R Notebook

Contents

```
rm(list=ls())
```

```
library(dplyr)
library(forcats)
library(survey)
library(tidyr)
library(car)
library(haven)
library(survey)
library(tidyr)
library(broom)
library(ggplot2)
library(ggeffects)
library(ggeffects)
library(Hmisc)
library(readxl)

# new 2023 ACS data
undoc23e11 <- read_dta("/Users/estellepan/Desktop/AIC/undocumented_student/undoc_2023_acs_02_27_2025.dta

# state-level enrollment & employment data
state_enroll <- read.csv("/Users/estellepan/Desktop/AIC/undocumented_student/state_enroll_long(Sheet1).
```

```
# Filter for undocumented youth aged 18-24
# who have completed high school (GED or diploma)
# but have NOT completed a college degree

undoc_youth <- undoc23e11 %>%
  filter(
    undoc2 == 1,                    # Undocumented immigrants (clean, NA-free flag)
    age >= 18 & age <= 24,          # Target age range per study definition
    educd %in% c(                   # Educational attainment codes (2023 ACS only):
      063,  # Regular high school diploma
      064,  # GED or alternative credential
      065,  # Some college, less than 1 year
      071,  # 1+ years of college credit, no degree
      081   # Associate's degree, type not specified
    )
  )

##  People who completed HS and are in college or have some college, exclude people who already have co
```

```
# dataset excluding people who already have college degrees
undoc23e1 <- undoc_youth %>%
  left_join(state_enroll, by = c("statefip" = "StateFIP"))
```

```
# remove labelled metadata
undoc23e1<-zap_labels(undoc23e1)
undoc23e1$state_employ_rate<-zap_formats(undoc23e1$state_employ_rate)
undoc23e1$state_enroll_rate<-zap_formats(undoc23e1$state_enroll_rate)
summary(undoc23e1$state_employ_rate)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.5170  0.5970  0.6020  0.6077  0.6260  0.6820
```

```r
summary(undoc23e1$state_enroll_rate)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.1970  0.2400  0.2670  0.2684  0.2950  0.3760
```

```r
# Explore Key Variables
summary(undoc23e1$age)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   18.00   20.00   21.00   21.16   23.00   24.00
```

```r
table(undoc23e1$undoc)
```

```
##
##    1
## 5704
```

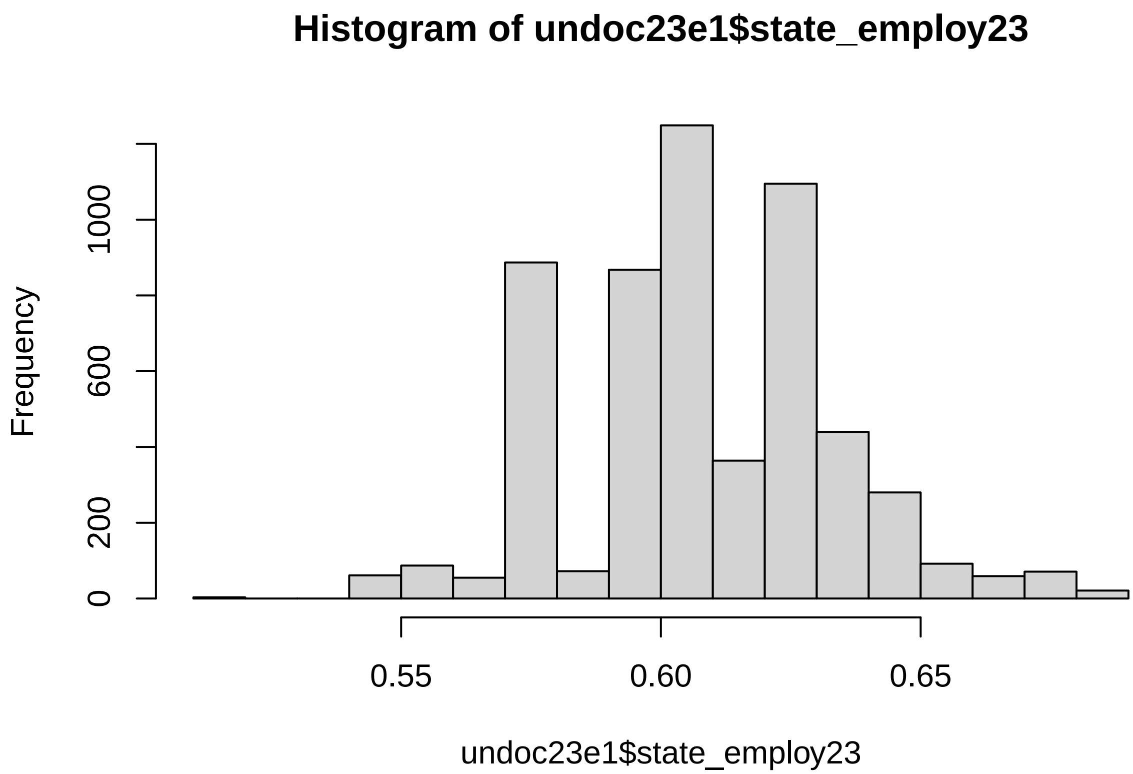
```r
colSums(is.na(undoc23e1))
```

```
##              year            sample            serial          cbserial
##                 0                 0                 0                 0
##           numprec              hhwt           cluster            region
##                 0                 0                 0                 0
##          statefip         countyfip              puma            strata
##                 0                 0                 0                 0
##                gq          hhincome          foodstmp            pernum
##                 0                 0                 0                 0
##             perwt            momloc            poploc             sploc
##                 0                 0                 0                 0
##           momloc2           poploc2            relate           related
##                 0                 0                 0                 0
##               sex               age          birthqtr             marst
##                 0                 0                 0                 0
##           birthyr            marrno            yrmarr              race
##                 0                 0                 0                 0
##             raced            hispan           hispand               bpl
##                 0                 0                 0                 0
##              bpld          ancestr1         ancestr1d          ancestr2
##                 0                 0                 0                 0
##         ancestr2d           citizen           yrnatur           yrimmig
##                 0                 0                 0                 0
##           yrsusa1           yrsusa2          speakeng           hcovany
##                 0                 0                 0                 0
##           hinstri          hinscaid          hinscare            hinsva
##                 0                 0                 0                 0
##            school              educ             educd          gradeatt
##                 0                 0                 0                 0
##         gradeattd          schltype          degfield         degfieldd
##                 0                 0                 0                 0
##         degfield2        degfield2d           empstat          empstatd 
##                 0                 0                 0                 0 
##          labforce          classwkr         classwkrd               occ 
##                 0                 0                 0                 0 
##           occ2010               ind          indnaics          uhrswork 
##                 0                 0                 0                 0 
##          wrklstwk          workedyr            inctot           incwage 
##                 0                 0                 0                 0 
##             incss          incwelfr           incsupp          migplac1 
##                 0                 0                 0                 0 
##          migpuma1           movedin           vetstat          vetstatd 
##                 0                 0                 0                 0 
##          qclasswk          qworkedy  citizen_original                fb 
##                 0                 0                 0                 0 
##           non_cit            sploc2          apartnum        sp_related 
##                 0                 0              5503              5503 
##        sp_citizen        sp_yrimmig          yearinus       sp_yearinus 
##              5503              5518                 0              5518 
##        yrsmarried               mex            sp_mex          natcheck 
##                 0                 0              5503                 0 
##            cit_or            cond_a           mom_cit          mom2_cit 
##                 0                 0              2943              5700 
##           pop_cit          pop2_cit          nativept            cond_b 
##              3315              5692                 0                 0 
##            cond_c       CHE_benefit    AFGHAN_benefit            cond_d 
##                 0                 0                 0                 0 
##            cond_e           refugee       refugeetype        ethnic_ref 
##                 0                 0              5703              5703 
##       mom_refugee      mom2_refugee       pop_refugee      pop2_refugee 
##              2943              5700              3315              5692 
##       mom_yrimmig      mom2_yrimmig       pop_yrimmig      pop2_yrimmig 
##              2976              5701              3424              5692 
##       dfyrimm_mom      dfyrimm_mom2       dfyrimm_pop      dfyrimm_pop2 
##              2976              5701              3424              5692 
##       marr_momarr       marr_poparr      marr_mom2arr      marr_pop2arr 
##              2976              3424              5701              5692 
##        age_momarr        age_poparr       age_mom2arr       age_pop2arr 
##              2976              3424              5701              5692 
##     child_refugee               siv           mom_siv          mom2_siv 
##                 0                 0              2943              5700 
##           pop_siv          pop2_siv         child_siv            cond_f 
##              3315              5692                 0                 0 
##            cond_g                z1         cond_g_1A            mom_g1 
##                 0                 0                 0              2943 
##           mom2_g1            pop_g1           pop2_g1         g1_parent 
##              5700              3315              5692                 0 
##         cond_g_1B         cond_g_1C                eu            cond_h 
##                 0              5503                 0                 0 
## attending_college           longUSA      parents_home         over20hrs 
##                 0                 0                 0                 0 
##       int_student            cond_i           cond_ai           x1_flag 
##                 0                 0                 0              5704 
##        sp_newcond          cond_all             legal             undoc 
##              5689                 0                 0                 0
##             undoc2             legal2              natur           citizen2 
##                  0                  0                  0                  0 
##          mom_undoc         mom2_undoc          pop_undoc         pop2_undoc 
##               2943               5700               3315               5692 
##        child_undoc        cond_c_daca                 x1                 x2 
##                  0                  0                  0                  0 
##        cond_e_daca       cond_ai_daca      cond_all_daca         legal_daca 
##                  0                  0                  0                  0 
##         undoc_daca         age31_yr12        age_arrival              us_16 
##                  0                  0                  0                  0 
##         lived_5yrs           cond_edu           daca_imm           daca_all 
##                  0                  0                  0                  0 
##     yrimmig_period              state state_enroll_rate state_employ_rate 
##                  0                  0                  0                  0
```

