analysis_i_variance_decomp

Hunter York

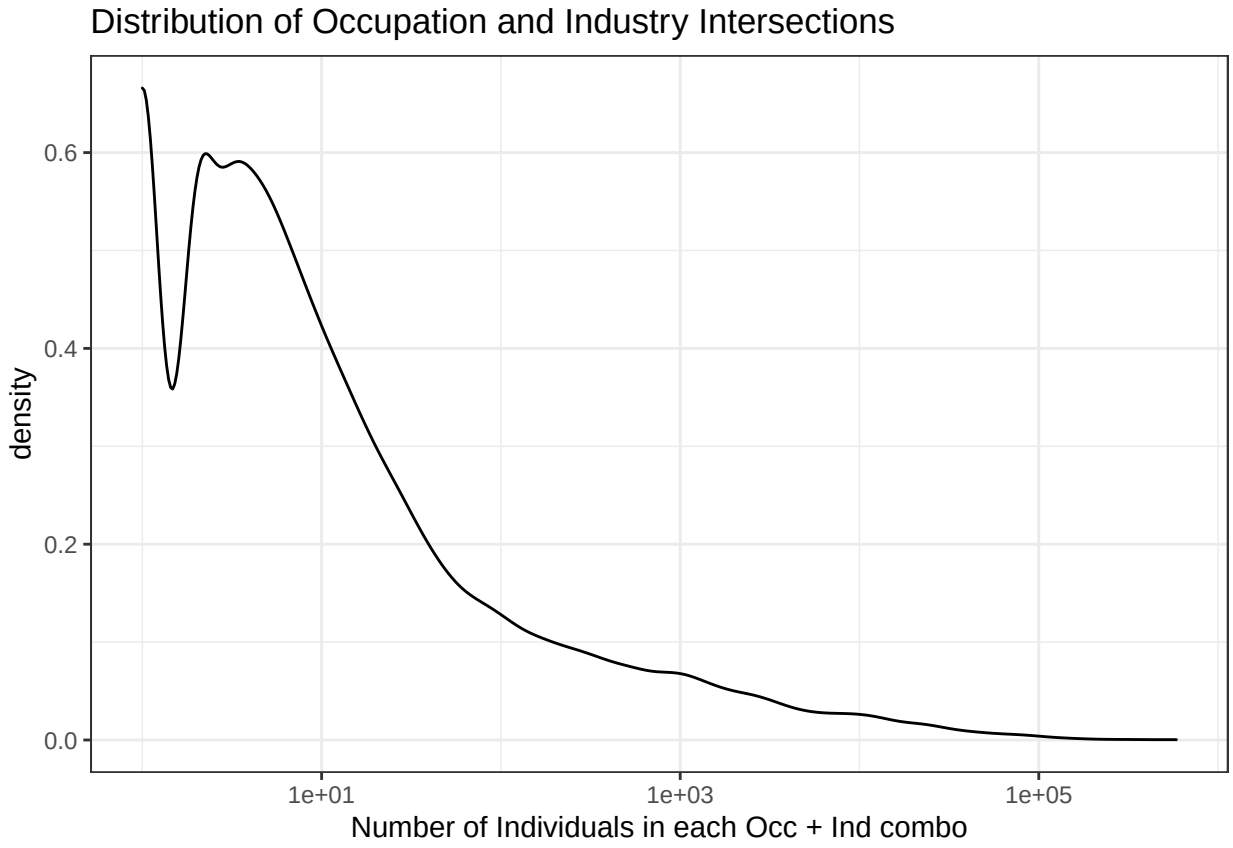
9/27/2020

Step 1 - Decompose variance by industry, by occupation, and then by their intersection

Step 1.2 - Quick descriptive table of Occupation + Industry intersection

There are 24942 intersections of occupation and industry. (230 occupations * 134 industry) - around 5000 missing combinations. About 5000 combos only have one respondent in the category. 4725 have 50 or more respondents, and 1397 have more than 1000. 83% of respondents are captured in all occ+ind combos with at least 5000 respondants, and 94% in all combos with at least 1000 respondents.

```
census_1940 %>%
  .[,.N, by = .(occ, ind)] %>%
  .[order(N)] %>%
  .[!is.na(N)] %>%
   ggplot(.) +
     geom_density(aes(x= N)) +
  scale_x_continuous(trans = "log10") +
  labs(x = "Number of Individuals in each Occ + Ind combo",
       title = "Distribution of Occupation and Industry Intersections")
```



