The Impact of Immigration Age on Wages: Insights from 2021 Canadian Census

Lucy Zhu

40752438

ECON 398

December 1st, 2023

Introduction and Background

Many parents choose to immigrate to a foreign country in the hope of providing their children access to better education, and some of them hold a belief that the younger the age of the children when immigrating, the easier the transition will be for them. The belief is in fact founded, especially when it comes to the acquisition of the language spoken in the receiving country: in an influential study conducted by Johnson and Newport (1989), they found that proficiency in English does not vary with age at arrival up to the age of 7, and then it deteriorates in the subsequent years. According to Corak (2012), this is because the early years of human lives are a critical sensitive period during which specific skills, such as speaking a second language, can be developed with the greatest ease.

“Skills beget skills”, possession of skills helps one to acquire more skills in the future, similarly, the early development of skills influences one’s ability to master related skills and hence have an impact on the later stage of their lives. Plenty of papers have studied immigrant children’s performance at school, and how their education outcome is linked to their origin country and parental characteristics (Levels, Dronkers & Kraaykamp, 2008), while few of them explored the connection between wages in adulthood and age at immigration, particularly in the context of Canada, therefore, my paper will contribute to fill this knowledge gap by answering the following question: is there any evidence from the 2021 Canadian census that shows immigration age has an impact on one’s wages?

In my paper, I used simple and multiple linear regressions to evaluate the average treatment effect of immigration age and found that immigration age does affect one's wage. The treatment effect still exists when I divided the sample based on whether they are visible minorities or not in the heterogeneity test.

Data

The data used in my paper come from the 2021 Canadian Census Public Use Microdata File (PUMF), which contains 980,868 records, representing 2.7% of the Canadian population (Statistics Canada, 2021).

Because my paper focuses on the impact of immigration age, that means, Canadian citizens at birth and non-permanent residents should be excluded in my study, hence the analysis sample is the immigrants in Canada.

The outcome variable in my study is wage, corresponding to "wages" in the file, it is encoded as a numeric variable and will be kept as that in my analysis. However, because the magnitude of "wages" is very large relative to the explanatory and control variables in my specifications, a log transformation is performed on "wages" to make the coefficients more interpretable.

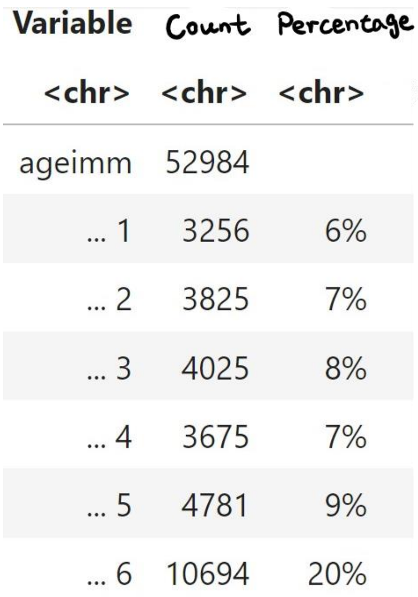
The key explanatory variable in my study is immigration age, corresponding to "ageimm" in the data file. It is encoded as a numeric variable, but according to the user guide, different numbers in "ageimm" represent different levels, for example, "1" means the age at immigration is 0 to 4 years old, hence I used the "mutate" function to make "ageimm" a categorical variable.

I also included "agegrp"(age group), "gender", "pob" (place of birth), "vismin" (visible minority), “kol” (knowledge of language), “hdgree” (highest degree), “locstud” (location of study), “pwpr” (place of work province), and “pkids” (indicates children’s presence in the census family) as my control variables since they are also very likely to have an impact on wage, and including them can reduce omitted variable bias. They are all encoded as numeric variables in the original dataset, similarly, I converted them to categorical variables due to the nature of their type.

