A Model of Worker Deaths and Firm Adjustment

Nathan Lazarus

Submitted to the Department of Economics
on August 12, 2022, in partial fulfillment of
the requirements for the degree of
Master of Applied Science in Data,
Economics, and Development Policy

Abstract

Here, I (as a component of a paper by Simon Jäger and Jörg Heining) develop and calibrate a dynamic model of wage-setting based on the model in Kline et al. (2019). Using the model, I estimate firms' costs of replacing a worker from empirical reactions to worker deaths in German Social Security data. Estimated replacement costs are quite large, on the order of two years of worker salaries. I show analytically that the rise in wages in response to a death is evidence for convex adjustment costs. However, I estimate the convexity of the hiring cost function and find an exponent of 1.09, far from the typical quadratic functional form. I also calibrate the model separately for thick and thin labor markets, finding that replacement costs are almost three times larger in thin markets. I then generalize the model to have two types of workers and estimate the elasticity of substitution between workers of different occupations, finding an elasticity of 6.5. I also estimate that only 30% of the observed earnings response is due to the intensive margin of hours, while the remaining 70% is due to increased hourly wages. Together, these findings imply a substantial degree of imperfect competition in the labor market, and provide evidence for the existence of rents from employment relationships due to costly replacement.

1 Introduction

This is a component of the paper "How Substitutable Are Workers? Evidence from Worker Deaths," by Simon Jäger and Jörg Heining (2022).[1] My main contribution to this paper was the design of a model and calibrating it to the empirical evidence, using the method of simulated moments. Through various calibrations, we use the model to estimate replacement costs, the convexity of the hiring cost function, and the elasticity of substitution between workers of different occupations. Additionally, as detailed in Section 4.2, by adding an intensive margin of labor supply calibrated to the elasticity preferred by the meta-analysis of Chetty et al. (2013), we estimate the extent of the adjustment that would take place on the intensive margin of hours. We cannot observe this empirically, we observe yearly earnings, not hourly wages, but the model offers a bound on the extent to which this could confound our results. I additionally contributed to the empirical section, updating the results and identifying and fixing econometric issues with the existing specifications.

2 Motivating Empirical Results

The empirical section uses matched employer-employee data on the universe of German workers and firms from social security records, which are linked to death records. For each worker who dies, we identify a matched control worker of the exact same age, gender, and education group, with similar earnings, at a firm with the exact same number of employees (coarsened exact matching, see Iacus et al., 2011). The main specifications compare hiring and coworker wages at firms that experienced a death to their matched control firms with a difference-in-differences estimator, giving quasi-experimental evidence about how firms respond to worker exits.

The empirical evidence identifies three margins of adjustment by which firms respond to a worker death: increased hiring, reduced scale, and increases in incumbent wages to increase retention. Importantly, we observe a positive wage response to worker deaths. In a competitive benchmark, with no frictions and identical workers, the firm would adjust entirely on the hiring margin: it can simply hire one additional worker at the same wage to replace the deceased worker. In the case of constant returns to scale in labor, workers' marginal product would remain constant, so the firm would not increase wages or hiring expenditures. The firm would adjust entirely on the margin of reducing scale. And with linear hiring costs and identical workers, if it was profitable for the firm to engage in positive hiring in steady-state, then the adjustment would also take place entirely on the hiring margin.

[1] This work was done as pre-doctoral research assistant for Simon Jäger. What follows is what I wrote and contributed, much of which is included in the latest revision of the paper, here, and the online appendix, here.

To understand the deviation from these benchmarks, and what the observed incumbent wage increase implies about the lack of ability to adjust on those other margins, we offer a model that follows Kline, Petkova, Williams, and Zidar (2019, henceforth KPWZ), in which incumbents and outside workers are imperfect substitutes due to convex hiring costs. The intuition, following the above, is that when the firm loses a worker, it first recognizes that workers are now more productive due to the decreasing returns to scale in production. It tries to get back to the previous optimal level of workers by increasing hiring, but runs into convex hiring costs. Facing an increased cost of hiring, the firm chooses to raise incumbent wages to make incumbents less likely to take outside job offers.

We argue, therefore, that the fact that some of the adjustment takes place on the incumbent retention margin provides evidence for these convex hiring costs, and the model we calibrate allows us to estimate the degree of convexity in the hiring cost function. We don't microfound these convex hiring costs, but they are the subject of a large literature on adjustment costs in labor markets reviewed in Hamermesh and Pfann (1996) and Bond and Van Reenen (2007). Convex adjustment costs can come from diseconomies of scale in a matching function, where a firm has to expend more than twice as much on vacancy posting to encounter twice as many workers, or diseconomies of scale in training.

In addition to learning about the hiring cost function, the model also allows us to structurally estimate the cost of replacing a worker from the quasi-experimental evidence. Given that adjustment takes time in the data, it's clear that the replacement cost isn't 0 (if it were, the firm would never deviate from its steady-state employment level). The model allows us to quantify these costs. One exercise that is useful in both approximating replacement costs and in motivating the need for numerical estimates from a model is to use the expenditure on incumbent retention to estimate the firms' willingness to pay for an additional worker. We document that, in response to a worker death, firms pay an average of €527 more to their incumbents and get 0.018 more worker-years from their retained workers (in total over a period of five years).[2] The implied expenses for retaining a full incumbent are hence equal to $527/0.018 = 28364$ or roughly one annual salary. This does not provide an upper bound on replacement costs though. We think the cost of retained workers is convex.[3] The fact that the firm was able to acquire 0.018 more worker-years for €28364 per worker-year does not mean that the firm could acquire 5 more worker-years to fully replace the deceased worker

[2]A threat to identification for this exercise is if firms increase wages for their workers as a consequence of deaths, but not for the purpose of increasing retention. In particular, one frequently offered rationale for our findings is that firms increase wages for incumbents because they promote them internally in response to a death. We look at effects on promotions and find small positive effects. We also restrict the sample to incumbents who were initially paid higher or in a higher paying occupation than the deceased and find a similar magnitude of wage increases.

