analysis.R

byrds

2025-05-07

```r
# Run "./data/new_data97-educational-data/new_data97-educational-data.R" first.
# This file runs analyses on two rounds of the NLSY97 data set.
setwd("C:/Users/byrds/Documents/rStudio_work/social_research")

library(MASS)
library(tidyverse)
```

```
## Warning: package 'ggplot2' was built under R version 4.4.3
```

```
## Warning: package 'purrr' was built under R version 4.4.3
```

```
## Warning: package 'lubridate' was built under R version 4.4.3
```

```
## -- Attaching core tidyverse packages ----------------------- tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.5
## v forcats   1.0.0     v stringr   1.5.1
## v ggplot2   3.5.2     v tibble    3.2.1
## v lubridate 1.9.4     v tidyr     1.3.1
## v purrr     1.0.4
## -- Conflicts ------------------------------------------- tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## x dplyr::select() masks MASS::select()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```r
library(extrafont)
```

```
## Registering fonts with R
```

```r
library(ggthemes)
```

```
## Warning: package 'ggthemes' was built under R version 4.4.2
```

```r
library(hrbrthemes)
library(thematic)
```

```
## Warning: package 'thematic' was built under R version 4.4.2
```

```r
library(colorspace)
library(addinslist)
library(gmodels)
```

```
## Warning: package 'gmodels' was built under R version 4.4.2
```

```r
library(RColorBrewer)
library(DescTools)
```

```
## Warning: package 'DescTools' was built under R version 4.4.3
```

```
## Registered S3 method overwritten by 'DescTools':
##   method         from
##   reorder.factor gdata
```

```r
library(viridis)
```

```
## Loading required package: viridisLite
```

```r
library(ggpmisc)
```

```
## Warning: package 'ggpmisc' was built under R version 4.4.2
```

```
## Loading required package: ggpp
```

```
## Warning: package 'ggpp' was built under R version 4.4.2
```

```