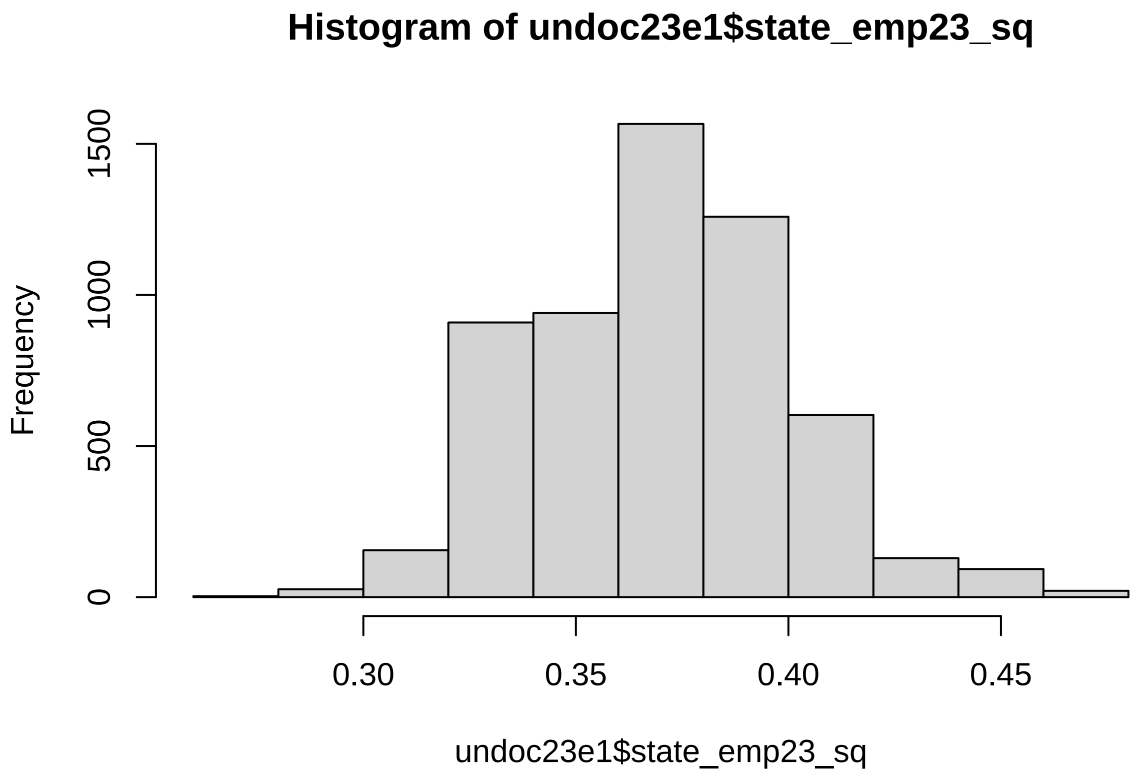
Cleaning variables again and doing some prelim stats before regressions

Re-making state employ into a factor incase of linearity of logit violations

```
undoc23e1$state_employ23 <- as.numeric(undoc23e1$state_employ_rate)
hist(undoc23e1$state_employ23)
```



```
# Why square it? To capture curvature in the relationship if state_employ_rate has a nonlinear effect o
undoc23e1$state_emp23_sq<-undoc23e1$state_employ23^2
hist(undoc23e1$state_emp23_sq)
```



```
summary(undoc23e1$state_employ23)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.5170  0.5970  0.6020  0.6077  0.6260  0.6820
```

```
undoc23e1$state_employ23_cat<-cut(undoc23e1$state_employ23,
                           breaks=3,
                           labels=c("1","2","3"))

table(undoc23e1$state_employ23_cat)
```

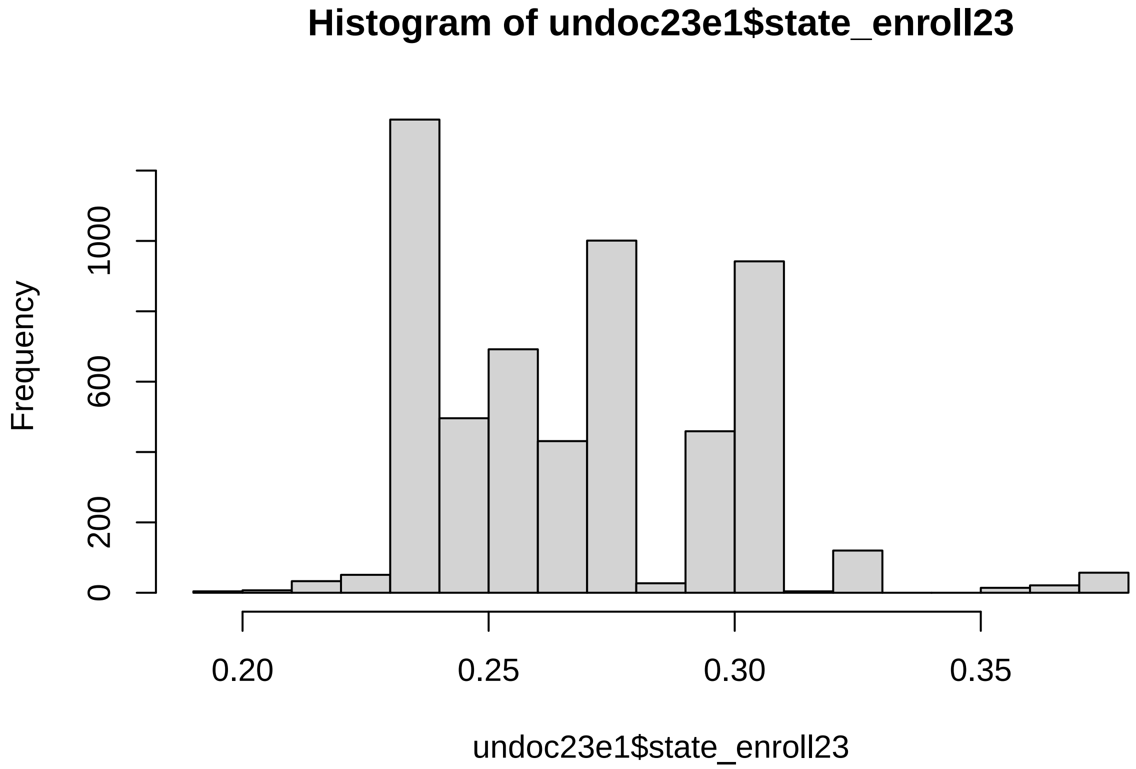
```
##
##    1    2    3
##  845 3896  963
```

```
undoc23e1%>%group_by(state_employ23_cat)%>%
  summarise(min=min(state_employ23),
            max=max(state_employ23))
```

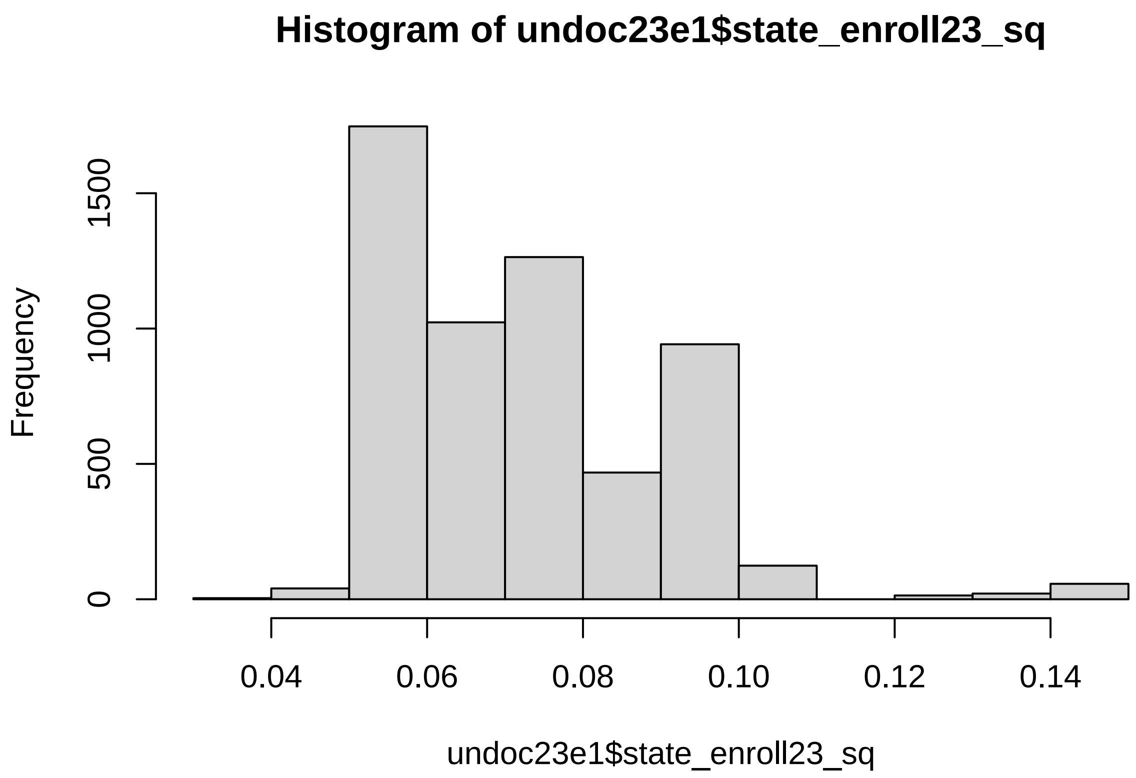
```
## # A tibble: 3 x 3
##   state_employ23_cat   min   max
##   <fct>              <dbl> <dbl>
## 1 1                  0.517 0.572
## 2 2                  0.574 0.626
## 3 3                  0.631 0.682
```

remaking state enrollment into a factor incase of linearity of logit violations

```
undoc23e1$state_enroll23 <- as.numeric(undoc23e1$state_enroll_rate)
hist(undoc23e1$state_enroll23)
```



```
undoc23e1$state_enroll23_sq<-undoc23e1$state_enroll23^2
hist(undoc23e1$state_enroll23_sq)
```



```
undoc23e1$state_enroll23_cat<-cut(undoc23e1$state_enroll23,
                           breaks=3,
                           labels=c("1","2","3"))

table(undoc23e1$state_enroll23_cat)
```

```
##
##    1    2    3
## 2167 3321  216
```

```
undoc23e1%>%group_by(state_enroll23_cat)%>%
  summarise(min=min(state_enroll23),
            max=max(state_enroll23))
```

```
## # A tibble: 3 x 3
##   state_enroll23_cat   min   max
##   <fct>              <dbl> <dbl>
## 1 1                  0.197 0.254
## 2 2                  0.259 0.301
## 3 3                  0.317 0.376
```