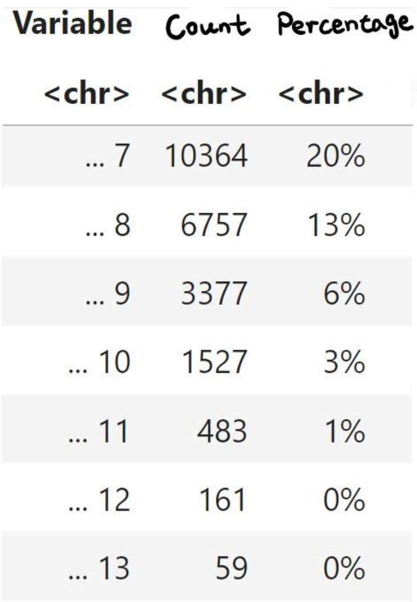
In addition to that, I also transformed “gender” and “pkids” into dummy variables: “gen_dum” takes the value of one if the unit of observation is female and zero if male; “kid_dum” takes the value 1 if there are kids present in the census family and zero if there are not. A third dummy variable, “vm_dum” was made to conduct a heterogeneity test in the extension section, it takes the value of one if the person is a visible minority and zero if they are not.

Summary Statistics

For the key explanatory variable “ageimm”:

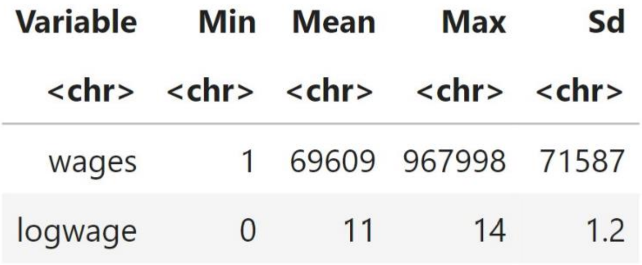


Variable	Count	Percentage
<chr>	<chr>	<chr>
ageimm	52984	
... 1	3256	6%
... 2	3825	7%
... 3	4025	8%
... 4	3675	7%
... 5	4781	9%
... 6	10694	20%



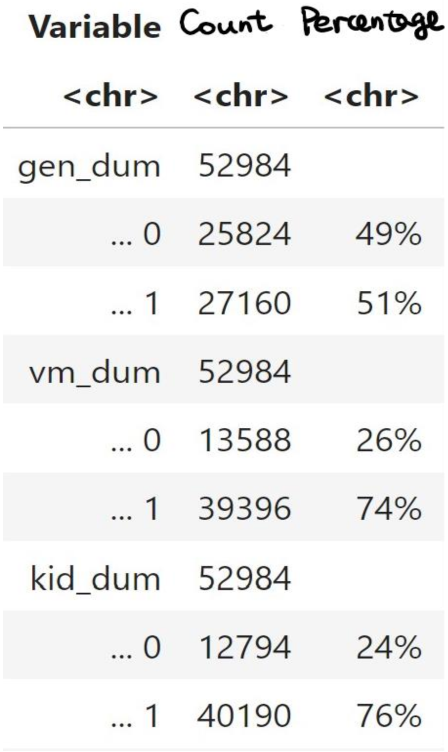
Variable	Count	Percentage
<chr>	<chr>	<chr>
... 7	10364	20%
... 8	6757	13%
... 9	3377	6%
... 10	1527	3%
... 11	483	1%
... 12	161	0%
... 13	59	0%

For the outcome variable “wages” as well as “logwage”:



Variable	Min	Mean	Max	Sd
<chr>	<chr>	<chr>	<chr>	<chr>
wages	1	69609	967998	71587
logwage	0	11	14	1.2

For the dummy variables of interest:



Variable	Count	Percentage
<chr>	<chr>	<chr>
gen_dum	52984	
... 0	25824	49%
... 1	27160	51%
vm_dum	52984	
... 0	13588	26%
... 1	39396	74%
kid_dum	52984	
... 0	12794	24%
... 1	40190	76%

Methodology

The methodology used in my study extends from a simple linear regression to multiple linear regressions.

The outcome variable is “logwage” in all regressions since we want to know how it is affected by the key explanatory variable, “ageimm”.