Step 2 - Between occupation vs within occupation variance, and then for industry and their intersection

This following section uses the following formula to calculate the statistics used in this section.

$$SS_{Total} = \sum_{j=1}^{k} \sum_{i=1}^{n_j} (x_{ij} - \bar{x}^2)$$

$$SS_{Between} = \sum n_j (\bar{x_j} - \bar{x})^2$$

$$SS_{Within} = SS_{Total} - SS_{Between}$$

$$MS_{Between} = \frac{SS_{Between}}{k-1}$$

$$MS_{Within} = \frac{SS_{Within}}{n-k}$$

$$F = \frac{MS_{Between}}{MS_{Within}}$$

Replication of Xie & Killewald

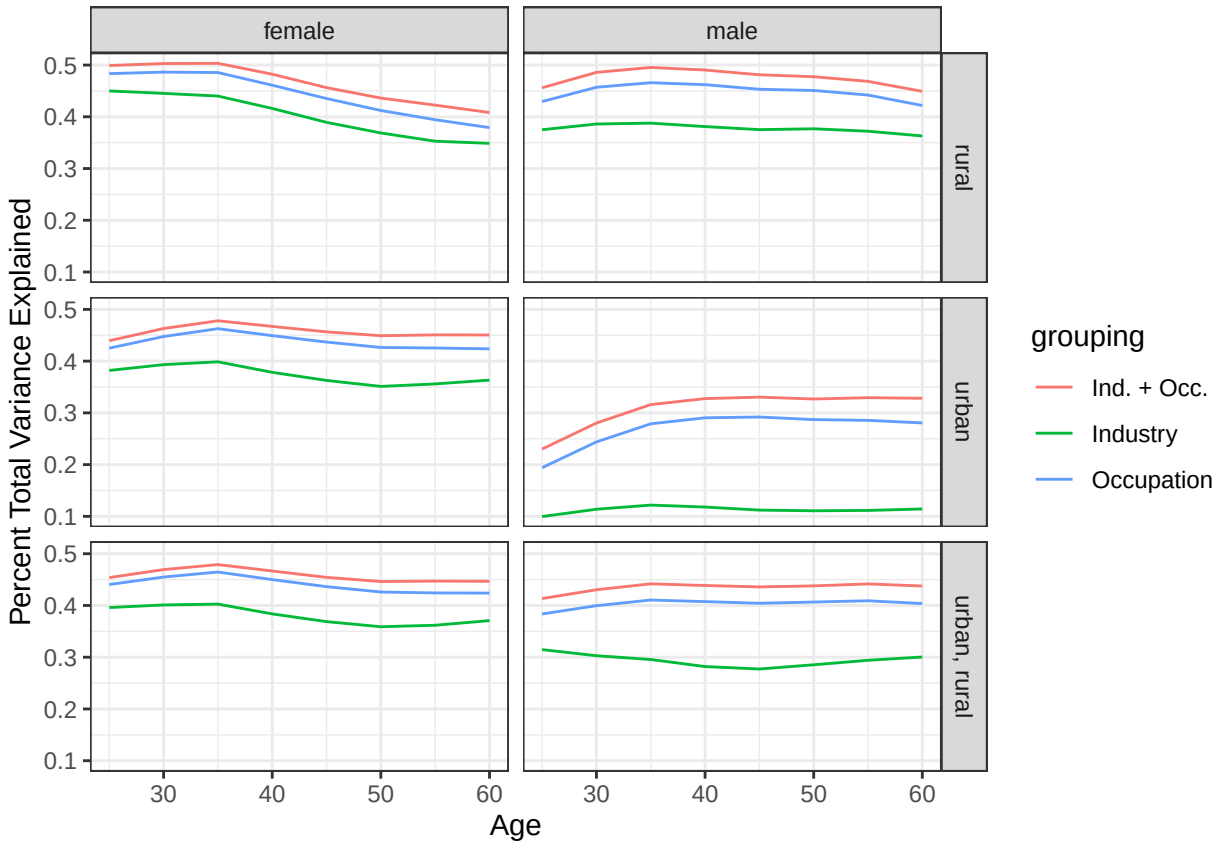
$$\eta^2 = \frac{SS_{Between}}{SS_{Total}}$$

