HKS SUP-135 Lab 2: Measuring Upward Mobility Using the National Longitudinal Survey

Matt Khinda

2/10/2023

```
# Load dataset found in the folder "datasets"
nlsy <- read_dta(here("datasets", "nlsy97.dta"))
```

Question 1: Subsetting Data

```
# Subset full dataset for children whose mothers have high school education or less
momEdu_under12 <- subset(nlsy, nlsy$mother_education <= 12)
# Calculate fraction of children who have a college degree in this subset
child_ed_mean <- mean(momEdu_under12$child_college, na.rm = T)

cat("The fraction of children whose mothers have a high school education
    or less that receive a college degree is ",child_ed_mean*100, "%", sep="")
```

1a: Fraction of Children Whose Mothers have a High School Education or Less Receiving College Degree

```
## The fraction of children whose mothers have a high school education
##     or less that receive a college degree is 18.18182%
```

1b: Judgement Given the small sample size of our data set, I would conclude that my result of 18.18% is fairly close to Chetty et. al's finding of 20.9%.

Question 2: Percentile Ranks

```
# Define pct_rank function
pct_rank <- function(x){
  #Catch NAs
  r <- ifelse(is.na(x),NA, rank(x,ties.method = "average"))
  #return percentile rank
  100*r/max(r,na.rm = TRUE)
}

# Calculate kid_inc_rank
nlsy$kid_inc_rank <- pct_rank(nlsy$kid_income)
```

2a: Calculate Child Income Percentile Rank

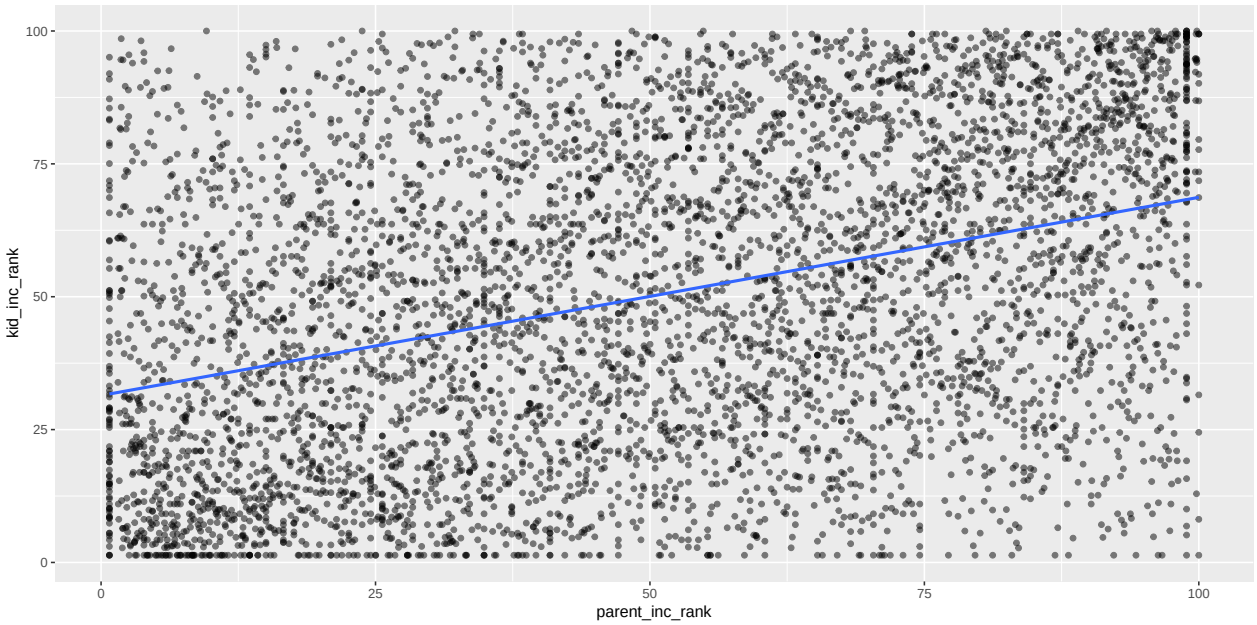
```
# Calculate parent_inc_rank
nlsy$parent_inc_rank <- pct_rank(nlsy$parent_inc)
```