In the multiple linear regressions, besides the key explanatory variable in the simple linear regression, additional control variables are incrementally added to see how the average treatment effect of "ageimm" would change and to further reduce the omitted variable bias.

In my regressions, the main treatment variable is "ageimm", and the coefficient on it will estimate its average treatment effect on the outcome variable "logwage".

$Y_i = \beta_0 + \beta_1 A_i + \epsilon_\mathrm{i}$ (1)

In the first simple linear regression, I would like to study how much the immigration age (A_i) on its own will affect the outcome variable, "logwage", represented by Y_i in the model. β_1 measures such average treatment effect of A_i.

$Y_i = \beta_0 + \beta_1 A_i + \boldsymbol{\gamma} \boldsymbol{V}_i + \epsilon_i$ (2)

The second specification includes $\boldsymbol{V_i}$, which is a vector of basic demographic controls: gen_dum (dummy for gender), vm_dum(dummy for visible minority), and kid_dum (dummy for the presence of kid in census family). γ represents a vector of coefficients on these control variables. After introducing these controls, I am interested in seeing how β_1 will differentiate from model (1).

$Y_i = \beta_0 + \beta_1 A_i + \boldsymbol{\gamma} \boldsymbol{V}_i + \boldsymbol{\lambda} \boldsymbol{X}_i + \epsilon_i$ (3)

In specification (3), I added more controls: "agegrp" (age group), "pob" (place of birth), and "kol" (knowledge of language), they are encapsulated into the vector $\boldsymbol{X_i}$, and their coefficients are represented by the vector $\boldsymbol{\lambda}$.

$Y_i = \beta_0 + \beta_1 A_i + \boldsymbol{\gamma} \boldsymbol{V}_i + \boldsymbol{\lambda} \boldsymbol{X}_i + \boldsymbol{\alpha} \boldsymbol{Z}_i + \epsilon_i$ (4)

In this final specification, variables that control for "hdgree" (highest degree), "locstud" (location of study), and "pwpr" (place of work province) are added, they are represented by vector $\boldsymbol{Z_i}$, and their coefficients are represented by vector $\boldsymbol{\alpha}$.

