Untitled

December 6, 2023

[1]:
```
library(tidyverse)
library(haven)
library(dplyr)
library(ggplot2)
#install.packages("stargazer")
library(stargazer)
```

```
  Attaching packages                           tidyverse
1.3.2
  ggplot2 3.4.2       purrr   1.0.1
  tibble  3.2.1       dplyr   1.1.1
  tidyr   1.3.0       stringr 1.5.0
  readr   2.1.3       forcats 0.5.2
  Conflicts
tidyverse_conflicts()
  dplyr::filter() masks stats::filter()
  dplyr::lag()    masks stats::lag()

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
```

[2]:
```
"My analysis of the 2021 census data was focused on understanding how education␣
  ↪qualifications influence wage levels among individuals. Through the␣
  ↪examination of the dataset, a clear trend emerged: individuals with higher␣
  ↪education qualifications, such as advanced degrees or specialized␣
  ↪certifications, consistently exhibited higher wages. This observation␣
  ↪underlined a distinct pattern where those with greater educational␣
  ↪achievements tended to command higher wages, providing robust empirical␣
  ↪evidence supporting the positive impact of education qualifications on wage␣
  ↪outcomes within the surveyed population."
```

'My analysis of the 2021 census data was focused on understanding how education qualifications influence wage levels among individuals. Through the examination of the dataset, a clear trend emerged: individuals with higher education qualifications, such as advanced degrees or specialized certifications, consistently exhibited higher wages. This observation underlined a distinct pattern where those with greater educational achievements tended to command higher wages, providing robust empirical evidence supporting the positive impact of education qualifications on wage outcomes within the surveyed population.'

[3]:
```
census_data = read_dta("cen_ind_2021_pumf_v2-2.dta")
glimpse(census_data)
```

```
Rows: 980,868
Columns: 17
$ agegrp  <dbl+lbl> 13, 11, 13, 16, 18, 16, 16, 16, 11, 12,
16, 13, 14, 18,  3…
$ fptwk   <dbl+lbl> 1, 2, 9, 9, 9, 9, 9, 9, 1, 9, 9, 1, 1, 1,
9, 9, 2, 2, 9, 9…
$ Gender  <dbl+lbl> 2, 1, 1, 2, 2, 1, 1, 1, 2, 2, 2, 1, 2, 2,
1, 1, 1, 1, 2, 1…
$ hdgree  <dbl+lbl>  7,  7,  2,  5,  2,  2,  9, 12,  6,  2,
6,  9,  2,  2, 99…
$ immstat <dbl+lbl> 88,  2,  1,  1,  1,  1,  1,  2,  1,  2,
1,  1,  2,  1,  1…
$ lfact   <dbl+lbl> 88,  3, 14, 13, 13, 13, 14, 13,  1,  1,
13,  1,  1,  1, 99…
$ locstud <dbl+lbl> 17, 16, 99,  9, 99, 99,  6, 14,  6, 99,
5,  6, 99, 99, 99…
$ lstwrk  <dbl+lbl> 8, 3, 4, 1, 1, 1, 4, 1, 3, 3, 1, 3, 3, 3,
9, 4, 3, 3, 9, 1…
$ naics   <dbl+lbl>  54,  48, 999, 999, 999, 999, 999, 999,
11,  48, 999,  56…
$ NOC21   <dbl+lbl> 88, 14, 99, 99, 99, 99, 99, 99, 25,  1,
99,  3,  6, 22, 99…
$ relig   <dbl+lbl> 88, 17, 22, 88,  6, 22,  7, 22, 22,  7,
2,  6, 22,  6, 22…
$ ssgrad  <dbl+lbl>  6,  6,  4,  6,  4,  4,  8, 11,  6,  4,
6,  8,  4,  4, 99…
$ TotInc  <dbl> 7.6e+04, 3.2e+04, 1.7e+04, 2.2e+04, 2.2e+04,
1.0e+03, 5.0e+03,…
$ vismin  <dbl+lbl> 88,  2,  1, 88,  1,  1,  1,  1,  1,  1,
1,  1,  7,  1,  1…
$ wkswrk  <dbl+lbl> 8, 5, 9, 9, 9, 9, 9, 9, 2, 0, 9, 6, 6, 6,
9, 9, 6, 3, 9, 9…
$ wrkact  <dbl+lbl> 88, 10,  1,  1,  1,  1,  1,  1,  3,  2,
1, 11, 11, 11, 99…
$ Wages   <dbl> 7.6e+04, 1.0e+08, 1.0e+08, 1.0e+08, 1.0e+08,
1.0e+08, 1.0e+08,…
```

[4]:

```
selected_vars <- c(
  gender = "Gender",
  visible_minority = "vismin",
  religion = "relig",
  wages = "Wages",
  college_education = "hdgree",
  age_group = "agegrp",
  Secondary_high_school_diploma_or_equivalency_certificate = "ssgrad",
  Immigration_status = "immstat",
  Full_time_or_part_weeks_worked = "fptwk",
  Occupation_major_group = "NOC21"
)
```

[5]:

```
"In my analysis, the independent variable of interest is education level, while␣
 ↪the dependent variable being investigated is wage level. To ensure a␣
 ↪comprehensive examination, several control variables, including age, gender,␣
 ↪religion, immigration status, and visible minority status, were incorporated.
 ↪ These control variables were considered crucial factors that might␣
 ↪influence wage levels alongside education, allowing for a more nuanced␣
 ↪exploration of the relationship between education level and wage outcomes␣
 ↪while accounting for various demographic and socio-economic characteristics."
```

'In my analysis, the independent variable of interest is education level, while the dependent variable being investigated is wage level. To ensure a comprehensive examination, several control variables, including age, gender, religion, immigration status, and visible minority status, were incorporated. These control variables were considered crucial factors that might influence wage levels alongside education, allowing for a more nuanced exploration of the relationship between education level and wage outcomes while accounting for various demographic and socio-economic characteristics.'