Household income categories

```
# Use quantile() with cut() to ensure balanced group sizes
undoc23e1$hhincome_cat <- cut(
  undoc23e1$hhincome,
  breaks = quantile(undoc23e1$hhincome, probs = c(0, 1/3, 2/3, 1), na.rm = TRUE),
  labels = c("1", "2", "3"),
  include.lowest = TRUE
)
table(undoc23e1$hhincome_cat)
```

```
##
##    1    2    3
## 1906 1898 1900
```

```
undoc23e1 %>%
  group_by(hhincome_cat) %>%
  summarise(
    min = min(hhincome, na.rm = TRUE),
    max = max(hhincome, na.rm = TRUE)
  )
```

```
## # A tibble: 3 x 3
##   hhincome_cat    min     max
##   <fct>         <dbl>   <dbl>
## 1 1              -300   67500
## 2 2             67600  139000
## 3 3            139100 9999999
```

```
summary(undoc23e1$age)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   18.00   20.00   21.00   21.16   23.00   24.00
```

Logistic Regression Set Up

setting up the svy object

```
undoc23e1$ids<-undoc23e1$serial+undoc23e1$pernum
```

final model we didn't use race

```
undoc23e1$race3 <- factor(undoc23e1$race,
                          levels = 1:9,
                          labels = c("White", "Black", "AIAN", "Chinese", "Japanese",
                                     "API_other", "Other", "TwoRaces", "ThreePlus"))
```

```r
#  collapsing "Asian or Pacific Islander (other)" into a broader "other" group
undoc23e1$race4<-fct_recode(undoc23e1$race3,
                           "other"="API_other")
# Sets "Black" as the reference group for modeling in race5
undoc23e1$race5<-relevel(undoc23e1$race4, ref="Black")
# using the original 9-category race3
undoc23e1$race6<-relevel(undoc23e1$race3, ref="Black")
```

```r
# helps detect or model non-linear effects in logit models for continuous predictors
undoc23e1$box_enroll23<-undoc23e1$state_enroll23*log(undoc23e1$state_enroll23)
undoc23e1$box_employ23<-undoc23e1$state_employ23*log(undoc23e1$state_employ23)
```

```r
#Create employ Variable
undoc23e1$employ <- ifelse(undoc23e1$empstat == 1, 1, 0)
```

```r
# Define state FIPS codes for in-state tuition access
# WA, OR, CA, NV, UT, AZ, CO, NM, NE, KS, OK, TX, HI, MN, IL, KY, VA, FL, NY, VT, MA, CT, RI, NJ, MD, D
isrt_states <- c(53, 41, 6, 32, 49, 4, 8, 35, 31, 20, 40, 48, 15,
                        27, 17, 21, 51, 12, 36, 50, 25, 9, 44, 34, 24, 11)

# Create binary indicator variable
undoc23e1$isrt <- ifelse(
  undoc23e1$statefip %in% isrt_states, 1, 0
)
```

```r
# Define state FIPS codes for states that allow driver's licenses for undocumented immigrants
drive_lic_states <- c(6, 8, 9, 10, 11, 15, 17, 24, 25, 27, 32, 34, 35, 36, 41, 44, 49, 50, 51, 53)

# Create binary indicator variable
undoc23e1$driveLic <- ifelse(undoc23e1$statefip %in% drive_lic_states, 1, 0)
```

combined Central America, Caribbean, and South America as 1 and the rest as 0.

```r
# Used `dpl` (summary birthplace variable) instead of `dpld` because:
# 1. `dpl` has full and consistent coverage across the ACS 2023 sample.
# 2. `dpld` is more granular but includes many missing or unavailable categories in the current dataset
# 3. `dpl` captures all major Latinx regional categories needed (e.g., Central America, Caribbean, Sout
# 4. Using `dpl` maintains consistency with prior studies' region-level groupings.
undoc23e1$birthplace1 <- case_when(
  undoc23e1$bpl %in% c(210, 250, 260, 299, 300) ~ "Cam Sam",  # Central America, Caribbean, South Ameri
  TRUE ~ "Other"
)
```

```r
# subset for lat but excluding Mexico , use this for further regression
lat_sub1 <- subset(undoc23e1, birthplace1=="Cam Sam")
```

```r
undoc23e1%>%group_by(state)%>%
  summarise(n=sum(perwt))%>%
  print(n=50)
```

```
## # A tibble: 51 x 2
##    state                     n
##    <chr>                 <dbl>
##  1 Alabama                4650
##  2 Alaska                 1200
##  3 Arizona               16898
##  4 Arkansas               3939
##  5 California           109066
##  6 Colorado              10423
##  7 Connecticut            9457
##  8 Delaware               1624
##  9 District of Columbia   1432
## 10 Florida               87017
## 11 Georgia               25951
## 12 Hawaii                 2098
## 13 Idaho                  2399
## 14 Illinois              26131
## 15 Indiana                7381
## 16 Iowa                   2764
## 17 Kansas                 5536
## 18 Kentucky               3695
## 19 Louisiana              5608
## 20 Maine                  1493
## 21 Maryland              14813
## 22 Massachusetts         13316
## 23 Michigan               9354
## 24 Minnesota              7649
## 25 Mississippi            3288
## 26 Missouri               5031
## 27 Montana                  47
## 28 Nebraska               4329
## 29 Nevada                 9380
## 30 New Hampshire          1049
## 31 New Jersey            39576
## 32 New Mexico             5643
## 33 New York              43837
## 34 North Carolina        16117
## 35 North Dakota            588
## 36 Ohio                   8122
## 37 Oklahoma               6133
## 38 Oregon                 5486
## 39 Pennsylvania          17105
## 40 Rhode Island           1870
## 41 South Carolina         7576
## 42 South Dakota           1284
## 43 Tennessee              9905
## 44 Texas                129111
## 45 Utah                   8705
## 46 Vermont                 117
## 47 Virginia              17491
## 48 Washington            15966
## 49 West Virginia           156
## 50 Wisconsin              4559
## # i 1 more row
```

```
save(undoc23e1, file="final_undoc23e.RData")
```

```
# svydesign
# creates a survey-weighted design object using perwt (person weight).
undoc_data1<-svydesign(id=~ids,
                         weights=~perwt,
                       data=undoc23e1)
```

```
lat_sub1<-svydesign(id=~ids,
                         weights=~perwt,
                       data=lat_sub1)
```

Running Regressions

```
# check following variables used in the prior regression model are not found in the current dataset:

model_vars <- c("attending_college", "isrt", "driveLic", "sex", "age",
                "birthplace1", "daca_imm", "employ", "hhincome_cat",
                "state_enroll23", "state_employ23", "box_enroll23", "box_employ23")

missing_vars <- model_vars[!(model_vars %in% names(undoc_data1$variables))]
missing_vars
```

```
## character(0)
```

Box Tidewell for state enrollment and employment in 2023

```
box_simple <- svyglm(attending_college ~ sex + age + hhincome_cat,
                     family = quasibinomial,
                     design = undoc_data1,
                     na.action = na.omit)
summary(box_simple)
```

```
##
## Call:
## svyglm(formula = attending_college ~ sex + age + hhincome_cat,
##     design = undoc_data1, family = quasibinomial, na.action = na.omit)
##
## Survey design:
## svydesign(id = ~ids, weights = ~perwt, data = undoc23e1)
##
```

```
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)    4.59525    0.43468  10.572  < 2e-16 ***
## sex            0.34102    0.07485   4.556 5.32e-06 ***
## age           -0.28674    0.01959 -14.635  < 2e-16 ***
## hhincome_cat2  0.11294    0.09233   1.223    0.221
## hhincome_cat3  0.67233    0.09000   7.471 9.20e-14 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## (Dispersion parameter for quasibinomial family taken to be 0.9976777)
##
## Number of Fisher Scoring iterations: 4
```

```
# survey-weighted logistic regression using the svyglm() function from the survey package
# It models the probability of attending college among undocumented youth using various predictors.
box1<-svyglm(attending_college~ isrt+ driveLic+ sex+age+birthplace1+daca_imm+
                      hhincome_cat+state_enroll23+state_employ23+
                      box_enroll23+box_employ23,
                    family=quasibinomial,
                    design=undoc_data1,
                    na.action = na.omit)
summary(box1)
```

```
##
## Call:
## svyglm(formula = attending_college ~ isrt + driveLic + sex +
##     age + birthplace1 + daca_imm + hhincome_cat + state_enroll23 +
##     state_employ23 + box_enroll23 + box_employ23, design = undoc_data1,
##     family = quasibinomial, na.action = na.omit)
##
## Survey design:
## svydesign(id = ~ids, weights = ~perwt, data = undoc23e1)
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      -14.28906   32.09629  -0.445    0.656
## isrt              -0.02013    0.11274  -0.179    0.858
## driveLic          -0.07724    0.10762  -0.718    0.473
## sex                0.34404    0.07584   4.537 5.84e-06 ***
## age               -0.28502    0.02028 -14.055  < 2e-16 ***
## birthplace1Other   0.69625    0.08752   7.955 2.14e-15 ***
## daca_imm          -0.02558    0.09503  -0.269    0.788
## hhincome_cat2      0.04624    0.09421   0.491    0.624
## hhincome_cat3      0.51442    0.09249   5.562 2.79e-08 ***
## state_enroll23     6.10088    5.42759   1.124    0.261
## state_employ23    12.86250   26.13012   0.492    0.623
## box_enroll23      -4.19303   19.60941  -0.214    0.831
## box_employ23     -25.23148   51.52884  -0.490    0.624
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## (Dispersion parameter for quasibinomial family taken to be 0.9994175)
##
```

```
## Number of Fisher Scoring iterations: 4
```