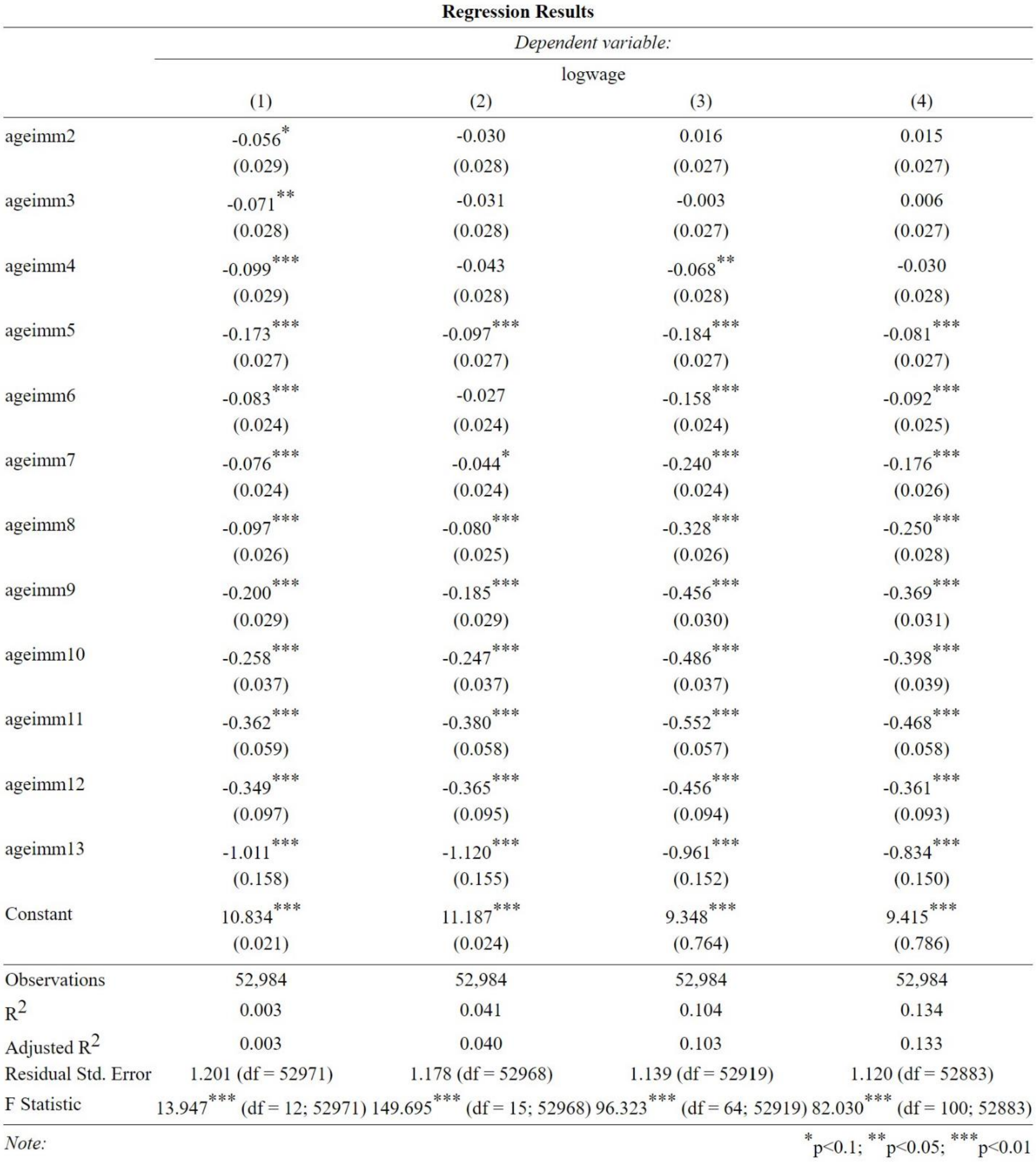
To make β_1 the average treatment effect, the key assumption is, there is no omitted variable bias in my models. For the simple linear regression (specification (1)), it implies the error term is not correlated with variation in A_i, and the expected value of the error term should be 0. For the multiple linear regression models (specifications (2), (3), and (4)), this means that, after adding the control variables, the error term ϵ_i has an expected value of 0 and is not correlated with the variation in the treatment variable A_i.

Alternatively, the regression models are causal when the CEFs they approximate are causal, to make CEFs causal, we need the same condition that makes the (conditional) comparison estimate causal. For specification (1), if the independence assumption is satisfied, meaning the potential outcome of person i is independent of their immigration age (the treatment), then there is no selection bias, so CEF approximated by specification (1) is causal, hence specification (1) is causal. Similarly, for the multiple linear regressions (specifications (2), (3), and (4)), if the conditional independence assumption is satisfied, meaning conditional on the control variables, individuals' potential outcomes are independent of their immigration age, then these regression models should also be causal.

Another noteworthy thing is that I don't have interaction terms, this assumes the treatment effect homogeneity, that is the average treatment effect of A_i does not vary across different values of each control variable.

Results

In all four specifications, the reference level is set to "ageimm1", the group with immigration age from 0 to 4 years old.



