Data Task: Spatial Mobility in the NLSY79

Yu Hui

2024-08-23

1. Summary Statistics

Report the count of moves across U.S. regions for each possible transition. Report the count of moves between urban and non-urban areas. Report mean wage income, mean employment, and mean educational attainment in each region and urban/non-urban bin. Comment on differences you might find.

1.1 Load Data

```
df <- read.csv("nlsy79-prepared.csv")
head(df,5)
```

```
##   i birth gender race region urban wage year educ
## 1 1    58      2    3      1     1 4620 1979   12
## 2 1    58      2    3      1     1 4620 1980   12
## 3 1    58      2    3      1     1 5000 1981   12
## 4 1    58      2    3     NA    NA   NA 1982   12
## 5 1    58      2    3     NA    NA   NA 1983   12
```

1.2 Count the moves for each possible transition pairs

```
df_new <- df %>%
  group_by(i) %>%
  arrange(year) %>%
  mutate(region = zoo::na.locf(region, na.rm = FALSE),
         urban = zoo::na.locf(urban, na.rm = FALSE)) %>%
  ungroup()

df_new <- df_new %>%
  group_by(i) %>%
  arrange(year) %>%
  mutate(region_lag = lag(region),
         urban_lag = lag(urban)) %>%
  ungroup()

# Count transitions between regions
region_moves <- table(df_new$region_lag, df_new$region, useNA = "no")
print(region_moves)
```

```
##
##          1      2      3      4
##   1  55124    306   1233    428
##   2    252  66599   1027    554
##   3    948    860 114501    929
##   4    381    478    954  59583
```

```
# Count transitions between urban and non-urban areas
urban_moves <- table(df_new$urban_lag, df_new$urban, useNA = "no")
print(urban_moves)
```

```
##
##          0      1      2
##   0  57390   6357    557
##   1   6577 220028    864
##   2    386    793   1098
```

1.3 Summary statistics for wage, income and education

```
# Define employment as non-null wages
df$employed <- ifelse(!is.na(df$wage), 1, 0)
# Group by region and urban status to calculate summary statistics
summary_stats <- df %>%
  group_by(region, urban) %>%
  summarise(mean_wage = mean(wage, na.rm = TRUE),
            mean_employment = mean(employed, na.rm = TRUE),
            mean_education = mean(educ, na.rm = TRUE))
```

```
## `summarise()` has grouped output by 'region'. You
## can override using the `.groups` argument.
```

```
knitr::kable(summary_stats)
```

region	urban	mean_wage	mean_employment	mean_education
1	0	18371.4885	0.8752937	10.716915
1	1	18381.8990	0.9024431	10.456225
1	2	31821.7391	0.1483871	10.580645
1	NA	1123.7356	0.9839400	11.382762
2	0	15618.1018	0.8681352	10.351162
2	1	15848.6236	0.8948306	10.440784
2	2	24035.3571	0.0853659	10.564024
2	NA	1965.6931	0.9855967	11.211934
3	0	12750.3276	0.8590190	9.938907
3	1	15758.6351	0.8851544	10.402961
3	2	16270.2743	0.0858663	9.993921
3	NA	1170.4968	0.9742006	11.412427
4	0	13773.2672	0.8045320	10.229935
4	1	17650.9121	0.9032408	10.329701
4	2	26716.5000	0.0903226	10.118535

region	urban	mean_wage	mean_employment	mean_education
4	NA	718.6612	0.9702923	11.341639
NA	0	5613.7222	0.5625000	10.609375
NA	1	4420.2195	0.5347826	10.386957
NA	NA	7844.1008	0.0258981	10.919742

```
# Define employment as non-null wages
df$employed <- ifelse(!is.na(df$wage), 1, 0)
summary_stats_r <- df %>%
  group_by(region) %>%
  summarise(mean_wage = mean(wage, na.rm = TRUE),
            mean_employment = mean(employed, na.rm = TRUE),
            mean_education = mean(educ, na.rm = TRUE))
knitr::kable(summary_stats_r)
```

