The association between education and earnings in the 1997 National Longitudinal Survey of Youth.

Adam J Batten

April 2019

```
library(tint)
library(dplyr)
library(survey)
library(tableone)
library(ggplot2)
library(ggthemes)
# invalidate cache when the package version changes
knitr::opts_chunk$set(tidy = FALSE, cache.extra = packageVersion('tint')
                      , comment = "")
options(htmltools.dir.version = FALSE)

# Data
load("nlsyModelMatrix.Rdata")
svyds <- svydesign(data = ds, ids = ~PUBID_1997, weights = ~WEIGHT)
```

Introduction

Explanation of Research Topic

Using data from the National Longitudinal Study of Youth, we analyze the relationship between education and poverty. We hypothesize that increased levels of education lead to less poverty. While many studies have already shown this to be true of earlier cohorts (Card, 1994), to our knowledge, there has not been a study to date that analyzes this relationship using the more recent 1997 cohort. This analysis is relevant as it has been rumoured (Chetty et al., 2018) that the offspring of the Baby Boomer generation are slated to earn less (on avearge) than there progenitors.

Design

We conducted a cross-sectional cohort study using the most recent responses to the NLSY follow-up surveys which were conducted in 2015. Our primary outcome is the ratio of household income to the poverty level in 2015. The primary exposure of interest is the highest level of education acheived. We model the ratio of income to poverty as a continuous variable using multivariate linear regression.

Data Source

The 1997 National Longitudinal Survey of Youth[1] consists of repeat observations on 8,984 U.S men and women born in the years 1980-84; respondents were ages 12-17 when first interviewed in 1997. This cohort has been interviewed several times since 1997 and most recently in 2015. While each survey in the NLSY has over 100,000 variables, the features we are interested in are income and education. In adjusted analysis, we consider a subset of features including: employment history, family background, marital status, childcare & fertility, health, attitudes, and crime & substance use.

[1] U.S. Bureau of Labor Statistics (BLS) website NLSY- 1997

Methods

Cohort & Exclusions

In the first step of our analysis we exclude any respondents that had no interviews in 2015. We can identify this with the `CVC_RND_XRND` feature which is an integer representing each year since 1997. Since we only want respondents to the 2015 wave we remove anyone with `CVC_RND_XRND < 17`. This resulted in an exclusion of 1,881 respondents from our cohort.

Outcome

There are several income related questions on the NLSY, including the below. In the current analysis we are primarily interested in the `CV_HH_POV_RATIO` features. We used the ratio of household income to the poverty level in 2015 as our main outcome. Adjusted analysis included the same metric from 1997 to account for the baseline wealth of the respondents household.

Select income features from the NLSY 1997.

Options | Refresh List | 500 | 1

	RNUM	QUESTION NAME	VARIABLE TITLE	YEAR
1	R0000100	PUBID	PUBID, YOUTH CASE IDENTIFICATION CODE	1997
2	R1204900	CV_HH_POV_RATIO	RATIO OF HOUSEHOLD INCOME TO POVERTY LEVEL	1997
3	U0008900	CV_INCOME_FAMILY	GROSS FAMILY INCOME	2015
4	U0009000	CV_HH_POV_RATIO	RATIO OF HOUSEHOLD INCOME TO POVERTY LEVEL	2015
5	U0956700	YINC-1400	R RECEIVE INCOME FROM JOB IN PAST YEAR?	2015
6	U0956900	YINC-1700	TOTAL INCOME FROM WAGES AND SALARY IN PAST YEAR	2015
7	U0957000	YINC-1800	ESTIMATED INCOME FROM WAGES AND SALARY IN PAST YEAR	2015
8	U0958200	YINC-2600	SPOUSES TOTAL INCOME FROM WAGES AND SALARY PAST YEAR	2015
9	U0958300	YINC-2700	ESTIMATED SPOUSES TOTAL INCOME FROM WAGES AND SALARY PAST YEAR	2015
10	Z9141400	CVC_HH_NET_WORTH_35	NET WORTH OF HOUSEHOLD AT AGE 35	XRND
11	Z9141800	CVC_ASSETS_FINANCIAL_35	FINANCIAL ASSETS AT AGE 35	XRND
12	Z9141900	CVC_ASSETS_NONFINANCIAL_35	NONFINANCIAL ASSETS AT AGE 35	XRND
13	Z9142000	CVC_ASSETS_DEBTS_35	DEBT AT AGE 35	XRND

