Econometrics Problem Set 4

Giacomo, Francesca, Neeharika, Noor, Kun

21 March 2022

A fundamental question in labor economics is whether a longer duration of unemployment benefits improve the quality of the job after finding employment. However, providing a causal answer to this question is difficult because unemployment duration is correlated with individual characteristics that may also affect job quality. In this exercise, we use a regression discontinuity design for causal identification. Under regular circumstances, an unemployed person receives unemployment benefits for 30 weeks. However, the benefit duration increases to 39 weeks if the person is at least 40 years old when he/she becomes unemployed. The dataset $data\_ps3.dta$ **is a sample of workers who have been laid off and subsequently found employment. The sample contains the following variables:** age **age at layoff** $nonemp$ **non-employment duration in weeks** $jobfind$ **a dummy that equals one if a person had a new job after 39 weeks** $lwage0$ **log monthly wage in previous job** $lwage1$ **log monthly wage in new job.**

1) Explain intuitively how this discontinuity in benefit duration can be exploited to estimate the causal effect of benefit duration on wages after re-employment.

In this regression discontinuity design, the running variable is the age of unemployed individuals, where the cut-off threshold of the sharp RDD is at age of 40. For individual falls above the threshold, the potential unemployment benefit duration will increase from 30 weeks to 39 weeks. Given that age is independent of subjective decision, it is less likely for one year increase in age to render dramatic variations that compared to age in the previous state. Also, age is an objective measure which cannot be changed at one's own will. Therefore, we expect a "smooth" continuous function for our running variable, and possibly no "jumps" around the threshold for variables of interest that might be caused by age alone. By this construct, we consider the continuity assumption of this sharp RDD is satisfied. A satisfied continuity assumption, RDD offers the complete overlap for treatment group and control group, of which the across group differences become infinitesimal close to their neighbourhoods at the limit around the cut-off point. And thus become comparable.

We notice that the observable, age, also our discontinuity running variable, is confounding both the treatment status as well as the potential outcomes. While age presumably is not a factor of how treatment of benefit duration affects the wage in re-employment. In other words, backdoor path from this confounding variable is closed by including the running variable in our model. Further, the running variable can be considered as independent of how the treatment would affect the potential outcome and the choice of the discontinuity cannot be endogenous to treatment assignment. Therefore, we limit the selection bias in the sense that the treatment is exogenous to the assignment rule, individuals will not self-select into the treatment with the on-and-off assignment rule.

To confirm that there is no self-selection, data should show no bunching around the cut-off threshold, which would indicate no evidence of manipulation around the discontinuity for running variable. With satisfied continuity assumption. In fact, it further implies that the age is not a source of heterogeneous treatment effect around the cut-off, and so the age is not deterministic based on the potential outcome on wage. Thus, with valid discontinuity, conditional independence is satisfied.

By construct, it is only the benefit duration that is changed at the cut-off threshold. So, with any dependent variable of interest which manifests "jump" from below to above the cut-off, should be considered as the treatment effect of receiving the longer benefit duration. When we consider to test the identification or to check the observable individual characteristics for potential sources of heterogeneity, with valid continuity assumption, RDD is informative. Particularly, the cut-off offers comparison of behavioural change in ante ex and ex post, as well as offers comparison of individual characteristics for those who aged below 40 and above. Valid continuity assumption also implies that there is no omitted variable bias, and also as no need to control for other variables.

Let age_i index each individual i for our running variable age, and denote the cut-off age eligibility for extended employment benefit duration as age_c with $age_c = 40$. The treatment status of having up to 39 weeks of unemployment benefit for each individual, D_i, is given by the treatment eligibility for this sharp RDD,

$$D_i = \begin{cases} 1 & \text{if } age_i \geq 40 \\ 0 & \text{if } age_i < 40 \end{cases} \tag{1}$$

By allowing variation in the maximum duration which is based on entitlement criterion of age of 40. We expect the potential outcomes differ. Given the treatment is the additional 9 weeks, this extension may exaggerate the magnitude and reveal the direction and the monotonicity of effects of unemployment benefit on wage in re-employment. Particularly by comparing the neighbourhoods which are below and just above the cut-off threshold.

Let X_i denote the discontinuity running variable of age_i, and Y_i denote the potential outcome of wage in re-employment, superscript denotes the treatment status,

$$Y_i^0 = \alpha + \beta X_i \tag{2}$$

$$Y_i^1 = Y_i^0 + \delta \tag{3}$$

where Y_i^0 is the potential outcome without treatment, Y_i^1 is the case that treatment occurs at the discontinuity, and δ is the treatment effect parameter at the discontinuity. Thus, for a linear relationship we have,

$$Y_i = Y_i^0 + (Y_i^1 - Y_i^0)D_i \tag{4}$$

$$Y_i = \alpha + \beta X_i + \delta D_i + \upsilon_i \tag{5}$$

and we have the local treatment at discontinuity,

$$\begin{aligned} \delta &= \lim_{X_i \to age_c} E[Y_i^1 | X_i = age_c] - \lim_{X_i \leftarrow age_c} E[Y_i^0 | X_i = age_c] && (6) \\ &= \lim_{X_i \to age_c} E[Y_i | X_i = age_c] - \lim_{X_i \leftarrow age_c} E[Y_i | X_i = age_c] && (7) \\ &= E[Y_i^1 - Y_i^0 | X_i = age_c] && (8) \end{aligned}$$

where the potential outcome is estimated by Wald Estimator, $\widehat{\delta}/\widehat{\Delta_Y}$ (equation (24)), and the treatment effect parameter is estimated by the reduced form (equation (16)). Estimates for Y_i indicate the treatment effect on wage of new job after being eligible for additional unemployment duration for an unemployed individual whose age is equal to or above 40.

To discuss the true benefit effect on wage in re-employment, first we need to acknowledge the conditions of the state of employment, by which we mean the underlying reasons that move individuals from employed to unemployed, or from unemployment to re-employment. Additionally, we also need to consider the decision-making preference on transition period between the two states of employment, namely, the unemployment duration. Thus, in order to establish the causal relationship for potential different unemployment benefit duration, we need to consider what are the conditions that changed by the potential difference in unemployment benefit.

We begin with the selectivity of employment. Suppose a business follows a pyramid-like organisational structure, where age is positively associated with seniority and experience. With an indiscriminate and proportionate layoff, younger the individuals higher the likelihood to lose job. Similarly for recruitment, individuals with more experience and better skilled are more likely to find a job. Furthermore, job matching for both employer and employee also matters in the sense of wage preference and vacancy availability. Altogether, the selectivity and job matching imply that the heterogeneity gives disproportionate shape of the population distributions across unemployed, employed, and the entire labour supply.

Next, we consider the employment decision, in other words, the willingness to work. First, we look at the preference over leisure-work trade-off. With any offering or potential available job, individual may prefer not to work given a high leisure-over-work utility ratio. Unemployment benefit then would serve as subsidy for such leisure-over-work preference, which leads to delay in re-employment whenever possible, so longer the benefit duration the further extended delay of re-employment.

Second, employment decision depends on the preference over job quality, particularly, compared to the previous job. Both selectivity and job matching are factors of such preference formation. Unemployment benefit is unlikely to have direct effect on improving the situation of selectivity and job matching, however, it allows more time for individuals to prepare or to search for a new job. Benefit then pays off the cost of such improvement, which provides opportunity for unemployed to find job with higher wage and to increase the probability of finding a matched job. This is more salient for individuals those who choose to leave the previous job, and for those who are more elastic with wage.

Finally, employment decision may depend on the expectation of the new wage, where the benefit pay will be taken as the reference point in the re-employment. In addition to preferences over leisure-work ratio and over job quality, we assume that individuals will compare the job offering with unemployment benefit and to maximise their utility. With a given amount of benefit, individuals who have wage below the benefit amount may update their wage preference closer to the benefit, and individuals who have wage above the benefit amount may further expect higher wage from the new job that is away from the benefit. This is particularly important for benefit recipients who will experience income drop after the benefit terminates.

Therefore, preferences on employment decision imply that the potential outcome of extended unemployment benefit duration is expected to have impact on unemployment duration, probability of finding new job, timing of finding a new job, wage differences between new and old, quality of job, and probability of having new wage being higher than the benefit amount.

Nevertheless, for each potential outcome of interest, with heterogeneity in employment selectivity and decision preference, the true treatment effect of benefit duration on wage outcome will be biased due to the violation of conditional independence assumption as well as the self-selection into treatment. Furthermore, without a experimental setup, we need to proxy the counterfactual for the average treatment effect (ATE) at sub-population where treatment group and control group should satisfy the overlap condition for extrapolation. That says, to estimate the causal effect, identification for this sharp RDD needs to satisfy conditional independence, complete overlap, and to eliminate the selection bias.