NB: This is equivalent to the method used by Xie & Killewald wherein η^2 is what would be returned if a linear model were fit and ANOVA was run on it.

```
# create function to get var from a variable
get_vars <- function(c.data, c.by_vars, c.var_interest){
  out_dt_1 <- c.data[,.(w_i_ss = var(get(c.var_interest)) * .N,
                        N = .N,
                        k = max(.GRP)),
                     by = c.by_vars]
  out_dt_1[,tot_ss := sum((c.data[,(get(c.var_interest))])^2) -
             (((sum(c.data[,(get(c.var_interest))]))^2)/
                length(c.data[,(get(c.var_interest))]))]
  out_dt_1 <- out_dt_1[!is.na(tot_ss)& !is.na(w_i_ss),
                       .(avg_within_var = sum(w_i_ss),
                         avg_total_var = mean(tot_ss),
                         avg_between_var = mean(tot_ss) -sum(w_i_ss),
                         N = sum(N),
                         k = length(unique(N[!is.na(w_i_ss)])))]
  return(out_dt_1)
}

#  create another function to loop over data and
#  calculate occ, ind, and occ + ind var

dem_var_gettr <- function(c.dat2,
                          c.age_cat,
                          c.sex,
                          c.urban){
  occ_only <- get_vars(c.dat2[age_cat %in% c.age_cat &
                                sex %in% c.sex &
                                urban %in% c.urban],
                       c.by_vars = "occ",
                       c.var_interest = "log_incwage")
  ind_only <- get_vars(c.dat2[age_cat %in% c.age_cat &
                                sex %in% c.sex&
                                urban %in% c.urban],
                       c.by_vars = "ind",
                       c.var_interest = "log_incwage")
  occ_ind <- get_vars(c.dat2[age_cat %in% c.age_cat &
                               sex %in% c.sex&
                               urban %in% c.urban],
                      c.by_vars = c("occ", "ind"),
                      c.var_interest = "log_incwage")

  out_dt <- rbindlist(list(occ_only, ind_only, occ_ind ))
  out_dt[,grouping := c("Occupation", "Industry", "Ind. + Occ.")]
  out_dt[, age_cat := paste(c.age_cat, collapse = ", ")]
  out_dt[, sex :=  paste(c.sex, collapse = ", ")]
  out_dt[, urban :=  paste(c.urban, collapse = ", ")]

  return(out_dt)
}

temp_male_urban <- lapply(unique(census_1940$age_cat), dem_var_gettr,
```

```
                          c.dat = census_1940,
                          c.sex = "male",
                          c.urban = "urban")
temp_male_rural <- lapply(unique(census_1940$age_cat), dem_var_gettr,
                          c.dat = census_1940,
                          c.sex = "male",
                          c.urban = "rural")

temp_female_urban <- lapply(unique(census_1940$age_cat), dem_var_gettr,
                            c.dat = census_1940,
                            c.sex = "female",
                            c.urban = "urban")
temp_female_rural <- lapply(unique(census_1940$age_cat), dem_var_gettr,
                            c.dat = census_1940,
                            c.sex = "female",
                            c.urban = "rural")

temp_female_both <- lapply(unique(census_1940$age_cat), dem_var_gettr,
                            c.dat = census_1940,
                            c.sex = "female",
                            c.urban = c("urban", "rural"))
temp_male_both <- lapply(unique(census_1940$age_cat), dem_var_gettr,
                            c.dat = census_1940,
                            c.sex = "male",
                            c.urban = c("urban", "rural"))

plot_dt <- rbindlist(list(rbindlist(temp_male_rural),
                          rbindlist(temp_female_rural),
                          rbindlist(temp_male_urban),
                          rbindlist(temp_female_urban),
                          rbindlist(temp_male_both),
                          rbindlist(temp_female_both)))

plot_dt[, f_stat := (avg_between_var/(k-1))/((avg_within_var)/(N-k))]
plot_dt[, ms_bw := (avg_between_var/(k-1))]
plot_dt[, ms_wi := (avg_within_var/(N-k))]

plot_dt[, age_start := as.numeric(substr(age_cat,1,2))]

plot_dt[, within_perc :=
          avg_within_var/
          (avg_within_var+avg_between_var)]
plot_dt[, between_perc :=
          avg_between_var/
          (avg_within_var+avg_between_var)]
plot_dt[, bw_wi_perc_ratio := between_perc/within_perc]

# ggplot(plot_dt)+
#   geom_line(aes(x = age_start, y = f_stat, color = grouping)) +
#   facet_grid(urban~sex) +
#   geom_hline(yintercept = 1, linetype = "dashed")

# cast long
```

```r
plot_dt_long <- melt(plot_dt, id.vars = c("age_cat",
                                          "urban",
                                          "grouping",
                                          "sex",
                                          "age_start"),
                     measure.vars = c("within_perc",
                                      "between_perc",
                                      "bw_wi_perc_ratio",
                                      "ms_wi",
                                      "ms_bw"))

ggplot(plot_dt_long[variable %like% "between_perc"]) +
  geom_line(aes(x = age_start, y = value,
                color = grouping))+
  facet_grid(urban ~ sex) +
  labs(x = "Age", y = "Percent Total Variance Explained")
```



