Chapter 21

Inequality

Per Krusell and Víctor Ríos-Rull

21.1 Introduction

Traditionally, inequality has simply not been a topic of its own within macroeconomics. Arguably, it has been present indirectly in traditional textbooks through the focus on unemployment, but then mostly as an indicator of how the degree of inefficiency of the aggregate economy moves up and down with the business cycle. Today, the situation is quite different. Since at least the turn of the millennium, inequality has risen and remained high on the macroeconomic research agenda, and macroeconomic policymakers, including central banks, pay significant attention to it.

There are several reasons for the current focus on inequality. One is that many, if not most, economists view significant inequality as a potential cause of concern, especially since many measures reveal a persistent and still ongoing increase in the concentration of earnings and wealth, since the last decades of the past century. At the same time, we recognize that inequality in many ways is a natural result of the working of an efficient market economy. Thus, at the very least we need to understand what explains the observed trends. Second, and relatedly, in many ways the determination of inequality is, by its nature, a macroeconomic phenomenon, i.e., one where general equilibrium interactions are important. For example, as we shall see, modern macroeconomic modeling naturally gives rise to theories of both wage and wealth inequality.

A third, and quite independent, reason for studying inequality in macroeconomics is that there is increasingly convincing evidence that it affects aggregates. How, for example, fiscal and monetary policy changes propagate through the economy critically hinges on the marginal propensities to consume, invest, and work throughout the population; this was evidenced in Chapter 11, where we learned that the propensity to consume varies in the full range between close to zero, for many consumers, and near one, for another non-trivial part of the population of households. A main reason behind the interest in heterogeneous-agent models is precisely that they admit this kind of heterogeneity in propensities. Moreover, they can be parameterized to match microeconomic data so as to maintain quantitative discipline when conducting counterfactual experiments. As a result, this new literature holds, the new generation of macroeconomic models can produce robust predictions and be very valuable for policymakers.

The goal of the present chapter is to briefly review some key data and then go over some of the main ways in which macroeconomic theory has addressed inequality. It is not a survey and therefore aims to keep references at a minimum and instead lay out the key facts and theories in a relatively compact way. It also omits some topics altogether. For example, how inequality influences aggregates through the political system is not addressed, even though it is arguably very important. There are two theory sections. The first one, Section 21.2, addresses the determinants of inequality: what macroeconomic theory predicts. In this section, the focus is first on the *skill premium*, i.e., channels through which highly educated workers earn more than the less educated. Then it presents the core macroeconomic theories of wealth (and consumption) inequality, what is broadly known as *heterogeneous-agent* models. Thereafter, Section 21.3 discusses how the presence of inequality affects the workings of macroeconomic aggregates: why inequality matters for macro.

21.1.1 Data

The idea in this section is to very briefly go over some evidence on inequality measures for the microeconomic equivalents of macroeconomic variables: labor income, wealth, hours worked, and consumption. As in most of the book, the focus is on the U.S., so we encourage the reader to search for the equivalent information for other countries. We begin with static descriptions and then describe some changes over time.

The cross-section

Beginning with the income distribution for U.S. households in 2022, consider Figure 21.1. It shows a histogram (the height of each equal-sized interval represents the fraction of households is such an interval). Its main feature is that it is *highly skewed*: a shape that rises steeply at 0, peaks below $50,000, and then slowly decreases, with a thick right tail. The right tail keeps increasing far beyond the maximum value on the x-axis ($800,000) and still has non-negligible mass at levels 1,000 times that. As a result of the significant skewness, mean income is far higher than median income, a feature that holds in all countries for which there is reliable data.

The same qualitative skewness can be recorded for labor earnings, as well as for wealth. To compare these distributions, it is useful to use a Lorenz curve and, based on it, compute Gini coefficients. A stylized Lorenz curve is depicted as the solid, convex curve for income in Figure 21.2. A point (x, y) on the curve describes the share y of overall income earned by the poorest x percent of the population. Perfect income equality would mean that the Lorenz curve is the 45-degree line (plotted as a dashed line). The Lorenz curve is always below the 45-degree line by construction, as the population is ranked in incomes from left to right. The most extreme income inequality would be a Lorenz curve that is a flat line (equal to the x-axis) up until 100, where it jumps to 100. The Gini coefficient is computed as the area A divided by the area A+B, i.e., the size of the full triangle under the 45-degree line. When there is perfect equality, the Gini coefficient is therefore 0; the Gini coefficient tends to 1 in the extreme case where one agent holds all income.

Turning now to the U.S. Lorenz curves, Figure 21.3 shows, simultaneously, the wealth and labor income distributions. The former is displayed for net wealth (which includes all

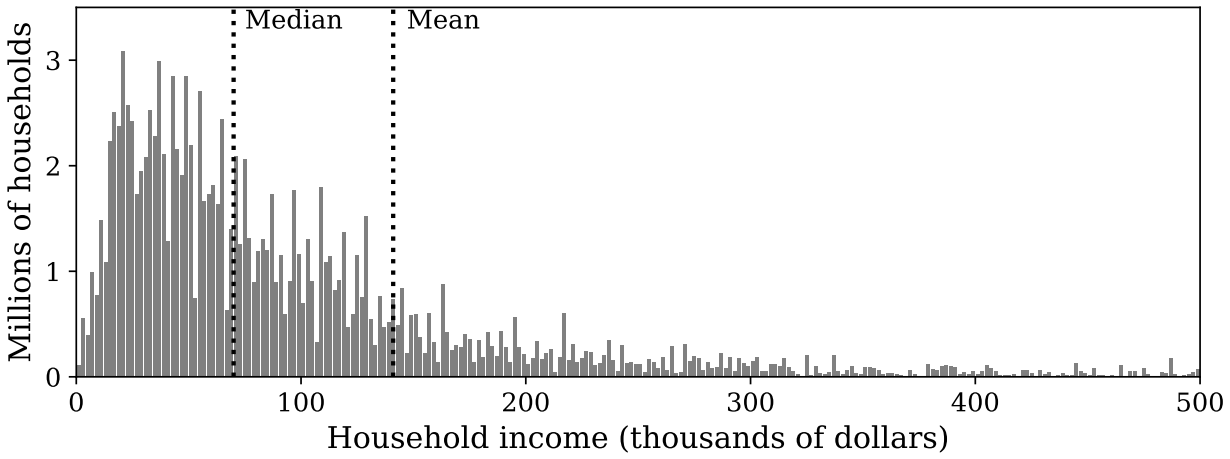

Figure 21.1: Histogram of the income distribution.

Source: Kuhn and Rios-Rull (2025) using the 2022 Survey of Consumer Finances.

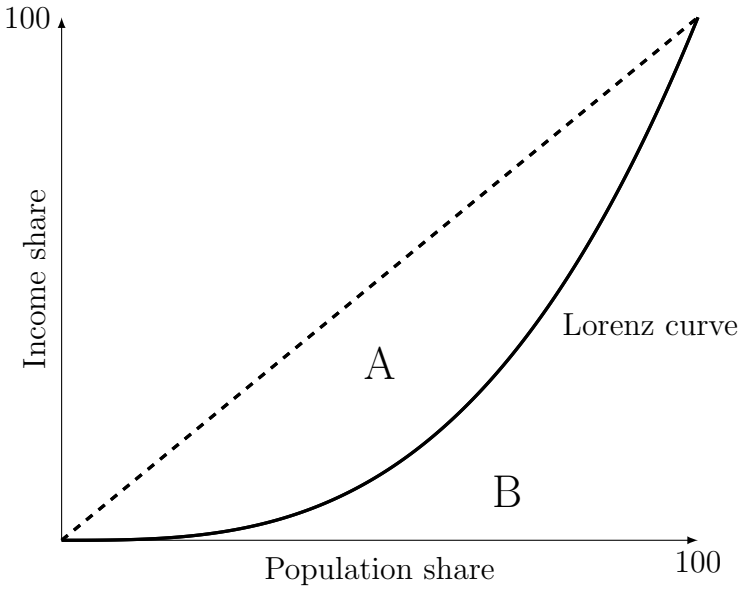

Figure 21.2: Lorenz curve.

assets net of debts) as well as for residential wealth only, and the latter in two versions as well (labor income includes a share of self-employed business income, whereas wage income simply has salaries). The Gini coefficients for each curve are displayed in the legend. We see, first, that the distribution for net worth, with a Gini at 0.83, is far more dispersed than those for the other distributions. Residential wealth is the least dispersed, with a Gini of 0.66, and the two labor income distributions are just slightly more dispersed, with Ginis around 0.68. The finding that wealth is more dispersed than labor income holds qualitatively over time and for all other economies for which we have seen studies.

Table 21.1 displays distributional information for earnings, total income, wealth, consumption, and hours worked. We learn that the earnings- and wealth-poorest 10 percent of

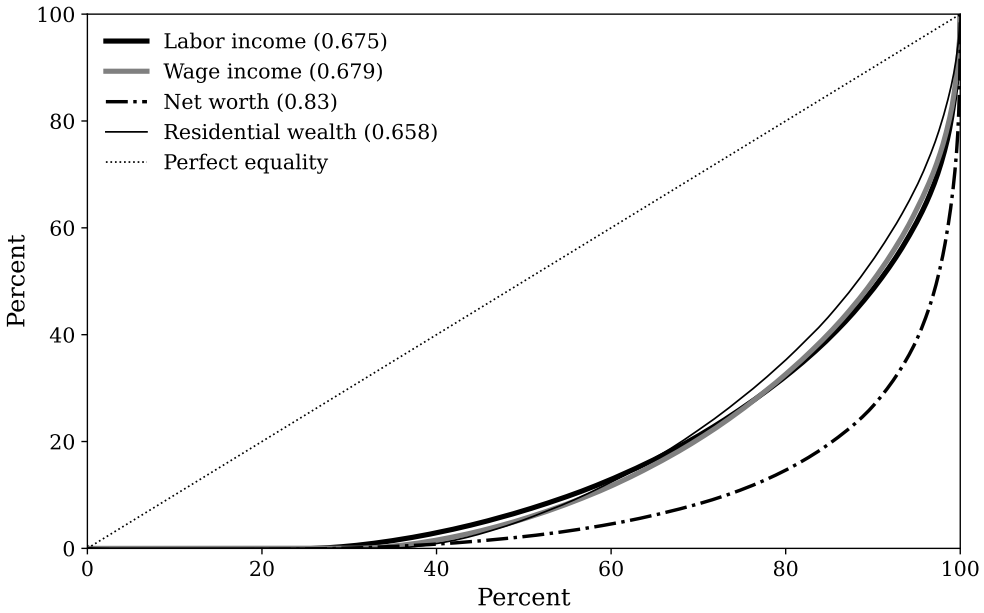

Figure 21.3: Lorenz curves of labor earnings, residential wealth, and net worth.

Source: SCF 2022. See Kuhn and Rios-Rull (2025).

the population of households have no earnings or wealth, respectively, while the 1-percent earnings- and wealth-richest have 19.4 and 35.1 percent of the total earnings and wealth, respectively. Especially the 35.1 number is staggering: over 1/3 of all wealth is held by the 1-percent wealthiest. The measurement of wealth is difficult, as there is no registry data on wealth and surveys offer a very partial account and have trouble sampling the very richest; for the richest groups, off-shore (unrecorded) wealth is another challenge. Furthermore, the taxation of wealth (either directly or via estate taxes) gives incentives to under-report the value of assets, especially for those businesses that are not publicly traded. One way to estimate wealth is from tax returns, by looking at capital income and inferring the stock from the income flow. Yet, in countries where wealth is taxed, registry data shows similar qualitative features to those in the table for the U.S.

Another characteristic often used in characterizing income and wealth distributions is the approximate fact that the right tails of the distributions are Pareto-shaped. A Pareto distribution for a variable x is characterized by a linear relation between the logarithm of x and the logarithm of one minus the cumulative distribution at x. The slope of this relationship is negative and defines, by its inverse, the thickness of the tail.[1] The wealth distribution has a significantly thicker right tail than the earnings distribution.

Table 21.1 also displays information on consumption and hours worked.

First of all, consumption is much less dispersed than earnings, income, and wealth.[2] The

[1]The cumulative function is $1-(x/\underline{x})^{\alpha}$, where $x \geq \underline{x}$, the minimum value for x, and $-\alpha$ is the slope referred to.

[2]This is in part a measurement issue: the distribution is trimmed at the top and the bottom percentiles,

Table 21.1: 2022 Per household shares of selected groups sorted by each variable.

	Bottom			Quintiles					Top		
	0-1	1-5	5-10	0-20	20-40	40-60	60-80	80-100	90-95	95-99	99-100
Earnings	0.00	0.00	0.00	−0.16	0.50	10	0.96	3.39	2.5	4.88	19.4
Income	0.00	0.08	0.12	0.14	0.30	0.50	0.82	3.24	2.5	4.45	22.4
Wealth	−0.2	−0.02	0	−0.01	0.05	0.19	0.50	4.70	2.48	6.48	35.1
Consumption				0.44	0.66	0.84	1.12	1.93			
Hours worked											
per household	0.07	1.05	2.25	9.31	12.64	16.84	23.54	37.66	8.18	8.04	7.65
per person	0.09	1.28	2.65	11.65	19.89	20.57	20.97	26.91	6.67	6.27	2.07

Note: Shares of the earnings, income, and wealth distribution in 2022. For earnings, income and wealth the source is Kuhn and Rios-Rull (2025). For consumption data from Table 1101, https://www.bls.gov/cex/tables/calendar-year/aggregate-group-share/cu-income-quintiles-before-taxes-2022.pdf.

low dispersion compared to wealth is also to be expected from standard permanent-income theory: consumption equals the flow equivalent of present-value earnings, which is the largest part for most people, plus the return on wealth. In addition, private insurance arrangements, as proposed by Krueger and Perri (2006), that are not measured in the data will also make consumption dispersion lower.

