THREE ESSAYS IN MACROECONOMICS

by

Jie Cao

A thesis submitted in conformity with the requirements for the degree of Doctor of Philosophy
Graduate Department of Economics
University of Toronto

© Copyright 2016 by Jie Cao
Abstract

Three Essays in Macroeconomics

Jie Cao
Doctor of Philosophy
Graduate Department of Economics
University of Toronto
2016

This thesis consists of three chapters. Chapter 1 studies a life-cycle pattern of female labor supply (hours per woman) in Japan. It exhibits an “M” shape with the second peak lower than the first one. Employing a micro-data set, I show that the pattern can be understood as a result of labor supply behavioral differences across different types of women and demographic composition changes along the life cycle. I then build a life-cycle model featuring transitions between heterogeneous types of women, human capital accumulation and childcare cost. The calibrated model accounts for the aggregate pattern well. Counterfactual experiments suggest that narrowing gender wage gap would have a larger positive effect on female labor supply than lowering childcare cost, a result mostly explained by the human capital channel.

Chapter 2 studies the links between idiosyncratic distortions and potential aggregate losses. Under the economic environment of Restuccia and Rogerson (2008) (RR), I characterize analytically the mappings from distortions to total factor productivity (TFP) and other aggregate measures. Using these mappings, I explain in a unified way three features emerged from RR’s numerical experiments, where endogenous exit of firms is assumed away and distortions are restricted so as to have no impact on capital accumulation. Additionally, I explain why these features disappear when distortions affect capital accumulation. I then extend RR’s study by introducing the endogenous exit margin and find that various aggregate losses can respond to this margin quite differently.

Chapter 3 studies financial frictions and resource misallocation in China’s manufacturing sector. I emphasize two aspects of the financial frictions in the Chinese context. One is credit discrimination: Unproductive state-owned enterprises (SOEs) have easy access to credit while productive non-SOEs are credit constrained. The other is that the credit constraint faced by non-SOEs is much tighter than those in financially developed countries. Using a firm-level data set, I find that a potential 24% TFP gain can be achieved if the credit constraint faced by non-SOEs is at a level similar to the one in the US, and that 53% of the gain can be attributed to improved allocations between SOEs and non-SOEs.
Acknowledgements

Foremost, I would like to express my sincere gratitude to my advisors, Xiaodong Zhu, Gueorgui Kam-bourov and Diego Restuccia, for their continuous support and guidance. I thank Andrés Erosa, Kevin Fawcett, Burhanettin Kuruscu, Rebecca Lindstrom, Ho-nam Mak, Serdar Ozkan, Carolyn Pitchik, Zhen-jie Qian, Ronald Wolthoff, Shouyong Shi, Aloysius Siow, and Joseph Steinberg for helpful discussions. This thesis also benefited from comments received from seminar participants at UofT ECO4060, UofT Macro workshop, 10th CIREQ Ph.D. students’ conference and CEA annual meetings (2014 & 2015).

I thank the Institute for Research on Household Economics for permission to access the Japanese Panel Survey of Consumers (JPSC). I also thank Dana Blumin at OECD and Mami Umezawa at Statistics Bureau Japan for kindly providing me the aggregate employment data. (Both JPSC and the employment data are used in Chapter 1.)

Lastly, and most importantly, I would like to thank my family. Special thanks go to my wife, Kumi Hoshino, for her encouragement, unconditional love and care. This journey would not have been possible without her support throughout the entire process. I am also thankful to my one-year old son Sean for bringing me happiness. I thank my parents, Kuixin Cao and Yunjie Qin, for all their love and support throughout my life.
Contents

1 Explaining Life-Cycle Female Labor Supply in Japan

  1.1 Introduction .................................................. 1
  1.2 Empirical Facts .............................................. 6
      1.2.1 The micro-data set .................................... 7
      1.2.2 Labor supply and marriage ............................. 7
      1.2.3 Labor supply and child-rearing ........................ 9
      1.2.4 Intensive and extensive margins ...................... 11
  1.3 Two simple examples ......................................... 11
      1.3.1 A static model ......................................... 12
      1.3.2 A two-period model ................................... 16
  1.4 A life-cycle model .......................................... 20
      1.4.1 Demographics .......................................... 20
      1.4.2 Preferences .......................................... 21
      1.4.3 Human capital accumulation ........................... 22
      1.4.4 Budget constraint ..................................... 22
  1.5 Calibration .................................................. 23
  1.6 Results ........................................................ 27
      1.6.1 The baseline economy .................................. 27
      1.6.2 The three determinants ................................. 29
          1.6.2.1 Childcare cost .................................. 30
          1.6.2.2 Initial gender wage gap .......................... 31
          1.6.2.3 Return to experience ............................. 32
          1.6.2.4 Summary ......................................... 33
  1.7 Conclusion .................................................... 33
Appendix 1.A More Facts ........................................... 35
    1.A.1 Life-cycle labor supply in contrast .................... 35
    1.A.2 The two margins by marital status ...................... 35
Appendix 1.B Data .................................................. 36
    1.B.1 Number of observations by age ........................ 36
    1.B.2 Data construction ...................................... 37
Appendix 1.C Computation .......................................... 39
    1.C.1 Value Function ......................................... 39
    1.C.2 Algorithm .............................................. 39
# Contents

<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bibliography</td>
<td></td>
<td>41</td>
</tr>
<tr>
<td><strong>2</strong></td>
<td>Idiosyncratic Distortions &amp; Aggregate Losses</td>
<td>43</td>
</tr>
<tr>
<td>2.1</td>
<td>Introduction</td>
<td>43</td>
</tr>
<tr>
<td>2.2</td>
<td>The model</td>
<td>46</td>
</tr>
<tr>
<td>2.2.1</td>
<td>Technology</td>
<td>46</td>
</tr>
<tr>
<td>2.2.2</td>
<td>Representative household</td>
<td>48</td>
</tr>
<tr>
<td>2.2.3</td>
<td>Equilibrium</td>
<td>48</td>
</tr>
<tr>
<td>2.3</td>
<td>Characterization of the equilibrium</td>
<td>49</td>
</tr>
<tr>
<td>2.3.1</td>
<td>Undistorted economy</td>
<td>49</td>
</tr>
<tr>
<td>2.3.2</td>
<td>Distorted economy</td>
<td>53</td>
</tr>
<tr>
<td>2.3.3</td>
<td>Aggregate losses</td>
<td>55</td>
</tr>
<tr>
<td>2.4</td>
<td>Results</td>
<td>55</td>
</tr>
<tr>
<td>2.4.1</td>
<td>Examining RR’s results</td>
<td>56</td>
</tr>
<tr>
<td>2.4.1.1</td>
<td>$c_f = 0$ and $K = K_d$</td>
<td>57</td>
</tr>
<tr>
<td>2.4.1.2</td>
<td>$c_f = 0$</td>
<td>63</td>
</tr>
<tr>
<td>2.4.2</td>
<td>$c_f &gt; 0$</td>
<td>65</td>
</tr>
<tr>
<td>2.5</td>
<td>Concluding remarks</td>
<td>70</td>
</tr>
<tr>
<td>Appendix 2.A</td>
<td>Algebra</td>
<td>71</td>
</tr>
<tr>
<td>2.A.1</td>
<td>Aggregation in the undistorted economy</td>
<td>71</td>
</tr>
<tr>
<td>2.A.2</td>
<td>Aggregate output in the distorted economy</td>
<td>72</td>
</tr>
<tr>
<td>Appendix 2.B</td>
<td>Proofs</td>
<td>72</td>
</tr>
<tr>
<td>2.B.1</td>
<td>Sufficient conditions for $TFP_r &lt; 1$ in the case of $c_f = 0$</td>
<td>72</td>
</tr>
<tr>
<td>2.B.2</td>
<td>Sufficient conditions for $TFP_r &lt; 1$ in the case of $c_f &gt; 0$ and $G_z(z)$ Pareto</td>
<td>72</td>
</tr>
<tr>
<td>Bibliography</td>
<td></td>
<td>73</td>
</tr>
<tr>
<td><strong>3</strong></td>
<td>Financial Frictions and Resource Misallocation</td>
<td>75</td>
</tr>
<tr>
<td>3.1</td>
<td>Introduction</td>
<td>75</td>
</tr>
<tr>
<td>3.2</td>
<td>Empirical evidence</td>
<td>77</td>
</tr>
<tr>
<td>3.2.1</td>
<td>Background of SOEs and non-SOEs</td>
<td>77</td>
</tr>
<tr>
<td>3.2.2</td>
<td>Credit discrimination</td>
<td>78</td>
</tr>
<tr>
<td>3.2.3</td>
<td>Productivity gap between SOEs and non-SOEs</td>
<td>79</td>
</tr>
<tr>
<td>3.2.4</td>
<td>Underdeveloped financial system</td>
<td>79</td>
</tr>
<tr>
<td>3.2.5</td>
<td>$K/L$ discrepancy between SOEs and non-SOEs</td>
<td>80</td>
</tr>
<tr>
<td>3.3</td>
<td>The model</td>
<td>80</td>
</tr>
<tr>
<td>3.3.1</td>
<td>Model Setup</td>
<td>80</td>
</tr>
<tr>
<td>3.3.2</td>
<td>Definition of equilibrium</td>
<td>82</td>
</tr>
<tr>
<td>3.3.3</td>
<td>Equilibrium characterization</td>
<td>83</td>
</tr>
<tr>
<td>3.4</td>
<td>Quantitative Analysis</td>
<td>87</td>
</tr>
<tr>
<td>3.4.1</td>
<td>Data</td>
<td>87</td>
</tr>
<tr>
<td>3.4.2</td>
<td>Calibration</td>
<td>88</td>
</tr>
<tr>
<td>3.4.3</td>
<td>Results</td>
<td>89</td>
</tr>
<tr>
<td>3.5</td>
<td>Concluding remarks</td>
<td>91</td>
</tr>
<tr>
<td>Appendix 3.A</td>
<td>Proof of proposition 1</td>
<td>93</td>
</tr>
<tr>
<td>Section</td>
<td>Page</td>
<td></td>
</tr>
<tr>
<td>-------------------------</td>
<td>------</td>
<td></td>
</tr>
<tr>
<td>Appendix 3.B Factor prices</td>
<td>94</td>
<td></td>
</tr>
<tr>
<td>Bibliography</td>
<td>96</td>
<td></td>
</tr>
</tbody>
</table>
Chapter 1

Explaining Life-Cycle Female Labor Supply in Japan

1.1 Introduction

Aggregate life-cycle female labor supply in Japan features an interesting pattern. In Figure 1.1, I plot age profiles of average weekly hours worked per woman for four cohorts of women born after WWII. These profiles reveal a life-cycle pattern with two main features. First, the life-cycle labor supply exhibits an “M” shape, with two peaks occurring when women are in their 20s and 40s, and a significant dip in between. Second, the two peaks are of different heights with the second one lower than the first. Consider, for example, the cohort born between 1958 and 1962. On average, they work the most in their early 20s for about 30 hours a week (the first peak). After that, they gradually work less until the number of weekly hours hits the bottom at about 19 hours per week in their early 30s. The drop in hours is substantial, more than one-third relative to the first peak. Although afterward they start to work more again, they never work as much as they do in their early 20s. At the second peak, they are in their late 40s and work about 25 hours per week. The weekly hours then trends downward for the rest of their working life.

The goal of this paper is to provide an explanation to this aggregate pattern. Specifically, I focus on understanding the life-cycle pattern after the first peak. The rise of labor supply before the first peak is mostly due to women of different educational attainments entering labor force at different ages, and hence of less interest.¹ The large dip between the two peaks and the uneven heights of the two peaks are more puzzling; therefore they are my focus.

Employing a survey data set, I start my analysis by documenting relevant facts underlying the aggregate pattern. First, I extract from the data set a panel of female cohort who were born between 1965 and 1969 and were followed from age 26 to 41. Although a short panel relative to women’s entire working life, it provides insights into female labor supply behavior during the period corresponding to the segment between the two peaks in the aggregate pattern. These insights help to understand the mechanisms that form the whole life-cycle pattern.

Next, I group the 1965-69 cohort by women’s marriage and fertility status, and, if they are mothers, by the age of their children. I then examine the underlying life-cycle pattern for each subgroup.

¹For later cohorts, another reason for the rise of labor supply is that employed teenagers mostly work part-time.
Chapter 1. Explaining Life-Cycle Female Labor Supply in Japan

Figure 1.1: Weekly Hours Worked per Woman by Cohort


Note: 1) Weekly hours worked per woman is constructed using data on employment rate and average weekly hours worked by employed women; 2) The age profiles are constructed using repeated annual aggregate data so the cohorts are “synthetic”.

that throughout the observed ages single women work relatively stable annual hours and on average they work much more than married women.\(^2\) Married women work less before their early 30s but they gradually work more afterward (though they still work much less than single women). Among married women, those who have kids work less than those who don’t, and those who have preschool kids (before elementary school) work less than those who have grown-up kids (elementary school or above).\(^3\) Further, perhaps surprisingly, married women who do not have kids still work much less than single women. The evidence indicates that marriage, fertility and child-rearing on average all reduce female labor supply.

From a life-cycle perspective, since most women get married and afterward soon have kids, the above observations suggest that most women change their labor supply behavior at different stages of their lives depending on their marital and fertility status. Meanwhile, in aggregate, the demographic composition of the cohort also changes along the life cycle. For example, the fraction of single women decreases as the cohort becomes older. Putting the two observations together, I show that the segment between the two peaks in the aggregate pattern can be understood as a result of both the demographic composition change and the labor supply behavioral differences across women of different types (or women at different stages of their lives). The fall in labor supply after the first peak is formed as more and more single women get married and afterward reduce their work hours due to either marriage itself or child rearing.

---

\(^2\) The annual hours refers to average annual hours, averaged over both workers and non-workers. Therefore, the labor supply behavior described is average behavior of a specified group of women such as single women, married women and etc.

\(^3\) In the paper, kids before entering first grade are referred to as preschool kids. They are mostly between age 0 and 5. Kids in elementary school or above are referred to as grown-up kids.
or both. The recovery after the drop is associated with mothers, who are the majority of married women, gradually increasing their work hours as their kids grow older. However, because married women on average work much less than single women and the fraction of married women is much larger when the cohort of women are in their 40s than in their 20s, the second peak in the M-shaped aggregate pattern is lower than the first one.

The empirical facts point to two potential reasons why married women on average work less than single women even in their 40s when majority of them, who are mothers, increase their work hours as their kids grow older. One is the pure marriage effect as evidenced by the fact that married women without kids work much less than single women. The other is that mothers with grown-up kids likely work less than they do before their kids are born. The indirect evidence for it is that on average married women with grown-up kids work less than married women without kids.

To account for the empirical facts, I then develop a life-cycle model of female labor supply. Comparing with a standard representative agent life-cycle model with saving and labor choices, my model has several extra ingredients. First, it features heterogeneous types of women: single women, married women without kids, married women with preschool kids and married women with grown-up kids. A typical woman starts as single and she may transit from one type to another during her life cycle. For example, a single woman may get married and then have kids depending on the exogenous marriage and fertility shocks she receives, and a married woman with preschool kids becomes a married woman with grown-up kids as her kids grow up. The transitions between types lead to changes in demographic composition of the cohort. In addition, they also induce women's labor supply behavioral changes. For example, as a single woman gets married, she forms a new family and receives a source of income from her husband. As a result, her labor supply changes accordingly.

Second, I introduce childcare costs. I assume that husbands always work, so mothers with small kids face a tradeoff: they can stay at home to take care of their kids and hence foregone wages but avoid paying the childcare cost, or they can work and earn wages but pay for childcare. This tradeoff essentially lowers the effective wages for mothers with small kids, and thus lowers their labor supply due to standard mechanisms. As the tradeoff disappears when their kids grow up, mothers start to work more. The childcare cost is thus one main ingredient of the model to account for the dip and the recovery of the hours worked in the aggregate life-cycle pattern.

Third, the model also features human capital accumulation. A woman enters labor market with an initial human capital and it then evolves according to a learning-by-doing process. In particular, human capital in the next period depends on current period human capital, hours worked and learning ability. Under this setting, labor choice today affects future wages through human capital accumulation. The learning-by-doing process provides a mechanism to explain why married women with grown-up kids work less than they do before their kids are born. As mothers with small kids work less hours, they accumulate less human capital. Consequently, after their kids grow up, they face relatively low wages, which discourage their labor supply. In aggregate, the result is a lower second peak in the life-cycle pattern. Thus, the human capital channel is essential to explain the uneven height of the two peaks.

Embedded in the model are three key determinants of female labor supply that I choose to focus on in this paper: initial gender wage gap, return to experience and childcare cost. After a woman gets married and forms a family, one key determinant of her labor supply choice is her wage relative to her husband's. Japan has a large gender wage gap. Although it has narrowed steadily for the past few decades, the

\[4\text{In this paper, mothers refer to married women with kids.}\]
gender wage gap measured as ratio of female to male hourly wages was still about 70% in 2000s, a number that is almost twice that of OECD average (see Abe, 2010 and OECD, 2014). Therefore, the wage gender gap is a factor that cannot be ignored in explaining the aggregate labor supply pattern. In a life-cycle model with human capital accumulation, gender wage gap evolves depending on the initial gap and rate of return to work experience. Hence they are the two determinants I focus on in this paper. The negative effect of childcare cost on labor supply of mothers with small kids is well-known and empirically shown in many countries. In Japan, Oishi (2002) finds a significant negative impact of nursery fees on labor force participation of mothers, an elasticity about $-0.60$. Her finding is consistent with what’s observed in the 1965-69 cohort panel. Women with preschool kids work the least comparing to single women, and mothers without kids or with grown-up kids. Since the period around the trough of the aggregate pattern largely corresponds to child-rearing ages of most mothers, it is natural to examine the effect of the childcare cost in my model as well.

The model is calibrated to match statistics of the 1965-69 female cohort. After the calibration, I assess how well the model performs. I show that the model can account the aggregate life-cycle pattern well. It also captures the main features of the underlying labor supply patterns for different types (or subgroups) of women. Overall, the theory built into the model is capable of explaining the life-cycle female labor supply in Japan.

Taking the calibrated model as a baseline, I then experiment on how changes in economic environment, specifically changes in the three key determinants, affect labor supply of different types of women and life-cycle labor supply in aggregate. The experiment has two purposes. First, it quantifies to what extent each of the determinants contribute to the formation of the aggregate pattern. Second, it serves as a potential guidance to authorities that wish to implement polices to promote female labor supply.

The experiment shows that lowering childcare cost reduces the magnitude of the fall in work hours during women's early 30s but it does not help too much on the recovery, i.e., it mainly makes the trough between the two peaks in the aggregate pattern shallower but it does not bring even the two peaks. Narrowing gender wage gap (through lowering the initial gap or increasing the rate of return to experience), on the other hand, do double duties. It alleviates the magnitude of the fall in hours as well as boosts the recovery, bringing the second peak relatively even to the first one. Comparing to lowering childcare cost, the positive effect of narrowing gender wage gap on labor supply also lasts longer, well into the final stage of women's working life. As a result, narrowing gender wage gap has a much larger effect on average female labor supply over the life-cycle than lowering childcare cost. In terms of the formation of the life-cycle pattern, the above results suggest that, relatively speaking, childcare cost is more associated with the “valley” part of in the “M” pattern, and gender wage gap is more responsible for the lower second peak.

The experiment results can be mainly explained by the human capital channel. Reducing initial gender wage gap provides an incentive for all women to work more in their early careers. More early human capital accumulated lead to higher wages in their future careers, which in turn give them further incentives to work more. Reducing initial gender wage gap thus takes full advantage of the cumulative nature of the learning-by-doing process. For married women, higher human capital and thus higher wages reduces the negative marriage effect on their labor supply. For mothers with small kids, higher

---

5 Similarly, using micro-data sets from 22 countries over the 1985-94 period, Blau and Kahn (2003) find that Japan has the largest gender pay gap among all the countries they studied.
6 See Kalb (2009) for a summary of results from some of the empirical studies in various countries.
7 The elasticity of average female labor supply over the life-cycle with respective to reducing gender wage gap is more than ten times larger than the one with respect to lowering childcare cost.
wages encourage them to use more childcare services instead of working less hours so as to take care of their kids. As a result, they suffer less human capital loss during the child-rearing years, and after their kids grow up, they face relatively high wages so they work more. In aggregate, we thus see a shallower trough and a much better recovery of hours. Similarly, increasing return to experience raises wage growth rate. It also encourages all women to work more at their early careers. The same argument implies that it also has a large positive effect on life-cycle female labor supply.

On the contrary, lowering childcare cost only applies to a particular group of women, namely women with small kids. Therefore, first, it helps little to increase labor supply for other groups of women. Second, most married women with small kids are already in their middle careers. Thus, it fails to utilize fully the cumulative nature of the learning-by-doing process. Third, for mothers with small kids, although both reducing childcare cost and narrowing gender wage gap raises their effective wages and thus encourage them to work more, in the latter case the high effective wages stem from the high human capital accumulated before their kids are born. It thus provides an extra incentive for mothers to work more because the high human capital means more effective learning and better future wages. Combining the above arguments, in aggregate, the effect of lowering childcare cost is a shallower trough but only a slightly better recovery.

The policy implication of these experiment results is clear. Policies that target on narrowing gender wage gap will have a large positive effect on aggregate female labor supply than polices that simply subside childcare.

Literature The topic of female labor supply in Japan has been extensively studied. Most of the studies take empirical micro approaches. Among them, most papers focus on examining how various economic forces affect aggregate female participation rate or hours worked, and they do not address life-cycle female labor supply behavior. A few empirical micro studies take a life-cycle approach. For example, using data on synthetic cohorts constructed from Employment Status Survey, Abe (2011) investigates the impact of equal employment opportunity law enacted in 1986 on female labor supply. Her main focus is on change of participation behavior across cohorts due to the new law, not on life-cycle pattern per se. Ueda (2007) studies marriage, childbearing and labor force participation of fertile-aged women in Japan. She estimates a life-cycle stochastic dynamic discrete choice model using a 5-year long panel data. Given the short panel, her focus is not on explaining the life-cycle pattern either. Fukuda (2006) applies a Bayesian cohort model to decompose life-cycle female labor participation rate into cohort, age and time effects. His main purpose is to identify the three effects confounded in the aggregate data. Since he uses only aggregate data, he ignores the life-cycle patterns at disaggregated levels.

Studies on female labor supply in Japan that use empirical macro approaches are surprisingly rare. Okada (2011) is one of a few examples. He studies the impact of childcare cost on female labor supply and fertility decision. He wants to explain the long term trend of rising female labor force participation and falling fertility seen in Japan since 1980. His focus is thus on the trend instead of the life-cycle pattern.

My paper contributes to the literature on female labor supply in Japan in several dimensions. First, I focus on understanding the life-cycle pattern of female labor supply. I use not only macro-data to

---

8The inter-temporal link implies that reducing childcare cost also affects labor supply of other women, not just women with small kids, but the effect is small in the calibrated model.

9For example, see Takahashi et al. (2009a) and references therein for a large literature on wage elasticity for married women, and see Sasaki (2002) for a study on family structure and female labor participation.
identify the aggregate pattern, but also micro-data to document empirical facts at dis-aggregate levels in order to discover the underlying causes of the aggregate pattern. Second, most previous studies are interested in female participation rate, but I focus on labor supply in terms of hours per woman. In section 2, I discuss the importance of using hours per woman as the labor supply measure in the context of Japan. Third, this paper takes an empirical macro approach, which complements the large body of micro studies in terms of methodology.

This paper also broadly relates to the large literature on female labor supply in the U.S. In particular, a strand of recent literature studies the causes of rising female labor supply in the U.S. during the last century, for example, see Attanasio et al. (2008), Olivetti (2006), Greenwood et al. (2005) and Jones et al. (2003). Among them, this paper is most related to Attanasio et al. (2008) and Olivetti (2006). I largely follow the modeling framework developed in these two papers. Attanasio et al. (2008) build a life-cycle model to study the determinants of the increase in female labor supply from the 1940s cohort to the 1950s cohorts. Since most of the change in labor supply between the two cohorts comes from the change in married women’s participation rate, they model only married women and their participation margin. Their model features saving, human capital accumulation, childcare cost and income uncertainty. The labor supply determinants they investigated are childcare cost, gender wage gap and return to experience. They find that a combination of reduction in childcare cost and gender wage gap is needed to account for the labor supply change between the cohorts. Olivetti (2006) builds a four-period life-cycle model to study the role of returns to experience in explaining the increase in hours worked by US women from 1970s to 1990s. She also only models married woman. Her model features saving, human capital accumulation, home production (of childcare) and child quality in utility. She finds that change in women’s returns to experience can account most of their increase in hours worked over the life cycle.

My paper takes a similar modeling framework as those two papers. In particular, the life-cycle model in this paper also features human capital accumulation, a key element both those papers emphasize. I study similar determinants on female labor supply: gender wage gap and return to experience. However, I also make several changes to suit my study. I model both single and married women including the transition from one to the other. Different from Attanasio et al. (2008) but similar to Olivetti (2006), I take hours per woman as my labor supply measure and I abstract away income uncertainty. Different from Olivetti (2006) but similar to Attanasio et al. (2008), I do not model labor choices of husbands, home production and child quality.

The paper proceeds as follows. In the next section, I document empirical facts underlying the aggregate pattern. In section 3, I present a static model and a two-period model to illustrate the main features and mechanisms that will be built into the full model. I then present a full life-cycle model in section 4. In section 5, I discuss the calibration, and section 6 the results. Section 7 concludes.

1.2 Empirical Facts

The goal of this section is to further document relevant empirical facts underlying the aggregate life-cycle pattern. To do so, I employ a national survey data set, Japan Panel Survey of Consumers (JPSC). In the following, after briefly describing the data set, I present facts obtained by relating women’s labor supply choices with their marriage, fertility and child-rearing behavior. I then show that the segment

\footnote{See Killingsworth and Heckman (1987) for a survey on the early literature.}
between the two peaks in the aggregate M-pattern can be viewed as a result of two dynamics along the life-cycle: demographic composition changes and labor supply behavioral differences across women of different types. The labor supply measure taken in this paper is hours worked per women. At the end of this section, I present evidence to support this choice.

1.2.1 The micro-data set

The micro-data set used in this study is Japan Panel Survey of Consumers (JPSC). It is administered by the Institute for Research on Household Economics in Japan. The Institute conducted its first wave of the survey in 1993 with a nationwide sample of 1500 women aged 24 to 34, i.e. women born between 1959 and 1969. These women have been followed annually since then, and information on various aspects of their lives is collected at each wave. In 1997, 2003 and 2008, three new samples of women in their middle to late 20s were added. At the time of this writing, 16 waves of data (from 1993 to 2008) are available.

For the purpose of this study, I only use a sub-sample of the JPSC. In particular, I extract from the JPSC a panel of women who were born between 1965 and 1969. I choose the 1965-69 cohort for three reasons. First, the choice allows me to construct a 16-year panel, the longest possible given the data I obtained. Second, in order to minimize the cohort effect and focus on the life-cycle perspective of labor supply, I restrict the sample to women that are born within a 5-year span. Third, the 1965-69 cohort are between age 24 and 28 in the first wave (1993) and between 39 and 43 in the last wave available (2008). This age span covers well the most interesting part of the aggregate pattern, the segment between the two peaks of the M-shaped curve.

Using this panel has several advantages. First, the panel follows the same sample of women over years, allowing me to study their true life-cycle labor supply behavior. Second, the detailed survey on their work life conditions enables me to correlate their labor supply behavior with their various characteristics, a necessary step to search for potential determinants of female labor supply. Third, the representative sample provides reliable aggregation in different dimensions. Thus, I can study underlying labor supply patterns for different subgroups of women. The panel also has its limitation. Its length is relative short from a life-cycle perspective. It covers an age span of 16 years, which is about one-third of a typical woman's working life.11 However, as I have already pointed out, it covers the most interesting segment of the life-cycle. Analyzing labor supply behavior during this part of their life can provide important clues to understand the whole aggregate life-cycle pattern.

To prepare the data for analysis, I group the chosen sample by age. This gives me a cohort of women with an age span from 24 to 43. Since there are too few observations at both ends of the age span, I drop two years of observations at each end and keep data for women between age 26 and 41. Table 1.5 in Appendix 1.B reports number of observations by age. Details of data construction are also discussed there.

1.2.2 Labor supply and marriage

The fall of the aggregate labor supply after the first peak corresponds to a period when most women get married. The transition from single to married may affect many aspects of women’s life including their

---

11Micro-data from annual census that covers long period of time are not publicly available. Otherwise, one may be able to construct a synthetic cohort that covers the whole life cycle.
labor supply behavior. I therefore start by relating women's labor supply choices to their marital status. Figure 1.2 displays three age profiles of average annual hours worked: one is averaged over all women, and the other two are averaged over single and married women respectively. Let’s start with the profile for all women. As expected, it exhibits a U-shape, consistent with the one between the two peaks in the aggregate M-pattern. On average, women work about 1100 hours annually at age 26. The annual hours then gradually drops and reaches the bottom at about 750 hours in their early 30s. Afterward, it starts to rise. At age 41, women again work about 1100 hours annually. Turning to the profiles for single and married women. Single women work about 1600 hours annually and the hours is quite stable for all ages observed. In sharp contrast, married women on average work much less. Their hours worked also has large changes over the ages observed. They only work about 500 hours annually between their middle 20s and early 30s. From then on, they gradually work more. Their annual hours worked rises to about 900 at age 41, although it’s still nowhere near that of the single women.

![Figure 1.2: Annual Hours Worked by Marital Status](image)

Data Source: JPSC 1965-69 Cohort
Note: The number next to the connected dots of the age profile of all women indicates the marriage rate.

At any given age, the average annual hours worked of all women is a weighted average of those of single and married women. The weight on the hours of married woman is the age-specific marriage rate defined as percentage of women married at a given age. The weight on the hours of single woman is one minus the marriage rate. In figure 1.2, the marriage rate is indicated next to the connected dots of the age profile for all women. As can be seen, most women get married before their early 30s and the marriage rate hardly changes after about age 33. Put together the evidence on the marriage rate and the difference in labor supply between single and married women, the fall and rise of the labor supply between the two peaks can be explained by the following mechanism. Most women are single in their early 20s, so more weight is on the high hours worked of single women. The hours of all women is,
therefore, relatively high. Gradually as more and more women get married, the weight shifts to the low hours worked of married women. Hence, the hours of all women falls. The fall continues in this manner until women are in their early 30s, around which the marriage rate begins to stabilize and the weight shifts little afterward. However, at about the same time, married women start to work more, driving up the hours of all women. In summary, one way to understand the segment between the two peaks in the aggregate pattern is to view it as a result of both demographic composition change and labor supply behavioral difference across single and married women. The fall is mainly associated with the increase in the number of married women who work much less than single women, and the rise is mainly associated with the increase in hours worked by married women.

Another feature in the aggregate pattern is the uneven heights of the two peaks: after the hours worked per woman bottoms out it never rises as high as it is at the first peak. Although unable to observe this feature in the micro-panel due to its length, the above discussion on labor supply during the observed ages points to the following potential explanation. At the time of the second peak, there are much more married women than at the time of the first peak. Thus, most of the weight is on the profile of married women. Although the hours of married women starts to rise when they are in early 30s, it is still much lower than that of single women. Therefore, from the first peak to the second one, the fall of the weighted hours of single women dominates the rise of the weighed hours of married women, resulting in a relatively lower second peak.

1.2.3 Labor supply and child-rearing

Why do married women on average work less? Most married women soon have children and child rearing may affect labor supply of mothers. For the 1965-69 cohort, about 84% of married women have kids. Among those who have kids, on average they have two kids and give birth to their first kid two years after marriage. When the kids are small, they require a lot of cares and mothers are usually the one that take care of them the most. In Japan, most husbands work long hours and spend little time with their kids on weekdays.\textsuperscript{12} To look for evidence on the impact of child rearing on female labor supply, I relate married women’s labor supply choices to their fertility status as well as their youngest kid’s age (if they have kids). Figure 1.3 displays age profiles of average annual hours worked for three groups of married women: married women without kid, married women whose youngest kid is at preschool age (preschool kid), and married women whose youngest kid is above preschool age (grown-up kid).\textsuperscript{13} Again, all annual hours are averaged over both workers and non-workers.

Among married women, on average married women with kids work less than married women without kid. Among married women with kids, those who have preschool kids work less than those who have grown-up kids. The evidence is consistent with the hypothesis that mothers work less to take care of their kids, especially when their kids are small.