## Registered S3 methods overwritten by 'ggpp':
##   method                  from
##   heightDetails.titleGrob ggplot2
##   widthDetails.titleGrob  ggplot2
##
## Attaching package: 'ggpp'
##
## The following object is masked from 'package:ggplot2':
##
##     annotate
```

```r
library(naniar)
```

```
## Warning: package 'naniar' was built under R version 4.4.3
```

```r
library(broom)
```

```
## Warning: package 'broom' was built under R version 4.4.3
```

```r
library(mice)
```

```
## Warning: package 'mice' was built under R version 4.4.3
```

```
##
## Attaching package: 'mice'
##
## The following object is masked from 'package:stats':
##
##     filter
##
## The following objects are masked from 'package:base':
##
##     cbind, rbind
```

```r
library(survey)
```

```
## Warning: package 'survey' was built under R version 4.4.3
```

```
## Loading required package: grid
## Loading required package: Matrix
##
## Attaching package: 'Matrix'
##
## The following objects are masked from 'package:tidyr':
##
##     expand, pack, unpack
##
## Loading required package: survival
##
## Attaching package: 'survey'
##
## The following object is masked from 'package:graphics':
##
##     dotchart
```

```r
source('data/nlsy97-educational-data/nlsy97-educational-data.R')

# Race key:
# 1 Black
# 2 Hispanic
# 3 Mixed Race (Non-Hispanic)
# 4 Non-Black

# Filter out non-responses

new_data <- new_data %>%
  mutate(degree_label = case_when(
    CV_HIGHEST_DEGREE_EVER_EDT_2017 == 0 ~ "None",
    CV_HIGHEST_DEGREE_EVER_EDT_2017 == 1 ~ "GED",
    CV_HIGHEST_DEGREE_EVER_EDT_2017 == 2 ~ "HS Diploma",
```

```r
    CV_HIGHEST_DEGREE_EVER_EDT_2017 == 3 ~ "AA",
    CV_HIGHEST_DEGREE_EVER_EDT_2017 == 4 ~ "BA",
    CV_HIGHEST_DEGREE_EVER_EDT_2017 == 5 ~ "MA",
    CV_HIGHEST_DEGREE_EVER_EDT_2017 == 6 ~ "PhD",
    TRUE ~ NA_character_
  ))

# Descriptive characteristics of respondents

# Degree attained

new_data <- new_data %>%
  mutate(degree_label = factor(degree_label,
                               levels = c("None", "GED", "HS Diploma",
