Lab2

2023-02-07

```
#filtering dataset by observations whose mother_education is less than or equal to 12, then calculating
nlsy97_momcollege <- nlsy97 |>
                     filter(mother_education <= 12)

mean(nlsy97_momcollege$child_college, na.rm = TRUE)
```

```
## [1] 0.1818182
```

The proportion of children whose mothers had a high school education or less who went on to receive a college degree or higher is 18.18%.

Question 1B

My estimate is close to the Census Bureau estimate of 20.9%. There is roughly a 2.5% difference between the two estimates, which is reasonable due to the different samples used.

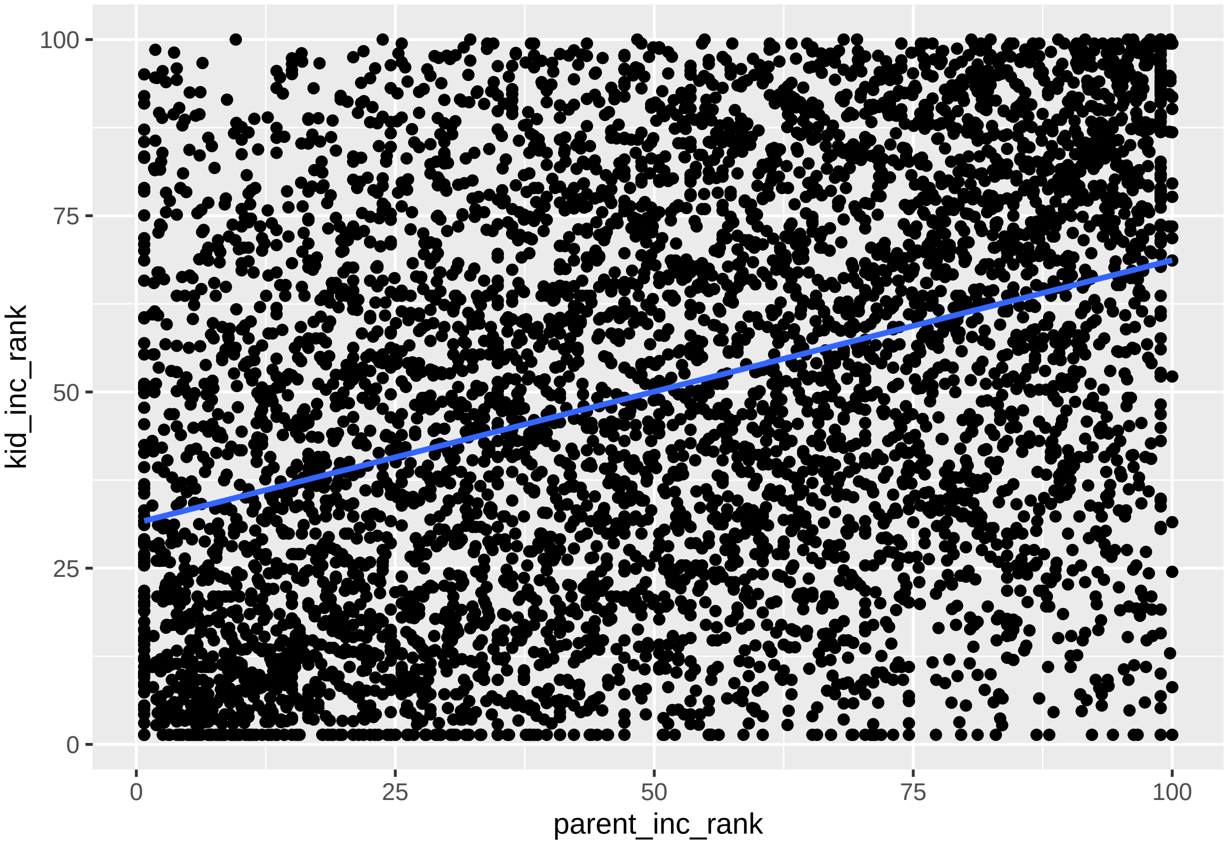
```
#2A & 2B: creating percentile ranks for kid_income and parent_inc, normalized to 100

nlsy97 <- nlsy97 |>
          mutate(ranked_kidincome = rank(kid_income),
                 max_rank = max(ranked_kidincome),
                 kid_inc_rank = (ranked_kidincome/max_rank)*100,
                 ranked_parentinc = rank(parent_inc),
                 max_rankparent = max(ranked_parentinc),
                 parent_inc_rank = (ranked_parentinc/max_rankparent)*100)
```

```
#3A: scatter plot of individual level data

scatterrank <- nlsy97 |>
  ggplot() +
  geom_point(aes(x = parent_inc_rank, y = kid_inc_rank)) +
  geom_smooth(aes(x = parent_inc_rank, y = kid_inc_rank), method = "lm", se = F)
scatterrank
```

```
## `geom_smooth()` using formula = 'y ~ x'
```