[3]It must be convex at some point, given that is upper-bounded by the number of incumbents. In the calibrated model, it becomes more costly to retain additional workers because the density of workers' reservation wages is declining. This is also the case in a model where preferences are drawn from an extreme value type 1 distribution like Card et al. (2018).

at that price. Convex costs would make further incumbents more costly to retain, so replacement costs could be much greater than €28364.

It's also possible that replacement costs could be lower because the average cost of adjusting through hiring could be lower; it is only the marginal costs that are equalized. We can't say anything about the relative magnitudes of these average costs without imposing a functional form.

In the model, we specify a functional form, letting N denote hiring, of $c(N) = \gamma \frac{1}{1+\lambda} \left(\frac{N}{I}\right)^{1+\lambda}$ which allows us to precisely estimate these costs. We estimate a value for λ of 0.09, which suggests that the cost function is convex, but not nearly as convex as the quadratic form often imposed for adjustment costs ($\lambda = 1$, see Hamermesh and Pfann (1996) and the references therein). In Section 3.4, I show how we derive our preferred replacement specification to use with the model.

3 Model

We present a model in which incumbent workers and newly hired workers are imperfect substitutes due to replacement costs. These hiring costs lead to rents from the employment relationship. We closely follow the model in KPWZ and add both dynamics and multiple types of workers to map the model more closely to the paper's empirical specifications. We also extend the baseline model to allow firms to set hours, and workers to have a disutility of supplying hours, to understand how much of the adjustment takes place on the intensive margin.

3.1 Baseline

We consider a representative firm solving a static problem, endowed with I incumbent workers at the beginning of the period. The firm cannot change the number of incumbents it has, but it can hire outsiders at market wage w^m. In addition to w^m, a firm that hires N new workers incurs a recruitment cost of $c(N/I)I$.[4]

Incumbents receive (exogenous) outside offers between w^m and the maximum wage $\overline{w}$, which are drawn from a distribution with CDF $G(\omega)$.

$$G(\omega) = \left(\frac{\omega - w^m}{\overline{w} - w^m}\right)^{\eta}, \omega \in [w^m, \overline{w}]$$

The firm chooses a wage w^I to pay its incumbents, and the timing is such that this wage is set before workers see their outside offer: firms can't respond to their workers' outside offers either individually (as in Postel-Vinay and Robin (2002)) or based on the deviation of the average draw from its expected value. Those who do not get an outside offer better than w^I stay.

$$\Pr(\text{outside offer} \leq w^I) = G(w^I)$$

[4]This CRS functional form ensures that marginal recruitment costs don't vary with firm size; a specification of $c(N)$ would lead a firm with 100 workers hiring 10 new workers to face much larger hiring costs than a firm with 10 workers hiring 1, given the convexity we find.

A firm therefore expects to retain $G(w^I)I$ workers. After the uncertainty in retention is resolved,[5] the firm chooses hiring of new workers. The total labor employed L is given by

$$L = G(w^I)I + N,$$

and total labor costs are given by

$$w^I G(w^I)I + w^m N + c\left(\frac{N}{I}\right) I$$

.

Production Q is linear in labor, $Q = TL$, and the firm faces downward sloping demand with elasticity $\epsilon > 1$ and demand shifter P^0, $Q^D(P) = \left(\frac{P}{P^0}\right)^{-\epsilon}$. Pricing is therefore given by $P(Q) = P^0 Q^{-\frac{1}{\epsilon}}$ and the marginal revenue product of labor at the firm is:

$$MRP(L) = \frac{dP(Q)Q}{dL} = \left(1 - \frac{1}{\epsilon}\right)\frac{P(Q)Q}{L}$$

Profits $\Pi(I, w^I, N)$ are therefore given by

$$\Pi(I, w^I, N) = P(Q)Q - \left(w^I G(w^I)I + w^m N + c\left(\frac{N}{I}\right) I\right)$$

Differentiating,

$$\frac{\partial \Pi}{\partial w^I} = \mathrm{MRP}(L)\frac{\eta}{w^I - w^m}G(w^I)I - \frac{w^I \eta}{w^I - w^m}G(w^I)I - G(w^I)I$$
$$\frac{\partial \Pi}{\partial N} = \mathrm{MRP}(L) - w^m - c'\left(\frac{N}{I}\right)$$

So the firm's first order conditions are

$$\left(1 - \frac{1}{\epsilon}\right)\frac{P^0(TL)^{\frac{\epsilon-1}{\epsilon}}}{L} = w^I + \frac{w^I - w^m}{\eta}$$
$$\left(1 - \frac{1}{\epsilon}\right)\frac{P^0(TL)^{\frac{\epsilon-1}{\epsilon}}}{L} = w^m + c'\left(\frac{N}{I}\right)$$

The above entirely follows the model in KPWZ. They take I to be exogenous, however. We will look at the firm's dynamic problem with I as a state variable.

[5] The timing here would be important, if the firm chose hiring before knowing how many workers it would retain, it would have to consider its expected marginal product across a range of possible sizes, similar to Stole and Zwiebel (1996). But everything we do takes $G(w^I)$ to be its expectation anyway.

3.2 Comparative Statics

First, we engage in an exercise of comparative statics, looking at what happens in the model when a worker dies. We take that to be an exogenous shock to I, and then look at the responses of wages and hiring $\frac{dw^I}{dI}$ and $\frac{dN}{dI}$, which are informative about the degree to which firms adjust on the hiring versus retention margins.