Model 2: logit 1b with employment and enrollment rates for 2023 (enrollment is categorical)

```
logit1b<-(svyglm(attending_college~isrt+driveLic+
                      sex+age+birthplace1+daca_imm+
                      employ+hhincome_cat+
                      state_enroll23_cat+state_employ23,
                    family=quasibinomial,
                    design=undoc_data1,
                    na.action = na.omit))
# tidy summary
tidy(logit1b)%>%
  mutate(estimate=round(estimate,2),
         std.error=round(std.error,2),
         statistic=round(statistic,2),
         or=round(exp(estimate),2),
         p.value=round(p.value,4))
```

```
## # A tibble: 13 x 6
##    term                estimate std.error statistic p.value     or
##    <chr>                  <dbl>     <dbl>     <dbl>   <dbl>  <dbl>
##  1 (Intercept)             4.92      1.1       4.45  0      137
##  2 isrt                   -0.03      0.11     -0.28  0.777    0.97
##  3 driveLic                0.07      0.1       0.69  0.490    1.07
##  4 sex                     0.31      0.08      4     0.0001   1.36
##  5 age                    -0.25      0.02    -12.0   0        0.78
##  6 birthplace1Other        0.68      0.09      7.53  0        1.97
##  7 daca_imm               -0.02      0.1      -0.2   0.838    0.98
##  8 employ                 -0.66      0.08     -8.24  0        0.52
##  9 hhincome_cat2           0.15      0.1       1.52  0.128    1.16
## 10 hhincome_cat3           0.6       0.1       6.35  0        1.82
## 11 state_enroll23_cat2     0.16      0.1       1.66  0.0973   1.17
## 12 state_enroll23_cat3     0.85      0.22      3.93  0.0001   2.34
## 13 state_employ23         -1.98      1.67     -1.19  0.234    0.14
```

Wald test for logit1b (using 2023 employment and enrollment data)

```
wald_test_full_logit1_23 <- regTermTest(logit1b, ~isrt+driveLic+
                      sex+age+birthplace2+daca_imm+
                      employ+hhincome_cat+
                      state_enroll23_cat+state_employ23)
print(wald_test_full_logit1_23)
```

```
## Wald test for isrt driveLic sex age birthplace2 daca_imm employ hhincome_cat state_enroll23_cat state
##  in svyglm(formula = attending_college ~ isrt + driveLic + sex +
##     age + birthplace1 + daca_imm + employ + hhincome_cat + state_enroll23_cat +
##     state_employ23, design = undoc_data1, family = quasibinomial,
##     na.action = na.omit)
## F =  32.048  on  11  and  5682  df: p= < 2.22e-16
```

```
# wald test for isrt for logit1

wald_test_isrtlog1<-regTermTest(logit1b, ~isrt)
print(wald_test_isrtlog1) ## significant nice
```

```
## Wald test for isrt
##  in svyglm(formula = attending_college ~ isrt + driveLic + sex +
##     age + birthplace1 + daca_imm + employ + hhincome_cat + state_enroll23_cat +
##     state_employ23, design = undoc_data1, family = quasibinomial,
##     na.action = na.omit)
## F =  0.08033123  on  1  and  5682  df: p= 0.77686
```

Predicted probs for logit1b

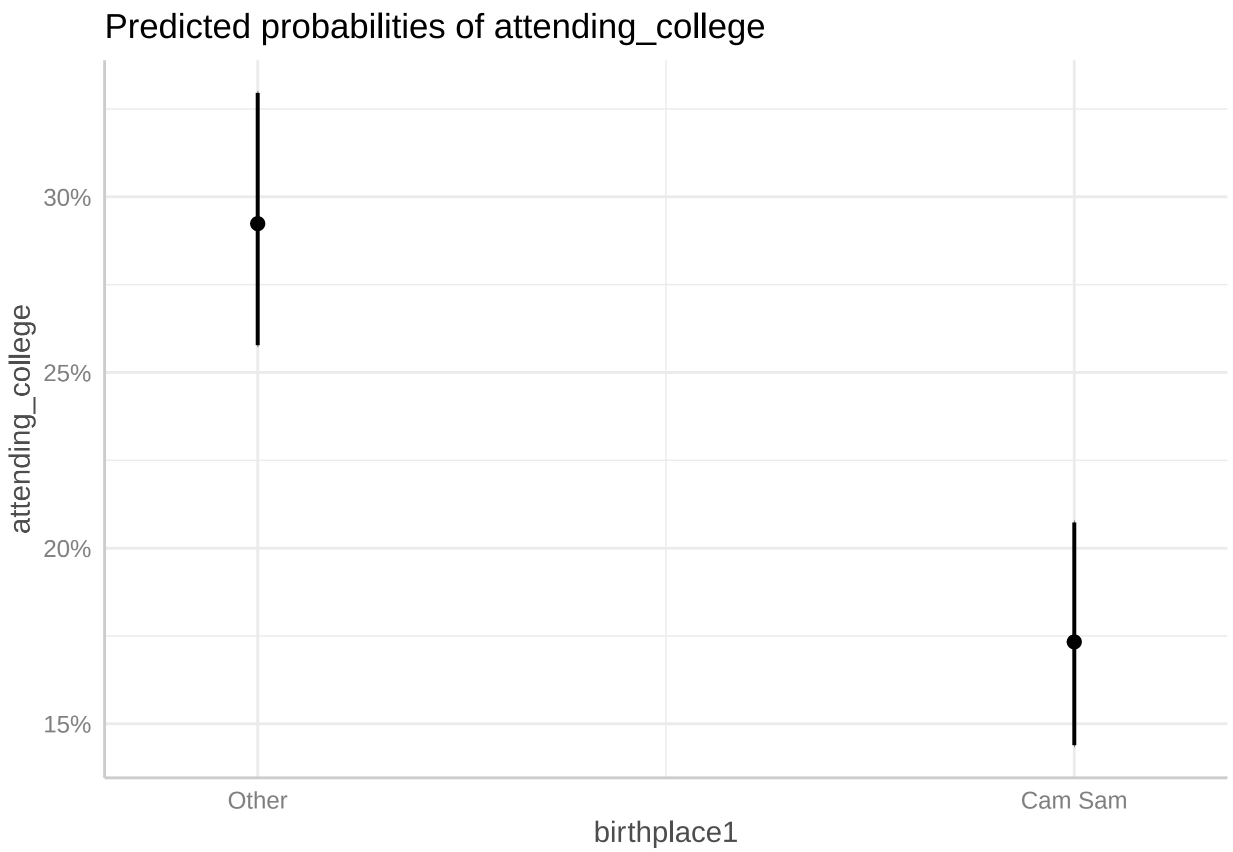
```
#visualize the predicted probabilities of college attendance based on birthplace1 from logit1b model.
log1b_pred_bpl<-ggpredict(logit1b, terms="birthplace1")
print(as.data.frame(log1b_pred_bpl))
```

```
##         x predicted  std.error  conf.low conf.high group
## 1   Other 0.2923613 0.08868401 0.2577305 0.3295792     1
## 2 Cam Sam 0.1733182 0.11272659 0.1438989 0.2072958     1
```

```
log1b_pred_bpl
```

```
## # Predicted probabilities of attending_college
##
## birthplace1 | Predicted |     95% CI
## ------------------------------------
## Other       |      0.29 | 0.26, 0.33
## Cam Sam     |      0.17 | 0.14, 0.21
##
## Adjusted for:
## *               isrt =  0.81
## *           driveLic =  0.47
## *                sex =  1.46
## *                age = 21.32
## *           daca_imm =  0.22
## *             employ =  0.65
## *       hhincome_cat =     1
## * state_enroll23_cat =     1
## *     state_employ23 =  0.61
```

```
plot(log1b_pred_bpl)
```