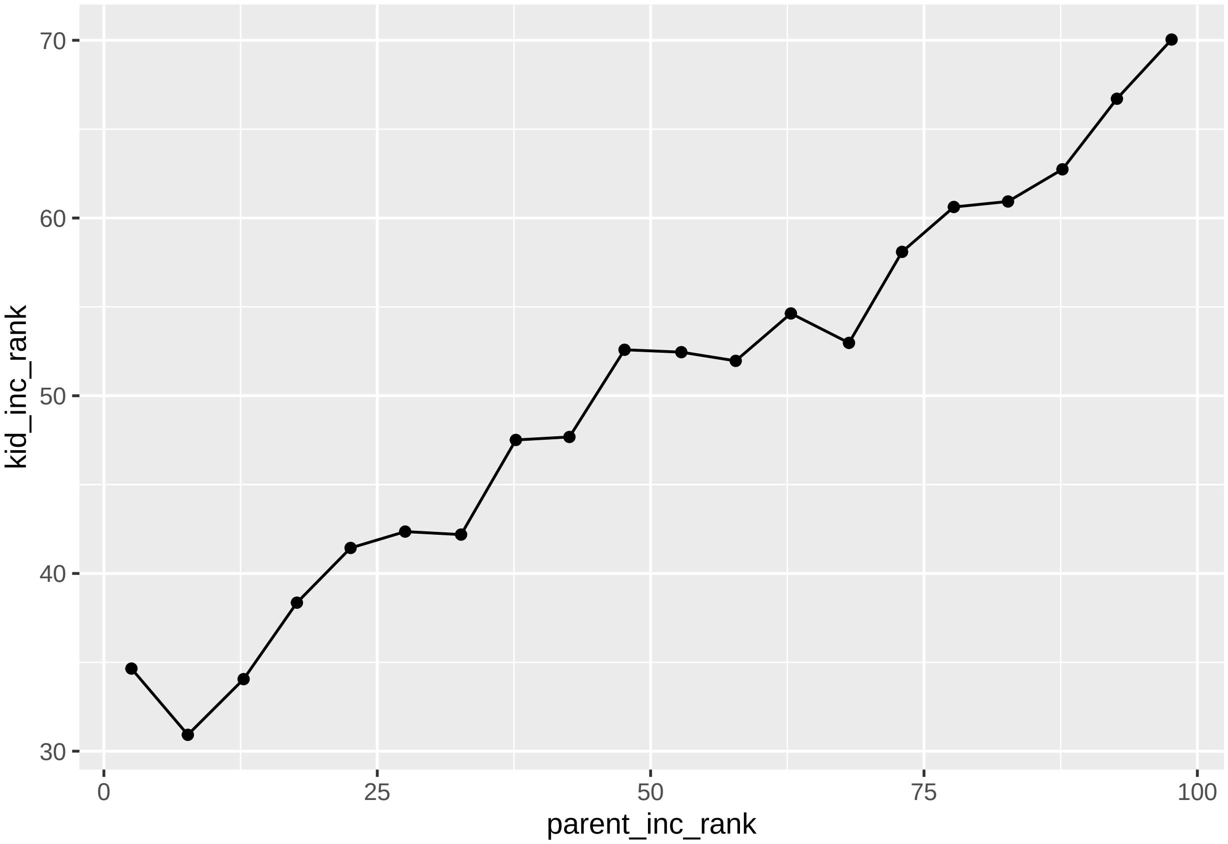


```r
#Save graph
ggsave("scatterrank.png")
```

```
## Saving 6.5 x 4.5 in image
## `geom_smooth()` using formula = 'y ~ x'
```

```r
#3B: binned scatter plot

binscatter <- nlsy97 |>
  ggplot(aes(x = parent_inc_rank, y = kid_inc_rank)) +
  stat_binmean(n = 20, geom = "point") +
  stat_binmean(n = 20, geom = "line")
binscatter
```



```
#save graph
ggsave("binscatter_connected.png")
```

```
## Saving 6.5 x 4.5 in image
```