From the FOCs,

$$w^I + \frac{w^I - w^m}{\eta} = MRP = \left(1 - \frac{1}{\epsilon}\right) \frac{P^0 (TL)^{\frac{\epsilon-1}{\epsilon}}}{L} = w^m + c'\left(\frac{N}{I}\right)$$

$$\frac{d}{dI}\left(\left(1 + \frac{1}{\eta}\right) w^I - \frac{w^m}{\eta}\right) = \frac{d}{dI}\left(w^m + c'\left(\frac{N}{I}\right)\right)$$

$$\left(1 + \frac{1}{\eta}\right)\frac{dw^I}{dI} = \frac{d^2 c(\frac{N}{I})}{d(\frac{N}{I})^2} \frac{I\frac{dN}{dI} - N}{I^2}$$

Using the more compact notation $c''\left(\frac{N}{I}\right) \equiv \frac{d^2 c(\frac{N}{I})}{d(\frac{N}{I})^2}$

$$c''\left(\frac{N}{I}\right)\left(\frac{dN}{dI} - \frac{N}{I}\right) = \frac{1+\eta}{\eta}\frac{dw^I}{dI} I$$

To clarify ideas, suppose c is linear, so $c''(\frac{N}{I}) = 0$. As Manning (2006) discusses for a more general class of models, all of the adjustment in response to a shock to I takes place on the hiring margin. The empirical evidence is that some response takes place on the retention margin, $\frac{dw^I}{dI} < 0$, which rejects the linearity of the cost function.

We can also calculate the marginal cost of recruiting a worker, which tells us about the degree of deviation from perfect competition.[6] Again, equating the FOCs,

$$w^I + \frac{w^I - w^m}{\eta} = w^m + c'\left(\frac{N}{I}\right)$$

$$c'\left(\frac{N}{I}\right) = \left(1 + \frac{1}{\eta}\right)(w^I - w^m) \tag{1}$$

$c'(\frac{N}{I})$ is the marginal cost of recruiting a worker. It's identifiable from equation (1) if we know the wages paid to incumbents and new hires and the shape parameter of the outside offer distribution η. Larger hiring costs will lead to higher incumbent wages because there are more rents from the employment relationship.

[6] In Section 3.4, I discuss other expressions that better correspond to the total costs of replacing a worker after a death. The marginal hiring cost is the cost of increasing I by some small amount and ignores the non-linearities; it also fails to account for the benefit to the firm from the fact that hired workers are paid a lower wage $w^m < w^I$. If the firm incurs a large cost of replacing a worker but faces inelastic labor supply and is therefore able to pass that on to the worker, there will be no distortion.

3.3 Dynamics

The model as presented so far is entirely static, and there are no dynamic benefits from retaining incumbent workers or hiring new workers to become incumbents. This case could correspond to a discount rate of $\beta = 0$. In reality, incumbent workers provide a flow of value to the firm in future periods. Further, the paper's empirical section identifies the dynamic effects of a death over the following five years. To make the model more realistic and map more closely to the empirical results, we add dynamics. Each period is a year, and the firm chooses wages and hiring every period. The firm's problem is defined by the Bellman equation:

$$V(I_t) = \max_{w_t^I, N_t} \Pi(I_t, w_t^I, N_t) + \beta V(I_{t+1}) \quad \text{s.t.} \qquad I_{t+1} = G(w_t^I) I_t + N_t,$$

The first order conditions look similar,

$$\left(1 - \frac{1}{\epsilon}\right) \frac{P_t^0 (T_t L_t)^{\frac{\epsilon-1}{\epsilon}}}{L_t} = w_t^I + \frac{w_t^I - w^m}{\eta} - \beta V'(G(w_t^I) I_t + N_t)$$
$$\left(1 - \frac{1}{\epsilon}\right) \frac{P^0 (T L_t)^{\frac{\epsilon-1}{\epsilon}}}{L_t} = w^m + c'\left(\frac{N_t}{I_t}\right) - \beta V'(G(w_t^I) I_t + N_t)$$

This gives the same decision rule as the static model for $\beta = 0$. For $\beta > 0$, holding the parameters fixed, the optimal L_t will be strictly greater because the value function is increasing (and concave) in I_t.

In estimating the model, we suppose the firm is in a steady state, and at time $t = 0$ it experiences an unexpected decrease in the number of its incumbent workers. Note that this is not an innocuous assumption, in fact, in the data we see both treated firms and matched controls are growing in their number of employees around the years of the death. We use the difference-in-differences estimates from the data and think about these shocks happening to a representative firm in steady state.

3.4 Replacement Costs

We calculate the replacement cost of an incumbent implied by our results. In the data and the model, the replacement cost is not simply the change in total costs after an incumbent death: the firm chooses a lower employment level, so its costs typically fall despite the expenditure on replacement. One statistic we report in Table 2 is the change in profits. Alternatively, we isolate the part of the change in total costs attributable to replacement costs. We decompose the difference in costs between period t and steady state as follows. Let $\mathcal{C}(I)$ denote the total costs incurred in a period given I incumbents, and let asterisks denote steady state values. Then the difference in costs is