region	mean_wage	mean_employment	mean_education
1	17574.229	0.9000555	10.52750
2	15667.879	0.8846927	10.42716
3	14436.019	0.8692339	10.29302
4	16496.301	0.8880071	10.36082
NA	7597.576	0.0279807	10.91778

```
# Define employment as non-null wages
df$employed <- ifelse(!is.na(df$wage), 1, 0)
summary_stats_u <- df %>%
  group_by(urban) %>%
  summarise(mean_wage = mean(wage, na.rm = TRUE),
            mean_employment = mean(employed, na.rm = TRUE),
            mean_education = mean(educ, na.rm = TRUE))
knitr::kable(summary_stats_u)
```

urban	mean_wage	mean_employment	mean_education
0	14167.869	0.8572268	10.15356
1	16714.443	0.8943640	10.40590
2	21191.864	0.0909894	10.14229
NA	2507.088	0.1117952	10.96051

1.4 Comments

Mean Wage Income:

Region 1 (Northeast) shows the highest income among four regions. Urban areas generally have higher mean wage incomes compared to non-urban areas across most regions. Region 1 shows a noticeable difference in mean wages between urban and non-urban areas, with urban areas having higher wages. The wage income for the missing data in urbanization status (coded as 2.0) in Region 1(Northeast) is particularly high, which could indicate a data quality issue or a specific subgroup.

Mean Employment:

Region 1 (Northeast) shows the employment rate income among four regions. Employment rates are generally higher in urban areas across all regions, indicating better job opportunities or labor market conditions in urban settings. Region 2(North Central) shows the highest mean employment rate in urban areas, with over 72% of individuals employed.

Mean Educational Attainment:

Educational attainment is relatively consistent across regions, with minor differences between urban and non-urban areas. In Region 1, non-urban areas show slightly higher educational attainment compared to urban areas, which is somewhat counterintuitive and might warrant further investigation.

2. Summarizing Data with Linear Regression

1)

```
# Create the indicator variable for the two groups of interest
df.2 <- df%>%
  mutate(group_indicator = ifelse(race == 3 & gender == 2 & urban == 1 & region == 2 & year >= 2004 & y
                                  ifelse((race == 1 | race == 2) & gender == 1 & urban == 0 & region !=

# Run the linear regression
model <- lm(wage ~ group_indicator, data=df.2)

# Summary of the model to get the coefficient and standard error
stargazer(model, type = "text")
```

1.1)

```
##
## ================================================
##                         Dependent variable:
##                     ---------------------------
##                                wage
## ------------------------------------------------
## group_indicator           -6,114.310***
##                            (1,155.701)
##
## Constant                  27,680.170***
##                             (973.578)
##
## ------------------------------------------------
## Observations                  1,853
## R2                            0.015
## Adjusted R2                   0.014
## Residual Std. Error   22,581.970 (df = 1851)
## F Statistic          27.990*** (df = 1; 1851)
## ================================================
## Note:               *p<0.1; **p<0.05; ***p<0.01
```

Compare with the group b) (indvidual with race-1-or-2 gender-1 non-urban workers who reside anywhere outside region 2) group a) 's wage is 6114.31 dollars lower. The standard error is 1155.7.

2)

```
df_reg <- df %>%
  mutate(group_indicator = ifelse(race == 2 & gender == 1 & urban == 0 & region == 3 & educ <=12 & educ

# Run the linear regression
model <- lm(wage ~ group_indicator -1, data = df_reg)

# Summary of the model to get the coefficient and standard error
stargazer(model, type = "text")
```

```
##
## ================================================
##                         Dependent variable:
##                     ---------------------------
##                                 wage
## ------------------------------------------------
## group_indicator            13,246.170***
##                             (1,210.181)
##
## ------------------------------------------------
## Observations                  216,891
## R2                             0.001
## Adjusted R2                    0.001
## Residual Std. Error  59,680.470 (df = 216890)
## F Statistic         119.806*** (df = 1; 216890)
## ================================================
## Note:               *p<0.1; **p<0.05; ***p<0.01
```

Average wage for "race-2 gender-1 non-urban workers in region 3 with an educational attainment in the 9-12 range" is 13246 dollars.