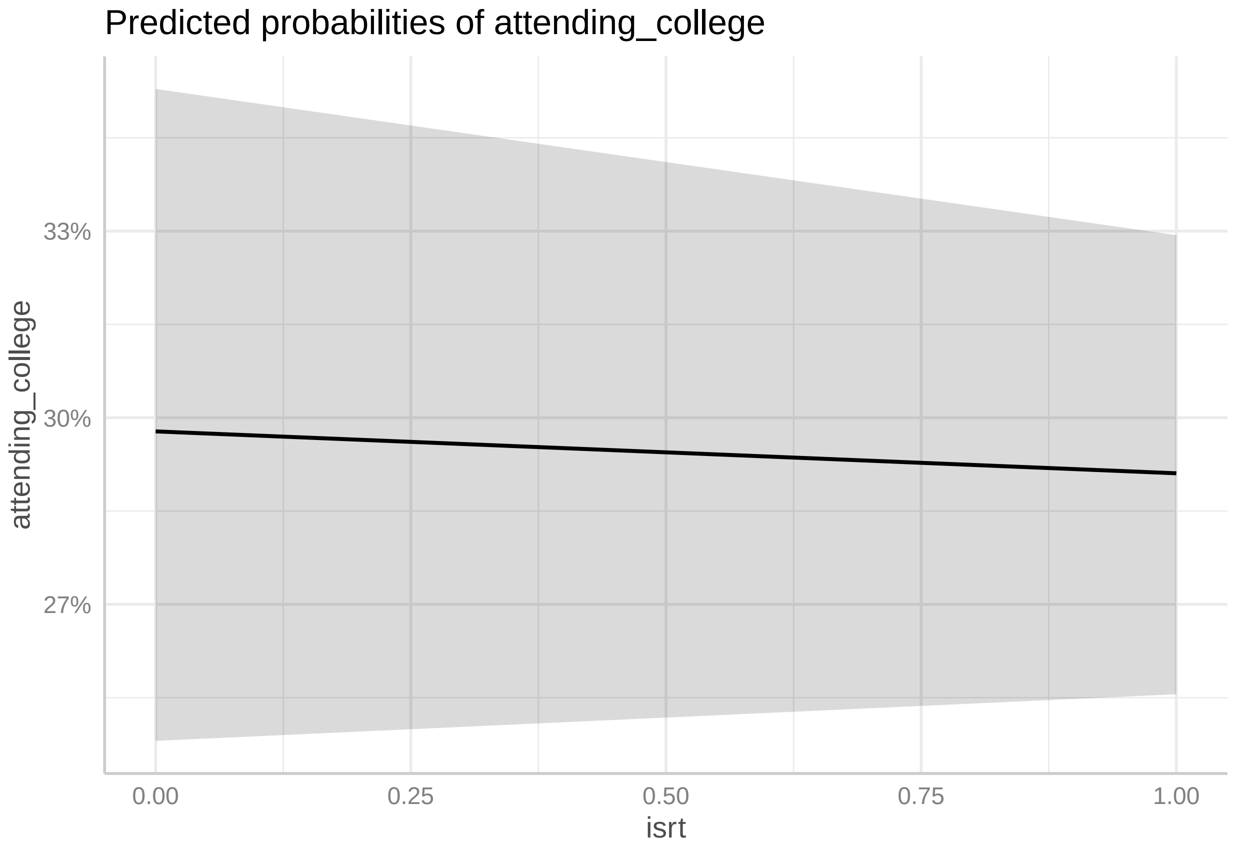
```r
log1b_pred_isrt<-ggpredict(logit1b, terms="isrt")
print(as.data.frame(log1b_pred_isrt))
```

```
##   x predicted  std.error  conf.low conf.high group
## 1 0 0.2978037 0.12816254 0.2480525 0.3528509     1
## 2 1 0.2910707 0.09133098 0.2555489 0.3293456     1
```

```r
log1b_pred_isrt
```

```
## # Predicted probabilities of attending_college
##
## isrt | Predicted |     95% CI
## -----------------------------
##    0 |      0.30 | 0.25, 0.35
##    1 |      0.29 | 0.26, 0.33
##
## Adjusted for:
## *           driveLic =  0.47
## *                sex =  1.46
## *                age = 21.32
## *        birthplace1 = Other
## *           daca_imm =  0.22
## *             employ =  0.65
## *        hhincome_cat =     1
## * state_enroll23_cat =     1
## *     state_employ23 =  0.61
```

```
plot(log1b_pred_isrt)
```



logit2: driver license and Isrt interaction

```
logit2<-(svyglm(attending_college~isrt+driveLic+
                      sex+age+birthplace1+daca_imm+
                      employ+hhincome_cat+
                      state_enroll23_cat+state_employ23+isrt:driveLic,
                    family=quasibinomial,
                    design=undoc_data1,
                    na.action = na.omit))
# tidy summary
tidy(logit2)%>%
  mutate(estimate=round(estimate,2),
         std.error=round(std.error,2),
         statistic=round(statistic,2),
         or=round(exp(estimate),2),
         p.value=round(p.value,4))
```

```
## # A tibble: 14 x 6
##    term                  estimate std.error statistic p.value     or
##    <chr>                  <dbl>     <dbl>     <dbl>   <dbl>  <dbl>
##  1 (Intercept)             4.93      1.1       4.46  0      138.
##  2 isrt                   -0.04      0.12     -0.33  0.738    0.96
##  3 driveLic               -0.33      0.77     -0.43  0.665    0.72
##  4 sex                     0.31      0.08      4     0.0001   1.36
##  5 age                    -0.25      0.02    -12.0   0        0.78
##  6 birthplace1Other        0.68      0.09      7.52  0        1.97
##  7 daca_imm               -0.02      0.1      -0.21  0.837    0.98
##  8 employ                 -0.66      0.08     -8.24  0        0.52
##  9 hhincome_cat2           0.15      0.1       1.52  0.129    1.16
## 10 hhincome_cat3           0.6       0.1       6.33  0        1.82
## 11 state_enroll23_cat2     0.16      0.1       1.67  0.0955   1.17
## 12 state_enroll23_cat3     0.85      0.22      3.92  0.0001   2.34
## 13 state_employ23         -1.99      1.66     -1.2   0.232    0.14
## 14 isrt:driveLic           0.41      0.78      0.53  0.598    1.51
```

logit 3: An employed and ISRT interaction

```
logit3<-(svyglm(attending_college~isrt+driveLic+
                      sex+age+birthplace1+daca_imm+
                      employ+hhincome_cat+
                      state_enroll23_cat+state_employ23+isrt:employ,
                    family=quasibinomial,
                    design=undoc_data1,
                    na.action = na.omit))
# tidy summary
tidy(logit3)%>%
  mutate(estimate=round(estimate,2),
         std.error=round(std.error,2),
         statistic=round(statistic,2),
         or=round(exp(estimate),2),
         p.value=round(p.value,4)) ## almost significant--interesting
```

```
## # A tibble: 14 x 6
##    term                estimate std.error statistic p.value     or
##    <chr>                  <dbl>     <dbl>     <dbl>   <dbl>  <dbl>
##  1 (Intercept)             4.8       1.11      4.31  0      122.
##  2 isrt                    0.1       0.16      0.66  0.508    1.11
##  3 driveLic                0.07      0.1       0.7   0.485    1.07
##  4 sex                     0.31      0.08      3.99  0.0001   1.36
##  5 age                    -0.25      0.02    -12.0   0        0.78
##  6 birthplace1Other        0.68      0.09      7.55  0        1.97
##  7 daca_imm               -0.02      0.1      -0.22  0.829    0.98
##  8 employ                 -0.48      0.18     -2.71  0.0067   0.62
##  9 hhincome_cat2           0.15      0.1       1.51  0.130    1.16
## 10 hhincome_cat3           0.61      0.1       6.39  0        1.84
## 11 state_enroll23_cat2     0.16      0.1       1.65  0.0991   1.17
## 12 state_enroll23_cat3     0.85      0.22      3.93  0.0001   2.34
## 13 state_employ23         -1.97      1.67     -1.18  0.238    0.14
## 14 isrt:employ            -0.23      0.2      -1.16  0.246    0.79
```

predicted probabilities for logit3–employed X isrt interaction

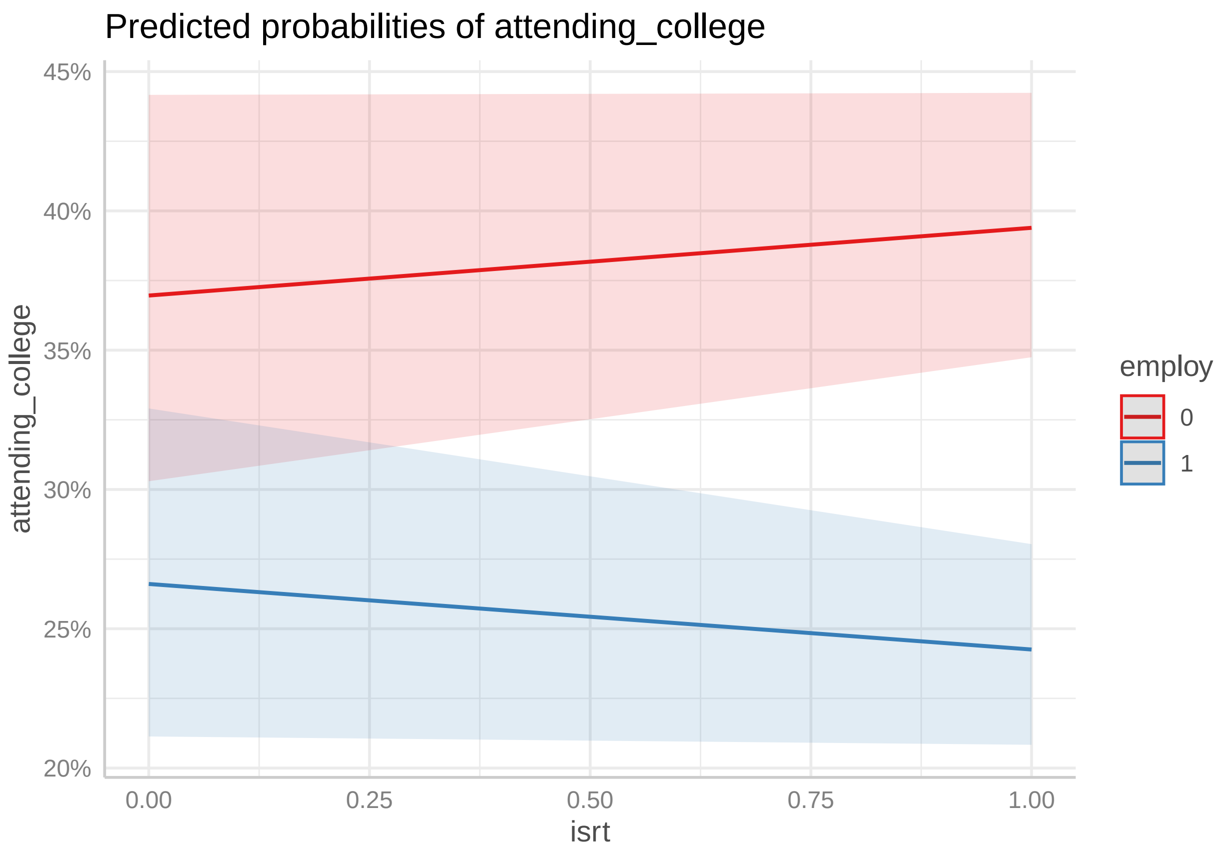
```
logit3_pred<-ggpredict(logit3, terms=c("isrt","employ"))
print(as.data.frame(logit3_pred))
```

```
##   x predicted std.error  conf.low conf.high group
## 1 0 0.3696105 0.1527250 0.3029505 0.4416450     0
## 2 0 0.2660744 0.1542061 0.2113287 0.3290844     1
## 3 1 0.3939172 0.1016873 0.3474602 0.4423747     0
## 4 1 0.2425461 0.1000807 0.2083389 0.2803806     1
```

```
logit3_pred
```

```
## # Predicted probabilities of attending_college
##
## employ: 0
##
## isrt | Predicted |     95% CI
## -----------------------------
##    0 |      0.37 | 0.30, 0.44
##    1 |      0.39 | 0.35, 0.44
##
## employ: 1
##
## isrt | Predicted |     95% CI
## -----------------------------
##    0 |      0.27 | 0.21, 0.33
##    1 |      0.24 | 0.21, 0.28
##
## Adjusted for:
## *           driveLic =  0.47
## *                sex =  1.46
## *                age = 21.32
## *        birthplace1 = Other
## *           daca_imm =  0.22
## *       hhincome_cat =     1
## * state_enroll23_cat =     1
## *     state_employ23 =  0.61
```

```
plot(logit3_pred)
```