[6]:

```
census_data <- census_data %>%
  select(all_of(selected_vars))
glimpse(census_data)
```

```
Rows: 980,868
Columns: 10
$ gender
<dbl+lbl> 2, 1, 1, …
$ visible_minority
<dbl+lbl> 88,  2,  …
$ religion
<dbl+lbl> 88, 17, 2…
$ wages
<dbl> 7.6e+04, 1.0e…
$ college_education
<dbl+lbl>  7,  7,  …
$ age_group
<dbl+lbl> 13, 11, 1…
$ Secondary_high_school_diploma_or_equivalency_certificate
```

```
<dbl+lbl>  6,  6,  …
$ Immigration_status
<dbl+lbl> 88,  2,  …
$ Full_time_or_part_weeks_worked
<dbl+lbl> 1, 2, 9, …
$ Occupation_major_group
<dbl+lbl> 88, 14, 9…
```

[7]:
```
"Here, the treatment group can be conceptualized as individuals across various␣
↪education levels (ranging from different degrees of education) within your␣
↪dataset. I am examining how different levels of education influence wage␣
↪levels among individuals, with education level being the primary variable of␣
↪interest.  While there isn't a distinct control group receiving no treatment␣
↪or intervention, I am considering individuals with advanced education levels␣
↪as a reference or baseline group for comparison against those with lower␣
↪education levels."
```

'Here, the treatment group can be conceptualized as individuals across various education levels (ranging from different degrees of education) within your dataset. I am examining how different levels of education influence wage levels among individuals, with education level being the primary variable of interest. While there isn\'t a distinct control group receiving no treatment or intervention, I am considering individuals with advanced education levels as a reference or baseline group for comparison against those with lower education levels.'

[8]:
```
census_data <- census_data %>% filter(college_education != 99)
census_data <- census_data %>% filter(college_education != 88)
census_data <- census_data %>% filter(wages != 99999999)
census_data <- census_data %>% filter(wages != 88888888)
```

[9]:
```
census_data <- census_data %>% mutate( degree_modified = case_when(␣
↪college_education == 1 ~ 'No_certificate_diploma_or_degree',␣
↪college_education == 2 ~ 'Secondary_Education', college_education == 3 ~␣
↪'Secondary_Education', college_education == 3 ~ 'Secondary_Education',␣
↪college_education == 4 ~ 'Secondary_Education', college_education == 5 ~␣
↪'Post_Secondary_Education', college_education == 6 ~␣
↪'Post_Secondary_Education', college_education == 7 ~␣
↪'Post_Secondary_Education', college_education == 8 ~␣
↪'Post_Secondary_Education', college_education == 9 ~␣
↪'Post_Secondary_Education', college_education == 10 ~␣
↪'Post_Secondary_Education', college_education == 11 ~ 'Advanced_Degrees',␣
↪college_education == 12 ~ 'Advanced_Degrees', college_education == 13 ~␣
↪'Advanced_Degrees'))
```

[10]:

```
"The education groupings encompass diverse academic achievements, from␣
↪individuals without any formal certification to those with advanced degrees.␣
↪It spans individuals who have not received any formal certificate or␣
↪diploma, covers secondary education such as high school diplomas, extends to␣
↪post-secondary education comprising short-term programs to university␣
↪degrees, and culminates in advanced degrees, including doctorates and␣
↪specialized fields like medicine. This classification system serves to␣
↪organize educational attainments comprehensively, aiding both institutions␣
↪in evaluating qualifications and individuals in presenting their educational␣
↪backgrounds concisely."
```