Hours worked are also much less dispersed than are earnings or wealth, especially as measured per person. Though we see that some people work many more hours than others (e.g., the 1-percent hardest working contribute almost 8 percent of the total working time), trying to see systematic patterns as to who works more and who works less is much more difficult. Figure 21.4 shows, for example, that across wage bins, the hours of work for men are very evenly distributed; these are *residualized* hours worked, i.e., some observables have been controlled for, namely, age, age squared, education, and race, as also discussed in Chapter 12.[3] That is, it is not that the most productive work significantly more, at least not if productivity is well approximated by wages. The figure also shows that the last two decades have seen a marked fall in the number of hours worked for the wage-poorest.

Given the limited dispersion in hours, and the fact that wages and hours are not strongly correlated, it follows—given the highly dispersed earnings—that wages exhibit significant dispersion. When discussing trends below, we will look at one measure of wage dispersion, namely, that between workers with different educational degrees, and its evolution over time. Wages of course differ within educational groups too, in part due to experience and the worker's age; in addition, there are observable characteristics like gender and race that also systematically influence wages. Whether the latter factors are due to discrimination of some sort is an important issue from a macroeconomic perspective: a society that does not allow all its individuals to flourish will under-perform in terms of efficiency. However, we do not discuss it further in this chapter.

as the Consumer Expenditure Survey does not include reliable measures at the extremes.

[3]The corresponding graph for women is very similar.

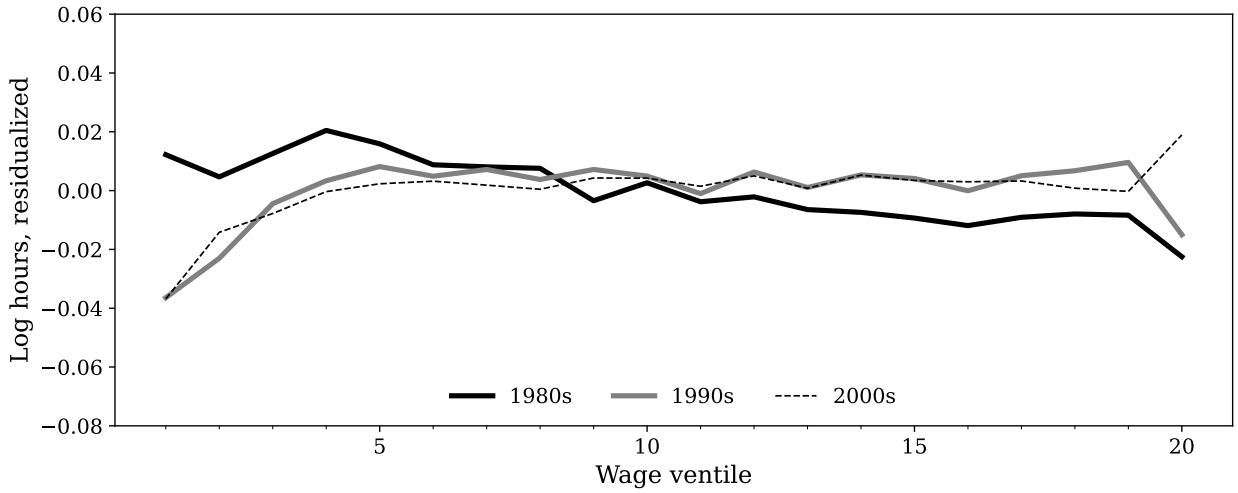

Figure 21.4: (Log) hours worked by wage ventiles.

Note: Ventiles are 5% bins. **Source**: Boppart et al. (2024).

We now look at the returns to capital—the per-dollar returns on the savings of households—and how they are dispersed in the population. We know that different assets give different returns and that these returns depend on their risk characteristics, as discussed in Chapter 16; for example, highly liquid savings give much lower average return than stock. Moreover, housing or land wealth delivers returns that are not financial—the services enjoyed by those who live in the house or use the land—but house and land prices also fluctuate significantly across locations and time.

At least for financial assets, one might have expected—from basic principles of portfolio management—that all consumers hold the same portfolios and enjoy the same returns on their portfolios, i.e., the return on "the market portfolio." Until relatively recently, little data was available on the portfolio holdings of individuals across the population. Fortunately, however, we now have data from countries where registry data (i.e., data for the whole population) is available, and increased efforts have also been made to unveil information implicit in capital incomes and from surveys. Thus, we now know more, and we know better. Figure 21.5 shows portfolio shares across the wealth distribution using US data. From the 25th until the 90th percentile, residential real estate is the biggest component of household assets. The further to the right in the distribution of wealth, the higher the share of risky financial wealth and, at the very top, private businesses.

The portfolio shares would perhaps be uninteresting if it were not for the fact that the different portfolios have different return characteristics. In Figure 21.6, which is a time-series average borrowed from Hubmer, Krusell, and Smith (2018), we display these return characteristics across different levels of wealth.[4]

[4]For any cell, we take the portfolio shares from Figure 2 of Bach, Calvet, and Sodini (2020), which is similar in spirit to Figure 21.5 here but using Swedish data, and apply a return to each component. The component derived from U.S. data, when available, and otherwise for the Swedish data from Bach et al. (2020). The U.S. data sources are Kartashova (2014), for public equity and businesses, Jordà, Knoll, Kuvshinov, Schularick, and Taylor (2019), for bonds, and the real estate return is based on the Case-Shiller index.

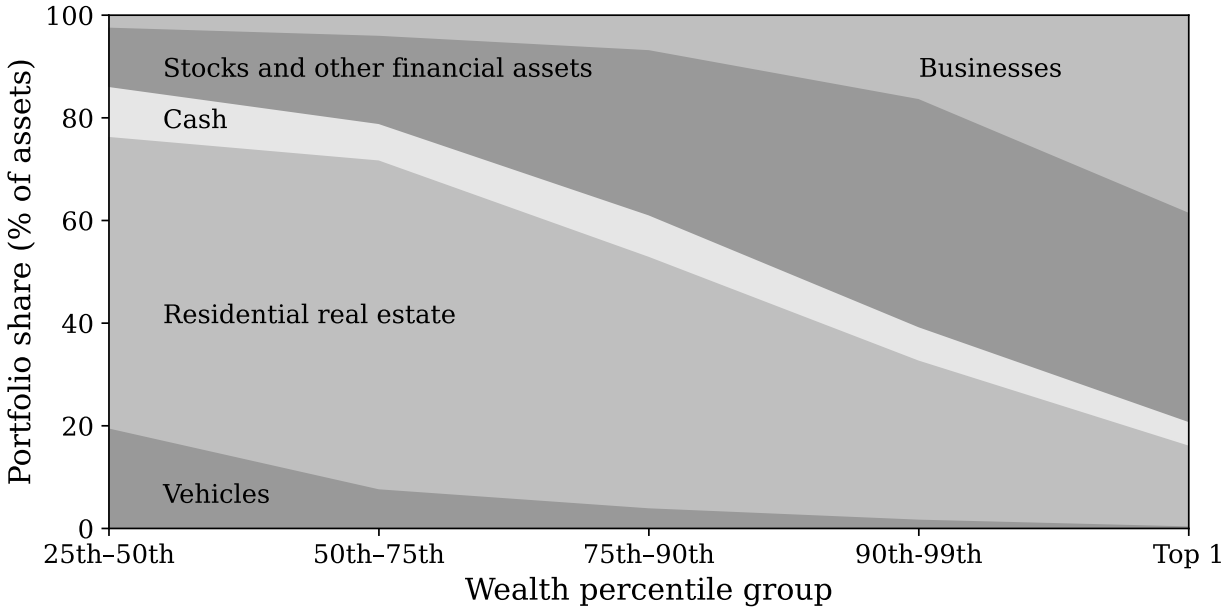

Figure 21.5: Portfolio shares from US Survey of Consumer Finance, 2022.

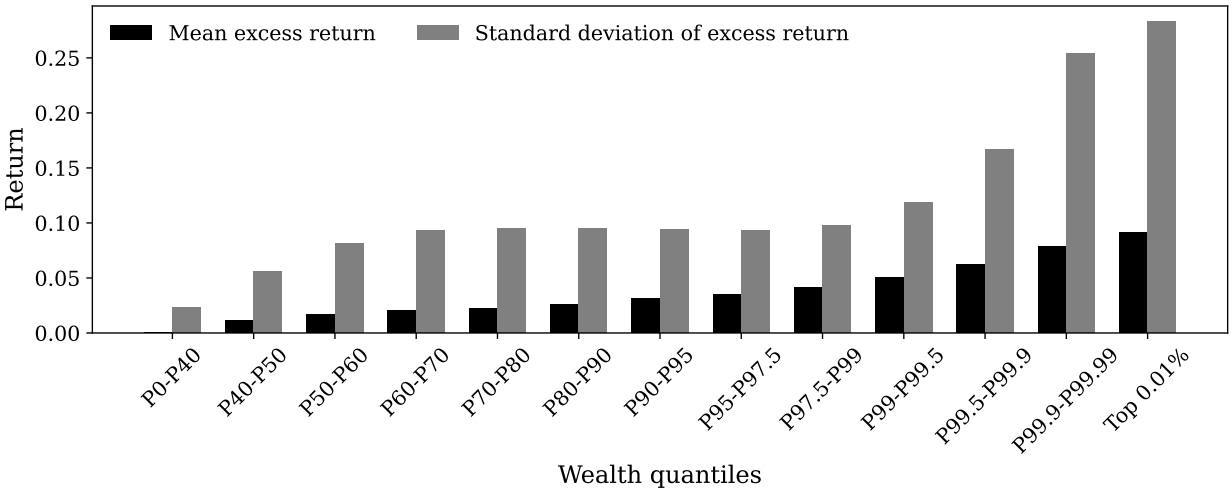

Figure 21.6: Mean and standard deviation for excess returns on portfolios.

The figure reveals strikingly large differences in mean returns—the darker bars—by wealth percentile: the higher up you are in the wealth rank, the higher is your return. Part of this comovement can of course be due to high returns causing high wealth, but an important part is surely just due to the systematic differences in returns across assets. Secondly, note that cash and deposits provide liquidity services by facilitating payments and these benefits are not captured in returns. Also, recall that the returns on housing do not include the housing service associated with owner occupancy. A second important feature of the figure is the significant standard deviation in returns and how, at the very top of wealth the distribution, the standard deviation shoots up significantly. Extremely high return out-

comes are a powerful source of wealth build-up, and we will discuss this mechanism in the theory section below.

Trends

We first look at the movements of capital and wealth relative to GDP (or income); these are relevant as capital and wealth are so concentrated among the wealthiest, whereas the large bulk of the population mostly rely on their flow income. Thus, Figure 21.7 shows that although the capital-GDP ratio has been very stable at around 3.5, there has been a marked upward trend in the wealth-to-income ratio, by around 50%. Note that wealth includes assets beyond capital, such as land and claims on the government.[5]

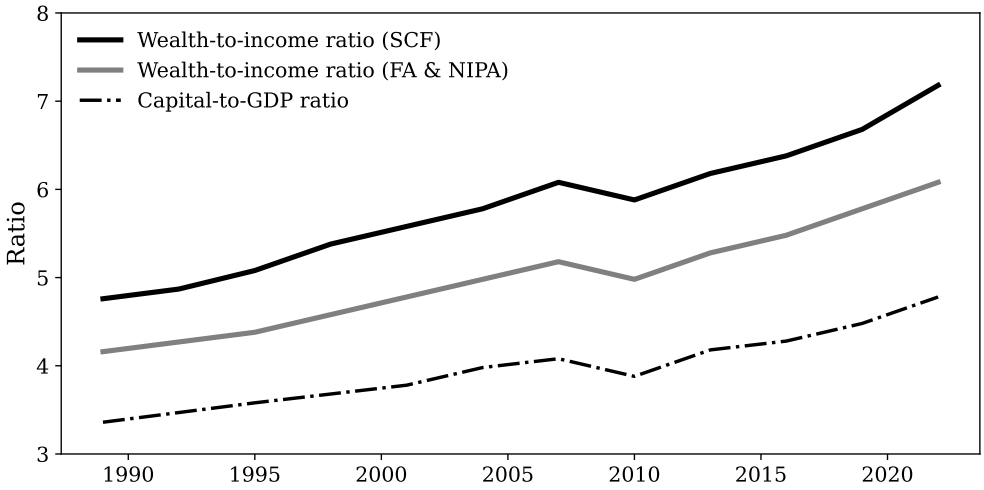

Figure 21.7: Wealth-to-income and capital stock-to-GDP ratios by SCF wave.

Note: Wealth-to-income ratio based on the SCF, wealth-to-income ratio using income from National Income and Product Accounts (NIPA), and wealth from Financial Accounts (FA), and capital stock-to-GDP ratio from Penn World Table 10.01. **Source**: Kuhn and Rios-Rull (2025).