2b: Calculate Parent Income Percentile Rank

Question 3: Visualizing

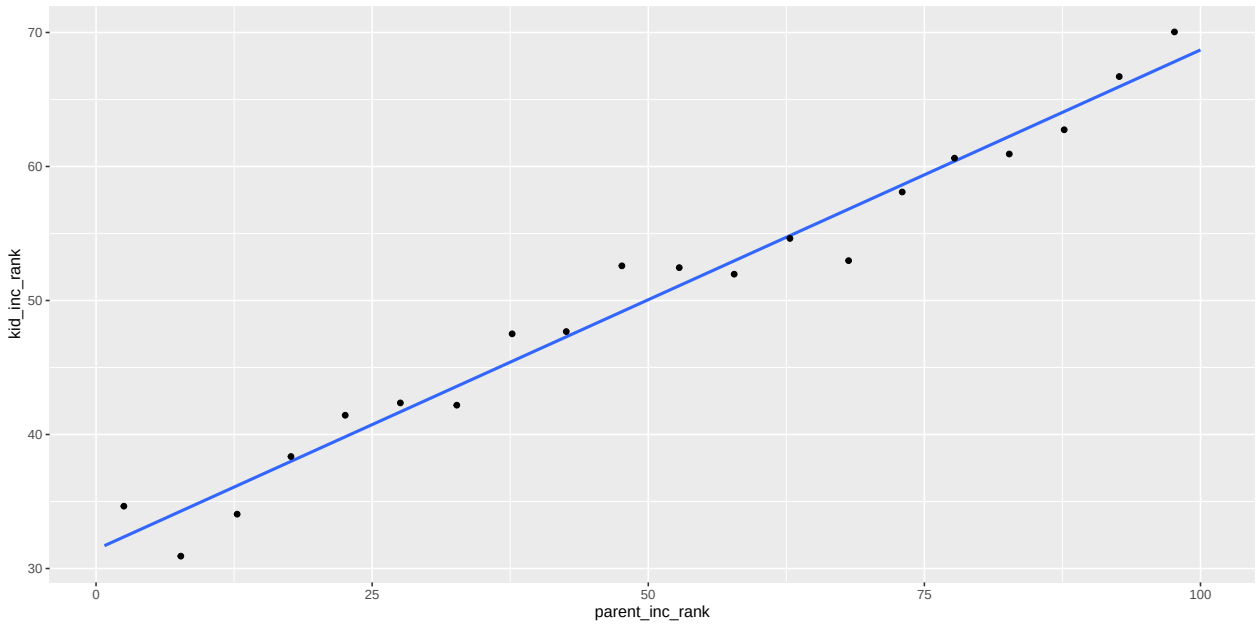
```
ggplot(nlsy, aes(x=parent_inc_rank, y=kid_inc_rank)) +
  geom_point(alpha = 0.5) +
  geom_smooth(method = 'lm', se = F)
```

3a: Individual-Level Scatter Plot




```
ggplot(nlsy, aes(x=parent_inc_rank, y=kid_inc_rank)) +
  geom_smooth(method = 'lm', se = F) +
  stat_binmean(n = 20, geom = "point")
```

3b: Binned Scatter Plot



By comparing the two plots above, we can see that the binned scatter plot provides a much clearer picture of the linear relationship between parent's income rank and the child's income rank at adulthood.

Question 4: Linear Regression

```
model_1 <- lm(kid_inc_rank ~ parent_inc_rank, data = nlsy)
summary(model_1)
```

```
##
## Call:
## lm(formula = kid_inc_rank ~ parent_inc_rank, data = nlsy)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -67.337 -22.709  -0.244  21.751  66.446
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)     31.41826    0.72467   43.35   <2e-16 ***
## parent_inc_rank  0.37279    0.01253   29.75   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 26.83 on 5484 degrees of freedom
## Multiple R-squared:  0.1389, Adjusted R-squared:  0.1388
## F-statistic: 884.8 on 1 and 5484 DF,  p-value: < 2.2e-16
```

Question 5: Measures of Upward Mobility

```r
absMobility25 <- 31.41826 + .37279*25
cat("The predicted income rank for someone born to parents at the
    25th percentile rank is", absMobility25)
```