Moreover, comparing the annual hours of married women without kid to that of single women (see figure 1.2), the hours of married women without kid is much lower, about two-thirds of that of single women. This evidence suggests a “pure” marriage effect on female labor supply, an effect that is not due

\textsuperscript{12}According to NHK time used survey in 2000, only 6% of adult men take physical care of their children for more than 15 minutes on a given weekday, while the number is 20% for adult women. The overall average time (including those not participating the activity) spent by men in taking physical care of their children is 5 minutes on a given weekday, while the number for women is 41 minutes.

\textsuperscript{13}Recall that in this paper I define preschool age as from birth to the last year before entering elementary school. I refer to kids in this age span as preschool kids, and kids above this age span as grown-up kids.
to child-rearing. Married women make labor supply choices in a family context. The marriage effect may be partially explained by their optimal choices taken into account the incomes from their husbands and the economies of scale within a family. Another explanation might be culture. Traditionally, women in Japan are expected to become full-time housewives once they get married regardless whether they have children. Although such culture aspect is largely abandoned in modern Japan, it may still linger among some families.

The age profile of married women in figure 1.2 can be viewed as a weighted average of the profiles of the three types of married women in figure 1.3. Similar as the previous analysis on age profile of all women, the age profile of married women can also be explained as a result of demographic changes and labor supply behavioral differences between the three types of married women. When the cohort of married women are in their late 20s, most weights are on married women with small kids who work less hours. Starting from their 30s, the weight gradually shifts to married women with grown-up kids who work relatively more hours. The result is a rising age profile of married women see in figure 1.2.

Recall that married women start to work more in their 30s but they still work much less than single women (see figure 1.2). The evidence above suggests two potential reasons. First, it’s the pure marriage effect as evidenced by the fact that married women without kids work much less than single women. Second, mothers with grown-up kids work less than they do before their kids are born. The indirect evidence for it is that married women with grown-up kids work less than married women without kids.

Each of the three age profiles also has its dynamics. For example, the profile for married women with
small kids rises over the observed ages. In this paper, I do not further investigate these dynamics. I emphasize only the level differences between the three age profiles.

1.2.4 Intensive and extensive margins

The age profiles plotted so far all take hours per woman (or per woman of a specified group) as the labor supply measure. This measure implicitly incorporates both extensive margin (female employment rate) and intensive margins (hours worked by employed women). A natural question to ask is how each margin contributes to the observed life-cycle pattern. In the following, I explore the facts on the two margins in aggregate.

In figure 1.4, for the same four cohorts of women as in figure 1.1, I plot two sets of age profiles. The left panel displays the age profiles of employment rate (extensive margin) and the right panel age profiles of average weekly hours worked by employed women (intensive margin). The figure reveals rich dynamics of the two margins. On the extensive margin, the main feature of the age profile is its M-shape. Comparing this “M” with the one in figure 1.1, what’s similar is the timing of the occurrence of the two peaks as well as the significance of the dip in between. What’s different is that this “M” has similar heights at the two peaks while the “M” in figure 1.1 doesn’t. In other words, the employment rate recovers fully after the dip while the hours worked per woman doesn’t. The reason for this lies in the intensive margin. On the intensive margin, the average weekly hours worked by employed women peaks when they are in their 20s and then decreases almost monotonically for the rest of their working life. Thus, while the full recovery of employment rate after its dip is quite remarkable, the recovery is dampened by the decrease in average hours worked by employed women. As a result of the dynamics from the two margins, we observe an “M” with a lower second peak in figure 1.1.

Because the two margins are very active, in order to understand the aggregate labor supply pattern, it is not adequate to study either margin alone. While it is attempting to have a theory that explicitly account for the two margins, doing so would require a rather complicated model. The hours per woman measure captures both margins in a convenient way. Moreover, the key mechanisms I intend to emphasize in this paper can be studied without explicitly modeling the two margins. Therefore, I choose hours per woman as the labor supply measure so as to keep things simple yet still suffice to capture and study the core of the life-cycle pattern.

1.3 Two simple examples

In this section, I lay out two simple examples in order to illustrate the main features that will be built into the coming full life-cycle model. The examples also help to build intuition on the key mechanisms that drive the results in the full model. I first present a static model to show how labor supply behavior may differ across different types of women characterized by their marital status and, if they are married, by whether they have small kids. I then extend the static model to a two-period setting to emphasize the inter-temporal tradeoff of labor supply and human capital accumulation.

---

14 A number of factors may contribute to the rise. For example, the number of mothers having more than one preschool kids may decrease over time, and the average age of preschool kids may also get larger over time. As a result, those mothers may have less burden of taking care of small kids and start to work more in later years.

15 In appendix 1A.2, I also discuss facts on the two margins by marital status.

16 This M-shaped curve is well known and it is often referred to as Japan’s “M-curve”.
1.3.1 A static model

In the static setting, I assume four types of women in the economy: single women, married women without kids, married women with a preschool kid, and married women with a grown-up kid (i.e., above preschool age). Each type has a representative woman. The mass of each type is not specified as it is not important for the purpose of this illustration. In the following, I first describe the utility maximization problem for each type, and then characterize and compare their optimal labor supply. I show that the static model can deliver most of the cross-sectional differences in labor supply behavior between different types of women seen in the data.

**Type 1: a single woman household** Consider a single woman who values her consumption and leisure. Her utility is given by

\[ u(c, l) = \ln c + \alpha \ln l, \]

where \( c \) is consumption and \( l \) is leisure. The parameter \( \alpha > 0 \) measures how she values leisure relative to consumption. Total time endowment is normalized to 1. Her budget constraint is thus

\[ c = w(1 - l), \]

where \( w \) is her wage, and \( 1 - l \) is her labor supply. The single woman maximizes her utility by choosing consumption and leisure, subject to the budget constraint. The problem has a simple solution. Her

---

17 I do not consider single mother in this paper as they are a tiny fraction in the data. Mothers in this paper refer to married women with kids.
optimal leisure choice is
\[
l = \frac{\alpha}{1 + \alpha}.
\]
The optimal labor supply is thus \(1 - l = \frac{1}{1 + \alpha}\). Note that it does not depend on wage. This is because under the log utility, we have a special case where income and substitution effects cancel out. The optimal labor supply only depends on how much she values her leisure, which is governed by \(\alpha\). A higher \(\alpha\) leads to a lower labor supply.

**Type 2: a married woman without kids** Now, consider a married woman who lives with her husband. The couple has no kids. To keep the analysis simple, I assume a unitary household framework, where the couple pool their resources to maximize a common family utility. The family utility of the married woman needs to be a natural extension to the utility of the single woman, otherwise a comparison of labor supply behavior between the single and married women would be rather arbitrary. To do so, I make a few assumptions on the family utility. First, it is a weighted sum of the wife and husband’s individual utility, where the wife’s utility is the same as the single woman’s and the husband’s utility has the same functional form as the wife’s. Second both the husband and wife value average household consumption (as opposed to their own consumption). Third, the husband always works full-time and his labor supply (or equivalently leisure) is exogenous. Under these three assumptions, the family utility is then given by

\[
u(C, l) = \theta \left( \ln \frac{C}{e(s)} + \alpha \ln l \right) + (1 - \theta) \left( \ln \frac{C}{e(s)} + \alpha_h \ln \bar{l}_h \right),
\]

where \(0 < \theta < 1\) is the weight on the wife’s utility, and \((1 - \theta)\) the weight on the husband’s. \(C\) is the total family consumption, and it is adjusted by an adult equivalent scale \(e(s)\), where \(s\) is the size of the household. For a couple without kids \(s = 2\), indicating that the family has two adults. In general, the equivalent scale depends on the number of adults and children in the household. It conveniently captures the economies of scale within a family. In this case, the equivalent scale is \(1 < e(2) < 2\), reflecting the economies of scale in consumption. Again, \(l\) is the wife’s leisure and \(\alpha\) measures how she values her leisure. Correspondingly, \(\bar{l}_h\) is the husband’s leisure, which is exogenous by assumption, and \(\alpha_h\) is a parameter measuring how the husband values his leisure.

Because I assume that the husband’s leisure is exogenous, for the purpose of utility maximization, the family utility can be further reduced to\footnote{Given the log utility, the equivalent scale \(e(2)\) can be dropped as well. I don’t do so as I will soon deviate from the log utility and show how the equivalent scale matters for the optimal labor supply.}

\[
u(C, l) = \ln \frac{C}{e(2)} + \theta \alpha \ln l.
\]

Define \(y = w_h(1 - \bar{l}_h)\) the husband’s earning (labor income), where \(w_h\) is the wage of the husband. The budget constraint for the family is

\[
C = w(1 - l) + y,
\]

that is, total family consumption must be equal to the sum of wife’s labor income and husband’s earning. Since both husband’s labor supply and wage are taken as given, his earning is also exogenous.

The family maximizes utility (1.1) subject to budget constraint (1.2). This problem is very similar to the one faced by the single woman. The differences in the objective function are the per-adult equivalent consumption \(\frac{C}{e(2)}\) and the family utility weight \(\theta\) on the leisure term. The parameter \(\theta \alpha\) can be viewed
as how a single woman values her leisure after she gets married. The difference in the constraint is the
exogenous earning of the husband. The optimal leisure choice is given by

\[
    l = \begin{cases} 
    \frac{\theta_\alpha}{1+\theta_\alpha} \left( 1 + \frac{y}{w} \right) & \text{if } 0 < y < \frac{w}{\theta_\alpha} \\
    1 & \text{if } y \geq \frac{w}{\theta_\alpha} 
    \end{cases}, \tag{1.3}
\]

Because the husband’s earning \( y > 0 \), corner solution is possible.

Similar as the labor supply of the single woman, the labor supply of the married woman without kid depends on how much she values leisure, which is now governed by \( \theta_\alpha \). A higher \( \theta_\alpha \) leads to less labor supply (assuming interior solution). Moreover, her labor supply also depends on the ratio between her husband’s earning and her wage. She works less if the ratio is larger. The ratio reflects a marriage income effect on the wife’s labor supply choice. It also relates to gender wage gap as it can be written as the wage ratio between husband and wife times the husband’s hours worked (which is exogenous), 

\[
    \frac{w}{w} = \frac{w}{w} \left( 1 - \bar{l}_h \right).
\]

As for a comparison, it can be easily show that the married woman with no kids works less than the single woman if and only if \( \frac{1-\lambda}{\theta(1+\lambda)} < \frac{w}{w} \). Therefore, the married woman without kids may work more or less than the single woman depending on her value of leisure (\( \alpha \)), the family utility weight (\( \theta \)) and the marriage income effect (\( \frac{w}{w} \)). All else equal, the larger is the value of leisure, the weight or the marriage income effect, the more likely that the married woman without kids works less.

**Type 3: a married woman with a preschool kid** Now consider a married woman who lives with her husband. The couple has one preschool kid. I assume that having a kid does not change the family utility except a change in equivalent consumption scale. Hence, the reduced family utility (equation 1.1) becomes

\[
    u(C,l) = \ln \frac{C}{e(3)} + \theta_\alpha \ln l,
\]

where \( e(3) \) is the adult equivalent scale for two adults and one kid. Since the kid is at preschool age and the husband works full-time (as in the type 2 case), if the woman chooses to work, she has to arrange childcare service for her kid.\(^{19}\) The price of the childcare is \( p \), so if the wife works for \( (1-l) \) unit of time, she pays \( p(1-l) \) for childcare. The budget constraint is then given by

\[
    C + p(1-l) = w(1-l) + y.
\]

Comparing with the woman without kids, the women with a preschool kid faces a similar budget constraint. The only difference is that her effective wage becomes \( w - p \) due to the childcare cost. Since the log utility renders the equivalent scale inessential in the optimization problem, the optimal leisure can be obtained by simply replacing the wage \( w \) with the effective wage \( w - p \) in the previous solution (equation 1.3). Assuming \( w - p > 0 \), the optimal leisure is thus

\[
    l = \begin{cases} 
    \frac{\theta_\alpha}{1+\theta_\alpha} \left( 1 + \frac{w}{w-p} \right) & \text{if } 0 < y < \frac{w-p}{\theta_\alpha} \\
    1 & \text{if } y \geq \frac{w-p}{\theta_\alpha} 
    \end{cases}.
\]

The weight \( \theta_\alpha \) and the ratio between her husband’s earning and her effective wage \( \frac{w}{w-p} \) affect her

\(^{19}\)I assume that all jobs are day time jobs.
labor supply in a similar way as in the type 2 case. She works less if the weight or the ratio are higher (assuming interior solution). Assuming wages and husbands’ earnings are the same in the type 2 and type 3 cases, the married woman with a preschool kid always works less than the married women without kids because of the former’s lower effective wage. If \( \frac{w_{\theta a}}{\theta a} \leq y < \frac{w}{\theta a} \) is satisfied, the former stays at home, but the latter works positive hours.

**Type 4: a married woman with a grown-up kid** A married woman who lives with her husband and has a grown-up kid (elementary school or above) faces the following maximization problem,

\[
\max_{C,l} \frac{C}{e(3)} + \theta a \ln l
\]

s.t. \( C = w(1-l) + y \).

The tradeoff between working and staying at home to take care of the kid disappears since the kid goes to school and does not need childcare service.

The only difference between this problem and the one faced by the married woman without kids (the type 2 woman) is the consumption equivalent scale. Because of the log utility, the difference does not lead to a different optimal solution. Therefore, if we assume that both types of women face the same wages and their husbands have the same earnings, their labor supply should be exactly the same. However, as seen in figure 1.3, on average mothers with youngest kid attending elementary school or above (\( \geq 6 \)) work less than women without kids. The current simple model is inconsistent with the data in this dimension.

The observation in the data may result from a few factors. The current model can be easily modified to incorporate them and resolve the inconsistency. First, working mothers with kids attending elementary school may still need to enroll them in after school care, so the trade-off between working and staying at home to take care of kids may still exist.\(^\text{20}\) Thus, similar as married women with preschool kids, it may not be too surprising to observe that on average mothers with youngest kid attending elementary school or above work less than married women without kids (especially in the micro-data the oldest mothers observed are only 41 and many of them still have their kids in early years of elementary school). Adding this factor into the model is easy. Suppose the woman with a grown-up kid also faces a childcare price \( p' \), but the price is lower than the one faced by the woman with a preschool kid, i.e., \( p > p' \). Under this assumption, she works less than the woman without kids and more than the woman with a preschool kid.

Second, in the data, women with grown-up kids may on average face lower wages than married women without kids, so they works less. The reason a woman with a grown-up kid may face a lower wage can be better understood in a dynamic setting with human capital accumulation. When her kid was at preschool age, she worked less than a married woman without kids. If human capital is accumulated through learning-by-doing, less hours worked means less human capital accumulated. Thus, when her kid grows up, her human capital stock would be lower than that of the married women without kids. Consequently her wage would be lower, discouraging her labor supply. To incorporate the human capital mechanism into the current static setting, I can simply assume that the woman with a grown-up kid faces a lower wage \( w' \) than the woman without kids. Under this assumption, the married woman with

---

\(^{20}\) There is no legal age limit for leaving children on their own in Japan, but few parents would want to leave their grade one kids unattended after school.
a grow-up kid works less than the married woman without kids.

Third, economies of scale may also be a reason that we observe women with grown-up kids on average work less than women without kids. As already noted, the log utility renders the equivalent scale irrelevant in the maximization problem. To add this factor into the model, suppose the reduced family utility has a somewhat general form,

$$u(C, l) = \left(\frac{C}{e(s)}\right)^{1-\gamma_c} + \theta\alpha \frac{l^{1-\gamma_l}}{1-\gamma_l},$$

where the curvatures $\gamma_c > 0$ and $\gamma_l > 0$.\(^{21}\) Again, $e(s)$ is the equivalent scale, where $s = 2$ for the married woman without kids and $s = 3$ for the married woman with a grown-up kid. Substituting the family consumption in the utility with the budget $C = w(1-l) + y$, the first order condition (FOC) with respect to $l$ (assuming interior) is

$$\theta\alpha l^{-\gamma_l} = e(s)^{\gamma_c-1}[w(1-l) + y]^{-\gamma_c} w.$$

The left-hand side (LFS) is the marginal benefit of enjoying an extra unit of leisure and the right-hand side (RHS) is the marginal cost of it, the forgone wage evaluated at marginal utility of consumption. Now, the equivalent scale clearly plays a role. Rewrite the FOC as

$$e(s)^{\frac{\gamma_c-1}{\gamma_c}} = [w(1-l) + y]^{\frac{1}{\gamma_c}} (\theta\alpha)^{\frac{1}{\gamma_c}} \left(\frac{1}{\theta\alpha}\right)^{\frac{1}{\gamma_l}} l^{\frac{\gamma_l}{\gamma_l}}.$$

Assuming $0 < \gamma_c < 1$, it is easy to see that the RHS is strictly decreasing in $l$ and the LFS is strictly decreasing in $e(s)$. Since $e(3) > e(2)$, at the optimal the married woman with a grown-up kid enjoys more leisure than the married woman without kids, and hence works less.

Taking stock Thus, I have shown that under reasonable assumptions the simple static model can deliver cross-sectional labor supply behavior of different types of women that are consistent with the facts in the micro-data. The static model also demonstrates the effect of many determinants of female labor supply. Specifically, it highlights the effect of family utility weight, economies of scale within family, marriage income effect (or gender wage gap), and childcare cost. Among them, the first three are related to the pure marriage effect as implied by the data. The childcare cost and gender wage gap are the key determinants I plan to focus on in this paper.

1.3.2 A two-period model

Having sorted out a single period model, I now move to a two-period one, where women may change types between the two periods. First, I study a two-period model with no human capital accumulation. I start with a woman who is single in the first period and married without kids in the second period. Using this setting, I demonstrate the inter-temporal tradeoff of labor supply and the marriage income effect along life cycle. I then briefly discuss other scenarios of transitions in types. Second, I add human capital accumulation into the simple two-period model to illustrate an additional effect due to the human capital channel.

\(^{21}\)I assume that the underlying individual utility of the husband and wife has the same parameter $\gamma_c$, so the family utility can again reduces to this simple form.
**A two-period model with no human capital accumulation** Consider a woman who is single in the first period but becomes married without kids in the second period. There is no uncertainty. Her life-time utility is her individual utility when she is single in the first period plus her discounted family utility (in reduced form) in the second period. The utility is thus given by

$$\ln c_1 + \alpha \ln l_1 + \beta (\ln \frac{C_2}{c(2)} + \theta \alpha \ln l_2),$$

where $\beta$ is the discount rate and the subscript indicates the period, either 1 or 2. All other variables are as explained in the static model. The budget constraint is

$$c_1 + \frac{C_2}{R} = w_1 (1 - l_1) + \frac{w_2}{R} (1 - l_2) + \frac{y_2}{R},$$

where $R - 1$ is the real interest rate. It states that, in present value, total consumption must equal total earnings. She maximizes the life-time utility subject to the budget constraint. Assuming interior solution, FOCs imply

$$\frac{1}{c_1} = \beta R \frac{1}{C_2},$$

$$\frac{1}{l_1 w_1} = \beta \theta \frac{1}{l_2 w_2},$$

$$\frac{1}{c_1} = \frac{\alpha}{l_1 w_1}.$$

The first two are inter-temporal Euler conditions for consumption and leisure, and the third is the intra-temporal condition between consumption and leisure in period 1. The optimal leisure (interior solution) can be solved as

$$l_1 = \frac{\alpha}{(1 + \alpha) + \beta (1 + \theta \alpha)} \left( 1 + \frac{w_2}{R} / w_1 + \frac{y_2}{w_1} / w_1 \right)$$

$$l_2 = \frac{\beta \theta \alpha}{(1 + \alpha) + \beta (1 + \theta \alpha)} \left( 1 + \frac{w_2}{R} / w_1 + \frac{y_2}{w_2} / w_2 \right).$$

I emphasize two inter-temporal links achieved by means of borrowing and saving. One is the inter-temporal trade-off of leisure (or labor supply) as demonstrated by the Euler equation of leisure. The Euler equation essentially states that the woman must be indifferent between enjoying one extra unit of leisure today or “saving” that unit and enjoying it tomorrow. Since wage is the “cost” of leisure, the relative leisure between the two periods depends on the relative wages between the two periods. If the relative wage $\frac{w_2}{w_1}$ increases, the woman would substitute leisure in the second period with the one in the first period, i.e. she would increase her relative labor supply in the second period. Another inter-temporal link is the marriage income effect along the life cycle – a two-period life cycle in this case. As seen in the expressions of the optimal leisure, her husband’s earning in the second period $y_2$ affects her optimal leisure in both periods. For a given period, the higher the ratio between her husband’s earning and her wage (in period 1, her husband’s earning is discounted to that period), the more she enjoys leisure and the less she works.

It’s easy to compare her labor supply behavior with that of a woman who stays single in both periods. The latter’s maximization problem is the same as the former’s with $y_2 = 0$ and $\theta = 1$. Thus, the latter’s
optimal leisure is just the former’s evaluated at $y_2 = 0$ and $\theta = 1$,

$$l_{1s}^* = \frac{\alpha}{(1 + \alpha) + \beta(1 + \alpha)} \left( 1 + \frac{w_2}{R} \right) \frac{w_1}{y_2}$$

$$l_{2s}^* = \frac{\beta \alpha}{(1 + \alpha) + \beta(1 + \alpha)} \left( 1 + \frac{w_1}{w_2} \right).$$

In the first period, the single woman who is going to get married in the second period works less than the woman who stays single in both periods (assuming they face the same wages). In the second period, the comparison depends on the parameter values. The married woman (who is single in the first period) is more likely to work less if the utility weight $\theta$ or the ratio $\frac{w_2}{w_2}$ is larger. The analysis is similar to the one discussed in the static model.

If a woman is single in the first period and married with a preschool kid in the second period, her maximization problem is again the same as the one faced by the single-to-married-without-kids woman except that her second period effective wage becomes $w_2 - p$ where $p$ is the price of the childcare service. So, her optimal leisure is

$$l_{1k}^* = \frac{\alpha}{(1 + \alpha) + \beta(1 + \theta \alpha)} \left( 1 + \frac{w_2 - p}{R} \right) \frac{w_1}{y_2}$$

$$l_{2k}^* = \frac{\beta \theta \alpha}{(1 + \alpha) + \beta(1 + \theta \alpha)} \left( 1 + \frac{w_1}{w_2} \right) \left( 1 + \frac{w_2 - p}{R} + \frac{y_2}{w_2} \right).$$

Comparing with the single-to-married-without-kids woman, she works more in the first period when she is single and less in the second period when she is married and have a preschool kid. The comparison with the single-single woman depends on parameter values.

I do not intend to exhaust all possible type switches between the two periods. The two-period setting mainly introduces two ingredients: the inter-temporal labor supply tradeoff, and the marriage income effect along the life cycle. Many of the results from the static model preserve under the two-period setting given reasonable parameter values.

**The marriage income effect.** In a full life-cycle model with no uncertainty, how large the marriage income effect depends on the life-time earnings of a husband. At a given age, a single woman who will get married in the future works less than a single woman who will stay single for her whole life time. Moreover, a single woman who will get married at an early age works less than a single woman who will get married at a later age because the former will receive a larger life-time earnings from her husband. Thus, the marriage income effect predicts a rising age profile of hours per single woman until the age after which no more single women get married. This prediction is at odd with what is observed in the data, where hours per single woman is relatively stable over the observed age range. The rising profile is not due to the assumption that the model has no marriage uncertainty. In particular, suppose that single women receive exogenous marriage shocks until a certain age. Comparing with older single women, younger single women have a higher probability of getting married in the future and thus have a higher expected life-time income from their potential husbands. Because the marriage income effect is stronger for younger singles than for older singles, younger singles work relatively less than older singles. The result is again a rising age profile of labor supply for single women.

In the full life-cycle model, I assume marriage uncertainty. Further, I assume that no borrowing is
allowed. Young single women would want to borrow and work less due to the marriage income effect, and hence the rising age profile. Assuming no borrowing increases young single women’s work hours and bring the age profile closer to what is observed in the data.

**A two-period model with human capital accumulation** Consider again a woman who is single in the first period and married without kids in the second period. Her utility is

\[ \ln c_1 + \alpha \ln l_1 + \beta \ln \left( \frac{c_2}{e^{C_2}} + \theta \alpha \ln l_2 \right). \]

Her budget constraint is

\[ c_1 + \frac{C_2}{R} = \pi h_1(1 - l_1) + \frac{\pi h_2}{R}(1 - l_2) + y_2, \]

where \( h_1 \) and \( h_2 \) are human capital in the first and second period, and \( \pi \) is the rental rate of human capital. The initial human capital \( h_1 \) is given but \( h_2 \) is endogenous. Human capital is accumulated through a process of learning-by-doing described by

\[ h_2 = H(h_1, 1 - l_1). \]

The function \( H(h_1, 1 - l_1) \) is increasing in both initial human capital and hours worked in period 1. Let \( \lambda \) be the Lagrangian multiplier for the budget constraint and \( \mu \) be the multiplier for the human capital accumulation equation. The FOCs w.r.t. \( c_1, c_2, l_1, l_2 \) and \( h_2 \) are

\[ c_1 : \quad \frac{1}{c_1} = \lambda \]
\[ c_2 : \quad \beta \frac{1}{c_2} = \lambda \frac{1}{R} \]
\[ l_1 : \quad \frac{\alpha}{l_1} = \lambda \pi h_1 + \mu H_2(h_1, 1 - l_1) \]
\[ l_2 : \quad \beta \frac{\theta \alpha}{l_2} = \lambda \frac{\pi h_2}{R} \]
\[ h_2 : \quad \lambda \frac{\pi}{R}(1 - l_2) = \mu \]

The FOC w.r.t. \( l_1 \) captures the effect of human capital in the model. The LHS is the marginal benefit of enjoying an extra unit of leisure, and the RHS is its marginal cost. In addition to the cost of forgone wage in the first period, with learning-by-doing, there is an extra cost, the cost of forgone increase in wage in the second period. An extra unit of leisure in the first period reduces, at margin, human capital in the second period by \( H_2(h_1, 1 - l_1) \). This loss of human capital evaluated at its marginal value in the second period, \( \mu \), is the extra cost in term of utility. With human capital accumulated through learning-by-doing, the woman has more incentive to work even when facing a relatively low wage.

Again, I do not intend to exhaust all the possible scenarios of type switch. Predictions under the static model preserve in this setting under reasonable parameter values. The human capital channel is important to explain why married women with grown-up kids work less than married women without kids. Consider a married woman with a small kid in period one and the kid grows up in period two. In the first period, because of the childcare cost and hence the tradeoff between working and staying at home to take care of her kid, she works less than a married woman without kids. This leads to her relatively low human capital stock in the second period when her kid grows up, and thus a lower wage.
than the married women without kids. As a result, although she works more in period two than period one, she still works less than the married women without kids.

\section*{1.4 A life-cycle model}

In this section, I develop a full life-cycle female labor supply model building on the simple models discussed in the previous section. Comparing to a standard life-cycle model with savings and labor supply choices, my model has three extra ingredients. It features heterogeneous types of women similar as those in the static setting and the transitions between them. It also features childcare cost and human capital accumulation. In what follows, I describe the model in details.

\subsection*{1.4.1 Demographics}

I model a cohort of large number of women. Consider a typical woman in this cohort. She begins her life and starts to work at age $b = 21$. Her mandatory retirement age is $r = 60$.\footnote{Since her labor supply is endogenous, she may choose to retire early. In the calibrated model, all women work until their mandatory retirement age.} She dies at age $T = 80$. She starts as a single woman, but she may get married in her early life. I do not model her marriage decision. Instead, I model marriage as exogenous shocks. Between age $s$ and $\bar{m} = 37$, if she is single she receives a marriage shock at the beginning of each period.\footnote{The last period of marriage shock $\bar{m} = 37$ is calibrated to match the data.} The probability of getting married depends on her age. Conditional on being single at age $t$, with probability $p_{m,t}$ she gets married, and with probability $1 - p_{m,t}$ she remains single. If she gets married, she receives no more marriage shocks and stays married for the rest of her life. Divorce is not allowed. If she remains single after she receives her last marriage shock at age $\bar{m}$, she stays single for the rest of her life.

I model men as simple as possible. A man starts to appear in the economy as a husband when a woman marries him, and “disappears” when the woman dies. I assume that a husband is always 2 years older than his wife.\footnote{The 2 year difference is the average number in the data. The difference will have implications on income gaps between the husband and wife.} Further, he works full-time and retires at age 62. Hence, if his wife retires at her mandatory retirement age (60), they retire at the same time, and the couple enjoy their retirement life for 20 years until they die at the same time (wife at age 80 and husband at age 82).

I assume that a single woman never has a kid and a married woman either have no kid or two kids. Again, for simplicity, I model fertility as shocks. At the time of marriage, a woman receives a fertility shock. The fertility shock determines whether the couple will have kids. With probability $p_k$, the couple will have two kids, and with probability $1 - p_k$ they will have no kid. Note that the fertility shock does not depend on her age and only occurs at the time of marriage. If the couple have two kids, the arrival of the kids is deterministic. Their first kid is born 2 year after their marriage and the second kid is born 2 years after the first one is born. I assume that kids become independent and leave the house when they are 19 years old.

Under this setting, at any given time, a woman is identified by her type, which is characterized by her marital status, and if married, at what age she gets married and whether she has kids. The number of types is finite at any given time, and it increases over life-cycle until the marriage and fertility shocks disappear (age $\bar{m}$). Labor supply behavior only differs between types but is the same within a type. Figure 1.5 illustrates a time line for a typical woman, who is married and has two children.
1.4.2 Preferences

A model period is a year. I normalize the initial age to be 1, i.e. age \( t = 1 \) in the model corresponds to age 21 in the data. A woman maximizes her expected lifetime utility,

\[
\max \mathbb{E}_1 \sum_{t=1}^{T-b} \beta^{t-1} u(C_t, l_t; s_t),
\]

where \( \beta \) is the discount factor and \( u(C_t, l_t; s_t) \) is the period utility function. The period utility is given by

\[
u(C_t, l_t; s_t) = \left( \frac{C_t}{e(s_t)} \right)^{1-\gamma_c} + \alpha \theta^{s_t \geq 2} \left( \frac{l_t^{1-\gamma_l}}{1-\gamma_l} \right).
\]

All notations are the same as those in the static model (except the age subscript \( t \)). As discussed in the static setting, the period utility depends on whether a woman is single or married. Under the assumptions discussed there, the period utility for a married woman can be reduced to a similar form as the one for a single woman, up to changes in equivalent scale \( e(s_t) \) and a family utility weight \( \theta \) attached to the leisure. This has been reflected in the above period utility. A single woman has a household of size one \( (s_t = 1) \), and a married woman has a household of size two to four depending on how many kids the couple have at age \( t \) \( (2 \leq s_t \leq 4) \). The indicator function \( \mathbb{1}(s_t \geq 2) \) equals one if \( s_t \geq 2 \) and zero otherwise. Thus, the weight attached to the leisure term depends on a woman’s marital status. If she is single, it is \( \alpha \). If she is married, it is \( \theta \alpha \).

I use the OECD specification for the equivalent scale. The OECD equivalent scale assigns a value of 1 to the first adult member, 0.7 to each additional adult and 0.5 to each kid. Given the demographic setting in this paper (only married couple may have kids and there is no divorce), the equivalent scale can be defined as a function of household size \( s_t \),

\[
e(s_t) = \begin{cases} 
1 & \text{if } s_t = 1 \\
1.7 & \text{if } s_t = 2 \\
1.7 + 0.5(s_t - 2) & \text{if } 2 < s_t \leq 4
\end{cases}
\]
1.4.3 Human capital accumulation

I assume that human capital is accumulated in a process of learning-by-doing. In particular, following Wallenius (2011), I assume that the law of motion of human capital is given by

\[ h_{t+1} = (1 - \delta) h_t + z e^{-gt} (1 - l_t)^\eta h_t, \]

where the initial human capital \( h_1 \) is given. It states that human capital at \( t + 1 \) is human capital at \( t \) net of depreciation (the first term) plus new human capital produced at time \( t \) (the second term). The depreciation rate is \( \delta \). The production of human capital depends on current hours worked \((1 - l_t)\) and current stock of human capital \((h_t)\). The term \( z e^{-gt} \) can be interpreted as productivity of the human capital production. The exponential part \((e^{-gt}) \) captures the fact that learning becomes harder as one gets older. The parameter \( z \) is a level shift.