3)

```
library(dplyr)

# Create the indicator variable for the two groups of interest
df.3 <- df %>%
  mutate(group_indicator = ifelse(race == 1 & gender == 1 & region == 2 & birth == 62, 1,
                                  ifelse(urban == 1 & region == 3 & educ >= 13 & educ <= 16 & year == 2

# Run the linear regression
model <- lm(wage ~ group_indicator, data = df.3)

# Summary of the model to get the coefficient and standard error
stargazer(model, type = "text")
```

```
##
## ================================================
##                         Dependent variable:
```

```
##                       ----------------------------
##                                    wage
## ---------------------------------------------------
## group_indicator              -23,248.130***
##                                (3,712.212)
##
## Constant                      45,442.170***
##                                (2,867.253)
##
## ---------------------------------------------------
## Observations                       409
## R2                                0.088
## Adjusted R2                       0.086
## Residual Std. Error    36,830.540 (df = 407)
## F Statistic           39.220*** (df = 1; 407)
## ===================================================
## Note:                *p<0.1; **p<0.05; ***p<0.01
```

Compare with the group b), group a) 's wage is 23248 dollars lower. The standard error is 3712.

3. Event Study

```
library(dplyr)
library(ggplot2)
library(tidyr)

# Step 1: Identify Movers
# Create lagged variables to identify changes in region and urban status
df.3 <- df %>%
  group_by(i) %>%
  arrange(year) %>%
  mutate(region_lag = lag(region),
         urban_lag = lag(urban),
         move_region = ifelse(region != region_lag & !is.na(region_lag), 1, 0),
         move_urban = ifelse(urban != urban_lag & !is.na(urban_lag), 1, 0),
         move = ifelse(move_region == 1 | move_urban == 1, 1, 0),
         move_year = ifelse(move == 1, year, NA)) %>%
  fill(move_year, .direction = "downup") %>%
  mutate(relative_time = year - move_year) %>%
  filter(relative_time >= -2 & relative_time <= 2)
```

```
df.3 <- df.3 %>%
  group_by(i) %>%
  mutate(moved_any_year_region = ifelse(any(move_region == 1), 1, 0),
         moved_any_year_urban = ifelse(any(move_urban == 1), 1, 0)) %>%
  ungroup()
```

```
# Step 2: Calculate Mean Wage Income by Relative Time for Regional Moves
mean_wage_by_time_region <- df.3 %>%
  group_by(relative_time) %>%
  filter(moved_any_year_region == 1) %>%
  summarise(mean_wage = mean(wage, na.rm = TRUE))
```

```r
# Step 3: Plot the Results for Regional Moves
ggplot(mean_wage_by_time_region, aes(x = relative_time, y = mean_wage)) +
  geom_line() +
  geom_point() +
  labs(title = "Event Study: Mean Wage Income Around Regional Moves",
       x = "Years Relative to Move",
       y = "Mean Wage Income") +
  theme_minimal()
```