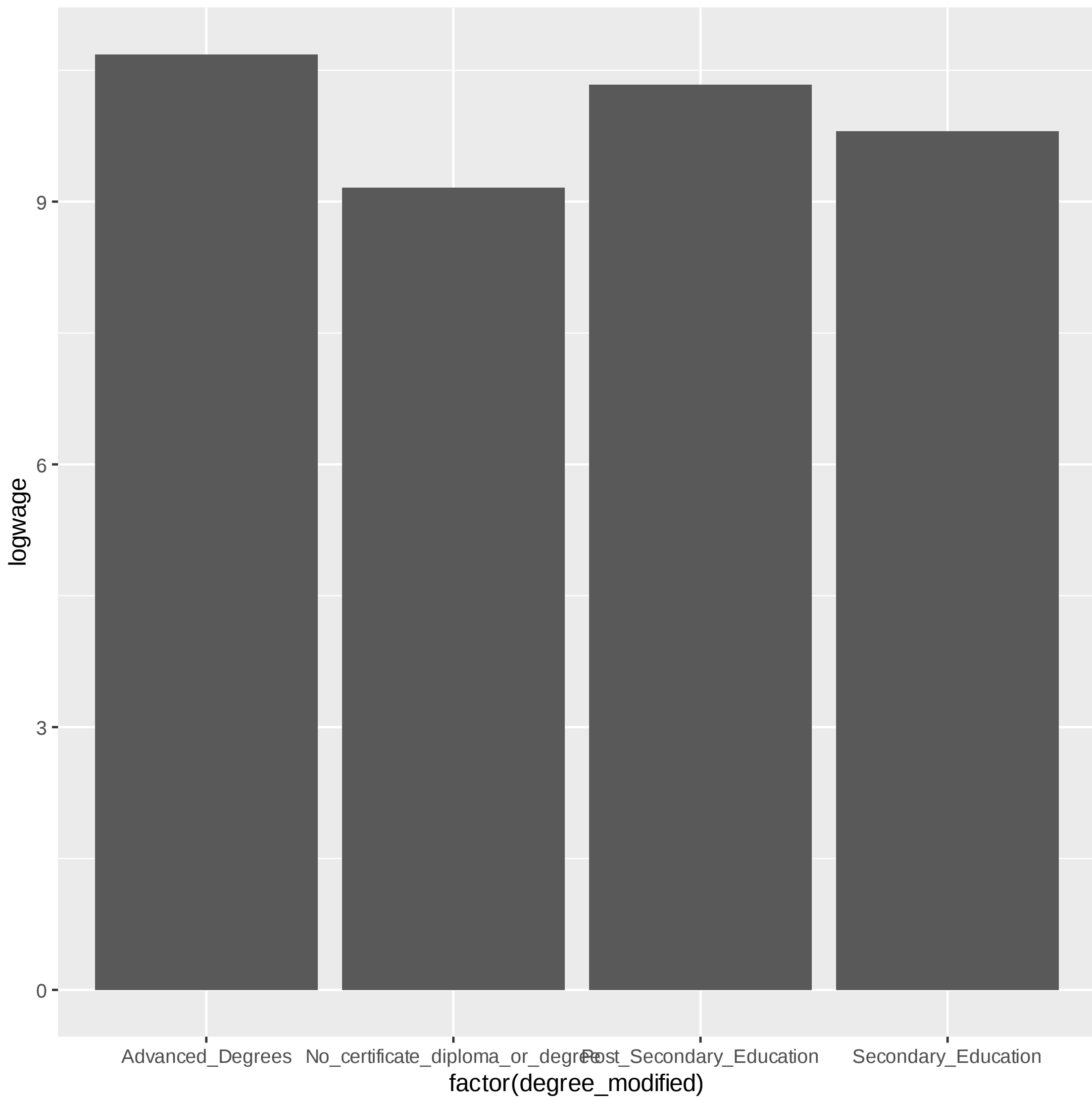
'The education groupings encompass diverse academic achievements, from individuals without any formal certification to those with advanced degrees. It spans individuals who have not received any formal certificate or diploma, covers secondary education such as high school diplomas, extends to post-secondary education comprising short-term programs to university degrees, and culminates in advanced degrees, including doctorates and specialized fields like medicine. This classification system serves to organize educational attainments comprehensively, aiding both institutions in evaluating qualifications and individuals in presenting their educational backgrounds concisely.'

[11]:
```
census_data <- census_data %>% mutate(logwage = log(wages)  )
```

[12]:
```
"Utilizing log transformations on wage data when studying the correlation␣
↪between education and wages is advantageous for statistical analysis. These␣
↪transformations normalize skewed wage distributions, establish a more linear␣
↪relationship between education and wages, reduce variance across educational␣
↪levels, and enhance the interpretability of wage changes concerning␣
↪educational attainment. Overall, employing log transformations aids in␣
↪addressing data irregularities and gaining clearer insights into the␣
↪education-wage relationship."
```

'Utilizing log transformations on wage data when studying the correlation between education and wages is advantageous for statistical analysis. These transformations normalize skewed wage distributions, establish a more linear relationship between education and wages, reduce variance across educational levels, and enhance the interpretability of wage changes concerning educational attainment. Overall, employing log transformations aids in addressing data irregularities and gaining clearer insights into the education-wage relationship.'

[13]:
```
ggplot(census_data, aes(x = factor(degree_modified), y=logwage)) +␣
↪geom_bar(stat = "summary", fun = "mean")
```



[14]:

```
summary_stats <- census_data %>%
  filter(wages != 99999999) %>%
  filter(wages != 88888888) %>%
  group_by(degree_modified) %>%
  summarize(max_wages = max(wages), mean_wages = mean(wages, na.rm = TRUE))



print(summary_stats)
```

```
# A tibble: 4 × 3
  degree_modified                  max_wages mean_wages
  <chr>
```

```
<dbl>      <dbl>
1 Advanced_Degrees
967998     87217.
2 No_certificate_diploma_or_degree
967998     27693.
3 Post_Secondary_Education
967998     59585.
4 Secondary_Education
967998     40587.
```

[15]:
```
"The data suggests a clear correlation between higher educational attainment␣
↪and increased earning potential. Individuals with advanced degrees tend to␣
↪command the highest average wages, followed by those with post-secondary␣
↪education. Conversely, individuals without certificates, diplomas, or␣
↪degrees and those with only secondary education exhibit notably lower mean␣
↪wages. This highlights the pivotal role of education in shaping income␣
↪levels, indicating that higher levels of education often translate to higher␣
↪earning capacities in the workforce."
```

'The data suggests a clear correlation between higher educational attainment and increased earning potential. Individuals with advanced degrees tend to command the highest average wages, followed by those with post-secondary education. Conversely, individuals without certificates, diplomas, or degrees and those with only secondary education exhibit notably lower mean wages. This highlights the pivotal role of education in shaping income levels, indicating that higher levels of education often translate to higher earning capacities in the workforce.'