I assume that the rental rate of human capital \((\pi)\) is constant. Hence, wage growth can be written as \( \frac{w_{t+1}}{w_t} = \frac{\pi h_{t+1}}{\pi h_t} = (1 - \delta) + z e^{-gt} (1 - l_t)^\eta \). Following Olivetti (2006), I define the rate of return to labor market experience as rate of wage growth with respect to hours worked.

\[
\frac{\partial \ln \left( \frac{w_{t+1}}{w_t} \right)}{\partial (1 - l_t)} = \frac{z e^{-gt} \eta (1 - l_t)^{\eta - 1}}{(1 - \delta) + z e^{-gt} (1 - l_t)^\eta}.
\] (1.4)

Therefore, taking \( \delta, g, \) and \( \eta \) as exogenous technology parameters, \( z \) can also be interpreted as a parameter that governs the rate of return to experience. In the quantitative exercise, I assess the effect of changes in return to experience on female labor supply by experimenting on changes in parameter \( z \).

1.4.4 Budget constraint

The period budget constraint of a woman depends on her type. I first describe her budget constraints before her mandatory retirement \((1 \leq t \leq r - b)\). If she is single, her budget constraint is

\[ C_t + a_{t+1} = \pi h_t (1 - l_t) + Ra_t, \]

where \( a_t \) is her asset at age \( t \) and \( R - 1 \) is the real interest rate. Her wage \( \pi h_t \) equals her human capital \( h_t \) times its constant rental rate \( \pi \). Thus, her labor income is \( \pi h_t (1 - l_t) \). The budget constraint states that her consumption \((C_t)\) and saving \((a_{t+1} - a_t)\) must come from her labor income and the return to her assets \((R - 1)a_t\).

Similarly, the budget constraint for a married woman without kids or with grown-up kids,

\[ C_t + a_{t+1} = \pi h_t (1 - l_t) + y_t + Ra_t, \]

where \( y_t \) is the earning of her husband at her age \( t \). The earnings of the husband over his life-cycle is assumed to be quadratic in his age \( t' \),

\[ y_{t'} = \varphi_0 + \varphi_1 t' + \varphi_2 t'^2, \quad 1 \leq t' \leq r - b + 2. \]

Note that \( y_t = y_{t+2} \) since I assume that a wife is two years younger than her husband. Once a woman gets married, her husband’s earning enters the family budget constraint. However, I assume that the husband brings in zero asset at the time of marriage.
The budget constraint for a woman with small kid(s) between age 0 to 8 is
\[ C_t + pt(1 - l_t) + a_{t+1} = \pi h_t(1 - l_t) + y_t + Ra_t, \]
where \( p_t \) is the childcare cost per unit of time at \( t \). It depends on how many kids a mother has and the childcare price per unit of time per kid. As discussed in section 1.3.1, one of the three reasons that we observe married women with youngest kids at or above the age of elementary school work less on average than married women without kids may be the potential after school care cost for small kids in elementary school. Therefore, I assume that kids before age 8 (grade 3) need childcare service if their mothers work. The price for childcare service per kid per unit of time depends on the kid’s age. Let \( p_k \) denote the price of childcare for a kid of age \( 0 \leq k \leq 8 \). Since kids arrive in a deterministic way after a woman receives a positive fertility shock, \( p_t \) can be easily derived from \( p_k \) and the woman’s marriage age.

The assumption that the childcare cost is divisible might seem rather unrealistic. However, since I model women by types, childcare cost averaged within a type can be treated as approximately divisible. Moreover, the model period is one year, and it is certainly possible for mothers to work and arrange childcare service for their kids for a certain period in the year, which makes childcare cost divisible from that perspective.

The budget constraints for a woman after the mandatory retirement age \((r - b < t \leq T - r)\) is (either single or married),
\[ C_t + a_{t+1} = Ra_t. \]
She (the couple) has no labor income and has to live on her (their) savings.

Last, I assume that all women start with zero assets \((a_1 = 0)\) and cannot be in debt when they die \((a_{T-b+1} \geq 0)\). Moreover, in light of the discussion on the marriage income effect in section 1.3.2, throughout, I assume that all women (or the couples) face borrowing constraint so that \( a_t \geq 0 \) for all \( t \).

### 1.5 Calibration

I calibrate the model to the 1965-69 cohort of women. The main data I use for the calibration is the sample I constructed from the JPSC data. However, because the sample is restricted to women between age 26 and 41, I also use other data sources for some parameters when calibration target is not available in the sample. Since my goal is to explain the aggregate pattern, I avoid targeting the dynamics of the age profiles in the calibration. In the following, I divide the parameters into categories and discuss the calibration in details. Table 2.3 summarizes the calibration results.

**Demographics** Most of the demographic related parameters are already set when describing the model: a woman starts at age \( b = 21 \), retires at age \( r = 60 \) and dies at age \( T = 80 \); she receives her last marriage shock at age \( \bar{m} = 37 \); if she ever marries, her husband is two years older than her and retires and dies at the same time as she does; if she has kids, the first kid is born two years after the marriage and the second kid is born two years after the first one is born. These parameters are set to match the average values in the data. Age parameter \( t \) is normalized so that the initial age is 1. Hence, a woman’s life span in the model is \( 1 \leq t \leq T - b \).

The remaining demographic parameters are the marriage shocks \( p_{m,t} \) and the fertility shock \( p_k \).
Table 1.1: Calibration

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value</th>
<th>Target</th>
</tr>
</thead>
<tbody>
<tr>
<td>marriage shocks</td>
<td>( p_{m,t} )</td>
<td>marriage rate at age ( t )</td>
</tr>
<tr>
<td>fertility shock</td>
<td>( p_k )</td>
<td>0.84 % of married couples that have kids</td>
</tr>
<tr>
<td>interest rate</td>
<td>( R )</td>
<td>avg. real rate in 90s and 00s</td>
</tr>
<tr>
<td>discount factor</td>
<td>( \beta )</td>
<td>0.9671 ( R\beta = 1 )</td>
</tr>
<tr>
<td>period utility</td>
<td>( \gamma_c )</td>
<td>0.72 avg. hrs. (mothers with youngest kid ( \geq 6 ))</td>
</tr>
<tr>
<td></td>
<td>( \gamma_l )</td>
<td>2.26 Frisch elasticity (Kuroda and Yamamoto, 2008)</td>
</tr>
<tr>
<td></td>
<td>( \alpha )</td>
<td>1.25 avg. hrs. (single women)</td>
</tr>
<tr>
<td></td>
<td>( \theta )</td>
<td>0.56 avg. hrs. (married women without kids)</td>
</tr>
<tr>
<td>human capital (h.c.)</td>
<td>( h_1 )</td>
<td>1 normalization</td>
</tr>
<tr>
<td></td>
<td>( z )</td>
<td>0.065 wage growth (single women)</td>
</tr>
<tr>
<td></td>
<td>( g )</td>
<td>0.018 as above</td>
</tr>
<tr>
<td></td>
<td>( \eta )</td>
<td>0.2 set; ref. Wallenius (2011)</td>
</tr>
<tr>
<td></td>
<td>( \delta )</td>
<td>0.03 as above</td>
</tr>
<tr>
<td>rental rate of h.c.</td>
<td>( \pi )</td>
<td>0.9 initial gender wage gap;</td>
</tr>
<tr>
<td>male earning profile</td>
<td>( \psi_0 )</td>
<td>0.4281 linear least square;</td>
</tr>
<tr>
<td></td>
<td>( \psi_1 )</td>
<td>0.0317 as above</td>
</tr>
<tr>
<td></td>
<td>( \psi_2 )</td>
<td>-0.0004 as above</td>
</tr>
<tr>
<td>childcare</td>
<td>( p^0 ) to ( p^2 )</td>
<td>0.32w_1 monthly childcare cost; female starting wage</td>
</tr>
<tr>
<td></td>
<td>( p^3 ) to ( p^5 )</td>
<td>0.25w_1 as above</td>
</tr>
<tr>
<td></td>
<td>( p^6 ) to ( p^8 )</td>
<td>0.1w_1 as above</td>
</tr>
</tbody>
</table>

directly estimate them from the data. The probability of getting married at age \( t \) conditional on being single \( p_{m,t} \) is calibrated to target the marriage rate (percentage of women who are married) at each age.\(^{25}\) Since women enter the economy at age 21 in the model, I assume that marriages before age 21 all occur at age 21. This assumption is inessential to the main results as the fraction of women married before age 21 is less than 7\%. The probability of having kids once married is set to target the percentage of married women who have kids in the data. This implies \( p_k = 0.84 \).

**Utility** Utility related parameters are the discount factor \( \beta \), the curvatures of consumption and leisure \((\gamma_c, \gamma_l)\), the weight attached to the leisure \( \alpha \) and the utility weight of married women \( \theta \). I also categorize the real interest rate \((R - 1)\) and the endowment of total annual hours as utility related parameters.

I set \((R - 1) = 0.034\), targeting the average real interest rate in Japan in the 1990s and 2000s. The discount factor \( \beta \) is set such that \( \beta R = 1 \). I assume that the discretionary time endowment per day is 14 hours, so the annual total hours endowed is \( 14 \times 365 = 5110 \).

With the chosen utility function, the theoretical Frisch elasticity of labor supply \((\epsilon_l)\) in a model of similar setting but no human capital accumulation is given by \( \epsilon_l = \frac{l}{1-\gamma_l} \), where \( l \) is leisure and \( \gamma_l \) is its curvature. My model features human capital accumulation. Nevertheless, I simply calibrate \( \gamma_l \) using the above theoretical Frisch elasticity. I set \( \gamma_l \) to match the Frisch elasticity estimated in the literature and the weekly hours worked per woman in the aggregate data. Kuroda and Yamamoto (2008) provide an estimation of female Frisch elasticity in Japan using data in the 1990s. The estimated elasticity on the extensive and intensive margins combined is between 1.3 and 1.5. I take the average value 1.4 as my target. To estimate the weekly hours worked per woman, I use aggregate data from the OECD and

\(^{25}\text{Although the sample of 1966-69 female cohort only covers an age range from 26 to 41. Marriage age is a retrospective question for married women at the first wave of the survey (1993) so I can calculate marriage rates at ages before 26.}
Statistics Bureau Japan because they cover a wider age range for the 1965-69 female cohort than my sample constructed from the JPSC data. Using data on employment rate and weekly hours worked per worker, I calculate the average weekly hours worked per woman to be 23.7.\textsuperscript{26} With the weekly hours endowment being 14 × 7 = 98, the normalized average weekly leisure is \( l = 1 - 23.7/98 \approx 0.76 \). Thus, \( \gamma_l \) is calibrated to be \( \gamma_l = \frac{l}{1-l} \frac{1}{\gamma} \approx 2.26 \).

The rest of the utility related parameters are calibrated to match the relevant moments in the data. The coefficient on leisure \( \alpha \) is calibrated to match the average annual hours worked by single women. The utility weight of married women \( \theta \) is calibrated to target the average annual hours worked by married women without kids. As discussed in the static setting, the curvature of consumption \( \gamma_c \) has an implication on the difference of annual hours worked between married women without kids and married women with grown-up kids. Correspondingly, \( \gamma_c \) is calibrated to match the average annual hours worked for married women with grown-up kids.\textsuperscript{27}

**Human capital** I normalize the initial female human capital to be 1, i.e., \( h_1 = 1 \). The parameters to be calibrated are human capital depreciation rate \( \delta \), learning productivity shift \( z \), learning ability decay rate \( g \) and production function curvature \( \eta \).

I calibrate these parameters to match wage growth of single women in the data. About 90\% of single women work, so targeting their wage growth minimizes the selection bias.\textsuperscript{28} However, in searching for the parameter values, I find it difficult to identify all the four parameters. Many values of these parameters can provide good fit. To proceed, I simply set \( \delta = 0.03 \) and \( \eta = 0.2 \) and calibrate only \( g \) and \( z \). The values of \( \delta \) and \( \eta \) are set within the ranges of possible estimates. They also lie in the ranges reported by Wallenius (2011), who finds similar identification issues when she calibrates the same set of parameters in her model (which has the same learning-by-doing specification).

**Budget** Budget related parameters are the female human capital rental rate \( \pi \) and the coefficients of the husband’s quadratic earning profile \((\varphi_0, \varphi_1, \varphi_2)\).

I calibrate \( \pi \) by targeting the initial gender wage gap. I assume that all men and women start working at the same age and have the same initial human capital; hence men also start working at age 21 and their initial human capital \( h_{1m} = 1 \). I define the initial gender wage gap as the ratio between female and male wages when they are at the age 21. Denote \( w_1^l \) the male wage at age 21 and \( \pi^m \) the rental rate of male human capital. The initial gender wage gap is given by \( \frac{w_1^l}{w_1^m} = \frac{\pi h_1^m}{\pi^m h_1^m} = \frac{\pi}{\pi^m} \). That is, it is simply the ratio between the rental rates of female and male human capital. I further normalize male initial wage to be 1, so I obtain the male human capital rental rate \( \pi^m = w_1^m / h_1^m = 1 \). I then calculate \( \pi \) by matching the ratio \( \frac{\pi}{\pi^m} \) to the initial gender wage gap in the data. Abe (2011) reports gender wage gap by cohort and age in Japan. For the cohorts born in 1950s, 1960s and 1970s, the gender wage gaps at early 20s are all about 90\%. I take it as my target and obtain \( \pi = 0.9 \).\textsuperscript{29}

\textsuperscript{26}This average is over women between age 21 and 44 using the latest data available for the 1965-69 cohort.

\textsuperscript{27}Clearly, all these parameters are determined jointly.

\textsuperscript{28}The selection here refers to selection to labor market. I assume that selection to marriage is less an issue, i.e., average wage profile of single women is not affected by their choices of being single.

\textsuperscript{29}To be precise, because only husbands enter the model and I assume that selection to marriage is less an issue, i.e., average wage profile of single women is not affected by their choices of being single.
The coefficients of the quadratic earning profile are estimated directly from the data using linear least square method followed by an adjustment so that the fitted value of earnings are consistent with the normalization on male initial human capital and wage. Specifically, the coefficients estimated from least square are multiplied by a factor so that the fitted initial earning \( \pi^m \) satisfies \( \pi^m h^m n^m = \), where \( n^m \) is the husband’s annual hours worked. The adjustment factor is easily obtained. We already have \( \pi^m = 1 \) and \( h^m = 1 \). The annual hours worked \( n^m \) is estimated using aggregate data from Statistics Bureau Japan.\(^{30}\)

**Childcare price** The childcare cost depends on how many kids a mother has and the age of each of her kids. The parameters to be calibrated are the childcare prices \( p^k \) for kids of ages \( 0 \leq k \leq 8 \) (childcare cost per kid of age \( k \) per unit of time). I assume three price levels for kids between age zero and eight: one for kids between zero and two years old (i.e. infants or toddlers, \( p^0 = p^1 = p^2 \)), one for kids between three and five years old (i.e. preschoolers, \( p^3 = p^4 = p^5 \)), and one for kids between six and eight years old (i.e. grade 1 to 3 kids, \( p^6 = p^7 = p^8 \)).

I calibrate the price for infants or toddler using data on monthly childcare cost. The first ten waves of JPSC collected information on childcare arrangement from women who gave birth on the survey year. The average monthly childcare cost reported by women who returned or planned to return to work and were using or planned to use childcare services is about 41.6 thousand yen (about 416 USD).

To stay consistent with previous normalizations, I express the childcare price in term of the female initial wage \( w_1 \). My sample cohort start at age 26, so I turn to data on starting wage reported by Japan’s Ministry of Health, Labor and Welfare (MHLW).\(^{31}\) I use the starting wage reported in 1988, in which year the 1965-69 female cohort are 21 years old on average. The starting wages are reported by educational attainment. The average starting monthly wage is 149 thousand yen for university graduates and 113.8 thousand yen for high school graduates. I take the average of the two to approximate the average starting wage for females.\(^{32}\) This implies \( p^0 = p^1 = p^2 = \frac{41.6}{(149+113.8)/2} w_1 \approx 0.32 w_1 \).\(^{33}\)

Next, I estimate the childcare price for preschoolers to be 20% lower. The estimate is obtained using data on: 1) learning expenditure per child in kindergarten reported by Japan’s Ministry of Education, Culture, Sports, Science and Technology (MEXT); and 2) licensed daycare cost and shares of mothers using kindergarten and daycare services reported by Oishi (2002). This gives \( p^3 = p^4 = p^5 = 0.25 w_1 \).

Since after-school childcare service usually provides care for about three hours, three-eighth of a full-time daycare, I approximate its monthly cost to be about three-eighth of the cost for a preschooler. This gives \( p^6 = p^7 = p^8 = 0.1 w_1 \).

**Computation** The life-cycle problem I set up has two continuous endogenous state variables, asset and human capital stock. Moreover, types of women arise over the life cycle due to marriage and fertility uncertainties. To solve the model efficiently, I extend the endogenous grid method (EGM) developed

---

\(^{30}\)In particular, I use weekly hours worked data for full-time male workers in 1986, in which year the 1963-67 male cohort are on average 21 years old. The data are in 5-year age bands, so I use the data on 20-24 age band to approximate the weekly hours worked at age 21. I assume 48 working weeks a year and obtain the annual hours worked at age 21.

\(^{31}\)The data on starting wage are constructed using the Basic Survey on Wage Structure (BSWS), an annual survey on earnings and hours conducted by MHLW. The micro-data set of this survey is not publicly available.

\(^{32}\)I assume that all new female graduates are employed. Otherwise, to calculate a better estimate of the average starting wage, I would need data on employment rate of new female graduates by educational attainment.

\(^{33}\)The average starting wage for female workers are likely overestimated. For the 1965-69 female cohort, less than 50% of female workers have educational attainment above high school and among them many are college graduates whose average starting wage is lower than that of the university graduates. Since childcare price is calibrated as a fraction of initial wage, it is then likely underestimated so as to be conservation.
by Carroll (2006) to a two-dimensional setting, much in the same spirit of Ludwig and Schöhn (2013). The advantage of the EGM in this case is that in each iteration I only need to solve a one-dimensional nonlinear equation as opposed to a two-dimensional one in the traditional method. The tradeoff is that the method also introduces an interpolation problem on irregular grids. Overall, the method speeds up the computation substantially. I leave the details of the computation to Appendix 1.C.

1.6 Results

In this section, I first assess the performance of the calibrated model by comparing its predicted age profiles of labor supply with the actual ones in the aggregate data and micro-panel data. I then investigate how each of the three determinants affects life-cycle labor supply of different types of women and to what extent it contributes to the aggregate M-shaped pattern.

1.6.1 The baseline economy

Figure 1.6 compares the model prediction with the actual age profiles of labor supply by marital status (as seen in figure 1.2). Overall, the predicted profiles match the ones in the data quite well. The predicted age profile for all women captures the U-shaped segment observed in the data. As shown separately in figure 1.7, it also captures the M-shaped aggregate pattern in the macro-data (the “M” after the first peak). The timings of the two peaks are comparable to the ones seen in the aggregate pattern. The magnitude of the fall and recovery of labor supply is also comparable.

Turning to the age profile for married women in figure 1.6, the predicted profile again does quite well in capturing the main pattern: for observed ages, hours worked by married women is low before their 30s but it rises gradually afterward. In terms of the details, married women work slightly more in the
model than in the data except for the last two observed ages (40 and 41). The hours worked for married women after 40 is trending slightly higher in the data than in the model.

The predicted age profile for single women fits data relatively less well. As discussed in section 1.3.2, in a model without borrowing constraint, the marriage income effect would predict a rising age profile for single women during the periods that they receive marriage shocks. The borrowing constraint flattens out the age profile for single women when it binds. In the calibrated model, the constraint binds for single women before age 31 so they supply more labor than they would be willing to if they were not constrained. As a result, their labor supply before age 31 is relatively flat. After age 31, the constraint no more binds and their labor supply rises until the marriage shocks end at age 37. Although the age profile of single women doesn’t fit the data perfectly, it nevertheless still captures the main feature that single women work much more than married women and they work relatively stable annual hours during their life cycle comparing with married women.

![Figure 1.7: Hours Worked (Model & Data)](image)

Note: The model predicted age profile is converted into a 5-year age group profile so as to be comparable with the ones from the macro-data. In addition, since the predicted age profile is measured in annual hours and the ones in the macro-data are measured in weekly hours, I scale up the age profile of weekly hours for the 1965-69 cohort so that its 20-24 age group works the same annual hours as does the model’s 21-24 age group. I then scale up profiles of other cohorts using the same factor. The scaling removes the discrepancy between the measurement of weekly hours and annual hours, and allows the comparison to be focused on the relative change along the life cycle. I also normalize the total annual hours to be one.

Figure 1.8 compares the model prediction with the data on labor supply of married women without kids, with youngest kid at preschool age and with youngest kid above preschool age. The predicted age profiles capture well the level differences between the three actual profiles but do not do a good job on matching their exact patterns. The two actual profiles for married women with kids rise faster than what’s predicted in the model. The V-shaped segment of the profile for married women without kids is completely missing in the model. As mentioned when presenting the facts on these profiles in section
1.2.3, I do not intend to examine the patterns (dynamics) of these profiles, so no specific mechanism is built into the model to account for them. The predicted profiles are thus not expected to match the exact pattern in the data. What’s important is that the key mechanisms in the model are clearly in action and their predictions are mostly consistent with the level differences observed in the data. Mothers with small kids face tradeoff between working and taking care of kids at home. Since they face lower effective wages, they work less. As their kids grow up, the childcare cost gradually decreases and eventually disappears so they start to work more. However, less hours worked during the child-rearing period leads to less human capital accumulation, so they face relatively low wages after their kids grow up. The low wages, together with the effect of economics of scale within a family, result in their lower labor supply relative to married women without kids. The transition of women from single to married changes the family setting where they make labor supply decisions (as captured in the model by the utility change). The family utility weight, the economies of scale within a family and the marriage income effect (or gender wage gap) result in a pure marriage effect so the model predicts that married women without kids work less than single women.

1.6.2 The three determinants

Given the calibrated model, I assess how each of the three key determinants, childcare cost, initial gender wage gap and return to experience, affects female labor supply. To do so, I take a comparative statics approach: I experiment on how labor supply reacts to small changes in each determinant. The experiment has two purposes. First, it quantifies to what extent each of the determinants contributes to the formation of the aggregate pattern. Second, it serves as a potential guide to implementing polices that promote female labor supply.
1.6.2.1 Childcare cost

Table 1.2 summarizes the impact of reducing childcare cost on female labor supply. The first column of the table lists the statistics of interest. I report percentage changes in aggregate labor supply and labor supply of different types of women relative to the baseline model. To assess how labor supply is affected along the life-cycle, I also report percentage changes in labor supply for four different age segments. The remaining three columns present the results when childcare price $p^k$ is lowered by 1, 5 and 10 percent.

Overall, the effect of decreasing childcare price on aggregate labor supply is small. Its elasticity is about 0.13. Married women, in particular married women with small kids, are affected the most by the change in childcare price. This is as expected since lowering childcare price benefits directly mothers with small kids by increasing their effective wages, which in turn encourages their labor supply. The inter-temporal link implies that changes in childcare price can also affect labor supply of single women and married women without kids. However, under the calibrated model, the effect is negligible. As seen in the table, labor supply of single women and married women without kids hardly change for any of the three reductions in childcare price.

From a life-cycle perspective, aggregate labor supply increases the most when women are in their 30s because most of the child-rearing happens in this age range. Labor supply also increases for women at 40s and 50s, but the increase is quite small. This increase in labor supply in late worklife can be mostly attributed to the increase in labor supply of married women with grown-up kids. Their increase in hours worked during the child-rearing years leads to less human capital loss and hence they face relatively high wages when their kids grow up, encouraging their labor supply.

The changes in labor supply by age segments also reveal that reducing childcare price can modestly alleviate the drop in aggregate labor supply during women’s 30s but can hardly boost its recovery. In other words, reducing childcare price can make the trough of “M” shallower but cannot bring even the two peaks. While the childcare cost is certainly an important factor contributing to the formation of the “valley” part of the M-pattern, the result suggests that it is not the main cause of the lower second peak.

Last, for all the reported statistics, percentage changes in labor supply relative to the baseline are approximately linear with respect to the percentage decrease in childcare cost. This result suggests that policies that promote female labor supply by lowering childcare cost should not expect too much diminishing to return within small change in childcare cost, although the overall effect of the policies
Table 1.3: Decreasing initial gender wage gap

<table>
<thead>
<tr>
<th>Δ% in hours</th>
<th>Δ% initial wage gap</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>-1%</td>
</tr>
<tr>
<td>All women</td>
<td>1.61</td>
</tr>
<tr>
<td>Single</td>
<td>0.16</td>
</tr>
<tr>
<td>Married</td>
<td>2.89</td>
</tr>
<tr>
<td>No Kids</td>
<td>1.69</td>
</tr>
<tr>
<td>Youngest Kid &lt; 6</td>
<td>5.51</td>
</tr>
<tr>
<td>Youngest Kid ≥ 6</td>
<td>3.24</td>
</tr>
<tr>
<td>All women (21 - 29)</td>
<td>0.74</td>
</tr>
<tr>
<td>All women (30 - 39)</td>
<td>1.85</td>
</tr>
<tr>
<td>All women (40 - 49)</td>
<td>1.73</td>
</tr>
<tr>
<td>All women (50 - 59)</td>
<td>2.19</td>
</tr>
</tbody>
</table>

wouldn’t be too big as the elasticity is quite small.

1.6.2.2 Initial gender wage gap

Table 1.3 summarizes the effect of narrowing initial gender wage gap on female labor supply. The columns of the table are arranged in the same way as those of table 1.2. The statistics of interest are also the same. A \( x \% \) decrease in initial gender wage gap is defined as an increase in the initial female to male wage ratio from its baseline value 0.9 to \( 0.9(1 + x/100) \). Thus, a 1%, 5% or 10% decrease in initial wage gap corresponds to an increase in the initial female-male wage ratio from 0.9 to 0.909, 0.945 or 0.99.

Comparing to reducing childcare price, aggregate labor supply is much more responsive to narrowing initial gender wage gap. The elasticity is about -1.61, more than ten times in absolute value than the previous one. Several factors contribute to this large effect. First, reducing initial gender wage gap applies to all women while reducing childcare price mainly applies to women with small kids. As seen in the table, both single women and married women without kids respond positively to a decrease in initial gender wage gap but they react little to a decrease in childcare price in the previous case. Second, reducing initial gender wage gap provides an incentive for all women to work more in their early careers. Since their learning ability is relatively high when they are young, it has a large positive effect on early human capital accumulation and labor supply. Third, the cumulative nature of the learning-by-doing process implies high human capital early on benefits future human capital accumulation, which in turn leads to better wage and more labor supply in women’s middle and late career. Indeed, as seen in table 1.3, among all four age segments, labor supply of women at their 50s are the most responsive. In the previous case, reducing childcare price applies to married women with small kids when most of them are already in their middle career. Therefore, even though the human capital channel is also present, its effect is little because its cumulative nature is not fully utilized.

Similar as in the previous case, among different types of women, labor supply of mothers with small kids respond the most. Moreover, in this case their labor supply increases much more. Two main factors contribute to this larger increase. The first one is the decrease in marriage income effect as women improve their relative wages to their husbands’ due to human capital accumulated before marriage. This can be seen by comparing the increases in labor supply of the married women without kids in the two cases. The second factor is the increase in effective wage, real wage minus the childcare price. This factor
dominates the last one as the difference in labor supply increase between married women without kids and married women with small kids is larger than the labor supply increase of married women without kids.

Note that both reducing initial gender wage gap and lowering childcare price increase effective wage. However, in the first case, it is achieved by an increase in human capital. Therefore, it gives additional incentive to work more because high human capital means more effective learning and high future human capital and wages. Comparing with the previous case, this human capital channel also leads to larger positive effect on labor supply of mothers when their kids grow up. The increase in labor supply of mothers with grown-up kids is about 50% as that of mothers with small kids, while in the previous case it’s less than 20%.

The changes in labor supply by women in different age segments contrast the previous case in that labor supply of women at 50s responds the most, followed by women at 30s and 40s. Labor supply of women at 20s responds the least. The fact that labor supply increases relatively more at women’s 40s than at their 20s implies that reducing initial gender wage gap not only alleviates the fall of labor supply at women’s 30s but also boosts its recovery. In terms of the aggregate pattern, this means reducing initial gender wage gap leads to a shallower trough of the “M” and a relatively even height of the two peaks. This suggests that, relative to childcare cost, the gender wage gap (the initial gap in this case) is more responsible for the lower second peak in the aggregate pattern.

### 1.6.2.3 Return to experience

As discussed in section 1.4.3, I define rate of return to labor market experience as rate of wage growth with respect to hours worked (equation 1.4). Since it depends on age and types of women, to study the effect of its change on labor supply I must take a reference point. I take the return to experience of single women at their initial age as my reference point. I then increase the value of parameter \( z \) such that single women’s return to experience at age 21 increases by 1%, 5% ad 10%.

Table 1.4 summarizes the effect of decreasing return to experience on female labor supply. All results are quite comparable with the case of reducing initial gender wage gap, and hence the comparison with the results from the case of reducing childcare price is also similar. The elasticity of aggregate labor supply with respect to return to experience is 1.47. Again, labor supply of single women and married women with no kids reacts the least, and labor supply of married women with small kids reacts the

<table>
<thead>
<tr>
<th>( \Delta% ) in hours</th>
<th>( \Delta% ) return to experience</th>
<th>1%</th>
<th>5%</th>
<th>10%</th>
</tr>
</thead>
<tbody>
<tr>
<td>All women</td>
<td>1.47</td>
<td>7.05</td>
<td>14.19</td>
<td></td>
</tr>
<tr>
<td>Single women</td>
<td>0.22</td>
<td>0.87</td>
<td>1.91</td>
<td></td>
</tr>
<tr>
<td>Married women</td>
<td>2.49</td>
<td>12.12</td>
<td>24.50</td>
<td></td>
</tr>
<tr>
<td>No Kids</td>
<td>1.53</td>
<td>7.46</td>
<td>15.33</td>
<td></td>
</tr>
<tr>
<td>Youngest Kid &lt; 6</td>
<td>4.61</td>
<td>23.29</td>
<td>47.96</td>
<td></td>
</tr>
<tr>
<td>Youngest Kid ≥ 6</td>
<td>2.97</td>
<td>14.24</td>
<td>28.06</td>
<td></td>
</tr>
<tr>
<td>All women (21 - 29)</td>
<td>0.34</td>
<td>1.96</td>
<td>4.43</td>
<td></td>
</tr>
<tr>
<td>All women (30 - 39)</td>
<td>1.36</td>
<td>6.24</td>
<td>12.89</td>
<td></td>
</tr>
<tr>
<td>All women (40 - 49)</td>
<td>1.66</td>
<td>7.93</td>
<td>15.58</td>
<td></td>
</tr>
<tr>
<td>All women (50 - 59)</td>
<td>2.59</td>
<td>12.33</td>
<td>24.46</td>
<td></td>
</tr>
</tbody>
</table>
most. Labor supply of mothers with grown-up kids also increases by a relatively large amount, about 50% of that of mothers with small kids. From a life-cycle perspective, the largest increase in aggregate labor supply comes from women at their 50s, followed by women at their 40s, 30s, and 20s. Note that their labor supply increases relatively more at 40s than at 20s as in the case of reducing initial gender wage gap. Therefore, the changes along the life cycle again make the trough of the M-shaped aggregate pattern shallower and bring the two peaks to a relatively even height. This result confirms that relative to childcare cost the gender wage gap (a consequence of return to experience in this case) is more associated with the lower second peak observed in the aggregate pattern.

These results are expected since increasing return to experience works through the human capital channel in a similar way as reducing initial gender wage gap. It increases the rate of wage growth for an extra hours worked, so it provides more incentive for human capital accumulation. The cumulative effect of the learning-by-doing process thus kicks in.

1.6.2.4 Summary

In sum, the three experiments show that narrowing gender wage gap (through reducing initial gap or increasing return to experience) has much larger effect on aggregate female labor supply than reducing childcare cost. Analyzing the labor supply of different types of women and women of different age segments highlights the important role of the human capital channel. The experiments also serve to evaluate the relative merit that each determinant contributes to the observed two features in the aggregate life-cycle pattern: the M-shape and the uneven heights of the two peaks. While childcare cost certainly contributes to the formation of the “valley” part of the M-shape, the results suggest it is not the main reason behind the uneven heights of the two peaks. Gender wage gap, on the other hand, is more responsible for the uneven heights. Again, the dynamic impact of the gender wage gap and childcare cost on life-cycle female labor supply is influenced by the human capital accumulation process.

These experiments also provide other insights. It shows that among different types of women, married women with preschool kids always respond the most to all three economic environment changes, and in contrast single women always respond the least. Moreover, both aggregate labor supply and labor supply of any types of women (even married women with preschool kids) are more affected by narrowing gender wage gap than by reducing childcare cost. These insights reinforce the importance of implementing policies to narrow gender wage gap as well as identify the group of women that are most and least likely to be affected by the policies.