                                          "AA", "BA", "MA", "PhD")))

new_data_rmNA <- new_data %>% dplyr::filter(!is.na(degree_label))

# Removes outliers
new_data_rmNA <- new_data_rmNA %>%
  filter(CV_HGC_RES_MOM_1997 <= 20 | is.na(CV_HGC_RES_MOM_1997)) %>%
  filter(CV_HGC_RES_DAD_1997 <= 20 | is.na(CV_HGC_RES_DAD_1997))


ggplot(new_data_rmNA, aes(x = degree_label)) +
  geom_bar(fill = "steelblue") +
  labs(title = "Highest Degree Attained",
       x = "Degree") +
  stat_count(geom = 'text',
             color = 'black',
             aes(label = after_stat(count)),
             position = position_stack(vjust = 1.05)) +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```
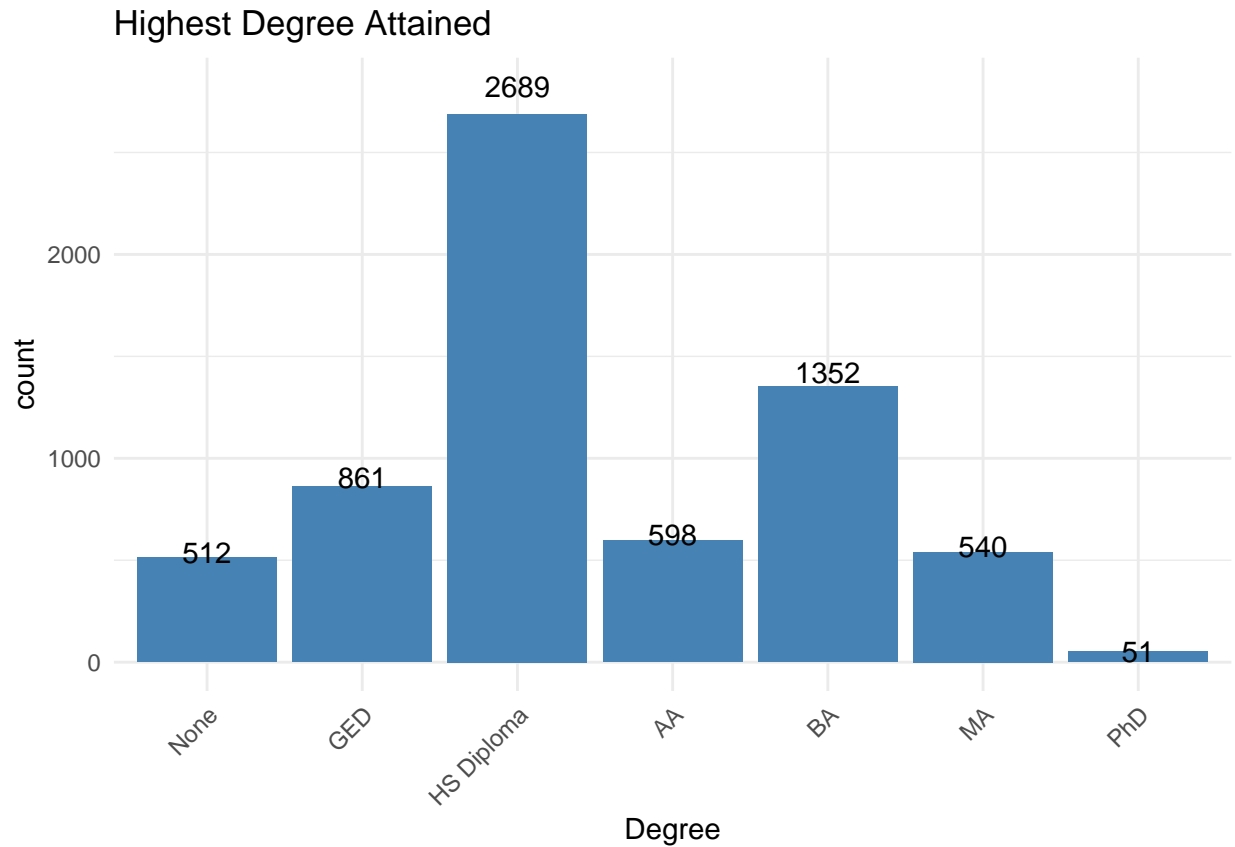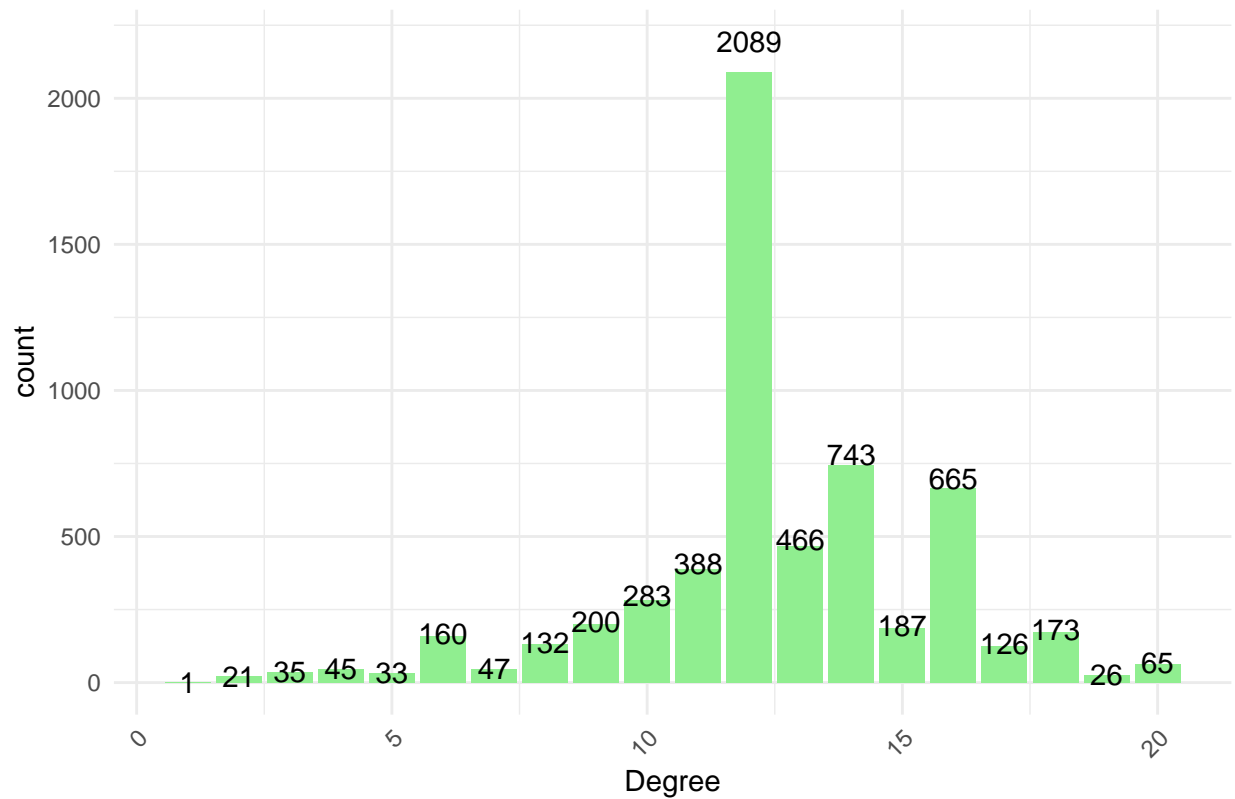
Highest Degree Attained


```
ggplot(new_data_rmNA, aes(x = CV_HGC_RES_MOM_1997)) +
  geom_bar(fill = "lightgreen") +
  labs(title = "Highest Degree Attained (Mother)",
       x = "Degree") +
  stat_count(geom = 'text',
             color = 'black',
             aes(label = after_stat(count)),
             position = position_stack(vjust = 1.05)) +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```

```
## Warning: Removed 718 rows containing non-finite outside the scale range
## ('stat_count()').
```

```
## Warning: Removed 718 rows containing non-finite outside the scale range
## ('stat_count()').
```

Highest Degree Attained (Mother)


```
ggplot(new_data_rmNA, aes(x = CV_HGC_RES_DAD_1997)) +
  geom_bar(fill = "lightpink") +
  labs(title = "Highest Degree Attained (Father)",
       x = "Degree") +
  stat_count(geom = 'text',
             color = 'black',
             aes(label = after_stat(count)),
             position = position_stack(vjust = 1.05)) +
  theme_minimal() +
  theme()
```

```
## Warning: Removed 2424 rows containing non-finite outside the scale range
## ('stat_count()').
```

```
## Warning: Removed 2424 rows containing non-finite outside the scale range
## ('stat_count()').
```