The binned scatter plot is a far more useful summary of the data because it is cleaner to visualize and the positively correlated relationship between parent_inc_rank and kid_inc_rank is more visible.

```
#run linear regression of kid income ranks on parent income ranks

mod1 = lm(kid_inc_rank ~ parent_inc_rank, data = nlsy97)
summary(mod1)
```

```
##
## Call:
## lm(formula = kid_inc_rank ~ parent_inc_rank, data = nlsy97)
##
## 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
```

In this linear regression, the intercept is 31.41, and the coefficient is 0.372. What this means is that the average parent income rank for child income ranks of 0 is 31.41, according to this regression. The coefficient tells us that for every increase of 1 parental income rank, the child income rank increases by 0.37.

```
#Statistic 1: predicted child income rank from the rank-rank regression for parental income rank 25th p

as.numeric(mod1$coefficients)
```

```
## [1] 31.4182627  0.3727907
```

```
intercept <- as.numeric(mod1$coefficients)[1]
slope <- as.numeric(mod1$coefficients)[2]

intercept + 25*slope
```

```
## [1] 40.73803
```

The predicted child income rank for those whose parents where in the 25th percentile of income rank according to the regression is 40.73, which is not too far from the full tax data predicted rank of 41.3

```
#Statistic 2: Relative mobility
100*slope
```

```
## [1] 37.27907
```

The predicted relative mobility using the linear model is 37.27, which is close to the full tax data relative mobility of 34.1

```
#Statistic 3: Probability that a child born to parents in the bottom fifth of the income distro reaches
nlsy97$top20 <- ifelse(nlsy97$kid_inc_rank >= 80, 1, 0)
nlsy97_bottom20 <- nlsy97 |>
                   filter(parent_inc_rank < 20)
mean(nlsy97_bottom20$top20)
```

```
## [1] 0.07370337
```

The probability that a child born to parents in the bottom fifth of the income distro reaches the top fifth according to the linear model is 7.37%, which is very close to the full tax data value of 7.5%.

```
#Statistic 4: Fraction of children who make more than their parents (inflation adjusted)

nlsy97$morethanparents <- ifelse(nlsy97$kid_income > 1.4767*nlsy97$parent_inc, 1, 0)
mean(nlsy97$morethanparents)
```

```
## [1] 0.5089318
```

The simple linear model suggests that 50.89% of children will make more in inflation adjusted dollars than their parents, which is very close to the 50% estimate from the full tax data.

```
#Black Men regression and upward mobility statistics
#Filtering the data for Black men
black_men <- nlsy97 |>
            filter(black == 1 & female == 0)

#Linear regression of kid income ranks on parent income ranks for Black men
mod1_blackmen = lm(kid_inc_rank ~ parent_inc_rank, data = black_men)
summary(mod1_blackmen)
```

```
##
## 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
```

```
#Upward mobility statistics for Black men

#Statistic 1: predicted child income rank from the rank-rank regression for parental income rank 25th p

as.numeric(mod1_blackmen$coefficients)
```

```
## [1] 25.7034768  0.2943199
```

```
intercept_bm <- as.numeric(mod1_blackmen$coefficients)[1]
slope_bm <- as.numeric(mod1_blackmen$coefficients)[2]

intercept_bm+ 25*slope_bm
```

```
## [1] 33.06148
```

```
#Statistic 2: Relative mobility
100*slope_bm
```

```
## [1] 29.43199
```

```
#Statistic 3: Probability that a child born to parents in the bottom fifth of the income distro reaches
black_men$top20 <- ifelse(black_men$kid_inc_rank >= 80, 1, 0)
blackmen_bottom20 <- black_men |>
                   filter(parent_inc_rank < 20)
mean(blackmen_bottom20$top20)
```

```
## [1] 0.05785124
```

```
#Statistic 4: Fraction of children who make more than their parents (inflation adjusted)

black_men$morethanparents <- ifelse(black_men$kid_income > 1.4767*black_men$parent_inc, 1, 0)
mean(black_men$morethanparents)
```

```
## [1] 0.4893314
```

```
#White Men regression and upward mobility statistics
#Filtering the data for White men
white_men <- nlsy97 |>
            filter(white == 1 & female == 0)
```

```r
#Linear regression of kid income ranks on parent income ranks for Black men
mod1_whitemen = lm(kid_inc_rank ~ parent_inc_rank, data = white_men)
summary(mod1_whitemen)
```

```
##
## 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
```

```r
#Upward mobility statistics for White men

#Statistic 1: predicted child income rank from the rank-rank regression for parental income rank 25th p

as.numeric(mod1_whitemen$coefficients)
```

```
## [1] 40.110220  0.266921
```

```r
intercept_wm <- as.numeric(mod1_whitemen$coefficients)[1]
slope_wm <- as.numeric(mod1_whitemen$coefficients)[2]

intercept_wm+ 25*slope_wm
```

```
## [1] 46.78325
```

```r
#Statistic 2: Relative mobility
100*slope_wm
```

```
## [1] 26.6921
```

```r
#Statistic 3: Probability that a child born to parents in the bottom fifth of the income distro reaches
white_men$top20 <- ifelse(white_men$kid_inc_rank >= 80, 1, 0)
whitemen_bottom20 <- white_men |>
                    filter(parent_inc_rank < 20)
mean(whitemen_bottom20$top20)
```

```
## [1] 0.1027027
```

```r
#Statistic 4: Fraction of children who make more than their parents (inflation adjusted)

white_men$morethanparents <- ifelse(white_men$kid_income > 1.4767*white_men$parent_inc, 1, 0)
mean(white_men$morethanparents)
```

```
## [1] 0.4826347
```