1.7 Conclusion

Using the empirical facts gathered from the JPSC micro-data set, I showed that the M-shaped aggregate life-cycle pattern can be understood as a result of demographic composition changes and labor supply behavioral differences across different types of women. To account for the facts, I built a life-cycle model featuring heterogeneous types of women and transitions between the types, human capital accumulation and childcare cost. The model prediction matches the data reasonably well. Counterfactual experiments suggest that narrowing the gender wage gap (through lowering initial gender wage gap or increasing return to experience) would have a larger positive effect on female labor supply than simply lowering the childcare cost, a result mostly explained by the human capital channel.
The model in this paper abstracts away a few factors that may be important to explain the aggregate life-cycle pattern. For example, marriage and fertility are taken as exogenous. The potential interaction between labor market behavior and marriage and fertility decisions is thus not considered. Tax is also not considered. Tax policy in particular the spousal tax deduction in Japan’s tax system may affect labor supply of the secondary earners (wives in most cases) in an important way.\textsuperscript{34} Wage uncertainty is another factor omitted. Attanasio et al. (2005) show that female labor supply can act as insurance against idiosyncratic wage shocks within the family. The reason that I do not consider these many other factors is to keep the model simple so as to focus on the ones that I believe are of first order importance. Extending the current model to incorporate some of the above mentioned factors is left for future research.

\textsuperscript{34} See, for example, Akabayashi (2006), Takahashi et al. (2009b), and Yamada (2011).
Appendix

1.A More Facts

1.A.1 Life-cycle labor supply in contrast

To further highlight the two features of the aggregate life-cycle female labor supply in Japan, I contrast it with two other aggregate life-cycle patterns: one is the life-cycle female labor supply in the US. (figure 1.9), and the other is the life-cycle male labor supply in Japan (figure 1.10).

![Average Weekly Hours Worked (Japan & USA)](image)

**Figure 1.9: Average Weekly Hours Worked (Japan & U.S.)**

*Data Source: OECD Labor Force Statistics & Statistics Bureau Japan*

*Note: The age profiles are constructed using repeated annual aggregate data, so the cohorts are “synthetic”.*

1.A.2 The two margins by marital status

The two margins are also very active for both married and single women. In figure 1.11, I plot by marital status age profiles of employment rate (left panel) and annual hours worked per employed women (right panel). The employment status used to construct the employment rate is derived from individual’s annual hours worked. A woman is classified as employed if she works at least 100 hours in a year.

On the extensive margin, the three age profiles of employment rate have similar dynamics as the corresponding age profiles of annual hours worked in figure 1.2. The age profile for all women also exhibits an U-shape. It’s consistent with the profile of employment rate obtained from the macro-data (the left panel of figure 1.4). The employment rate is at 60% when women are at age 26. It quickly drops to about 48% at age 30, and then gradually climbs back to a little above 70% at age 41. Turning to the employment rate of single and married women, we see much more single women work than married
women. While the employment rate for single women stays around 85% for the entire observed ages (from age 26 to 41), only about 35% of married woman work before their 30s. Although afterward the employment rate of married women gradually rises, reaching to about 65% when they are at 41, it is still much lower than that of the single women.

On the intensive margin, the profile for all women (i.e., the profile of average annual hours worked by employed women) is mostly decreasing during the observed ages. The decrease is largely due to the fall in hours worked by employed married women, especially before their 30s. Overall, employed married women work much less than employed single women.

The evidence on the two margins makes it clear that the large difference between average annual hours worked per married woman and per single woman seen in figure 1.2 is a result of both margins in action: less married women work and working married women work less hours on average.

Figure 1.11 also reveals that while more women work at age 41 relative to age 26, they work less hours on average. The reason is that although many married women re-enter labor market starting around their 30s they work less hours on average.

1.B Data

1.B.1 Number of observations by age

See table 1.5.
1.B.2 Data construction

Cohort To construct the 1965-69 female cohort, I take samples of women born between 1965 and 1969 in the JPSC data and group them by age. The cohort has 750 women in total. They were between 24 to 29 at the first wave of the survey (1993), and between 39 to 44 at the sixteenth wave of the survey (2008). After grouping them by age, the cross-sectional sample size is small at the two ends of the age profiles (age 24 to 25 and age 42 to 43). Hence, I drop the two ends and obtain samples with age profiles between 26 to 41.

Annual hours worked There is no direct question on annual hours worked in the JPSC. I construct annual hours worked of each individual based on the following two questions.

1) How many hours do you work per week in general?
Answer choices: 1) Under 15 hrs.; 2) 15 to 21 hrs.; 3) 22 to 34 hrs.; 4) 35 to 42 hrs.; 5) 43 to 45 hrs.; 6) 46 to 48 hrs.; 7) 49 to 54 hrs.; 8) 55 to 59 hrs.; 9) 60 to 64 hrs. 10) Over 64 hrs.

2) How many days did you work really in the past year?
Answer choices: 1) Under 50 days; 2) 50 to 99 days; 3) 100 to 149 days; 4) 150 to 174 days; 5) 175 to 199 days; 6) 200 to 224 days; 7) 225 to 249 days; 8) 250 to 274 days; 9) 275 to 299 days; 10) 300 days or above.

To construct annual hours worked, I do the following.
Step I: convert the choices of weekly hours band and annual days band into weekly hours and annual days. For those bands with lower and upper bounds, I take the average of the two bounds. For example, choice 2) of question 1 is “15 to 21 hrs”. I convert it into $(15 + 21)/2 = 18$ hours worked per week. For
Table 1.5: Number of Observations by Age

<table>
<thead>
<tr>
<th>Age</th>
<th>Observations</th>
</tr>
</thead>
<tbody>
<tr>
<td>26</td>
<td>426</td>
</tr>
<tr>
<td>27</td>
<td>522</td>
</tr>
<tr>
<td>28</td>
<td>637</td>
</tr>
<tr>
<td>29</td>
<td>602</td>
</tr>
<tr>
<td>30</td>
<td>583</td>
</tr>
<tr>
<td>31</td>
<td>561</td>
</tr>
<tr>
<td>32</td>
<td>535</td>
</tr>
<tr>
<td>33</td>
<td>511</td>
</tr>
<tr>
<td>34</td>
<td>492</td>
</tr>
<tr>
<td>35</td>
<td>477</td>
</tr>
<tr>
<td>36</td>
<td>454</td>
</tr>
<tr>
<td>37</td>
<td>441</td>
</tr>
<tr>
<td>38</td>
<td>423</td>
</tr>
<tr>
<td>39</td>
<td>411</td>
</tr>
<tr>
<td>40</td>
<td>402</td>
</tr>
<tr>
<td>41</td>
<td>393</td>
</tr>
</tbody>
</table>

Choices 1) of questions 1 and 2, both of which have no explicit lower bounds, I assume weekly hours of 10 and annual days of 30 respectively. For choices 10) of questions 1 and 2, both of which have no upper bounds, I assume weekly hours of 70 and annual days of 306 respectively.

Step II: estimate weeks worked per year from days worked per year obtained in Step I. To do so, I need an estimate of days worked per week. I assume that those who reply 1) to 7) in question 2 work five days a week and those who reply 8) to 10) work six days a week. Weeks worked per year is then calculated as days worked per year divided by days worked per week.

Step III: obtain annual hours worked as the product of weeks worked per year and hours worked per week.

Annual hours worked of individuals from wave 2 (1994) to wave 16 (2008) are constructed as above. The construction of annual hours worked for wave 1 (1993) has to be handled differently because a different set of survey questions related to hours worked is asked in that wave. In wave 1, depending on how often a worker is paid, different questions are asked. If a worker is paid monthly or weekly, hours per week and days per month are asked. If a worker is paid daily or by hourly rate, hours per day and days per month is asked. In the first case, I assume 48.5 work weeks per year, so annual hours worked is calculated as hours per week times 48.5. In the second case, I first calculate hours worked per month as hours per day multiplied by days per month. Assuming 11.3 work months a year, I then calculate annual hours worked as hours worked per month times 11.3.

While JPSC do have a survey question on number of days off per week, which I could use to obtain days worked per week, the question is not asked in wave 11, 12 and 13. In order to construct the annual hours worked in a consistent manner, I choose not to use the information from this question.

Considering two week-long national holidays in May and August, an almost week-long national holiday around New Year’s day, and a possible annual week-long paid vacation (other national holidays are spread across the weeks), I estimate that a typical worker in Japan works about 48-49 weeks a year. Given 5 work days a week on average, a typical full-time worker would choose option 7) (225 to 249 days) to question 2. Using the typical worker as a benchmark, I further assume that those who work more than 249 days per year work 6 days a week, and those who work less than 225 days per year work 5 days a week. Note that I assume all workers including part-time workers work at least 5 days a week.

The 48.5 work weeks per year is obtained as 52 weeks minus about 3 weeks of week-long holiday day and a potential week-long annual paid leave. See the previous footnote for details.

48-49 work weeks per year is about 11.3 work month per year.
The annual hours worked can also be constructed in yet an alternative way. JPSC collects time allocation data on the respondent and her husband’s daily living schedule for a typical work day and a typical day off. One can then construct annual hours worked by utilizing these time allocation data and the days worked per year data collected in the job related section of the survey. As a robust check, I also construct annual hours worked in this way. I find that the numbers obtained are usually larger than those using the above method, although the shape of the age profiles of average annual hours worked are very similar. In this paper, I construct annual hours worked using the previous method. Calculating annual hours from reported weekly hours should give a more accurate measure.

1.C Computation

1.C.1 Value Function

The value function for a woman’s maximization problem is given by

$$V_t(a_t, h_t, m_t) = \max_{C_t, l_t, a_{t+1}, h_{t+1}} \{ u(C_t, l_t; s_t) + \beta \mathbb{E}_t V_{t+1}(a_{t+1}, h_{t+1}, m_{t+1}) \}$$

where the period utility function, the budget constraint and law of motion of human capital stock are all as specified in section 1.4. The value function has three state variables, asset $a_t$, human capital stock $h_t$ and marital and fertility status $m_t$. The state $m_t$ takes finite values: if a woman is single, $m_t = 0$; if a woman is married with a negative fertility draw, $m_t = 10$; if a woman is married with a positive fertility draw, $m_t = 100 + x$, where $1 \leq x < \bar{m} - b + 1 = 17$ is her marriage age. Family size $s_t$ and childcare cost (which depends on number of kids and their ages) can be easily derived from the state $m_t$. The state $m_t$ also determines the appropriate budget constraint at a given age $t$.

Substitute $C_t$ using the budget constraint, and let $\lambda_t$ be the Lagrangian multiplier of the law of motion of human capital. The FOCs (assuming interior solution) are given by

$$a_{t+1}: \quad \frac{\partial u(C_t, l_t; s_t)}{\partial C_t} = \beta \mathbb{E}_t \frac{\partial V_{t+1}(a_{t+1}, h_{t+1}, m_{t+1})}{\partial a_{t+1}}$$

$$l_t: \quad \frac{\partial u(C_t, l_t; s_t)}{\partial C_t} \pi h_t = \frac{\partial u(C_t, l_t; s_t)}{\partial l_t} - \lambda_t \frac{\partial H(h_t, l_t)}{\partial l_t}$$

$$h_{t+1}: \quad \beta \mathbb{E}_t \frac{\partial V_{t+1}(a_{t+1}, h_{t+1}, m_{t+1})}{\partial h_{t+1}} = \lambda_t$$

$$\lambda_t: \quad h_{t+1} = (1 - \delta)h_t + H(h_t, l_t)$$

The envelope conditions implies

$$a_t: \quad \frac{\partial V_t(a_t, h_t, m_t)}{\partial a_t} = \frac{\partial u(C_t, l_t; s_t)}{\partial C_t} R$$

$$h_t: \quad \frac{\partial V_t(a_t, h_t, m_t)}{\partial h_t} = \frac{\partial u(C_t, l_t; s_t)}{\partial C_t} \pi (1 - l_t) + \lambda_t \left[ (1 - \delta) + \frac{\partial H(h_t, l_t)}{\partial h_t} \right]$$

1.C.2 Algorithm

Since the problem has finite horizon, the value function is solved by backward induction. Solving the value function for $r < t \leq T$ is standard. Solving it for $t = r$ is only slight different from solving it for
Therefore, in what follows, I only describe the algorithm for \( b \leq t < r \). The value function has two endogenous state variables, and they are both continuous. To solve the model efficiently, I extend the endogenous grid method (EGM) developed by Carroll (2006) to a two dimensional setting, much in the same spirit of Ludwig and Schön (2013). The algorithm only requires a 1-dimensional root finding at each iteration so it speeds up the computation.

Define grids of \((a_t, h_t, m_t)\) as \(G_a \times G_h \times G_{m_t}\). Note that \(G_{m_t}\) is naturally defined as the set of values taken by \(m_t\) and the set depends on age \(t\). Let \(t = r-1\). From the last iteration, we obtain the values of \(\frac{\partial V_{t+1}(a_{t+1}, h_{t+1}, m_{t+1})}{\partial a_{t+1}}\) and \(\frac{\partial V_{t+1}(a_{t+1}, h_{t+1}, m_{t+1})}{\partial h_{t+1}}\) and policy functions of \(C_{t+1}, a_{t+2}, h_{t+1}\) and \(h_{t+2}\) on each of the grid points.

Step I, for each \(m_t\) on the grid of \(G_{m_t}\), do the following. For each \((a_{t+1}, h_{t+1})\) on the grid of \(G_a \times G_h\), using the budget constraint, FOCs and envelop conditions, solve the endogenous \((a_{t+1}^{\text{end}}, h_{t+1}^{\text{end}})\) such that \((a_{t+1}, h_{t+1})\) is the optimal choice. Along solving \((a_{t+1}^{\text{end}}, h_{t+1}^{\text{end}})\), obtain \(C_{t+1}^{\text{end}}\) and \(m_{t+1}^{\text{end}}\) as well. This step involves a 1-dimensional root finding.

Step II, Obtain the policy function \(C_t(a_t, h_t, m_t)\) by interpolating \(C_{t+1}^{\text{end}}(a_{t+1}^{\text{end}}, h_{t+1}^{\text{end}}, m_{t+1})\) on the original grid of \((a_t, h_t, m_t)\). Similarly, obtain \(a_{t+1}(a_t, h_t, m_t)\) by interpolating \(a_{t+1}(a_{t+1}^{\text{end}}, h_{t+1}^{\text{end}}, m_{t+1})\) on the grid of \((a_t, h_t, m_t)\). The grid of \((a_{t+1}^{\text{end}}, h_{t+1}^{\text{end}})\) is irregular, so use a triangulation-based scattered data interpolation (Delaunay triangulation). Solve the policy functions for \(l_t\) and \(h_{t+1}\) using \(C_t, a_{t+1}\), the budget constraint and law of motion of human capital.

Step III, obtain \(\frac{\partial V_t}{\partial a_t}\) and \(\frac{\partial V_t}{\partial h_t}\) using the envelop conditions. Go to Step I.

Note that above I only describe the core of the algorithm. Details such as checking for potential corner solutions are omitted.
Bibliography

URL: http://ideas.repec.org/a/eee/asieco/v21y2010i2p142-155.html

URL: http://www.sciencedirect.com/science/article/pii/S0889158310000316

URL: http://ideas.repec.org/a/kap/reveho/v4y2006i4p349-378.html


URL: http://www.aeaweb.org/articles.php?doi=10.1257/aer.98.4.1517

URL: http://www.jstor.org/stable/10.1086/344125

URL: http://ideas.repec.org/a/eee/ecolet/v91y2006i3p312-320.html

URL: http://dx.doi.org/10.1007/s11150-006-0013-4

URL: http://ideas.repec.org/a/oup/restud/v72y2005i1p109-133.html

URL: http://ideas.repec.org/p/fip/fedmsr/317.html

URL: http://ideas.repec.org/a/bla/ausere/v42y2009i3p276-299.html

URL: http://ideas.repec.org/h/eee/labchp/1-02.html

URL: http://ideas.repec.org/a/eee/jjieco/v22y2008i4p566-585.html


URL: http://ideas.repec.org/p/kls/series/0065.html


URL: http://ideas.repec.org/a/red/issued/05-95.html


URL: http://www.jstor.org/stable/3069654


URL: http://dx.doi.org/10.1111/j.1468-5876.2007.00432.x


URL: http://ideas.repec.org/a/red/issued/10-106.html


URL: http://ideas.repec.org/a/eee/labeco/v18y2011i4p539-546.html
Chapter 2

Idiosyncratic Distortions & Aggregate Losses

2.1 Introduction

Differences in total factor productivity (TFP) is an important factor to explain the large differences in income per capita between rich and poor countries.\(^1\) Seeking what causes the TFP gap, a growing strand of literature argues that resource misallocation across firms may potentially be the answer.\(^2\) The basic idea is that firms with different productivity levels face policy distortions that create wedges between marginal products of factor inputs and their prices. As a result, marginal products are not equalized across firms and hence the misallocation of resources across firms and the aggregate inefficiency.

One of the important contributions in this literature is Restuccia and Rogerson (2008) (RR henceforth). Under a neoclassical growth model that incorporates heterogeneous firms as developed in Hopenhayn (1992) and Hopenhayn and Rogerson (1993), they study how general distortions modeled as firm-specific tax or subsidy wedges can affect aggregate efficiency. RR refer to these distortions as idiosyncratic distortions to emphasize the fact that firms under these distortions effectively face different factor prices and hence are distorted differently from each other. They show that idiosyncratic distortions cause inefficient resource allocations across firms, and consequently can lead to sizable decreases in TFP and output. Further, they show that such effect can occur even when distortions have no impact on aggregate prices and capital accumulation.

RR establish their results mainly through numerical exercises.\(^3\) In this paper, we take the same economic environment as in RR but we characterize analytically the mappings from distortions to potential aggregate losses in TFP, capital stock, output, consumption and number of firms. Our first goal is to provide a transparent view on the links between distortions and aggregate losses. In particular, using these analytical mappings, we are able to open the “black box” of RR’s numerical exercises and examine their precise nature. We therefore shed light on how various assumptions of the model may lead to different implications for aggregate losses.

In their main numerical exercises, RR consider policy distortions in the form of tax or subsidy wedges

---

\(^1\) See Caselli (2005); Hall and Jones (1999) and Klenow and Rodríguez-Clare (1997).

\(^2\) For a review on this literature, see Restuccia and Rogerson (2013) and Hopenhayn (2014).

\(^3\) Indeed, as pointed by Hopenhayn (2014b), most papers in this literature are quantitative in nature. See, for example, Guner et al. (2008), Barrieu et al. (2013) and Alfaro et al. (2009).
on firm-level output. Moreover, the distortions are restricted such that steady-state capital stock in the distorted economy remains the same as the one in the undistorted economy. The point they want to make is that distortions can cause large losses in TFP and output even when aggregate prices and capital accumulation are not affected. The numerical results under these assumptions display three features: 1) taxing more firms leads to larger TFP loss; 2) taxing at a higher rate leads to larger TFP loss; and 3) as opposed to random distortions, correlated distortions in the sense of taxing productive firms and subsidizing less productive ones leads to larger TFP loss. These features may look intuitive as taxing more firms, taxing at a higher rate, or taxing productive firms all seem to worsen the distortions. However, as also shown in RR, when distortions are not restricted to keep capital stock unchanged, these features don’t hold any more. The result raises doubt upon our intuition. Their numerical exercises provide little insight on what delivers those features and why they break down as the model assumption changes. In this paper, we present the exact mappings between distortions and aggregate losses for their numerical exercises. The mappings allow us to explain in a unified way the three features from their results: they all correspond to mean preserving spreads of the distortions under a transformed probability measure. We also show exactly how distortions affect capital accumulation and why the features don’t hold when capital accumulation is allowed to respond to distortions.

Another key assumption maintained by RR is that firms face zero fixed production cost. In the context of the model, this assumption implies that all new entrants choose to operate so there is no endogenous exit of firms. Although they assume that firms exit at a constant rate, the rate applies to all firms independent of their productivity levels and the distortions they face. In other words, their exit margin is exogenous and is more of a technical assumption to support the steady-state equilibrium.

Endogenous exit may be a potentially important selection margin, so a natural question to ask is how introducing this margin may affect the implications of the distortions on aggregate losses. The second goal of this paper is to investigate this question. However, when the fixed production cost is assumed to be positive so as to have an active endogenous exit margin, our analytical mappings become too complicated to provide us useful insights. We therefore resort to numerical method. After calibrating the model with positive fixed production cost to the US economy, we experiment on varying the properties of distortions to see how the aggregate losses are affected with and without the endogenous exit margin.

Under our calibration, we find that aggregate losses can respond to the endogenous exit margin very differently. We show that for given distribution of firm-level productivity and distortions, TFP and output losses can be insensitive to the endogenous exit margin while welfare loss can be very sensitive to it. For example, when moving from a zero fixed production cost to a positive one, our results show that in some cases TFP loss increases by only 1%, but welfare loss can decrease more than 60%. In other words, while the endogenous exit margin may have little impact on TFP loss, it can substantially mitigate the welfare loss. Quite related, we also find that in general TFP loss is a poor indicator to how large the welfare loss is. Another finding of our experiments is that the loss in the number of operating firms can respond to the changes in correlation between distortions and firm-level productivity differently with respect to the endogenous exit margin. As the correlation increases (large negative correlation moving to large positive correlation), the loss in the number of operating firms decreases when there is no endogenous exit, but increases when there is endogenous exit.

Our results suggest that empirical studies on resource misallocation should evaluate carefully the presence of fixed production cost as it implies an endogenous exit margin that can potentially affect quantitative results, especially on welfare loss. Similarly, capital accumulation shouldn’t be ignored.
either as it also has important implications on aggregate losses. Further, while TFP loss is an important consequence of misallocation, losses in other aggregate measures, in particular, the welfare loss, should also be studied since TFP loss is usually a poor measure for other aggregate losses.

**Related literature**  This paper is closely related to Hopenhayn (2014b), who provides an elegant theoretical analysis on the mapping between distortions and aggregate productivity. In a static model of heterogeneous firms with a single input factor (labor), he defines a generalized distortion as the firm specific employment ratio between the distorted and the undistorted ones. He shows that the TFP loss can be characterized as the integral of a strictly concave function with respect to an employment-weighted measure of the distortions. This result can be generalized to models with more than one non-cumulative input. Similar to this paper, his characterization explains all three features of RR’s numerical results as mean preserving spreads of the new measure of distortions. Although essentially equivalent, we explain the three features under our characterization of the TFP loss, which we derive and interpret in a different way. Our work therefore complements Hopenhayn (2014b) in that we explain the features of RR’s result from a new angle. Further, more importantly, we also study a setting of the model with capital accumulation and endogenous selection of firms, both of which are not considered in Hopenhayn (2014b). By allowing for capital accumulation, we are able to characterize the mappings in RR’s exercises and explain exactly why the much simpler setting of Hopenhayn (2014b) can explain the features of RR’s results. In addition, we also can explain why those features break down when capital accumulation responses to distortions. By allowing for endogenous selection, in particular, the endogenous exit margin, we are able to show much richer implications of distortions on aggregate losses.

This paper also relates to a set of studies in the misallocation literature that emphasize the role of endogenous entry and exit margins. In a model of heterogeneous firms together with firm dynamics and endogenous entry and exit, Fattal Jaef and Hopenhayn (2012) examine whether, for given distortions, the welfare loss due to resource misallocation is a result of the inefficiency response of the economy. First, they compare a competitive equilibrium under distortions with an equilibrium of a social planner who is also constrained by the same distortions. They show that the competitive equilibrium is inefficient at the entry and exit margins. Next, they quantify the welfare gains by moving from the competitive equilibrium to the planner’s equilibrium. Our work differs from theirs in that we study how properties of distortions affect aggregate losses with and without the endogenous exit margin between the distorted and undistorted economies, not the efficiency loss due to the selection between a competitive equilibrium and a planner’s.

Fattal Jaef (2014) studies how entry and exit margin affect aggregate losses in TFP and welfare due to misallocation. He builds a monopolistic competition setting into the firm dynamic framework of Hopenhayn (1992), and shows that removing distortions leads to a small TFP gain due to an offsetting effect in the selection margin, but a large welfare gain when considering the transition dynamics. Our study puts more emphasis on the endogenous exit margin and we compare different implications of distortions on aggregate losses with and without the margin. In addition, our model is different from his in that we consider a homogeneous good environment so our selection margin do not connect to variety of goods as in a monopolistic competition setting. We also do not incorporate firm-level productivity shocks. We only study steady-state equilibrium and do not consider transition dynamics. Moreover, our model incorporates an accumulative input factor, capital, which is not studied in Fattal Jaef (2014).

In a monopolistic competition environment similar to Fattal Jaef (2014), Yang (2014) structurally
estimates the TFP loss due to misallocation using plant-level data from Indonesia. He shows that
considering the extensive margin empirically magnifies the welfare loss, and it also increases TFP loss by
40%. Both his model and ours have endogenous entry and exit, however, he emphasizes the effect of both
margins while we focus only on the effect of the endogenous exit margin. In our paper, we show that
introducing the endogenous exit margin can have small negative effect on TFP loss, but it can mitigate
the welfare loss substantially. Again, our model with the homogeneous good environment has no variety
effect as in his monopolistic competition setting. In addition, we consider capital accumulation but he
takes capital as a fixed factor.

The rest of the paper is organized as follows. In the next section, we describe the economic model. In
section 3, we analytically characterize the equilibrium and obtain mappings between distortions and
aggregate losses. We present our results in section 4. First, using the analytical mappings, we provide
a unified view on results from two sets of RR’s numerical experiments. Next, we extend the model
to include features beyond their numerical exercises, re-calibrate it to the US data, and present our
numerical results. We conclude in section 5.

2.2 The model

Our economic environment is identical to that of Restuccia and Rogerson (2008), up to some minor
change of variables for ease of analytical expressions. In what follows, we describe the model and define
the steady-state equilibrium. Throughout the paper, we only focus on steady-state analysis.

2.2.1 Technology

**Incumbent firms** Incumbent firms produce a homogeneous good. The production technology is
decreasing returns to scale (DRS) and it’s given by

\[ f(z, k, l) = z^{1-\gamma} (k^\alpha l^{1-\alpha})^{\gamma}, \quad \gamma, \alpha \in (0, 1), \]

where \( z \) is firm-level productivity and \( k \) and \( l \) are capital and labor input. The curvature parameter \( \gamma \)
governs the degree of DRS and is also known as the span-of-control parameter (Lucas, 1978). Capital
and labor shares are given by parameters \( \alpha \gamma \) and \( (1-\alpha)\gamma \).

Firms are heterogeneous in their productivity levels. As in RR, we abstract from the stochastic nature
of the firm-level productivity and assume that \( z \) is constant over time. Firms also face idiosyncratic
distortions \( t \), which act as tax or subsidy wedges on output. A firm in this economy is thus characterized
by its productivity and the distortion it faces, \((z, t)\).

Time is discrete. In each period, an operating firm must pay a fixed production cost \( c_f \) measured in
units of output. The profit maximization problem for a firm is static in each period and it’s given by

\[ \pi(z, t) = \max_{k, l \geq 0} \left\{ (tz)^{1-\gamma} (k^\alpha l^{1-\alpha})^{\gamma} - Rk - wl - c_f \right\}, \]

where \( R \) and \( w \) are capital rental rate and wage. Note that the distortion \( t \geq 0 \) is modeled as a
transformation of the output tax or subsidy wedge.\(^4\) The corresponding tax or subsidy wedge \( \tau \) is

\(^4\)This specification allows me to obtain simpler expressions in the coming analysis.
simply $1 - t^{1-\gamma}$. A distortion $0 < t < 1$ corresponds to a tax rate $0 < \tau < 1$, and a distortion $t > 1$ corresponds to a subsidy rate $\tau < 0$.

Conditional on operating, optimal capital and labor inputs are easily solved as

$$\bar{k}(z, t) = \theta_k R^{-\frac{1-(1-\alpha)}{\alpha\gamma}} w^{-\frac{(1-\alpha)\gamma}{1-\gamma}} zt,$$  \hspace{1cm} (2.1)

$$\bar{l}(z, t) = \theta_l R^{-\frac{\alpha\gamma}{1-\gamma}} w^{-\frac{1-\alpha\gamma}{1-\gamma}} zt,$$  \hspace{1cm} (2.2)

where $\theta_k = \gamma^{\frac{1}{1-\gamma}} \alpha^{-\frac{1-(1-\alpha)\gamma}{\alpha\gamma}} (1-\alpha)^{\frac{(1-\alpha)\gamma}{1-\gamma}}$ and $\theta_l = \gamma^{\frac{1}{1-\gamma}} \alpha^{-\frac{\alpha\gamma}{1-\gamma}} (1-\alpha)^{\frac{1-\alpha\gamma}{1-\gamma}}$. Substituting the optimal factor inputs into the objective function of the maximization problem, we obtain the profit function

$$\pi(z, t) = (1-\gamma)\theta_\pi R^{-\frac{\alpha\gamma}{1-\gamma}} w^{-\frac{(1-\alpha)\gamma}{1-\gamma}} zt - c_f,$$  \hspace{1cm} (2.3)

where $\theta_\pi = \gamma^{\frac{1}{1-\gamma}} \alpha^{-\frac{\alpha\gamma}{1-\gamma}} (1-\alpha)^{\frac{(1-\alpha)\gamma}{1-\gamma}}$.

Incumbent firms may exit exogenously. Specifically, in each period after the production takes place, all incumbent firms face an exogenous probability of death $\lambda \in (0, 1)$.

**Entering firms** To create a new firm, an entrant must first incur an entry cost $c_e$ measured in terms of output. Next, it draws a firm-level productivity and a distortion from a distribution $G(t, z)$ with $z \in [\bar{z}, \bar{z}]$, $\bar{z} \geq 0$, and $t \geq 0$. Denote $g(z, t)$ the pdf of the distribution $G(t, z)$. Note that under this specification, distortions can be potentially correlated with firm-level productivity. For later use, let $G_z(z)$ and $G_t(t)$ be the marginal distributions of productivity and distortions. Let $g_z(z)$ and $g_t(t)$ be their corresponding pdfs.

As is evident from the profit function (2.3), due to the fixed production cost, a firm that cannot break even will choose to shut down the operation. Moreover, since the firm-level productivity and distortion are constant over time, a firm that chooses to shut down will simply exit and this decision is made at entering stage right after it draws its productivity and distortion. Let $x \in \{0, 1\}$ denote a firm’s decision on whether to operate or exit, where $x = 1$ means that the firm operates and $x = 0$ means that the firm exits. It is clear that the optimal decision follows a cutoff rule on productivity level $z$ and the cutoff depends on the distortion $t$. Denote $\hat{z}(t)$ the productivity cutoff function for firms facing distortion $t$. The cutoff function $\hat{z}(t)$ satisfies

$$\pi(\hat{z}(t), t) = 0.$$  \hspace{1cm} (2.4)

Thus, the optimal operation decision is given as

$$\bar{x}(z, t) = \begin{cases} 1 & \text{if } z \geq \hat{z}(t) \\ 0 & \text{otherwise} \end{cases}.$$  \hspace{1cm} (2.5)

The value of an potential entrant is the present value of its expected profit minus the entry cost

$$W_e = \int_{(z, t)} \bar{x}(z, t) \frac{\pi(z, t)}{1-\rho} dG(z, t) - c_e,$$  \hspace{1cm} (2.6)

where $\rho = \frac{1-\lambda}{1+\nu}$ is the firm’s effective discount factor incorporating both the interest rate $r$ and the exogenous exit probability $\lambda$. Free entry implies $W_e = 0.$
**Invariant distribution of firms** Let $\mu(z, t)$ denote the distribution of the operating firms over firm-level productivity and distortions in this period. Let $\mu'(z, t)$ denote the corresponding distribution in the next period. The evolution of the distribution is

$$\mu'(z, t) = (1 - \lambda)\mu(z, t) + \bar{x}(z, t)g(z, t)E,$$

where $E$ is the mass of entrants. Note that under this specification $\mu(z, t)$ in general does not integrate to one. In the steady state, $\mu(z, t) = \mu'(z, t)$. Hence, the invariant distribution is given by

$$\mu(z, t) = E\frac{1}{\lambda}\bar{x}(z, t)g(z, t). \tag{2.7}$$

### 2.2.2 Representative household

There is an infinitely-lived representative household whose life-time utility is

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

where $C_t$ is consumption at period $t$ and $\beta \in (0, 1)$ is the discount factor. The period utility function $u(\cdot)$ is assumed to be differentiable, strictly increasing and strictly concave. The household is endowed with $L$ units of labor in each period and an initial capital $K_0$. Since the household does not value leisure, labor supply is inelastic. The household maximizes lifetime utility by choosing consumptions and savings subject to a period-by-period budget constraint given by

$$C_t + K_{t+1} - (1 - \delta)K_t \leq w_tL_t + R_tK_t + \Pi_t - T_t \quad \forall t \geq 0,$$

where $\delta$ is the capital depreciation rate, $K_t$ and $L_t = L$ are aggregate capital and labor, $\Pi_t$ is the total profit from the operating firms, and $T_t$ is the lump sum transfer from the government. $T_t > 0$ is therefore a lump sum tax and $T_t < 0$ a lump sum subsidy.