Figure 1: Income Variables

Primary exposure

The primary exposure of interest in this analysis was the highest degree received as of 2015 (`CVC_HIGHEST_DEGREE_EVER`). This feature has the following possible allowed responses:

```
0 None
1 GED
2 High school diploma (Regular 12 year program)
3 Associate/Junior college (AA)
4 Bachelor's degree (BA, BS)
5 Master's degree (MA, MS)
6 PhD
7 Professional degree (DDS, JD, MD)
```

Adjustment features

The NLSY has hundreds of demographic and socio-economic data on their respondents. In the current analysis we decided to focus on a subset of these, including: sex, race, marital status, whether the respondent resides in a metropolitan statistical area (MSA), the number of biological children, and the health status of respondent.

We selected sex to account for the differences in earnings potential across genders, and similar reasoning was followed when including race. We included marital status to account for the potential increase in household wealth resulting from a potential spouses income. MSA was included to accouunt for spatial differences in income. We included the number of children to adjust for the fact that children are very expensive. Finally we included the health status of the respondents as illness can impact longterm earnings due to extended absences from work.

Statistical Analysis

In analyzing this data the BLS recommends using their post-stratification survey weights. Thus we use survey weighted least squares regression to assess the relationship between the poverty ratio and education. Specifically our primary model can be parameterized as[2]

$$R \sim \beta_0 + \beta_1 X_{Education}$$

Where R is the ratio of household income to the poverty level in 2015. For the adjusted analysis we considered the following model:

$$\begin{aligned} R \sim \beta_0 + \beta_1 X_{Education} + \beta_2 X_{Female} + \beta_3 X_{Race} + \beta_4 X_{HealthStatus} \\ + \beta_5 X_{MaritalStatus} + \beta_6 X_{MSA} + \beta_7 X_{zChildren} \end{aligned}$$

[2] All statistical analysis was conducted with the survey package (Lumley, 2004) in R (R Core Team, 2019). This report was rendered in markdown (Allaire et al., 2018) using the tint template (Eddelbuettel and Gilligan, 2019).

Results

Univariate Statistics

Looking at our table of demographics we can get a better idea of our sample. The majority of the sample is White (70.4%), in good to very good health (66%), married (48%), live in the suburbs (56%), have on average 1.4 children, are relatively wealthy and earn ~4 times the poverty rate, and 40% have some post-High School education.

```
kableone(svyCreateTableOne(data = svyds
          , vars = c("Female", "Race", "HealthStatus"
             , "MaritalStatus", "MSA", "Children"
             , "PovertyRatio1997", "PovertyRatio2015"
             , "HighestDegreeCat"))
         , caption = 'Table 1: Patient demographics, poverty, and education.')
```

Table 1: Table 1: Patient demographics, poverty, and education.

	Overall
n	1937845395.00
Female (mean (SD))	0.49 (0.50)
Race (%)	
Non-Black / Non-Hispanic	1364176820.0 (70.4)
Black	298706421.0 (15.4)
Hispanic	249228863.0 (12.9)
Mixed Race (Non-Hispanic)	25733291.0 (1.3)
HealthStatus (%)	
Excellent	427569314.0 (22.1)
Very good	708926429.0 (36.6)
Good	569948794.0 (29.5)
Fair	195855146.0 (10.1)
Poor	32709511.0 (1.7)
MaritalStatus (%)	
Married	931275521.0 (48.3)
Never Married	784853466.0 (40.7)
SepaDivoWido	211743832.0 (11.0)
MSA (%)	
Not in CBSA	86808761.0 (4.5)
In CBSA, not in central city	1077689180.0 (55.7)
In CBSA, in central city	745809844.0 (38.5)
In CBSA, not known	4149829.0 (0.2)
Not in country	20835996.0 (1.1)
Children (mean (SD))	1.38 (1.36)

	Overall
PovertyRatio1997 (mean (SD))	3.15 (2.80)
PovertyRatio2015 (mean (SD))	4.11 (3.75)
HighestDegreeCat (%)	
None	140074265.0 (7.2)
GED	230946962.0 (12.0)
High School	765319880.0 (39.6)
Associate	167835390.0 (8.7)
Bachelor	428004573.0 (22.1)
Master	155251148.0 (8.0)
PhD	13658359.0 (0.7)
ProfessionalDegree	31230839.0 (1.6)