Turning to the inequality within earnings and wealth, Figure 21.8 contains two panels and allows us to make a number of observations.

While average real earnings have risen—the solid line in the left panel of the figure—we see that the median and the 30th percentile have seen no real earnings growth since 1990. Thus, the top earners have had significant earnings growth, as evidenced by the 90th percentile earnings growth rising from 150,000 to over 200,000 over this period (in 2022 USD). That is, earnings inequality has grown. We will show one aspect of this increase in dispersion, the rising skill premium, just below.

In the right panel of Figure 21.8, we see that real wealth has more than doubled on average over the same time period, but again with very limited changes for the median or in the lower percentiles: the wealth buildup has occurred at the top, from around 800,000

[5]The figure shows the data from two sources, yielding a similar qualitative message but a slightly higher growth rate from the SCF survey measure.

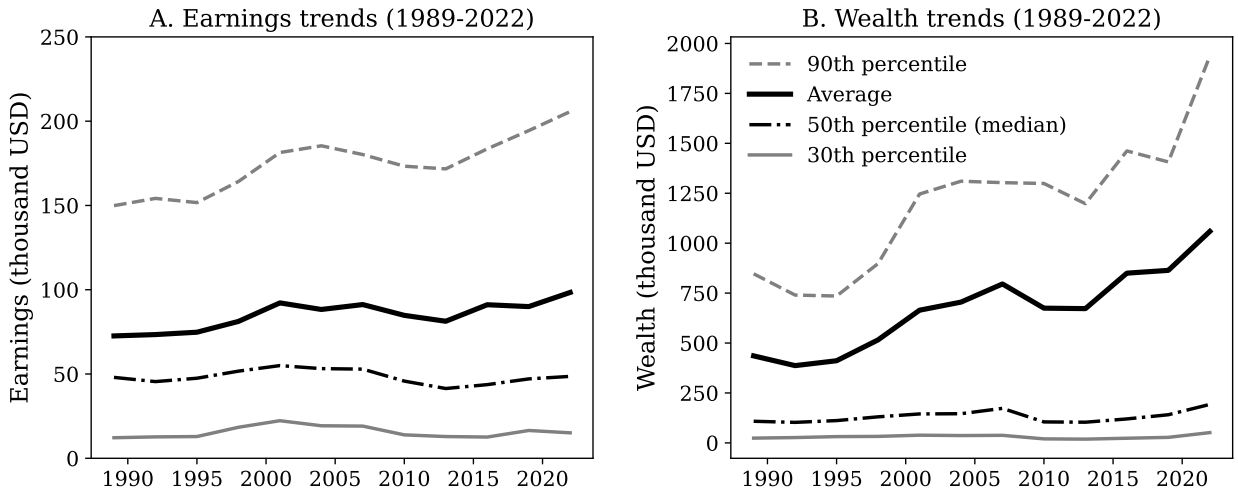

Figure 21.8: Inequality statistics for the evolution of earnings and wealth 1989–2022.

Note: These figures show the evolution of earnings and wealth across different percentiles of the distribution using data from the Survey of Consumer Finances (SCF). Panel (A) shows earnings trends and panel (B) shows wealth trends. All values are in thousands of 2022 USD.

to almost 2,000,000 (in 2022 USD). In sum, both earnings and wealth inequality have seen increases, mainly through rapid growth at the top of the distribution and none at the bottom.

Returning to the trends in earnings inequality, let us begin with trends in hours. Overall, hours are stable in the U.S. over the postwar period, but there are movements under the surface. An important determinant of hours worked is the extent of unemployment; data on that, revealing no long-run trend but large fluctuations over the business cycle, was discussed in Chapter 20. Another component is labor-force participation, and we displayed data on that in Chapter 12 and documented, in particular, that women's employment rate has risen dramatically. Today female labor force participation is only 10 percentage points below the value for males; in the middle of the 20th century, the difference was 50 percentage points. These factors, however, are minor in importance in comparison with the wage paid (on average) per hour: these differ drastically across the population and is the main reason why earnings are so highly dispersed. Figure 21.9 shows the development of the skill premium—the wage gap between those with a college degree and those without—over time.

In 1963, the premium was around 50%. The time series in the figure, normalized to be 1 in 1963, shows the growth since then. We see that the premium has risen by 50%, rather steadily since 1980, after a dip during the 1970s. This increase is particularly striking given that the number of college graduates has risen at a high rate during a number of decades during the second half of the twentieth century (but less so in recent decades). Of course there is also wide dispersion within the group with a college degree and within the one without; these dispersions have also increased over time.[6]. A source of wage dispersion that has gone the other way is the wage wedge between men and women (for otherwise

[6]A graduate student may be especially interested in learning that the graduate school premium also has risen; for an early study, see Eckstein and Nagypal (2004)

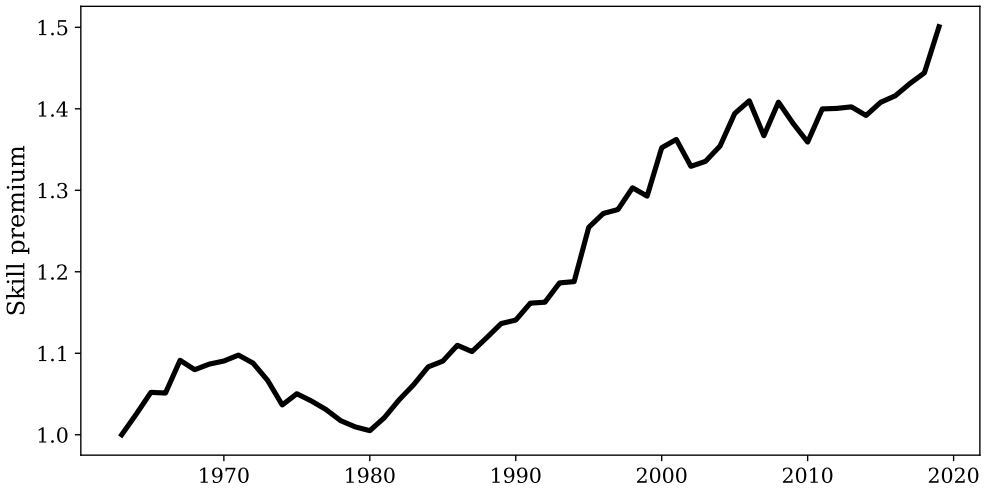

Figure 21.9: Evolution of the skill premium over time.

Source: Ohanian et al. (2023).

observationally equivalent workers); it is now less than 10 percent.[7]

21.2 Theory: macroeconomics and inequality

The present section and the next sections address the facts in the previous section using macroeconomic models. We first look at macroeconomic factors that determine important parts of the observed inequality. In particular, macroeconomic models are useful because the extent and shape of inequality are often affected by general-equilibrium interactions. Since the macroeconomic models we use have explicit microeconomic foundations, they also allow us to neatly separate partial- from general-equilibrium effects. Second, we look at how the presence of, and possible movements in, inequality affects macroeconomic aggregates.

We begin the present section by addressing some basic income inequality facts and then move to wage inequality. We finally discuss wealth inequality and, in that context, briefly also touch on inequality in hours worked.

21.2.1 The labor share and the capital-output ratio

From the perspective of the labor share mainly going to "workers" and the capital share to "capitalists", or "rentiers", it becomes relevant to study the behavior of these shares, as well as of the ratio of capital to output.

A first observation is then that our basic, neoclassical models have implications for the labor share of income. The most common assumption in the applied literature is that the aggregate production function is of the Cobb-Douglas variety, $A_t k_t^{\alpha} h_t^{1-\alpha}$. Here, TFP moves over time but the output elasticities with respect to capital and labor are constant, α and $1-\alpha$, respectively, and will equal the corresponding income shares under perfect competition.

[7] See, e.g., Blau and Kahn (2017). Men are still vastly over-represented among the very highest earners.

Therefore, the income shares are time-invariant, also off the balanced growth path. We saw, however, in Chapter 2 that the observed labor share has been declining in many countries in recent times. Can such a development occur in the neoclassical model?

Let us examine the labor share for a more general, constant-returns-to-scale production function $F(A_{k,t}k_t, A_{h,t}h_t)$, where the shape of F is time-invariant and technological change occurs via the capital- and labor-augmenting factors A_k and A_h. Under the assumption of perfectly competitive input markets, the share becomes

$$\frac{r_t k_t}{y_t} = \frac{A_{k,t}F_1\left(A_{k,t}k_t, A_{h,t}h_t\right)k_t}{F\left(A_{k,t}k_t, A_{h,t}h_t\right)} = \frac{F_1\left(1, \frac{A_{h,t}h_t}{A_{k,t}k_t}\right)}{F\left(1, \frac{A_{h,t}h_t}{A_{k,t}k_t}\right)},$$

where $F_i(\cdot,\cdot)$ is the partial derivative with respect to the ith argument, and we have used the homogeneity of degree 1 of F. Clearly, the share only depends on $(A_{h,t}h_t)/(A_{k,t}k_t)$, which is constant on a balanced growth path but off the balanced path it is changing and how the share then moves depends on the shape of F. Assume that it is a CES function, with elasticity ρ.[8] Then

$$\frac{r_t k_t}{y_t} = \frac{1}{1 + \left(\frac{1-\alpha}{\alpha}\right)^{\frac{1}{\rho}}\left(\frac{A_{h,t}h_t}{A_{k,t}k_t}\right)^{\frac{\rho-1}{\rho}}}.$$

We see that if $\rho > 1$, i.e., under stronger substitutability than Cobb-Douglas, a decrease in $(A_{h,t}h_t)/(A_{k,t}k_t)$ increases the capital share. A decrease in $(A_{h,t}h_t)/(A_{k,t}k_t)$, furthermore, is generated by the capital input growing faster than the labor input, which in turn results if there is investment-specific technological change, as described in Chapter 13. Hence, investment-specific technological change can cause a lower labor share; Karabarbounis and Neiman (2014) advance this theory, with an estimate of ρ slightly above 1 along with a decline in the relative price of capital.

The labor share can also be affected by departures from perfect competition, because firms' product-market power changes, affecting firms' margins,/ or because the power has shifted between employers and employees (monopsony vs. union power, respectively). We do not provide an analysis of these mechanisms here but they have attracted significant research over the last decade.[9]

Turning to k/y, the value of the stock of capital relative to output, we note that along a balanced growth path,

$$sy_t = (\delta + g)k_t \quad \Rightarrow \quad \frac{k_t}{y_t} = \frac{s}{\delta + g},$$

where g is the (labor-augmenting) growth rate, s is the saving rate, and δ the depreciation rate. Thus, a higher long-run growth rate makes k/y fall, but g is small relative to δ, at least based on historically recorded levels.[10] As for transitional dynamics, $k/y = k/F(k,h)$

[8] That is, we have $F(x,y) = \left(\alpha^{\frac{1}{\rho}}x^{\frac{\rho-1}{\rho}} + (1-\alpha)^{\frac{1}{\rho}}y^{\frac{\rho-1}{\rho}}\right)^{\frac{\rho}{\rho-1}}$.

[9] See, e.g., De Loecker, Eeckhout, and Unger (2020).

[10] Piketty (2014) formulates a so-called "Second Fundamental Law of Capitalism" where, based on a non-standard version of Solow's growth model where as g goes to 0, $k/(y-\delta k)$ goes to infinity. The non-standard aspect is in the formulation of saving: the saving rate is constant in net terms, i.e., net investment (the increase in the capital stock) is a constant fraction of net output (output net of capital depreciation). See Krusell and Smith (2014) for details.

is an increasing function of k, so k/y will tend to increase (decrease) toward its steady-state value if k is initially below (above) its steady-state value.

The k/y ratio can also be thought of as the wealth-to-output ratio in the standard neoclassical model, at least when wealth is thought of as an asset: the asset is k.[11] As we saw above, in the data, assets, broadly defined, have several components: land (a very large component), liquid, low-return assets (narrowly defined as cash and liquid deposits), bonds (some risky, some less so), and stock (both publicly traded and private).[12] As we saw in our introductory Chapter 2, the k/y is roughly 3 for annual data, but assets more broadly defined is at least twice that. From the perspective of inequality, the broader measure is interesting for several reasons. First, some of the additional assets are concentrated particularly among the wealthiest and therefore are very important for a full assessment of inequality. Second and relatedly, the prices, as well as the returns, on some of these assets are highly variable over time. Third, from our perspective here, general-equilibrium modeling of the broadest view on assets requires a model that goes beyond the simple neoclassical model: (i) it needs land and housing modeled explicitly; (ii) it needs a formulation of the value of firms that involves adjustment costs of capital, or non-competitive practices, producing something more like a Lucas tree model of firms, for which values and returns fluctuate greatly; and (iii) it needs a financial sector. Such a treatment goes well beyond the present text and, moreover is challenging to put together, since it also needs confront the asset pricing puzzles discussed in Chapter 16. Below, we will therefore treat return processes as exogenous, when they are discussed.

21.2.2 Wage inequality

We now turn to earnings inequality, in particular that driven by differences in wages. The labor market is highly complex both in terms of how the overall market works and, given the market structure and prices, how individual workers fare relative to each other. We will focus mainly on general-equilibrium effects here and on markets with perfect competition. Of course, the labor market is not characterized by perfect competition and we will briefly comment on departures from it at the end of the section, but we still believe that the perfect-competition perspective gives a very useful benchmark and will, in many cases, give highly relevant practical insights.