2) Write down an estimating equation for a sharp regression discontinuity design whereby you control for the running variable with a second-order polynomial that is allowed to differ

above and below the discontinuity. State and explain the identifying assumption that is necessary to interpret your coefficient of interest as causal.

Now suppose a nonlinear relationship, let Y_i now denote all the dependent variable of potential unemployment benefit effects on individual,

$$E[Y_i^0|X_i] = f(X_i) \tag{9}$$

and we have,

$$Y_i = f(X_i) + \delta D_i + \eta_i \tag{10}$$

which is valid when the smoothness of $f(X_i)$ is satisfied. Note that this is the nonlinear counterpart of a linear relationship in equation (5).

However, given the treatment will affect the age profile of outcome on each treated individual, we do not observe the potential outcome for the same treated individuals if there were no treatment (i.e., the counterfactual is in range of $X_i > age_c$). By employing higher-order polynomials, we extrapolate the age profile with $f(X_i)$ for two different X_i functions at values above and below the cut-off with their corresponding interaction terms with treatment status D_i. We rearrange the two second-order polynomials on each side of cut-off threshold to obtain,

$$E[Y|X] = E[Y^0|X] + (E[Y^1|X] - E[Y^0|X])D \tag{11}$$

$$E[Y_i^0|X_i] = f_0(X_i) = \lambda + \beta_{01}\ \widetilde{X}_i\ + \beta_{02}\ \widetilde{X}_i^2 \tag{12}$$

$$E[Y_i^1|X_i] = f_1(X_i) = \lambda + \beta_{11}\ \widetilde{X}_i\ + \beta_{12}\ \widetilde{X}_i^2\ + \delta \tag{13}$$

$$Y_i = \lambda + \beta_1\ \widetilde{X}_i\ + \beta_2\ \widetilde{X}_i^2\ + \gamma_1\ \widetilde{X}_i\ D_i + \gamma_2\ \widetilde{X}_i^2\ D_i + \delta D_i + \epsilon_i \tag{14}$$

$$Y_i = \lambda + f^m(X_i) + f^n(X_i)D_i + \delta D_i + \epsilon_i \tag{15}$$

where f^m and f^n are second-order polynomials, $\widetilde{X}_i$ is the running variable that recentred at $X_i - age_c$ with change in coefficients of intercept term. Coefficient estimates of interest at discontinuity give $\gamma_1 = \beta_{11} - \beta_{01}$, and $\gamma_2 = \beta_{12} - \beta_{02}$, where subscript p and q ($p = 0, 1$; $q = 1, 2$) in β_{pq} and γ_{pq} denote the treatment status and the index for higher-order coefficient parameter, respectively.

Equation (15) is reduced to (5), when assuming the age profile $f(X_i)$ is linear or is the same for values above and below the cut-off. Namely, $\beta_{11} = \beta_{01}$ and $\beta_{12} = \beta_{02}$, so $\gamma_1 = 0$ and $\gamma_2 = 0$. Additionally, when identification assumption holds, ϵ_i is expected to be indifferent at discontinuity.

Assuming perfect compliance, when every unemployed claims their benefit fully and accordingly based on age. Our interest of intention-to-treat (ITT) lies upon the additional benefit duration on outcomes, which is for the compliers who were 40 when being laid-off,

$$\begin{aligned} Reduced\,Form &= E[Y_i|D_i = 1, i \in complier] - E[Y_i|D_i = 0, i \in complier] \\ &= \widehat{\delta} \end{aligned} \tag{16}$$

From previous section, we state that the individual characteristics are not just affect their state of employment, but also the unemployment duration. Heterogeneity then gives the treatment in unemployment benefit with three subcompliant population: always-takers, whose age is equal or greater than 40 and being unemployed for more than 30 weeks; never-takers, whose age is less than 40 when laid-off; compliers, who were 40 at when laid-off, claiming the standard duration if stay unemployed equal or less than 30 weeks, or claiming the extended duration if stay unemployed for more than 30 weeks. The reduced form identifies the treatment effect on individuals who are around the threshold of age cut-off, by looking at the outcomes between those who close to age 40 and just above age 40.

Let us denote the non-employment duration in weeks, $nonemp_i$; the new job finding indication dummy, $jobfind_i$; the log monthly wage in previous job, $lwage0_i$; the log monthly wage in new job, $lwage1_i$; the unemployment benefit payment amount, $benefit_i$. The followings characterise the First Stage, $\widehat{\Delta_Y}$, for each dependent variable of interest from discussion in the previous section. We begin with the treatment effect on non-employment duration,

$$\begin{aligned} First\ Stage &= E[nonemp_i|D_i = 1] - E[nonemp_i|D_i = 0] \\ &= \widehat{\Delta_1} \end{aligned} \tag{17}$$

For probability of finding new job,

$$\begin{aligned} First\ Stage &= E[jobfind_i|D_i = 1] - E[jobfind_i|D_i = 0] \\ &= \widehat{\Delta_2} \end{aligned} \tag{18}$$

For wage differences between new and old,

$$\begin{aligned} First\ Stage &= E[lwage1_i - lwage0_i|D_i = 1] - E[lwage1_i - lwage0_i|D_i = 0] \\ &= \widehat{\Delta_3} \end{aligned} \tag{19}$$

For quality of job,

$$\begin{aligned} First\ Stage &= E[lwage1_i|D_i = 1] - E[lwage1_i|D_i = 0] \\ &= \widehat{\Delta_4} \end{aligned} \tag{20}$$

For probability of having new wage being higher than the benefit amount,

$$\begin{aligned} First\ Stage &= E[lwage1_i > benefit_i|D_i = 1] - E[lwage1_i > benefit_i|D_i = 0] \\ &= \widehat{\Delta_5} \end{aligned} \tag{21}$$

For timing of finding a job within 30 weeks of unemployment,

$$\begin{aligned} First\ Stage &= E[jobfind_i * nonemp_i(\leq 30)|D_i = 1] - E[jobfind_i * nonemp_i(\leq 30)|D_i = 0] \\ &= \widehat{\Delta_6} \end{aligned} \tag{22}$$

For timing of finding a job within 39 weeks of unemployment,

$$\begin{aligned} First\ Stage &= E[jobfind_i * nonemp_i(\leq 39)|D_i = 1] - E[jobfind_i * nonemp_i(\leq 39)|D_i = 0] \\ &= \widehat{\Delta_7} \end{aligned} \tag{23}$$

Given the discontinuity design and variables of interest, our treatment effect of interest is to look at individuals whose local treatment effect at discontinuity divided by difference in the actual treatment. We estimate with the Wald estimator, from (16) and the first stage, $\widehat{\Delta_Y}$, we have,

$$Wald\ Estimator = \frac{E[Y_i|D_i = 1, i \in complier] - E[Y_i|D_i = 0, i \in complier]}{First\ \ Stage} = \frac{\widehat{\delta}}{\widehat{\Delta_Y}} \tag{24}$$

For each dependent variable of interest, from $\widehat{\Delta_1}$ to $\widehat{\Delta_7}$ (equation (17) to (23)), we have the Wald Estimators as follows, respectively,

$$\frac{\widehat{\delta}}{\widehat{\Delta_1}} = \lim_{age_i \rightarrow age_c} E[Y_i|nonemp_i] \ - \lim_{age_i \leftarrow age_c} E[Y_i|nonemp_i] \tag{25}$$

$$\frac{\widehat{\delta}}{\widehat{\Delta_2}} = \lim_{age_i \rightarrow age_c} E[Y_i|jobfind_i] \ - \lim_{age_i \leftarrow age_c} E[Y_i|jobfind_i] \tag{26}$$

$$\frac{\widehat{\delta}}{\widehat{\Delta_3}} = \lim_{age_i \rightarrow age_c} E[Y_i|lwage1_i - lwage0_i] \ - \lim_{age_i \leftarrow age_c} E[Y_i|lwage1_i - lwage0_i] \tag{27}$$

$$\frac{\widehat{\delta}}{\widehat{\Delta_4}} = \lim_{age_i \rightarrow age_c} E[Y_i|lwage1_i] \ - \lim_{age_i \leftarrow age_c} E[Y_i|lwage1_i] \tag{28}$$

$$\frac{\widehat{\delta}}{\widehat{\Delta_5}} = \lim_{age_i \rightarrow age_c} E[Y_i|lwage1_i > benefit_i] \ - \lim_{age_i \leftarrow age_c} E[Y_i|lwage1_i > benefit_i] \tag{29}$$

$$\frac{\widehat{\delta}}{\widehat{\Delta_6}} = \lim_{age_i \rightarrow age_c} E[Y_i|jobfind_i * nonemp_i(\leq 30)] \ - \lim_{age_i \leftarrow age_c} E[Y_i|jobfind_i * nonemp_i(\leq 30)] \tag{30}$$

$$\frac{\widehat{\delta}}{\widehat{\Delta_7}} = \lim_{age_i \rightarrow age_c} E[Y_i|jobfind_i * nonemp_i(\leq 30)] \ - \lim_{age_i \leftarrow age_c} E[Y_i|jobfind_i * nonemp_i(\leq 30)] \tag{31}$$

3) Carry out two tests for the validity of the RD design:

1) plot the density of age at layoff

Before testing the validity, we observed the data. We noticed that *lwage1* has a lot of NAs. Furthermore, we noticed that the distribution of NAs of *lwage1* covers most of the NAs of *lwage0* and *nonemp*. We checked how *lwage1* NAs are distributed in the dataset and we noticed that there is no statistical difference between the *lwage1* NAs and the rest of the distribution. For this reason, we dropped the observations without concerns.