The first order conditions of the household problem yields the standard Euler equation. Further imposing the steady-state conditions, we obtain a constant capital rental rate

$$R = \frac{1}{\beta} - (1 - \delta).$$

The real interest rate is thus given by $r = R - \delta$.

### 2.2.3 Equilibrium

A **steady-state competitive equilibrium** consists of prices $(R, w)$, a lump sum transfer $T$, a mass of entry $E$, an invariant distribution of operating firms $\mu(z, t)$, value functions $\pi(z, t)$ and $W_e$, policy functions $\bar{k}(z, t)$, $\bar{l}(z, t)$ and $\bar{x}(z, t)$, and aggregate consumption $C$ and capital $K$ such that

1. Firm optimization: Given prices $(R, w)$, the value functions $\pi(z, t)$ and $W_e$ solve incumbent and entering firm’s problem and $\bar{k}(z, t)$, $\bar{l}(z, t)$, $\bar{x}(z, t)$ are policy functions.

2. Free entry: $W_e = 0$.

3. Consumer optimization: $R = \frac{1}{\beta} - (1 - \delta)$. 
4. Market clearing:

(a) capital
\[ \int \bar{k}(z,t)d\mu(z,t) = K \]

(b) labor
\[ \int \bar{l}(z,t)d\mu(z,t) = L \]

(c) goods
\[ C + \delta K + Ec_e = \int [f(z,\bar{k}(z,t),\bar{l}(z,t)) - c_f]d\mu(z,t) \]

5. Government budget balance:
\[ T + \int (1 - t^{1-\gamma})f(z,\bar{k}(z,t),\bar{l}(z,t))d\mu(z,t) = 0 \]

6. Invariant distribution of operating rms
\[ \mu(z,t) = E_0^{1/\lambda} x(z,t) g(z,t). \]

### 2.3 Characterization of the equilibrium

In this section, we first characterize the equilibrium in the undistorted economy and then the one in the distorted economy. In each economy, we derive the aggregate production function, and obtain an expression for the TFP. We also solve for other interested aggregate variables: capital stock, output, consumption, and number of entrants and operating firms. At the end of this section, we obtain mappings from distortions to aggregate losses. The aggregate losses are defined as ratios of corresponding aggregate variables in the two economies.

#### 2.3.1 Undistorted economy

An undistorted economy is simply defined by a degenerated marginal distribution of distortions:
\[ g_t(t = 1) = 1. \]

To characterize the equilibrium, we start by pinning down the cutoff productivity. Let \( \hat{z} \) be the cutoff productivity defined by the equation (2.4). Using the profit function (2.3), we obtain
\[ (1 - \gamma)\theta_x R^{-\frac{\alpha}{1-\gamma}} w^{\frac{(1-\alpha)\gamma}{1-\gamma}} \hat{z} - c_f = 0 \] (2.8)

In the case that the fixed production cost is zero \( (c_f = 0) \), it is trivial that the cutoff \( \hat{z} = 0 \). Hence, all entrants will operate regardless the productivity levels they draw.\(^5\) The extensive margin thus comes from the free entry and the exogenous exit governed by the firm death rate \( \lambda \).

In the case that the fixed production cost is positive \( (c_f > 0) \), the cutoff productivity \( \hat{z} > 0 \) so endogenous exit may occur depending on whether the cutoff is binding for some entrants. For example,

\(^5\)If \( \hat{z} = 0 \), we assume that \( G_z(\hat{z}) = 0 \) so that the mass of firms with \( z = 0 \) has zero measure.
if $c_f > 0$ and the lower bound of the support of the productivity distribution $\bar{z} = 0$, the cutoff is binding for some entrants who draw productivity levels $\bar{z} \leq z < \hat{z}$, and hence they choose to exit without producing any output. However, if $c_f > 0$ and $\hat{z} > 0$, the cutoff may not be binding for any firms, i.e., depending on parameter values we may have $\hat{z} < \bar{z}$. In this case, there is no endogenous exit even though firms face a positive fixed production cost.

To ease some of the coming expressions, we define

\[
M = \int_{z \geq \hat{z}} dG_z(z), \quad \hat{M} = \int_{z \geq \hat{z}} zdG_z(z),
\]

where $M$ is the fraction of entrants that choose to operate, and $\frac{\hat{M}}{M} = \mathbb{E}[z \geq \hat{z}]$ is the average productivity of those operating entrants. Since incumbent firms exit at the same exogenous rate $\lambda$, the distribution of steady-state operating firms is the same as that of the operating entrants, a left truncated distribution derived from $G_z(z)$ and the cutoff $\bar{z}$. Therefore, $\mathbb{E}[z \geq \hat{z}]$ is also the average productivity of operating firms at the steady state.

Plugging the profit function (2.3) and the optimal operational decision (2.5) into the value of a potential entrant (2.6), the free entry condition $W_e = 0$ then implies

\[
(1 - \gamma)eR^{-\alpha\gamma / \alpha - \gamma} w^{-(1-n)\gamma / \alpha - \gamma} \hat{M} - c_f M = (1 - \rho)c_e. \quad (2.9)
\]

Substituting out the expression $(1 - \gamma)eR^{-\alpha\gamma / \alpha - \gamma} w^{-(1-n)\gamma / \alpha - \gamma}$ using equation (2.8), we obtain an equation that implicitly pins down the cutoff $\hat{z}$ when $c_f > 0$,

\[
M \left( \frac{\mathbb{E}[z \geq \hat{z}]}{\hat{z}} - 1 \right) = \frac{(1 - \rho)c_e}{c_f}, \quad (2.10)
\]

where the effective discount rate $\rho = \frac{1 - \lambda}{1 + \tau} = \beta(1 - \lambda)$ since $R = \frac{1}{\beta} - (1 - \delta)$ from household optimization and the real interest rate $r = R - \delta$.

To see how $\hat{z}$ is uniquely pinned down, it is useful to rewrite the equation (2.10) as

\[
\int_{z \geq \hat{z}} (\frac{z}{\hat{z}} - 1)dG_z(z) = \frac{(1 - \rho)c_e}{c_f}. \quad (2.11)
\]

Define the left-hand-side of equation (2.11) as a function of the cutoff productivity, i.e., define

\[
H(x) = \int_{z \geq x} (\frac{z}{x} - 1)dG_z(z),
\]

where $x \in (0, \bar{z}]$. $H(x)$ is strictly decreasing. In addition, $\lim_{x \to 0} H(x) = \infty$ and $H(\bar{z}) = 0$. The right-hand-side of equation (2.11) is a constant greater than zero. Hence, the cutoff $\hat{z}$ exists and it is unique. See figure (2.1) for an illustration.

Note that if $H(\bar{z}) = \frac{\bar{z}}{\hat{z}} - 1 \geq \frac{(1 - \rho)c_e}{c_f}$, we have $\hat{z} \geq \bar{z}$. Thus, the cutoff is binding for some entrants and we have endogenous exit in the equilibrium. Otherwise $\hat{z} < \bar{z}$ so the cutoff is not binding for any entrants and there is no endogenous exit in the equilibrium.

The cutoff depends on the ratio of entry cost and fixed production cost, $\frac{c_f}{c_e}$. An increase in the ratio leads to a decrease in the cutoff $\hat{z}$. Consequently, it also lowers the average productivity of operating firms faced by the distribution is simply $G_z(z)$ itself. 

\footnote{In case of $\hat{z} < \bar{z}$, the distribution is simply $G_z(z)$ itself.}
firms in the steady state provided that the cutoff is binding for some entrants initially.

Once we characterize the cutoff $\hat{z}$, equation (2.9) solves $w$. Aggregating labor demand of each firm (2.2) over the invariant distribution (2.7), we obtain total demand for labor. Labor market clearing then implies

$$\frac{E}{\lambda} \theta_t R^{-\frac{\alpha \gamma}{1-\gamma}} w^{-\frac{1-\alpha \gamma}{1-\gamma}} \hat{M} = L.$$ 

This equation solves the mass of entry $(E)$. Similarly, we obtain total demand for capital by aggregating firm-level capital demand (2.1). Capital market clearing then implies

$$\frac{E}{\lambda} \theta_k R^{-\frac{1-\gamma}{1-\alpha}} w^{-\frac{1-\alpha \gamma}{1-\gamma}} \hat{M} = K,$$

which solves the steady-state capital stock $K$. Other equilibrium objects can then be solved easily. We omit the details.

**Aggregation** The aggregate output $Y$ can be obtained as

$$Y = \int y(z) d\mu(z),$$

where $y(z) = f(z, \bar{k}(z), \bar{l}(z))$ is the output at firm level. With some algebra (see appendix 2.A.1), it can be shown that

$$Y = Z K^\alpha L^{1-\alpha},$$

where $Z$ is the TFP, and $K$ and $L$ are aggregate capital and labor. Thus, the aggregate production exhibits constant returns to scale (CRS) in capital and labor. For notational convenience, define

$$A = \frac{(1-\gamma)E[z|z \geq \hat{z}]}{(1 - \beta(1 - \lambda))c_e/M + c_f}.$$ 

The aggregate TFP is given by

$$Z = A^{\frac{1-\alpha}{1-\gamma}}, \quad (2.12)$$

and the aggregate capital

\[ \text{Figure 2.1: Pinning down the cutoff productivity when } c_f > 0 \]
\[ K = (\gamma \alpha)^{\frac{1}{1-\alpha}} A^{\frac{1-\gamma}{1-\alpha}} R^{\frac{1}{1-\alpha}} L. \]  

(2.13)

In this undistorted economy, the capital rental rate and wage are corresponding aggregate marginal products adjusted by the curvature parameter \( \gamma \),

\[ R = \gamma \alpha ZK^{\alpha-1}L^{1-\alpha} \]

\[ w = \gamma (1-\alpha)ZK^{\alpha}L^{-\alpha}. \]

Thus, the capital and labor shares are \( \gamma \alpha \) and \( \gamma (1-\alpha) \). The remaining \( 1-\gamma \) fraction of the output goes to the total profits.

The expression of the TFP (equation 2.12) allows us to interpret it in an intuitive way: it is just the average productivity of operating firms \( (E[z|z \geq \hat{z}]) \) adjusted by the entry and fixed cost \( ((1-\beta)(1-\lambda))c_e/M + c_f) \).

Alternatively, the TFP can also be written as

\[ Z = \left( \frac{EM}{\lambda} E[z|z \geq \hat{z}] K^{-\alpha}L^{-(1-\alpha)} \right)^{1-\gamma}, \]

where \( E \) is the mass of entrants given by

\[ E = (\gamma \alpha)^{\frac{1}{1-\alpha}} M^{\frac{\gamma}{1-\alpha}} R^{\frac{1}{1-\alpha}} A^{\frac{1-\gamma}{1-\alpha}}. \]  

(2.14)

This expression of the TFP allows us to view it as a composition of three elements: (1) the total mass of operating firms, \( \frac{EM}{\lambda} \); (2) the average productivity of those firms, \( E[z|z \geq \hat{z}] \); and (3) the steady-state capital, represented by the term \( K^{-\alpha}L^{-(1-\alpha)} \).

In the case that \( c_f > 0 \), equation (2.10) implies \( (1-\gamma)\hat{z}/c_f = A \). Thus, the aggregate TFP can also be written as

\[ Z = \left[ (1-\gamma)\hat{z}/c_f \right]^{\frac{1-\gamma}{1-\alpha}}. \]

In this case, the cutoff productivity and the fixed production cost are sufficient statistics to characterize the aggregate TFP.

The effects of entry cost and fixed production cost on \( Z, K, E \) and number of operating firms \( N = EM/\lambda \) are clear from inspecting equations (2.12), (2.13) and (2.14). If \( c_f = 0 \), the cutoff \( \hat{z} = 0 \). An increase in \( c_e \) leads to decreases in \( Z, K, E \) and \( N \). If \( c_f > 0 \), as already discussed, an increase in \( c_e/c_f \) leads to a decrease in \( \hat{z} \). Hence, an increase in \( c_e \) leads to a decrease in \( \hat{z} \), and if \( \hat{z} \) is binding for some firms initially, it consequently leads to decreases in \( Z, K \) and \( E \) but the effect on \( N \) is ambiguous. The comparative statics on \( c_f \) is ambiguous.

\(^7\)If we take firms as a factor of production, the aggregate output can be written as

\[ Y = E[z|z \geq \hat{z}] \left( \frac{EM}{\lambda} \right)^{1-\gamma} (K^{\alpha}L^{1-\alpha})^\gamma. \]

It is thus constant returns to scale in the number of firms and the compound input \( K^{\alpha}L^{1-\alpha} \). The TFP is given by \( E[z|z \geq \hat{z}] \). In this paper, we do not take this approach.
2.3.2 Distorted economy

The equilibrium in a distorted economy is characterized in a similar way. In what follows, we lay out the main steps. Again, we first characterize the cutoff function $\hat{z}(t)$. Analog to the undistorted economy, define

$$M_d = \int_t \int_{z \geq \hat{z}(t)} dG(z,t), \quad \dot{M}_d = \int_t \int_{z \geq \hat{z}(t)} tzdG(z,t),$$

where $M_d$ is the fraction of entrants that choose to operate. Define $E[z|z \geq \hat{z}(t)] = \frac{\dot{M}_d}{M_d}$. It is the average effective productivity (productivity adjusted by the distortion) of those entrants who choose to operate. It is also the average effective productivity for operating firms in the steady state for the same reason discussed in the undistorted economy.

If $c_f = 0$, we have $\hat{z}(t) = 0$ for all $t$. If $c_f > 0$, the cutoff $\hat{z}(\tilde{t})$ for a distortion $\tilde{t}$ are characterized by

$$M_d \left( E[z|z \geq \hat{z}(t)] - 1 \right) \frac{(1 - \rho)c_e}{c_f} = \frac{1}{\hat{z}(\tilde{t})}, \quad \forall \tilde{t}.$$

Just as in the undistorted economy, $\hat{z}(\tilde{t})$ can be uniquely solved. Also note that $\hat{z}(t)t$ is constant for all $t$.

Given the cutoff function $\hat{z}(t)$, equation (2.9) solves for the wage $w_d$. Labor market clearing condition gives mass of entry $E_d$ and capital market clearing gives aggregate capital $K_d$. Other equilibrium objects can then be easily solved.

**Aggregation**  The aggregate output is given by

$$Y_d = \int g(z,t)d\mu(z,t),$$

where $g(z,t) = f(z,k(z,t),\bar{l}(z,t))$. The aggregate production in a distorted economy again exhibits CRS,

$$Y_d = Z_d K_d^{1-\alpha} L^{1-\alpha}.$$

For notational convenience, define

$$A_d = \frac{(1 - \gamma)E[z|z \geq \hat{z}(t)]}{(1 - \beta(1 - \lambda))c_e/M_d + c_f}, \quad D = \frac{E[z^{1-\gamma}|z \geq \hat{z}(t)]}{E[z|z \geq \hat{z}(t)]}.$$

The aggregate TFP is given by

$$Z_d = A_d^{1-\gamma} D,$$

and the steady-state aggregate capital stock is

$$K_d = (\gamma\alpha)^{\frac{1}{1-\alpha}} A_d^{\frac{1-\gamma}{1-\alpha}} R^{-\frac{1}{1-\alpha}} L.$$

---

8 I use a subscript $d$ to denote corresponding variables in the distorted economy that may take different values from the undistorted ones.
The analogy between the expressions of TFP and capital stock in the distorted economy and those in the undistorted economy is obvious. Indeed, if we view \( zt \) as the effective productivity of firms, the equilibrium in the distorted economy is equivalent to the undistorted one in terms of locations.\(^9\) The term in the expression of \( Z_d \) that has no counterpart in the undistorted economy is \( D \), which I refer to as the distortion factor.

The capital rental rate and wage are aggregate marginal products adjusted by \( \gamma D^{-1} \),

\[
R = \gamma \alpha D^{-1} Z_d K_d^{\alpha-1} L^{1-\alpha}
\]

\[
w = \gamma (1-\alpha) D^{-1} Z_d K_d^{\alpha} L^{-\alpha}.
\]

Therefore, the government’s tax or subsidy wedge revenue accounts fraction \((1 - D^{-1})\) of the total output.\(^10\) The rest of the output is divided into total profit, capital income and labor income in fractions \((1-\gamma)\), \( \gamma \alpha \) and \( \gamma(1-\alpha) \). The term \( D \) would drop out of the above expressions if the distortions are such that total tax revenue from firms equals total subsidy to firms, i.e. \( D = 1 \) so that the lump-sum tax/subsidy to the representative consumer equals zero.

The expression of TFP (equation 2.15) can be interpreted in an intuitive way as well. It is the average effective productivity of operating firms adjusted by the entry and fixed cost as represented by the term \( A_d \), multiplied by the distortion factor \( D \).

Similar as in the undistorted case, \( Z_d \) can also be expressed as

\[
Z_d = \left[ \frac{E_d M_d}{\lambda} \mathbb{E}[zt|z \geq \hat{z}(t)] K_d^{-\alpha} L^{-(1-\alpha)} \right]^{1-\gamma} D,
\]

where \( E_d \) is the mass of entry in the distorted economy given by

\[
E_d = (\gamma \alpha) \frac{\lambda L}{M_d} R^{-\frac{\alpha}{\gamma}} A_d^{\frac{1-\alpha}{\gamma}}.
\]

From this expression, the TFP can be viewed as consisting of four components: (1) the mass of the operating firms \( \frac{E_d M_d}{\lambda} \); (2) the average effective productivity of the operating firms \( \mathbb{E}[zt|z \geq \hat{z}(t)] \); (3) capital stock, represented by the term \( K_d^{-\alpha} L^{-(1-\alpha)} \); and (4) the distortion factor, represented by the term \( D \). The first three are counterparts to those in the undistorted economy, and the forth one is new.

Also similar to the undistorted case, if \( c_f > 0 \), then \( \hat{z}(t) > 0 \) for all \( t \) and \( Z_d \) can be expressed as

\[
Z_d = \left[(1-\gamma)\hat{z}(t)\lambda/c_f \right]^{\frac{1-\alpha}{\gamma}} D.
\]

The comparative statics of \( Z_d \), \( K_d \), \( E_d \) and the number of operating firms \( N_d = E_d M_d/\lambda \) with respect to \( c_e \) and \( c_f \) are slightly different from the undistorted ones. If \( c_f = 0 \), an increase in \( c_e \) still leads to a decrease in \( Z_d \), \( K_d \), \( E_d \) and \( N_d \). If \( c_f > 0 \), an increase in \( c_e \) leads to a decrease in \( \hat{z}(t) \) for all \( t \), hence a decrease in \( K_d \) and \( E_d \) if \( \hat{z}(t) \) is binding for some firms. Its effect on \( Z_d \) and \( N_d \) is ambiguous. The comparative statics of \( c_f \) is also ambiguous.

\(^9\)The equilibrium in the distorted economy can also be characterized by utilizing the fact that it’s locational equivalent to the one in the undistorted economy. See Hopenhayn (2014b).

\(^10\)A negative fraction would mean that in aggregate the government lump-sum taxes the consumer and subsidizes the output.
2.3.3 Aggregate losses

We define a relative aggregate measure as the ratio of corresponding aggregate variables between the distorted and undistorted economies. An aggregate loss is then one minus the relative measure.\footnote{The word “loss” may sometimes be misleading as the relative measure [i.e. the ratio] may potentially be greater than one. We will be explicit about it whenever confusion may arise. In most cases, we deal with the relative measures directly. Note that a larger value in a relative measure [less than one] means a smaller loss.} The main aggregate variables we are interested in are TFP, capital stock, and output. Their relative measures are given by, \(r\) indicates relative\)

\[
\text{TFP}_r = \frac{Z_d}{Z} = \left(\frac{A_d}{A}\right)^{\frac{1-\gamma}{\gamma}} D, \quad (2.16)
\]

\[
K_r = \frac{K_d}{K} = \left(\frac{A_d}{A}\right)^{\frac{1-\gamma}{1-\alpha}}, \quad (2.17)
\]

\[
Y_r = \frac{Y_d}{Y} = \left(\frac{A_d}{A}\right)^{\frac{1-\gamma}{1-\alpha}} D. \quad (2.18)
\]

In the case of \(c_f > 0\) we are also interested in mass of entrants, mass of operating firms and aggregate consumption.

\[
E_r = \frac{E_d}{E} = \frac{\dot{M}}{M_d} \left(\frac{A_d}{A}\right)^{\frac{1-\alpha}{1-\alpha\gamma}}, \quad (2.19)
\]

\[
N_r = \frac{E_d M_d}{EM/\lambda} = \frac{E[z|z \geq \hat{z}]}{E[z|z \geq \hat{z}(t)]} \left(\frac{A_d}{A}\right)^{\frac{1-\alpha}{1-\alpha\gamma}}, \quad (2.20)
\]

\[
C_r = \frac{C_d}{C} = \frac{Y_d - \delta K_d - c_e E_d - c_f E_d M_d/\lambda}{Y - \delta K - c_e E - c_f EM/\lambda}.
\]

The above expressions are in their most general forms and may look daunting. Once we impose further assumptions on the model, the expressions can be greatly simplified. We do so in the next section.

2.4 Results

We have set the stage to analyze the mappings between the distortions and aggregate losses. We start by examining some of the main findings in RR. In section 4.1, we provide a unified view on three features from RR’s numerical results, where they maintain two assumptions: 1) fixed production cost is zero \((c_f = 0)\) so there is no endogenous exit; and 2) distortions are such that steady-state capital stock in the distorted economy remains the same as the one in the undistorted economy \((K = K_d)\). We also discuss the result from another set of RR’s numerical exercise where capital stock responds to distortions. In section 4.2, we go beyond their exercises and study aggregate losses when fixed production cost is positive \((c_f > 0)\) so the endogenous exit margin is active. We re-calibrate the model under the new assumption and analyze how the endogenous exit margin may change the implications of distortions for the aggregate losses.
Table 2.1: Relative TFP - three sets of experiments in RR

<table>
<thead>
<tr>
<th>Fraction of firms taxed</th>
<th>Uncorrelated $\tau_t$</th>
<th>Correlated $\tau_t$</th>
<th>Exempt $\tau_t$</th>
</tr>
</thead>
<tbody>
<tr>
<td>90%</td>
<td>0.84 0.74</td>
<td>0.66 0.51</td>
<td>0.85</td>
</tr>
<tr>
<td>50%</td>
<td>0.96 0.92</td>
<td>0.80 0.69</td>
<td>0.78</td>
</tr>
<tr>
<td>10%</td>
<td>0.99 0.99</td>
<td>0.92 0.86</td>
<td>0.85</td>
</tr>
</tbody>
</table>

Source: Restuccia and Rogerson (2008)

Note: (1) In all experiments, $c_f = 0$; (2) In the first two sets of experiments, subsidy $\tau_s$ to the remaining firms is chosen such that $K = K_d$.

2.4.1 Examining RR’s results

We briefly review the main results from three sets of numerical experiments conducted by RR. In all these experiments, the fixed production cost is assumed to be zero ($c_f = 0$). Further, distortions are tax or subsidy wedges on output. In the first two sets of their experiments, firms are divided into two groups: one group being taxed at rate $\tau_t$ and the remaining subsidized at rate $\tau_s$. Further, the tax and subsidy rate are such that capital stock in the distorted economy remains the same as the one in the undistorted economy ($K = K_d$). The two sets of experiments are configured to vary along three dimensions: (1) what fraction of the firms are taxed, (2) at what rate they are taxed, and (3) whether or not the distortions and firm-level productivity are independent (referred to as uncorrelated v.s. correlated in RR). In the case that distortions and firm-level productivity are dependent, they are perfect correlated in the sense that firms with high productivity are taxed and firms with low productivity are subsidized. In the third set of their experiments, only a fraction of high productivity firms are taxed and the remaining firms are exempt. As a result, the capital stock responds to the distortions and in general it differs from the one in the undistorted economy.

Table 2.1 summarizes the experiment results on relative TFP reported by RR. The three sets of experiments are referred to as uncorrelated, correlated and exempt in the table. The fraction of firms taxed are 90%, 50% and 10%. The tax rate considered are 0.2 and 0.4. The numbers reported are relative TFP between distorted and undistorted economy, i.e. $TFP_r$ as defined in this paper, so a larger value implies a smaller TFP loss.

Three features stand out from the first two sets of experiments. First, for a given tax rate, the larger the fraction of firms taxed, the larger the TFP loss. Second, for a given fraction of firms taxed, the larger the tax rate, the larger the TFP loss. Third, for given fraction of firms taxed and the tax rate, TFP loss is larger when distortions are correlated. These features may look intuitive. Taxing more firms, taxing at a higher rate and taxing productive firms all seem to make the distortion more “severe” so they should aggravate the misallocation problem and cause larger efficiency loss. However, inspecting the results from the third set of experiments where capital accumulation are affected by the distortions, we see that it is no more true that TFP loss always gets larger when more productive firms are taxed. This prompts us to investigate what precisely drives the results. In the following, we provide a unified view on the above three features and we also explain why we see the result from the third set of experiments.

12This is the same table as in Hopenhayn (2014) where he also reviews the results from Restuccia and Rogerson (2008).
### 2.4.1.1 \( c_f = 0 \) and \( K = K_d \)

In the first two sets of experiments, RR assume that fixed production cost is zero and the distortions do not affect steady-state capital stock. Before explaining the three features above, we first derive the precise mappings between the distortions and aggregate losses under the two assumptions. First, the assumption of zero fixed production cost implies that the cutoff productivity levels in both undistorted and distorted economies are zero, i.e., \( \hat{z} = 0 \) and \( \hat{z}(t) = 0 \) for all \( t \). Therefore, all entering firms choose to operate and there is no endogenous exit (\( M = M_d = 1, \mathbb{E}(z) = M \) and \( \mathbb{E}(zt) = M_d \)). Further, examining the expressions for the conditional distribution of distortions \( G(tz) = 1 \). In other words, maintaining this assumption restricts the set of distortions one can consider. Another implication of this assumption is that aggregate losses in output and TFP are identical (or equivalently, \( TFP_r = Y_r \)). Moreover, inspecting the expressions of \( TFP_r \) and \( K_r \) (equations 2.16 and 2.17), we see that \( TFP_r \) just equals \( D \), the distortion factor. Recall that \( (1 - D^{-1}) \) is the fraction of tax wedge revenue out of total output. Thus, relative TFP also equals the inverse of one minus the fraction of government transfer out of output.

Combining the two assumptions, equation (2.17) of relative capital implies that the restriction on distribution of distortions must be such that \( \mathbb{E}(z) = \mathbb{E}(tz) \), i.e., the mean productivity of firms must equal the mean effective productivity of firms. This restriction further implies that \( K_r, E_r \), and \( N_r \) all equal one. As a result, changes in \( TFP_r \) and \( Y_r \) due to changes in distortions will be in the same direction as changes in relative welfare \( (C_r) \) because investment \( (\delta K) \), total entry cost \( (c_f E) \) and total fixed production cost \( (c_f EM/\lambda) \) are the same in both undistorted and distorted economies. However, this does not imply that the magnitude of TFP loss and output loss are good indicators to the magnitude of welfare loss.

In what follows we only present the \( TFP_r \) since it is the key aggregate statistic under the current assumptions. We do so for two cases, \( t \) and \( z \) are independent, and \( t \) and \( z \) are dependent.

**t and z are independent.** In addition to the two assumptions on fixed production cost and capital stock, if the productivity and distortions are independent (uncorrelated \( \tau_k \)), relative TFP is given by a simple expression together with a constraint on the distribution of distortions,

\[
TFP_r = \mathbb{E}(t^\gamma) \quad s.t. \mathbb{E}(t) = 1. \tag{2.21}
\]

Note that \( \mathbb{E}(t^\gamma) < [\mathbb{E}(t)]^\gamma = 1 \) by Jensen’s Inequality. As just mentioned above, \( TFP_r = D \) and \( (1 - D^{-1}) \) is the fraction of tax wedge revenue out of total output. The fact that \( TFP_r < 1 \) implies that in aggregate government subsidizes the firms.

In this independent case, the TFP loss is independent of the distribution of productivity. What is the intuition? First, when there is no fixed production cost, the cutoff productivity is just zero regardless of the distribution of productivity. Second, because the distortions and productivity levels are independent, if we group firms by their productivity levels, each group will face the same distribution of distortions. In other words, the distribution of distortions is sufficient to summarize the TFP loss, and the distribution

\[\text{We assume that the distribution of } t \text{ is not degenerated so the equality case is ruled out.}\]
of productivity level does not matter.

Note that since $\gamma < 1$, the expression of TFP loss can be viewed as the expected utility of a risk averse agent with Bernoulli utility $u(t) = t^\gamma$. It therefore follows from the standard consumption theory under uncertainty that TFP loss would increase under a mean preserving spread of the distribution of $t$. Note that the constraint holds under a mean preserving spread of the distribution of $t$.

$t$ and $z$ are dependent If productivity and distortions are dependent, the relative TFP is given by

$$TFP_r = \frac{\mathbb{E}(t^\gamma z)}{\mathbb{E}(z)} \quad \text{s.t. } \mathbb{E}(tz) = \mathbb{E}(z).$$

In this case, TFP loss depends on the distribution of productivity. It is not immediately obvious that $TFP_r < 1$. However, it can be shown that under a transformed probability measure, $TFP_r$ can be written in a similar expression as the one in the independent case. Hence Jensen’s Inequality again applies and $TFP_r < 1$. We state this result as a proposition and prove it. The proof also provides clues to understand the three features discussed above.

**Proposition 2.1.** Under the model assumption that $c_f = 0$ and $K = K_d$, $TFP_r < 1$.

**Proof.** The proof starts with a change of probability measure, and then applying Jensen’s Inequality.\(^{14}\) Let $X = \frac{\mathbb{E}(z(t))}{\mathbb{E}(z)}$. Note that $X$ is a positive random variable (r.v.) on some probability space $(\Omega, F, \mathbb{P})$ and $\mathbb{E}[X] = 1$. As a result, it can be shown that $\mathbb{P}(A) = \int_A X(\omega)d\mathbb{P}(\omega)$ for $A \in F$ is a probability measure. Since $t$ is a positive r.v., we have $\mathbb{E}(t) = \mathbb{E}[tX] = \frac{\mathbb{E}(tz)}{\mathbb{E}(z)}$, and $\mathbb{E}(t^\gamma) = \mathbb{E}[t^\gamma X] = \frac{\mathbb{E}(t^\gamma z)}{\mathbb{E}(z)}$, where $\mathbb{E}$ is the expectation under the probability measure $\mathbb{P}$. Hence relative TFP can be re-written as

$$TFP_r = \frac{\mathbb{E}(t^\gamma)}{\mathbb{E}(t)} \quad \text{s.t. } \mathbb{E}(t) = 1. \tag{2.22}$$

Since $\gamma \in (0, 1)$, by Jensen’s Inequality, $TFP_r = \frac{\mathbb{E}(t^\gamma)}{\mathbb{E}(t)} < \left[\frac{\mathbb{E}(t)}{\mathbb{E}(t)}\right]^\gamma = 1$. We assume that the distribution of $t$ is not degenerated so as to rule out the equality case. \(\square\)

The key step in the proof is the change of probability measure. The random variable $X$ relates the original and new probability measures $\mathbb{P}$ and $\mathbb{P}$. It is the ratio of two average productivity levels, one for firms conditional on distortions and one for all firms. Thus, it measures the deviation between average productivity of firms conditional on distortions to the average productivity of all firms. If $X > 1$, the original probability is revised upward. If $X < 1$, the original probability is revised downward. Relative TFP is thus a weighted sum of a strictly concave function of distortion, $t^\gamma$. The weights are the mass of firms affected by the distortion $t$ adjusted by $X$. An alternative way to write equation (2.22) is

$$TFP_r = \int t^\gamma dT(t) \quad \text{s.t. } \int t \cdot dT(t) = 1,$$

where $dT(t) = \frac{\mathbb{E}(z(t))}{\mathbb{E}(z)} dG_t(t)$ is the “weight”. Note that $\int dT(t) = 1$.