Skill-biased technological change

The literature on wage inequality gained momentum as a result of the significant increase in the "skill premium" that started in the mid-1970s, as first documented by Katz and Murphy (1992). The general idea in this literature is that, given an aggregate production function $F_t(k, u, s)$ where u is unskilled and s skilled labor and the t subscript on F denotes a general form of technical change, one simply computes the relative wage of skilled and unskilled workers as the ratio of partial derivatives, $F_{s,t}/F_{u,t}$, evaluated at the aggregate quantities of labor and of capital. The literature has made different assumptions on the shape of F and the way in which technological change enters, leading to different interpretations of the data.

[11] One can alternatively define wealth to incorporate human wealth, present and future.

[12] When computing net assets, of course, various forms of debt, such as mortgage loans, must be subtracted.

Katz and Murphy proposed an aggregate production function $F(k, G(A_u u, A_s s))$, i.e., a function where there is a sub-nesting G that only involves the two kinds of labor and where G has constant shape over time while technological change occurs through the As: skill-augmenting factors. The sub-nesting in particular means that the aggregate capital stock does not influence relative wages. With competitive input pricing, we obtain a skill premium

$$\frac{w_s}{w_u} = \frac{A_s G_s(A_u u, A_s s)}{A_u G_u(A_u u, A_s s)} = \frac{A_s}{A_u} \frac{G_s\left(1, \frac{A_s s}{A_u u}\right)}{G_u\left(1, \frac{A_s s}{A_u u}\right)}.$$

We see that both A_s/A_u—"skill-biased technological change"—and s/u—the relative quantities of the two inputs—matter. With G being a CES function, as Katz and Murphy further assume, one obtains

$$\log(w_s/w_u) = -\frac{1}{\rho}\log(s/u) + \frac{\rho - 1}{\rho}\log(A_s/A_u) + \text{ constant}.$$

As the relative supply of skilled workers increases, we see that the skill premium falls. We also see that skill-biased technological change makes the skill premium rise if and only if the substitution elasticity between the two skill types is above one. This is a standard result: when factor-augmenting technological change is biased toward one input factor, its relative productivity will fall if the two inputs have a sufficiently low degree of substitutability: due to the complementarity, the other factor becomes in higher "need". (With a Cobb-Douglas function, factor-augmenting technological change cannot affect the relative input prices as the direct productivity effect and the need effect cancel exactly.)

Katz and Murphy further write $A_s/A_u = (1+\gamma)^t$, i.e., interpreting there to be a constant rate of skill-biased technological change. With $\log(A_s/A_u) = t\log(1+\gamma)$, they then run the regression

$$\log(w_s/w_u) = -\beta_1 \log(s/u) + \beta_2 t + \text{ constant} + \epsilon.$$

Here, $\hat{\beta}_1$ allows us to back out an estimate of the substitution elasticity ρ and, based on this estimate, $\hat{\beta}_2$ can then be used to estimate of the rate of skill-biased technological change. To implement this procedure, one needs to define the notion of skill; Katz and Murphy use educational groups and add up to two labor inputs, one with college graduates and one with the remaining workers.[13] The supply of high-skilled workers increased quite fast during the period under study, pushing the high-skilled wages down. The regression estimate then implies a ρ around 1.4 and a counteracting demand force of around 10 percent per year, i.e., a very rapid rate of skill-biased technological change.[14]

Endogenous skill-biased technological change

The hypothesis of skill-biased technological change can be explored further, in part by trying to understand its determinants. One idea is that of *endogenous, directed technological change*,

[13] They use CPS data 1967–1987 and create education cells, each associated with an average wage within the cell. Then aggregation over cells uses these wages as weights, thus delivering a notion of efficiency units of labor input for skilled and unskilled labor, respectively.

[14] The value 1.4 was in line with the pre-existing literature of a "$\sqrt{2}$"-rule of thumb. However, Katz and Murphy also try other values for ρ and back out a skill-biased technology path year by year, yielding similar qualitative conclusions.

where the input-augmenting technology levels are derived as a function of changes in the environment. One such change is the relative supplies of labor themselves. Acemoglu (1998) proposes such a theory, along the lines of the endogenous growth literature discussed in Chapter 13 where purposeful R&D develops patents that are used to improve the productivity of skilled workers. This group having become larger thus constitutes the motivation for patent developers, who can then increase their profits by targeting that growing market. A simple version of this idea can be implemented as follows. Suppose the overall technology in society is given by

$$\max_{\{A_s,A_u\}} \left[(A_s s)^\sigma + (A_u u)^\sigma\right]^{1/\sigma} \text{ s.t. } \left[\lambda A_s^\phi + (1-\lambda) A_u^\phi\right]^{1/\phi} = 1,$$

where $\phi > 1$ to ensure an interior solution. The constraint describes the choice of the direction of technology: it builds in a trade-off between A_s and A_u, and this can be thought of as a reduced form of having to allocate a given amount of researchers to two different activities. Here, the elasticity of substitution between the two inputs for *given* input-saving technologies, ρ, equals $1/(1-\sigma)$. The first-order condition becomes

$$\frac{A_s}{A_u} = \frac{\lambda}{1-\lambda}\left(\frac{s}{u}\right)^{\sigma/(\phi-\sigma)}.$$

So if $\sigma \in (0,1)$ (more substitutability than Cobb-Douglas), this formulation delivers that a higher relative skill supply attracts more R&D. This means that a higher relative supply of skilled labor will, through endogenous directed technology, lead to a counteracting positive effect on the skill premium. Can this effect overturn the direct ("neoclassical") effect?

Substitute the expression for relative efficiencies into formula for relative wage (MPL_s/MPL_u) to obtain

$$\log\left(\frac{w_s}{w_u}\right) \propto \frac{\sigma - \phi(1-\sigma)}{\phi - \sigma} \log\left(\frac{s}{u}\right).$$

We see that the skill premium is *increasing* in $\frac{s}{u}$ as long as $\rho - 1 = \frac{\sigma}{1-\sigma} > \phi > 1$. I.e., with a large enough elasticity of substitution between skilled and unskilled labor (at least above 2), an increase in the relative supply of skilled labor, such as during a college attendance boom, can in fact increase the relative wage of college graduates.

Capital-skill complementarity

A related hypothesis is that skilled labor and unskilled labor play distinct roles in production and interact with capital in different ways. The hypothesis that goes back to Griliches (1969), who noted systematic differences in the average level of education and capital intensity across industries. Thus, the behavior of the skill premium over time could potentially be linked to an observable, namely, the behavior of capital, rather than be measured residually. Krusell, Ohanian, Ríos-Rull, and Violante (2000) pointed out that there had been especially fast investment-specific technological change (see Chapter 13) during the second half of the period studied by Katz and Murphy, precisely when the wage skill premium started rising.[15]

[15] Their study focused on equipment capital, as opposed to structures. For simplicity, we use a general notion of capital here.

These ideas can be formalized by adopting a slightly different production function, nesting capital asymmetrically with the two types of labor: conceptually, $F(u, G(k, s))$ allows us to use a higher substitutability in the G nest than in the F nest. Thus, rapid growth in k can make the relative wage of skilled labor go up without technology factors playing a role.

To understand the logic, consider the simple example $F(k, u, s) = u + s^{\nu}k^{1-\nu}$. Clearly, higher capital raises the marginal product of skilled labor, whereas the unskilled marginal product is unaffected (and equal to 1). Generalizing to both F and G being of the CES kind, one can show that if the elasticity in F is greater than that in G, then the skill premium rises as k goes up. The account provided by Krusell et al. (2000) does suggest capital-skill complementarity has played an important role.[16] By this account, where the growth of equipment capital raises the productivity of skilled labor more than that of unskilled labor, we are closer to direct measurement of the source behind skill-biased technological change.

A different nesting, $F(s, G(k, u))$, allows a similar interpretation but some different qualitative properties. In particular, a rise in the capital stock now allows the real wage of unskilled workers, w_u, to fall (e.g., consider $s^{\nu}(k + u)^{1-\nu}$). This feature, which cannot be obtained under the previous nesting, is in line with the stagnant wages of parts of the population over several decades beginning in the 1970s.

Finally, capital has also been argued to be especially complementary with human capital in times of rapid technological progress, as in Greenwood and Yorukoglu (1997). The idea here is that human capital increases a worker's ability to adapt to new circumstances.

Human capital accumulation

The previous sections emphasize skill differences as important determinants of wages, in particular as measured by formal education. More generally, we conceptualize "human capital" as a key determinant of wages. What is human capital, and how is it accumulated?

There is a vast literature covering many aspects of human capital relevant for labor markets. First, human capital can be a thought of as a general skill, useful in many tasks, occupations, and industries in generating more output per unit of time. Such a skill can be accumulated during education as well as while working. But skills can also be quite specific and not easily transferable across occupations or jobs. More generally, human capital is multi-dimensional, in which case it is natural that workers sort across jobs and tasks. The Roy model (Roy, 1951) expresses this clearly: each individual has a skill vector describing the amounts of different types of skills, and different jobs give different returns to the different skill types; the model then describes, given a distribution of individuals across skill vectors, how individuals sort into jobs to maximize their income given their skill combinations. Furthermore, even when skills are one-dimensional there can be a non-trivial pattern by which individuals match to different job features; for example, the skill can be complementary with capital, whereby markets would push toward high-skilled individuals matching with firms that have advanced capital equipment. This hypothesis is closely related to the capital-skill complementarity hypothesis earlier in this section.[17]

[16] Ohanian et al. (2023), with recent data, estimate the higher elasticity to be slightly below 2 and the lower one slightly above 0.5.

[17] High-skilled individuals can also be complementary with other high-skilled individuals, such as in the O-ring theory of Kremer (1993).

There is also a rich set of models of the accumulation of human capital. These models are typically framed in a partial-equilibrium setting but are often imported into macroeconomic general-equilibrium settings. A basic model of education was provided in Mincer (1974), where the key choice is how many years of schooling to obtain. Under some conditions, the framework delivers the well-known Mincer equation. This equation characterizes the log wage of an individual as linear in the number of years of schooling and in the years of work experience; when applied econometrically, it delivers a "return to schooling" (the coefficient on the years of schooling) of around 0.1 in a large number of contexts. Ben-Porath (1967) gives us a basic, more general model of human capital accumulation, where one's time can be divided into working or human capital accumulation (education or training). The framework naturally gives rise to a specialization on human capital accumulation early on and a specialization on working only later in life, when investments in human capital no longer pay off given the short remaining working life; midlife, there is an interior solution with both work and training.

Both the models of human capital accumulation just noted are particularly simple in that they assume a fixed amount of time available (for either work or study), along with perfect credit markets. This means that the objective function can be phrased as present-value lifetime income: there is a separation between this problem and the smoothing of consumption over time. If credit markets are not perfect, the human capital accumulation decision, as discussed briefly in Section 21.3.1 below, becomes intertwined with the consumption decision. Hence, the individual's wealth level matters, and a worker can then make decisions that appear suboptimal in a present-value sense—e.g., they fail to educate themselves or they take jobs involving no training—because they lack liquidity and need cash for consumption purposes; even low-skill traps are conceivable. See Griffy (2021) for a recent study.

Task-based models

The approach based on aggregate production functions skips the many details of actual production processes and how different individuals sort into different occupations and tasks to be performed across the different sectors of the economy. A recent line of research offers an alternative abstraction: a production process is defined by a set of tasks to be performed, and the tasks are all complementary (see, e.g., Acemoglu and Autor, 2011). Then different inputs, such as labor of different skill types and capital, have different abilities to produce different tasks. For each task, the different inputs are substitutable; in the simplest case, they are perfect substitutes, with coefficients that differ by task and input. In a competitive environment, the different inputs are then allocated optimally to different tasks. A full model description and solution is beyond the scope here; suffice it to say that there are specific assumptions on the task/input coefficients such that one can derive a closed-form expression for overall production as a function of the amount of inputs (capital and the different labor inputs). Thus, an aggregate production function is derived endogenously and its properties depend on the underlying task/input assumptions.

The task model is appealing in that it can give a concrete expression for concepts such as "automation". A prime example is the "hollowing out" of the employment distribution in the U.S. during the last century's last decades, when many middle-manager tasks, according

to a common account, were automated.[18] That is, specific worker skills were made obsolete. A topic of current interests is the adoption of robots: a large number of tasks currently performed by labor may instead performed by capital (say, because capital has become more abundant, or because the task/input features changed). Again, it is a worker's type of skill that determines whether it is easy to replace by machines or not. The task model also allows one to analyze *outsourcing*, i.e., the idea that some tasks in a production process is performed by workers abroad, where wages may be lower. When the model is applied empirically, one needs to identify tasks and usually, the O*NET data—Occupational Information Network (a U.S. government initiative), linking skills to job requirements—is then used; the task descriptions "abstract", "routine", and "manual" are then mapped into how easily they can be performed by different labor (or capital) inputs.

Labor markets in practice

The above discussion is based on aggregate production functions and the assumption that wages equal marginal products. In practice, labor markets of course have many features that influence wages through other mechanisms. We now briefly mention examples of such mechanisms.