The RDD is valid if three assumptions hold. First of all there must not be manipulation of units around the discontinuity. In other words we want that distribution of units around the threshold does not present bunching.

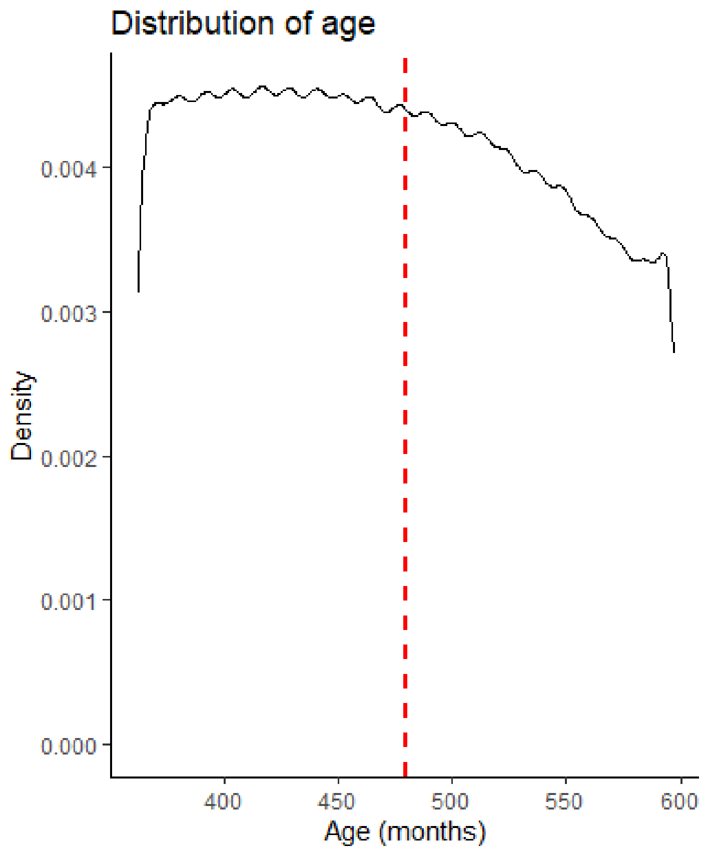


Figure 1: Figure showing the density of age at layoff

Figure 1 plot the distribution of age (considered in months). As we can see at 480 months there is no bunching and the distribution is smooth across the discontinuity. The figure shows the distribution of age grouped in bins whose width is 4 months each.

2) plots the log previous wage against the age at layoff. For each test, produce a scatter plot with bin size 4 months (i.e. each point summarizes the average value on the vertical axis for workers whose age at layoff falls within that bin). For easier visual inspection, the plot should contain a vertical line at the discontinuity and lines for the second-order polynomials

above and below the cutoff. What do those graphs tell us about the validity of the RD design?

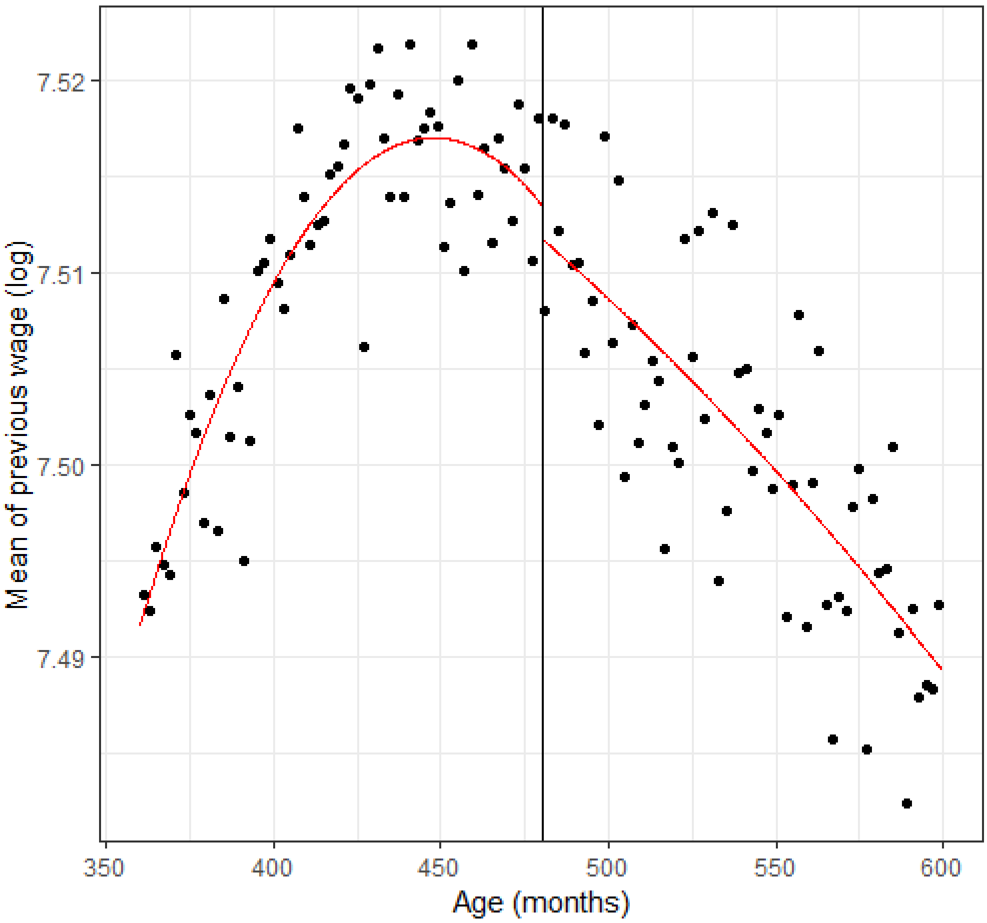


Figure 2: Polynomial of Log of previous wages at layoff

Figure 2 plot the age at layoff against the mean of the wage in the previous job. Each point represent the value for a 4-months-age bin. We derived a second-order polynomial fitting the data, allowing its parameters to vary below and above the threshold. As we can see from the plot, we cannot state there is a significant discontinuity around the threshold if we plot the second-order polynomial.

4) Produce the main results graphically. We focus here on three outcomes, namely non-employment duration, the probability of finding a job within 39 weeks and log wages in the new job. Use the same binned scatters as in exercise 3). Interpret your results.