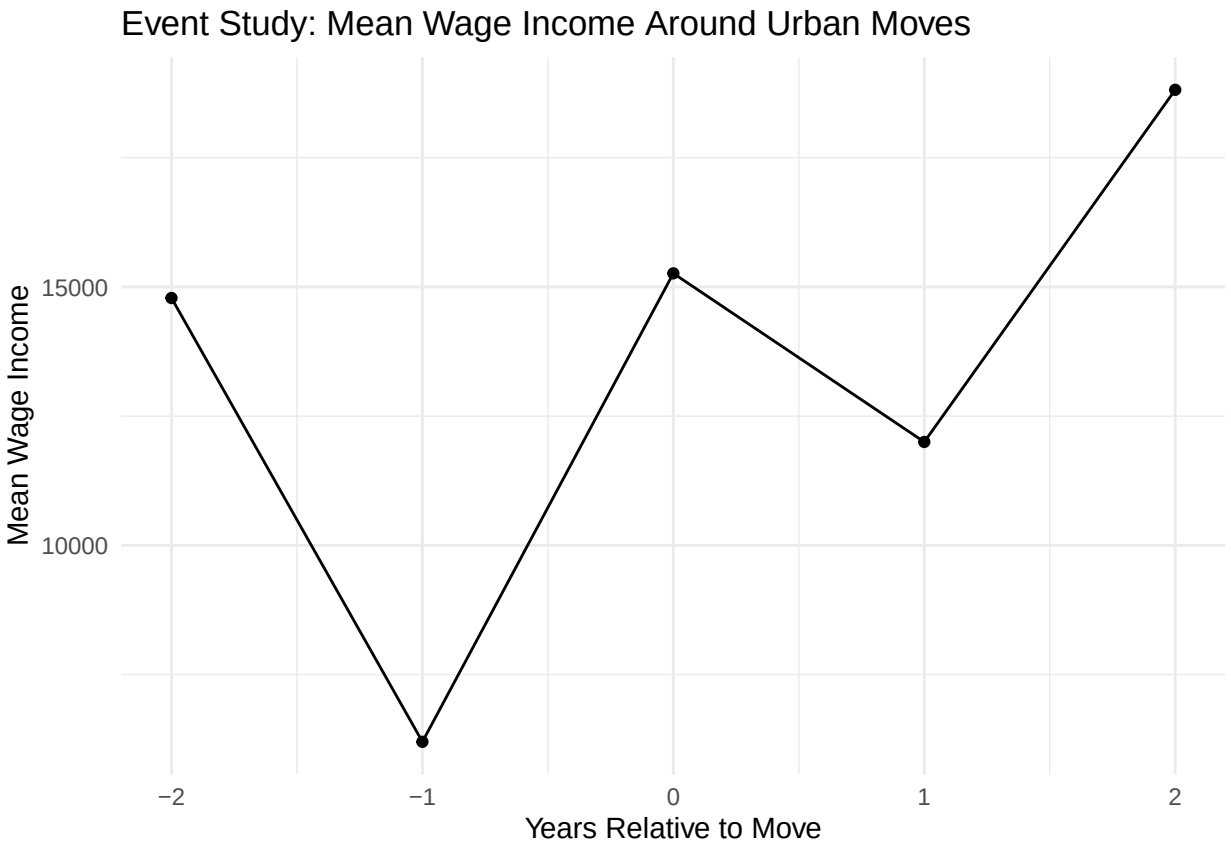
Interpretation:

Pre-Move Period (-2 to -1): The mean wage income decreases from -2 to -1, indicating that the period just before the move might be associated with lower wages, possibly due to job insecurity or other factors related to preparing for the move. At the Time of the Move (0): There's a significant increase in wage income at the time of the move, which could suggest that moving regions is associated with securing higher-paying jobs or better opportunities. Post-Move Period (1 to 2): After the move, the wage income fluctuates slightly but generally remains higher compared to the pre-move period, especially at +2 years, where the income is at its peak. This trend suggests that the move had a positive long-term effect on wages.

```r
# Step 4: Calculate Mean Wage Income by Relative Time for Urban Moves
mean_wage_by_time_urban <- df.3 %>%
  filter(moved_any_year_urban == 1) %>%
  group_by(relative_time) %>%
  summarise(mean_wage = mean(wage, na.rm = TRUE))

# Step 5: Plot the Results for Urban Moves
ggplot(mean_wage_by_time_urban, aes(x = relative_time, y = mean_wage)) +
```

```
  geom_line() +
  geom_point() +
  labs(title = "Event Study: Mean Wage Income Around Urban Moves",
       x = "Years Relative to Move",
       y = "Mean Wage Income") +
  theme_minimal()
```



Pre-Move Period (-2 to -1): Similar to regional moves, there's a noticeable decrease in wage income from -2 to -1, potentially indicating challenges or disruptions faced before moving to a new urban environment. At the Time of the Move (0): The wage income significantly increases at the time of the move, reflecting the potential benefits of relocating to a different urban setting, such as better job opportunities or increased demand for certain skills. Post-Move Period (1 to 2): Post-move, the income briefly dips at +1 year but then rises sharply by +2 years, indicating a delayed but substantial benefit from moving to a new urban area.