$$
\begin{aligned}
&\mathcal{C}(I_t) - \mathcal{C}(I_*)\\
&= w^m N_t - w^m N_* + c(N_t/I_t)I_t - c(N_*/I_*)I_*\\
&\quad + w_t^I G(w_t^I)I_t - w_*^I G(w_*^I)I_*\\
&= w^m(N_t - N_*) + (c(N_t/I_t) - c(N_*/I_*))I_t + c(N_*/I_*)(I_t - I_*)\\
&\quad + w_t^I G(w_t^I)I_t[-w_t^I G(w_*^I)I_t + w_t^I G(w_*^I)I_t][-w_*^I G(w_*^I)I_t + w_*^I G(w_*^I)I_t] - w_*^I G(w_*^I)I_*\\
&= w^m(N_t - N_*) + (c(N_t/I_t) - c(N_*/I_*))I_t + c(N_*/I_*)(I_t - I_*)\\
&\quad + w_t^I(G(w_t^I) - G(w_*^I))I_t + (w_t^I - w_*^I)G(w_*^I)I_t + w_*^I G(w_*^I)(I_t - I_*)\\
&= (c(N_t/I_t) - c(N_*/I_*))I_t + c(N_*/I_*)(I_t - I_*) + (w_t^I - w_*^I)G(w_*^I)I_t + w_t^I(G(w_t^I) - G(w_*^I))I_t\\
&\quad - w_*^I(G(\overline{w}_*^I)(I_* - I_t) - (N_t - N_*)) - (w_*^I - w^m)(N_t - N_*)\\
&= \underbrace{(c(N_t/I_t) - c(N_*/I_*))I_t}_{\text{Change in Hiring Rate}} - \underbrace{c(N_*/I_*)(I_* - I_t)}_{\text{Hiring Scale Effect}}\\
&\quad + \underbrace{(w_t^I - w_*^I)G(w_*^I)I_t}_{\text{Inframarginal Incumbents}} + \underbrace{(w_t^I - w_*^I)(G(w_t^I) - G(w_*^I))I_t}_{\text{Marginal Incumbents}}\\
&\quad - \underbrace{w_*^I(G(w_*^I)(I_* - I_t) - (N_t - N_*) - (G(w_t^I) - G(w_*^I))I_t)}_{\text{Savings from Net Employment Level Decline}} - \underbrace{(w_*^I - w^m)(N_t - N_*)}_{\text{Savings from New Hires}}
\end{aligned}
$$

Suppose an incumbent dies, leading I_t to be less than I_*. The first term captures the increase in hiring costs from convexity, and the second term captures the lower hiring costs due to fewer incumbents. The third and fourth terms capture the larger earnings paid to inframarginal incumbents, who would have been retained given steady-state earnings, and to marginal incumbents, who are newly retained due to the earnings response in t. The fifth term captures savings from a lower employment level, and the sixth term captures savings from using new hires rather than incumbents. Using this decomposition, we define the replacement cost differential in period t, denoted by RCD_t, as the difference in costs relative to steady state excluding the hiring scale effect and savings from the net employment level decline. In other words,

$$
\begin{aligned}
RCD_t = &\text{ Convex Hiring} + \text{Inframarginal Incumbents}\\
&+ \text{Marginal Incumbents} - \text{Savings from New Hires}
\end{aligned}
$$

To arrive at a replacement cost for one full incumbent, we sum the replacement cost differential over time after an incumbent death and normalize by the number of incumbents marginally recruited by the firm. The normalization is necessary because a firm can passively replace an incumbent by sticking with its steady-state choices of incumbent wages and hiring (it will converge back to the steady-state level of employment). Thus, the marginal number of incumbents added from changing w_t^I and N_t is not the total change in incumbents after an incumbent death, but the number of additional incumbents added relative to choosing $w_t^I = w_*^I$ and $N_t = N_*$. Let

$$MI_t \equiv (G(w_t^I) - G(w_*^I))I_t + N_t - N_* \tag{2}$$

denote the number of incumbents marginally recruited by the firm. Then the total replacement cost of an incumbent is the sum of the retention costs across τ periods:

$$RC \equiv \frac{\sum_{t=0}^{\tau} RCD_t}{\sum_{t=0}^{\tau} MI_t} \tag{3}$$

Alternatively, we can compute the recruitment expenditure, defined as just the costs of acquiring a new worker through increasing hiring or retention, ignoring the benefit from newly hired workers earning less than the worker they replaced. We report this as well in Table 2 (for $\tau = 3$, the horizon over which almost all of the replacement takes place).

$$\begin{aligned} RE_t &= \text{Convex Hiring} + \text{Inframarginal Incumbents} \\ &\quad + \text{Marginal Incumbents} \\ RE &\equiv \frac{\sum_{t=0}^{\tau} RE_t}{\sum_{t=0}^{\tau} MI_t} \end{aligned} \tag{4}$$

If the firm were to choose $w_t^I = w_*^I$ and $N_t = N_* + \varepsilon$, $\varepsilon > 0$, it would, as mentioned, gradually converge back to steady state. For small ε,

$$\frac{RE_t}{\varepsilon} = \frac{(c((N_* + \varepsilon)/I_*) - c(N_*/I_*))I_*}{\varepsilon} = c'(N_*/I_*)$$

. That is, the marginal hiring cost is the expenditure per replacement worker for small changes in hiring. Table 2 therefore reports this marginal hiring cost, the replacement cost measures in (3) and (4), lost profits (the cumulative difference in profits due to the shock), and the share of replacement expenditure on hiring (versus increased incumbent wages).