In figure 3 we see the unemployment benefits duration against the age. We see a sharp rise in the distribution of the bins. We fitted the data using again a second-order polynomial. We can observe that there is a sharp rise in the average duration for people above 40 years old. Moreover, we see that the benefits duration increase monotonically with the age, then a second-polynomial is not even necessary to fit the data. In figure 4, instead, is reported the discontinuity for the value of the new wage, found after the layoff. As we can see, the similarities with figure 2 are many. This allows us to state that the

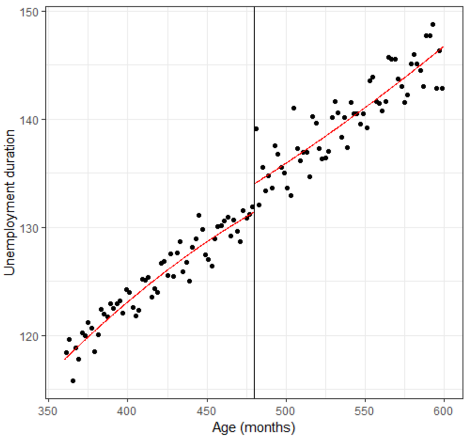


Figure 3: Unemployment benefits duration discontinuity

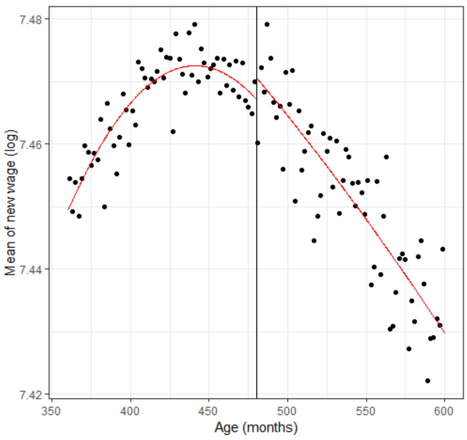


Figure 4: New wage (log) discontinuity

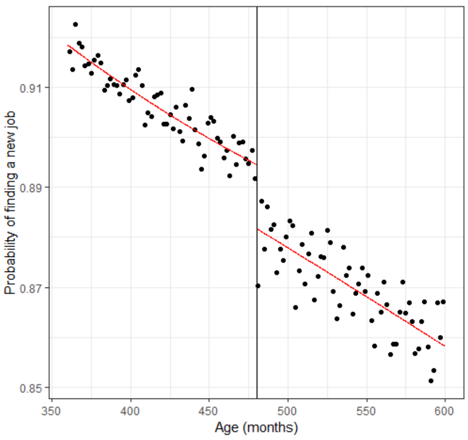


Figure 5: Probability of finding a job discontinuity

discontinuity may be not exploitable. Finally, figure ?? plots the age at layoff against the probability of finding a new job. There is negative quadratic relationship. After, the discontinuity, probability of finding the new job sharply deceases with age.

We estimate the following outcomes with our estimation model

$$Y_i = \lambda + f^m(X_i) + f^n(X_i)D_i + \delta D_i + \epsilon_i$$

for non-employment duration, the probability of finding a job within 39 weeks and log wages in the new job, respectively,

$$\frac{\widehat{\delta}}{\widehat{\Delta}_1} = \lim_{age_i \rightarrow age_c} E[Y_i|nonemp_i] - \lim_{age_i \leftarrow age_c} E[Y_i|nonemp_i] \tag{32}$$

$$\frac{\widehat{\delta}}{\widehat{\Delta}_4} = \lim_{age_i \rightarrow age_c} E[Y_i|lwage1_i] - \lim_{age_i \leftarrow age_c} E[Y_i|lwage1_i] \tag{33}$$

$$\frac{\widehat{\delta}}{\widehat{\Delta}_7} = \lim_{age_i \rightarrow age_c} E[Y_i|jobfind_i * nonemp_i(\leq 39)] - \lim_{age_i \leftarrow age_c} E[Y_i|jobfind_i * nonemp_i(\leq 39)] \tag{34}$$

5) Focusing on the same three outcomes, report the coefficient of interest (i.e. the coefficient of a dummy for being above or below the discontinuity) of the regression outlined in exercise 2). Produce a regression table with five panels

- **The reduced-form effect at the discontinuity in the full sample using the regression from exercise**

- **The reduced-form effect at the discontinuity with a bandwidth of ±5 years in age at layoff.**

- **The reduced-form effect at the discontinuity in the full sample using a linear control for the running variable, allowing for different slopes above and below the discontinuity.**

- **The reduced-form effect at the discontinuity in the full sample, controlling for the running variable with a fourth-order polynomial and, allowing for different parameters above and below the discontinuity**

- **The reduced-form effect at the discontinuity based on the optimal bandwidth computed based on the procedure by Calonico et al. (2014). You can use their Stata/R package *rdrobust*. The package offers multiple procedures to calculate the bandwidth. Use the default procedure. Report the estimated bandwidth along with each coefficient. Interpret and discuss the differences between the panels.**

Table 1: Table of estimates of the discontinuity on new wage (log)

	Dependent variable:				
	lwage1				
	(1)	(2)	(3)	(4)	(5)
cutoff	-0.015^{***}	-0.010^{***}	0.230^{***}	68.962	0.005
	(0.001)	(0.001)	(0.011)	(52.792)	(0.003)
Observations	1,189,446	619,039	1,189,446	1,189,446	638,623
R^2	0.0003	0.0001	0.001	0.001	0.001
Adjusted R^2	0.0003	0.0001	0.001	0.001	0.001
Residual Std. Error	0.433 (df = 1189444)	0.435 (df = 619037)	0.433 (df = 1189442)	0.433 (df = 1189436)	
F Statistic	338.892^{***} (df = 1; 1189444)	82.665^{***} (df = 1; 619037)	275.892^{***} (df = 3; 1189442)	98.132^{***} (df = 9; 1189436)	
Note:					*p<0.1; **p<0.05; ***p<0.01

This tables report the estimates of the discontinuity on *jobfind* in different settings. In (1) we see the difference of the means above and below the discontinuity across all the sample. In (2) we see the difference of the means above and below the discontinuity in the ±5 bandwidth. In (3) the running variable *age* is added as linear control across all the sample. In (4) two fourth-order polynomials of the running variable *age* are added above and below the threshold. In (5) we see the estimate when the choice of the bandwidth followed the procedure of Calonico (2014). Standard errors are shown in parentheses.

The table 1 is a summary of the RDD design with the estimates of the discontinuity on the new wages measured in logs. We see that our estimates in columns 1,2 and 3 are statistically significant. In column 1, we have the dummy of the treatment, we see that there is a decrease of 1.5% in the wages in the next period. In other words, with the discontinuity at the age 40, the wages in the next employment are 1.5% significantly lower. In column 2, we have the regression design as the dummy treatment but with a bandwidth, and within this bandwidth we see that wages earned in the next employed are 1% significantly lower. There is a difference of of 0.5 units earned. Therefore we can say that the individuals within this bandwidth earn slight higher wages as opposed to the 1st model. Column 3 gives us the estimate with a linear control. We now focus on the estimates in column 4 and 5, there are statistically insignificant. Column 4 gives the values of the 4th order polynomial. While on the other hand, we have in column 5 the model of Calonico with optimal bandwidth selection. The optimal bandwidth here is estimated as ±2.565. In the last model we added *lwage0* as a covariates, as a proxy of workers unobserved characteristics. The non-significance of the model is consistent with the intuitions in figure 16.

Table 2: Table of estimates of the discontinuity on probability of finding a new job

	Dependent variable:				
	jobfind				
	(1)	(2)	(3)	(4)	(5)
cutoff	-0.035^{***}	-0.024^{***}	-0.013	34.651	-0.012^{***}
	(0.001)	(0.001)	(0.008)	(38.165)	(0.002)
Observations	1,189,446	619,039	1,189,446	1,189,446	638,623
R^2	0.003	0.001	0.004	0.004	0.004
Adjusted R^2	0.003	0.001	0.004	0.004	0.004
Residual Std. Error	0.313 (df = 1189444)	0.315 (df = 619037)	0.313 (df = 1189442)	0.313 (df = 1189436)	
F Statistic	$3{,}731.028^{***}$ (df = 1; 1189444)	884.104^{***} (df = 1; 619037)	$1{,}432.560^{***}$ (df = 3; 1189442)	477.954^{***} (df = 9; 1189436)	
Note:					*p<0.1; **p<0.05; ***p<0.01

This tables report the estimates of the discontinuity on *jobfind* in different settings. In (1) we see the difference of the means above and below the discontinuity across all the sample. In (2) we see the difference of the means above and below the discontinuity in the ±5 bandwidth. In (3) the running variable *age* is added as linear control across all the sample. In (4) two fourth-order polynomials of the running variable *age* are added above and below the threshold. In (5) we see the estimate when the choice of the bandwidth followed the procedure of Calonico (2014). Standard errors are shown in parentheses.

Table 2 reports the estimates for the same regression design for the probability of finding a new job after 39 weeks from layoff. In the first two columns, the discontinuity is negative and significant, similarly to what we saw in figure 5. While in 3 and in 4 is not significant anymore, it is significant again column 5. The model estimated using the procedure suggested by Calonico (2014) actually provides the lowest estimate for the coefficient, being half of the one estimated with the bandwidth of ±5 and one third of the full model. In this case, it ends up that the optimal bandwidth is ±3.894. In other words, as we reduce the bandwidth, from model (2) to model (5), individuals share more similar characteristics, and the treatment effect decreases. Though it remains negatively significant.