4. Comparing movers to stayers

```
library(dplyr)
library(ggplot2)
library(lmtest)
```

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

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

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

```
library(sandwich)

# Step 1: Identify Movers and Stayers
# Create an indicator for movers
df.4 <- df %>%
  group_by(i) %>%
  arrange(year) %>%
  mutate(region_lag = lag(region),
         urban_lag = lag(urban),
         move_region = ifelse(region != region_lag & !is.na(region_lag), 1, 0),
         move_urban = ifelse(urban != urban_lag & !is.na(urban_lag), 1, 0),
         move = ifelse(move_region == 1 | move_urban == 1, 1, 0),
         move_year = ifelse(move == 1, year, NA)) %>%
  fill(move_year, .direction = "downup") %>%
  mutate(relative_time = year - move_year)

df.4 <- df.4 %>%
  group_by(i) %>%
  mutate(moved_any_year = ifelse(any(move_region == 1), 1, 0)) %>%
  ungroup()
# transform age var
df.4<- df.4%>% mutate(age = 2024-1900+birth)
# Create a new column for the original region of each moved individual
df.4 <- df.4 %>%
  group_by(i) %>%
  mutate(original_region = first(region)) %>%
  ungroup()
# Create a column for the last region of each moved inidvidual
df.4 <- df.4 %>%
  group_by(i) %>%
  mutate(last_region =last(region)) %>%
  ungroup()
# Step 2: Create the Relative Time Variable (Already done in previous steps)
```

```
# Step 3: Run Regression Models
# a) Origin Region
# Regression to compare wage income in the origin region
model_origin <- lm(wage ~ moved_any_year * factor(original_region) + age + educ + race + gender, data =
summary_origin <- summary(model_origin)
stargazer(model_origin, type = "text")
```

```
##
## ===============================================================
##                                         Dependent variable:
##                                     ---------------------------
##                                                wage
## ---------------------------------------------------------------
## moved_any_year                            -5,680.796***
##                                             (665.546)
```

```
##
## factor(original_region)2                        -3,590.101***
##                                                   (599.517)
##
## factor(original_region)3                        -4,061.702***
##                                                   (551.130)
##
## factor(original_region)4                        -2,120.160***
##                                                   (638.847)
##
## age                                              1,895.646***
##                                                   (102.158)
##
## educ                                             3,181.820***
##                                                   (114.775)
##
## race                                             1,586.935***
##                                                   (207.026)
##
## gender                                          -9,897.069***
##                                                   (294.173)
##
## moved_any_year:factor(original_region)2          3,884.079***
##                                                   (879.831)
##
## moved_any_year:factor(original_region)3          2,956.433***
##                                                   (847.491)
##
## moved_any_year:factor(original_region)4            691.605
##                                                   (979.694)
##
## Constant                                        -351,020.500***
##                                                   (19,722.960)
##
## ---------------------------------------------------------------------
## Observations                                        156,463
## R2                                                    0.013
## Adjusted R2                                           0.013
## Residual Std. Error                        57,793.930 (df = 156451)
## F Statistic                               192.181*** (df = 11; 156451)
## =====================================================================
## Note:                                      *p<0.1; **p<0.05; ***p<0.01
```

```
# b) Destination Region
# Regression to compare wage income in the destination region
model_destination <- lm(wage ~ moved_any_year * factor(last_region) + age + educ + race + gender, data =
summary_destination <- summary(model_destination)
stargazer(model_destination, type = "text")
```

```
##
## ============================================================
##                                       Dependent variable:
##                                   ---------------------------
##                                              wage
```

```
## ------------------------------------------------------------
## moved_any_year                              172.151
##                                            (998.352)
##
## factor(last_region)2                     -3,708.829***
##                                            (643.823)
##
## factor(last_region)3                     -3,800.819***
##                                            (590.962)
##
## factor(last_region)4                      -1,573.165**
##                                            (687.617)
##
## age                                       1,729.269***
##                                            (116.507)
##
## educ                                      3,265.097***
##                                            (130.688)
##
## race                                      2,314.668***
##                                            (235.295)
##
## gender                                   -10,712.350***
##                                            (336.641)
##
## moved_any_year:factor(last_region)2        -2,288.677*
##                                           (1,252.953)
##
## moved_any_year:factor(last_region)3          -449.131
##                                           (1,141.459)
##
## moved_any_year:factor(last_region)4         1,076.061
##                                           (1,294.605)
##
## Constant                                 -321,828.700***
##                                            (22,475.200)
##
## ------------------------------------------------------------
## Observations                                140,144
## R2                                            0.014
## Adjusted R2                                   0.014
## Residual Std. Error                  62,621.420 (df = 140132)
## F Statistic                        178.411*** (df = 11; 140132)
## ============================================================
## Note:                               *p<0.1; **p<0.05; ***p<0.01
```