Bivariate Relationships

Looking at some bivariate relationships between poverty rate and our adjustment features we see some evidence towards confirmation of our hypothesis. The increasing trend in wealth with increasing education is made apparent in Fig. 2.

```
par(mfrow = c(1,2))
svyboxplot(PovertyRatio2015 ~ HighestDegreeCat
           , design = svyds, horizontal = TRUE
           , las = 2, main = "PovertyRatio2015 ~ HighestDegreeCat")
svyboxplot(PovertyRatio2015 ~ HealthStatus
           , design = svyds, horizontal = TRUE
           , las = 2, main = "PovertyRatio2015 ~ HealthStatus")
```

```
par(mfrow = c(1,2))
svyboxplot(PovertyRatio2015 ~ MaritalStatus
           , design = svyds, horizontal = TRUE
           , las = 2, main = "PovertyRatio2015 ~ MaritalStatus")
svyboxplot(PovertyRatio2015 ~ factor(Female)
           , design = svyds
           , main = "PovertyRatio2015 ~ Female")
```

Main Effect

We begin modeling our relationship with the single predictor first. Since we are dealing with survey weighted data we run a survey weighted regression in addition to the standard OLS model. Wald tests for the significance of this regression model are highly significant, providing evidence towards a relationship between education and wealth.

Figure 2: Boxplots reveal a nice linear trend with more education and health.

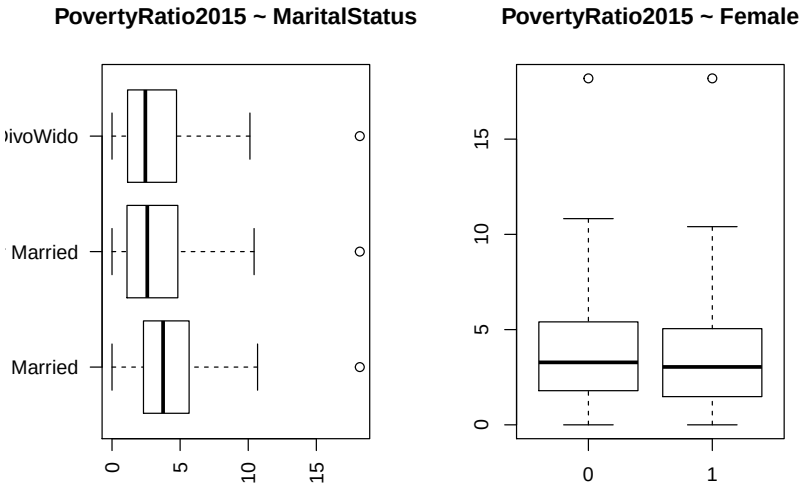

Figure 3: Boxplots reveal slight decreases in poverty across sex and marital status (SepaDivoWido is the separated, divorced, or widowed category).

```
# Survey weighted model
svym01 <- svyglm(PovertyRatio2015 ~ HighestDegreeCat, design = svyds)
# Ignoring the weights
m01 <- lm(PovertyRatio2015 ~ HighestDegreeCat, data = ds)
# Wald test
regTermTest(svym01, ~HighestDegreeCat)
```

```
Wald test for HighestDegreeCat
 in svyglm(formula = PovertyRatio2015 ~ HighestDegreeCat, design = svyds)
F =  133.6054  on  7  and  6292  df: p= < 2.22e-16
```

```
anova(m01)
```

```
Analysis of Variance Table

Response: PovertyRatio2015
                   Df Sum Sq Mean Sq F value
HighestDegreeCat    7  12956 1850.91  173.64
Residuals        6292  67070   10.66
                    Pr(>F)
HighestDegreeCat < 2.2e-16 ***
Residuals
---
Signif. codes:
  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1  ' ' 1
```

To see how close the effect sizes are for the survey weighted and standard models we can quickly plot the coefficient and 95% confidence intervals. From Fig. 4 we can see that the coefficient estimates are quite close whether we decide to include the survey weights or not. Interestingly, while the estimated effect due to PhD is significant it is highly unstable (high variability) and does not appear to be significantly different from the effect of Master's.

```
# Create dataFrame of coefficient estimates
coefMain <- rbind.data.frame(confint(m01), confint(svym01))
names(coefMain) <- c("LB", "UB")
coefMain$Estimate <- c(coef(m01), coef(svym01))
coefMain$Term <- gsub("HighestDegreeCat","", gsub("\\d+","", rownames(coefMain)))
coefMain$Model <- c(rep("M01", length(coef(m01))), rep("SVYM01", length(coef(m01))))
# Order the terms by effect size
foo <- aggregate(Estimate ~ Term, coefMain, mean)
coefMain$Terms <- factor(coefMain$Term, levels = foo$Term[order(foo$Estimate)])
ggplot(coefMain, aes(x = Terms, y = Estimate, group = Model, colour = Model)) +
  geom_errorbar(aes(ymin=LB,ymax=UB), position = "dodge") +
```

```
  coord_flip() +
  theme_tufte()
```