In table 3, we present the summary of the estimates of discontinuity on unemployment benefits duration. We can see by comparing the estimates in columns 1, 2 3 are statistically significant at 99%, as it appeared from figure 3. In 5, however, the magnitude of the effect decreases to 2.64. The optimal bandwidth in this case is ±3.000.

Table 3: Table of estimates of the discontinuity on unemployment benefits duration

	Dependent variable:				
	nonemp				
	(1)	(2)	(3)	(4)	(5)
cutoff	14.574^{***}	8.245^{***}	5.427^{*}	−2,540.438	2.635^{**}
	(0.226)	(0.314)	(3.206)	(14,986.970)	(1.043)
Observations	1,189,446	619,039	1,189,446	1,189,446	638,623
R^2	0.003	0.001	0.004	0.004	0.004
Adjusted R^2	0.003	0.001	0.004	0.004	0.004
Residual Std. Error	122.907 (df = 1189444)	123.459 (df = 619037)	122.849 (df = 1189442)	122.848 (df = 1189436)	
F Statistic	$4,157.099^{***}$ (df = 1; 1189444)	689.555^{***} (df = 1; 619037)	$1,763.696^{***}$ (df = 3; 1189442)	588.803^{***} (df = 9; 1189436)	

Note: *p<0.1; **p<0.05; ***p<0.01

This tables report the estimates of the discontinuity on *nonemp* in different settings. In (1) we see the difference of the means above and below the discontinuity across all the sample. In (2) we see the difference of the means above and below the discontinuity in the ±5 bandwidth. In (3) the running variable *age* is added as linear control across all the sample. In (4) two fourth-order polynomials of the running variable *age* are added above and below the threshold. In (5) we see the estimate when the choice of the bandwidth followed the procedure of Calonico (2014). Standard errors are shown in parentheses.

Final File _V3

Kun, Giacomo, Francesca, Neeharika and Noor

21/03/2022

```
rm(list = ls())

library(ggplot2)
library(foreign)

## Warning: package 'foreign' was built under R version 4.1.2

library(stargazer)

## Warning: package 'stargazer' was built under R version 4.1.2

##
## Please cite as:

##  Hlavac, Marek (2022). stargazer: Well-Formatted Regression and Summary
Statistics Tables.

##  R package version 5.2.3. https://CRAN.R-project.org/package=stargazer

library(haven)
library(rlang)

## Warning: package 'rlang' was built under R version 4.1.2

library(tidyverse)

## ── Attaching packages ─────────────────────────────────────── tidyverse
1.3.1 ──

## ✓ tibble  3.1.6     ✓ dplyr   1.0.8
## ✓ tidyr   1.2.0     ✓ stringr 1.4.0
## ✓ readr   2.1.2     ✓ forcats 0.5.1
## ✓ purrr   0.3.4

## Warning: package 'tidyr' was built under R version 4.1.2

## Warning: package 'readr' was built under R version 4.1.2

## Warning: package 'dplyr' was built under R version 4.1.2

## ── Conflicts ──────────────────────────────────────────
tidyverse_conflicts() ──
## x purrr::%@%()         masks rlang::%@%()
## x purrr::as_function() masks rlang::as_function()
## x dplyr::filter()      masks stats::filter()
## x purrr::flatten()     masks rlang::flatten()
## x purrr::flatten_chr() masks rlang::flatten_chr()
## x purrr::flatten_dbl() masks rlang::flatten_dbl()
## x purrr::flatten_int() masks rlang::flatten_int()
## x purrr::flatten_lgl() masks rlang::flatten_lgl()
## x purrr::flatten_raw() masks rlang::flatten_raw()
## x purrr::invoke()      masks rlang::invoke()
## x dplyr::lag()         masks stats::lag()
## x purrr::splice()      masks rlang::splice()
```

```
library(finalfit)
library(rdrobust)
library(ggplot2)
library(binsreg)
```

```
library(rddtools)
```

```
## Warning: package 'rddtools' was built under R version 4.1.2
```

```
## Loading required package: AER
```

```
## Loading required package: car
```

```
## Loading required package: carData
```

```
## Warning: package 'carData' was built under R version 4.1.2
```

```
##
## Attaching package: 'car'
```

```
## The following object is masked from 'package:dplyr':
##
##     recode
```

```
## The following object is masked from 'package:purrr':
##
##     some
```

```
## Loading required package: lmtest
```

```
## Loading required package: zoo
```

```
##
## Attaching package: 'zoo'
```

```
## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric
```

```
## Loading required package: sandwich
```

```
## Loading required package: survival
## Loading required package: np

## Nonparametric Kernel Methods for Mixed Datatypes (version 0.60-11)
## [vignette("np_faq",package="np") provides answers to frequently asked
questions]
## [vignette("np",package="np") an overview]
## [vignette("entropy_np",package="np") an overview of entropy-based methods]

##
## Please consider citing R and rddtools,
## citation()
## citation("rddtools")

data = read_dta("data_ps4.dta")

##Data screening

data$age_month = data$age*12
summary(data)

##       age            lwage0             nonemp          jobfind
##  Min.   :30.00   Min.   :-0.8335   Min.   :  1.0    Min.   :0.0000
##  1st Qu.:34.86   1st Qu.: 7.3066   1st Qu.: 49.0    1st Qu.:1.0000
##  Median :39.60   Median : 7.5786   Median : 82.0    Median :1.0000
##  Mean   :39.72   Mean   : 7.5169   Mean   :114.3    Mean   :0.8375
##  3rd Qu.:44.49   3rd Qu.: 7.7937   3rd Qu.:125.0    3rd Qu.:1.0000
##  Max.   :50.00   Max.   :13.8387   Max.   :729.0    Max.   :1.0000
##                  NA's   :3054      NA's   :149609
##      lwage1         age_month
##  Min.   : 1.1     Min.   :360.0
##  1st Qu.: 7.2     1st Qu.:418.4
##  Median : 7.5     Median :475.2
##  Mean   : 7.5     Mean   :476.6
##  3rd Qu.: 7.7     3rd Qu.:533.9
##  Max.   :11.7     Max.   :600.0
##  NA's   :549341

#Let's observe how NAs match inside the dataset
table(is.na(data$lwage1), is.na(data$nonemp)) #All nonemp NAs are also lwage1
NAs

##
##             FALSE    TRUE
##   FALSE 1189446       0
##   TRUE   399732  149609

table(is.na(data$lwage1), is.na(data$lwage0)) #Not all lwage0 NAs are also
lwage1 NAs

##
##             FALSE    TRUE
##   FALSE 1187476    1970
##   TRUE   548257    1084
```

```
#Observing if lwage1 NAs (the most numerous) are distributed uniformely
across the dataset

Diff_mean=mean(data[is.na(data$lwage1),]$lwage0, na.rm=T)-
mean(data[!is.na(data$lwage1),]$lwage0, na.rm=T)
Var=sqrt(var(data[is.na(data$lwage1),]$lwage0,
na.rm=T)+var(data[!is.na(data$lwage1),]$lwage0, na.rm=T))
2*(1-pnorm(abs(Diff_mean/Var)))
```

```
## [1] 0.9588284
```

```
#Mean of lwage0 in the part of population where we do not observe the lwage1
is not statistically different
#from the part of the population where we observe it

Diff_mean=mean(data[is.na(data$lwage1),]$age_month, na.rm=T)-
mean(data[!is.na(data$lwage1),]$age_month, na.rm=T)
Var=sqrt(var(data[is.na(data$lwage1),]$age_month,
na.rm=T)+var(data[!is.na(data$lwage1),]$age_month, na.rm=T))
2*(1-pnorm(abs(Diff_mean/Var)))
```

```
## [1] 0.9306214
```

```
#Distribution of age_month in the sample of population where we do not
observe the lwage1 is not statistically different
#from the sample of the population where we observe it

Diff_mean=mean(data[is.na(data$lwage1),]$nonemp, na.rm=T)-
mean(data[!is.na(data$lwage1),]$nonemp, na.rm=T)
Var=sqrt(var(data[is.na(data$lwage1),]$nonemp,
na.rm=T)+var(data[!is.na(data$lwage1),]$nonemp, na.rm=T))
2*(1-pnorm(abs(Diff_mean/Var)))
```

```
## [1] 0.5892496
```