logit 4: birthplace and isrt interaction

```
logit4<-(svyglm(attending_college~isrt+driveLic+
                      sex+age+birthplace1+daca_imm+
                      employ+hhincome_cat+
                      state_enroll23_cat+state_employ23+isrt:birthplace1,
                    family=quasibinomial,
                    design=undoc_data1,
                    na.action = na.omit))
# tidy summary
tidy(logit4)%>%
  mutate(estimate=round(estimate,2),
         std.error=round(std.error,2),
         statistic=round(statistic,2),
         or=round(exp(estimate),2),p.value=round(p.value,4))
```

```
## # A tibble: 14 x 6
##    term                     estimate std.error statistic p.value     or
##    <chr>                       <dbl>     <dbl>     <dbl>   <dbl>  <dbl>
##  1 (Intercept)                  4.79      1.11      4.31  0       120.
##  2 isrt                         0.16      0.22      0.74  0.457    1.17
##  3 driveLic                     0.08      0.1       0.8   0.422    1.08
##  4 sex                          0.31      0.08      3.96  0.0001   1.36
##   5 age                      -0.25      0.02    -12.1   0         0.78
##   6 birthplace1Other          0.91      0.22      4.08  0         2.48
##   7 daca_imm                 -0.02      0.1      -0.21  0.834     0.98
##   8 employ                   -0.66      0.08     -8.23  0         0.52
##   9 hhincome_cat2             0.15      0.1       1.51  0.131     1.16
##  10 hhincome_cat3             0.6       0.09      6.32  0         1.82
##  11 state_enroll23_cat2       0.15      0.1       1.52  0.128     1.16
##  12 state_enroll23_cat3       0.83      0.22      3.82  0.0001    2.29
##  13 state_employ23           -2         1.66     -1.2   0.23      0.14
##  14 isrt:birthplace1Other    -0.29      0.24     -1.21  0.228     0.75
```

predicted probabilities for isrt X birthplace interaction

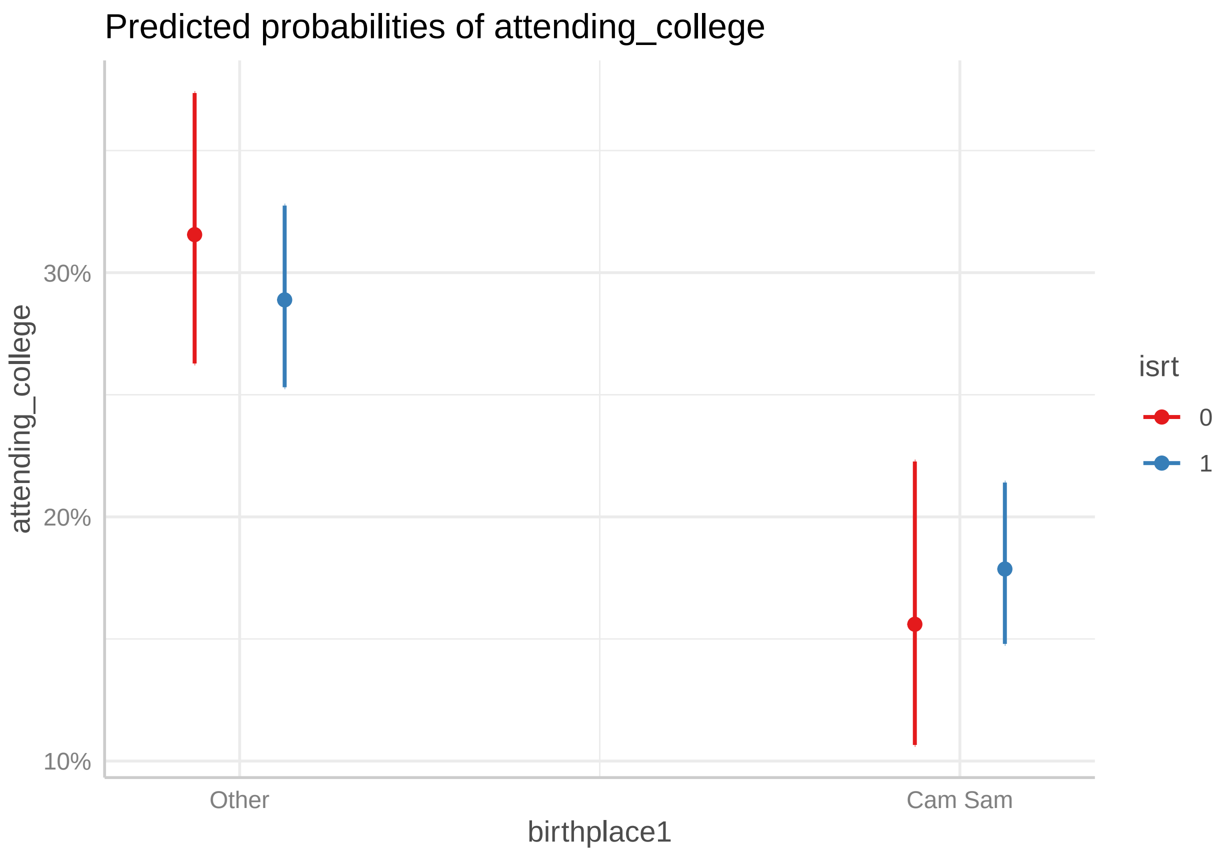
```
logit4_pred<-ggpredict(logit4, terms=c("birthplace1","isrt"))
logit4_pred
```

```
## # Predicted probabilities of attending_college
##
## isrt: 0
##
## birthplace1 | Predicted |     95% CI
## ------------------------------------
## Other       |      0.32 | 0.26, 0.37
## Cam Sam     |      0.16 | 0.11, 0.22
##
## isrt: 1
##
## birthplace1 | Predicted |     95% CI
## ------------------------------------
## Other       |      0.29 | 0.25, 0.33
## Cam Sam     |      0.18 | 0.15, 0.21
##
## Adjusted for:
## *           driveLic =  0.47
## *                sex =  1.46
## *                age = 21.32
## *           daca_imm =  0.22
## *             employ =  0.65
## *       hhincome_cat =     1
## * state_enroll23_cat =     1
## *     state_employ23 =  0.61
```

```
print(as.data.frame(logit4_pred))
```

```
##         x predicted  std.error  conf.low conf.high group
## 1   Other 0.3155668 0.13123255 0.2627960 0.3735646     0
## 2   Other 0.2888457 0.09249868 0.2530654 0.3274670     1
## 3 Cam Sam 0.1560160 0.22342021 0.1065797 0.2226679     0
## 4 Cam Sam 0.1786263 0.11478975 0.1479572 0.2140556     1
```

```r
plot(logit4_pred)
```



Logit 5: Race instead of birthplace

```r
logit5<-(svyglm(attending_college~isrt+driveLic+
                      sex+age+race3+daca_imm+
                      employ+hhincome_cat+
                      state_enroll23_cat+state_employ23,
                    family=quasibinomial,
                    design=undoc_data1,
                    na.action = na.omit))
# tidy summary
tidy(logit5)%>%
  mutate(estimate=round(estimate,2),
         std.error=round(std.error,2),
         statistic=round(statistic,2),
         or=round(exp(estimate),2),p.value=round(p.value,4)) # isrt isn't significant
```

```
## # A tibble: 20 x 6
##    term                 estimate std.error statistic p.value     or
##    <chr>                   <dbl>     <dbl>     <dbl>   <dbl>  <dbl>
##  1 (Intercept)              4.74      1.12      4.23  0       114.
##  2 isrt                     0.08      0.12      0.67  0.503    1.08
##  3 driveLic             0         0.11     -0.01  0.988    1
##  4 sex                  0.32      0.08      4.06  0.0001   1.38
##  5 age                 -0.25      0.02    -11.7   0        0.78
##  6 race3Black           0.43      0.17      2.49  0.013    1.54
##  7 race3AIAN           -0.65      0.26     -2.5   0.0125   0.52
##  8 race3Chinese         2.29      0.27      8.56  0        9.87
##  9 race3Japanese        2.07      0.75      2.74  0.0062   7.92
## 10 race3API_other       1         0.14      6.96  0        2.72
## 11 race3Other          -0.76      0.12     -6.44  0        0.47
## 12 race3TwoRaces       -0.64      0.12     -5.35  0        0.53
## 13 race3ThreePlus      -0.09      0.44     -0.21  0.838    0.91
## 14 daca_imm             0.46      0.1       4.68  0        1.58
## 15 employ              -0.55      0.08     -6.53  0        0.58
## 16 hhincome_cat2        0.18      0.1       1.84  0.0655   1.2
## 17 hhincome_cat3        0.54      0.1       5.34  0        1.72
## 18 state_enroll23_cat2  0.07      0.1       0.75  0.455    1.07
## 19 state_enroll23_cat3  0.55      0.23      2.41  0.016    1.73
## 20 state_employ23      -0.88      1.68     -0.52  0.600    0.41
```