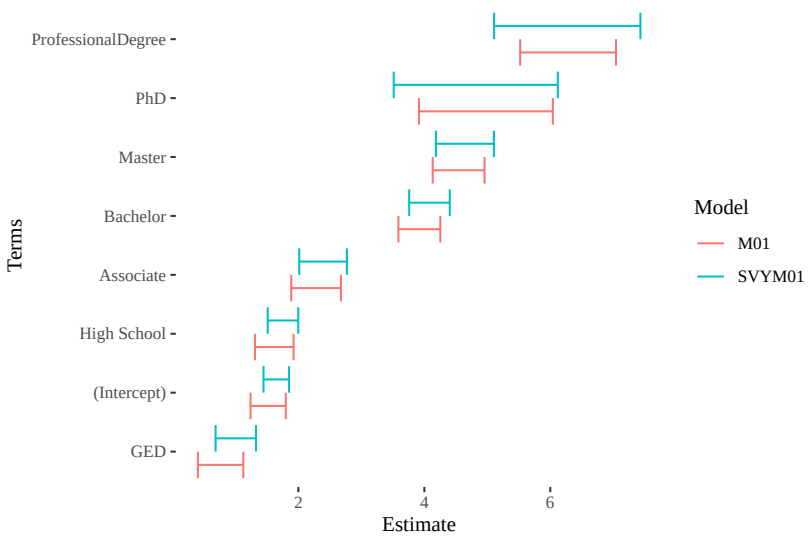

Figure 4: 95 confidence intervals for the model coefficient estimates for both the non-survey weighted regression (M01) and the survey weighted regression (SVYM01)

Multivariate Model

In the next step we add the adjustment variables and test for significance of regression. The Wald test and Ftest reveal that education is still a significant predictor of wealth even after accounting for sex, race, health, marital status, location, number of children, and historic poverty.

```
svym02 <- update(svym01, . ~ . + Female + Race + HealthStatus
        + MaritalStatus + MSA + zChildren + PovertyRatio1997)
m02 <- update(m01, . ~ . + Female + Race + HealthStatus
        + MaritalStatus + MSA + zChildren + PovertyRatio1997)
regTermTest(svym01, ~HighestDegreeCat)
```

```
Wald test for HighestDegreeCat
 in svyglm(formula = PovertyRatio2015 ~ HighestDegreeCat, design = svyds)
F =  133.6054  on  7  and  6292  df: p= < 2.22e-16
```

```
anova(m02)
```

```
Analysis of Variance Table

Response: PovertyRatio2015
                   Df Sum Sq Mean Sq
```

```
HighestDegreeCat    7   8724 1246.30
Female              1    337  336.89
Race                3   1760  586.51
HealthStatus        4    351   87.76
MaritalStatus       2    179   89.35
MSA                 4    122   30.58
zChildren           1   1938 1938.12
PovertyRatio1997    1    953  952.61
Residuals        4606  43741    9.50
                  F value    Pr(>F)
HighestDegreeCat 131.2383 < 2.2e-16 ***
Female            35.4750 2.776e-09 ***
Race              61.7614 < 2.2e-16 ***
HealthStatus       9.2415 1.956e-07 ***
MaritalStatus      9.4085 8.361e-05 ***
MSA                3.2200   0.01195 *
zChildren        204.0894 < 2.2e-16 ***
PovertyRatio1997 100.3120 < 2.2e-16 ***
Residuals
---
Signif. codes:
  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1  ' ' 1
```