Search frictions In Chapter 20, we saw that wages for identical workers can differ due to search frictions. It is hard to assess the importance of this channel as the notion of "identical workers" receiving different wages is hard to implement in practice. Measures based on "wage residuals", i.e., wage regressions using all available observables still deliver large residual dispersion, but personal characteristics such as "diligence" or "ability to cooperate" may still differ greatly within the set of individuals with the same observable characteristics. Significant wage dispersion due to search frictions alone also raise the issue of why some employers pay so much more than others.

Compensating differentials Another important feature of labor markets is that workers value other aspects of jobs than just the earnings: they take into account the *amenities* of different jobs. Today, a large number of job amenities are in principle available for many employees (just google "job amenity" and you will see a large number of examples!), who then may accept jobs at lower wages than otherwise; the possibility to work at home is one that is very popular today but only seems to have become common after the Covid pandemic. Similarly, some jobs have negative amenities like lack of job safety or highly irregular work hours, but what appears like a plus to some workers may be a minus to others, so there is also selection to take into account.[19] From a welfare perspective, we note that wage dispersion due to the amenity heterogeneity of jobs is a natural, and arguably desirable, feature: it would seem undesirable to have equal wages at two jobs that are identical but differ markedly in the attractiveness of their amenities.

[18] See, e.g., Acemoglu and Autor (2011).

[19] As increasing amounts of micro data on job descriptions has become available recently, empirical research on amenities is currently very active.

Monopsony and union power Another currently very active research area is to assess the role of monopsony power in labor markets: firms' abilities to lower wages below marginal products because, for one reason or another, it is costly for workers to change jobs. Like the presence of job amenities, monopsony power is multi-faceted and hard to measure; the key from the perspective of understanding wage inequality is how different firms benefit from different degrees of monopsony power and the observation of rising firm concentration over the last several decades has directed the attention of researchers to this possibility.

Another departure from wages equaling marginal products is made possible by workers having a degree of market power vis-à-vis firms. Unions are a central component of labor markets in Europe, while much less so in the U.S. today; in the U.S., union membership peaked in the 1950s at a little over 30% of workers being unionized, whereas the number today is below 10%. Moreover, in Europe also many non-unionized workers are covered by so-called collective agreements, i.e., contracts negotiated between unions and employer federations stipulating a range of features of the labor contract, including wage floors and amenities. Changes in the degree of influence that unions exert on labor markets are broadly viewed to be important determinants of wages and wage inequality, though hard evidence in the form of controlled experiments is only scant. Before the search and matching theory became the dominant framework for analyzing labor markets and unemployment, theories based on union wage determination were in focus. Today, we see a certain resurgence of such theory, at least partly because we are now able to access more data on wage contracts and union membership.

21.2.3 Wealth inequality

In this section we will look at a number of determinants of wealth inequality. We will focus on the long run, i.e., we will use a number of models evaluated at their steady states as a way of relating to the data described in the earlier sections.[20] This seems reasonable as the wealth distribution has certain properties—it is highly skewed and it is significantly more dispersed than the earnings distribution—that hold true over time and across countries. There has been significant research in macroeconomics over the last decades aiming at accounting for this feature of the data; as we shall see, it is a challenging problem.

Recall from above that the relative shares of capital and labor income can be analyzed straightforwardly using our aggregate production function. Here, however, the focus is on differences in capital holdings between households. In the most commonly used model of wealth inequality today—the heterogeneous-agent model—consumers are subject to idiosyncratic, uninsurable shocks, and they accumulate and decumulate their asset position in part to self-insure against shocks. For a thorough understanding of this class of models, however, it is important to first understand how a model without shocks works. We therefore first look at deterministic models. We will begin with the simplest possible model and then gradually introduce more elements.

[20] How wealth inequality varies over time, out of steady state and in response to exogenous aggregate shocks, is not addressed in this section but are briefly commented on below.

Deterministic models

We begin with the frictionless dynastic model, and we will see that this model has very particular long-run predictions for asset inequality: any long-run wealth distribution is possible, and which one will occur is entirely dependent on the time-zero wealth distribution. We then discuss various extensions that can potentially break this result.

The benchmark frictionless model We assume that all consumers have the same preferences but begin with different wealth levels. They also have different earnings in the form of different efficiency units of labor; labor supply is assumed to be inelastically supplied (the measure of consumers is 1 and each consumer supplies 1 labor unit). There is no uncertainty. In a steady state, we thus have consumer i maximizing

$$\sum_{t=0}^{\infty} \beta^t u(c_{i,t})$$

subject to

$$c_{i,t} + a_{i,t+1} = (1 + r - \delta)a_{i,t} + \epsilon_i w$$

for all t.[21] Here, ϵ measures the number of efficiency units of labor and we assume that $\sum_i \epsilon_i = 1$. Hence, total labor input also equals 1. We assume a standard neoclassical model, for simplicity without growth, with a perfectly competitive firm sector. Aggregate capital at t is given by $k_t = \sum_i a_{i,t}$ at all times given that we assume a closed economy and as capital is the only asset in positive net supply.

The steady-state level of capital is found by evaluating the consumers' Euler equations. They all face the same interest rate and they have the same discount factor so, given $c_{i,t} = c_{i,t+1}$ for all i, the Euler equation is identical and satisfied for all consumers when $\beta(1+r-\delta) = 1$. The firm's first-order conditions for profit maximization deliver r and w as a function of aggregate capital, which allows us to pin down the steady-state level of capital, k. This, in turn determines w as well. What remains is to find the distribution of consumption and asset levels for consumers. The only remaining equations to use are consumers' budget equations, which read $c_i = (r - \delta)a_i + \epsilon_i w$ for each i. We note that for each agent there is one equation but two unknowns, c_i and a_i; hence any combination is possible, so long as $c_i \geq 0$. We thus have steady-state *indeterminacy* in the wealth and consumption distributions. In particular, labor earnings are not connected to wealth or consumption.

To understand why we obtain indeterminacy, recall that we demonstrated at various points earlier in the text, that we have a standard permanent-income setting: the consumer maintains the initial asset level forever, thus consuming earnings and the net interest income off of the wealth. Higher initial wealth thus simply produces higher consumption, with a marginal propensity $r - \delta$.

We can imagine that consumers also receive shocks to earnings. Below we will look at the effects of such shocks when they are not fully insurable. But assume now that they are fully insurable and idiosyncratic, without any associated aggregate risk. Then steady-state indeterminacy would obtain again: whatever it is, the initial wealth distribution stays constant

[21] We also assume the standard no-Ponzi-game restriction.

over time, provided it sums to steady-state k, and the consumption distribution follows mechanically. Budget constraints are more complicated due to the presence of insurance instruments but consumption remains constant and insulated from shocks.[22]

Departures from a dynastic setting Consider now an overlapping-generations model, where agents live for a finite number of periods and do not give bequests (because they do not value their offspring). We know from earlier in the book that the steady-state interest rate is nontrivially pinned down; its value depends on the life-time earnings profiles of agents and their implied savings needs given their desire to smooth consumption and, in the presence of capital accumulation and production, in conjunction with the properties of the production function. All consumers save zero in the final period and if they all start with zero wealth (which is natural given the absence of bequests), the relative wealth holdings during their lifetimes will depend on the timing and size of their earnings; agents with high earnings that occur early save a lot and become the wealthiest, but if these agents's earnings occur late then they will borrow and become the wealth-poorest. In sum, the wealth distribution will be determinate and will reflect the earning profiles of agents during their lifetimes.

Suppose now that, in the same overlapping-generations model, we add altruism: each agent values their offspring, as modeled by an additional term in utility multiplied by $\beta < 1$, thus assigning a smaller value for children. With two-period-lived agents and life utility $U(c_{y,t}, c_{o,t+1})$ for an agent who is young at t, we obtain an indirect utility function

$$V(a_t) = \max_{c_{y,t}, c_{o,t+1}, a_{t+1}} U(c_{y,t}, c_{o,t+1}) + \beta V(a_{t+1}) \quad \text{subject to}$$

$$c_{y,t} + \frac{c_{o,t+1}}{1+r-\delta} + a_{t+1} = a_t(1+r-\delta) + w(\epsilon_y + \frac{1}{1+r-\delta}\epsilon_o),$$

where we find it convenient to define the bequest from generation t to generation $t+1$ as a_{t+1}, an amount given at t already and received at $t+1$, and where we have assumed steady state and that all generations have the same earnings profile. This problem can be rephrased as

$$V(a_t) = \max_{R_{y,t}, a_{t+1}} u(R_{y,t}) + \beta V(a_{t+1}) \quad \text{subject to}$$

$$R_{y,t} + a_{t+1} = a_t(1+r-\delta) + w\hat{\epsilon},$$

where $u(R) \equiv \max_{c_y, c_o} U(c_y, c_o)$ subject to $c_y + c_o/(1+r-\delta) = R$ and $\hat{\epsilon} = (\epsilon_y + \frac{1}{1+r-\delta}\epsilon_o)$. This is now an entirely standard permanent-income model, and hence any conclusions we obtained for the dynastic household apply here as well. In particular, the steady-state wealth distribution is indeterminate. We thus see that the assumption of altruism plays a crucial role.

There are also other formulations that involve bequests. One is the "warm-glow" setting, where an agent receives utility from giving bequests—a joy of giving—according to an exogenous function; let us denote it by $\hat{V}$. Notice that this is in sharp contrast with V above, which is endogenous and would change, say, if the returns to saving or policy changed. V would also change in response to differences in earnings across cohorts; earnings-rich parents

[22] If the initial capital stock is not at steady state, the initial asset distribution still determines the long-run distribution but it evolves non-trivially over time. A full characterization is available in Chatterjee (1994).

might want to give more bequests to earnings-poor children, for example. Thus, the warm glow formulation is non-standard in that it simply enters an action—an amount of saving—into the utility function as a primitive, unlike a consumption good where prices influence demand.[23]

Discount factor heterogeneity

Still for a frictionless framework, suppose consumer i has a discount factor β_i, with $\beta_i \neq \beta_{i'}$ for some (i, i'). Then the long-run wealth distribution is determinate and very special. First of all, the consumer (i^*) with the highest discount factor has constant consumption and the capital stock is determined by $\beta_{i^*}(1 + r - \delta) = 1$, while all other consumers' consumption levels are converging to zero, i.e., they do not have constant consumption. Thus, all the other consumers have their wealth levels given by $0 = (r - \delta)a_i + \epsilon_i w$, i.e., a negative level such that consumption equals zero.

Intuitively, more patient consumers save more and, in the limit, though convergence will be slow to the extent that the differences between discount factors are small, the most patient consumer consumes the economy's entire output.[24] We also note that the extreme long-run inequality is a result of choice. Thus, it is hard to argue that it is "unfair": those who consume very little later on have simply chosen to consume more early. From this perspective, wealth inequality per se should not be seen as undesirable.

An assumption of exogenous and permanently different discount factors—or permanent differences in other preference parameters—is arguably not an appropriate assumption for a dynastic model. A reasonable alternative is to allow randomness that can capture how preferences are rather stable during a person's lifetime but less so between parents and children.

Distortions, credit-market restrictions, and imperfect asset markets

We now briefly look at deterministic environments with distortions or restrictions on choice. Beginning with distortions, suppose there is a tax on capital income (net of costs of capital) that is progressive: gross capital income is $a\,(1 + (r - \delta)(1 - \tau(a)))$, where $\tau(a)$ is the tax rate schedule as a function of the amount saved. So suppose $\tau'(a)$ is positive and strictly increasing. Then the Euler equation, which now involves $\tau(a)$ and $\tau'(a)$, cannot be met for a constant (steady-state) consumption path at the same time for all agents, to the extent they

[23] As a non-generic outcome, it is possible that $\hat{V}$ and V coincide for two economies that are otherwise identical. However, a change in policy or, say, earnings structures, would change V but not $\hat{V}$, and they would no longer coincide. In particular, the warm-glow model does not map into Arrow-Debreu, and hence standard welfare theorems do not apply; see Chapter 6 for an example with dynamic inefficiency. Consequently, what is assumed about $\hat{V}$ becomes critical in terms of the long-run implications for wealth inequality. In particular, what is important is the relative curvature of $\hat{V}$ relative to the parent's utility function of consumption. Taking prices as given, a relatively less curved $\hat{V}$ translates into higher future wealth inequality given current wealth inequality; however, in general equilibrium prices may counteract this initial effect.

[24] It is convenient for these reasons to assume that there is a large number of consumers with each discount factor; otherwise the assumption of price-taking is not appropriate. Within each discount factor group, then, the wealth distribution is still indeterminate.

have different levels of a.[25] In this sense, the situation is quite like that under discount-rate heterogeneity. However, the outcome here is the reverse, i.e., a wealth distribution that converges to full equality over time: in steady state, $a_i = a_{i'}$ for all i and i'. Intuitively, progressive taxes on capital income simply slowly eat away the wealth of savers, and the more so the higher the level of savings. Proportional taxes on capital income also lower saving, but not differentially across saving levels.

The case of a progressive tax on capital income, or one on wealth, is likely important in practice for understanding the long-run evolution of inequality. Hubmer et al. (2018) in particular argue that Reagan's tax cuts and decreases in the degree of tax progressivity, which subsequently have not been reversed, constitute a key factor behind the slow further build-up in wealth concentration.

