analysis.R

byrds

2025-05-06

```
# 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 error
```

```
library(extrafont)
```

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

```
library(ggthemes)
```

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

```
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
```

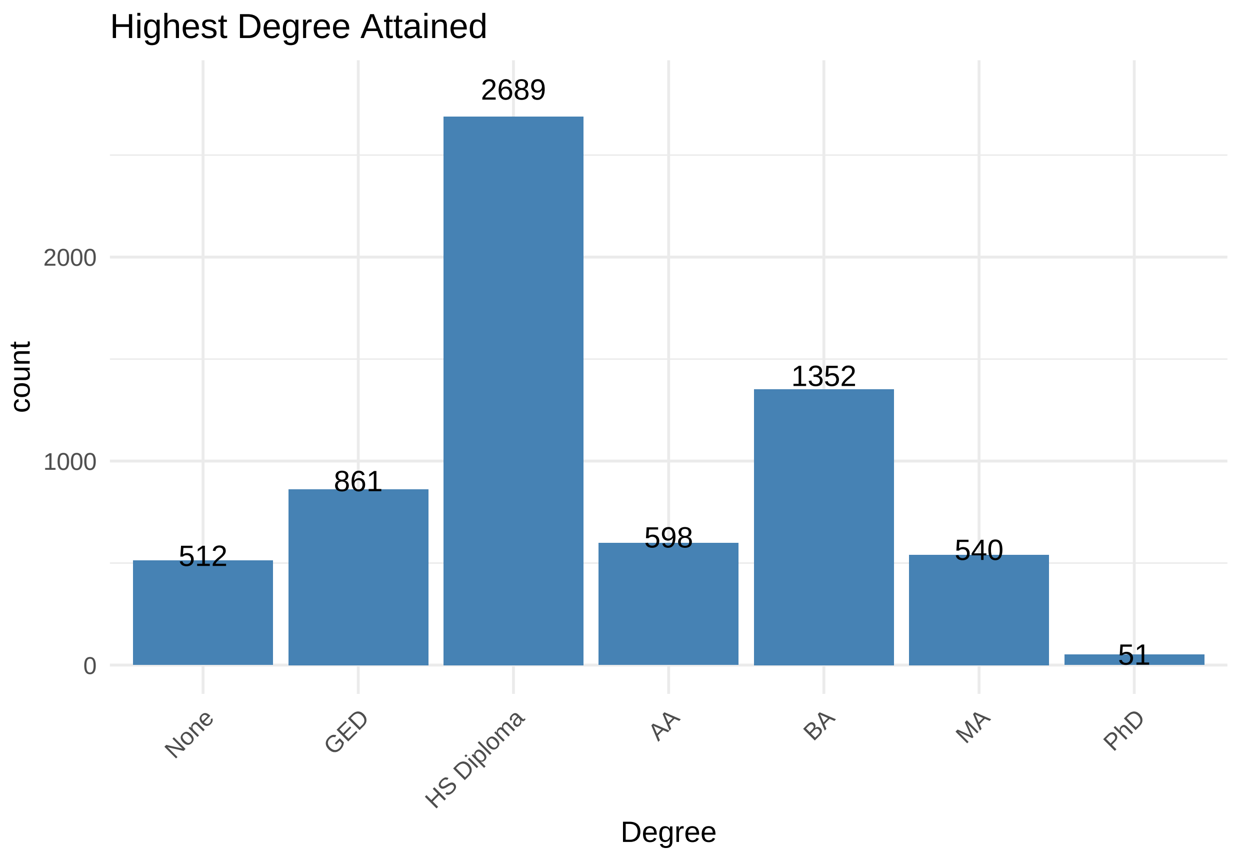
```
##
## Attaching package: ’mice’
##
## The following object is masked from ’package:stats’:
##
##     filter
##
## The following objects are masked from ’package:base’:
##
##     cbind, rbind
```

```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
```