Highest Degree Attained (Father)
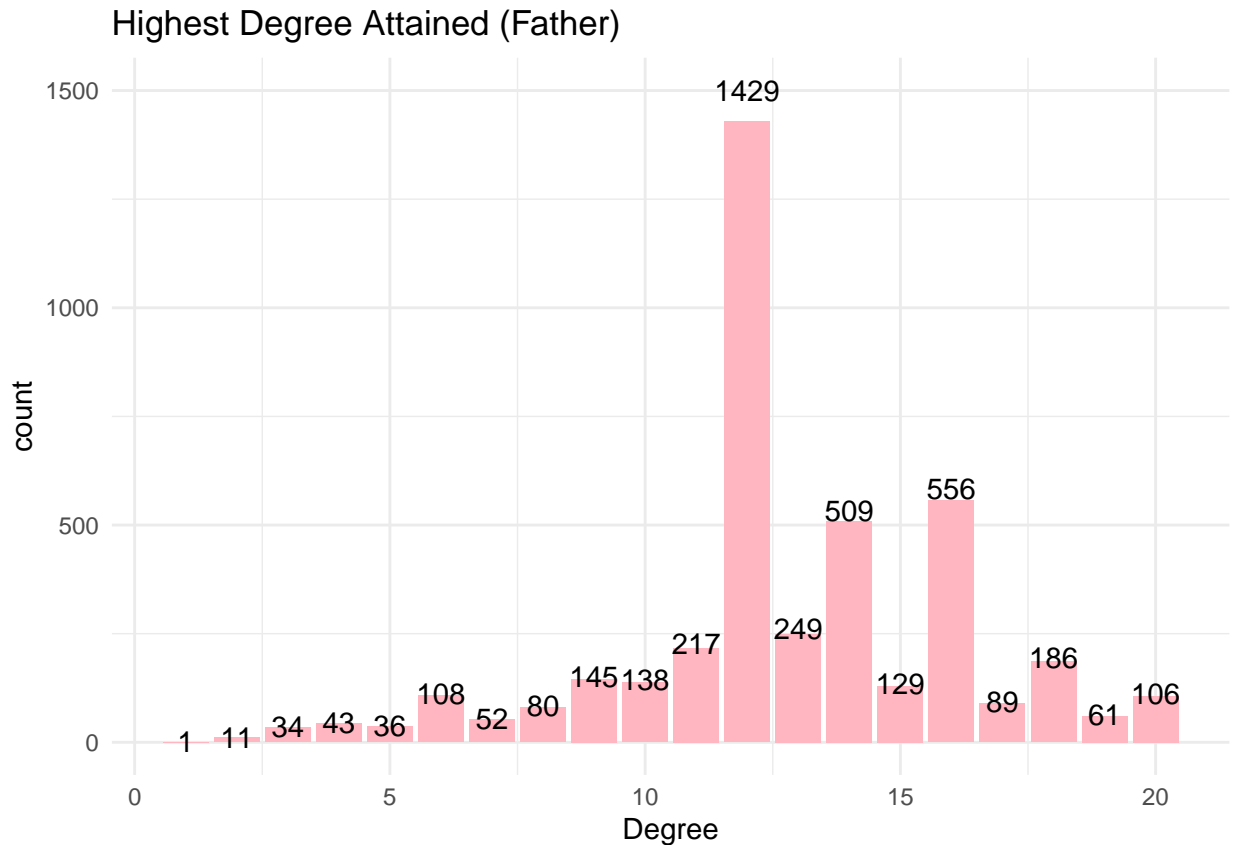

```r
########## Start of missing data analysis ##########
# Looking at missing-ness by race/ethnicity
new_data_rmNA %>%
  dplyr::select(KEY_RACE_ETHNICITY_1997, CV_HGC_RES_MOM_1997, CV_HGC_RES_DAD_1997) %>%
  mutate(
    mom_missing = is.na(CV_HGC_RES_MOM_1997),
    dad_missing = is.na(CV_HGC_RES_DAD_1997)
  ) %>%
  group_by(KEY_RACE_ETHNICITY_1997) %>%
  summarise(
    n = n(),
    mom_missing_pct = mean(mom_missing) * 100,
    dad_missing_pct = mean(dad_missing) * 100
  )
```

```
## # A tibble: 4 x 4
##   KEY_RACE_ETHNICITY_1997     n mom_missing_pct dad_missing_pct
##                     <int> <int>           <dbl>           <dbl>
## 1                       1  1808            14.7            59.9
## 2                       2  1391            12.4            40.2
## 3                       3    62            27.4            43.5
## 4                       4  3342            7.90            22.6
```

```r
# Testing for patterns in missing-ness

gg_miss_upset(new_data_rmNA)
```
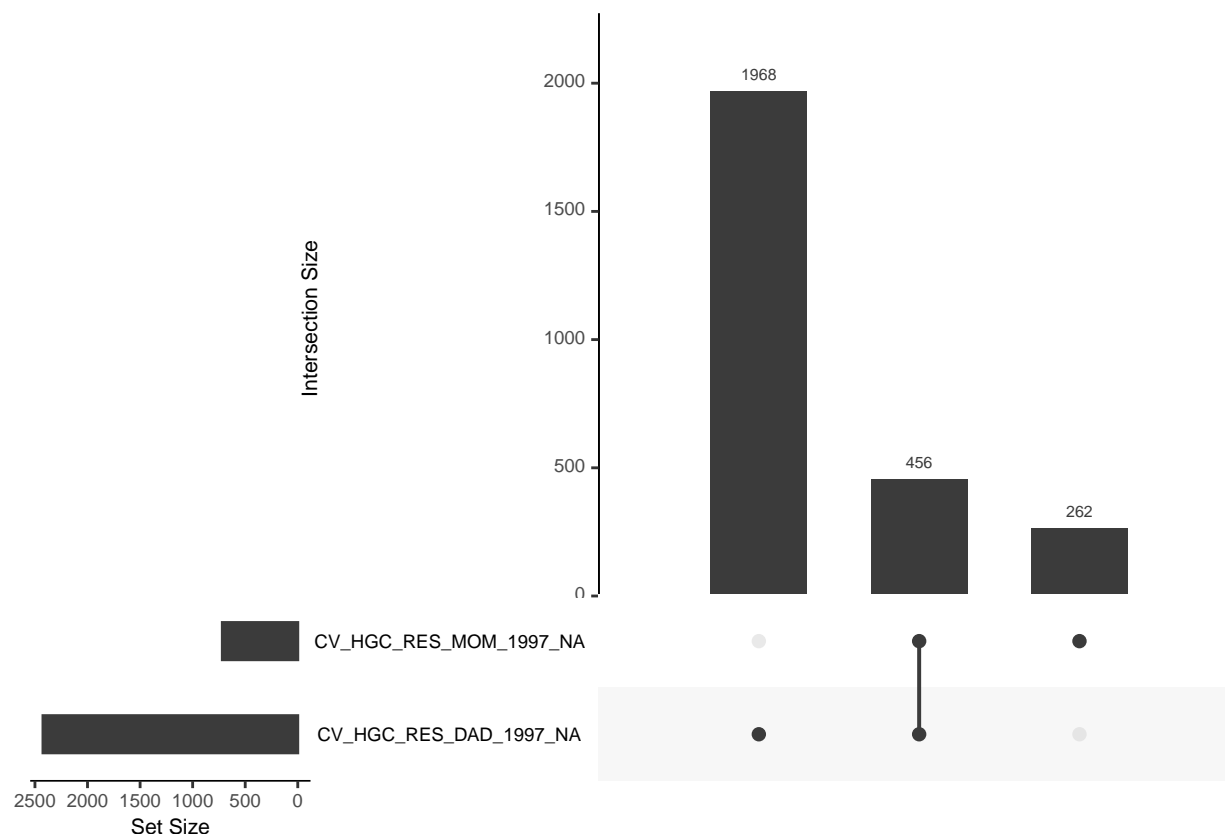
```r
mcar_data <- new_data_rmNA %>%
  dplyr::select(CV_HIGHEST_DEGREE_EVER_EDT_2017,
        CV_HGC_RES_MOM_1997,
        CV_HGC_RES_DAD_1997)

mcar_test(mcar_data)
```

```
## # A tibble: 1 x 4
##   statistic    df p.value missing.patterns
##       <dbl> <dbl>   <dbl>            <int>
## 1      364.     5       0                4
```

```r
##########  Imputations ##########
# Multiple imputations

imp_data <- new_data_rmNA %>%
  dplyr::select(CV_HIGHEST_DEGREE_EVER_EDT_2017,
        KEY_RACE_ETHNICITY_1997, CV_HGC_RES_MOM_1997,
        CV_HGC_RES_DAD_1997, SAMPLING_WEIGHT_CC_2017)

imp <- mice(imp_data, m = 5, method = 'pmm')
```

```
##
##  iter imp variable
##   1   1  CV_HGC_RES_MOM_1997  CV_HGC_RES_DAD_1997
```

```