To see which levels of the factor variables are contributing to the model fit we again plot the 95% CI for the coefficient estimates. We include a vertical bar at 0 to highlight those effects that are not significantly different from zero. From Fig. 5 we see that

```
# Create dataFrame of coefficient estimates
coefMain2 <- rbind.data.frame(confint(m02), confint(svym02))
names(coefMain2) <- c("LB", "UB")
coefMain2$Estimate <- c(coef(m02), coef(svym02))
coefMain2$Term <- gsub("HighestDegreeCat","", gsub("\\d+","", rownames(coefMain2)))
coefMain2$Model <- c(rep("M02", length(coef(m02))), rep("SVYM02", length(coef(m02))))
# Order the terms by effect size
foo <- aggregate(Estimate ~ Term, coefMain2, mean)
coefMain2$Terms <- factor(coefMain2$Term, levels = foo$Term[order(foo$Estimate)])
ggplot(coefMain2, aes(x = Terms, y = Estimate, group = Model, colour = Model)) +
  geom_errorbar(aes(ymin=LB,ymax=UB), position = "dodge") +
  coord_flip() +
  geom_hline(yintercept = 0, alpha = 0.4) +
  theme_tufte()
```

Listing the predictors that are significant we see that all of the education levels are significant and increasing with higher levels. This confirms our

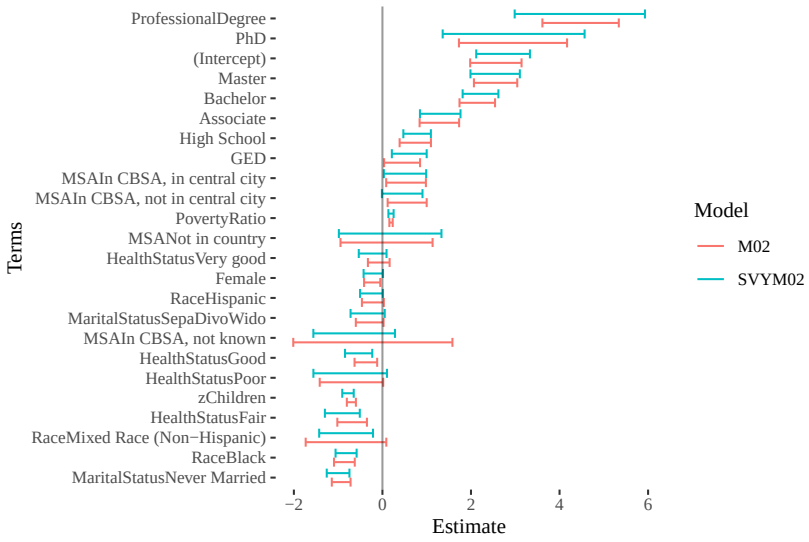

Figure 5: 95 confidence intervals for the model coefficient estimates from the adjusted models: non-survey weighted regression (M01) and the survey weighted regression (SVYM01)

hypothesis. However, the model fit is quite bad (adjusted R^2 = 0.243441) and only accounts for 24% of the variability in the data.

```
sigEffects <- unique(coefMain2[!(coefMain2$LB<=0 & coefMain2$UB>=0), 'Term'])
knitr::kable(coefMain2[coefMain2$Term%in%sigEffects & coefMain2$Model=="M02"
                       , c("Terms", "Estimate", "LB", "UB")]
             , row.names = FALSE)
```

Terms	Estimate	LB	UB
(Intercept)	2.5617613	1.9818103	3.1417124
GED	0.4438528	0.0379522	0.8497535
High School	0.7418823	0.3883218	1.0954428
Associate	1.2855236	0.8391990	1.7318482
Bachelor	2.1420104	1.7407255	2.5432952
Master	2.5552903	2.0659845	3.0445962
PhD	2.9488943	1.7279398	4.1698487
ProfessionalDegree	4.4755803	3.6130538	5.3381068
Female	-0.2320873	-0.4150012	-0.0491734
RaceBlack	-0.8597644	-1.0941832	-0.6253456
RaceMixed Race (Non-Hispanic)	-0.8231866	-1.7337739	0.0874008
HealthStatusGood	-0.3742507	-0.6296899	-0.1188115
HealthStatusFair	-0.6843471	-1.0185820	-0.3501121
MaritalStatusNever Married	-0.9316330	-1.1432985	-0.7199675
MSAIn CBSA, not in central city	0.5584102	0.1176660	0.9991545
MSAIn CBSA, in central city	0.5331697	0.0848783	0.9814612
zChildren	-0.7017609	-0.8038454	-0.5996764

