Lab2

2023-02-07

```r
#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.

```r
#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)
```

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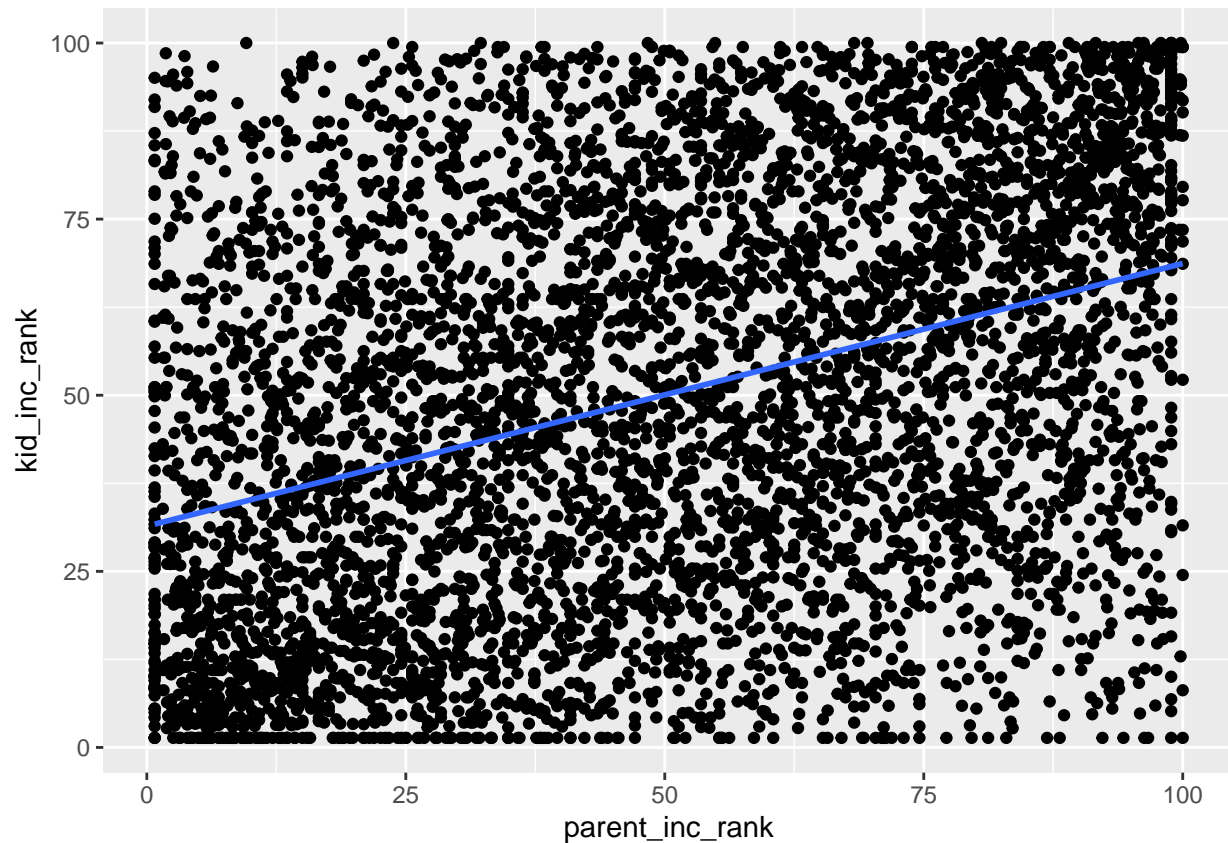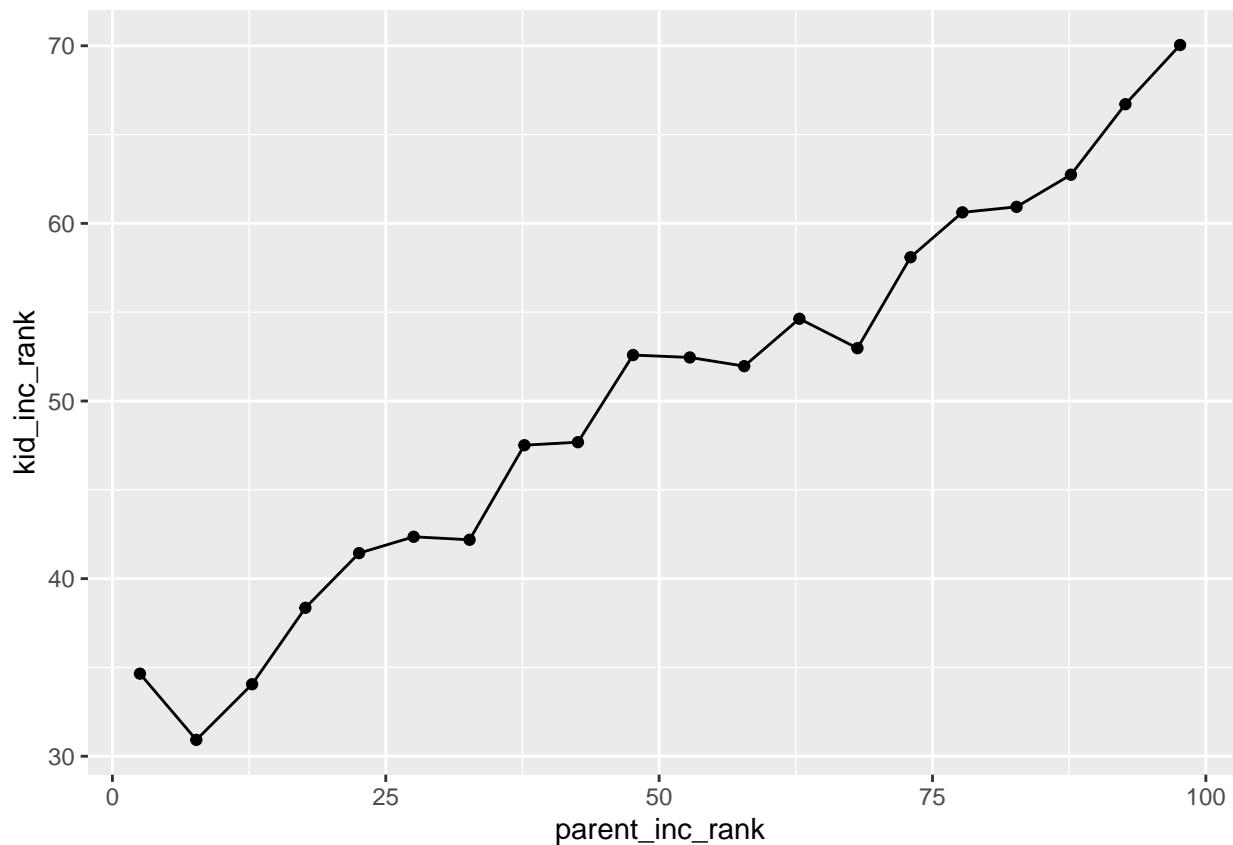
```
#Save graph
ggsave("scatterrank.png")
```

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

```
#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
```

```r
#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.

```r
#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)
```

```
#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
```

```
#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
```

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

intercept_wm+ 25*slope_wm
```

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

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

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

```
#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
```

```
#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
```