## 1   2  CV_HGC_RES_MOM_1997  CV_HGC_RES_DAD_1997
## 1   3  CV_HGC_RES_MOM_1997  CV_HGC_RES_DAD_1997
## 1   4  CV_HGC_RES_MOM_1997  CV_HGC_RES_DAD_1997
## 1   5  CV_HGC_RES_MOM_1997  CV_HGC_RES_DAD_1997
## 2   1  CV_HGC_RES_MOM_1997  CV_HGC_RES_DAD_1997
## 2   2  CV_HGC_RES_MOM_1997  CV_HGC_RES_DAD_1997
## 2   3  CV_HGC_RES_MOM_1997  CV_HGC_RES_DAD_1997
## 2   4  CV_HGC_RES_MOM_1997  CV_HGC_RES_DAD_1997
## 2   5  CV_HGC_RES_MOM_1997  CV_HGC_RES_DAD_1997
## 3   1  CV_HGC_RES_MOM_1997  CV_HGC_RES_DAD_1997
## 3   2  CV_HGC_RES_MOM_1997  CV_HGC_RES_DAD_1997
## 3   3  CV_HGC_RES_MOM_1997  CV_HGC_RES_DAD_1997
## 3   4  CV_HGC_RES_MOM_1997  CV_HGC_RES_DAD_1997
## 3   5  CV_HGC_RES_MOM_1997  CV_HGC_RES_DAD_1997
## 4   1  CV_HGC_RES_MOM_1997  CV_HGC_RES_DAD_1997
## 4   2  CV_HGC_RES_MOM_1997  CV_HGC_RES_DAD_1997
## 4   3  CV_HGC_RES_MOM_1997  CV_HGC_RES_DAD_1997
## 4   4  CV_HGC_RES_MOM_1997  CV_HGC_RES_DAD_1997
## 4   5  CV_HGC_RES_MOM_1997  CV_HGC_RES_DAD_1997
## 5   1  CV_HGC_RES_MOM_1997  CV_HGC_RES_DAD_1997
## 5   2  CV_HGC_RES_MOM_1997  CV_HGC_RES_DAD_1997
## 5   3  CV_HGC_RES_MOM_1997  CV_HGC_RES_DAD_1997
## 5   4  CV_HGC_RES_MOM_1997  CV_HGC_RES_DAD_1997
## 5   5  CV_HGC_RES_MOM_1997  CV_HGC_RES_DAD_1997
```

```r
imp <- complete(imp, action = "long", include = TRUE)

imp$CV_HIGHEST_DEGREE_EVER_EDT_2017 <- factor(
  imp$CV_HIGHEST_DEGREE_EVER_EDT_2017,
  levels = 0:6,
  labels = c("None", "GED", "HS", "AA", "BA", "MA", "PhD"),
  ordered = TRUE
)

# Re-convert to mids object
imp <- as.mids(imp)

# Runs ordinal log regr on imp data
pom_imp <- with(imp, polr(
  CV_HIGHEST_DEGREE_EVER_EDT_2017 ~ CV_HGC_RES_MOM_1997 +
    KEY_RACE_ETHNICITY_1997 + CV_HGC_RES_DAD_1997,
  Hess = TRUE
))

# Pool the results
pom_pooled <- pool(pom_imp)
summary(pom_pooled)
```

```
##                    term  estimate  std.error statistic        df
## 1     CV_HGC_RES_MOM_1997 0.1124828 0.01316745  8.542487  26.12942
## 2 KEY_RACE_ETHNICITY_1997 0.1573360 0.01925781  8.169982 198.95235
## 3     CV_HGC_RES_DAD_1997 0.1464721 0.01376937 10.637533  12.98876
## 4             None|GED 0.9374694 0.11677139  8.028245 713.19784
## 5              GED|HS 2.1534857 0.11516447 18.699219 515.45461
```