5a: Absolute Mobility at 25th Percentile

```
## The predicted income rank for someone born to parents at the
##     25th percentile rank is 40.73801
```

```r
relMobility <- (31.41826 + .37279*100)-(31.41826 + .37279*0)
cat("The relative mobility is", relMobility)
```

5b: Relative Mobility

```
## The relative mobility is 37.279
```

```r
nlsy$top_20 <- ifelse(nlsy$kid_inc_rank > 80, 1,0)
horatioAlger <- with(subset(nlsy, parent_inc_rank < 20), mean(top_20, na.rm = T))
cat("The probability that a child born to parents in the bottom income quintile
    reaches the top income quintile is ",horatioAlger*100, "%", sep="")
```

5c: Horatio Alger Measure

```
## The probability that a child born to parents in the bottom income quintile
##     reaches the top income quintile is 7.370337%
```

```r
#set inflation adjustment factor
infl_adj <- 1.4767

nlsy$kid_inc_greater <- ifelse(nlsy$kid_income > infl_adj*nlsy$parent_inc, 1, 0)
absMobility <- mean(nlsy$kid_inc_greater)
cat("The fraction of children earning more than their parents in
    inflation-adjusted dollars is ",absMobility*100, "%", sep="")
```

5d: Absolute Mobility

```
## The fraction of children earning more than their parents in
##     inflation-adjusted dollars is 50.89318%
```

Question 6: Racial Disparities in Measures of Upward Mobility

```r
#subset data by race and gender
black_men <- subset(nlsy, female == 0 & black ==1)
white_men <- subset(nlsy, female == 0 & white ==1)

black_men_model <- lm(kid_inc_rank ~ parent_inc_rank, data = black_men)
white_men_model <- lm(kid_inc_rank ~ parent_inc_rank, data = white_men)
summary(black_men_model)
```

6a: Absolute Mobility at 25th Percentile

```
##
## Call:
## lm(formula = kid_inc_rank ~ parent_inc_rank, data = black_men)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -49.102 -22.725  -4.398  20.356  69.048
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)     25.70348    1.71292  15.006  < 2e-16 ***
## parent_inc_rank  0.29432    0.03714   7.925 8.96e-15 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 26.85 on 701 degrees of freedom
## Multiple R-squared:  0.08223,    Adjusted R-squared:  0.08093
## F-statistic: 62.81 on 1 and 701 DF,  p-value: 8.959e-15
```

```r
summary(white_men_model)
```

```
##
## Call:
## lm(formula = kid_inc_rank ~ parent_inc_rank, data = white_men)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -64.453 -20.644   1.727  21.700  57.949
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)     40.11022    1.52681   26.27   <2e-16 ***
## parent_inc_rank  0.26692    0.02379   11.22   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 26.28 on 1668 degrees of freedom
## Multiple R-squared:  0.07019,    Adjusted R-squared:  0.06963
## F-statistic: 125.9 on 1 and 1668 DF,  p-value: < 2.2e-16
```

```
black_men_absMobility25 <- 25.70348 + 0.29432*25
white_men_absMobility25 <- 40.11022 + 0.26692*25

cat("The predicted income rank for Black men born to parents at the
    25th percentile rank is", black_men_absMobility25, "while for white men it is", white_men_absMobili
```

```
## The predicted income rank for Black men born to parents at the
##     25th percentile rank is 33.06148 while for white men it is 46.78322
```

```
black_men_relMobility <- (25.70348 + 0.29432*100)-(25.70348 + 0.29432*0)
white_men_relMobility <- (40.11022 + 0.26692*100)-(40.11022 + 0.26692*0)

cat("The relative mobility measure for Black men is", black_men_relMobility,
    "while for white men it is", white_men_relMobility)
```

6b: Relative Mobility

```
## The relative mobility measure for Black men is 29.432 while for white men it is 26.692
```

```
black_men_horatioAlger <- with(subset(black_men, parent_inc_rank < 20), mean(top_20, na.rm = T))
white_men_horatioAlger <- with(subset(white_men, parent_inc_rank < 20), mean(top_20, na.rm = T))

cat("The probability that a Black man born to parents in the bottom income
    quintile reaches the top income quintile is ",black_men_horatioAlger*100,
    "% while for a white man it is ", white_men_horatioAlger*100, "%",sep="")
```

