz, L. F. (2009). *The race between education and technology*. Harvard University Press.
- Kasy, M. (2014). Who wins, who loses? tools for distributional policy evaluation. *working paper*.
- Milanovic, B. (2015). Global Inequality of Opportunity: How Much of Our Income Is Determined by Where We Live? *The Review of Economics and Statistics*, 2(97):452–460.
- Piketty, T. (2014). *Capital in the 21st Century*. Harvard University Press, Cambridge.
- Rawls, J. (1973). *A theory of justice*. Harvard University Press, Cambridge.
- Roemer, J. E. (1998). *Theories of distributive justice*. Harvard University Press, Cambridge.
- Saez, E. (2001). Using elasticities to derive optimal income tax rates. *The Review of Economic Studies*, 68(1):205–229.
- Sen, A. (1995). *Inequality reexamined*. Oxford University Press, Oxford.
- Wright, E. O. (2005). *Approaches to class analysis*. Cambridge University Press, Cambridge.