Regression Results

	Dependent variable:			
	logwage			
	(1)	(2)	(3)	(4)
ageimm2	-0.056*	-0.030	0.016	0.015
	(0.029)	(0.028)	(0.027)	(0.027)
ageimm3	-0.071**	-0.031	-0.003	0.006
	(0.028)	(0.028)	(0.027)	(0.027)
ageimm4	-0.099***	-0.043	-0.068**	-0.030
	(0.029)	(0.028)	(0.028)	(0.028)
ageimm5	-0.173***	-0.097***	-0.184***	-0.081***
	(0.027)	(0.027)	(0.027)	(0.027)
ageimm6	-0.083***	-0.027	-0.158***	-0.092***
	(0.024)	(0.024)	(0.024)	(0.025)
ageimm7	-0.076***	-0.044*	-0.240***	-0.176***
	(0.024)	(0.024)	(0.024)	(0.026)
ageimm8	-0.097***	-0.080***	-0.328***	-0.250***
	(0.026)	(0.025)	(0.026)	(0.028)
ageimm9	-0.200***	-0.185***	-0.456***	-0.369***
	(0.029)	(0.029)	(0.030)	(0.031)
ageimm10	-0.258***	-0.247***	-0.486***	-0.398***
	(0.037)	(0.037)	(0.037)	(0.039)
ageimm11	-0.362***	-0.380***	-0.552***	-0.468***
	(0.059)	(0.058)	(0.057)	(0.058)
ageimm12	-0.349***	-0.365***	-0.456***	-0.361***
	(0.097)	(0.095)	(0.094)	(0.093)
ageimm13	-1.011***	-1.120***	-0.961***	-0.834***
	(0.158)	(0.155)	(0.152)	(0.150)
Constant	10.834***	11.187***	9.348***	9.415***
	(0.021)	(0.024)	(0.764)	(0.786)
Observations	52,984	52,984	52,984	52,984
R^2	0.003	0.041	0.104	0.134
Adjusted R^2	0.003	0.040	0.103	0.133
Residual Std. Error	1.201 (df = 52971)	1.178 (df = 52968)	1.139 (df = 52919)	1.120 (df = 52883)
F Statistic	13.947*** (df = 12; 52971)	149.695*** (df = 15; 52968)	96.323*** (df = 64; 52919)	82.030*** (df = 100; 52883)
Note:				*p<0.1; **p<0.05; ***p<0.01

In the first specification where we have no controls, all immigration age groups have a negative coefficient. For example, the coefficient of "ageimm2" (5 to 9 years old) is -0.056, that means, moving from “ageimm1” to “ageimm2”, there is a 0.056 unit decrease in the outcome variable “logwage”, in another word, the average treatment effect of “ageimm2”, using “ageimm1” as a reference group, is a -0.056 unit change in “logwage”. In the context of my research question, this means that, compared with being in the 0-to-4-year-old immigration age group, being in the 5-to-9-year-old immigration age group decreases the wage by 5.6%. Similarly, the coefficient of “ageimm13” (60 years and over) is -1.011, which means that compared with being in the 0-to-4-year-old immigration age group, being in the 60-year-old-and-over immigration age group decreases the wage by 101.1%. The same logic applies to the other 3 specifications in which I incrementally added control variables to them.

Interestingly, we have some positive coefficients in specifications (3) and (4). For instance, in specification (3), the coefficient of “ageimm2” suggests that when we control for gender, visible minority status, presence of kids in the census family, age group, place of birth, and knowledge of language, being in the 5-to-9-year-old immigration age group increases wages by 1.6% compared to being in the 0-to-4-year-old immigration age group. Due to the number of immigration age groups in each specification, I will not extensively interpret them one by one.

Alternatively, we can compare the coefficients within and across specifications. I will use specification (4) as an example: when we move from younger to older immigration age groups, the coefficients generally become more negative, this suggests the older the immigration age, the larger the decrease in wage relative to immigrating at 0-4 years old (“ageimm1”). When we compare coefficients between specifications (3) and (4), we can find the average treatment effect of being in each immigration age group is less negative in specification (4), where we added in more control variables.