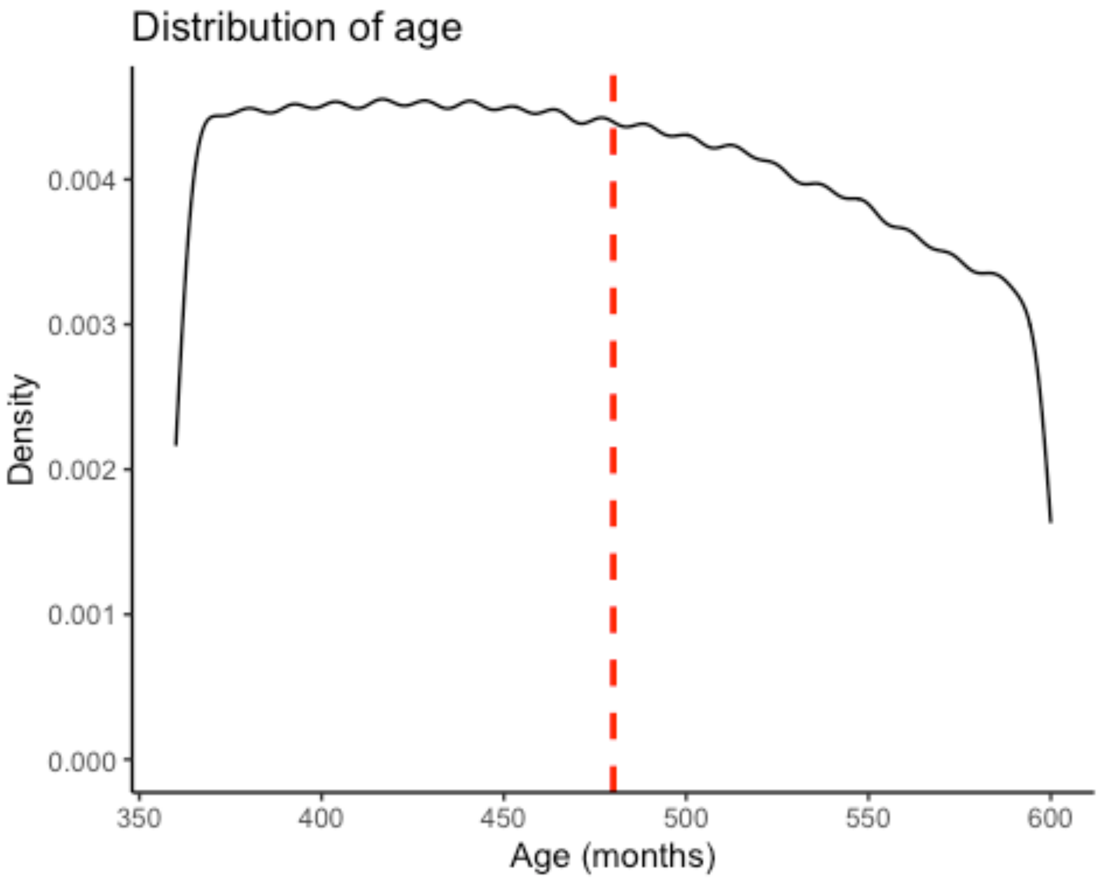
```
#Distribution of nonemp in the part of population where we do not observe the
lwage1 is not statistically different
#from the part of the population where we observe it

Diff_mean=mean(data[is.na(data$lwage1),]$jobfind, na.rm=T)-
mean(data[!is.na(data$lwage1),]$jobfind, na.rm=T)
Var=sqrt(var(data[is.na(data$lwage1),]$jobfind,
na.rm=T)+var(data[!is.na(data$lwage1),]$jobfind, na.rm=T))
2*(1-pnorm(abs(Diff_mean/Var)))
```

```
## [1] 0.7625978
```

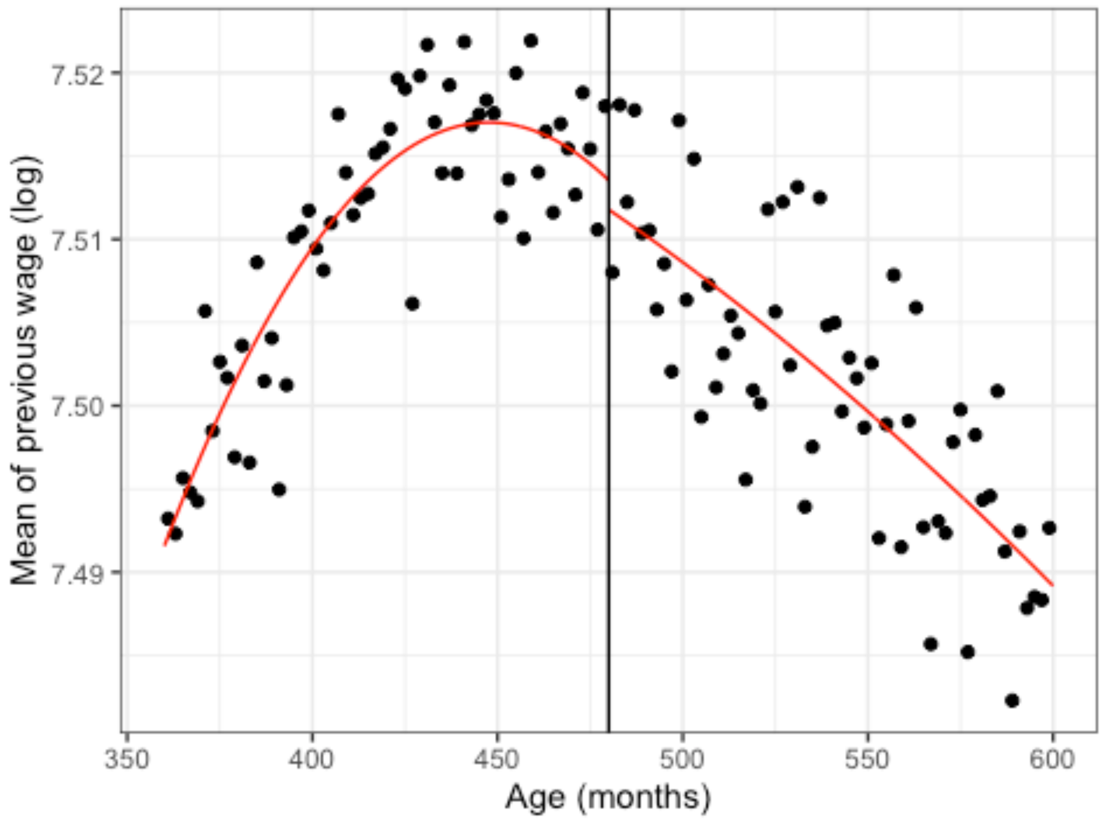
```
#Distribution of jobfind in the part of population where we do not observe
the lwage1 is not statistically different
#from the part of the population where we observe it

#We can drop out NAs of lwage1 without excessive concerns

data = subset(data, !is.na(data$lwage1))
summary(data)
```

```
##       age            lwage0            nonemp         jobfind      
##  Min.   :30.00   Min.   :-0.8335   Min.   :  1.0   Min.   :0.0000  
##  1st Qu.:34.65   1st Qu.: 7.2954   1st Qu.: 60.0   1st Qu.:1.0000  
##  Median :39.31   Median : 7.5595   Median : 95.0   Median :1.0000  
##  Mean   :39.50   Mean   : 7.5065   Mean   :131.9   Mean   :0.8896  
##  3rd Qu.:44.19   3rd Qu.: 7.7728   3rd Qu.:150.0   3rd Qu.:1.0000  
##  Max.   :50.00   Max.   :13.8387   Max.   :729.0   Max.   :1.0000  
##                  NA's   :1970                                      
##      lwage1         age_month    
##  Min.   : 1.094   Min.   :360.0  
##  1st Qu.: 7.245   1st Qu.:415.9  
##  Median : 7.521   Median :471.7  
##  Mean   : 7.460   Mean   :474.0  
##  3rd Qu.: 7.738   3rd Qu.:530.2  
##  Max.   :11.737   Max.   :600.0  
## 
```

```
##RDD diagnostics
Density <- ggplot(data, aes(x = age_month)) +
  geom_density() +
  geom_vline(aes(xintercept = 480),
             color = "red", linetype = "dashed", size = 1)+
  labs(x="Age (months)", y="Density", title="Distribution of age")+
  theme_classic()
Density
```