The expression of relative TFP implies that similar as in the independent case, a mean preserving spread of the distribution of distortions under the new probability measure would result in a larger TFP loss. In addition, for the same reason as in the independent case, the fact that $TFP_r < 1$ implies that in aggregate government subsidizes the firms.

\(^{14}\) See Shreve (2004) Theorem 1.6.1 for change of measure and other related results used in the proof. I thank Weichi Wu and Yichun Chi for pointing out this proof.
Chapter 2. Idiosyncratic Distortions & Aggregate Losses

**RR’s results explained** Now, we are ready to study the three features emerged from RR’s first two sets of experiments. We show that all of them corresponds to a mean preserving spread of the distribution of distortions under the new probability measure.

We start with the first feature: the larger the fraction of firms taxed, the larger the TFP loss. To understand it, consider the case of correlated distortions where a fraction \( \phi \) of the productive firms are taxed at rate \( \tau_t \) and the rest less productive firms are subsided at rate \( \tau_s \). Recall that the distortion \( t \) is just a transformation of the tax rate \( \tau \): \( t = (1 - \tau)^{\frac{1}{\tau_t}} \). Let the corresponding tax and subsidy distortions be \( t_s \) and \( t_s \). The distribution of distortions is then \( \{(t_t, \phi), (t_s, 1 - \phi)\} \). Denote \( x(\phi, t_t) = \frac{E[z | t = t_t]}{E(z)} \), the average productivity of taxed firm relative to the average productivity of all firms. Similarly, denote and \( x(\phi, t_s) = \frac{E[z | t = t_s]}{E(z)} \), the average productivity of subsidized firm relative to the average productivity of all firms. Following the change of probability measure in the proof, the distribution of distortions under the transformed probability measure is \( \{(t_t, \phi x(\phi, t_t)), (t_s, (1 - \phi)x(\phi, t_s))\} \). The expectation of the distortions under the new probability measure must be one.

Now consider an increase in \( \phi \) to \( \phi' \) such that more fraction of productive firms are taxed at the same rate \( \tau_t \). The remaining firm are subsidized at rate \( \tau_s' \). Let \( (t_t, t_s') \) be the corresponding tax and subsidy distortions. Note that since more fraction of productive firms are taxed and \( E(t_z) = E(z) \) must hold, we have \( t_s' > t_s \). Define \( x(\phi', t_t) = \frac{E[z | t = t_t]}{E(z)} \) and \( x(\phi', t_s') = \frac{E[z | t = t_s]}{E(z)} \). The distribution of the new distortions under its transformed probability measure is \( \{(t_t, \phi' x(\phi', t_t)), (t_s', (1 - \phi')x(\phi', t_s'))\} \). The expectation of the distortions under the new probability measure must also be one. Since a smaller fraction of less productive firms are subsidized, we have \( (1 - \phi') < (1 - \phi) \) and \( x(\phi', t_s') \leq x(\phi, t_s) \). Hence comparing the distributions of the two sets of distortions, it must be \( (1 - \phi')x(\phi', t_s') < (1 - \phi)x(\phi, t_s) \) and \( \phi' x(\phi', t_s') > \phi x(\phi, t_s) \).

Figure 2.2 plots the transformed distributions of the two sets of distortions. Since the two sets of distortions have the same expected value, the areas of the rectangles A and B must be the same. It then follows from the graph that the distribution of \( (t_t, t_s') \) second order stochastic dominates the distribution of \( (t_t, t_s) \), or equivalent, the former is a mean preserving spread of the latter.

The argument for the uncorrelated case is the same because it is just a special case under the uncorrelated one where \( x(\cdot, \cdot) = 1 \).

![Figure 2.2: Two Distribution (larger fraction is taxed at the same tax rate)](image)

The second feature of the results is that the larger the tax rate, the larger the TFP loss. It can also be understood as a mean-preserving spread of the distribution of distortions. The argument is very similar. We again consider the case of correlated distortions where a fraction \( \phi \) of the productive firms
are taxed at rate $\tau_t$ and the rest less productive firms are subsided at rate $\tau_s$. In the same way (and using the same notation) we obtain the distribution of distortions under the transformed probability measure as $\{(t_t, \phi x(\phi, t_t)), (t_s, (1 - \phi)x(\phi, t_s))\}$. The expectation of the distortions under the new probability measure is one.

Now consider an increase in tax rate from $\tau_t$ to $\tau'_t$. Correspondingly the tax distortion decreases from $t_t$ to $t'_t$. Let $t'_s$ be the new subsidy distortion. Since the fraction of productive firms taxed is unchanged and $E(tz) = E(z)$ must hold, the subsidy distortion must increase, $t'_s > t_s$. Define $x(\phi, t'_t) = \frac{E(z|t=t'_t)}{E(z)}$ and $x(\phi, t'_s) = \frac{E(z|t=t'_s)}{E(z)}$. Note that $x(\phi, t'_t) = x(\phi, t_t)$ and $x(\phi, t'_s) = x(\phi, t_s)$. After changing the probability measure, the distribution of the distortions is $\{(t'_t, \phi x(\phi, t'_t)), (t'_s, (1 - \phi)x(\phi, t'_s))\}$. The expectation of the distortions under the new probability measure is one. Since $\phi x(\phi, t_t) = \phi x(\phi, t'_t)$ and $(1 - \phi)x(\phi, t_s) = (1 - \phi)x(\phi, t'_s)$, the distribution is obviously a mean preserving spread of the original one.

Figure 2.3 plots the two distributions under the transformed probability measure and illustrate the above argument. Since the two sets of distortions have the same expected value, the areas of the rectangle $A$ and $B$ must be equal. It then follows from the graph that the distribution of $(t'_t, t'_s)$ second order stochastic dominates the distribution of $(t_t, t_s)$. Equivalent, the former is a mean preserving spread of the latter.

![Figure 2.3: Two Distributions (same fraction is taxed at a higher rate)](image)

The third feature, larger TFP loss when distortions are correlated with productivity, can also be shown to correspond to a mean preserving spread of the distribution of distortions. Start again with the case of correlated distortions where a fraction $\phi$ of the productive firms are taxed at rate $\tau_t$ and the rest less productive firms are subsided at rate $\tau_s$. The distribution of the distortions under the transformed probability measure is $\{(t_t, \phi x(\phi, t_t)), (t_s, (1 - \phi)x(\phi, t_s))\}$. Note that $x(\phi, t_t) > 1$ and $x(\phi, t_s) < 1$. The expectation of the distortion is one.

Now suppose the distortion is independent of the firm-level productivity, i.e., instead of taxing $\phi$ fraction of the productive firms, firms are taxed at the random and $\phi$ fraction of firms are taxed. Let $t'_s$ be the subsidy distortion under random taxation. Since firms are now taxed at random and $E(tz) = E(z)$ must hold, we have $t'_s < t_s$. Random taxation also implies $\frac{E(z|t)}{E(z)} = 1$ for $t = t_t, t'_s$. As a result the distribution of distortions under the transformed probability measure is $\{(t_t, \phi), (t'_s, (1 - \phi))\}$. The expectation of the distortions under the new distribution is one. Figure 2.4 plots the distributions of the two sets of distortions. Same argument shows that the distribution of the correlated distortions is a mean preserving spread of the uncorrelated one.
We thus have provided a unified explanation to all the three features: larger TFP losses when varying along the three dimensions of the experiments all correspond to mean preserving spreads of the distribution of distortions under a transformed probability measure.

Other than the three features, two things related to the correlated distortions are worth noting. First, a negative correlation between productivity and distortions is not necessary to get a large TFP loss. In general, a negative correlation means that more productive firms are more likely to be taxed and less productive firms are more likely to be subsidized. As is evident from the expression of relative TFP when productivity and distortions are independent, equation (2.21), TFP loss can get arbitrarily large as the dispersion of the distortion gets large. Second, in the exercise moving from the uncorrelated distortions to the correlated ones, the marginal distribution of the distortion is not kept unchanged. Ideally, one would want to hold marginal distribution of distortions and productivity unchanged, and to check if TFP loss is larger when distortions and productivity are more negatively correlated. Unfortunately, such experiment is not possible under the assumptions of $c_f = 0$ and $K = K_d$. In the correlated case, we have that the correlation

$$corr(t, z) = \frac{cov(t, z)}{\sqrt{var(t)var(z)}} = \frac{E(tz) - E(t)E(z)}{\sqrt{var(t)var(z)}} = \frac{E(z) [1 - E(t)]}{\sqrt{var(t)var(z)}}.$$ 

The last equality follows from the restriction $E(tz) = E(z)$. Hence, when $corr(t, z)$ changes, variance or mean of $t$ or $z$ has to change, implying that we cannot hold the marginal distribution of $t$ or $z$ unchanged.

**Hopenhayn (2014b)** In a static model of heterogeneous firms with a single input factor (labor), Hopenhayn (2014b) provides a similar explanation to the three features above. One may wonder how a much simpler model – one input factor and no selection margin – can explain results from a rather complicated model as RR’s. The reason lies in the three assumptions of RR’s numerical experiment. First, the distortions are only output wedges. The assumption keeps the effect of the distortions symmetric to the two input factors as opposed to, for example, an assumption to put wedges on both output and one of the input, which will create asymmetry. Viewing capital and labor as a compound input, the
symmetry makes it possible to analyze RR’s results in a single input setting. Second, the distortions are restricted so as to have no impact on capital accumulation. Indeed, under the model of Hopenhayn (2014b), the expression for relative TFP is given by (in our notation)

$$TFP_r = \frac{\mathbb{E}(t^\gamma z)}{[\mathbb{E}(tz)]^{\gamma} [\mathbb{E}(z)]^{1-\gamma}}. \quad (2.23)$$

As will be discussed in the next section, the expression for relative TFP under RR’s experiment but without restriction on distortions is given by

$$TFP_r = \left[\frac{\mathbb{E}(tz)}{\mathbb{E}(z)}\right]^{\frac{1-\gamma}{\gamma}} \frac{\mathbb{E}(t^\gamma z)}{\mathbb{E}(tz)}.$$

The two expressions are quite different. However, once we impose the restriction on the distortions, which is $\mathbb{E}(tz) = \mathbb{E}(z)$. The two expressions becomes the same. Third, fixed cost of production is assumed to be zero so there is no endogenous exit. Without the endogenous exit margin, steady-state analysis of RR’s results thus can be done equivalently in a static setting. In summary, these three assumptions combined render the mapping from distortions to TFP loss in RR’s exercises exact the same as the one in Hopenhayn (2014b).

Since Hopenhayn (2014b) also provides a unified explanation to the three features, another thing one may wonder is the difference between his explanation and ours. To explain the three features, Hopenhayn (2014b) first show that the expression of relative TFP given by 2.23 can be re-written as

$$TFP_r = \frac{\int \theta \cdot dL(\theta)}{L},$$

where $\theta$ is a general firm level distortion and $L(\theta)$ is a measure over $\theta$. The general distortion $\theta$ is defined as the ratio of labor inputs between distorted and undistorted economies, $\theta = \frac{\ell(tz)}{\ell(z)}$. The measure $L(\theta)$ is defined such that $\int_{\Theta} dL(\theta)$ gives the number of employment in the undistorted economy that is affected by the distortions $\theta \in \Theta$. Therefore, it must satisfy $L = \int dL(\theta)$. In addition, feasibility implies that it also must satisfy $L = \int \theta dL(\theta)$. Given the defined general distortions and their measure, he then shows that all three features from RR’s result correspond to mean-preserving spreads of this measure.

Inspecting the above expression of relative TFP and ours as of equation 2.22, we see that both of them summarize TFP loss in a simple and similar way. Of course, the mapping implied by these two expressions must be the same.\footnote{Otherwise, the same distortions would lead to two different TFP losses, which is absurd.} However, the two expressions are different and they are derived and interpreted differently. Our work therefore complements Hopenhayn (2014b) in that we take a different approach to derive the mapping from distortions to TFP loss under the assumptions of $c_f = 0$ and $K = K_u$. Because of the different approach, we are able to view the mapping from a new angle and gain new insights.

More importantly, we will extend our analysis in the next step by removing the restriction on distortions and allowing positive fixed production cost, both of which are not considered in Hopenhayn (2014b). The slightly complicated model we take from RR and the mappings we derived in the last section give us the flexibility the push the analysis further.
2.4.1.2 \( c_f = 0 \)

Now we examine the third set of RR’s experiment, in which case the fixed production cost is zero but the distribution of distortions is not restricted to keep capital stock in the distorted economy the same as the one in the undistorted economy. The implications of zero fixed production cost is already discussed. In particular, one of them is \( K_r = E_r = N_r \). Allowing capital stock to respond to distortions implies that \( TFP_r \) and \( Y_r \) are in general different. In what follows, we therefore present relative measures in TFP, capital and output. We do so for the case that \( t \) and \( z \) are independent and the case that they are dependent.

**t and z are independent** When distortions and productivity levels are independent, relative TFP, capital and output are given by

\[
TFP_r = \left[ \frac{E(t)}{E(t')} \right]^{1-\gamma} \frac{E(t')}{E(t)}, \quad K_r = \left[ \frac{E(t)}{E(t')} \right]^{1-\gamma} \frac{1}{\gamma} \frac{E(t')}{E(t)}, \quad Y_r = \left[ \frac{E(t)}{E(t')} \right]^{1-\gamma} \frac{1}{\gamma} \frac{E(t')}{E(t)}
\]

Again, in the independent case, the aggregate losses do not depend on the distribution of productivity. However, different from the case when \( K = K_d \) is enforced, in general, \( TFP_r \) may not be less than 1. In other words, there might be a gain in TFP. While distortions certainly lowers the welfare (consumption) and make the representative consumer worse off, it may not be reflected in the TFP.

Since \( E(t') < [E(t)']^{\gamma} \) by Jensen’s, \( TFP_{loss} = \left[ E(t) \right]^{1-\gamma} \frac{E(t')}{E(t)} < \left[ E(t) \right]^{1-\gamma} \frac{1}{\gamma} \frac{E(t')}{{E(t)}} \). A sufficient condition for \( TFP_r \) to be less than 1 is that \( E(t) \leq 1 \), i.e. on average, firms are not being subsidized. When \( \gamma \in (0.5, 1) \), which is the case when the parameter \( \gamma \) is calibrated, a weaker sufficient condition is that \( E(t') \leq 1 \). (See Appendix 2.B.1 for a proof.)

Capital loss only depends on the mean of distortions. Since \( K_r < 1 \) if and only if \( E(t) < 1 \), the more firms are on average taxed, the larger the capital loss. If firms are on average subsidized, capital stock in the distorted economy will be larger than the one in the undistorted economy.

**t and z are dependent** When distortions and productivity levels are dependent, relative TFP, capital and output are given by

\[
TFP_r = \left[ \frac{E(tz)}{E(z)} \right]^{1-\gamma} \frac{E(t'z)}{E(tz)}, \quad K_r = \left[ \frac{E(tz)}{E(z)} \right]^{1-\gamma} \frac{1}{\gamma} \frac{E(t'z)}{E(tz)}, \quad Y_r = \left[ \frac{E(tz)}{E(z)} \right]^{1-\gamma} \frac{1}{\gamma} \frac{E(t'z)}{E(tz)}
\]

Similar as in the independent case, \( TFP_r \) may not be less than 1. A sufficient condition for \( TFP_r \) to be less than 1 is that \( \frac{E(tz)}{E(z)} \leq 1 \), i.e. the average effective productivity is less than the average productivity. When \( \gamma \in (0.5, 1) \), a weaker sufficient condition is that \( \frac{E(t'z)}{E(tz)} \leq 1 \). (See Appendix 2.B.1 for a proof.)

The loss of capital stock depends on the ratio between average effective productivity and average productivity. Even when \( TFP_r < 1 \), it’s still possible that \( K_r > 1 \).

**RR’s results explained** The results from the third set of RR’s experiments do not follow the features discussed above. From the last column of the table 2.1, we see that the relative TFP is not monotonic decreasing when more fraction of productive firms are taxed. For example, taxing 90% of productive firms results in less TFP loss than taxing 50% of them.\(^{16}\)

\(^{16}\)Recall that less TFP loss corresponds to a larger value of relative TFP in the table.
Table 2.2: Relative TFP, Capital and Output ($\tau_t = 0.4$)

<table>
<thead>
<tr>
<th>Relative Measures</th>
<th>TFP</th>
<th>Capital</th>
<th>Output</th>
</tr>
</thead>
<tbody>
<tr>
<td>90%</td>
<td>0.85</td>
<td>0.42</td>
<td>0.66</td>
</tr>
<tr>
<td>50%</td>
<td>0.78</td>
<td>0.53</td>
<td>0.65</td>
</tr>
<tr>
<td>10%</td>
<td>0.85</td>
<td>0.75</td>
<td>0.78</td>
</tr>
</tbody>
</table>

Source: Restuccia and Rogerson (2008)
Note: (1) $c_f = 0$; (2) A fraction of productive firms are taxed at rate $\tau_t = 0.4$ and the rest of firms are exempt.

To understand these results, table 2.2 reports, in addition to relative TFP, relative capital and output from the experiments. All the numbers are taken from RR. We see that unlike relative TFP, relative capital monotonically decreases as more fraction of productive firms are taxed. However, relative output is similar to relative TFP, it first decreases and then increases.

The analytical mappings allows us to better understand these results. As more productive firms are taxed, the average effective productivity of firms decreases, i.e. $E(tz)$ decreases. Since $K_r$ increases in $E(tz)$, it also decreases. However, the term $\frac{E(tz)}{E(t)}$ in $TFP_r$ and $Y_r$ increases because its numerator does not decrease as much as its denominator. Therefore, $TFP_r$ and $Y_r$ are non-monotone, first decreases and then increases.

When we allow capital stock to respond to distortions, we can hold marginal distributions of $t$ and $z$ unchanged, and investigate whether a negative correlation between $t$ and $z$ results in larger TFP loss than the case that $t$ and $z$ are independent, and whether more negative correlation results in larger TFP loss. We are not able to answer this question in general. In what follows, we show that the answer to both questions are yes under the joint log normal case.

Assuming joint log normal of distortions and productivity levels. Precisely, let $x = \ln(t)$ and $y = \ln(z)$. We assume that

\[
(x, y) \sim N\left(\begin{pmatrix} \mu_x \\ \mu_y \end{pmatrix}, \begin{pmatrix} \sigma_x^2 & \sigma_{xy} \\ \sigma_{xy} & \sigma_y^2 \end{pmatrix}\right).
\]  

In the independent case, we get

\[
\ln TFP_r = \frac{1 - \gamma}{\gamma} \left(1 - \gamma\right) \left(\mu_x + \frac{1}{2} \sigma_x^2\right) - \frac{\gamma^2}{2} \sigma_x^2.
\]

In the correlated case, we get

\[
\ln TFP_r = \frac{1 - \gamma}{\gamma} \left(1 - \gamma\right) \left(\mu_x + \frac{1}{2} \sigma_x^2 + \sigma_{xy}\right) - \frac{\gamma^2}{2} \sigma_x^2.
\]

In the log normal case, it is easy to see that $TFP_r$ may be greater than one. A sufficient condition for relative TFP to be less than one in the independent case is that $\mu_x + \frac{1}{2} \sigma_x^2 < 0$, or equivalently $E(t) < 1$, i.e. firms are on average taxed. A sufficient condition for relative TFP to be less than one in the dependent case is that $\mu_x + \frac{1}{2} \sigma_x^2 < 0$ and $\sigma_{xy} < 0$, or equivalently, $E(t) < 1$ and $\text{cov}(t,z) < 0$, i.e. firms are on average taxed and distortions and productivity levels are negatively correlated. Since $\text{cov}(t, z) = E(t)E(z) (e^{2\sigma_{xy}} - 1)$, for given marginal distributions of distortions and productivity levels, i.e. fixing $\mu_x$, $\sigma_x^2$, $\mu_y$ and $\sigma_y^2$, a negative correlation between $t$ and $z$ leads to a larger TFP loss than the


### Table 2.3: Calibration

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value</th>
<th>Target</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\gamma$</td>
<td>0.85</td>
<td>taken from RR</td>
</tr>
<tr>
<td>$\alpha$</td>
<td>0.3333</td>
<td>capital income share</td>
</tr>
<tr>
<td>$\beta$</td>
<td>0.96</td>
<td>real rate of return (4%)</td>
</tr>
<tr>
<td>$\delta$</td>
<td>0.072</td>
<td>capital output ratio (2.5)</td>
</tr>
<tr>
<td>$c_e/c_f$</td>
<td>93.2510</td>
<td>productivity cutoff</td>
</tr>
<tr>
<td>$\lambda$</td>
<td>0.0549</td>
<td>annual exit rate (0.1)</td>
</tr>
<tr>
<td>$\mu_y$</td>
<td>0.2478</td>
<td>establishment size distribution</td>
</tr>
<tr>
<td>$\sigma_y^2$</td>
<td>3.9768</td>
<td>same as above ($\tilde{z}$ normalized to 1)</td>
</tr>
</tbody>
</table>

Note: Data on size distribution are from Restuccia and Rogerson (2008).

case that $t$ and $z$ are independent, and the larger the negative correlation, the larger the TFP loss.

#### 2.4.2 $c_f > 0$

In this section, we go beyond RR’s experiments. We assume that fixed production cost $c_f > 0$ so firms may endogenously exit. We also impose no restriction on distortions so in general capital stock in the distorted economy is different from the one in the undistorted economy. When there is endogenous exit, distortions can have an important effect on the selection margin: firms that exit in the undistorted economy may stay operate in the distorted economy, and firms that operate in the undistorted economy may be forced to exit in the distorted economy.

A natural question is how aggregate losses are affected by the distortions when the endogenous exit margin is active. Our second goal in this paper is to examine this question. However, under the assumption that $c_f > 0$, the mappings between distortions and aggregate losses are in their most general form and give us little analytical insights. To make progress, we further assume that productivity and distortions are joint log normal, and we solve the mappings numerically.\footnote{We can obtain sharper analytical mappings if we assume that firm-level productivity is Pareto distributed, and distortions and productivity are independent or perfectly correlated. However, to study the effect of correlation between distortions and productivity on the aggregate losses, we would still need to resort to numerical method.} We first calibrate the undistorted economy to the US data. We then design and perform experiments to study how distortions may affect aggregate losses with and without the endogenous exit margin.

**Calibration** Assuming $(t, z)$ are log normal as in the last section, the parameters to calibrate in the undistorted economy are $(\gamma, \alpha, \beta, \delta, c_e/c_f, \lambda, \mu_y, \sigma_y^2)$. We identify up to $c_e/c_f$ because all interested relative aggregate measures depend only on the ratio of entry cost and fixed production cost. Table 2.3 summarizes the calibrated values and their targets.

The first three parameters $(\gamma, \alpha, \beta)$ are calibrated in the same way as in RR. The curvature parameter is taken from RR to be 0.85. RR argue that estimates of production function and various calibration procedures in the literature point to this value. The parameter $\alpha$ is calibrated to target the capital income share, which is 1/3. Targeting a real rate of return of 4%, we obtain the discount factor $\beta = 0.96$.

The rest five parameters $(\delta, c_e/c_f, \lambda, \mu_y, \sigma_y^2)$ are calibrated differently from RR. We discuss them in details below. We calibrate the depreciation rate to be 0.072 by targeting the US capital output ratio of 2.5. Depreciation rate are usually calibrated to match investment to output ratio. In the model,
investment consists of not only capital investment but also total entry cost and fixed production cost. We find that using the investment to output ratio as a target yields unreasonable low depreciation rate and high capital output ratio in the model. The implied investment to output ratio under the current calibration is 0.27, only slightly higher than the US number, which is about 0.20-0.22.

We use the employment distribution of establishments to calibrate the distribution of firm-level productivity. Under our model specification, the relative demand for labor between two firms in the undistorted economy is simply their relative productivity.

\[
\frac{l_i}{l_j} = \frac{z_i}{z_j}.
\]

Therefore, the model implies a left-truncated log-normal employment distribution. We normalize the cutoff productivity to be 1, \( \hat{z} = 1 \). In other words, the lower bound of the productivity distribution after the selection is normalized to be one. We then fit a left-truncated log-normal distribution to the employment distribution in the data. This implies \( \mu_y = 0.2478 \) and \( \sigma^2_y = 3.9768 \). Figure 2.5 displays the fit.

![Figure 2.5: Establishment Size Distributions – Model vs. Data](image)

After calibrating the productivity distribution, equation (2.10) implies \( \frac{c_e}{c_f} = 93.2510 \). At the steady state, the number of operating firms is \( N = EM/\lambda \) and the number of exiting firms per period is \( E \). Targeting at an annual exit rate of 0.1, we obtain \( \lambda = 0.0549 \).

**Experiments & results**  With the calibrated undistorted economy, we experiment on how distortions may affect the aggregate losses under the assumption of positive or zero fixed cost. In particular,
under our joint log normal specification, we set baseline parameters of distortions to be \((\mu_z, \sigma_z^2, \rho_{zy}) = (-2, 1, 0)\), where \(\mu_z, \sigma_z^2\) and \(\rho_{zy}\) are as specified in 2.24. The baseline parameters imply a tax wedge of 0.25 on average.\(^{20}\) Note that \(\mu_z\) and \(\sigma_z^2\) govern the mean and variance of distortions, \(\mathbb{E}(t)\) and \(\text{Var}(t)\). Together with \(\rho_{zy}\), they also determine the correlation between the distortions and productivity, \(\text{corr}(t, z)\).\(^{21}\) In the experiment, we vary \(\mathbb{E}(t), \text{Var}(t)\), and \(\text{corr}(t, z)\) one at a time to examine how they affect aggregate losses. We perform the experiment for \(c_f > 0\) first. We then set \(c_f = 0\) and repeat the same experiment. Note that we do not re-calibrate the model when setting \(c_f = 0\). The reason is that our goal is not to assess the bias of model mis-specification, i.e., \(c_f > 0\) is the correct model, but one takes \(c_f = 0\) and calibrates the model under the incorrect assumption. Our goal is to study how the assumption of \(c_f > 0\) may affect the aggregate losses comparing with the assumption of \(c_f = 0\) for given distortions, firm-level productivity and other parameter values.

Figure 2.6 demonstrates how distortions may affect the decision of entrants on whether to operate.\(^{22}\) The figure is plotted under the calibrated parameters and the baseline parameters of distortions. The x-axis is the log productivity of entrants, and the y-axis is the log distortions they face. The scattered plot of the “+” signs illustrates the distribution of entrants over their productivity levels and the distortions they face. The solid vertical line is the log cutoff productivity in the undistorted economy. Firms to the right of this line are surviving firms in the undistorted economy. The dash line with a negative slope is the log cutoff function in the distorted economy. As the distortion \(t\) increases, the cutoff function \(\hat{c}(t)\) decreases. Firms right to this dash line are surviving firms in the distorted economy. The distortions, therefore, alter the operating versus exit decisions of two sets of firms. The first set of firms are those lie in the upper triangle in the figure. They exit in the undistorted economy but operate in the distorted one. The second set of firms are those lie in the lower triangle. They operate in the undistorted economy but are forced to exit in the distorted one.\(^{23}\)

Figures 2.7, 2.8 and 2.9 report the relative aggregate measures when we vary \(\mathbb{E}(t), \text{Var}(t)\), and \(\text{corr}(t, z)\) respectively. The solid line is for the case \(c_f > 0\) and the dash line is for the case \(c_f = 0\). Before discussing our results, we emphasize that these results are obtained under specific assumptions on joint distribution of distortions and productivity. Therefore, they only indicate what distortions can do to the aggregate losses, and we should be cautious about drawing general conclusions from these results. Nevertheless, understanding these results serves as a guidance for studies that model the resource misallocation and quantify its impact on aggregate losses.

Let us first focus on the results of varying \(\mathbb{E}(t)\). In this experiment, we adjust \((\mu_z, \sigma_z^2)\) such that \(\mathbb{E}(t)\) increases from its initial value of 0.22 to 1.79 while \(\text{Var}(t)\) and \(\text{corr}(t, z)\) = 0 remain unchanged. Correspondingly, firms move from being taxed at an average rate of 0.25 to being subsidized at an average rate of 0.08. From figure 2.7, we see that within the specified range of \(\mathbb{E}(t)\), all relative measures increases as \(\mathbb{E}(t)\) increases. Almost all of them, except the relative consumption under \(c_f = 0\), reach values greater than one at some point, which means that the corresponding aggregate measures can be higher under the distortions. Perhaps it’s a bit surprising that consumption, i.e., consumer welfare, can be higher in the distorted economy when \(c_f > 0\), which implies that the steady-state competitive

\(^{20}\)The average wedge is given by \(\mathbb{E}(\tau) = \mathbb{E}(1 - e^{1-\gamma}) = 1 - e^{(1-\gamma)\mu_x + \frac{1}{2}(1-\gamma)^2\sigma_x^2}\). A positive average implies a tax wedge on average, and a negative average implies a subsidy wedge on average.

\(^{21}\)\(\mathbb{E}(t) = e^{\mu_x + \sigma_x^2/2}\) and \(\text{Var}(t) = \left(e^{\sigma_x^2} - 1\right)e^{2\mu_x + \sigma_x^2}\). In addition, we have \(\mathbb{E}(tz) = e^{\mu_x + \mu_y + \left(\sigma_x^2 + \sigma_y^2 + 2\rho_{xy}\sigma_x\sigma_y\right)/2}\), and \(\text{corr}(t, z)\) can then be derived.

\(^{22}\)Yang (2014) present a similar graph under a monopolistic competition setting.

\(^{23}\)Yang (2014) calls the first set of firms zombies and the second set shadows, which summarizes the nature of the two sets of firms quite nicely.
equilibrium under \( c_f > 0 \) is not Pareto optimal. However, note that welfare can still improve if we take into account the welfare accrued during the transition period after removing the distortions. In a similar economic environment to ours but with monopolistic competition production side and no capital, Fattal Jaef (2014) also finds that welfare decreases in steady state when distortions are removed, but taking into account the transition period revert the result. Fattal Jaef and Hopenhayn (2012) have a very similar model to ours except that firms produce without capital input. They show that the competitive equilibrium in their model is Pareto optimal. Their result suggests that the capital accumulation channel could be the key behind our finding on steady-state welfare effect.

Comparing the \( C_r \) and \( TFP_r \) plots, we see that the magnitude of TFP loss is often not a good indicator to welfare loss. Welfare loss can be large while TFP loss is moderate. This is especially true when \( c_f = 0 \). Moreover, we see that TFP loss is less responsive to the endogenous exit margin than welfare loss. For example, moving from \( c_f = 0 \) to \( c_f > 0 \) under our baseline parameters where \( \mathbb{E}(t) = 0.22 \), welfare loss \( (1 - C_r) \) decreases more than 60%, while TFP loss \( (1 - TFP_r) \) almost doesn’t change.\(^{24}\) In fact, except relative consumption, all other aggregate measures are not responsive to the endogenous exit margin under this experiment.

Let us move to the results of varying \( Var(t) \). Similarly, we adjust \((\mu_x, \sigma^2_x)\) such that \( Var(t) \) increases from its initial value of 0.09 to 0.94 while \( \mathbb{E}(t) \) and \( Corr(t, z) = 0 \) remain unchanged. Correspondingly, firms move from being taxed at an average rate of 0.25 to 0.34. In this case, as \( Var(t) \) increases, relative measures do not change in the same direction. For example, independent of \( c_f \), \( TFP_r, Y_r \) and \( C_r \) all decreases as \( Var(t) \) increases. When \( c_f > 0 \), \( K_r \) and \( E_r \) increases and \( N_r \) decreases. As expected, when \( c_f = 0 \), \( K_r, E_r \) and \( N_r \) stay unchanged since these measures only depend on \( \mathbb{E}(t) \) when \( t \) and \( z \) are independent. Similar to the previous case, TFP and output losses almost do not respond to the

\(^{24}\)TFP loss increases by 1%
endogenous exit margin, but welfare loss is quite sensitive to it. We see again that the magnitude of TFP loss is not a good indicator for welfare loss. In addition, examining the Y-axis scales of $K_r$ and $E_r$, we see that the responses of relative capital and entrants to the endogenous exit margin is also minimum.