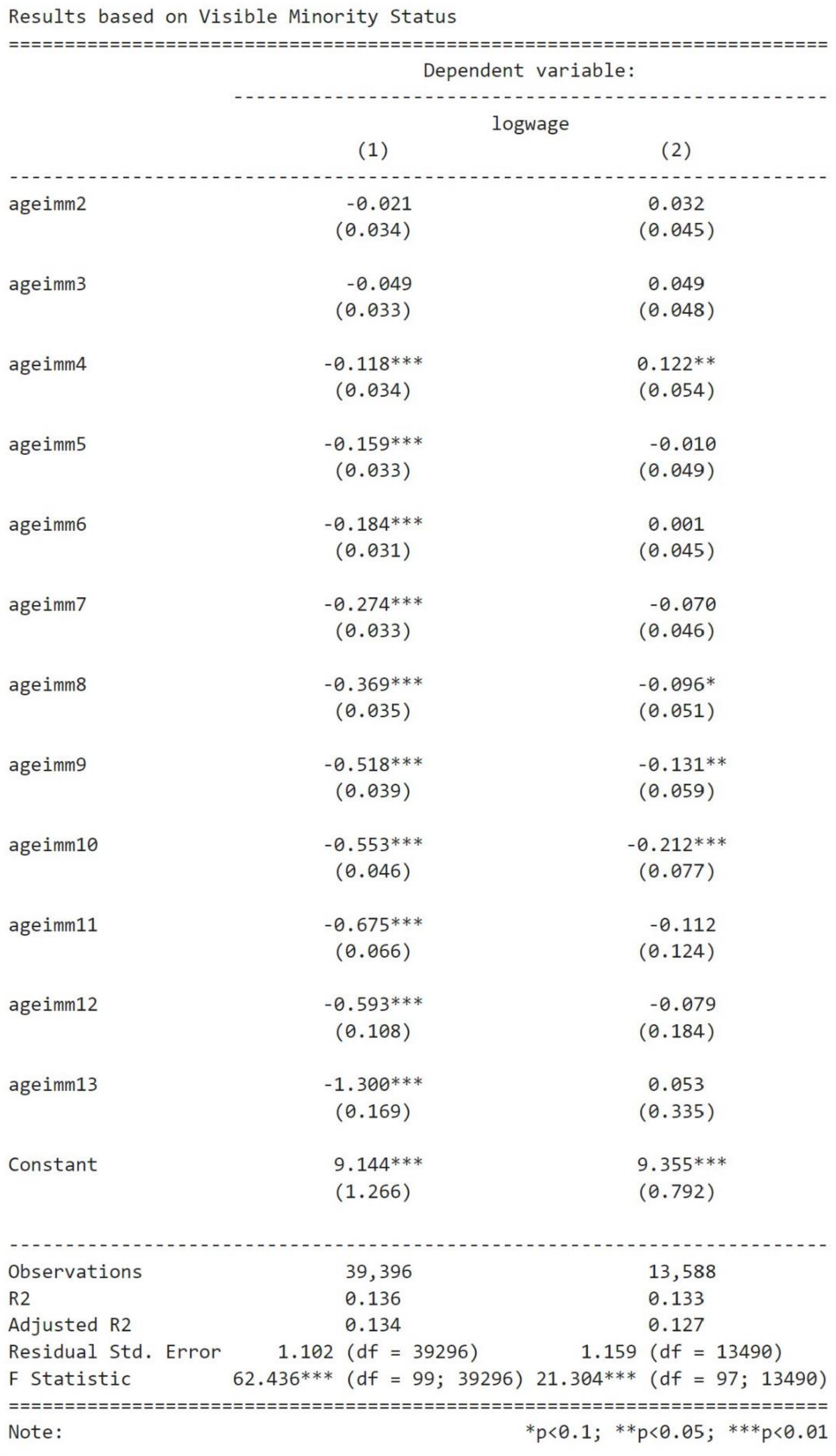
The strength of my multiple linear regression models, especially specification (4), comes from the numerous control variables included in them. By introducing controls, the omitted variable bias should be consequently reduced. One weakness of my results comes from the lack of interaction terms in my regression model because I assumed treatment effect homogeneity to simplify the models. Using the dummy for visible minority as an example, the average treatment effect of being in a particular “ageimm” group relative to being in “ageimm1” is assumed to be the same for both visible minority and non-visible minority, which is hard to believe in real life.

Discussion and Extension

The drawback of the treatment effect homogeneity assumption motivates a heterogeneity test by dividing the sample into subgroups based on their visible minority status, which, as we will see, further solidifies the causal relationship in my analysis.

I chose to use specification (4) in this heterogeneity test, except I excluded “vm_dum” from the control variables.