4 Further Extensions

4.1 Two Types

We further generalize the model to allow for multiple types of workers. A key empirical result in the paper is the difference in wage and hiring responses between workers by occupation. The empirical results are that deaths in the same occupation lead to wage increases that are 50% larger for workers in the same occupation as the deceased as for workers in other occupations (€239 to €162), and hiring responses that are 8 times as large. To model this heterogeneity, suppose a firm employs two different types of workers and arrives into the period with I_k incumbents of type $k \in \{A, B\}$, Workers of type A and B combine to produce output according to the CES production function

$$Q = (\alpha(A_A L_A)^\rho + (1-\alpha)(A_B L_B)^\rho)^{\frac{1}{\rho}}$$
$$L_{kj} = G(w^I_{kj})I_k + N_k.$$

The remainder of the model is otherwise the same. The value function now is defined over both types,

$$V(I_A, I_B) = \max_{w^I, N} \Pi(I_A, I_B, w^I_A, N_A, w^I_B, N_B) + \beta V(I^+(I_A, I_B, w^I_A, N_A, w^I_B, N_B))$$
$$\text{s.t.} \quad I^+(I_A, I_B, w^I_A, N_A, w^I_B, N_B) = \{G(w^I_A)I_A + N_A, G(w^I_B)I_B + N_B\}$$

If a worker in type A dies, marginal products and therefore wages will rise for type A workers. If workers are perfect substitutes ($\rho = 1$), wages will rise by the same amount for type B workers. If we find $\rho < \frac{\epsilon-1}{\epsilon}$, wages will fall for workers of type B in response to a death of a type A worker.

4.2 Hours

We also extend the model with disutility of hours. New workers earn w^m, and work a typical market level h^m, which we take to be the average in our data. Outside offers also ask the worker to work h^m. The firm can adjust h^I, its hours requirements for incumbent workers, and receive more hours of labor from them. The tradeoff here is that workers get disutility from working additional hours, and might be willing to accept an outside offer for lower earnings if the firm is demanding $h^I > h^m$. Specifically, there will be a compensating differential, where demanding a higher h^I is equivalent to offering a lower w^I in the worker's decision function. We discipline the model by calibrating it to empirical evidence on the intensive margin of labor supply in response to tax changes from Chetty et al. (2013). Therefore we also introduce a tax τ on labor income so that we can identify the model using this quasi-experimental evidence from tax changes.

Define

$$r(w^I, h^I) = (1-\tau)w^I - \frac{\chi}{1+\psi}((h^I)^{1+\psi} - \phi^{1+\psi}) \tag{5}$$

to be an incumbent's "reservation earnings level", where τ is the effective tax rate, w^I is now interpreted as a worker's earnings rather than wage, and the parameters χ and ψ capture a worker's disutility from labor. The disutility is zero when hours equal the steady-state level. Incumbents receive offers at other firms drawn from the distribution:

$$G(\omega) = \left(\frac{\frac{\omega}{1-\tau} - w^m}{\overline{w} - w^m}\right)^\eta. \tag{6}$$

The division of ω by $1-\tau$ indicates that ω is the post-tax level of earnings. Unlike before, incumbents accept any offer if $\omega \geq r(w^I, h^I)$ rather than $\omega \geq (1-\tau)w^I$.

Equilibrium is now characterized by the profit function

$$\Pi(I, w^I, h^I) = P^0 Q^{\frac{\epsilon-1}{\epsilon}} - c\left(\frac{N}{I}\right) I - w^m N - w^I G(r(w^I, h^I))I$$
$$Q = T\left(N + \frac{h^I}{\phi} G(r(w^I, h^I))I\right),$$

and the four following equilibrium conditions.

$$MRP_t + \beta V'(I_{t+1}) = \frac{\epsilon-1}{\epsilon} P_0 T^{\frac{\epsilon-1}{\epsilon}} \left(N_t + \frac{h_t^I}{\phi} G(r(w_t^I, h_t^I))I_t\right)^{-1/\epsilon}$$
$$MRP_t + \beta V'(I_{t+1}) = w^m - c'\left(\frac{N_t}{I_t}\right)$$
$$MRP_t + \beta V'(I_{t+1}) = \left(\frac{h_t^I}{\phi}\right)^{-1} \left(\frac{(w_t^I - \frac{\chi}{1+\psi}(h_t^I)^{1+\psi}) - (w^m - \frac{\chi}{1+\psi}\phi^{1+\psi})}{\eta} + w_t^I\right)$$
$$MRP_t + \beta V'(I_{t+1}) = w_t^I \left(\frac{h_t^I}{\phi}\right)^{-1} \frac{\frac{\eta\chi}{1+\psi}(h_t^I)^{1+\psi}}{\frac{\eta\chi}{1+\psi}(h_t^I)^{1+\psi} - ((w_t^I - \frac{\chi}{1+\psi}(h_t^I)^{1+\psi}) - (w^m - \frac{\chi}{1+\psi}\phi^{1+\psi}))}$$

To estimate χ and ψ, we target $h^I = 31.55$ (the average level in the data) before the incumbent shock and an intensive-margin Hicksian elasticity of 0.33 following Chetty et al. (2013). Since our model is dynamic, the Hicksian elasticity is the appropriate choice when using a steady-state tax change. The intensive-margin elasticity is calculated by computing the elasticity of hours to a permanent decrease in the effective tax rate by 1%, as in Chetty et al. (2013).

We fix the remaining parameters. We calibrate the remaining parameters. We set $\tau = 0.15$ and $\phi = 31.55$ so that there is no penalty for choosing the steady-state level of hours. As in the baseline model, we target $w^m = 17163$.

5 Results

I report parameter estimates in Table 1 and replacement cost estimates in Table 2. Column 1 reports results based on short-run effects (one year after a worker death) and column 2 based on the overidentified specification using the long-run results for all five years (see Appendix B for a detailed description).