Original region stayer and mover comparison: What can we learned from the regression is that movers, on average, earn $5680.8 less than stayers in the origin region, controlling for other factors. This negative coefficient suggests that movers might experience a decrease in wages when initially moving from their origin region. When we look into interaction terms, we can see Region 2 (3884.1): Movers from Region 2 earn 3884.1 more than movers from Region 1, but still less than stayers in Region 1 (considering the negative main effect of moved_any_year). The result is significant at 1%. Region 3 (2956.4): Similar to Region 2, movers from Region 3 earn more than those from Region 1 but less than stayers. The result is also significant at 1%. Region 4 (691.6): The difference is not statistically significant (p-value = 0.480226), indicating no

strong evidence of a wage difference for movers from Region 4 compared to Region 1.

Destination region stayer and mover comparison: What we can see for the distination region is that the coefficient before moving indicator is positive but small and not statistically significant (p-value = 0.8631), suggesting that there is no significant difference in wages between movers and stayers in the destination region on average. Region 2 (-2288.7): Movers to Region 2 earn $2288.7 less than stayers in Region 1, but this effect is marginally significant (p-value = 0.0678). Region 3 (-449.1): The difference is not statistically significant (p-value = 0.6940), indicating no strong evidence of a wage difference for movers to Region 3 compared to Region 1. Region 4 (1076.1): The difference is not statistically significant (p-value = 0.4059), indicating no strong evidence of a wage difference for movers to Region 4 compared to Region 1.

```
# b) Destination Region
# Regression to compare wage income in the destination region
model_destination <- lm(wage ~ moved_any_year * factor(original_region)*relative_time + age + educ + ra
summary_destination <- summary(model_destination)
summary_destination
```

```
##
## Call:
## lm(formula = wage ~ moved_any_year * factor(original_region) *
##     relative_time + age + educ + race + gender, data = df.4)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
##  -48905  -12040   -4099    5746 4122648
##
## Coefficients:
##                                                      Estimate
## (Intercept)                                        -332872.07
## moved_any_year                                       -8679.91
## factor(original_region)2                             -2567.93
## factor(original_region)3                             -4464.23
## factor(original_region)4                             -4442.80
## relative_time                                          716.56
## age                                                   1826.82
## educ                                                  2927.94
## race                                                   746.33
## gender                                               -9231.45
## moved_any_year:factor(original_region)2               2228.15
## moved_any_year:factor(original_region)3               1801.78
## moved_any_year:factor(original_region)4               3190.26
## moved_any_year:relative_time                             2.75
## factor(original_region)2:relative_time                -144.25
## factor(original_region)3:relative_time                -272.79
## factor(original_region)4:relative_time                -311.68
## moved_any_year:factor(original_region)2:relative_time  203.12
## moved_any_year:factor(original_region)3:relative_time  534.43
## moved_any_year:factor(original_region)4:relative_time  227.45
##                                                     Std. Error
## (Intercept)                                          25485.16
## moved_any_year                                         920.28
## factor(original_region)2                               944.44
## factor(original_region)3                               885.60
## factor(original_region)4                              1063.10
```

```