Terms	Estimate	LB	UB
PovertyRatio	0.1941398	0.1561384	0.2321413

Is a PhD worth it? Digging into our discovery above about the overlap in effects between the effects due to Master's and PhD's we test the linear hypothesis that $\beta_{PhD} = \beta_{Master's}$ using the (Torsten et al., 2010) package. The large p-value on the below tests reveals that under the assumptions in this model, PhD's and Master's are on average equivalently wealthy (or poor, depending on your point of view).

```
summary(multcomp::glht(m02
      , linfct = c("HighestDegreeCatMaster - HighestDegreeCatPhD = 0")))
```

```
     Simultaneous Tests for General Linear Hypotheses

Fit: lm(formula = PovertyRatio2015 ~ HighestDegreeCat + Female + Race +
    HealthStatus + MaritalStatus + MSA + zChildren + PovertyRatio1997,
    data = ds)

Linear Hypotheses:
                                                  Estimate
HighestDegreeCatMaster - HighestDegreeCatPhD == 0  -0.3936
                                                  Std. Error
HighestDegreeCatMaster - HighestDegreeCatPhD == 0     0.6181
                                                  t value
HighestDegreeCatMaster - HighestDegreeCatPhD == 0  -0.637
                                                  Pr(>|t|)
HighestDegreeCatMaster - HighestDegreeCatPhD == 0    0.524
(Adjusted p values reported -- single-step method)
```

Just to verify that a Master's degree is worth the effort we can also test the hypothesis that $\beta_{Bachelor's} = \beta_{Master's}$. While the p-value is significant, it is borderline.

```
summary(multcomp::glht(m02
      , linfct = c("HighestDegreeCatMaster - HighestDegreeCatBachelor = 0")))
```

```
     Simultaneous Tests for General Linear Hypotheses

Fit: lm(formula = PovertyRatio2015 ~ HighestDegreeCat + Female + Race +
    HealthStatus + MaritalStatus + MSA + zChildren + PovertyRatio1997,
    data = ds)
```

```
Linear Hypotheses:
                                                       Estimate
HighestDegreeCatMaster - HighestDegreeCatBachelor == 0   0.4133
                                                       Std. Error
HighestDegreeCatMaster - HighestDegreeCatBachelor == 0     0.1984
                                                       t value
HighestDegreeCatMaster - HighestDegreeCatBachelor == 0   2.083
                                                       Pr(>|t|)
HighestDegreeCatMaster - HighestDegreeCatBachelor == 0   0.0373

HighestDegreeCatMaster - HighestDegreeCatBachelor == 0 *
---
Signif. codes:
  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1  ' ' 1
(Adjusted p values reported -- single-step method)
```

Conclusions

This study analyzed the relationship between education and wealth using the National Longitudinal Survey of Youth from 1997. In our analysis we found a strong association between education and wealth as measured by the proportion of earnings above the poverty line. This relationship held after adjusting for respondent demographics, location, and baseline poverty.

While we originally expected a linear trend, with higher education leading to higher wealth, we discovered that while respondents with Master's degrees are wealthier than respondents with Bachelor's, it is not the case that respondents with PhD's have higher earnings than Master's (on average).

The major limitation in this study is the overall poor fit of the models. Further analysis should be done using more advanced methods to identify those factors that influence the relationship between education and wealth.

References

JJ Allaire, Yihui Xie, Jonathan McPherson, Javier Luraschi, Kevin Ushey, Aron Atkins, Hadley Wickham, Joe Cheng, Winston Chang, and Richard Iannone. *rmarkdown: Dynamic Documents for R*, 2018. URL https://CRAN.R-project.org/package=rmarkdown. R package version 1.11.

David Card. Earnings, schooling, and ability revisited. Working Paper 4832, National Bureau of Economic Research, August 1994. URL http://www.nber.org/papers/w4832.

Raj Chetty, Nathaniel Hendren, Maggie R Jones, and Sonya R Porter. Race and economic opportunity in the united states: An intergenerational perspective. Working Paper 24441, National Bureau of Economic Research, March 2018. URL http://www.nber.org/papers/w24441.

Dirk Eddelbuettel and Jonathan Gilligan. *tint: 'tint' is not 'Tufte'*, 2019. URL https://CRAN.R-project.org/package=tint. R package version 0.1.2.

Thomas Lumley. Analysis of complex survey samples. *Journal of Statistical Software*, 9(1):1–19, 2004. R package verson 2.2.

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2019. URL https://www.R-project.org/.

Hothorn Torsten, Frank Bretz, Peter Westfall, Richard M. Heiberger, Andre Schuetzenmeister, and Susan Scheibe. *multcomp: Simultaneous Inference in General Parametric Models*, 2010. URL http://www.crcpress.com/product/isbn/9781584885740. R package version 1.4-10.