Several clear results emerge that are consistent across specifications and point towards substantial replacement costs. First, we find high values of γ, the parameter determining the steady-state marginal hiring costs, with values ranging from €76,000 to €98,000. Second, we find moderate convexity of hiring costs, with λ being 0.09. A result of $\lambda = 0$ would have implied that all adjustment to a worker death occurred on the hiring (rather than retention) margin. Third, we find low values for η, the elasticity of incumbent retention to the incumbent wage premium, ranging between 0.2 and 0.3. These can be transformed into retention elasticities and are consistent with the reduced-form retention elasticity of 0.62. The estimate is at the lower end, but within the range of estimates for the retention elasticity surveyed in meta-analyses (Manning, 2021; Sokolova and Sorensen, 2021).

As a summary measure, we calculate the implied marginal replacement cost $c'(\frac{N}{I})$ for firms in our sample and find values ranging between €65,000 and €84,000. We compare these to the wages of incumbents in our worker death sample. This calculation reveals a marginal replacement cost between 2.3 and 3 annual salaries of an incumbent. Our estimates of replacement costs are substantially higher than standard estimates in the literature based on firm surveys (see, e.g., Manning, 2011). An important distinction of our results from ones based on firm surveys is that our results draw on actual employment and wage responses of firms in response to worker exits. Our results are in line with the results in Kline et al. (2019, 2021), which point to marginal replacement costs of 1.27 times the annual earnings of an incumbent and who use a similar framework but different empirical strategy with identification stemming from wage differences between new workers and incumbents and rent sharing elasticities.

We further gauge the plausibility of the results of the model by tracing the paths of hiring and incumbent wages implied by our parameter estimates. In Figure 1, we compare them to reduced-form findings. Panel (a) shows that the model almost perfectly replicates the observed short-run employment response to a worker death in the first three years after a worker death. In the subsequent years, we see a slight divergence, with model employment fully converging while observed employment remains slightly lower. However, the difference between the model prediction and the data is not statistically significant. Panel (b) reports results for hiring in the model and the data. The model matches the overall pattern of hiring responses very well, with a sharp increase in the year after the worker death and a subsequent decline. Again, the long-run differences are not statistically distinguishable even though the point estimates for hiring in the data remain slightly elevated compared to the model. Finally, we show the wage response in panel (c). Here, we see a perfect match in the first year after the event (and the response in period 0 for wages is muddled due to taking annual averages). However, we see a divergence in years two through four, where the observed wage response in the data remains more elevated while wages in the model converge more quickly.

Two potential hypotheses for the divergence are (i) that it might take more than one period for new workers to become incumbents because incumbents cannot perfectly substitute for new workers in production (so that the effective number of incumbents remains depressed for longer), or (ii) there could be frictions in wage setting, e.g., wage rigidity, so that a firm cannot easily take back raises it granted. We find larger and more persistent gains in more specialized occupations and thicker labor markets, which suggests that new hires may in fact be imperfect replacements.

Additionally, we calibrate the model to thick and thin labor markets separately. We define labor market thickness by the relative agglomeration of jobs in the given worker's five-digit occupation in the region; that is, the local labor market's share of workers in that occupation relative to average. Thick labor markets are defined as ones above the median share. In thick labor markets, the wage response is more muted (€139 compared to €207 in thin markets) and the retention response is stronger (retention increases by 0.46 percentage points after a death, compared to matched control firms, whereas that is 0.24 percentage points in thin labor markets). These findings consistent with thick labor markets having more elastic labor supply, fewer frictions and less costly replacement. (See the retention elasticity values in Table 2.) We calibrate the one dimensional model separately for each labor market and report the parameter estimates in columns (3) and (4) of Tables 1 and 2. Replacement costs are much larger in thin labor markets.

Table 2 shows the results of the replacement cost decomposition described in Section 3.4. Hiring costs explain the vast majority of replacement costs, 93–95% across specifications. Almost all of the rest, quantitatively, is explained by the additional wages paid to inframarginal incumbents. The costs spent on obtaining marginal incumbents are minimal since the retention probability does not increase by much in comparison with the stock of workers (we target an increase in retention from 82.6% to 82.9% that we observe in the data). As described in Section 3.4, replacement costs are lower than recruitment costs due to the significant savings from paying new hires less than incumbents, that is, the firm captures some of its expenditure on recruitment through wage markdowns.

Figure 2 shows the results of the hours calibration described in Section 4.2. We find that about 70% of the observed increase in total earnings is due to an increase in hourly wages, whereas 30% of the increase in earnings is actually a consequence of increased hours. This still leaves the central results about imperfect competition intact, but it reduces the estimated replacement costs because the firm has another margin on which to adjust. Estimated replacement costs fall by more than half, as shown in column (5) of Table 2.

6 Conclusion

We are able to use the model to shed light on the intensive margin of hours adjustment, which is a threat to our interpretation of the wage effects as imperfect substitutability with outside workers. In another robustness check, we instead bring in administrative data on hours from the German Statutory Accident Insurance, and estimate the treatment effect on hours. But these data are too noisy to rule out the positive wage effects being driven entirely by hours adjustment. With the model, we formalize the intuition that there must be costs to adjusting hours, otherwise optimizing firms would already have demanded more labor supply on the intensive margin.

The model quantifies replacement costs across thick and thin markets, showing that labor markets with more jobs in a worker's occupation have higher elasticities of labor supply to the firm and therefore lower costs of replacing workers. This informs a growing literature showing that labor market structure matters for wage-setting (Azar et al., 2022; Jarosch et al., 2019). The model predicts larger markdowns of the wages of new hires due to increased replacement costs in thick markets.

Finally, the model provides a bridge between the reduced-form estimates of the responses to the shock of a worker death and the structural hiring cost function. We find evidence for convex adjustment costs and are able to estimate the degree of that convexity based on the size of the wage response. The parameter λ nests both the frictionless world ($\lambda = 0$, linear costs), and the case of quadratic adjustment costs, $\lambda = 1$. We both reject linearity and the extreme convexity of the quadratic cost function.