[16]:
```
summary_stats_by_group <- census_data %>%
  group_by(degree_modified) %>%
  summarise(
    min_wages = min(wages, na.rm = TRUE),
    max_wages = max(wages, na.rm = TRUE),
    sd_wages = sd(wages, na.rm = TRUE)
  )

print(summary_stats_by_group)
```

```
# A tibble: 4 × 4
  degree_modified                    min_wages max_wages sd_wages
  <chr>
<dbl>     <dbl>
<dbl>
1 Advanced_Degrees                           1
967998    98007.
2 No_certificate_diploma_or_degree           1
967998    34499.
3 Post_Secondary_Education                   1
967998    65102.
4 Secondary_Education                        1
```

967998   44268.

[17]:
```
"The table depicting wage statistics categorized by different education levels␣
↪demonstrates consistency in minimum and maximum wages across all educational␣
↪categories. However, the notable variance lies in the standard deviation␣
↪(sd_wages) within these categories. Higher standard deviations in wages are␣
↪evident among individuals with Advanced Degrees and those in Post-Secondary␣
↪Education, indicating greater variability in earnings within these groups.␣
↪Conversely, individuals without formal certificates or diplomas and those␣
↪with Secondary Education exhibit lower variability in wages. This data␣
↪underscores the significant diversity in earnings distribution among various␣
↪educational levels. The higher standard deviations in Advanced Degrees and␣
↪Post-Secondary Education imply a wider range of earnings, suggesting a more␣
↪extensive spectrum of income potential among individuals with higher␣
↪educational attainment compared to those with lower education levels."
```

'The table depicting wage statistics categorized by different education levels demonstrates consistency in minimum and maximum wages across all educational categories. However, the notable variance lies in the standard deviation (sd_wages) within these categories. Higher standard deviations in wages are evident among individuals with Advanced Degrees and those in Post-Secondary Education, indicating greater variability in earnings within these groups. Conversely, individuals without formal certificates or diplomas and those with Secondary Education exhibit lower variability in wages. This data underscores the significant diversity in earnings distribution among various educational levels. The higher standard deviations in Advanced Degrees and Post-Secondary Education imply a wider range of earnings, suggesting a more extensive spectrum of income potential among individuals with higher educational attainment compared to those with lower education levels.'

[18]:
```
"No we come to the regression analysis of this project. Analyzing without␣
↪control variables initially helps isolate and understand the direct impact␣
↪of specific factors, like education levels, on the outcome variable, here␣
↪log wages. It allows for assessing the independent influence of education on␣
↪wages before considering other potential factors that might affect this␣
↪relationship. Later, introducing control variables helps capture a more␣
↪comprehensive picture of how various factors collectively influence wages."
```

'No we come to the regression analysis of this project. Analyzing without control variables initially helps isolate and understand the direct impact of specific factors, like education levels, on the outcome variable, here log wages. It allows for assessing the independent influence of education on wages before considering other potential factors that might affect this relationship. Later, introducing control variables helps capture a more comprehensive picture of how various factors collectively influence wages.'

[19]:
```
regression1 = lm(data = census_data, logwage ~ degree_modified)
summary(regression1)
```

```
Call:
```

```
lm(formula = logwage ~ degree_modified, data = census_data)

Residuals:
     Min       1Q   Median       3Q      Max 
-10.6807  -0.4611   0.5028   1.0596   4.6224 

Coefficients:
                                                Estimate Std. Error t value
(Intercept)                                    10.680661   0.009638 1108.22
degree_modifiedNo_certificate_diploma_or_degree -1.520093   0.013014 -116.81
degree_modifiedPost_Secondary_Education        -0.343910   0.010447  -32.92
degree_modifiedSecondary_Education             -0.878118   0.010694  -82.11
                                               Pr(>|t|)    
(Intercept)                                      <2e-16 ***
degree_modifiedNo_certificate_diploma_or_degree   <2e-16 ***
degree_modifiedPost_Secondary_Education          <2e-16 ***
degree_modifiedSecondary_Education               <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.983 on 518725 degrees of freedom
Multiple R-squared:  0.04161,	Adjusted R-squared:  0.04161 
F-statistic:  7507 on 3 and 518725 DF,  p-value: < 2.2e-16
```

[20]:

```
"The regression model analyzes the relationship between log wages and different educational categories without considering any control variables. Heres a summary of the regression analysis. The model explores the association between log wages and various educational categories: No certificate diploma or degree, Post Secondary Education, and Secondary Education. Each educational category coefficient signifies the difference in log wages compared to the reference category (Advanced Degrees). Intercept(advanced degrees)The intercept stands at 10.6807, indicating the log wages for individuals in the reference category (Advanced Degrees). Coefficients for each educational category: No certificate diploma or degree: This category is associated with an average decrease of approximately 1.52 in log wages compared to Advanced Degrees. Post secondary education:Individuals in this category have an average decrease of around 0.34 in log wages compared to Advanced Degrees. Secondary education: This category experiences an average decrease of roughly 0.88 in log wages compared to advanced degrees. Each coefficient demostrates statistical significance (p<0.001) suggesting that all educational categories significantly relates to log wages compared to the reference category. The adjusted R-squared pf 0.0416 indicates that approximatley 4.16% of the variation in log wages can be explained by these educational categories in the absence of control varibales.The regression model provides insights into how different educational levels correlate with log wages independently of other factors."
```