6c: Horatio Alger Measure

```
## The probability that a Black man born to parents in the bottom income
##     quintile reaches the top income quintile is 5.785124% while for a white man it is 10.27027%
```

```
black_men_absMobility <- mean(black_men$kid_inc_greater)
white_men_absMobility <- mean(white_men$kid_inc_greater)

cat("The fraction of Black men earning more than their parents in
    inflation-adjusted dollars is ",black_men_absMobility*100, "% while
    the fraction of white men is ", white_men_absMobility*100, "%",sep="")
```

6d: Absolute Mobility

```
## The fraction of Black men earning more than their parents in
##     inflation-adjusted dollars is 48.93314% while
##     the fraction of white men is 48.26347%
```

Of the four measures of mobility presented above, three of them (absolute mobility at the 25th percentile, relative mobility, and the Horatio Alger measure) show a clear disparity in outcomes for white and Black men. However, because absolute mobility is nearly the same for white and Black men it would be misleading to make a determination on that statistic alone. Therefore it does matter which measure you use to determine if there is a racial disparity.

Question 7: Regression vs Arithmetic Mean

```
# Subset full dataframe for parent income rank between 20 and 30
subset_btwn_20_30 <- subset(nlsy, parent_inc_rank > 20 & parent_inc_rank < 30)
# Calculate mean child income rank based on that subset
arithmetic_mean_20_30 <- mean(subset_btwn_20_30$kid_inc_rank)
cat("The average (mean) income rank for someone born to parents at the
    25th percentile rank is", arithmetic_mean_20_30)
```

7a: Arithmetic Mean

```
## The average (mean) income rank for someone born to parents at the
##     25th percentile rank is 41.93234
```

```
# Set random seed based on HUID number
HUID <- 41531460
set.seed(HUID)

# Generate list of random values between 0 and 1
random_values <- runif(nrow(nlsy))

# Assign random values to rows in dataframe
nlsy$rand_val <- random_values[nlsy$id_num]

# Create new dataframe with top 50 highest random values
nlsy_randSubset <- nlsy %>%
  arrange(desc(rand_val)) %>%
  slice(1:50)

# Run linear regression on subset
randSubset_model <- lm(kid_inc_rank ~ parent_inc_rank, data = nlsy_randSubset)
summary(randSubset_model)
```

7b: Random Sample — Linear Regression

```
##
## Call:
## lm(formula = kid_inc_rank ~ parent_inc_rank, data = nlsy_randSubset)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
```

```
## -52.662 -17.667   3.574  18.878  48.607
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)      33.1733     8.3652   3.966 0.000243 ***
## parent_inc_rank   0.4604     0.1220   3.773 0.000443 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 23.89 on 48 degrees of freedom
## Multiple R-squared:  0.2287, Adjusted R-squared:  0.2127
## F-statistic: 14.23 on 1 and 48 DF,  p-value: 0.0004435
```

```
# Calculate absolute mobility at 25th percentile
randSubset_absMobility25 <- 33.1733 + .4604*25
cat("The predicted income rank for someone born to parents at the
    25th percentile rank in this random sample (n=", nrow(nlsy_randSubset),
    ") is ", randSubset_absMobility25, sep = "")
```

```
## The predicted income rank for someone born to parents at the
##     25th percentile rank in this random sample (n=50) is 44.6833
```

```
randSubset_btwn_20_30 <- subset(nlsy_randSubset, parent_inc_rank > 20 & parent_inc_rank < 30)
randSubset_arithmetic_mean_20_30 <- mean(randSubset_btwn_20_30$kid_inc_rank)
cat("The average (mean) income rank for someone born to parents at the
    25th percentile rank in this random sample (n=",nrow(randSubset_btwn_20_30),
    ") is ", randSubset_arithmetic_mean_20_30, sep ="")
```

7c: Random Sample — Arithmetic Mean

```
## The average (mean) income rank for someone born to parents at the
##     25th percentile rank in this random sample (n=4) is 60.98585
```

7d: Random Sample — Comparing Measures As shown in the calculations above on the random sample, the predicted income rank based on the linear regression is much closer than the arithmetic mean to the full population estimate of 41.3 calculated by Chetty et al.