Turning to credit-market restrictions, we shall see that these can, when they restrict borrowing for consumption, lower asset inequality. Consider a setting with two representative dynastic consumers, A and B, equal in numbers, where consumer A has endowment $\bar{y}$ in even periods and $\underline{y} < \bar{y}$ in odd periods; for consumer B the situation is the reverse (high endowment in odd periods, low endowment in even periods). Each agent's budget constraint is $c_t + q_t a_{t+1} = y_t + a_t$. Total consumption equals total endowments in each period; there is no production. With a standard utility function $\sum_{t=0}^{\infty} \beta^t u(c_t)$ for both agents, where u is strictly concave, let us find the equilibrium interest rate and the resulting pattern of borrowing and lending. First, then when there are no restrictions on borrowing, there will be full consumption smoothing and $q = \beta$ (the gross interest rate will equal $1/\beta$). This is achieved by the currently endowment-rich agent lending to the endowment-poor agent each period, so that each consumer is a borrower one period and a lender the next period.[26] This outcome can be prevented by the existence of borrowing constraints. Suppose $a_{t+1} \geq \underline{a}$ at all times, with $\underline{a} < 0$ close enough to zero that the constraint binds at all times. Then in equilibrium, consumers are not able to fully smooth, because they are not allowed to let their assets move freely. That is, asset inequality is a sign of an imperfect asset market, and consumers are worse off. The interest rate from periods 1 and on will be determined by

$$qu'(\bar{y} + \underline{a}(1+q)) = \beta u'(\underline{y} - \underline{a}(1+q)),$$

which is one equation and one unknown, given that $\underline{a}$ is exogenous.[27]

In the data, as we have seen above, there is a significant fraction of households with negative financial asset holdings. This is in part due to consumption loans, in a manner similar to that just described, but it can also be due to borrowing to fund investment (e.g., in housing, via mortgage loans). In the case of investment loans, restrictions would not affect asset inequality much; they would simply distort the portfolio choice, which in turn can have real consequences, such as for how and where to live.

[25] The gross return on saving now equals $1+(r-\delta)(1-\tau(a)-\tau'(a)/a)$. Thus, in steady state this expression times β must equal 1.

[26] The exact amount of borrowing depends on the present-value wealth for the two consumers, which in turn depends on the time-zero asset position. If $a_0 = 0$ for both agents, then agent A is richer in present-value terms since they receive the high endowment first, at time 0, and will therefore permanently have a somewhat higher consumption that agent B.

[27] In the very first period, the interest rate is different if the initial asset position is $a_0 = 0$ for both agent.

Finally, a potentially very important determinant of wealth inequality is in place to the extent that different consumers simply obtain different returns on their wealth in asset markets. This phenomenon is challenging to fully understand without market imperfections, asymmetric information, commitment problems, or behavioral components. At this stage, the macroeconomic literature has only begun to explore this channel, as there is no quantitative, off-the-shelf model of return heterogeneity. At the same time, with increased access to individual data on asset holdings and asset returns, we see that return heterogeneity is quantitatively significant. Some of this heterogeneity is due to different households holding different kinds of portfolios (e.g., housing, liquid savings in the form of bank deposits, publicly traded stock, private equity, or cryptocurrency) but some is due to heterogeneity within asset class; for example, investments in individual stock, as opposed to a market index fund, is commonplace and, clearly, contribute to wealth inequality. Clearly, those who invested early in cryptocurrency made a rapid climb upward the wealth distribution and, more generally, rapid movements in a household's relative wealth position are often explained through risky asset investments.

The simplest model of return differences in terms of the model above is to assume, ad hoc, that different consumers receive permanently different deterministic returns $r_i - \delta$ on saving, where i denotes a specific consumer. One can close the model by assuming that $\sum_i r_i a_i = r \sum_i a$, where r is the marginal product of capital in production, evaluated at $k = \sum_i a_i$.[28] Simple inspection of the Euler equations of consumers tells us that the agent with the highest r_i must hold the entire capital stock plus an amount of lending to the remaining consumers, who have zero consumption and negative asset holdings.

Another, more structural description of return heterogeneity is an environment where an "entrepreneur" has access to a high-returning investment project but credit markets are restricted, along the lines of Chapter 19. If markets worked perfectly—if the entrepreneur could obtain funds at a frictionless credit market—then others could share in obtaining the high return and hence the project would just be part of the overall production possibility set of the economy; there would be no sense in which the entrepreneur could obtain a higher return on saving than anyone else. So imagine instead that such capitalization of the investment project were not possible but instead the entrepreneur had to self-finance, provided a minimum investment amount is attained from the investor's own saving. Such a setting, which was analyzed by Quadrini (2000), generates high wealth accumulation among those with enough money and opportunities to invest. Which projects are funded by banks, by private equity, by bond finance, and through public stock exchanges, varies greatly across time and countries; the point here is that the efficiency and specific nature of these markets matters for wealth inequality.

Models with idiosyncratic, uninsurable shocks

We now turn to the class of heterogeneous-agent models of wealth inequality that also form a new core of much of modern macroeconomic analysis. We begin with the simplest such

[28] For the summing up of income to work out, the r_is need to be endogenously connected to the asset distribution, as the sum of capital incomes needs to match $r \sum_i a_i$; clearly, one needs to understand just how the return differences materialize in order for a complete, general-equilibrium understanding of this phenomenon.

framework and then discuss extensions.

Idiosyncratic earnings shocks and precautionary saving If earnings shocks are idiosyncratic but not insurable, except for the possibility of saving in a riskfree asset, then we have the settings analyzed by Aiyagari (1994) and Huggett (1993) and described in Chapter 11. The first paper in this literature, however, was Imrohoroğlu (1989): she demonstrated how business cycles affect the welfare costs of business cycles through the effects on consumption inequality.[29]

Relative to the results above, in the present section we note that the presence of uninsurable, idiosyncratic shocks makes the steady-state distribution of wealth determinate. This is most easily seen using the Huggett version of the model, where we recall that the budget constraint reads

$$c_t + q_t a_{t+1} = \epsilon_t + a_t,$$

where ϵ is idiosyncratic and random and there is also a borrowing constraint: $a_{t+1} \geq \underline{a}$. There is no production or aggregate storage, so assets always sum to zero in the population. Given the possibility of borrowing, some agents will want to borrow and others lend, so long as ϵ is mean-reverting, such as an AR(1) process with autocorrelation strictly less than 1. We saw the model applied to a special case in Chapter 17: the case of $\underline{a} = 0$, where borrowing is not allowed and the equilibrium is autarky, with a real interest rate determined in closed form based on the Euler equation for the non-constrained agent. This example makes the point that very unevenly spread out wealth can reflect something positive: households are able to share risk through borrowing and lending. The extreme opposite case—the autarky outcome, which has perfect equality in asset holdings—allows no consumption smoothing at all. The most extreme asset inequality is obtained when $\underline{a}$ is set at the natural borrowing limit: the most generous level such that debt can always be paid back. It also shows that the level of the interest rate is a key variable in the macroeconomic determination of asset inequality.

In an Aiyagari model, there is also aggregate saving, with returns influenced by a neoclassical production function. From the perspective of the individual's need to insure, this offers an additional possibility, as explained in Chapter 11. As a consequence, with more risk to be insured, or higher risk aversion, the steady-state capital stock will rise, reflecting the increased need to save.

Quantitatively, a simple AR(1) process for earnings generates wealth inequality that is more dispersed than earnings inequality, but not by a sufficient amount to match high wealth concentration at the top as observed in the data. The reason for this is that the richest do not value insurance so much—by virtue of being rich, as their accumulated saving provide a very good buffer against bad earnings shocks. In addition, as we saw in Chapters 11 and 17, in this class of economies, the riskfree rate is below the discount rate, which means that well-insured agents will *decumulate* wealth. Thus, the right tail of the wealth distribution becomes limited: the precautionary need dies off at higher wealth levels and the low return on savings dominates.

[29] Imrohoroğlu (1989) assumes that the real interest rate is exogenous and constant; Aiyagari (1994) and Huggett (1993) derive the interest rate endogenously, but do not have aggregate shocks.

If, on the other hand and as in Castañeda et al. (2003), the earnings process has a combination of (i) very large upside shocks while (ii) still allowing a non-trivial risk to fall far down from the top of the earnings distribution, then the model generates highly a skewed wealth distribution. Very large earnings shocks and consumption smoothing together makes it possible to accumulate large amounts of wealth possible, and the decumulation at high levels is hampered by the precautionary motive still being active, since earnings can drop precipitously.

Idiosyncratic shocks to discount factors or asset returns We saw above that permanent differences in discount factors generate a determinate, and extreme, wealth distribution. Now consider random movements in discount factors, say, at the frequency of cohorts; e.g., as the dynasty wealth is passed on from one cohort to the next, the new dynasty head may be more, or less, patient than the previous one, and in a random manner.[30] Suppose these shocks are idiosyncratic and independent across dynasties, possibly with some persistence, but drawn from the same distribution; then the steady-state wealth distribution will be more dispersed than in the case of common discount factors.

Next, suppose individuals receive shocks to the returns of their portfolios, again in an idiosyncratic manner but drawn from the same distribution; these shocks could also feature some persistence. Then again the wealth distribution would spread out and the higher is the variance and persistence of the shocks, the more the steady-state wealth distribution will spread out.

Let us now take a very brief detour from the structural models in focus in this section: suppose we simply have a framework of "random growth", as described in Kesten (1973):

$$a_{t+1} = s_t a_t + \epsilon_t,$$

where s_t and ϵ_t are i.i.d. Then under some conditions on the primitives of this process, it turns out that a_t converges in probability to a random variable A that satisfies $\lim_{a\to\infty} Prob(A > a) \propto a^{-\zeta}$, i.e., the right tail of the stationary distribution has a Pareto shape.[31] This result is remarkable in that the processes for s and ϵ are unspecified, and yet the limit distribution is of a specific shape, namely Pareto, which is also a very good approximation to the right-most tail of wealth, as well as earnings, in the data.

It is possible to connect the random growth model to our structural model, allowing for idiosyncratic shocks to earnings, discount factors, and returns. It turns out—for details, see Hubmer et al. (2018)—that if one assumes CRRA utility, then for large asset positions, the random-growth formula is a very good approximation to optimal behavior of saving.[32] In particular, the randomness in s has two components, one deriving from the discount factor shock while the other is a return shock: intuitively, the money saved is multiplied by a gross

[30] Formally, in a model with constant (geometric) discounting, the discount factor applied to utils at t in terms of utils at 0 is β^t; here, they would be $\prod_{s=0}^{t-1} \beta_s$, where β_t is random, e.g., a first-order Markov process. This kind of discounting is time-consistent, i.e., the consumer would not want to change a contingent plan made in advance when the future arrives.

[31] The result follows if there exists a $\zeta > 0$ with $\mathbb{E}[s^\zeta] = 1$, with $\mathbb{E}[\epsilon^\zeta] < \infty$. For a nice exposition, see Gabaix (2009).

[32] The linearity applies quite well also lower down in the distribution, except for the lowest levels of wealth where the asset evolution is noticeably non-linear with a slight convexity.

return, which is random, to which a marginal saving rate, also random due to discount-factor heterogeneity, is then applied. The ϵ, then, contains the earnings shock. Note that the Pareto shape will not apply unless s is random, so earnings shocks per se do not generate Pareto tails, unless they are Pareto distributed themselves.[33]

Finally, notice that the two mechanisms behind a right Pareto tail—random discount factors and random returns—are quite different from a welfare perspective. In the former case, the wealth distribution reflects conscious choice: poor dynasties are those that, on average at least, had low discount factors earlier on. Thus, when born into a dynasty with low wealth, high current consumption is not an option but high consumption was probably what occurred earlier in time for this dynasty. Moreover, if offered insurance, a dynasty would not necessarily want to hedge a low (or high) future discount factor, as preferences throughout have been assumed to be time-consistent.

In contrast, random returns fundamentally lead to undesired wealth dispersion: those with high wealth are wealthy because they were lucky in the asset market. If given the opportunity to insure, the consumer would highly value such insurance. Thus, a currently poor dynasty is highly likely poor because the ancestors were unlucky in the returns on their savings.

Hours worked In the discussion above, earnings were taken to be exogenous. If individuals can choose how much to work, then this margin can affect wealth inequality. First, labor supply can be used as additional insurance vehicle in response to shocks. For some shocks, such as those to wages, then a positive shock can of course also lead to higher hours worked, in this case enhancing earnings variability. The interaction of hours worked and wealth accumulation also allows this class of models to address the data on the distribution of hours worked. While there is considerable variation in hours worked across individuals, it is difficult to understand the determinants of this variation since the correlation of hours with observable worker characteristics tends to be low. For example, the correlation between hours and wages or financial wealth is curiously close to zero, despite a common perception that productive or rich people on average work harder. A model with wage shocks and labor supply featuring strong income effects, in line with the discussion in Chapter 12, can make sense of a zero, or weak, correlation: income effects make the wealthier want to work less, but intertemporal substitution of labor supply at the same time generates high hours worked by those with high current wages and who also save in order to smooth consumption. In this kind of environment, tax policy distorting labor decisions, will also influence the observed correlations. For details, see Domeij and Floden (2006) and Pijoan-Mas (2006).