'The regression model analyzes the relationship between log wages and different educational categories without considering any control variables. Heres a summary of the regression analysis. The model explores the association between log wages and various educational categories: No certificate diploma or degree, Post Secondary Education, and Secondary Education. Each educational category coefficient signifies the difference in log wages compared to the reference category (Advanced Degrees). Intercept(advanced degrees)The intercept stands at 10.6807, indicating the log wages for individuals in the reference category (Advanced Degrees). Coefficients for each educational category: No certificate diploma or degree: This category is associated with an average decrease of approximately 1.52 in log wages compared to Advanced Degrees. Post secondary education:Individuals in this category have an average decrease of around 0.34 in log wages compared to Advanced Degrees. Secondary education: This category experiences an average decrease of roughly 0.88 in log wages compared to advanced degrees. Each coefficient demostrates statistical significance ($p<0.001$) suggesting that all educational categories significantly relates to log wages compared to the reference category. The adjusted R-squared pf 0.0416 indicates that approximatley 4.16% of the variation in log wages can be explained by these educational categories in the absence of control varibales.The regression model provides insights into how different educational levels correlate with log wages independently of other factors.'

[21]:
```
regression2 = lm(data = census_data, logwage ~ degree_modified + age_group)
summary(regression2)
```

```
Call:
lm(formula = logwage ~ degree_modified + age_group, data = census_data)

Residuals:
     Min       1Q   Median       3Q      Max
-10.7206  -0.4655   0.4973   1.0653   5.2307

Coefficients:
                                                Estimate Std. Error t value
(Intercept)                                   10.7892873  0.0114212  944.67
degree_modifiedNo_certificate_diploma_or_degree -1.5224720  0.0130103 -117.02
degree_modifiedPost_Secondary_Education       -0.3465899  0.0104452  -33.18
degree_modifiedSecondary_Education            -0.8807468  0.0106922  -82.37
age_group                                     -0.0085867  0.0004848  -17.71
                                              Pr(>|t|)
(Intercept)                                     <2e-16 ***
degree_modifiedNo_certificate_diploma_or_degree   <2e-16 ***
degree_modifiedPost_Secondary_Education         <2e-16 ***
degree_modifiedSecondary_Education              <2e-16 ***
age_group                                       <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.983 on 518724 degrees of freedom
Multiple R-squared:  0.04219,	Adjusted R-squared:  0.04218
F-statistic:  5712 on 4 and 518724 DF,  p-value: < 2.2e-16
```

[22]:
```
"The regression model (regression2) examines the association between log wages␣
↪and educational categories, incorporating age groups as a control variable.␣
↪The results reveal notable findings: educational levels beyond Advanced␣
↪Degrees show reduced log wages. Moreover, for each unit increase in age␣
↪(regardless of the age unit in the dataset), theres an average decrease of␣
↪approximately 0.0086 in log wages, holding educational level constant. All␣
↪coefficients exhibit high statistical significance (p < 0.001), indicating␣
↪their impact on log wages. However, the adjusted R-squared value of 0.04218␣
↪suggests that only about 4.22% of the variation in log wages is explained by␣
↪both education and age in this model. Age serves as a critical control␣
↪variable in wage analyses due to its substantial influence on earnings. As␣
↪individuals advance in their careers, their wages often increase due to␣
↪experience and seniority. Incorporating age as a control helps consider how␣
↪age-related changes in earnings might affect the link between education␣
↪levels and wages. By accounting for age, the analysis better isolates and␣
↪evaluates the specific impact of education on wages while minimizing␣
↪potential confounding effects stemming from age-related wage trends. As␣
↪people age, retirement wages and age-specific unemployment benefits might␣
↪affect earnings independently of educational qualifications. Introducing age␣
↪as a control variable helps isolate the true impact of education on wages by␣
↪considering these age-related financial aspects, ensuring a more accurate␣
↪evaluation of the relationship between education levels and income."
```

'The regression model (regression2) examines the association between log wages and educational categories, incorporating age groups as a control variable. The results reveal notable findings: educational levels beyond Advanced Degrees show reduced log wages. Moreover, for each unit increase in age (regardless of the age unit in the dataset), theres an average decrease of approximately 0.0086 in log wages, holding educational level constant. All coefficients exhibit high statistical significance ($p < 0.001$), indicating their impact on log wages. However, the adjusted R-squared value of 0.04218 suggests that only about 4.22% of the variation in log wages is explained by both education and age in this model. Age serves as a critical control variable in wage analyses due to its substantial influence on earnings. As individuals advance in their careers, their wages often increase due to experience and seniority. Incorporating age as a control helps consider how age-related changes in earnings might affect the link between education levels and wages. By accounting for age, the analysis better isolates and evaluates the specific impact of education on wages while minimizing potential confounding effects stemming from age-related wage trends. As people age, retirement wages and age-specific unemployment benefits might affect earnings independently of educational qualifications. Introducing age as a control variable helps isolate the true impact of education on wages by considering these age-related financial aspects, ensuring a more accurate evaluation of the relationship between education levels and income.'