Comparing the regression-based statistics for Black men and White men, some patterns emerge. The regression slope for the Black men-only dataset is 0.29, while for White men, it is 0.266. For Black men whose parents were in the 25th income percentile, the regression predicts that they will be in 33rd percentile, while for White men, roughly the 47th percentile. Relative mobility is slightly higher for Black men, at 29.43 compared to 26.69. The probability that Black men born to parents in the bottom fifth of the income distribution will reach the top fifth is 5.7%, compared to 10.2% for white men. Finally, the fraction of children who make more in (inflation adjusted) dollars than their parents is similar between Black and White men, 48.9% and 48.2%, respectively.

Taken together, there is slight evidence that mobility is higher for White men than Black men. White men are predicted to be in a higher income rank if their parents were in the 25th income percentile, and crucially, are twice as likely than Black men to make it to the top fifth of the income distribution if they were born to parents in the bottom fifth. Other indicators of mobility, such as relative mobility and fraction of children who make more in inflation adjusted dollars than their parents are more balanced. This suggests that more robust studies of upward mobility need to take other factors, such as social capital, into account.

```
#7A: Simple arithmetic mean of child income rank for those born to parents between the 20th and 30th pe
nlsy_7a <- nlsy97 |>
           filter(parent_inc_rank > 20 & parent_inc_rank < 30)

mean(nlsy_7a$kid_inc_rank)
```

```
## [1] 41.93234
```

The simple mean of child income rank for those born to parents between the 20th and 30th percentile is 41.9. This is slightly higher than the regression predicted binned average, which is 40.7.

```
#7B: Calculating absolute mobility at the 25th percentile with linear regression for a random sample of

set.seed(21519588)
sample50 <- sample_n(nlsy97, 50, replace = TRUE)
modsample = lm(kid_inc_rank ~ parent_inc_rank, sample50)
summary(modsample)
```

```
##
## Call:
## lm(formula = kid_inc_rank ~ parent_inc_rank, data = sample50)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -46.860 -24.763  -1.121  21.353  56.063
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)      32.2283     7.1416   4.513 4.14e-05 ***
## parent_inc_rank   0.3572     0.1211   2.949  0.00491 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 27.08 on 48 degrees of freedom
## Multiple R-squared:  0.1534, Adjusted R-squared:  0.1357
## F-statistic: 8.697 on 1 and 48 DF,  p-value: 0.004914
```

```
as.numeric(modsample$coefficients)
```

```
## [1] 32.2283017  0.3572043
```

```
interceptsample <- as.numeric(modsample$coefficients)[1]
slopesample <- as.numeric(modsample$coefficients)[2]

interceptsample + 25*slopesample
```

```
## [1] 41.15841
```

```
#7C: Mean of child income rank for parents between 20th and 30th percentile from random sample

mean(sample50$kid_inc_rank[sample50$parent_inc_rank > 20 & sample50$parent_inc_rank <30])
```

```
## [1] 47.64491
```

From a random sample of 50 observations, the absolute mobility at the 25th percentile is 41.15. The simple mean of child income rank for those born to parents with income ranked between the 20th and 30th percentiles from the random sample is higher, at 47.64. The sample predicted mean using the linear regression is much closer to the full population estimate of 41.3. This is because linear regression is a better tool for providing estimated values of the relationship between two variables.