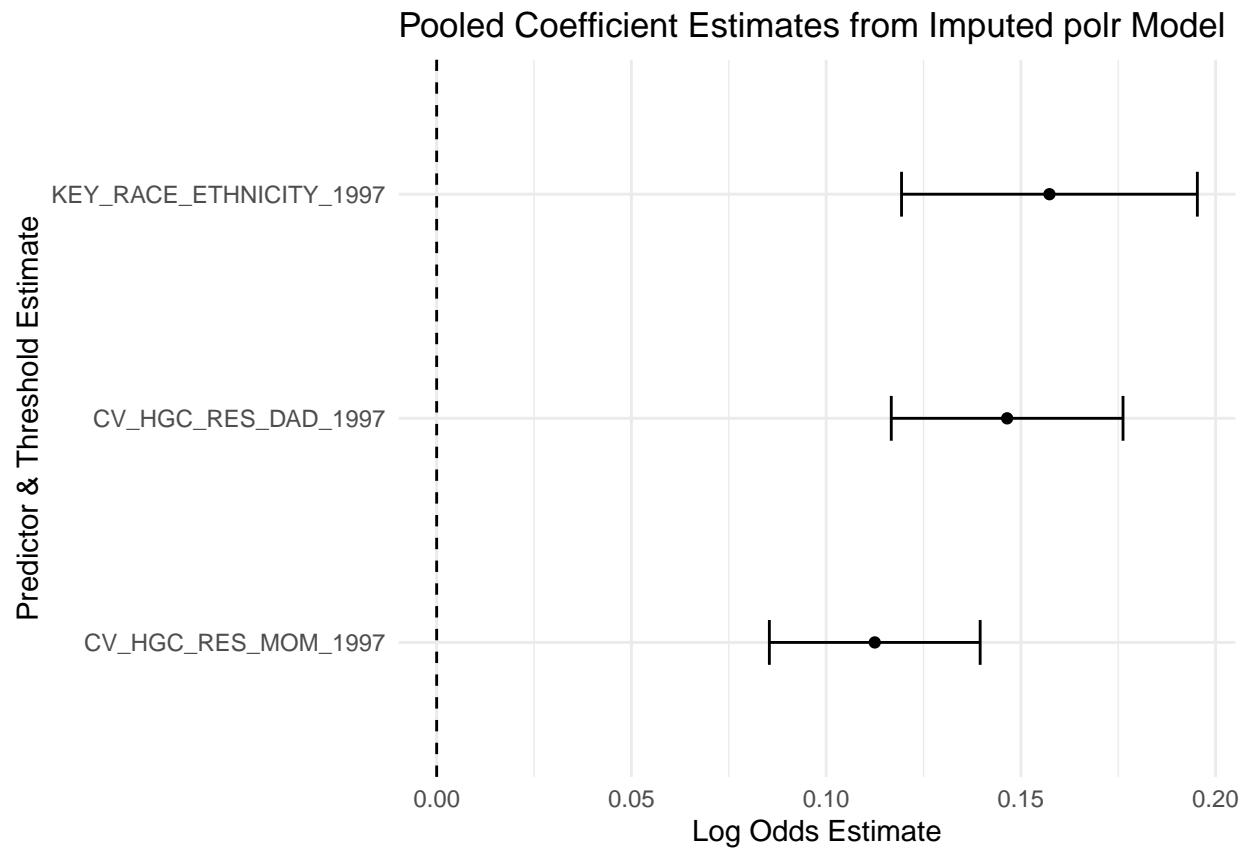
```
## 6                     HS|AA 4.1927522 0.12375241 33.880165   513.50049
## 7                     AA|BA 4.6604037 0.12656825 36.821271   492.42738
## 8                     BA|MA 6.2598687 0.13769654 45.461336   478.72429
## 9                    MA|PhD 8.8617108 0.19426315 45.617045 1371.14852
##        p.value
## 1  4.857036e-09
## 2  3.525224e-14
## 3  8.814373e-08
## 4  4.078123e-15
## 5  6.060005e-60
## 6 5.048183e-133
## 7 1.562305e-143
## 8 8.059260e-176
## 9 3.401583e-277
```

```r
# Convert pooled polr results to tidy format
pooled_summary <- summary(pom_pooled)

# Add term names
tidy_pooled <- tidy(pom_pooled, conf.int = TRUE, conf.level = 0.95)

tidy_pooled_sub <- subset(tidy_pooled, tidy_pooled$estimate <= 0.8)

# Plot
ggplot(tidy_pooled_sub, aes(x = estimate, y = reorder(term, estimate))) +
  geom_point() +
  geom_errorbarh(aes(xmin = conf.low, xmax = conf.high), height = 0.2) +
  geom_vline(xintercept = 0, linetype = "dashed") +
  labs(
    title = "Pooled Coefficient Estimates from Imputed polr Model",
    x = "Log Odds Estimate",
    y = "Predictor & Threshold Estimate"
  ) +
  theme_minimal()
```

Pooled Coefficient Estimates from Imputed polr Model


```
tidy_pooled_sub <- subset(tidy_pooled, tidy_pooled$estimate > 0.8)

# Plot
ggplot(tidy_pooled_sub, aes(x = estimate, y = reorder(term, estimate))) +
  geom_point() +
  geom_errorbarh(aes(xmin = conf.low, xmax = conf.high), height = 0.2) +
  geom_vline(xintercept = 0, linetype = "dashed") +
  labs(
    title = "Pooled Coefficient Estimates from Imputed polr Model",
    x = "Log Odds Estimate",
    y = "Predictor & Threshold Estimate"
  ) +
  theme_minimal()
```
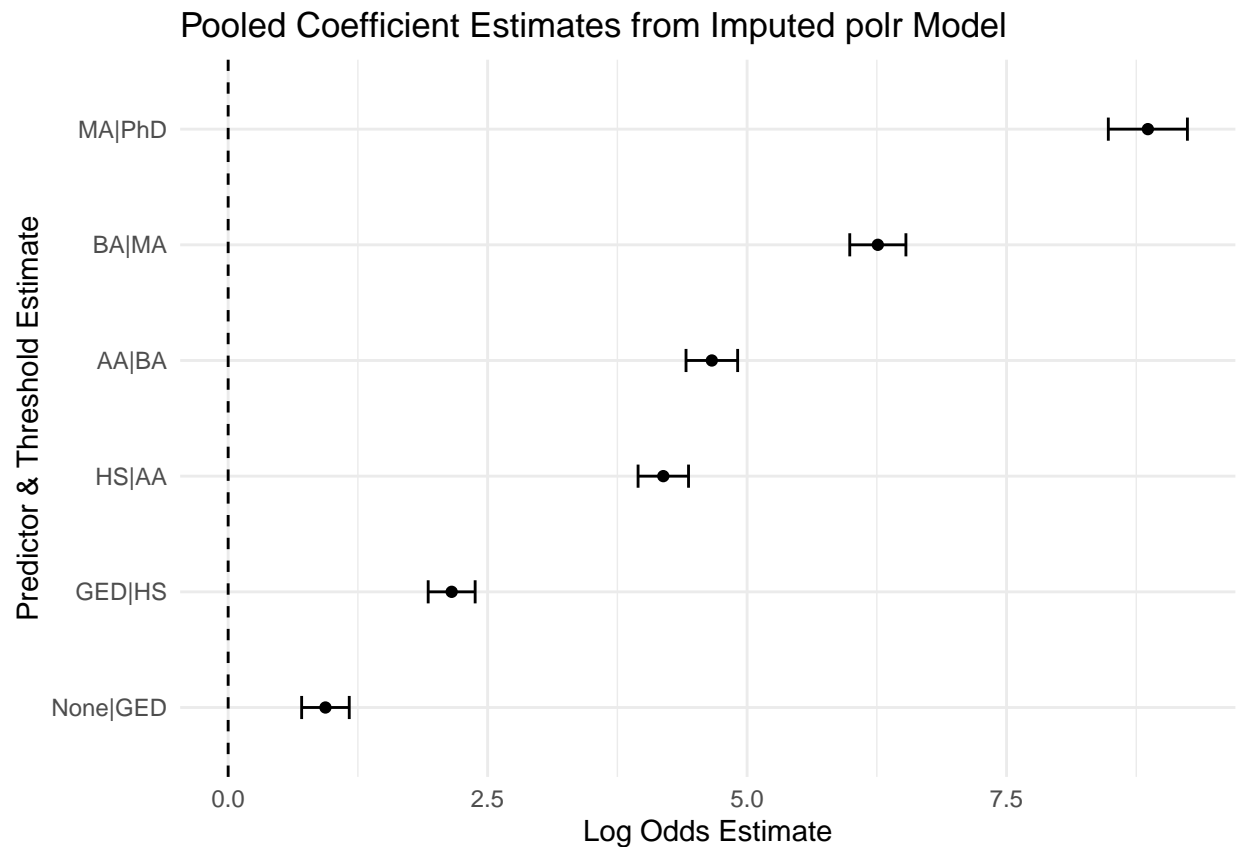
Pooled Coefficient Estimates from Imputed polr Model


```r
# Selects an imp
completed_data <- complete(imp, action = 5L)

# Weighted GLM
completed_data$degree_num <- as.numeric(