[23]:
```
regression3 = lm(data = census_data, logwage ~ degree_modified + age_group +␣
↪Immigration_status)
summary(regression3)
```

```
Call:
lm(formula = logwage ~ degree_modified + age_group + Immigration_status,
    data = census_data)

Residuals:
     Min       1Q   Median       3Q      Max
-10.7230  -0.4649   0.4973   1.0644   5.2301

Coefficients:
                                                Estimate Std. Error  t value
(Intercept)                                   10.7949228  0.0114918  939.357
degree_modifiedNo_certificate_diploma_or_degree -1.5236569  0.0130129 -117.088
degree_modifiedPost_Secondary_Education       -0.3474166  0.0104467  -33.256
degree_modifiedSecondary_Education            -0.8821015  0.0106964  -82.467
age_group                                     -0.0085489  0.0004849  -17.631
Immigration_status                            -0.0035717  0.0008074   -4.424
                                              Pr(>|t|)
(Intercept)                                    < 2e-16 ***
degree_modifiedNo_certificate_diploma_or_degree  < 2e-16 ***
degree_modifiedPost_Secondary_Education        < 2e-16 ***
degree_modifiedSecondary_Education             < 2e-16 ***
age_group                                      < 2e-16 ***
Immigration_status                            9.69e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.983 on 518723 degrees of freedom
Multiple R-squared:  0.04223,	Adjusted R-squared:  0.04222
F-statistic:  4574 on 5 and 518723 DF,  p-value: < 2.2e-16
```

[24]:

```
"The regression analysis (regression3) includes immigration status alongside␣
 ↪age groups as additional control variables in examining the relationship␣
 ↪between log wages and educational categories. The results show that, after␣
 ↪considering education and age, immigration status still plays a role, as␣
 ↪evidenced by its statistically significant coefficient. The coefficients for␣
 ↪educational categories and age groups remain consistent with previous␣
 ↪findings, demonstrating a significant impact on log wages. The adjusted␣
 ↪R-squared of 0.04222 suggests that around 4.22% of the variation in log␣
 ↪wages can be explained by these factors collectively in this model. Overall,␣
 ↪immigration status, along with education and age, contributes to␣
 ↪understanding the dynamics of wages, emphasizing its relevance in␣
 ↪determining earnings.An example is , skilled immigrants arriving in Canada␣
 ↪often take low-income jobs initially, despite their high education levels,␣
 ↪aiming to gain Canadian work experience for Permanent Residency (PR). They␣
 ↪may work in roles like driving, retail, or customer service to accumulate␣
 ↪local experience needed for better jobs. When analyzing education levels and␣
 ↪wages, these choices can distort income levels compared to qualifications.␣
 ↪Including immigration status in the analysis helps consider how seeking␣
 ↪Canadian experience affects earnings, offering a clearer view of the impact␣
 ↪of education on wages among immigrants."
```

'The regression analysis (regression3) includes immigration status alongside age groups as additional control variables in examining the relationship between log wages and educational categories. The results show that, after considering education and age, immigration status still plays a role, as evidenced by its statistically significant coefficient. The coefficients for educational categories and age groups remain consistent with previous findings, demonstrating a significant impact on log wages. The adjusted R-squared of 0.04222 suggests that around 4.22% of the variation in log wages can be explained by these factors collectively in this model. Overall, immigration status, along with education and age, contributes to understanding the dynamics of wages, emphasizing its relevance in determining earnings.An example is , skilled immigrants arriving in Canada often take low-income jobs initially, despite their high education levels, aiming to gain Canadian work experience for Permanent Residency (PR). They may work in roles like driving, retail, or customer service to accumulate local experience needed for better jobs. When analyzing education levels and wages, these choices can distort income levels compared to qualifications. Including immigration status in the analysis helps consider how seeking Canadian experience affects earnings, offering a clearer view of the impact of education on wages among immigrants.'

[25]:
```
regression4 = lm(data = census_data, logwage ~ degree_modified + age_group +␣
 ↪Immigration_status + gender)
summary(regression4)
```

```
Call:
lm(formula = logwage ~ degree_modified + age_group + Immigration_status +
    gender, data = census_data)

Residuals:
     Min       1Q   Median       3Q      Max
```

```
-10.8996  -0.4417   0.5049   1.0594   5.1211

Coefficients:
                                               Estimate Std. Error  t value
(Intercept)                                  10.2866838  0.0140784  730.671
degree_modifiedNo_certificate_diploma_or_degree -1.5619424  0.0129796 -120.339
degree_modifiedPost_Secondary_Education      -0.3313280  0.0104114  -31.824
degree_modifiedSecondary_Education           -0.9151867  0.0106703  -85.770
age_group                                    -0.0089173  0.0004831  -18.457
Immigration_status                           -0.0035156  0.0008044   -4.371
gender                                        0.3439047  0.0055433   62.039
                                             Pr(>|t|)
(Intercept)                                   < 2e-16 ***
degree_modifiedNo_certificate_diploma_or_degree  < 2e-16 ***
degree_modifiedPost_Secondary_Education       < 2e-16 ***
degree_modifiedSecondary_Education            < 2e-16 ***
age_group                                     < 2e-16 ***
Immigration_status                           1.24e-05 ***
gender                                        < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.975 on 518722 degrees of freedom
Multiple R-squared:  0.04928,	Adjusted R-squared:  0.04927
F-statistic:  4481 on 6 and 518722 DF,  p-value: < 2.2e-16
```