As we can see in the table below, the treatment effect still exists across all immigration age groups in both subgroups, but the coefficients are different between groups 1 (visible minority) and 2 (non-visible minority). For example, compared to immigrating at age 0 to 4, immigrating at age 5 to 9 decreases the wage by 2.1% for



```
Results based on Visible Minority Status
========================================================================
                                    Dependent variable:
                    ----------------------------------------------------
                                          logwage
                               (1)                        (2)
------------------------------------------------------------------------
ageimm2                       -0.021                     0.032
                             (0.034)                    (0.045)

ageimm3                       -0.049                     0.049
                             (0.033)                    (0.048)

ageimm4                      -0.118***                   0.122**
                             (0.034)                    (0.054)

ageimm5                      -0.159***                   -0.010
                             (0.033)                    (0.049)

ageimm6                      -0.184***                   0.001
                             (0.031)                    (0.045)

ageimm7                      -0.274***                   -0.070
                             (0.033)                    (0.046)

ageimm8                      -0.369***                  -0.096*
                             (0.035)                    (0.051)

ageimm9                      -0.518***                  -0.131**
                             (0.039)                    (0.059)

ageimm10                     -0.553***                 -0.212***
                             (0.046)                    (0.077)

ageimm11                     -0.675***                   -0.112
                             (0.066)                    (0.124)

ageimm12                     -0.593***                   -0.079
                             (0.108)                    (0.184)

ageimm13                     -1.300***                   0.053
                             (0.169)                    (0.335)

Constant                      9.144***                  9.355***
                             (1.266)                    (0.792)

------------------------------------------------------------------------
Observations                  39,396                     13,588
R2                            0.136                      0.133
Adjusted R2                   0.134                      0.127
Residual Std. Error     1.102 (df = 39296)         1.159 (df = 13490)
F Statistic         62.436*** (df = 99; 39296) 21.304*** (df = 97; 13490)
========================================================================
Note:                                        *p<0.1; **p<0.05; ***p<0.01
```

visible minorities but increases the wage by 3.2% for people who are not visible minorities. It is also worth noting that, for the visible minority group, the general trend is older the immigration age, the bigger the decrease in wage (“ageimm12” is an exception), however for the non-visible minority group, the trend seems to be less clear across immigration age groups. Most importantly, the impact of being in a particular immigration age group on wages, relative to being in “ageimm1”, is more negative for visible minorities than non-visible minorities at all immigration age levels.

Albeit the heterogeneity test result contradicts my treatment effect homogeneity assumption, this assumption is generally innocuous in my causal analysis, except that β_1 in my proposed specifications, instead of measuring the average treatment effect, is measuring a weighted average of individual-specific treatment effects.

Conclusion

Using the data from the 2021 Canadian census, I found that immigration age indeed has an impact on wages, and this impact varies across different specifications where different control variables were used. In general, compared with immigrating at age 0 to 4, immigrating after early adolescence has a negative impact on wages, because the signs of coefficients are all negative for “ageimm4” (15-to-19-year-old group) and older groups in all four specifications from the “result” section. By conducting a heterogeneity test, I found that immigration age still has an impact on wages, but the impact is different for people who are visible minorities and those who are not.

One potential issue left unsolved is immigration age might not be totally exogenous, this endogeneity issue can be alleviated by introducing instrumental variables to further assess the average treatment effect of immigration age on wages in future studies.

References

Corak, M. (2012). Age at immigration and the education outcomes of children. *Realizing the potential of immigrant youth*, 90-116.

Johnson, J. S., & Newport, E. L. (1989). Critical period effects in second language learning: The influence of maturational state on the acquisition of English as a second language. *Cognitive Psychology,* 21(1), 60–99.

Levels, M., Dronkers, J., & Kraaykamp, G. (2008). Immigrant Children's Educational Achievement in Western Countries: Origin, Destination, and Community Effects on Mathematical Performance. *American Sociological Review*, 73(5), 835-853.

Statistics Canada. (2023). *2021 Census Public Use Microdata File*. Statistics Canada. https://abacus.library.ubc.ca/dataset.xhtml?persistentId=hdl:11272.1/AB2/1WTDOP