  completed_data$CV_HIGHEST_DEGREE_EVER_EDT_2017)

svy_design <- svydesign(
  ids = ~1,
  weights = ~SAMPLING_WEIGHT_CC_2017,
  data = completed_data
)

svy_model <- svyglm(
  degree_num ~ CV_HGC_RES_MOM_1997 +
    KEY_RACE_ETHNICITY_1997 + CV_HGC_RES_DAD_1997,
  design = svy_design
)

summary(svy_model)
```

```
##
## Call:
## svyglm(formula = degree_num ~ CV_HGC_RES_MOM_1997 + KEY_RACE_ETHNICITY_1997 +
##     CV_HGC_RES_DAD_1997, design = svy_design)
##
```

```
## Survey design:
## svydesign(ids = ~1, weights = ~SAMPLING_WEIGHT_CC_2017, data = completed_data)
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)              0.600724   0.084230   7.132 1.10e-12 ***
## CV_HGC_RES_MOM_1997      0.099820   0.008536  11.694  < 2e-16 ***
## KEY_RACE_ETHNICITY_1997  0.089688   0.013247   6.770 1.39e-11 ***
## CV_HGC_RES_DAD_1997      0.110427   0.007290  15.148  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 1.635078)
##
## Number of Fisher Scoring iterations: 2
```

```
# Visualize pooled regression coefficients of imps (not done yet)

library(effects)
```

```
## Warning: package 'effects' was built under R version 4.4.3
```

```
## Loading required package: carData
```

```
## Warning: package 'carData' was built under R version 4.4.3
```

```
## lattice theme set by effectsTheme()
## See ?effectsTheme for details.
```

```
# Use one completed data set as demonstration
effs <- Effect(c("KEY_RACE_ETHNICITY_1997", "CV_HGC_RES_MOM_1997"),
               svy_model)

# For predicted probabilities
as.data.frame(effs)
```

```
##    KEY_RACE_ETHNICITY_1997 CV_HGC_RES_MOM_1997      fit         se    lower
## 1                      1.0                 1.0 2.229896 0.10281198 2.028351
## 2                      1.8                 1.0 2.301646 0.10217012 2.101360
## 3                      2.5                 1.0 2.364428 0.10250789 2.163479
## 4                      3.2                 1.0 2.427209 0.10367725 2.223968
## 5                      