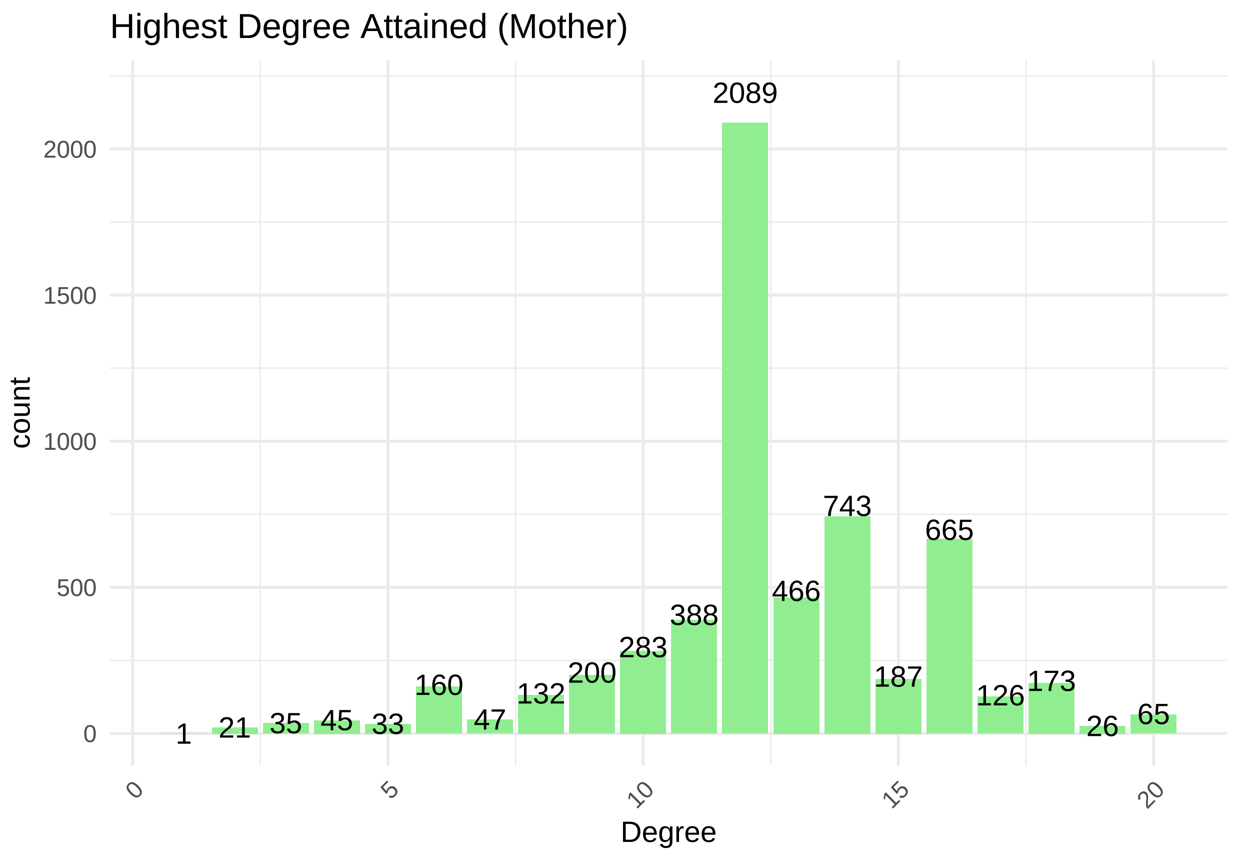
```r
source('data/nlsy97-educational-data/nlsy97-educational-data.R')

# 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",
    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))
```



```r
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()`).
```