Logit6: ISRT X state employment rate interaction

```
logit6<-(svyglm(attending_college~isrt+driveLic+
                      sex+age+race3+daca_imm+
                      employ+hhincome_cat+
                      state_enroll23_cat+state_employ23+state_enroll23_cat:isrt,
                    family=quasibinomial,
                    design=undoc_data1,
                    na.action = na.omit))
# tidy summary
tidy(logit6)%>%
  mutate(estimate=round(estimate,2),
         std.error=round(std.error,2),
         statistic=round(statistic,2),
         or=round(exp(estimate),2),p.value=round(p.value,4)) # isrt isn't significant
```

```
## # A tibble: 22 x 6
##    term           estimate std.error statistic p.value    or
##    <chr>             <dbl>     <dbl>     <dbl>   <dbl> <dbl>
##  1 (Intercept)        4.01      1.13      3.54  0.0004 55.2
##  2 isrt              -0.17      0.14     -1.24  0.214   0.84
##  3 driveLic          -0.1       0.11     -0.94  0.345   0.9
##  4 sex                0.33      0.08      4.08  0       1.39
##  5 age               -0.25      0.02    -11.8   0       0.78
##  6 race3Black         0.4       0.17      2.33  0.0199  1.49
##  7 race3AIAN         -0.68      0.26     -2.63  0.0086  0.51
##  8 race3Chinese       2.3       0.27      8.51  0       9.97
##  9 race3Japanese      2.03      0.76      2.68  0.0074  7.61
## 10 race3API_other     1         0.14      6.96  0       2.72
## # i 12 more rows
```

predicted probs for logit6 (isrt X state enroll interaction)

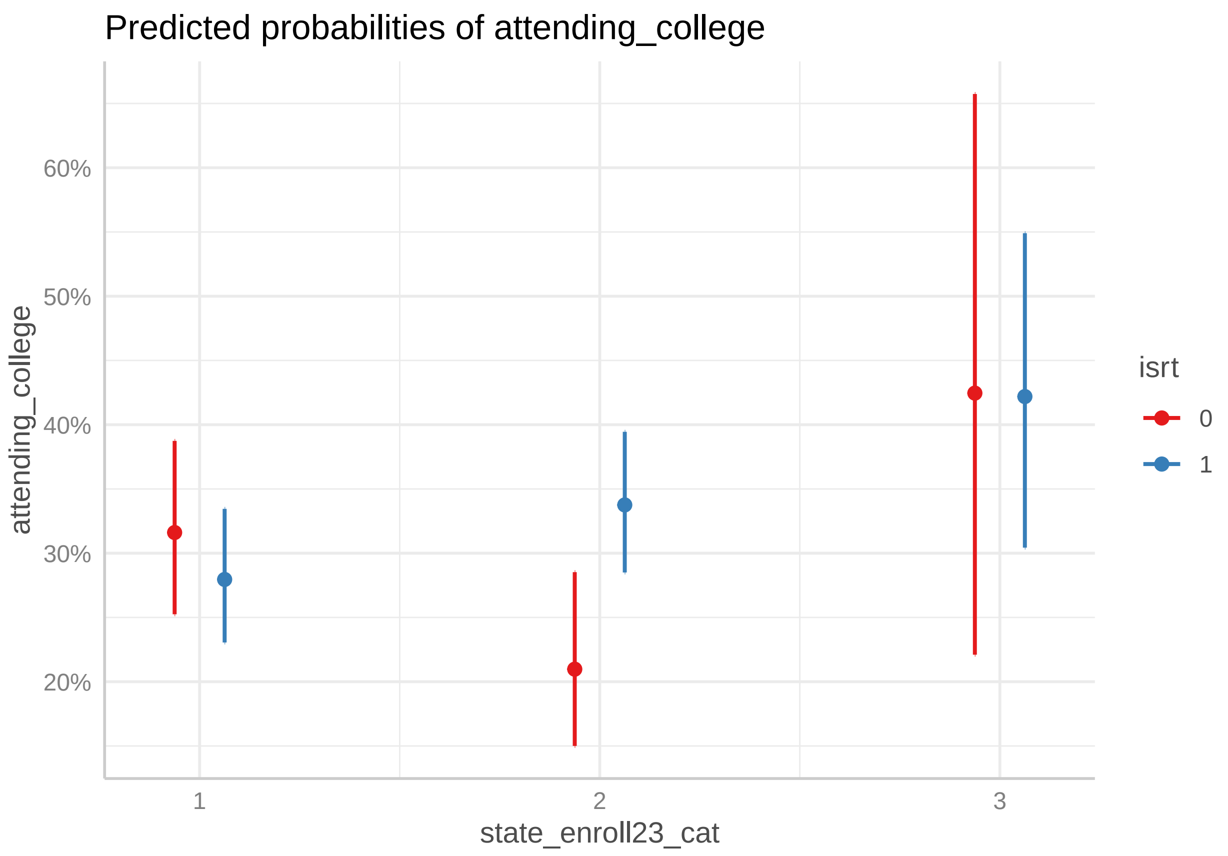
```
log6_pred<-ggpredict(logit6, terms=c("state_enroll23_cat", "isrt"))
log6_pred
```

```
## # Predicted probabilities of attending_college
##
## isrt: 0
##
## state_enroll23_cat | Predicted |     95% CI
## ------------------------------------------
## 1                  |      0.32 | 0.25, 0.39
## 2                  |      0.21 | 0.15, 0.29
## 3                  |      0.42 | 0.22, 0.66
##
## isrt: 1
##
## state_enroll23_cat | Predicted |     95% CI
## ------------------------------------------
## 1                  |      0.28 | 0.23, 0.33
## 2                  |      0.34 | 0.28, 0.39
## 3                  |      0.42 | 0.30, 0.55
##
## Adjusted for:
## *       driveLic =  0.47
## *            sex =  1.46
## *            age = 21.32
## *          race3 = White
## *       daca_imm =  0.22
## *         employ =  0.65
## *    hhincome_cat =     1
## * state_employ23 =  0.61
```

```
print(as.data.frame(log6_pred))
```

```
##   x predicted std.error  conf.low conf.high group
## 1 1 0.3160925 0.1599769 0.2524848 0.3874196     0
## 2 1 0.2795633 0.1319931 0.2305185 0.3345067     1
## 3 2 0.2097390 0.2081509 0.1500059 0.2852750     0
## 4 2 0.3375718 0.1253299 0.2849928 0.3944991     1
## 5 3 0.4245832 0.4875202 0.2210234 0.6574023     0
## 6 3 0.4219004 0.2610776 0.3043245 0.5490505     1
```

```r
plot(log6_pred)
```



Logistic Regressions with the Latinx sub-set sample

log_latx1 with region birthplace 3 and 2023 employment and enrollment data

```r
log_latx1<-(svyglm(attending_college~isrt+driveLic+
                      sex+age+daca_imm+
                      employ+hhincome_cat+
                      state_enroll23_cat+state_employ23_cat,
                    family=quasibinomial,
                    design=lat_sub1,
                    na.action = na.omit))

tidy(log_latx1)%>%
  mutate(estimate=round(estimate,2),
         std.error=round(std.error,2),
         statistic=round(statistic,2),
         or=round(exp(estimate),2),
         p.value=round(p.value,4))
```

```
## # A tibble: 13 x 6
##    term                estimate std.error statistic p.value    or
##    <chr>                  <dbl>     <dbl>     <dbl>   <dbl> <dbl>
##  1 (Intercept)             3.22      0.8       4.01  0.0001 25.0
##  2 isrt                    0.07      0.24      0.28  0.780   1.07
##  3 driveLic                0.16      0.26      0.62  0.537   1.17
##  4 sex                     0.39      0.13      2.91  0.0036  1.48
##  5 age                    -0.24      0.04     -6.75  0       0.79
##  6 daca_imm                0.72      0.21      3.4   0.0007  2.05
##  7 employ                 -0.18      0.14     -1.22  0.221   0.84
##  8 hhincome_cat2           0.42      0.16      2.65  0.0082  1.52
##  9 hhincome_cat3           0.33      0.17      1.87  0.0615  1.39
## 10 state_enroll23_cat2     0.13      0.21      0.63  0.532   1.14
## 11 state_enroll23_cat3     0.81      0.36      2.25  0.0243  2.25
## 12 state_employ23_cat2    -0.38      0.25     -1.53  0.125   0.68
## 13 state_employ23_cat3    -0.51      0.3      -1.69  0.0905  0.6
```

Wald Test for effect of ISRT with Latinx sub-sample

```
wald_full_latx1 <- regTermTest(log_latx1, ~isrt+driveLic+
                      sex+age+daca_imm+
                      employ+hhincome_cat+
                      state_enroll23+state_employ23)
print(wald_full_latx1)
```

```
## Wald test for isrt driveLic sex age daca_imm employ hhincome_cat state_enroll23 state_employ23
##  in svyglm(formula = attending_college ~ isrt + driveLic + sex +
##     age + daca_imm + employ + hhincome_cat + state_enroll23_cat +
##     state_employ23_cat, design = lat_sub1, family = quasibinomial,
##     na.action = na.omit)
## F =  9.363365  on  8  and  2004  df: p= 9.1854e-13
```

```
# wald test for isrt for logit1

wald_isrt_latx1<-regTermTest(log_latx1, ~isrt)
print(wald_isrt_latx1) ## significant nice
```

```
## Wald test for isrt
##  in svyglm(formula = attending_college ~ isrt + driveLic + sex +
##     age + daca_imm + employ + hhincome_cat + state_enroll23_cat +
##     state_employ23_cat, design = lat_sub1, family = quasibinomial,
##     na.action = na.omit)
## F =  0.07817957  on  1  and  2004  df: p= 0.77981
```