Finally, we discuss the results on varying $\text{Corr}(t,z)$. In this experiment, we adjust $\rho_{xy}$ such that $\text{Corr}(t,z)$ increases from its initial value of -0.09 to 0.55. The corresponding change in $\rho_{xy}$ is from -0.96 to 0.91. Note that the attainable range of $\text{Corr}(t,z)$ is very different from $\rho_{xy}$ due to our joint log normal assumption. As $\text{Corr}(t,z)$ or $\rho_{xy}$ varies, the marginal distribution of distortions does not change, so firms are always taxed at an average rate of 0.25. The most notable feature in this case is that as $\text{Corr}(t,z)$ increases how relative number of operating firms change depends on whether the endogenous exit margin is active. If it is active, the relative number of operating firms declines, while if it is not, the opposite is true. Other than this feature, all relative measures increases just like in the first experiment although in this case $\text{TFP}_r, Y_r$ and $C_r$ all stay less than one. The fact that $\text{TFP}_r$ increases in $\text{Corr}(t,z)$ is consistent with the findings in the literature that more negative correlation between distortions and firm productivity generally leads to larger TFP loss. In this experiment $C_r$ is again sensitive to the endogenous exit margin, but the response of $\text{TFP}_r, K_r, Y_r$ and $E_r$ to the endogenous exit margin is negligible. Again, TFP loss is a bad measure for welfare loss, especially when $c_f = 0$.

Comparing across three experiments, we summarize a few key features. 1) For the same set of distortions, TFP, capital and output losses can be insensitive to the endogenous exit margin while welfare loss can be very sensitive to it; 2) TFP loss is generally a bad measure for welfare loss; 3) Whether or not the endogenous exit margin is considered, larger dispersion of distortions and larger negative correlation between distortions and firm-level productivity can lead to larger losses in TFP, output and welfare. The first two features deliver the bottom-line message of our experiments: Empirical studies on impact of misallocation should assess not only TFP loss but also welfare loss. In addition, to correctly measure welfare loss, a careful evaluation of the presence of fixed production cost can be necessary.
2.5 Concluding remarks

In this paper, we characterized the precise mappings between distortions and aggregate losses under RR’s economic environment. Using the mappings, we provided an analytical study on RR’s numerical results. We explained three features from RR’s main numerical results as well as why those features do not hold when capital accumulation responds to distortions. We also extended their work to allow for the endogenous exit margin and we showed that this margin can change the implications of distortions on aggregate losses. Our work suggests that empirical studies on misallocation should take into account the effect of capital accumulation and should also carefully evaluate the presence of fixed production cost and the endogenous exit margin.

Our model has no firm-level productivity shocks. Therefore, the endogenous exit of firms only takes place at time of their entrance into the market. Adding firm-level productivity shocks into the current model would allow us to study the effect of the selection margin in a more realistic environment. Further, our model features a rather restrictive production setting: a homogeneous good production with the same returns to scale. Extending the model to allow for non-homogeneous goods and non-uniform returns to scale production can be an interesting future research topic.\footnote{This point is also made by Hopenhayn (2014b).}
Appendix

2.A Algebra

2.A.1 Aggregation in the undistorted economy

\[ Y = \int y(z) d\mu(z) \]
\[ = \frac{E}{\lambda} \int_{z \ge \hat{z}} y(z) dG(z) \]
\[ = \frac{E}{\lambda} \theta_\pi R^{-\alpha} w^{-(1-\alpha)l} \hat{M} \]
\[ = \frac{E}{\lambda} \theta_\pi \left( \frac{\lambda R K_u}{\theta_k EM} \right)^\alpha \left( \frac{\lambda w L}{\theta_l EM} \right)^{1-\alpha} \hat{M} \]
\[ = \theta_\pi \left( \frac{R}{\theta_k} \right)^\alpha \left( \frac{w}{\theta_l} \right)^{1-\alpha} K^\alpha L^{1-\alpha} \]
\[ = \theta_\pi \theta_k^{-\alpha} \theta_l^{1-\alpha} \left[ \frac{(1-\gamma)\theta_k \hat{M}}{(1-\rho)c + Mcf} \right]^{1-\gamma} K^\alpha L^{1-\alpha} \]
\[ = \left[ \frac{(1-\gamma)\mathbb{E}[z \ge \hat{z}]}{(1-\beta(1-\lambda)c/M + cf)} \right]^{1-\gamma} K^\alpha L^{1-\alpha} \]
The algebra to see that $Z$ can be written as $Z = \left( \frac{EM}{\lambda} \mathbb{E}[z|z \geq \hat{z}] K^{-\alpha} L^{-(1-\alpha)} \right)^{1-\gamma}$ is similar.

$$Y = \int y(z) d\mu(z)$$
$$= \frac{E}{\lambda} \int_{z \geq \hat{z}} y(z) dG(z)$$
$$= \frac{E}{\lambda} \theta_{x} R^{-\frac{\alpha}{1-\gamma}} w^{-(1-\frac{\alpha}{1-\gamma})} M$$
$$= \frac{E}{\lambda} \theta_{x} \left( \frac{\lambda K_{u}}{\theta_{E} EM} \right)^{\frac{\alpha}{1-\gamma}} \left( \frac{\lambda L}{\theta_{L} EM} \right)^{(1-\frac{\alpha}{1-\gamma})} M$$

$$= \left( \frac{EM}{\lambda} \right)^{1-\gamma} \theta_{x} \theta_{k}^{-\alpha} \theta_{l}^{-(1-\frac{\alpha}{1-\gamma})} (K^{\alpha} L^{1-\alpha})^{\gamma}$$
$$= \left( \frac{EM}{\lambda} \mathbb{E}[z|z \geq \hat{z}] K^{-\alpha} L^{-(1-\alpha)} \right)^{1-\gamma} K^{\alpha} L^{1-\alpha}$$

2.A.2 Aggregate output in the distorted economy

The algebra is similar as the one above.

2.B Proofs

2.B.1 Sufficient conditions for $TFP_r < 1$ in the case of $c_f = 0$

In this proof, we write distortions and firm-level productivity in capital letter to indicate they are random variables. The relative TFP when distortions and productivity are dependent is given by $TFP_r = \left[ \frac{E(Z)}{E(Z)} \right]^{1-\gamma} \frac{E(TZ)}{E(T)}$. To see that $\frac{E(TZ)}{E(Z)} \leq 1$ is a sufficient condition for $TFP_r < 1$, first re-write the expression of relative TFP as $TFP_r = \left[ \frac{E(TZ)}{E(Z)} \right]^{(1-\gamma)^2} \frac{E(TZ)}{E(TZ)/E(Z)^{1-\gamma}}$. Then, for $\gamma \in (0, 1)$, we have $\left( \frac{E(TZ)}{E(Z)} \right)^{(1-\gamma)^2} \leq 1$ (by the sufficient condition) and $\frac{E(TZ)}{E(TZ)/E(Z)} < 1$ (by the argument in the proof of proposition 2.1).

When $\gamma \in (0.5, 1)$, we obtain a weaker sufficient condition $\frac{E(TZ)}{E(Z)} \leq 1$. Indeed, if $\frac{E(TZ)}{E(Z)} \leq 1$, $TFP_r < 1$ by the first sufficient condition. If $\frac{E(TZ)}{E(Z)} > 1$, re-write the expression of relative TFP as $TFP_r = \left[ \frac{E(TZ)}{E(Z)} \right]^{\frac{1}{\gamma} - 2} \frac{E(TZ)}{E(Z)}$, and note that $\left[ \frac{E(TZ)}{E(Z)} \right]^{\frac{1}{\gamma} - 2} < 1$ since $\gamma \in (0.5, 1)$. This sufficient condition is weaker because $\frac{E(TZ)}{E(Z)} < \left( \frac{E(TZ)}{E(Z)} \right)^{\gamma}$ so $\frac{E(TZ)}{E(Z)} \leq 1$ implies that $\frac{E(TZ)}{E(Z)} \leq 1$. The two sufficient conditions $E(T) \leq 1$ and $E(T) \leq 1$ for the case when distortions and productivity are independent follow from the same proof.

2.B.2 Sufficient conditions for $TFP_r < 1$ in the case of $c_f > 0$ and $G_z(z)$ Pareto

Since $\frac{1-1+\gamma}{\eta} \in (0, 1), \mathbb{E}(t^{\eta-1+\gamma}) = \mathbb{E}(t^{\eta}) \frac{\eta^{\eta-1+\gamma}}{\eta^{\eta}} \leq \mathbb{E}(t^{\eta}) \frac{\eta^{1+\gamma}}{\eta^{\eta}}$ by Jensen’s Inequality. Thus, $TFP_r = \mathbb{E}(t^{\eta}) \frac{1}{\eta^{\eta}} \mathbb{E}(t^{\eta-1+\gamma}) \leq \mathbb{E}(t^{\eta}) \frac{\eta^{1+\gamma}}{\eta^{\eta}}$. Because $\frac{1}{\eta^{\eta}} \geq \frac{\eta^{1+\gamma}}{\eta^{\eta}}$, a sufficient condition for $TFP_r < 1$ is that $\mathbb{E}(t^{\eta}) \leq 1$. 


Bibliography


URL: http://ideas.repec.org/h/nbr/nberch/8244.html


URL: http://www.aeaweb.org/articles.php?doi=10.1257/aer.103.1.305


URL: http://ideas.repec.org/h/eee/grochp/1-09.html


URL: http://EconPapers.repec.org/RePEc:ed:se012:1103


URL: http://ideas.repec.org/a/red/issued/07-73.html


URL: http://qje.oxfordjournals.org/content/114/1/83.abstract


URL: http://ideas.repec.org/a/eee/emetrp/v60y1992i5p1127-50.html


URL: http://www.nber.org/papers/w20404


URL: http://ideas.repec.org/a/ucp/jpolec/v101y1993i5p915-38.html


URL: http://qje.oxfordjournals.org/content/124/4/1403.abstract
URL: http://ideas.repec.org/h/nbr/nberch/11037.html

URL: http://www.jstor.org/stable/3003596


URL: http://ideas.repec.org/a/red/issued/13-0.html


Chapter 3

Financial Frictions and Resource Misallocation in China’s Manufacturing Sector

3.1 Introduction

Differences in total factor productivity (TFP) is one of the main reasons behind the large gaps in income per worker between rich and poor countries.¹ Seeking what causes the TFP differences, a recent strand of literature highlights the role of resource misallocation.² Restuccia and Rogerson (2008) show that idiosyncratic distortions at firm level can lead to an inefficient allocation of factors across firms, which in turn results in sizable aggregate TFP losses. Hsieh and Klenow (2009) present empirical evidence that such misallocation is substantial in China’s manufacturing sector. They calculate that China’s manufacturing TFP can improve by 30%-50% when resources are hypothetically reallocated efficiently to the extent of the US level. In their analysis, however, instead of identifying any specific frictions behind the misallocation, they model frictions as generic abstract “wedges” between marginal products of factors and their prices. Therefore, what remains an open question is: What are the underlying frictions that generate the misallocation in China’s manufacturing sector? It is an important question as identifying and understanding the frictions is a necessary first step for potential policy remedies.

This paper attempts to address this question by proposing financial frictions as a source of misallocation in China’s manufacturing sector. We focus on financial frictions because evidence suggests that it is present and it may be an acute problem. In particular, we emphasize two aspects of the financial frictions in the context of the credit market in China. The first one is credit discrimination, a phenomenon specific to China. Banks, most of which are state-owned, favor state-owned enterprises (SOEs) when giving out loans. As a result, SOEs have easy access to credit while non-SOEs are often credit constrained. The credit discrimination together with the fact that SOEs are on average much less productive than non-SOEs creates a source of misallocation between the two types of firms. The second aspect of the financial frictions is the usual one studied in the literature. Due to underdeveloped

¹ See, for example, Caselli (2005); Hall and Jones (1999) and Klenow and Rodríguez-Clare (1997).
² See Restuccia and Rogerson (2013) and Hopenhayn (2014) for extensive reviews on this literature.
financial markets, firms, in our case particularly non-SOEs, face a tight collateral constraint: they can only borrow up to a small fraction of their net worth. Consequently, productive but low net worth firms are unable to borrow optimally to finance production. This aspect of the financial frictions therefore leads to misallocation across heterogeneous firms.

We start our analysis by presenting evidence on financial frictions in China. Since misallocation between SOEs and non-SOEs is a major consequence of financial frictions, we first briefly overview the past and current status of the two types of firms. The main observation is that since China started its economic reform in 1978, relative to non-SOEs, SOEs have become smaller in aggregate but larger on average. We view this observation as a sign of possible resource misallocation. We then present three pieces of evidence that motivate our study, namely, credit discrimination, productivity gap between SOEs and non-SOEs, and underdeveloped financial market with limited access to credit.

Next, we develop a model of financial frictions by extending a static version of Moll (2010). Our model features a representative SOE and a continuum of non-SOEs operated by entrepreneurs who are heterogeneous in wealth and productivity. The SOE is not financially constrained, but the entrepreneurs face a collateral constraint: they can only borrow up to a fraction of their wealth. Under this setting, financial frictions lead to resource misallocation between the SOE and non-SOEs as well as across non-SOEs. To better isolate the effect of financial frictions on misallocation between SOEs and non-SOEs, we also introduce two wedges on prices of capital and labor faced by non-SOEs. These wedges capture, in an abstract way, other frictions that also distort the allocations between the two types of firms so that we can quantitatively assess the relative impact of financial frictions against other frictions that are not explicitly modelled. Admittedly, the two wedges cannot capture potential firm-specific distortions that can cause misallocation across non-SOEs. We do not go further to model idiosyncratic distortions at firm level so as to keep the model simple.

The model we build is tractable in aggregate, allowing us to first study analytically the aggregate features of the model. We then take the model to a microdata set on manufacturing firms in China. Under our calibration and counter-factual experiments, we find that a potential 24% TFP gain can be achieved if the credit constraint faced by non-SOEs is at a reasonable level similar to the one in the US. Moreover, 53% of the gain can be attributed to improved allocations between SOEs and non-SOEs. By comparing these numbers to the TFP gains obtained when removing the capital and labor wedges, we argue that financial frictions are likely the main reason behind the resource misallocation in China’s manufacturing sector.

Related literature This paper contributes to the literature that studies the impact of resource misallocation on aggregate TFP (see, for example, Restuccia and Rogerson, 2008, Hsieh and Klenow, 2009, Alfaro et al., 2008 and Guner et al., 2008). In particular, as already mentioned, our work directly complements Hsieh and Klenow (2009) as we focus on a specific kind of frictions that generates the misallocation in China’s manufacturing sector studied in their paper.

This paper also complements Brandt and Zhu (2010) and Brandt et al. (2013), both of which find significant capital misallocation between SOEs and non-SOEs in China’s non-agriculture sector, but neither of them model any specific frictions. Brandt and Zhu (2010) study the sources of China’s rapid economic growth in the past three decades. In addition to their main finding that TFP growth in non-state and non-agriculture sector is the main driver of the growth, they also find strong evidence of capital misallocation: a significant portion of fixed investment goes to the less efficient state sector.
They quantify that if capital had been allocated efficiently, China could have achieved the same growth performance without any increase in the rate of aggregate investment, which reached 40% of GDP in 2007 from only 21% in 1978. Brandt et al. (2013) measure TFP loss due to resource misallocation across sectors and provinces in China’s non-agriculture sector between 1985 and 2007. They find that factor misallocation reduced aggregate TFP by about 30% on average. In addition, they find that TFP loss initially declined but started to increase since the mid-1990s. They argue that the increase can be mostly attributed to capital misallocation between state and non-state sectors within provinces, and that such capital market distortions are related to government policies that encourage investments in the state sector at the expense of investments in the more productive non-state sector. In our paper, misallocation between SOEs and non-SOEs is a direct consequence of financial frictions. Our findings on TFP loss in the manufacturing due to financial frictions may suggest that they are also a main source of misallocation behind their findings in the non-agriculture sector.

This paper is also related to a large body of literature on financial frictions, capital misallocation and aggregate TFP loss (for example, see Jeong and Townsend, 2007, Buera et al., 2011, Moll, 2010 and Midrigan and Xu, 2010). We contribute to this literature by focusing on the case of China’s manufacturing sector. We also emphasize an aspect of the financial frictions that is specific to China, namely, credit discrimination.

The rest of the paper is organized as follows. After motivating our analysis by documenting key empirical evidence in section 2, we build a model and present our analytical findings in section 3. In section 4, we calibrate the model using a firm-level data set from China’s manufacturing sector and we report our quantitative findings. Section 5 concludes.

### 3.2 Empirical evidence

In this section, we first briefly overview the background of SOEs and non-SOEs in China as misallocation between the two types of firms is one of the main consequences of financial frictions emphasized in this paper. We then document three pieces of key empirical evidence that motivates our study. First, credit availability is not uniform across firms: SOEs have a much better access to credit market than do non-SOEs, i.e. credit discrimination. Second, SOEs are on average less productive than non-SOEs. Third, the financial system in China, particularly its credit system, is still underdeveloped despite undergoing substantial reforms. We draw most of the evidence from the literature, but we also complement it with indirect evidence obtained from our data. We report a large discrepancy in capital-labor ratios between SOEs and non-SOEs, a possible sign of financial frictions.

#### 3.2.1 Background of SOEs and non-SOEs

Before China started its reform in 1978, it has a centrally planned economy, in which SOEs played a dominant role, especially in industrial sector. Industrial output of SOEs accounted for 80.8% of the total industrial output in 1978 (Lin, 1998). However, since the reform, China has been gradually switching to a market-oriented economy and the presence of SOEs has been steadily declining. The expansion of non-SOE sector accelerated after the government officially endorsed the “socialist market economy” in late 1992. By the end of 2008, according to the Second National Economic Census conducted in that year, there were 1,903,000 enterprises in total in industrial sector, and among them SOEs only account
for 1.5%, about 28,000. Moreover, of 117.38 million employment in industrial enterprises, only 9.2% were employed by SOEs.

This paper concerns the manufacturing sector in China.\(^3\) Using a firm-level data set, the Chinese Industrial Enterprise Database (1998-2007), we confirm the substantial aggregate decline of SOEs.\(^4\) Capital, employment and value added shares of SOEs in manufacturing sector declined substantially over the sample years, from 53%, 46% and 35% in 1998 to just 14%, 7% and 7% in 2007. These measures indicate that in aggregate SOEs gradually became a smaller part of the manufacturing sector relative to non-SOEs.

Despite the decline of SOEs in aggregate, their average size is still much larger than non-SOEs. From 1998 to 2004, average employment of SOEs are about twice that of non-SOEs. Since 2005, the size gap has widened. By 2007, the last year we have data, SOEs are on average 3.6 times the size of non-SOEs.

The contrast between aggregate and average size of SOEs and non-SOEs is an interesting phenomenon and it may be a sign of serious resource misallocation. Next, we present three pieces of evidence that suggest financial frictions may be the main reason behind the phenomenon.

### 3.2.2 Credit discrimination

Allen et al. (2012) provides a comprehensive overview of China’s financial system. Among many of their findings, they show that even with the entrance and growth of many domestic and foreign banks and financial institutions in recent years, China’s banking system is still mainly controlled by the four largest state-owned banks. Possibly due to government policy or better connection between SOEs and state-owned banks, SOEs tends to have much easier access to credit market than do non-SOEs. Huang (2004) argues that allocation of credit in China has a “political pecking order”. SOEs are considered politically important than non-SOEs and hence are often given advantage over non-SOEs in accessing to credit. Similarly, Boyreau-Debray and Wei (2005) argue that the state-dominated financial system severely retards the efficient allocation of capital. The system tends to allocate capital systematically away from more productive regions toward less productive ones to favor inefficient SOEs.

In a recent empirical study, using a firm-level data set from China over the period 1998-2005, Poncet et al. (2010) present evidence that private Chinese firms are credit constrained while SOEs and foreign-owned firms are not. Moreover, they show that geographical and sectoral presence of SOEs aggravates financial constraints for private Chinese firms.\(^5\) Similarly, using a firm-level data set from 2000 to 2007, Guariglia et al. (2011) find that non-SOEs are discriminated against by financial institutions. As a result, SOEs finance their investments more through bank loans than do non-SOEs.

In the macroeconomic literature, Song et al. (2011) report that more than 30% of SOEs’ investments are financed through bank loans comparing to less than 10% for non-SOEs. Dollar and Wei (2007) study returns of capital in Chinese firms. They show similar evidence that non-SOEs rely less on bank loans but more on retained earnings.

---

\(^3\) In this paper, manufacturing sector consists of all industrial sectors excluding mining and utilities.

\(^4\) See section 4 for a detailed description of the data and how variables are constructed.

\(^5\) In this paper, non-SOEs include both private Chinese firms and foreign firms. Thus, the credit constraint faced by the non-SOEs discussed in this paper can be seen as an average constraint faced by firms of different ownership within non-SOEs.
3.2.3 Productivity gap between SOEs and non-SOEs

If SOEs are more productive than non-SOEs, credit discrimination may not necessarily lead to large resource misallocation between the two types of firms. However, numerous studies find that SOEs are on average much less productive than non-SOEs. For example, in a comprehensive study of firm-level TFP in China’s manufacturing sector between 1998 and 2007, Brandt et al. (2011) show that SOEs were on average 27% less productive than non-SOEs and their TFP growth rate was 4.6% lower. Hsieh and Klenow (2009) also find that SOEs exhibit 41% lower revenue TFP than non-SOEs in the manufacturing sector. In non-agriculture sector, Brandt and Zhu (2010) show that both TFP and TFP growth rate of non-SOEs are significantly higher than those of SOEs.

3.2.4 Underdeveloped financial system

Besides resource misallocation between SOEs and non-SOEs, financial frictions can also lead to misallocation across non-SOEs. Because credit available to non-SOEs are often limited and depends on the collateral they can provide, productive but low net worth non-SOEs often do not have enough collateral to borrow optimally. Although China’s financial system has undergone significant reforms since early 1990s, it still lags far behind developed economies. This paper concerns the credit market. To get a general picture on how well the credit market functions in China relative to advanced economies such as the US, we compare two credit market indexes between China and the US. The first index is domestic credit to private sector as percentage of GDP, and the second one is domestic credit provided by financial sector as percentage of GDP. Both indexes are from the Global Financial Development Database (GFDD) maintained by the World Bank. Figure 3.1 plots the indexes since 1990. We see that China is at about two thirds of the US level for both indexes.

Figure 3.1: Two Credit Market Indexes

---

See Cihak et al. (2012) for details of data construction.
3.2.5 \( K/L \) discrepancy between SOEs and non-SOEs

Financial frictions can cause factors to be disproportionately allocated across firms with heterogeneous productivity.\(^7\) Given the productivity gap between SOEs and non-SOEs, we can look for indirect evidence of financial frictions by comparing the aggregate capital-labor ratio between the two types of firms. Using our firm-level data set, in figure 3.2 we plot the ratio of capital-labor ratios between SOEs and non-SOEs from 1998 to 2007. We see a large discrepancy between capital-labor ratios of the two types of firms. Not only the capital-labor ratio of SOEs is consistently larger than that of non-SOEs, but the gap also widened over the years. By 2007, the last year of the sample, the capital-labor ratio of SOEs is about twice larger than that of non-SOEs.

3.3 The model

In this section, we develop a model of financial frictions. Our model builds on the static version of Moll (2010), who investigates whether self-finance can undo capital misallocation caused by financial frictions. We first provide an overview of our economic environment, followed by a detailed problem set-up. We then define the competitive equilibrium. Last, we characterize the equilibrium, and derive and discuss aggregate variables of interest.

3.3.1 Model Setup

A final good is produced by two types of firms: SOEs and non-SOEs. We model SOEs as a representative firm. Its ownership is inessential under our setting as it makes zero profit in equilibrium. On the other hand, non-SOEs are owned by a continuum of entrepreneurs, who are heterogeneous in their wealth and productivity. An entrepreneur chooses to operate a non-SOE if it’s profitable, or otherwise becomes a pure lender. All firms produce under a constant returns to scale (CRS) technology with capital and labor as inputs. Capital is financed through an imperfect financial market. First, the market favors the

---

\(^7\)In our model of financial frictions, where firms are heterogeneous in productivity and produce under constant returns to scale technology (with the same factor shares), a less productive firm has a higher capital-labor ratio.
representative SOE and discriminates against entrepreneurs running non-SOEs. In particular, the SOE faces no borrowing constraint when taking out loans, but entrepreneurs can only borrow up to a fraction of their wealth, which acts as their collateral. Second, due to financial underdevelopment, the collateral constraint faced by entrepreneurs is relatively tight comparing to those in advanced financial markets.

In addition to the financial frictions, we introduce two abstract wedges on the prices of capital and labor faced by non-SOEs to capture other frictions that also distort allocations between the SOE and non-SOEs. This approach allows us to better isolate the impact of financial frictions on misallocation between the two types of firms when performing quantitative exercises. We do not model firm-specific frictions that can cause misallocation within SOEs for simplicity. Workers in this economy are homogeneous and supply their labor inelastically. In the following, we set up the model in details.

**SOEs** The representative SOE produces output $Y_s$ using capital $K_s$ and labor $L_s$. The production technology is constant returns to scale, $(Z_s K_s)^{\alpha} L_s^{1-\alpha}$, where $0 < \alpha < 1$ is the capital share and $Z_s$ is the productivity. The SOE maximizes profit taking as given output price (normalized to be 1), wage $w$ and capital rental rate $R$. It therefore solves the following maximization problem.

$$\max_{K_s, L_s \geq 0} (Z_s K_s)^{\alpha} L_s^{1-\alpha} - w L_s - RK_s \quad (3.1)$$

**Entrepreneurs and non-SOEs** There is a continuum of entrepreneurs with a total mass normalized to 1. They are heterogeneous in their wealth $a$ and productivity $z$. Denote $G(a, z)$ the joint distribution of their wealth and productivity with positive support, and $F(z)$ the marginal distribution of productivity. To avoid uninteresting cases, we assume that SOEs’ productivity $Z_s$ is in the interior of the support of $F(z)$. That is, some entrepreneurs are more productive than SOEs, but others are not. For later use, denote $f(z)$ the density of $F(z)$.

An entrepreneur chooses whether to stay inactive or to operate a non-SOE. There is no entry cost to operate a firm. The production technology is CRS with the same capital share as that of the representative SOE, $(zk)^{\alpha} L^{1-\alpha}$. Entrepreneurs face a collateral constraint when borrowing to finance capital input. They can borrow up to $\lambda a$ of their wealth, where $\lambda \geq 1$ is an exogenous parameter that captures the imperfection of the credit market. In a perfect market, $\lambda = \infty$, and in the absent of credit market, $\lambda = 1$. The use of $\lambda$ to capture the credit market imperfection can be derived from a model where contracts between lenders and borrowers cannot be perfectly enforced. In our setting, in addition to imperfect enforcement, it also reflects the credit discrimination faced by non-SOEs. While using their wealth as collateral, entrepreneurs can also deposit their wealth in financial intermediaries and earn a rate of return $r$. Those entrepreneurs that don’t operate firms become pure lenders of their wealth.

In addition to the financial frictions, operating entrepreneurs pay different prices for labor and capital from those paid by the SOE due to unspecified distortions. Specifically, they hire workers at a wage rate $(1 + \tau_l)w$, and rent capital at a rental price $(1 + \tau_k)R$. The price wedges $(\tau_l, \tau_k)$ are modeled as labor and capital taxes, and the revenue collected will be returned to workers (not entrepreneurs) as a lump-sum subsidy to be specified below.$^8$ The wedges may take negative values, in which case they are subsidies, and are financed through lump-sum taxes on workers. Modeling the wedges as taxes or subsidies is for

---

$^8$Whether the tax revenue is returned to workers or entrepreneurs is not essential in our static model. In a dynamic extension of the current model, which I will briefly discuss at the end of the paper, the assumption is needed to keep the model tractable.
exploratory convenience. They capture distortions other than the financial frictions explicitly modeled, and they are not necessary taxes or subsidies in the real world.

Given the above setting, an entrepreneur with wealth and productivity pair \((a, z)\) solves the following profit maximization problem.

\[
\pi(a, z) = \max_{k, l \geq 0} \left\{ (zk)^{\alpha}l^{1-\alpha} - (1 + \tau_k)Rk - (1 + \tau_l)wl \quad \text{s.t.} \quad k \leq \lambda a \right\}
\]  

(3.2)

Note that the entrepreneur can choose to stay inactive, in which case the optimal inputs are \(k = 0\) and \(l = 0\). Consumption of the entrepreneur is simply \(c(a, z) = \pi(a, z) + (1 + r)a\). Under our static setting, preferences of entrepreneurs are inessential, and hence they are left unspecified.

**Workers**  
The total mass of worker is \(L\). Each worker is endowed with 1 unit of time, which he supplies inelastically. Workers earn wages and consume their earnings. Again, preferences of workers are inessential. A worker’s consumption is simply \(c^w = w + s\), where \(s\) is the lump sum subsidy or tax depending on its sign.

**Credit and capital rental market**  
Financial intermediaries receive deposits and provide loans to entrepreneurs. Under the assumption of a competitive financial market and zero transaction cost, both the deposit and loan have the same rate of return \(r\). Entrepreneurs rent capital at a capital market with the rate \(R\). The market is also assumed to be competitive. Hence, zero profit condition implies that \(R = r + \delta\), where \(\delta\) is the depreciation rate of capital. As already described above, the financial intermediaries treats SOEs and entrepreneurs/non-SOEs differently. Borrowing by entrepreneurs are collateral constrained, but no such constraint is imposed on SOEs.

The formulation of our problem may seem rather simple without explicitly spelling out who owns the capital. Moll (2010) shows that the current setting is equivalent to a model where entrepreneurs own capital and trade risk free bonds. Therefore, our formulation is for exploratory convenience.

### 3.3.2 Definition of equilibrium

A **competitive equilibrium** consists of prices \(\{w, R\}\) and allocations \(Y, \{K_s, L_s\}, \{k(a, z), l(a, z), c(a, z)\}\), \(c^w\), and \(s\) such that

1. Given \(\{w, R\}\), the representative SOE solves 3.1. \(\{K_s, L_s\}\) are the optimal capital and labor allocations.

2. Given \(\{w, R\}\), an entrepreneur with wealth and productivity pair \((a, z)\) solves 3.2. \(\{k(a, z), l(a, z)\}\) are the optimal factor inputs and \(c(a, z) = \pi(a, z) + (1 + r)a\) is the optimal consumption.

3. Workers consume their earnings and the lump sum tax or subsidy, \(c^w = w + s\).

4. Capital market makes zero profit: \(R = r + \delta\).

5. Government balance budget: \(sL = \tau_k \int k(a, z)dG(a, z) + \tau_l \int l(a, z)dG(a, z)\)

6. Markets clear:

   (a) capital

   \[
   \int k(a, z)dG(a, z) + K_s = \int adG(a, z) = K
   \]
(b) labor
\[ \int l(a, z)dG(a, z) + L_s = L \]

### 3.3.3 Equilibrium characterization

Depending on parameter values, two kinds of equilibria are possible: one in which the representative SOE operates, and the other it doesn’t. Both equilibria are solved in a similar way. Under the second equilibrium, the model collapses to the static version of Moll (2010).

**SOEs**  
Assuming the representative SOE operates in equilibrium, its maximization problem 3.1 implies the following FOCs,

\[ R = \alpha Z_s K_s^{\alpha-1} L_s^{1-\alpha} \quad (3.3) \]
\[ w = (1 - \alpha)(Z_s K_s)^\alpha L_s^{-\alpha} \quad (3.4) \]

**Non-SOEs / Entrepreneurs**  
We start with the profit maximization problem 3.2. For a fixed capital \( k \), optimal labor is

\[ l(a, z) = \left[ \frac{1 - \alpha}{(1 + \tau_l)w} \right]^{\frac{\alpha}{1-\alpha}} z^{\lambda a} \quad (3.5) \]

Substituting it back to the profit function, we have

\[ \pi(a, z) = \max_k \left\{ z^{\alpha} \left[ \frac{1 - \alpha}{(1 + \tau_l)w} \right]^{\frac{1-\alpha}{1-\alpha}} - (1 + \tau_k)R \right\} \text{ s.t. } k \leq \lambda a \quad (3.6) \]

The expression suggests that the decision on whether to operate a non-SOE follows a cutoff strategy. Moreover, if an entrepreneur decides to operate a non-SOE, the linearity of profit function in \( k \) implies that the optimal capital is at corner, i.e., \( k = \lambda a \). We summarize these observations in the following lemma.

**Lemma 3.1.** There exists a cutoff productivity \( z_c \) such that an entrepreneur chooses to operate a non-SOE if \( z > z_c \). An operating entrepreneur’s factor demands and profit are linear in wealth:

\[ k(a, z) = \lambda a, \quad l(a, z) = \left[ \frac{1 - \alpha}{(1 + \tau_l)w} \right]^{\frac{\alpha}{1-\alpha}} z^{\lambda a} \]
\[ \pi(a, z) = z^{\alpha} \left[ \frac{1 - \alpha}{(1 + \tau_l)w} \right]^{\frac{1-\alpha}{1-\alpha}} - (1 + \tau)\tilde{R} \lambda a \]

If an entrepreneur chooses to stay inactive, his factor demands and profit are all zero: \( k(a, z) = 0 \), \( l(a, z) = 0 \), and \( \pi(a, z) = 0 \).