Why are so many so poor? In the models above, the focus was mostly on mechanisms through which the right tail of the wealth distribution becomes thick, as it is in the data. A different, but no less important, question is to understand the wealth formation, or lack thereof, of the very poorest. As we saw earlier in the chapter, the 40% poorest only hold 1% of total financial wealth, and in addition there is poor coverage of people in this part of the distribution.

[33] In this case, we obtain a Pareto tail for assets equaling that for earnings.

A benchmark Aiyagari model predicts many too few individuals in the left tail of the wealth distribution, simply because they save their way out given that it is painful to have very low consumption. A key missing element is social security support—provided either by government in the form of transfers, food stamps, and other free goods and services, or by family/friends—that effectively constitutes a consumption floor, hence making it much less costly not to have wealth.

Many low earners may also lack effective means of saving, if they have needs to conceal wealth from others due to informal sharing arrangements within multi-person households or social networks. Yet others, who have defaulted on loans but have been unable to file for bankruptcy, may be able to save but their savings are then typically garnishable by creditors.

Clearly, many individuals may also suffer from mental conditions, addiction problems, or simply elements of irrationality not captured by the standard utility functions used here. Another element missing in our models is how crime, both as an activity and through incarceration, shapes the earnings and wealth of those involved. The very poorest are generally understudied in economics, including in macroeconomic analyses.

Quantitative analysis We now briefly illustrate the above points by presenting results from steady states given a set of extensions of the Aiyagari model. First off, a standard Aiyagari model, calibrated to PSID data using an AR(1) income shock, as in Aiyagari (1994), delivers an income Gini coefficient of 0.37 and a wealth Gini of 0.67, with a steady-state interest rate of 2 percent and a ratio of capital to annual GDP of 3. Thus, the model does not generate nearly as much wealth inequality as we see in the data (recall that it is significantly above 0.8); the percentage of wealth held by the 1 percent richest is 9%, as opposed to nearly 40% in the data. The average MPC is 0.17, but the bottom 10% of the distribution has an MPC of 0.78. Let us now look at a number of extensions.

A superstar earnings process: assume earnings are in line with Castañeda et al. (2003), such that the Gini coefficient for wealth equals 0.8; this involves an income Gini of 0.67, with the same interest rate and k/y ratio as in the standard model. The share of wealth held by the 1 percent richest is now over 0.4, i.e., it even overshoots. This model, however, while delivering highly dispersed wealth, does not change the MPC distribution in the population more than very marginally.

Stochastic discount factors: assume that β takes on 3 possible values randomly. This process is persistent, with expected duration of any given value of β of around 75 years, and has 10 percent of the population at each of the extreme values and 80 percent in the middle. The three β values are then chosen to match a wealth Gini of 0.8. This model generates more moderate wealth concentration at the top—the 1 percent richest have only 19% of all wealth—but depresses the real interest rate to around 0.5 percent (and a k/y slightly above 3): this rate is highly influenced by the most patient, who now have a high β. Here the MPC distribution is moved up significantly, to an average of 0.29; the poorest have much higher MPCs, while the richest still have very low MPCs.

Random returns to saving: assume that the returns are iid with a standard deviation chosen to match the 0.8 wealth Gini. Now the (average) real interest rate is 3 percent, with a k/y ratio a little below 3. The richest 1 percent hold a fraction right in between those of the previous two model versions (so around 0.3) but the MPC distribution is again back at

roughly the same level as for the basic Aiyagari model.

Comparing all these models we see that different theories of wealth inequality imply very different MPC distributions. There is, as of yet, no firm consensus as to which mix of assumptions is most appropriate: the research is very much ongoing.[34]

A note on welfare comparisons in heterogeneous-agent models

We end this section by briefly discussing a conceptual challenge—one that is as natural as it is important—when welfare comparisons are made in economies that are dynamic and inhabited by a heterogeneous population.

Transition First, we already know from Chapter 6 that in a dynamic representative-agent model, comparing steady-state welfare (say, resulting from two alternative policies) is not sufficient: the transition path needs to be taken into account. For example, let the steady-state capital stock in a frictionless benchmark be k^*. Then a small, time-independent subsidy to capital income would take us to a new steady state in which capital, and hence welfare, are higher.[35] But taking the transition path into account, the welfare of the representative agent, at time 0 as of the introduction of the policy, must be lower, since the initial economy is efficient. Intuitively in this case, the higher capital accumulation requires consumption to be lower during the initial phase of the transition, outweighing the benefits from the higher long-run consumption due to higher capital.

With heterogeneous agents, it is of course in addition important to take into account how different individuals are affected by a counterfactual experiment. In the benchmark model we will discuss below, consumers are also hit by various shocks that are not fully insurable; hence, they also move around in the distribution over time. This poses another conceptual challenge. Say that we are again interested in the effects on welfare of adopting a subsidy to saving, and say that the initial position is one of a steady state. Then the correct welfare comparison is arrived at by (i) solving for a full transition path given the new policy and (ii) comparing present-value utility for all agents, as of time 0 (taking transition into account). This may result in some agents gaining, and others losing, from adopting the subsidy. Such a result could be reported to a policymaker wanting to evaluate the policy. The policymaker may at this point want to add their own way of weighing the different agents' outcomes together—by applying a specific social welfare function—or not. A commonly adopted social welfare function in the literature is the equally-weighted utilitarian function: one where the present-value utility functions of all agents at time zero is just summed up. Such a social welfare function is normative—the particular function is one that implicitly puts a high weight on equality—and not cannot generally be justified otherwise.[36,37]

[34] See Ozkan, Hubmer, Salgado, and Halvorsen (2023) for recent advances.

[35] Recall that in the standard dynastic model, steady-state welfare is not maximized: due to discounting, it involves slightly lower capital than that maximizing steady-state consumption.

[36] Equal weights will want the planner to distribute consumption so as to equalize the marginal utilities of all agents. When utility is additive in consumption, this implies equal consumption for all agents.

[37] The equal-weight utilitarian welfare function, applied at time 0 for present-value utility, can be justified normatively if, at time zero, all agents are in identical situations—but may differ ex post due to shocks. This measure is referred to as one "behind the veil of ignorance", a notion introduced in Rawls (1971).

The nature of contracts The heterogeneous-agent literature, where wealth inequality is a natural outcome, fundamentally builds on incomplete markets. The canonical model has no insurance but riskless saving, along with an exogenous borrowing limit, with the loose empirical motivation that many risks in life appear uninsurable, except via saving and only limited borrowing. Yet a number of insurance markets do exist, life insurance, property insurance, and health insurance being prominent examples. In addition, most countries have some degree of social security and publicly provided health services at low cost; also note that in some countries, such as the U.S., there is personal bankruptcy protection, providing further insurance. Insurance against professional failure (captured by "wage risk" in the model), or divorce, are examples where the complete lack of formal insurance markets seems a more appropriate assumption.[38]

Given all this, what is appropriate modeling of the nature of contracts? The literature tends to adopt assumptions that are easy to implement and seem like reasonable descriptions of reality. Many models include a description of the publicly provided safety net. Examples of important and non-trivial extensions to the canonical model can be found in the literature that incorporates personal bankruptcy protection into the canonical setting, as pioneered by Chatterjee, Corbae, and Rios-Rull (2008) and Livshits, MacGee, and Tertilt (2007), and models how publicly available information on individuals (e.g., credit scores) is used as a determinant of borrowing contracts, such as in Chatterjee, Corbae, Dempsey, and Ríos-Rull (2023). Common to all these frameworks is a difficulty of conducting policy analysis: one would expect the nature of contracts to react to any policy change, and if the model does not endogenize the source of market incompleteness assumed, it is vulnerable to a Lucas critique.

To be concrete, in a canonical model it seems model-feasible to enact a policy that entirely insures against wage/earnings risk, and such a policy would improve welfare in an ex-ante sense.[39] If enacted in the real world, it would likely lower average income significantly, as efforts and investments in human capital would likely be significantly reduced: after all, such a policy would be "socialism pure", a system which few economists would argue in favor of. Most likely, the reason why markets do not provide insurance for earnings shocks is a significant degree of private information leading to moral hazard and adverse selection, as well as limited ability of consumers to commit.[40] One approach would thus be to formulate models with explicit information/commitment frictions and derive the optimal contract given these frictions; such models would be "Lucas-proof". This approach is technically challenging, however, and often lead to contract types that do not resemble what we observe in reality.[41] Thus, there remains a tension between descriptive accuracy and Lucas-proof modeling.

[38] Informal insurance markets, through family and social networks, may still be active.

[39] That is, suppose agents are all identical at time zero, with equal wealth and without having experienced any shocks yet. Then they are all identical and would benefit from full insurance.

[40] Mirrlees (1971) and a large follow-up literature studies optimal insurance contracts in private-information economies. Note also that publicly supported bankruptcy regulation is often an imperfect solution to a commitment problem; see Mateos-Planas, McCrary, Ríos-Rull, and Wicht (2025).

[41] See Allen (1985) and Cole and Kocherlakota (2001) for progress in this regard.

21.3 Reasons why inequality matters for aggregates

The previous sections described inequality, along with a number of theories of different dimensions of inequality. Here we briefly address how the presence of inequality affects macroeconomic aggregates in important ways.

21.3.1 Long-run channels

We only briefly mention the channels through which inequality is important for growth and development. The brief mention is not for lack of importance, however; since long, many connections between inequality and growth/development have been studied in the macroeconomic literature. The discussion here will mainly be qualitative.

Incentives

There is a view that inequality is "good": not by itself, but because it may reflect the presence of incentives to work hard and to innovate and accumulate. Tax and transfer systems aimed at equalizing consumption would then normally generate lower aggregate output, and likely worsen welfare for a large set of agents. Take, for example, an Aiyagari model with variable labor supply. A system that taxes earnings and capital income at proportional rates while transferring the revenue in lump-sum, equal amounts to all agents would indeed make the distributions of disposable income and consumption less dispersed. (Formulate such a model and use our program package to solve it for a steady state and you will see!) However, such a system would generate a lower capital stock in steady state as well as fewer hours worked. On average, people will most likely be worse off under such a system, even as measured from time 0, taking the transition into account. The qualification "most likely" is necessary since redistribution, even when distortionary, can improve welfare when markets for insurance are missing: consumption smoothing across states is desirable but markets do not, by assumption in this case, allow it.[42] Of course, ideally, then, a policy that improves on market performance, if such a policy is feasible, would be desirable. However, deeper frictions, such as private information (moral hazard or adverse selection) or lack of commitment, might make such improvements impossible.[43] I.e., no government policy would then allow us to achieve better outcomes.

A recent literature looks at the incentives for innovation and patent creation in rich countries. Can significant inequality hamper these activities? To the extent insurance markets are poor and the potential innovators, whose projects involve risks, cannot insure against these risks, they may benefit from social insurance schemes or from joining larger companies where the payoffs are somewhat smoothed out across states. How and where—in what market forms—innovation activities occur is therefore an important element behind productivity growth. Arguably, in many rich countries the need for insurance may be limited. The outgrowth of unicorn companies and the remarkable wealth accumulation associated to them

[42] A policy that is not distortionary but that distributes from the lucky, i.e., those with high earnings or wage shocks, to the unlucky, would of course be even better.

[43] That is, even the distortionary tax system just described may not be feasible.

is a phenomenon that is not just observed in countries like the United States: they are observed, along with high Gini coefficients for wealth, in Scandinavian countries as well, where the social insurance schemes are much more developed. It is conceivable, however, that the even more extreme wealth accumulation outcomes observed in the United States can be a result of lower government involvement and freer markets, reflecting incentive effects that generate more innovative activity.

Credit-market restrictions and inequality traps

A basic, and very important, issue concerns human capital accumulation and how inequality may hinder it.[44] Education is to some extent a consumption good but the consensus in the literature is still that human capital is key for production and economic development. The phenomenal developments within AI recently is an example of technology developments that surely would be impossible without highly educated innovators. The "access to education", which clearly is a prerequisite for advanced development of an economy, can be interpreted as "education involves a fixed cost". I.e., if there are no schools in one's country, or one's neighborhood, or if there are but they involve significant tuition fees, then one's wealth level becomes a key determinant of the possibility of attending school/university, unless it is possible to borrow to finance the education. Suppose a worker can work as unskilled and earn a present-value lifetime labor income of w_u or attain education and obtain the skilled lifetime income $w_s > w_u$, and suppose the education costs F. Then if $w_s - w_u > F$, it would pay off for the individual to obtain education. They would thus choose to become educated, unless the cost of education is paid upfront and the income gains materialize later and the individual cannot self-finance education or borrow to attain it. Without any possibility of borrowing, their own assets would need to be at least above F for education to be a feasible choice. Thus, the fraction of individuals who would choose education given these income levels equals the fraction of people with wealth above F in the wealth distribution; hence, the wealth distribution matters for educational attainment. A full general-equilibrium treatment of this idea would endogenize the earnings w_u and w_s, perhaps with an aggregate production function like that studied earlier in this chapter; in such a setting, the fraction of people who obtain education along with the earnings distribution would be affected by the initial wealth distribution. Poor individuals can also end up in "trapped dynasties" where the next generation will not obtain an education either, due to the low earnings of the present generation. On the level of an entire economy, there can potentially also be a trap, with low overall GDP due to credit constraints preventing a large part of the population from becoming educated. To what extent this mechanism is an important factor behind inequality within a country, or behind the differences in productivity across countries, are still open issues.