[26]:

```
"In regression4, the model explores the relationship between log wages and␣
↪educational categories alongside age groups, immigration status, and gender␣
↪as control variables. The results reveal that gender, when considered␣
↪alongside education, age, and immigration status, shows a significant impact␣
↪on log wages. The coefficients for educational categories, age groups, and␣
↪immigration status remain consistent with previous findings, continuing to␣
↪influence log wages significantly. The adjusted R-squared of 0.04927␣
↪suggests that around 4.927% of the variation in log wages is explained by␣
↪these combined factors in this model. Including gender alongside education,␣
↪age, and immigration status provides a more comprehensive understanding of␣
↪the wage dynamics, indicating its relevance in determining earnings.Wage␣
↪disparities between men and women, despite possessing equivalent educational␣
↪qualifications and performing the same job, are indicative of persistent␣
↪gender-based inequality prevalent in various sectors. Numerous studies and␣
↪statistical analyses have consistently shown that women, even with similar␣
↪educational attainment and job roles, often face lower wages compared to␣
↪their male counterparts. This wage discrepancy is a consequence of systemic␣
↪biases and gender-based pay gaps entrenched in certain industries or␣
↪workplaces. Therefore, the existence of wage differentials between men and␣
↪women despite identical educational levels is a clear demonstration of␣
↪gender-based inequality prevailing in the workforce"
```

'In regression4, the model explores the relationship between log wages and educational categories alongside age groups, immigration status, and gender as control variables. The results reveal that gender, when considered alongside education, age, and immigration status, shows a significant impact on log wages. The coefficients for educational categories, age groups, and immigration status remain consistent with previous findings, continuing to influence log wages significantly. The adjusted R-squared of 0.04927 suggests that around 4.927% of the variation in log wages is explained by these combined factors in this model. Including gender alongside education, age, and immigration status provides a more comprehensive understanding of the wage dynamics, indicating its relevance in determining earnings.Wage disparities between men and women, despite possessing equivalent educational qualifications and performing the same job, are indicative of persistent gender-based inequality prevalent in various sectors. Numerous studies and statistical analyses have consistently shown that women, even with similar educational attainment and job roles, often face lower wages compared to their male counterparts. This wage discrepancy is a consequence of systemic biases and gender-based pay gaps entrenched in certain industries or workplaces. Therefore, the existence of wage differentials between men and women despite identical educational levels is a clear demonstration of gender-based inequality prevailing in the workforce'

[27]:
```
regression5 = lm(data = census_data, logwage ~ degree_modified + age_group␣
↪+Immigration_status + gender+ visible_minority)
summary(regression5)
```

```
Call:
lm(formula = logwage ~ degree_modified + age_group + Immigration_status +
    gender + visible_minority, data = census_data)
```

```
Residuals:
     Min       1Q   Median       3Q      Max
-10.9068  -0.4435   0.5043   1.0588   5.1067

Coefficients:
                                              Estimate Std. Error  t value
(Intercept)                                 10.2941566  0.0141010  730.032
degree_modifiedNo_certificate_diploma_or_degree -1.5650303  0.0129829 -120.545
degree_modifiedPost_Secondary_Education     -0.3330187  0.0104122  -31.984
degree_modifiedSecondary_Education          -0.9182271  0.0106746  -86.020
age_group                                   -0.0088022  0.0004833  -18.214
Immigration_status                          -0.0024043  0.0008134   -2.956
gender                                       0.3435649  0.0055430   61.982
visible_minority                            -0.0016270  0.0001779   -9.147
                                            Pr(>|t|)
(Intercept)                                  < 2e-16 ***
degree_modifiedNo_certificate_diploma_or_degree  < 2e-16 ***
degree_modifiedPost_Secondary_Education      < 2e-16 ***
degree_modifiedSecondary_Education           < 2e-16 ***
age_group                                    < 2e-16 ***
Immigration_status                           0.00312 **
gender                                       < 2e-16 ***
visible_minority                             < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.975 on 518721 degrees of freedom
Multiple R-squared:  0.04943,	Adjusted R-squared:  0.04942
F-statistic:  3854 on 7 and 518721 DF,  p-value: < 2.2e-16
```

[28]:

```
"The latest regression analysis, regression5, delves into the association
 between log wages and multiple factors, now incorporating 'visible minority'
 alongside degree_modified, age_group, Immigration status, and gender. The
 coefficient for 'visible minority' (-0.0016) denotes a reduction in log
 wages linked to being part of a visible minority, even when considering
 education, age, immigration status, and gender. Significant coefficients for
 education, age, immigration status, and gender maintain their implications
 similar to the previous model. This suggests that visible minority status
 contributes to wage differences, shedding light on disparities faced by
 these groups. The adjusted R-squared value of 0.04942 implies that
 approximately 4.942% of the log wage variation is explained by this combined
 model. Additionally, among university-educated individuals in Canada who are
 Canadian-born and belong to a visible minority, their average earnings are
 approximately 87.4 cents for every dollar earned by individuals of Caucasian
 ethnicity with similar education levels, indicating an income disparity
 despite similar educational qualifications."
```