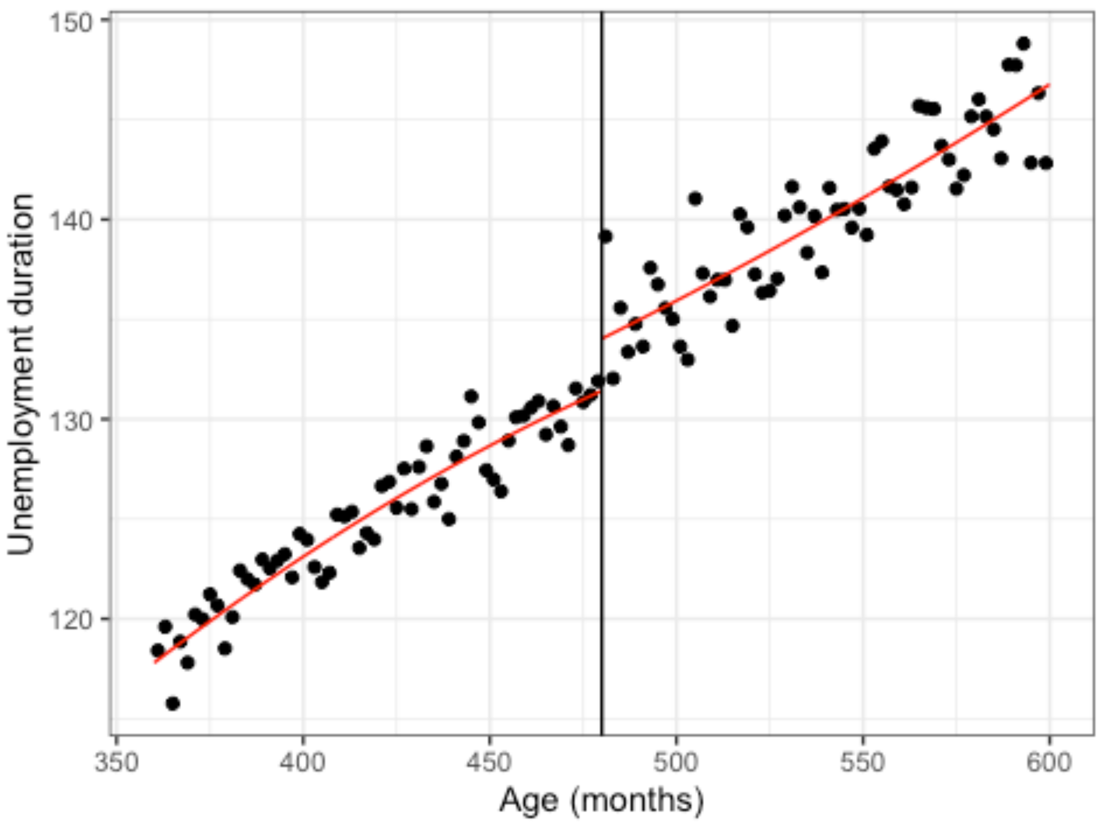


```
Prev_wage = rdplot(data$lwage0, data$age_month, c=480, p=2, nbins=60,
col.dots = "black",
                   x.label = "Age (months)", y.label="Mean of previous wage
(log)", title = "" )

## [1] "Mass points detected in the running variable."
```

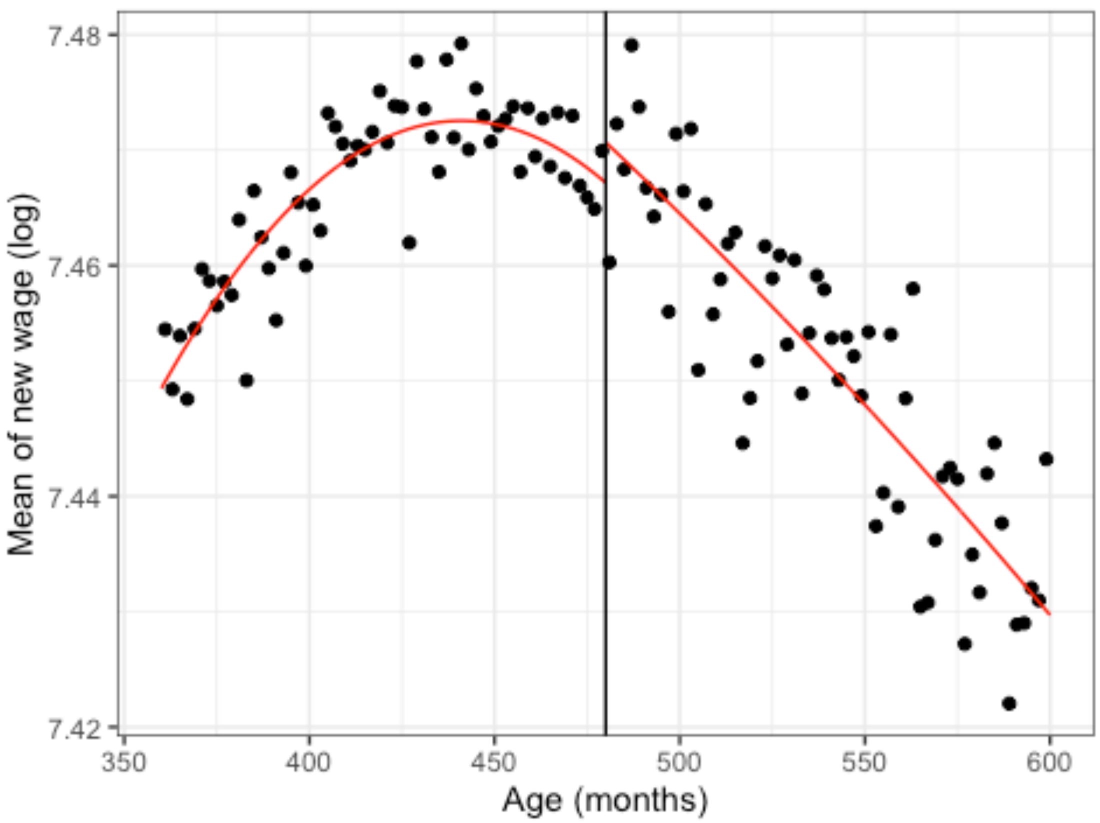


```
##RDD estimates graph

unemp_RDD <- rdplot(data$nonemp, data$age_month, c=480, p=2, nbins=60,
col.dots = "black",
                    x.label = "Age (months)", y.label="Unemployment
duration", title = "" )
```