21.3.2 The business cycle

The role of inequality for understanding business cycles has received significant attention in macroeconomic research over the last decades. In particular, heterogeneous-agent models—

[44] For an early treatment, see Galor and Tsiddon (1997).

essentially like those discussed at several points in the book and in detail in the previous section of this chapter—with one or more sources of aggregate fluctuations, including with nominal frictions, are now viewed to be an important part of our macroeconomic toolkit. A key reason why this literature has had impact in macroeconomics is that it allows the marginal propensity of consumption, MPC, to rise relative to that coming out of a representative-agent, dynastic model. In the latter, recall that the permanent income effects of transfers are very small (on the order of the interest rates); in heterogeneous-agent models many consumers have very high MPCs (those who are literally borrowing-constrained have a marginal propensity of 1).[45] With larger MPCs, transfer policies, such as typical fiscal policy, will have larger effects on overall demand and hence make such stabilization policy more powerful.

The how-to

In principle, an endogenous wealth distribution is a high-dimensional study object, though as we have seen earlier in this chapter, it is straightforward to solve for the distribution in steady state. In particular, in the basic Aiyagari model, it suffices to guess on a steady-state capital stock and then, given this guess, be able to solve a dynamic programming problem with one endogenous state (asset holdings, or cash on hand), find the stationary distribution of asset holdings generated by the implied decision rules and iterate until the sum of asset holdings in the stationary distribution matches the capital stock assumed. This numerical procedure finds a solution within seconds. However, to study aggregate uncertainty requires also determining how the distribution evolves stochastically over time. In principle, the distribution of wealth will affect prices, so each agent's dynamic programming must have as a state variable the distribution of wealth, which is a high-dimensional object: solving dynamic programming problems is explosively more costly as more states are included. The agent also needs to know the mapping from the wealth distribution to prices. Overall, it simply seems infeasible to study such economies, even with fast computers.

The literature has addressed this in two ways. One has been to lower the ambition level in terms of heterogeneity and assume that there are two, or a very small number, of representative agents in the economy. The leading, simplest example of this approach would have one agent who is a worker and cannot save. Hence, this agent is "hand-to-mouth", i.e., consumes all current income and thus has an MPC of 1. The other agent, then, would be a pure "rentier" who does not work but simply owns and rents out capital and consumes from the capital income. Such an agent would have an MPC much closer to zero, as in a standard representative-agent case. The fraction of consumers of each type can then be selected so as to obtain an aggregate MPC of, say, 0.5. Such a version of a New Keynesian setting, often referred to as a "TANK" (Two-Agent NK) model, is used at many policy institution.

The other approach in the literature has been to confront the computational challenge head on. Krusell and Smith (1998) showed one path forward, and the following decades have added numerous elaborations and alternatives, with the result that methods for solving heterogeneous-agent models with aggregate shocks are now taught in graduate programs and considered standard second-year material, and they also used at policy institutions. One

[45] The Keynesian consumption function, in particular, was often estimated to feature MPCs of 0.5 or higher.

method that is particularly accessible conceptually and computationally relies on a presumption that, around a steady state, the behavior of aggregates can be well approximated as a linear function of the shocks hitting the economy.[46] The procedure boils down to computing the transitional deterministic behavior of the economy in response to a one-time unexpected shock of a given size; then random, repeated shocks of different sizes can be computed using linearity.[47] Further improvements are being added at a rapid rate, including for cases where second-moment effects and nonlinearity play a central role.

Insights and relevance

As already stated, the heterogeneous-agent approach to business cycles allows us to derive MPCs that are more in line with data. Consider for example an Aiyagari-style model, augmented so that it matches the wealth distribution in steady state. Then there will be MPCs within the distribution of agents ranging from 0.01 or so to 1, and depending on the exact form of the distribution, the average MPC will be 0.1 or quite a bit higher. These features are captured in the stylized Figure 21.10 depicting the decision rule for saving.

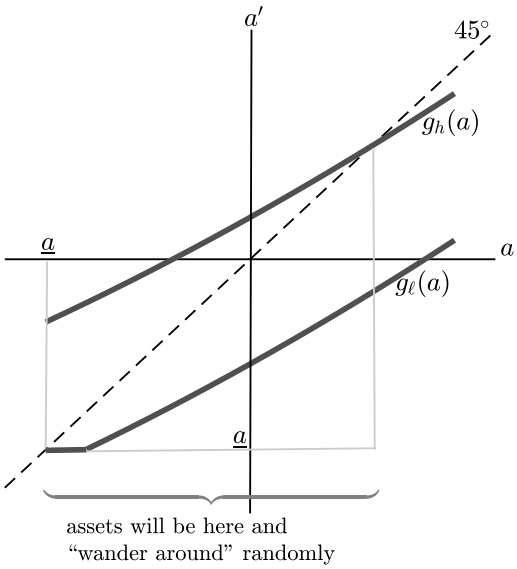

Figure 21.10: MPCs high for low asset values (1 when constraint binds).

A particular recent version of Aiyagari models also includes the so-called "wealthy hand-to-mouth" agents. These are individuals who, despite being relatively wealthy, have high MPCs due to their asset portfolios being rather illiquid.[48] A particularly simple, ad hoc way of generating this mechanism is to assume that, independently of the agent's current asset position a, there is an iid random event whereby $a' \geq a$ is imposed: i.e., if the event occurs,

[46] See Boppart et al. (2018) and Auclert, Bardóczy, Rognlie, and Straub (2021).

[47] Linearity means that responses can be scaled by shock size and added up, including for a vector of different kinds of shocks. It also means that the responses to shocks will be identical whether they are random or deterministic: recall from Chapter 7 that certainty equivalence applies. Hence also the effects of random shocks can just be added up.

[48] See Kaplan and Violante (2010).

the agent cannot decumulate assets. Thus, in case this event occurs in a period where the non-asset income is low and the agent would like to draw down on asset holdings in order to consume, the agent will be constrained and have an MPC equal to 1. With this additional mechanism, obviously the average MPC rises even further.

As a result of higher average MPCs, fluctuations in aggregate consumption are brought closer to the data: the correlation between consumption and output rises and the excess sensitivity puzzle (see Chapter 11) can be explained: aggregate income shocks, even recent ones, move consumption directly through their effects on borrowing-constrained agents. However, the barebones Aiyagari model, even with an added mechanism that generates MPC heterogeneity, cannot contribute much to increased output movements, because the "demand side" of the model is rudimentary: output is given by the production function, where a productivity shock is the main reason for fluctuations. Models where demand plays a central role in determining output, at least in the short run, are thus needed. One example is to be found in the literature combining the Aiyagari-style model with New Keynesian frictions (so-called HANK models). Let us, however, look at an another very simple example where demand matters, similar to that alluded to in Chapter 3.5.1. Suppose output is given by $\hat{A}_t k_t^{\alpha}$, that is, labor is not a variable factor, but where $\hat{A}_t = A_t c_t^{\omega}$, where A_t is exogenous TFP and c_t is consumption at time t, with $\omega \geq 0$: there is a positive externality to consumption. That is, if people demand higher consumption, output rises, everything else equal: output is, to this extent, demand-determined.[49] Suppose the model is otherwise neoclassical, of the Aiyagari (1994) variety, but with the addition of a wealthy-hand-to-mouth mechanism as the one just discussed above: for each agent, the probability that saving is not allowed to decrease between t and $t+1$, γ, is iid. Figure 21.11 shows the results of the impact response of consumption and output of a one-percent shock to the exogenous part of TFP. If $\omega = \gamma = 0$, output (in the right-hand-side panel) rises mechanically by 1 percent, whereas consumption rises by 0.3 percent. If γ is high (0.6), so that many agents have a high MPC, consumption rises by over 1.1 percent, but output is unaffected: investment responds negatively, as total resources available for consumption and investment is fixed. However, with a significant externality, the strong consumption response also generates a significant rise in output. We see that the two mechanisms—high MPCs and a demand channel—reinforce each other and generate strong propagation from TFP shocks, even in the complete absence of an endogenous hours channel.

A further insight is that the effects of government policy, such as fiscal transfers or monetary policy, depend on which population groups are affected. In order to boost demand, for example, effective policy instruments are those that target high MPC consumers. Moreover, the marginal propensities to work and invest in risky vs. non-risky assets will also differ across the population; i.e., policymakers more generally must assess the distribution of marginal propensities when comparing different policy options.

Models with aggregate fluctuations and heterogeneous agents also allow policymakers to look at the differential impacts of policy on the welfare of different agents. Traditional macroeconomic models have only allowed us to look at the unemployment dimensions of policy—how effects on the employed differ from those of the unemployed—but unemployment

[49] For models of this sort, see Krueger et al. (2016) and for more elaborate frameworks generating reduced forms similar to this framework, see Bai, Ríos-Rull, and Storesletten (2025) and Huo and Ríos-Rull (2015).

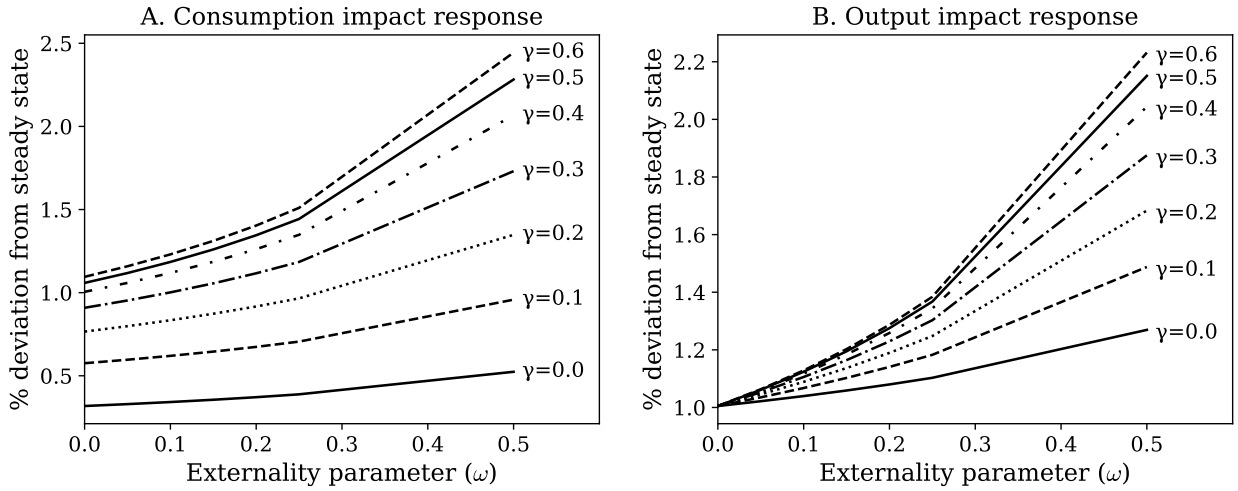

Figure 21.11: Impact responses to a 1 percent exogenous TFP shock.

is of course only one determinant of inequality, with effects through the full wage distribution playing an arguably important role. Thus, models with rich, endogenous wage inequality accompanied by wealth inequality and less than full insurance also offer many new avenues for policy analysis.

21.3.3 From micro to macro: more heterogeneity

The previous two subsections briefly discuss how macroeconomic aggregates may behave differently, in the long as well as in the short run, due to the presence of inequality in incomes and wealth. Let us finally mention a number of important lines of macroeconomic research that further explore heterogeneity, of different kinds, on the microeconomic level. All macroeconomic models fall short of modeling the full complexity at the microeconomic level and there is always a question of whether this abstraction is distorting our analysis. One motivation behind this kind of research is a desire to check that conclusions from our are robust when we allow for the main sources of difference across households. As an example, we have just seen that incorporating inequality in incomes and wealth on the microeconomic level will change our standard dynastic model in ways that can be important in addressing some questions, such as stabilization policy.

Age differences between individuals have already been discussed in our text, as a natural element of overlapping-generations models. The age structure of an economy can matter for aggregate saving, but also when studying social security and pensions. Currently, fertility is low in many countries, which means that an increasing burden will be placed in the future on the working individuals to provide for the elderly: aggregate labor supply will be significantly affected. Saving and labor supply are also affected by the household structure. Taking marriage and cohabitation into account can be important, for example because the response of labor supply to transfers and to changes in labor market conditions can become strong for "second earners" in a household.[50] In addition, cohabitation appears to be countercyclical—

[50]Married women have been documented to have a more elastic labor supply and a less strong attachment

presumably to cut costs when incomes are low—and have a downward trend.

A large literature explores labor markets from a macroeconomic perspective and there, heterogeneity in many dimensions, such as individual's education, occupation, work experience, health status, or location of residence, is often incorporated into the analysis. Sometimes, such models involve search and matching markets (as in Chapter 20), they pose additional computational challenges: the current distribution of agents across matches becomes relevant for any given individual and this distribution can sometimes be captured by a simple statistic, such as labor market tightness, but sometimes it cannot.

Heterogeneous-agent macroeconomics is expanding rapidly, in part because computational power and methods are advancing at a rapid pace, allowing our macroeconomic analyses to benefit from detailed microeconomic foundations. Finally, more realistic microeconomic foundations allow us to connect the model with microeconomic data and microeconomic studies. By relating the model to a wider range of data, we can bring more information to bear on our research questions.

to the labor force.