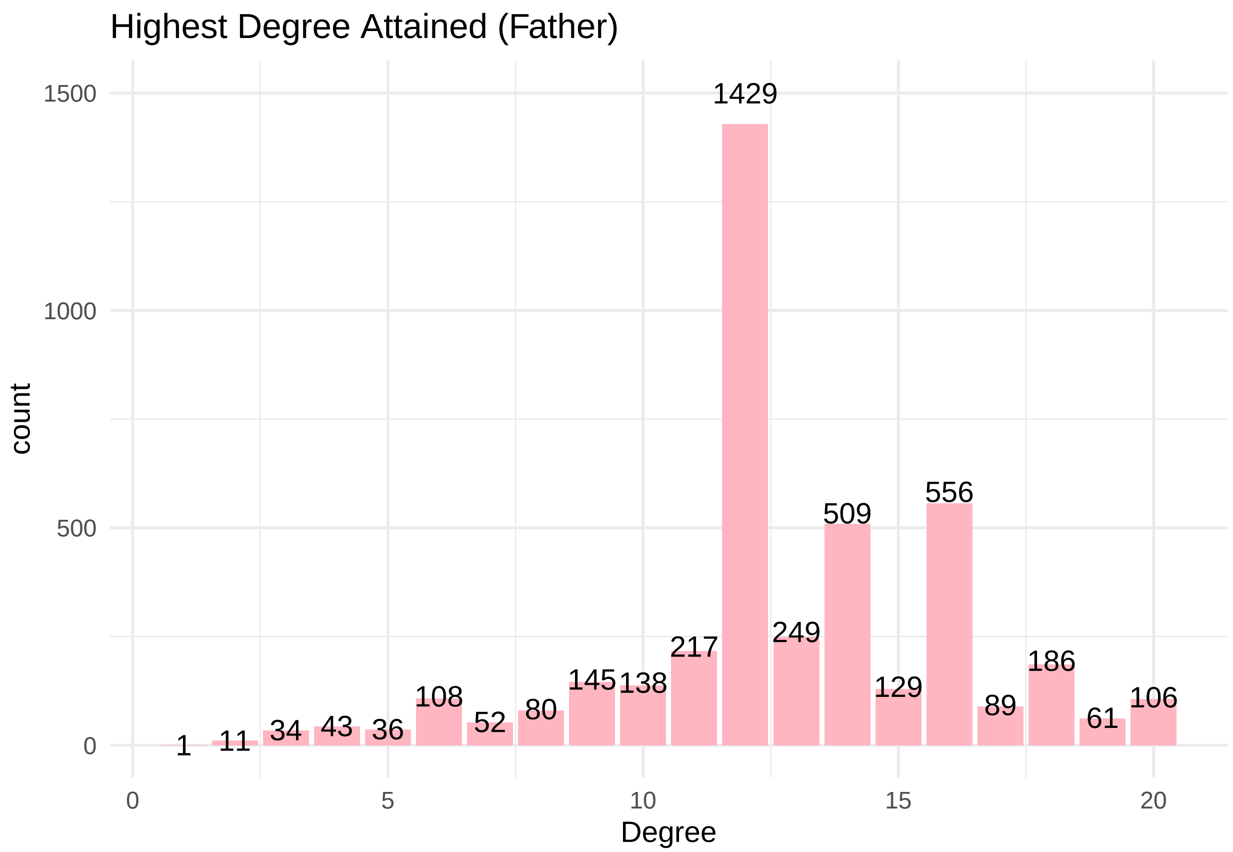
```
## Warning: Removed 718 rows containing non-finite outside the scale range
## (`stat_count()`).
```



```r
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()‘).
```

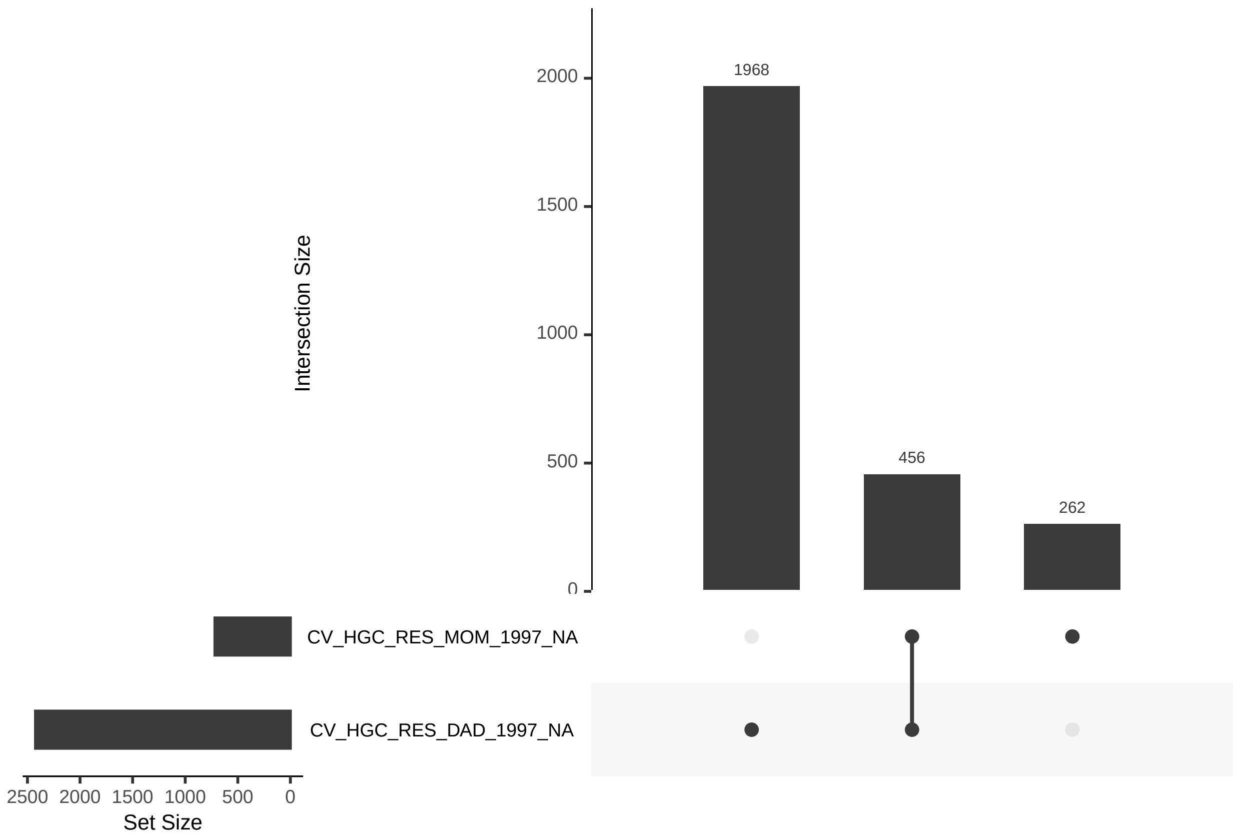
```
## Warning: Removed 2424 rows containing non-finite outside the scale range
## (‘stat_count()‘).
```



```
##########  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
```

```
# Testing for patterns in missing-ness

gg_miss_upset(new_data_rmNA)
```



```
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
```