## relative_time                                              70.14
## age                                                       131.58
## educ                                                      147.05
## race                                                      265.66
## gender                                                    366.16
## moved_any_year:factor(original_region)2                  1175.10
## moved_any_year:factor(original_region)3                  1128.03
## moved_any_year:factor(original_region)4                  1323.66
## moved_any_year:relative_time                               90.25
## factor(original_region)2:relative_time                     85.02
## factor(original_region)3:relative_time                     79.19
## factor(original_region)4:relative_time                     90.00
## moved_any_year:factor(original_region)2:relative_time     115.67
## moved_any_year:factor(original_region)3:relative_time     113.46
## moved_any_year:factor(original_region)4:relative_time     126.08
##                                                          t value
## (Intercept)                                              -13.061
## moved_any_year                                            -9.432
## factor(original_region)2                                  -2.719
## factor(original_region)3                                  -5.041
## factor(original_region)4                                  -4.179
## relative_time                                             10.216
## age                                                       13.884
## educ                                                      19.911
## race                                                       2.809
## gender                                                   -25.211
## moved_any_year:factor(original_region)2                    1.896
## moved_any_year:factor(original_region)3                    1.597
## moved_any_year:factor(original_region)4                    2.410
## moved_any_year:relative_time                               0.030
## factor(original_region)2:relative_time                    -1.697
## factor(original_region)3:relative_time                    -3.445
## factor(original_region)4:relative_time                    -3.463
## moved_any_year:factor(original_region)2:relative_time      1.756
## moved_any_year:factor(original_region)3:relative_time      4.710
## moved_any_year:factor(original_region)4:relative_time      1.804
##                                                         Pr(>|t|)
## (Intercept)                                              < 2e-16
## moved_any_year                                           < 2e-16
## factor(original_region)2                                0.006549
## factor(original_region)3                                4.64e-07
## factor(original_region)4                                2.93e-05
## relative_time                                            < 2e-16
## age                                                      < 2e-16
## educ                                                     < 2e-16
## race                                                    0.004965
## gender                                                   < 2e-16
## moved_any_year:factor(original_region)2                 0.057943
## moved_any_year:factor(original_region)3                 0.110204
## moved_any_year:factor(original_region)4                 0.015946
## moved_any_year:relative_time                            0.975694
## factor(original_region)2:relative_time                  0.089764
## factor(original_region)3:relative_time                  0.000572
## factor(original_region)4:relative_time                  0.000534
```

```
## moved_any_year:factor(original_region)2:relative_time 0.079082
## moved_any_year:factor(original_region)3:relative_time 2.47e-06
## moved_any_year:factor(original_region)4:relative_time 0.071221
##
## (Intercept)                                           ***
## moved_any_year                                        ***
## factor(original_region)2                              **
## factor(original_region)3                              ***
## factor(original_region)4                              ***
## relative_time                                         ***
## age                                                   ***
## educ                                                  ***
## race                                                  **
## gender                                                ***
## moved_any_year:factor(original_region)2               .
## moved_any_year:factor(original_region)3
## moved_any_year:factor(original_region)4               *
## moved_any_year:relative_time
## factor(original_region)2:relative_time                .
## factor(original_region)3:relative_time                ***
## factor(original_region)4:relative_time                ***
## moved_any_year:factor(original_region)2:relative_time .
## moved_any_year:factor(original_region)3:relative_time ***
## moved_any_year:factor(original_region)4:relative_time .
## ---
## Signif. codes:
## 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 60200 on 109538 degrees of freedom
##   (207592 observations deleted due to missingness)
## Multiple R-squared:  0.02209,    Adjusted R-squared:  0.02192
## F-statistic: 130.2 on 19 and 109538 DF,  p-value: < 2.2e-16
```

running out of time but plan to do some hetergenity analysis on different time periods relative to the move. Already create the relative_time variable for analysis.