References

Azar, J., I. Marinescu, and M. Steinbaum (2022): "Labor Market Concentration," *Journal of Human Resources*, 57, S167–S199.

Bond, S. and J. Van Reenen (2007): "Microeconometric Models of Investment and Employment," *Handbook of Econometrics*, 6, 4417–4498.

Card, D., A. R. Cardoso, J. Heining, and P. Kline (2018): "Firms and Labor Market Inequality: Evidence and Some Theory," *Journal of Labor Economics*, 36, S13–S70.

Chetty, R., A. Guren, D. Manoli, and A. Weber (2013): "Does Indivisible Labor Explain the Difference between Micro and Macro Elasticities? A Meta-Analysis of Extensive Margin Elasticities," *NBER Macroeconomics Annual*, 27, 1–56.

Hamermesh, D. S. and G. A. Pfann (1996): "Adjustment Costs in Factor Demand," *Journal of Economic Literature*, 34, 1264–1292.

Iacus, S. M., G. King, and G. Porro (2011): "Causal Inference Without Balance Checking: Coarsened Exact Matching," *Political Analysis*.

Jarosch, G., J. S. Nimczik, and I. Sorkin (2019): "Granular Search, Market Structure, and Wages," Working Paper 26239, National Bureau of Economic Research.

Kline, P., N. Petkova, H. Williams, and O. Zidar (2019): "Who Profits from Patents? Rent-Sharing at Innovative Firms," *The Quarterly Journal of Economics*, 134, 1343–1404.

——— (2021): "Corrigendum to 'Who Profits from Patents?'," *Working Paper*.

Manning, A. (2006): "A Generalised Model of Monopsony," *The Economic Journal*, 116, 84–100.

——— (2011): "Imperfect Competition in the Labor Market," *Handbook of Labor Economics*, 4, 973–1041.

——— (2021): "Monopsony in Labor Markets: A Review," *ILR Review*, 74, 3–26.

Postel-Vinay, F. and J.-M. Robin (2002): "Equilibrium Wage Dispersion with Worker and Employer Heterogeneity," *Econometrica*, 70, 2295–2350.

Sokolova, A. and T. Sorensen (2021): "Monopsony in Labor Markets: A Meta-Analysis," *ILR Review*, 74, 27–55.

Stole, L. A. and J. Zwiebel (1996): "Intra-firm Bargaining under Nonbinding Contracts," *The Review of Economic Studies*, 63, 375–410.

Su, C.-L. and K. L. Judd (2012): "Constrained Optimization Approaches to Estimation of Structural Models," *Econometrica*, 80, 2213–2230.

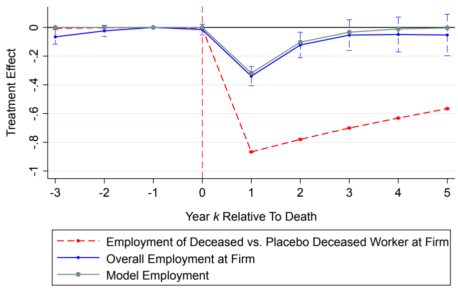
A Figures and Tables

Figure 1: Model Prediction vs. Reduced-Form Effects

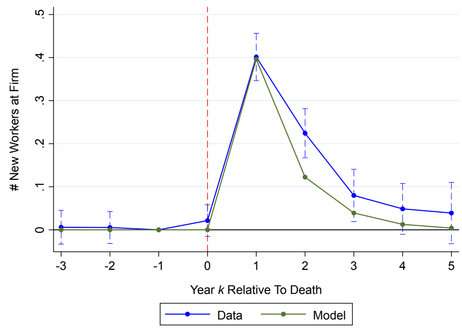
Note: The figure displays effects of worker deaths on several firm and incumbent worker outcomes. The blue lines report the measured effect in the data. The gray and green lines report the model predictions.

Figure 2: Model Prediction vs. Reduced-Form Effects

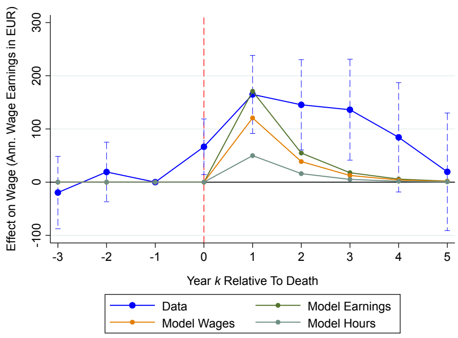
(a) Employment



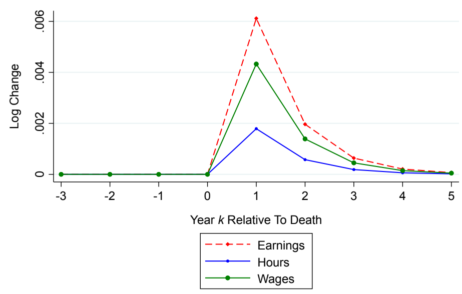
(b) Hiring



(c) Incumbent Wages



(d) Change in Earnings, Wages, and Hours



Note: The figure displays effects of worker deaths on several firm and incumbent worker outcomes. The blue lines report the measured effect in the data. The gray and green lines report the model predictions. Panel (c) shows model-implied earnings path if either hours did not move ("Model Wages") or wages did not move ("Model Hours") after an incumbent death. Panel (d) shows how the observed change in earnings can be decomposed into an intensive margin effect and a wage effect.