predicted probabilities for log_latx1

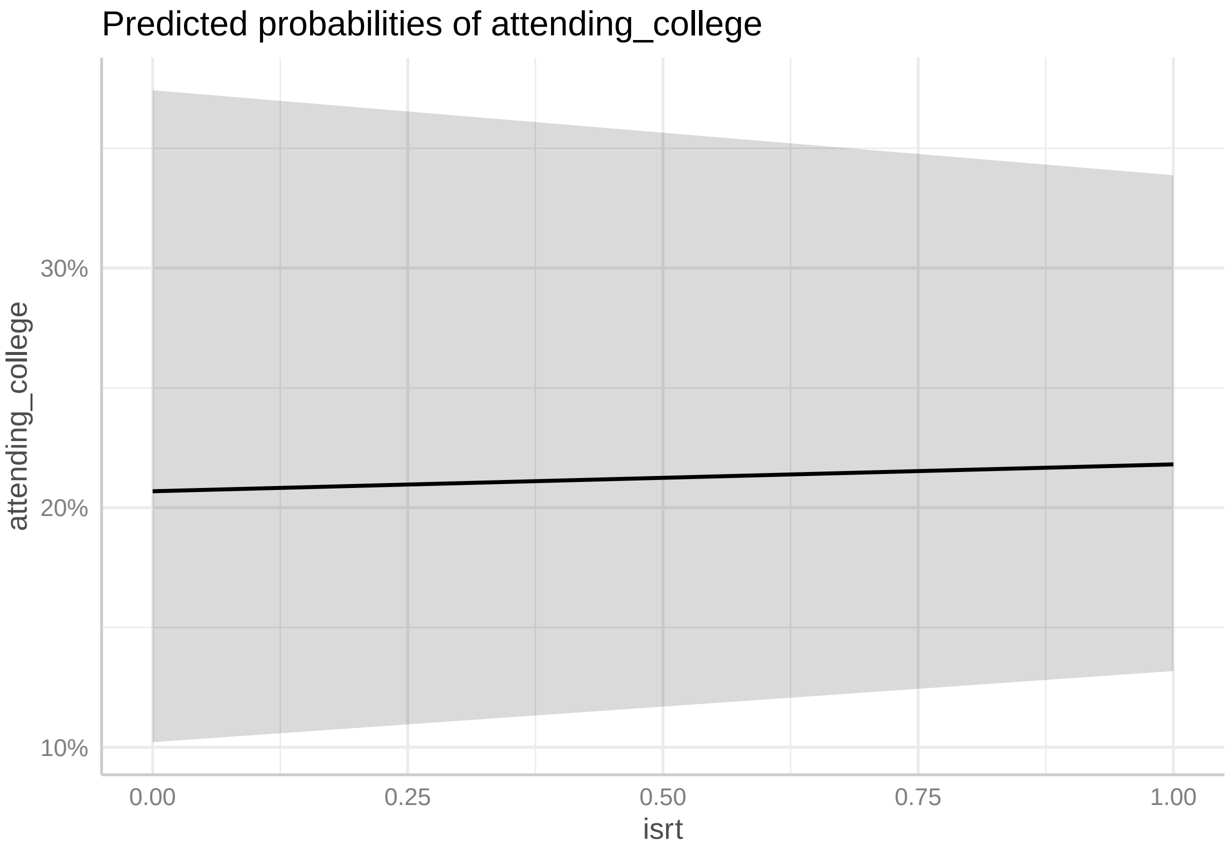
```
log_latx1_pred1<-ggpredict(log_latx1, terms=c("isrt"))
log_latx1_pred1
```

```
## # Predicted probabilities of attending_college
##
## isrt | Predicted |     95% CI
## ----------------------------
##    0 |      0.21 | 0.10, 0.37
##    1 |      0.22 | 0.13, 0.34
##
## Adjusted for:
## *           driveLic =  0.44
## *                sex =  1.46
## *                age = 21.40
## *           daca_imm =  0.09
## *             employ =  0.69
## *        hhincome_cat =     1
## * state_enroll23_cat =     1
## * state_employ23_cat =     1
```

```
print(as.data.frame(log_latx1_pred1))
```

```
##   x predicted std.error conf.low conf.high group
## 1 0 0.2068340 0.4232272 0.102098 0.3742318     1
## 2 1 0.2180502 0.3101526 0.131779 0.3387634     1
```

```
plot(log_latx1_pred1)
```



```r
# the effect of DACA status on college attendance differs by household income level.
logit_daca_income <- svyglm(
  attending_college ~ daca_imm * hhincome_cat +
    sex + age + isrt + driveLic + employ +
    state_enroll23_cat + state_employ23_cat,
  design = undoc_data1,
  family = quasibinomial,
  na.action = na.omit
)
summary(logit_daca_income)
```

```
##
## Call:
## svyglm(formula = attending_college ~ daca_imm * hhincome_cat +
##     sex + age + isrt + driveLic + employ + state_enroll23_cat +
##     state_employ23_cat, design = undoc_data1, family = quasibinomial,
##     na.action = na.omit)
##
## Survey design:
## svydesign(id = ~ids, weights = ~perwt, data = undoc23e1)
##
## Coefficients:
##                        Estimate Std. Error t value Pr(>|t|)
## (Intercept)             4.28846    0.46212   9.280  < 2e-16 ***
## daca_imm                0.21744    0.17057   1.275  0.20242
## hhincome_cat2           0.21732    0.10826   2.007  0.04475 *
## hhincome_cat3           0.72795    0.10609   6.862 7.52e-12 ***
## sex                     0.30377    0.07670   3.961 7.56e-05 ***
## age                    -0.25266    0.02035 -12.417  < 2e-16 ***
## isrt                   -0.10218    0.11459  -0.892  0.37261
## driveLic                0.20040    0.11155   1.797  0.07246 .
## employ                 -0.69975    0.08041  -8.703  < 2e-16 ***
## state_enroll23_cat2     0.03169    0.09601   0.330  0.74135
## state_enroll23_cat3     0.65942    0.21364   3.087  0.00203 **
## state_employ23_cat2    -0.03690    0.12551  -0.294  0.76877
## state_employ23_cat3    -0.22481    0.16036  -1.402  0.16098
## daca_imm:hhincome_cat2 -0.08859    0.22689  -0.390  0.69623
## daca_imm:hhincome_cat3 -0.09391    0.22198  -0.423  0.67227
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## (Dispersion parameter for quasibinomial family taken to be 0.9985401)
##
## Number of Fisher Scoring iterations: 4
```

```r
undoc_data_18_21 <- subset(undoc_data1, age >= 18 & age <= 21)
logit_age1821 <- svyglm(
  attending_college ~ isrt + driveLic + sex + age + race3 + daca_imm +
  employ + hhincome_cat + state_enroll23_cat + state_employ23_cat,
  design = undoc_data_18_21,
  family = quasibinomial
)
summary(logit_age1821)
```

```
##
## Call:
## svyglm(formula = attending_college ~ isrt + driveLic + sex +
##     age + race3 + daca_imm + employ + hhincome_cat + state_enroll23_cat +
##     state_employ23_cat, design = undoc_data_18_21, family = quasibinomial)
##
## Survey design:
## subset(undoc_data1, age >= 18 & age <= 21)
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)          1.27447    1.01569   1.255 0.209653
## isrt                 0.20394    0.15240   1.338 0.180912
## driveLic            -0.05155    0.15871  -0.325 0.745330
## sex                  0.31274    0.10363   3.018 0.002566 **
## age                 -0.09654    0.04999  -1.931 0.053520 .
## race3Black           0.26091    0.22068   1.182 0.237171
## race3AIAN           -0.45977    0.34034  -1.351 0.176819
## race3Chinese         2.26000    0.36170   6.248 4.72e-10 ***
## race3Japanese        1.65426    1.12840   1.466 0.142744
## race3API_other       0.96252    0.19586   4.914 9.37e-07 ***
## race3Other          -1.03418    0.15427  -6.704 2.40e-11 ***
## race3TwoRaces       -0.80511    0.15279  -5.269 1.46e-07 ***
## race3ThreePlus      -0.25768    0.61099  -0.422 0.673240
## daca_imm             0.45669    0.13823   3.304 0.000965 ***
## employ              -0.65668    0.10860  -6.047 1.66e-09 ***
## hhincome_cat2        0.31676    0.12830   2.469 0.013606 *
## hhincome_cat3        0.73845    0.12833   5.754 9.54e-09 ***
## state_enroll23_cat2  0.07001    0.13401   0.522 0.601413
## state_enroll23_cat3  0.86759    0.31668   2.740 0.006186 **
## state_employ23_cat2 -0.01620    0.17571  -0.092 0.926530
## state_employ23_cat3 -0.11849    0.21934  -0.540 0.589085
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## (Dispersion parameter for quasibinomial family taken to be 0.9946631)
##
## Number of Fisher Scoring iterations: 5
```

```
lat_sub1_age1821 <- subset(lat_sub1, age >= 18 & age <= 21)
log_latx1_age1821 <- svyglm(
  attending_college ~ isrt + driveLic + sex + age + daca_imm +
  employ + hhincome_cat + state_enroll23_cat + state_employ23_cat,
  design = lat_sub1_age1821,
  family = quasibinomial
)
summary(log_latx1_age1821)
```

```
##
## Call:
## svyglm(formula = attending_college ~ isrt + driveLic + sex +
##     age + daca_imm + employ + hhincome_cat + state_enroll23_cat +
##     state_employ23_cat, design = lat_sub1_age1821, family = quasibinomial)
##
```

```
## Survey design:
## subset(lat_sub1, age >= 18 & age <= 21)
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)           1.17796    1.63984   0.718 0.472714
## isrt                  0.33469    0.30259   1.106 0.268948
## driveLic              0.19052    0.36222   0.526 0.599014
## sex                   0.41215    0.16803   2.453 0.014343 *
## age                  -0.15007    0.07974  -1.882 0.060112 .
## daca_imm              0.62161    0.29402   2.114 0.034742 *
## employ               -0.26235    0.17608  -1.490 0.136542
## hhincome_cat2         0.73546    0.19252   3.820 0.000141 ***
## hhincome_cat3         0.43970    0.22563   1.949 0.051599 .
## state_enroll23_cat2  0.08729    0.30112   0.290 0.771954
## state_enroll23_cat3  1.11501    0.49840   2.237 0.025491 *
## state_employ23_cat2 -0.50110    0.35714  -1.403 0.160895
## state_employ23_cat3 -0.56558    0.42699  -1.325 0.185605
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## (Dispersion parameter for quasibinomial family taken to be 0.9934321)
##
## Number of Fisher Scoring iterations: 4
```