The linearity of the factor demands and profit function are a result of the CRS technology. It is one of the important features of the model that delivers the analytical tractability. Due to this linearity, if we remove the credit constraint (\( \lambda = \infty \)), only the most productive entrepreneurs will be active.\(^9\) In particular, SOEs cannot survive if the credit constraint is removed.

\(^9\)One can think of the equilibrium with \( \lambda = \infty \) as the limit of the model economy when \( \lambda \to \infty \). For such equilibrium to exist, the support of \( z \) must be finite, and \( \max\{z\} \) must exist and have positive mass.
Chapter 3. Financial Frictions and Resource Misallocation

The cutoff productivity \( z_c \) depends on whether the representative SOE operates in equilibrium. In the case that it does, equations 3.3 and 3.4 (the SOE’s FOCs) imply that

\[
R = Z_s \alpha \left( \frac{1 - \alpha}{w} \right)^{\frac{1-\alpha}{\alpha}}.
\]

Using the above to replace the term \( \alpha \left( \frac{1 - \alpha}{w} \right)^{\frac{1-\alpha}{\alpha}} \) in the profit function (see lemma 3.1), we obtain

\[
\pi(a, z) = R \left[ \frac{z}{(1 + \tau_l)^{\frac{1-\alpha}{\alpha} Z_s}} - (1 + \tau_k) \right] \lambda a.
\]

An entrepreneur chooses to operate if he makes a positive profit. Thus, the cutoff productivity in this case is

\[
z_c = (1 + \tau_l)^{\frac{1-\alpha}{\alpha}} (1 + \tau_k) Z_s. \quad (3.7)
\]

In the case that the representative SOE does not operate in equilibrium, the cutoff \( z_c \) can be obtained using the market clear conditions to be discussed next.

**Market clearing** To simplify some of the coming expressions, let’s first define a few notations. Define the share of wealth held by entrepreneurs with productivity \( z \) by

\[
\omega(z) \equiv \frac{1}{K} \int adG(a, z).
\]

As in Moll (2010), the \( \omega(z) \) will play a role of a density. Hence, define the corresponding cumulative distribution by

\[
\Omega(z) \equiv \int_0^z \omega(z) d(z).
\]

Note that if wealth and productivity are independent, we have \( \omega(z) = f(z) \) and \( \Omega(z) = F(z) \). That is, the wealth share of entrepreneurs with productivity \( z \) equals the mass of the entrepreneurs with that productivity. Let us denote the fraction of wealth owned by entrepreneurs who choose to operate in equilibrium by

\[
M(z_c) = 1 - \Omega(z_c).
\]

Now, consider the capital market clearing conditions. Lemma 3.1 implies that optimal capital demand for entrepreneurs who choose to operate is linear in their wealth and at the corner \( \lambda a \). Hence the total capital demanded by operating entrepreneurs is simply \( \lambda KM \), where \( KM \) is the total wealth held by operating entrepreneurs. The capital market clearing condition thus simplifies to

\[
\lambda KM(z_c) + K_s = K. \quad (3.8)
\]

If the representative SOE operates in equilibrium, \( z_c \) is given by equation 3.7, so the above expression pins down \( K_s \). On the other hand, if the representative SOE does not operate in equilibrium, then \( K_s = 0 \) and equation 3.8 instead pins down the cutoff \( z_c \).

Equation 3.8 also implies a necessary and sufficient condition for the SOE to operate in equilibrium: \( \lambda M(z_c) < 1 \). This condition is intuitive. If the financial market is well developed, i.e., the collateral

\[10\text{Recall that } F(z) \text{ and } f(z) \text{ are marginal distribution and density of } G(a, z) \text{ for productivity } z.\]
constraint is not too tight (λ is large), or if the fraction of wealth owned by operating entrepreneurs is large (M is large) due to either low productivity of SOE (Zs is small) or small capital or labor price wedges (τk or τl is small), then even with the credit discrimination, the SOE won’t be able to compete with the non-SOEs to stay in operation. The following lemma summarizes the above discussion.

**Lemma 3.2.** If and only if \( \lambda M (z_c) < 1 \) where \( z_c = (1 + \tau_l)^{\frac{1-\alpha}{\alpha}}(1 + \tau_k) Z_s \), the representative SOE operates in equilibrium and \( z_c \) is the cutoff productivity for entrepreneurs. Otherwise, only non-SOEs operate in equilibrium and the cutoff productivity is defined by \( \lambda M (z_c) = 1 \).

An entrepreneur’s decision on whether to operate is of particular interest because it determines how market frictions distort the extensive margin of non-SOEs. In the case that the representative SOE operates in equilibrium, the decision depends on the productivity of SOEs (Zs) and the labor and capital price wedges (τl, τk). Let’s focus on the empirically relevant case where \((1 + \tau_l)^{\frac{1-\alpha}{\alpha}}(1 + \tau_k) > 1 \). In this case, due to the credit discrimination and other unspecified frictions captured by the two wedges, only entrepreneurs more productive than the representative SOE choose to operate. Hence, it’s an endogenous result that in equilibrium non-SOEs are on average more productive than the SOE. In the case that the SOE doesn’t operate in equilibrium, the decision depends on how tight the collateral constraint is. A tighter collateral constraint (a smaller λ) implies a smaller \( z_c \), which in turn leads to lower aggregate productivity as we will see soon.

To characterize other equilibrium variables, we turn to labor market clearing condition. Again for notational convenience, let’s first define

\[
M' (z_c) = \int_{z_c}^{\infty} zd\Omega(z) .
\]

Note that \( \frac{M'}{M} = \frac{\int_{z_c}^{\infty} zd\Omega(z)}{1-\Omega(z_c)} \) is just the expected productivity of entrepreneurs conditional on they choose to operate. Denote the expected productivity by

\[
E_{\omega}[z|z > z_c] = \frac{M'}{M} = \frac{\int_{z_c}^{\infty}zd\Omega(z)}{1-\Omega(z_c)}. 
\]

Using the optimal labor expression from Lemma 3.1, we simplify the labor market clearing condition to

\[
\left[ \frac{1 - \alpha}{(1 + \tau_l)w} \right]^\frac{1}{2} \lambda K M' (z_c) + L_s = L \tag{3.9}
\]

If the representative SOE operates in equilibrium, the above equation together with equation 3.4 solve for \( w \) and \( L_s \). If only non-SOEs operate in equilibrium \((L_s = 0)\), the above equation solves for \( w \). In both cases, we can pin down \( R \) by using \( \pi(a, z_c) = 0 \) for any \( a \) (see lemma 3.1).

**Aggregate economy** This model delivers a tractable aggregate economy. We summarize it in the following proposition. The proof can be found in appendix 3.A.\(^{11}\)

**Proposition 3.1.** The aggregate production technology is

\[
Y = Z K^\alpha L^{1-\alpha} \tag{3.10}
\]

\(^{11}\)In the appendix, I also compare factor prices with their corresponding aggregate marginal products.
where $K$ and $L$ are aggregate capital and labor. Under the equilibrium in which the representative SOE operates (eqbm 1),

$$Z = \frac{(1 + \tau_l)^{-\frac{1-\alpha}{\alpha}} \lambda ME_\omega[z > z_c] + (1 - \lambda M)Z_s}{\left[(1 + \tau_l)^{-\frac{1}{\alpha}} \lambda ME_\omega[z > z_c] + (1 - \lambda M)Z_s\right]^{1-\alpha}}$$  

(3.11)

is the aggregate TFP. Under the equilibrium where only non-SOEs operate (eqbm 2), the TFP is

$$Z = \left(\lambda ME_\omega[z > z_c]\right)^{\alpha}.$$  

(3.12)

In terms of aggregate output, the economy can be represented by one with an aggregate CRS production function given by 3.10. The aggregate TFP is endogenously determined (see 3.11 and 3.12). Under eqbm 2, it’s simply the expected productivity of operating non-SOEs (to the power of $\alpha$). In other words, it’s a weighted average of the firm-level productivity with the weight being the conditional density. Under eqbm 1, the expression of TFP may look daunting. To take a closer look, let’s first assume $\tau_l = 0$, i.e. no wage on wage. In this case, the TFP becomes

$$Z = \left(\lambda ME_\omega[z > z_c] + (1 - \lambda M)Z_s\right)^{\alpha}. $$

It’s again a weighted average of firm-level productivity. The weight attached to the productivity of SOE is the share of total capital rented by SOEs, $1 - \lambda M$ (see equation 3.8). Similarly, the weight attached to the average productivity of non-SOEs ($E_\omega[z > z_c]$) is the share of total capital rented by non-SOEs, $\lambda M$. Now, if $\tau_l \neq 0$, Z is still a weighted average in disguise with the weight on the SOE adjusted by $\tau_l$ on both the numerator and denominator.

Under eqbm 2, since $z_c$ increases in $\lambda$, so does aggregate TFP. Therefore, worse financial frictions (smaller $\lambda$) lead to larger aggregate TFP loss. Note that under eqbm 2 the capital and labor wedges doesn’t affect the aggregate production of the economy. This is intuitive. Because the wedges only distort the allocations between the SOE and non-SOEs, they have no effect on real production allocations across non-SOEs when the SOE does not operate in equilibrium. Instead, prices ($r$ and $w$) adjust accordingly to undo the effect of the wedges. Under eqbm 1, the comparative statics on $\lambda$, $\tau_k$, and $\tau_l$ are ambiguous.

**Capital-labor ratio** We presented empirical evidence on capital-labor ratio discrepancy between SOEs and non-SOEs. Here, we solve an expression for the ratio of capital-labor ratio between the two types of firm. Details of algebra are again left to appendix 3.A. We can show that

$$\left(\frac{K_s}{L_s}\right) / \left(\frac{K_n}{L_n}\right) \approx (1 + \tau_l)^{-\frac{1}{\alpha}} \frac{E_\omega[z > z_c]}{Z_s}.$$  

The ratio of capital-labor ratio between the SOE and non-SOEs is just their relative productivity adjusted by the labor wedge. This expression is as expected since

$$\frac{K_s}{L_s} \propto \frac{1}{Z_s}, \quad \frac{k(a,z)}{l(a,z)} \propto \frac{1}{z(1 + \tau_l)^{\frac{1}{\alpha}}}$$  

from equation 3.4 and 3.5. That is, capital-labor ratio is proportional to the inverse of firm-level productivity for SOEs as well as for non-SOEs adjusted by the labor wedge.
3.4 Quantitative Analysis

In this section, we take the model to the data to quantitatively assess the effect of financial friction on aggregate TFP. We first describe the data set. We then discuss our calibration strategy. Last, we present the result of our counter-factual exercises.

3.4.1 Data

The main data used in this paper is extracted from Chinese Industrial Enterprise Database (1998 - 2007), which is based on Annual Surveys of Industrial Production conducted by National Bureau of Statistics of China (NBS). The survey covers non-SOEs with more than 5 million RMB annual sales (about $600,000) and all SOEs. Firms included in the survey account for 90% of total industrial value added. The consistency of this data set has been extensively checked by Brandt et al. (2011). They show that aggregate statistics derived from this firm-level data set are largely consistent with the ones reported in Chinese Statistical Yearbooks. (See table 1 in Brandt et al. (2011) for summary statistics of the data.)

For the purpose of our study, we only use data in the manufacturing sector. In addition, since the model is static, we choose the 2004 data for model calibration. Because China conducted its economic census in 2004, choosing the 2004 data allows us to double check the aggregate consistency of our data with the 2004 China Economic Census Yearbook and to compare the result with what's reported in Brandt et al. (2011). We find no obvious inconsistency in our data. The 2004 data also provide slightly better accounting information than other years, which benefits our calibration exercise.

The main variables we extract from the data set are annual measures of output, intermediate inputs, value added tax, capital stock at original purchased prices, employment, industry code, ownership, total debt, accounts payable (and other short-term payable accounting entries) and owner's equity. To construct real value added, we adopt the procedure developed by Brandt et al. (2011). Output and intermediate inputs are deflated separately using output and input deflators constructed by Brandt et al. (2011). Value added tax is deflated using the average of input and output deflators. Real value added is then obtained as output minus intermediate inputs plus value added tax. The construction of real capital stock series is difficult as we have data only on capital stock at original purchased prices. Moreover, the data set is cross sectional over years, but not a panel. That is, the same firm surveyed across years is not readily identified. Brandt et al. (2011) develop a procedure in which one first identifies and links the same firms in the data set across years, and then estimates their initial capital stocks, and finally uses a perpetual inventory method to obtain capital stocks in subsequent years. We completely follow their method.

We categorize a firm to be SOE if the state holds 50% or more shares of its registered capital. Another common approach in the literature to categorize SOEs is to use the ownership information from the firm's registration. Hsieh and Song (2015) show that the annual revenue shares of SOEs aggregated from the data are closer to the ones reported in the China Statistical Yearbooks when the shareholder approach is used. Hence, we adopt that approach.

12Data from 1998 to 2004 are also needed to construct firm-level real capital stocks.
3.4.2 Calibration

To quantify TFP loss, we need to calibrate the following parameters: capital share $\alpha$, depreciation rate $\delta$, productivity of the representative SOE $Z_s$, collateral constraint parameter $\lambda$, and labor and capital wedges $(\tau_l, \tau_k)$. In addition, we also need to specify the distribution $G(a, z)$ and calibrate its parameters. Note that $\delta$ is needed to construct firm-level real capital stocks although it does not enter the aggregate TFP expression.

The capital share $\alpha$ is obtained as one minus labor share of income in US manufacturing sector. We do not use labor share in China’s manufacturing sector because market distortions may cause endogeneity problem. According to the National Income and Product Accounts (NIPA), labor income share in US manufacturing in 2004 is about 0.6. Hence, we obtain $\alpha = 0.4$. We take $\delta = 0.09$ following Brandt et al. (2011).

We calibrate TFP of the representative SOE as $\hat{Z}_s = Y_s/K_s L_s^{1-\alpha}$, where $Y_s, K_s$ and $L_s$ are real value added, real capital stock and employment of SOEs aggregated from the data. Thus, productivity $Z_s$ as specified in our model is given by $(\hat{Z}_s)^{1/\alpha}$.

The model predicts that an operating entrepreneur borrows $(\lambda - 1)a$ and his debt to wealth ratio is thus $\lambda - 1$. To calibrate $\lambda$, we take loan to equity ratio as the corresponding ratio in the data. Since data on loan are not available in our microdata set, we use total debt net of short-term payable as a proxy.\footnote{Total loan is calculated as total debt minus accounts payable, wages and benefits payable, and taxes payable. These short-term payable accounting entries are the ones available in our data.}

For each non-SOE, we first calculate a firm-specific borrowing constraint using its loan to equity ratio. We then take the average of the firm-specific constraints to obtain $\lambda \approx 2.7$.

The wedges $(\tau_l, \tau_k)$ are calibrated to match $K_s/K$ and $L_s/L$ in the data. We obtain $\tau_l = 1.73$ and $\tau_k = 0.97$. Therefore, on top of the financial frictions explicitly modeled, our calibration suggests that non-SOEs face extra tax wedges on labor and capital.

Now, we need to specify the joint distribution $G(a, z)$. We first check whether $a$ and $z$ are independent. If it’s the case, we have the special case of $\Omega(z) = F(z)$ and it’s sufficient to just calibrate the distribution of $z$ for the interested aggregate measure of TFP. Thus, we first obtain TFP of each firm as $\hat{z} = y/k^\alpha l^{1-\alpha}$, where $y, k,$ and $l$ are real value added, real capital and employment taken from the data. The productivity $z$ as specified in the model is given by $z = \hat{z}^{1/\alpha}$. We then check for linear correlation between $z$ and firm’s equity and find it almost equal to zero ($\approx 0.02$). A scatter plot also shows no particular relation between the two. Moreover, since the model predicts that if $a$ and $z$ are correlated then a firm’s real capital stock should also correlate with its productivity (as $k = \lambda a$). We thus also check for the correlation between $z$ and $k$, and find it almost equal to zero ($\approx -0.003$). A scatter plot also confirms no particular dependence between the two. Given there is little evidence that $a$ and $z$ are dependent, we focus on calibrating the distribution of $z$.

For convenience, we work with the distribution $F(\hat{z})$ instead of $F(z)$. In particular, we assume that $F(\hat{z})$ follows a Gamma distribution, $F(\hat{z}) \sim Gamma(a, s)$, where $a$ and $s$ are the shape and scale parameters. The model thus predicts a left-truncated Gamma distribution for operating non-SOEs. Taking the model seriously, we should estimate the parameters by fitting a left-truncated Gamma. However, we don’t observe a clear truncation for $\hat{z}$ in the data. This could be due to measurement error as discussed in Moll (2010). Hence, one way to proceed is to model the measurement error and include it in the estimation. However, it is not clear what measurement error one should specify. To make progress,\footnote{We obtain $\hat{Z}_s = 9.7$, so $Z_s = 9.7^{(1/0.4)}$.}
we decide to ignore the truncation and simply fit $F(\hat{z})$ to the observed $\hat{z}$. If the measurement error has a low variance, the bias introduced by this compromise should be minimum. We obtain maximum likelihood estimates $a = 1.5$ and $s = 10.6$.

### 3.4.3 Results

In this paper, we propose financial frictions as a source of resource misallocation in China’s manufacturing sector. After calibrating the model, we are ready to evaluate the impact of financial frictions on TFP. We perform four counter-factual experiments. In the first two experiments, we measure potential TFP gain from reducing financial market frictions, and decompose the gain into two components, the part attributed to improved factor allocations between SOEs and non-SOEs and the part attributed to improved factor allocations across non-SOEs. In the last two experiments, we measure TFP gain from reducing capital wedge or labor wedge. We compare the TFP gains between the experiments to gauge the relative impact of financial frictions on resource misallocation. In the following, we present our results.

In the first experiment, we gradually relax financial constraint $\lambda$ while holding the capital and labor wedges unchanged, and we measure the percentage TFP gains (relative to the original calibrated TFP). Figure 3.3 presents the result. As we know from the analysis in the last section, the representative SOE may not survive in equilibrium if the financial constraint is not too tight. The dash line indicates the threshold value of $\lambda = 4.43$, above which the SOE doesn’t operate in equilibrium. Recall that once the SOE does not operate in equilibrium, the labor and capital wedges do not distort the aggregate production as factor prices adjust. For this reason, the TFP gain curve has a kink at $\lambda = 4.43$. The dotted line in the figure indicates the $\lambda$ value for the US. Using external finance to GDP ratio, Moll (2010) calibrates the $\lambda$ for US to be 4.15.

Our experiment shows that as $\lambda$ increases, TFP level continues to increase. The TFP gain curve is nearly piecewise linear, indicating that the speed of increase in TFP is almost constant with a slight drop at $\lambda = 4.43$. As $\lambda$ increases, up to 28% of the TFP gain can be achieved under an equilibrium

\[\text{Precisely, the TFP gain curve is strictly concave.}\]
where the representative SOE operates. However, further TFP gain requires a value of $\lambda$ under which the SOE cannot survive. When $\lambda$ is at the value comparable to the one in the US, TFP increases about 24%.

The TFP gain obtained from reducing finance frictions can be decomposed into two parts. One is the gain generated from better resource allocations between the SOE and non-SOEs, and the other is the gain from better allocations across non-SOEs. In the second experiment, we gradually increase $\lambda$ but adjust capital and labor wedges such that the capital and labor allocations for the SOE remain unchanged. We then calculate the TFP gain, which represents the part of the gain due to better allocations across non-SOEs.

Figure 3.4 panel 3.4a presents the result from the second experiment. We re-plot the TFP gain as shown in figure 3.3 together with the new TFP gain obtained when capital and labor wedges are adjusted accordingly. We see that roughly a half of the TFP gain can be attributed to better allocations across non-SOEs for any given increase in $\lambda$.

To emphasize the relative importance of resource reallocation between the representative SOE and non-SOEs as $\lambda$ increases, in panel 3.4b we plot the percentage of TFP gain that can be attributed to better allocations between the two types of firms, which is simply the difference between the TFP gain and TFP gain adjusted (see panel 3.4a) as a fraction of the TFP gain. We see that it increases in $\lambda$ if $\lambda$ is less than 4.43, the threshold value for the SOE to operate in equilibrium, and decreases in $\lambda$ otherwise. This property implies that under the equilibrium where the SOE operates, the fraction of TFP gain due to better allocations between the two types of firms increases as the financial market improves.\footnote{Overall, the increase is small though, from 47\% to 53\%.} However, under the equilibrium where the SOE does not operate, since reallocation across non-SOEs is the only source of TFP gain as the financial market further improves, the fraction of TFP gain due to better allocations between the two types of firms inevitably declines. Overall, as the financial market improves, the contribution to TFP gain coming from reallocations between the SOE and non-SOEs is roughly the same as the one coming from reallocations across non-SOEs. When $\lambda$ is at the level of the

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{figure3_4.png}
\caption{TFP Gain Decomposition}
\end{figure}
US, about 53% of TFP gain can be attributed to better allocations between the two types of firms.

To further evaluate the relative impact of financial frictions on TFP, we move to the next two experiments. In the third experiment, we measure TFP gain as we gradually reduce the labor wedge from its calibrate value to zero but hold financial frictions unchanged. Figure 3.5 panel 3.5a presents the result. For a similar reason as we increase the value of $\lambda$, the representative SOE may not survive in equilibrium for small value of $\tau_l$. The dash line indicates the threshold value of $\tau_l = 0.6$, below which the SOE does not operate in equilibrium. From the graph, we see that TFP increases as $\tau_l$ decreases till the threshold value. It then stays constant as $\tau_l$ further decreases to zero. This is because real allocation no more changes in $\tau_l$ when the SOE does not operate in equilibrium. Overall, when the labor wedge is removed, TFP rises about 10%, a number slightly less than half of the gain achieved in the first experiment when $\lambda$ increases to the US level.

Our last experiment is similar to the third one, but this time we gradually decrease $\tau_k$ from its calibrated value to zero. Figure 3.5 panel 3.5b presents the result. Unlike the third experiment, the representative SOE stays in equilibrium for all values of $\tau_k$ we experimented. We see that TFP gain increases as $\tau_k$ decreases. When the capital wedge is removed, the TFP gain is about 8%, a third of what’s achieved in the first experiment when $\lambda$ increases to the US level.

Our comparison of TFP gains between the results from experiments one, three and four suggests that financial frictions might be the main reason behind resource misallocation and TFP loss in China’s manufacturing sector.

3.5 Concluding remarks

In this paper, we investigated the impact of financial frictions on resource misallocation in China’s manufacturing sector. We found that sectoral TFP can gain about 24% if credit market development in China is at a level similar to the US. Among the TFP gain, 53% is attributed to better factor allocations between SOEs and non-SOEs, and the rest is attributed to better allocations across non-SOEs. Relative to other unspecified frictions between SOEs and non-SOEs as summarized by the two abstract capital or labor wedges, our result suggests that financial frictions may be the main reason behind the resource
misallocation in China’s manufacturing sector.

A few extensions to the current work are worth pursuing and are left for future research. First, this paper emphasizes the fact that productive non-SOEs are discriminated by banks so they have limited access to credit. Recent evidence suggests that because of the credit discrimination, some non-SOEs turn to other sources for credit, for example, informal financial institutions, foreign investors, input suppliers and trade partners (see Hale and Long, 2010 and references therein). Alternative financing sources alleviate the problem of financial frictions. Since they are absent in our model, we run the risk of miscalculating the impact of financial frictions on misallocation. Future work should attempt to control for the effect of alternative financing strategies when quantifying the impact of financial frictions. Second, our model abstracts away the dynamics of the economy. Importantly, productive entrepreneurs may self-finance to undo financial frictions as shown by Moll (2010), i.e., they may save and use internal fund for capital investment. An extension to a dynamic model would allow us to assess the effect of financial frictions both during the transition and at the steady state, which can be very different due to the self-financing mechanism (see, for example, Moll, 2014). Since China’s economy is in its transition, a dynamic model could reveal further insights.
Appendix

3.A Proof of proposition 1

We solve aggregate production for the case that the representative SOE operates in equilibrium. The other case that only non-SOEs operate in equilibrium is solved in a similar way, and hence we omit the algebra here. The total output is

\[ Y = \int_{z_c}^{\infty} \int_{0}^{\infty} (z\lambda a)^{\alpha} \left[ \left( \frac{1 - \alpha}{(1 + \tau_l)w} \right)^{\frac{1}{\alpha}} z\lambda a \right] \frac{1 - \alpha}{(1 + \tau_l)w} g(a, z) \frac{1}{\alpha} g(a, z) \frac{1 - \alpha}{(1 + \tau_l)w} 1 - \frac{1}{\alpha} (1 + \tau_l)^{-\frac{1}{\alpha}} \left( \left( 1 - \lambda M' \right) K + Z_s K_s \right) \]

Equality 3.13 uses the optimal factor demands \( k(a, z) \) and \( l(a, z) \) in Lemma 3.1. Equality 3.14 uses the representative SOE’s FOC for labor, equation 3.4. Equality 3.15 uses the definition of \( M' \). The labor market clearing condition 3.9 together with the SOE’s FOC for labor 3.4 implies

\[ \left( \frac{1 - \alpha}{w} \right)^{\frac{1}{\alpha}} \left[ (1 + \tau_l)^{-\frac{1}{\alpha}} \lambda KM' + Z_s (1 - \lambda M) K \right] = L. \]

Using the capital market clearing condition 3.8, we substitute \( K_s \) in the above expression by

\[ K_s = (1 - \lambda M) K. \]

We get

\[ \left( \frac{1 - \alpha}{w} \right)^{\frac{1}{\alpha}} \left[ (1 + \tau_l)^{-\frac{1}{\alpha}} \lambda KM' + (1 - \lambda M) K Z_s \right] = L. \]

Equality 3.13 uses the optimal factor demands \( k(a, z) \) and \( l(a, z) \) in Lemma 3.1. Equality 3.14 uses the representative SOE’s FOC for labor, equation 3.4. Equality 3.15 uses the definition of \( M' \). The labor market clearing condition 3.9 together with the SOE’s FOC for labor 3.4 implies

\[ \left( \frac{1 - \alpha}{w} \right)^{\frac{1}{\alpha}} \left[ (1 + \tau_l)^{-\frac{1}{\alpha}} \lambda KM' + Z_s (1 - \lambda M) K \right] = L. \]

Using the capital market clearing condition 3.8, we substitute \( K_s \) in the above expression by

\[ K_s = (1 - \lambda M) K. \]

We get

\[ \left( \frac{1 - \alpha}{w} \right)^{\frac{1}{\alpha}} \left[ (1 + \tau_l)^{-\frac{1}{\alpha}} \lambda KM' + (1 - \lambda M) K Z_s \right] = L. \]

Substitute the term \( \left( \frac{1 - \alpha}{w} \right)^{\frac{1}{\alpha}} \) in 3.15 using 3.18, we reach equality 3.16. TFP is thus

\[ Z = \left( \frac{(1 + \tau_l)^{-\frac{1}{\alpha}} \lambda M E \omega | z > z_c} + (1 - \lambda M) Z_s \right) \left[ \left( 1 + \tau_l \right)^{-\frac{1}{\alpha}} \lambda M E \omega | z > z_c} + (1 - \lambda M) Z_s \right]^{1 - \alpha}. \]
Next, we solve capital-labor ratios for the SOE and non-SOEs. The SOE’s FOC for labor 3.4 implies

\[ L_s = \left( \frac{1-\alpha}{\alpha} \right)^{\frac{1}{\alpha}} Z_s K_s. \]

Using 3.18 to replace the term \( \left( \frac{1-\alpha}{\alpha} \right)^{\frac{1}{\alpha}} \), we get

\[ \frac{K_s}{L_s} = \frac{(1+\tau_l)^{-\frac{1}{\alpha}} \lambda \mu E_\omega[z > z_c] + (1-\lambda M) Z_s}{Z_s} \left( \frac{K}{L} \right). \]  

(3.19)

Substituting out \( K_s \) using 3.17, we get

\[ L_s = \frac{(1-\lambda M) Z_s}{(1+\tau_l)^{-\frac{1}{\alpha}} \lambda \mu E_\omega[z > z_c] + (1-\lambda M) Z_s} L. \]

Thus total labor for non-SOEs is \( L_n = L - L_s \) and we obtain

\[ L_n = \frac{(1+\tau_l)^{-\frac{1}{\alpha}} \lambda \mu E_\omega[z > z_c]}{(1+\tau_l)^{-\frac{1}{\alpha}} \lambda \mu E_\omega[z > z_c] + (1-\lambda M) Z_s} L. \]

Since \( K_n = K - K_s = \lambda M K \), we obtain

\[ \frac{K_n}{L_n} = \frac{(1+\tau_l)^{-\frac{1}{\alpha}} \lambda \mu E_\omega[z > z_c] + (1-\lambda M) Z_s}{(1+\tau_l)^{-\frac{1}{\alpha}} \lambda \mu E_\omega[z > z_c]} \left( \frac{K}{L} \right). \]

Thus

\[ \left( \frac{K_s}{L_s} \right) / \left( \frac{K_n}{L_n} \right) = (1+\tau_l)^{-\frac{1}{\alpha}} \frac{\mu E_\omega[z > z_c]}{Z_s}. \]

3.B Factor prices

Due to the distortions, factor prices in this economy deviate from aggregate marginal products. Under the equilibrium where the representative SOE operates, the SOE’s FOCs (3.4 and 3.3) imply

\[ w = (1-\alpha) Z_s^\alpha \left( \frac{K_s}{L_s} \right)^\alpha, \quad R = \alpha Z_s^\alpha \left( \frac{K_s}{L_s} \right)^{\alpha^{-1}}. \]

Substituting out the capital-labor ratios using 3.19, we obtain

\[ w = (1-\alpha) \zeta_l ZK^\alpha L^{-\alpha}, \quad R = \alpha \zeta_k ZK^{\alpha-1}L^{1-\alpha}, \]

where

\[ \zeta_l = \frac{(1+\tau_l)^{-\frac{1}{\alpha}} \lambda \mu E_\omega[z > z_c] + (1-\lambda M) Z_s}{(1+\tau_l)^{-\frac{1}{\alpha}} \lambda \mu E_\omega[z > z_c] + (1-\lambda M) Z_s}, \]

\[ \zeta_k = \frac{Z_s}{(1+\tau_l)^{-\frac{1}{\alpha}} \lambda \mu E_\omega[z > z_c] + (1-\lambda M) Z_s}. \]

Under the equilibrium where only non-SOEs operate, \( w \) can be derived using the labor market clearing condition 3.9 with \( L_s = 0 \), and \( R \) can be derived by evaluating the profit function \( \pi(a, z) \) in lemma 3.1.
at $z = z_c$, The details are omitted. In this case,

$$
\zeta_l = (1 + \tau_l)^{-1}, \quad \zeta_k = (1 + \tau_k)^{-1} \frac{z_c}{Z^{1/\alpha}}.
$$
Bibliography

URL: http://ideas.repec.org/p/nbr/nberwo/14060.html

URL: http://ideas.repec.org/p/nbr/nberwo/17828.html


URL: http://ideas.repec.org/a/red/issued/11-95.html


Brandt, L. and Zhu, X. (2010), Accounting for china’s growth, Working Papers tecipa-394, University of Toronto, Department of Economics.


URL: http://ideas.repec.org/h/eeegro/chp/1-09.html

URL: http://ideas.repec.org/p/p/wb/wps/wb06175.html

URL: http://www.nber.org/papers/w13103

URL: http://ideas.repec.org/a/see/deveco/v96y2011i1p79-94.html

URL: http://ideas.repec.org/a/red/issued/07-73.html

**URL:** [http://qje.oxfordjournals.org/content/114/1/83.abstract](http://qje.oxfordjournals.org/content/114/1/83.abstract)


**URL:** [http://ideas.repec.org/p/nbr/nberwo/21006.html](http://ideas.repec.org/p/nbr/nberwo/21006.html)


**URL:** [http://ideas.repec.org/a/spr/joecth/v32y2007i1p179-221.html](http://ideas.repec.org/a/spr/joecth/v32y2007i1p179-221.html)

**URL:** [http://ideas.repec.org/h/nbr/nberch/11037.html](http://ideas.repec.org/h/nbr/nberch/11037.html)

**URL:** [http://dx.doi.org/10.4324/9780203027332.ch4](http://dx.doi.org/10.4324/9780203027332.ch4)


**URL:** [http://ideas.repec.org/a/cee/chierec/v21y2010i3p411-422.html](http://ideas.repec.org/a/cee/chierec/v21y2010i3p411-422.html)


URL: http://ideas.repec.org/a/red/issued/13-0.html