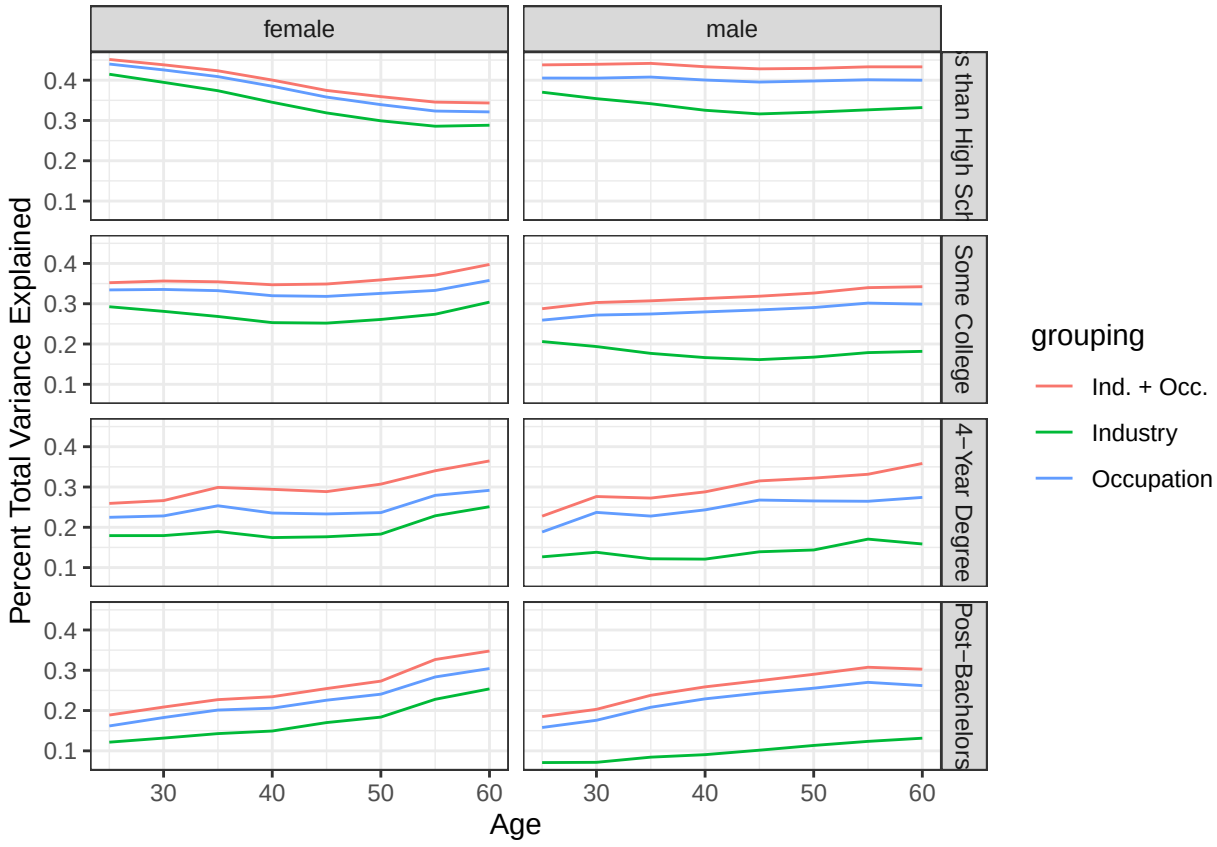

```r
#
# ggplot(plot_dt_long[variable %in% c("ms_wi", "ms_bw")]) +
#   geom_line(aes(x = age_start, y = value,
#                 color = grouping,
#                 linetype = variable))+
#   facet_grid(urban ~ sex)
```