'The latest regression analysis, regression5, delves into the association between log wages and multiple factors, now incorporating \'visible minority\' alongside degree_modified, age_group, Immigration status, and gender. The coefficient for \'visible minority\' (-0.0016) denotes a reduction in log wages linked to being part of a visible minority, even when considering education, age, immigration status, and gender. Significant coefficients for education, age, immigration status, and gender maintain their implications similar to the previous model. This suggests that visible minority status contributes to wage differences, shedding light on disparities faced by these groups. The adjusted R-squared value of 0.04942 implies that approximately 4.942% of the log wage variation is explained by this combined model. Additionally, among university-educated individuals in Canada who are Canadian-born and belong to a visible minority, their average earnings are approximately 87.4 cents for every dollar earned by individuals of Caucasian ethnicity with similar education levels, indicating an income disparity despite similar educational qualifications.'

[29]:
```
stargazer(regression1, regression2, regression3, regression4, regression5,␣
↪title="Comparison of Regression Results",
          align = TRUE, type="text", keep.stat = c("n","rsq"))
```

```
Comparison of Regression Results
================================================================================
=================
                                                                Dependent
variable:
-------------------------------------------------
                                                                      logwage
                                                    (1)       (2)       (3)
(4)       (5)
--------------------------------------------------------------------------------
-----------------
degree_modifiedNo_certificate_diploma_or_degree -1.520*** -1.522*** -1.524***
-1.562*** -1.565***
                                                  (0.013)   (0.013)   (0.013)
(0.013)   (0.013)

degree_modifiedPost_Secondary_Education         -0.344*** -0.347*** -0.347***
-0.331*** -0.333***
                                                  (0.010)   (0.010)   (0.010)
(0.010)   (0.010)

degree_modifiedSecondary_Education              -0.878*** -0.881*** -0.882***
-0.915*** -0.918***
                                                  (0.011)   (0.011)   (0.011)
(0.011)   (0.011)

age_group                                                 -0.009*** -0.009***
-0.009*** -0.009***
                                                           (0.0005)  (0.0005)
(0.0005)  (0.0005)
```

```
Immigration_status                                                     -0.004***
-0.004*** -0.002***
                                                                        (0.001)
(0.001)   (0.001)

gender
0.344***  0.344***
(0.006)   (0.006)

visible_minority
-0.002***
        (0.0002)

Constant                                         10.681*** 10.789*** 10.795***
10.287*** 10.294***
                                                  (0.010)   (0.011)   (0.011)
(0.014)   (0.014)

--------------------------------------------------------------------------------
-----------------
Observations                                      518,729   518,729   518,729
518,729   518,729
R2                                                 0.042     0.042     0.042
0.049     0.049
================================================================================
=================
Note:                                                                   *p<0.1;
**p<0.05; ***p<0.01
```

[30]:

```
"The regression analysis highlights several pivotal factors influencing␣
↪logwages. Education emerges as a critical determinant, showcasing a␣
↪substantial wage gap between individuals without certificates, diplomas, or␣
↪degrees compared to those with higher educational achievements. Moreover, a␣
↪discernible descending trend in logwages is observed as educational␣
↪attainment decreases from Post-Secondary to Secondary Education levels. Age␣
↪plays a nuanced role, with older individuals experiencing slightly lower␣
↪logwages, indicated by the negative coefficient for age_group. Immigration␣
↪status is shown to have a negative association with wages, suggesting that␣
↪being an immigrant correlates with lower logwages. Intriguingly, the␣
↪positive coefficient for gender reveals that, within this dataset, women␣
↪tend to exhibit higher logwages compared to men. Additionally, membership in␣
↪a visible minority may potentially correspond with slightly lower logwages.␣
↪However, despite these insightful correlations, the models modest R-squared␣
↪values of 0.042 to 0.049 underscore that the identified variables explain␣
↪only a fraction of the overall variance in logwages. This implies the␣
↪existence of unaccounted factors that contribute to the complexities and␣
↪nuances of wage differences among individuals, prompting further exploration␣
↪beyond the variables considered in this analysis."
```

'The regression analysis highlights several pivotal factors influencing logwages. Education emerges as a critical determinant, showcasing a substantial wage gap between individuals without certificates, diplomas, or degrees compared to those with higher educational achievements. Moreover, a discernible descending trend in logwages is observed as educational attainment decreases from Post-Secondary to Secondary Education levels. Age plays a nuanced role, with older individuals experiencing slightly lower logwages, indicated by the negative coefficient for age_group. Immigration status is shown to have a negative association with wages, suggesting that being an immigrant correlates with lower logwages. Intriguingly, the positive coefficient for gender reveals that, within this dataset, women tend to exhibit higher logwages compared to men. Additionally, membership in a visible minority may potentially correspond with slightly lower logwages. However, despite these insightful correlations, the models modest R-squared values of 0.042 to 0.049 underscore that the identified variables explain only a fraction of the overall variance in logwages. This implies the existence of unaccounted factors that contribute to the complexities and nuances of wage differences among individuals, prompting further exploration beyond the variables considered in this analysis.'

[ ]:

```

```