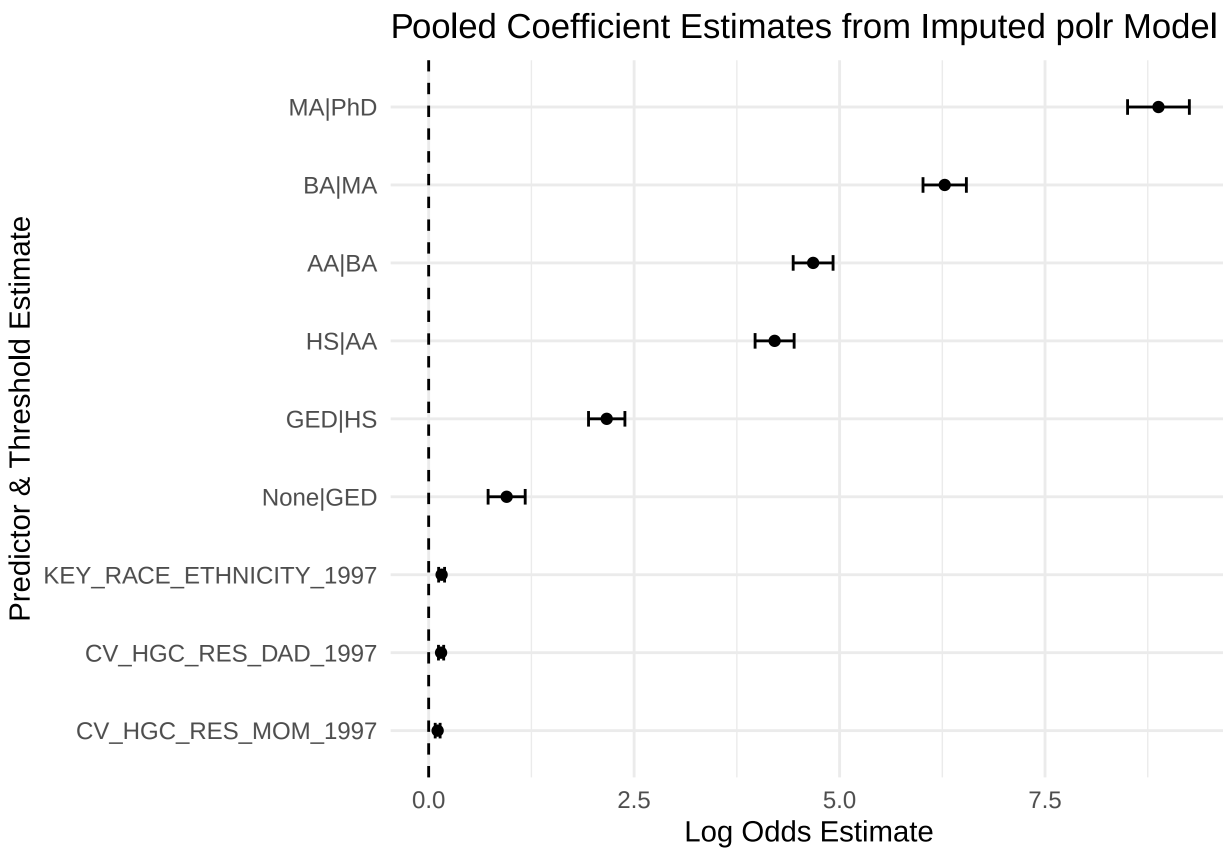
```
##########  Imputations ##########
# Multiple imputations

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 0.1092281 0.01416971  7.708563   18.16440
## 2 KEY_RACE_ETHNICITY_1997 0.1574861 0.01838871  8.564280  990.78938
## 3     CV_HGC_RES_DAD_1997 0.1509109 0.01448973 10.415028   11.49436
## 4                None|GED 0.9487487 0.11504080  8.247063 1452.32795
## 5                  GED|HS 2.1664620 0.11285281 19.197236 1297.90354
## 6                   HS|AA 4.2094128 0.12105771 34.771953 1515.82311
## 7                   AA|BA 4.6779871 0.12375460 37.800511 1469.96994
## 8                   BA|MA 6.2789331 0.13453656 46.670833 1493.44378
## 9                  MA|PhD 8.8801355 0.19169006 46.325487 3511.27825
##          p.value
## 1  3.899334e-07
## 2  4.109718e-17
## 3  3.363792e-07
## 4  3.594821e-16
## 5  1.699098e-72
## 6 2.783467e-195
## 7 5.151217e-219
## 8 5.102963e-294
## 9  0.000000e+00
```

```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)

# Plot
ggplot(tidy_pooled, 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()
```



```
# 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)             0.622069   0.083500   7.450 1.05e-13 ***
## CV_HGC_RES_MOM_1997     0.098013   0.008429  11.628  < 2e-16 ***
## KEY_RACE_ETHNICITY_1997 0.088522   0.013140   6.737 1.75e-11 ***
## CV_HGC_RES_DAD_1997     0.110790   0.007047  15.722  < 2e-16 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## (Dispersion parameter for gaussian family taken to be 1.637104)
##
## 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 2.254542 0.10158480 2.055403
## 2                      1.8                 1.0 2.325360 0.10085714 2.127648
## 3                      2.5                 1.0 2.387326 0.10111652 2.189105
## 4                      3.2                 1.0 2.449291 0.10220636 2.248934
## 5                      4.0                 1.0 2.520109 0.10443452 2.315383
## 