Table 1: Estimation of Model Parameters

	(1) Short-Run Estimation	(2) Long-Run Estimation	(3) Thick Labor Markets (Short-Run)	(4) Thin Labor Markets (Short-Run)
γ	76054	97944	45692	115518
λ	0.09	0.09	0.07	0.10
η	0.20	0.15	0.35	0.13
$\overline{w}$	45448	56310	35307	64726
ϵ	1.33	1.31	1.87	1.01
P^0	1147740	1394109	297754	61400900
Retention Elasticity Reduced Form	0.621		1.111	0.391

Note: The first column is estimated to match the wage, retention, and employment responses in the first year after a worker death. The second column matches the entire path of responses over a five-year horizon; see Appendix B for more information. Columns (3) and (4) split the sample by labor market thickness and report replacement costs for both specifications separately.

Table 2: Estimated Replacement Costs

	(1) Short-Run Estimation	(2) Long-Run Estimation	(3) Thick (Short-Run)	(4) Thin (Short-Run)	(5) Intensive Margin
Replacement Cost	291%	394%	163%	455%	122%
Replacement Expenditure	327%	431%	196%	491%	156%
Change in Profit	-239%	-329%	-128%	-383%	-90%
Marginal Cost of New Hires $c'\left(\frac{N}{I}\right)$	65449	84203	40402	96502	37727
As % of incumbent wage	(236%)	(303%)	(146%)	(345%)	(134%)
Share of replacement expenditure on hiring	94%	95%	93%	95%	91%

Note: The table displays various measures of the replacement costs described in Section 3.4. Most are expressed as a percentage of an incumbent's average annual earnings (€27770 overall, €27655 in thick markets and €27955 in thin markets).

Table 3: Estimation of Model Parameters
Extensions to the Baseline Model

A. Extension with intensive margin:		
	Baseline Estimation	Intensive Margin
γ	76054	53145
λ	0.09	0.04
η	0.20	0.49
$\overline{w}$	45448	33383
ϵ	1.33	4.96
P^0	1147740	69973
Marginal Cost of New Hires $c'(\frac{N}{I})$	65449	37727
(Expressed as % of incumbent salary)	(232%)	(134%)
B. Extension to two worker types (by occupation):		
	Occupation Calibration	
$\gamma_{\text{same occ}}$	68826	
$\lambda_{\text{same occ}}$	0.08	
$\eta_{\text{same occ}}$	0.21	
$\overline{w}_{\text{same occ}}$	43544	
$\gamma_{\text{other occ}}$	117651	
$\lambda_{\text{other occ}}$	0.21	
$\eta_{\text{other occ}}$	0.16	
$\overline{w}_{\text{other occ}}$	47634	
$A_{\text{other occ}}$	1.17	
ρ	0.85	
ϵ	[1.5]	
P^0	1831712	
Marginal Cost of New Hires $c'(\frac{N}{I})$	59504	
(Expressed as % of incumbent salary)	(211%)	

Note: The table replicates the specification in Table 1 in column 1. The intensive-margin column reports estimation results when allowing for an hours response (see Section 4.2). The occupation calibration draws on the two-type model and reports results for the substitutability of workers across occupational boundaries.

B Computational Strategy

To implement the estimation, we adopt the mathematical program with equilibrium constraints (MPEC) approach proposed by Su and Judd (2012). The typical approach for estimating equilibrium models is the following procedure.

1. Solve the model accurately given a fixed set of parameters.
2. Use an optimization algorithm (either derivative-free or with finite difference approximations to the derivatives) to update the parameters.
3. Iterate until a solution is found.

The issue with this approach is that step 1 is usually time-consuming. MPEC bypasses this issue by reframing the estimation problem as a constrained optimization problem. The targeted moments comprise the objective to minimize while equilibrium conditions are imposed as constraints. This approach speeds up computation by only solving the model accurately for the final set of parameters. Most algorithms for constrained optimization problems allow constraints to be violated during the parameter search and are robust to these violations. As a result, the algorithm does not waste time repeatedly solving the model for parameters that are not close to hitting the targeted moments. For the extension to two types, the model is estimated in the same way, with the value function being two-dimensional over both types of workers.

To implement the method of simulated moments, we follow Su and Judd (2012) and minimize squared deviations from the targeted moments subject to the model's equilibrium conditions holding as constraints. The dynamic equilibrium conditions are infinite-dimensional because the Bellman equation has to hold at every point. We obtain a finite-dimensional representation of the problem by approximating the value function with polynomials.[7] To confirm the accuracy of our method, we also use value function iteration to check the moment function at the solution.

In the baseline model, we estimate six parameters and have six moments, hence our model is exactly identified. The six parameters to estimate are γ, λ, η, w^m, $\overline{w}$, and P^0. We normalize labor productivity to $T = 1$ since it is not separately identified from P^0. We fix $\beta = 0.96$ to match a 4% annual discount rate, which is standard in the literature. And we fix w^m to the average wages of newly hired workers in the data, €17163 (and €17169 in thick markets and €17156 in thin markets).[8]

[7] Interestingly, Chebyshev polynomials perform worse here because they create non-convexities in the moment function with respect to the coefficients, which we're optimizing over.

[8] In the two types model, these are fixed to the average wages of new hires in the same occupation as the deceased, €17802, and average wages of new hires in other occupations, €16433. These differences are likely reflective of the deceased workers being older than average and therefore in higher paying 1-digit occupations.

These six moments are defined based on values from the data in the year just before and just after the death. But in fact we have, in the empirical event studies, information from the five years following the death. The natural generalization is to create the same moment conditions in these later years and be overidentified. We estimate the same model with the additional three moment conditions in each of the four later years,[9] and report this as the "Long-Run Estimation" in column (2) of Table 1.

[9] We weight each moment equally, a one-step approach, for clarity. Note that parameter estimates are scaled by their standard errors, giving a unit-free moments function.