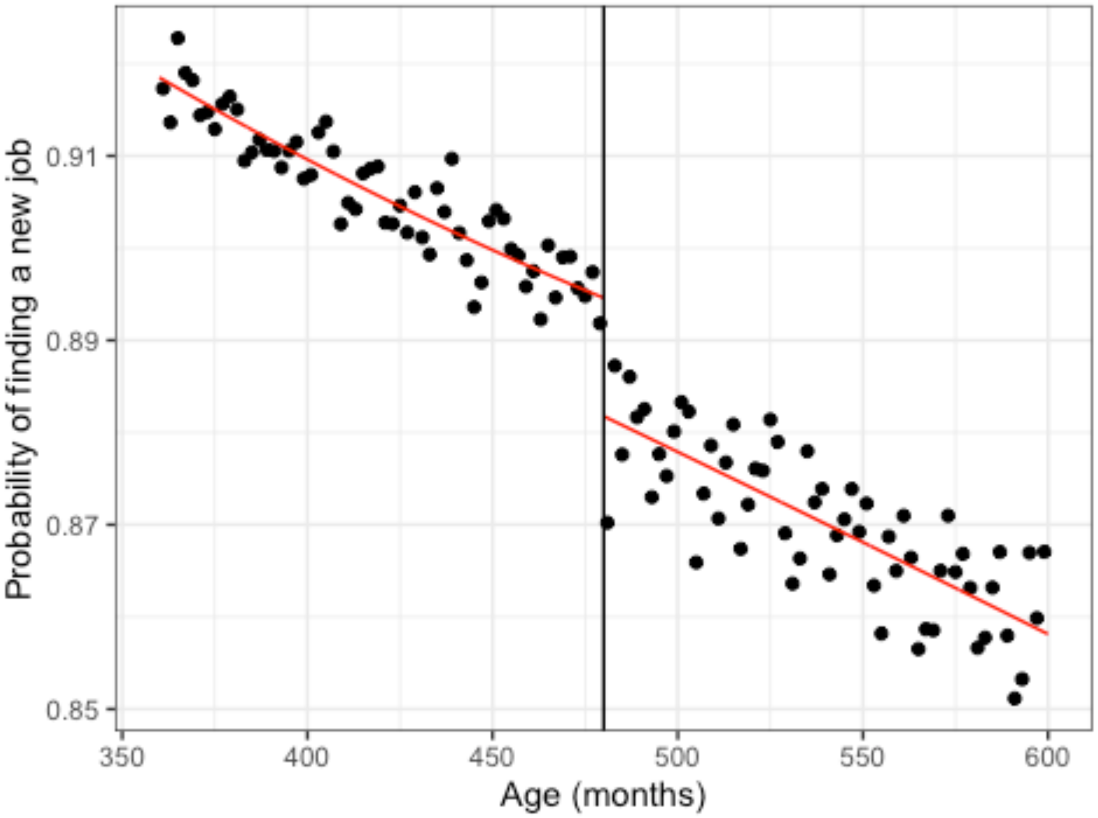
```
## [1] "Mass points detected in the running variable."
```



```
lwage1_RDD <- rdplot(data$lwage1, data$age_month, c=480, p=2, nbins=60,
col.dots = "black",
                          x.label = "Age (months)", y.label="Mean of new wage
(log)", title = "" )

## [1] "Mass points detected in the running variable."
```



```
jobfind_RDD <- rdplot(data$jobfind, data$age_month, c=480, p=2, nbins=60,
col.dots = "black",
                      x.label = "Age (months)", y.label="Probability of
finding a new job", title = "" )

## [1] "Mass points detected in the running variable."
```



```
##Regression Table

data$cutoff = ifelse(data$age_month>=480, 1, 0)

reg1A<-lm(lwage1~cutoff, data = data)
reg2A<-lm(lwage1~cutoff, data = data, subset = data$age>=35&data$age<=45)
reg3A<-lm(lwage1~cutoff*age, data = data)
reg4A<-lm(lwage1~cutoff*age+cutoff*I(age^2)+cutoff*I(age^3)+cutoff*I(age^4),
data = data)
#reg5A<-rdrobust(data$lwage1, data$age, c=40, covs=data$lwage0)

reg1B<-lm(jobfind~cutoff, data = data)
reg2B<-lm(jobfind~cutoff, data = data, subset = data$age>=35&data$age<=45)
reg3B<-lm(jobfind~cutoff*age, data = data)
reg4B<-lm(jobfind~cutoff*age+cutoff*I(age^2)+cutoff*I(age^3)+cutoff*I(age^4),
data = data)
#reg5B<-rdrobust(data$jobfind, data$age, c=40, covs=data$lwage0)

reg1C<-lm(nonemp~cutoff, data = data)
reg2C<-lm(nonemp~cutoff, data = data, subset = data$age>=35&data$age<=45)
reg3C<-lm(nonemp~cutoff*age, data = data)
reg4C<-lm(nonemp~cutoff*age+cutoff*I(age^2)+cutoff*I(age^3)+cutoff*I(age^4),
data = data)
#reg5C<-rdrobust(data$nonemp, data$age, c=40, covs=data$lwage0)

stargazer(reg1A, reg2A, reg3A, reg4A, type = "text")

##
##
=============================================================================
=======================================================
##
Dependent variable:
##                     ------------------------------------------------------
----------------------------------------------------------
##
lwage1
##                                 (1)                          (2)
(3)                          (4)
## --------------------------------------------------------------------------
----------------------------------------------------------
## cutoff                        -0.015***                   -0.010***
0.230***                     68.962
##                                (0.001)                     (0.001)
(0.011)                     (52.792)
##
## age
0.002***                     -1.772
##
(0.0002)                     (1.938)
##
## I(age2)
0.075
##
(0.083)
##
## I(age3)
-0.001
##
(0.002)
##
## I(age4)
0.00001
##
(0.00001)
##
## cutoff:age
-0.006***                     -5.786
##
(0.0003)                     (4.874)
##
## cutoff:I(age2)
```

```
0.178
##
(0.171)
##
## cutoff:I(age3)
-0.002
##
(0.003)
##
## cutoff:I(age4)
0.00001
##
(0.00002)
##
## Constant                      7.467***                    7.471***
7.405***                     22.969
##                               (0.001)                     (0.001)
(0.007)                     (16.837)
##
## ---------------------------------------------------------------------------
----------------------------------------------------------
## Observations                 1,189,446                    619,039
1,189,446                    1,189,446
## R2                             0.0003                      0.0001
0.001                        0.001
## Adjusted R2                    0.0003                      0.0001
0.001                        0.001
## Residual Std. Error     0.433 (df = 1189444)        0.435 (df = 619037)
0.433 (df = 1189442)        0.433 (df = 1189436)
## F Statistic         338.892*** (df = 1; 1189444) 82.665*** (df = 1;
619037) 275.892*** (df = 3; 1189442) 98.132*** (df = 9; 1189436)
##
=============================================================================
=======================================================
## Note:
*p<0.1; **p<0.05; ***p<0.01
```

```
stargazer(reg1B, reg2B, reg3B, reg4B, type = "text")
```

```
##
##
=============================================================================
=============================================================
##
Dependent variable:
##                       ------------------------------------------------------
----------------------------------------------------------------
##
jobfind
##                                    (1)                            (2)
```

```
(3)                           (4)
## ---------------------------------------------------------------------------
-----------------------------------------------------------------
## cutoff                        -0.035***                      -0.024***
-0.013                        34.651
##                                (0.001)                        (0.001)
(0.008)                       (38.165)
##
## age
-0.002***                       -0.128
##
(0.0001)                      (1.401)
##
## I(age2)
0.004
##
(0.060)
##
## I(age3)
-0.0001
##
(0.001)
##
## I(age4)
0.00000
##
(0.00001)
##
## cutoff:age
0.00003                        -3.141
##
(0.0002)                      (3.524)
##
## cutoff:I(age2)
0.107
##
(0.124)
##
## cutoff:I(age3)
-0.002
##
(0.002)
##
## cutoff:I(age4)
0.00001
##
(0.00001)
##
## Constant                       0.906***                      0.900***
0.990***                       2.348
```

```
##                                 (0.0004)                       (0.001)
(0.005)                       (12.172)
##
## ---------------------------------------------------------------------------
----------------------------------------------------------------
## Observations                1,189,446                      619,039
1,189,446                     1,189,446
## R2                            0.003                          0.001
0.004                           0.004
## Adjusted R2                   0.003                          0.001
0.004                           0.004
## Residual Std. Error      0.313 (df = 1189444)          0.315 (df = 619037)
0.313 (df = 1189442)          0.313 (df = 1189436)
## F Statistic         3,731.028*** (df = 1; 1189444) 884.104*** (df = 1;
619037) 1,432.560*** (df = 3; 1189442) 477.954*** (df = 9; 1189436)
##
==============================================================================
==============================================================
## Note:
*p<0.1; **p<0.05; ***p<0.01
```

```
stargazer(reg1C, reg2C, reg3C, reg4C, type = "text")
```

```
##
##
==============================================================================
==============================================================
##
Dependent variable:
##                       ------------------------------------------------------
----------------------------------------------------------------
##
nonemp
##                                   (1)                            (2)
(3)                           (4)
## ---------------------------------------------------------------------------
----------------------------------------------------------------
## cutoff                        14.574***                      8.245***
5.427*                      -2,540.438
##                                 (0.226)                        (0.314)
(3.206)                     (14,986.970)
##
## age
1.362***                      222.962
##
(0.053)                      (550.304)
##
## I(age2)
-8.977
##
```

```
(23.696)
##
## I(age3)
0.161
##
(0.452)
##
## I(age4)
-0.001
##
(0.003)
##
## cutoff:age
-0.091                       221.593
##
(0.079)                     (1,383.794)
##
## cutoff:I(age2)
-6.944
##
(48.653)
##
## cutoff:I(age3)
0.091
##
(0.776)
##
## cutoff:I(age4)
-0.0004
##
(0.005)
##
## Constant                      125.205***                   128.621***
77.528***                     -1,967.381
##                                (0.154)                        (0.218)
(1.875)                     (4,779.888)
##
## --------------------------------------------------------------------------
----------------------------------------------------------------
## Observations                  1,189,446                      619,039
1,189,446                     1,189,446
## R2                              0.003                          0.001
0.004                         0.004
## Adjusted R2                     0.003                          0.001
0.004                         0.004
## Residual Std. Error     122.907 (df = 1189444)        123.459 (df =
619037)        122.849 (df = 1189442)        122.848 (df = 1189436)
## F Statistic         4,157.099*** (df = 1; 1189444) 689.555*** (df = 1;
619037) 1,763.696*** (df = 3; 1189442) 588.803*** (df = 9; 1189436)
##
```

```
=============================================================================
=============================================================
## Note:
*p<0.1; **p<0.05; ***p<0.01

#reg5A[["coef"]]
#reg5A[["se"]]
#reg5A[["z"]]
#reg5A[["pv"]]
#reg5A[["bws"]]

#reg5B[["coef"]]
#reg5B[["se"]]
#reg5B[["z"]]
#reg5B[["pv"]]
#reg5B[["bws"]]

#reg5C[["coef"]]
#reg5C[["se"]]
#reg5C[["z"]]
#reg5C[["pv"]]
#reg5C[["bws"]]
```