4.0                 1.0 2.498960 0.10599574 2.291174
## 6                      1.0                 5.8 2.709034 0.06464845 2.582302
## 7                      1.8                 5.8 2.780784 0.06272719 2.657818
## 8                      2.5                 5.8 2.843566 0.06248859 2.721068
## 9                      3.2                 5.8 2.906347 0.06361540 2.781640
## 10                     4.0                 5.8 2.978098 0.06648226 2.847771
## 11                     1.0                10.0 3.128279 0.03663149 3.056470
## 12                     1.8                10.0 3.200030 0.03159326 3.138097
## 13                     2.5                10.0 3.262811 0.02969215 3.204605
## 14                     3.2                10.0 3.325593 0.03061186 3.265584
## 15                     4.0                10.0 3.397343 0.03480013 3.329124
```

```
## 16                     1.0            15.0 3.627381 0.03627416 3.556272
## 17                     1.8            15.0 3.699132 0.02922695 3.641837
## 18                     2.5            15.0 3.761913 0.02519067 3.712531
## 19                     3.2            15.0 3.824695 0.02422596 3.777204
## 20                     4.0            15.0 3.896445 0.02725787 3.843011
## 21                     1.0            20.0 4.126483 0.07023698 3.988796
## 22                     1.8            20.0 4.198233 0.06598307 4.068885
## 23                     2.5            20.0 4.261015 0.06349005 4.136554
## 24                     3.2            20.0 4.323797 0.06229108 4.201686
## 25                     4.0            20.0 4.395547 0.06259716 4.272836
##       upper
## 1  2.431441
## 2  2.501933
## 3  2.565376
## 4  2.630450
## 5  2.706746
## 6  2.835766
## 7  2.903750
## 8  2.966063
## 9  3.031054
## 10 3.108424
## 11 3.200089
## 12 3.261963
## 13 3.321017
## 14 3.385602
## 15 3.465563
## 16 3.698490
## 17 3.756426
## 18 3.811295
## 19 3.872185
## 20 3.949879
## 21 4.264170
## 22 4.327582
## 23 4.385476
## 24 4.445907
## 25 4.518258
```