6                      1.0                 5.8 2.725006 0.06389674 2.599748
## 7                      1.8                 5.8 2.795824 0.06191330 2.674454
## 8                      2.5                 5.8 2.857789 0.06161329 2.737007
## 9                      3.2                 5.8 2.919755 0.06267646 2.796889
## 10                     4.0                 5.8 2.990573 0.06547167 2.862227
## 11                     1.0                10.0 3.136662 0.03621550 3.065668
## 12                     1.8                10.0 3.207480 0.03118617 3.146345
## 13                     2.5                10.0 3.269445 0.02927884 3.212049
## 14                     3.2                10.0 3.331411 0.03018483 3.272239
## 15                     4.0                10.0 3.402228 0.03434933 3.334893
## 16                     1.0                15.0 3.626728 0.03579088 3.556567
## 17                     1.8                15.0 3.697546 0.02890529 3.640882
## 18                     2.5                15.0 3.759512 0.02504024 3.710425
## 19                     3.2                15.0 3.821477 0.02424275 3.773953
## 20                     4.0                15.0 3.892295 0.02738571 3.838610
## 21                     1.0                20.0 4.116795 0.06930469 3.980935
## 22                     1.8                20.0 4.187612 0.06520109 4.059797
## 23                     2.5                20.0 4.249578 0.06284508 4.126381
## 24                     3.2                20.0 4.311543 0.06178208 4.190431
## 25                     4.0                20.0 4.382361 0.06223187 4.260367
##       upper
## 1  2.453682
## 2  2.523073
## 3  2.585547
## 4  2.649649
## 5  2.724834
## 6  2.850264
## 7  2.917194
## 8  2.978571
## 9  3.042621
## 10 3.118918
## 11 3.207656
## 12 3.268615
## 13 3.326841
## 14 3.390583
## 15 3.469564
## 16 3.696890
## 17 3.754210
## 18 3.808599
## 19 3.869001
## 20 3.945980
## 21 4.252654
## 22 4.315428
## 23 4.372775
## 24 4.432656
## 25 4.504356
```