Step 2 - Try to recreate Xie, Killewald, and Near 2016

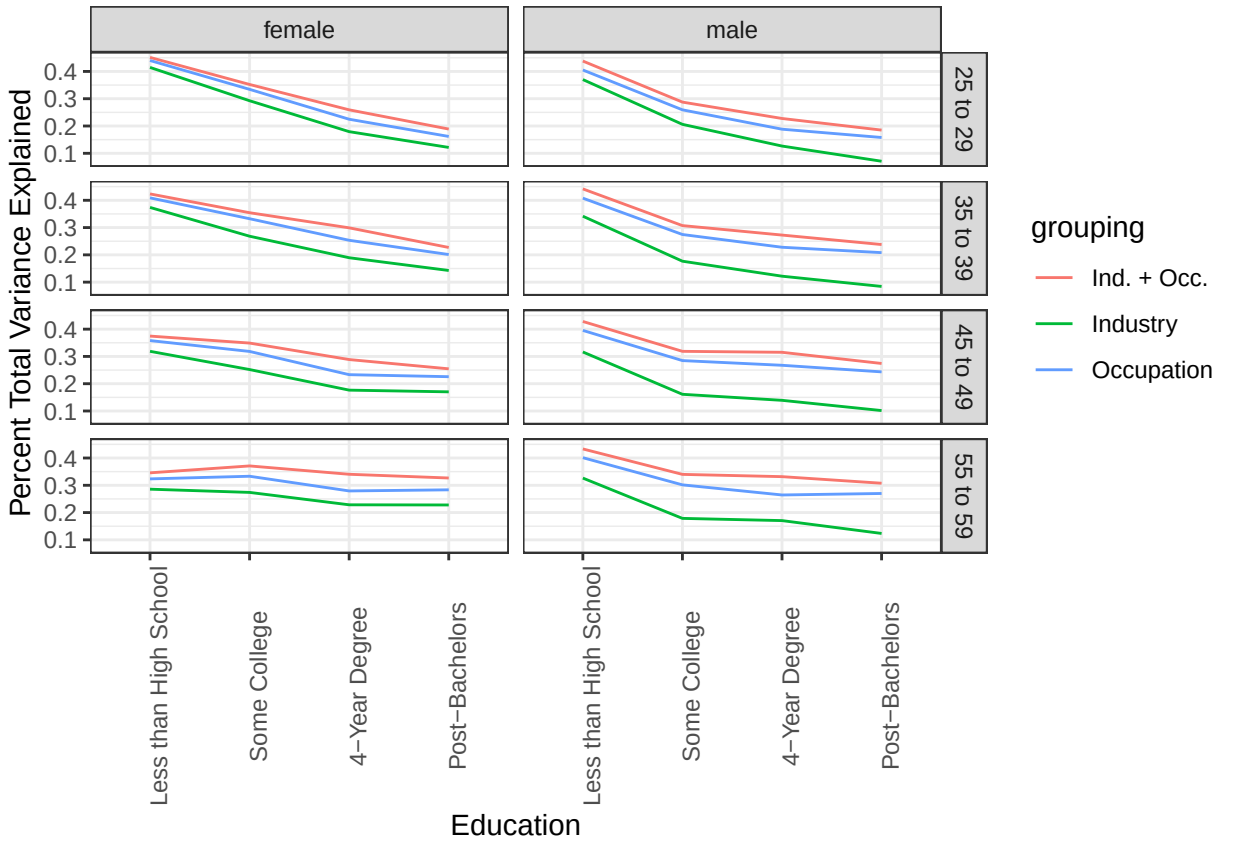
Step 1 - Bin by Education and repeat above analysis

```
# create education category
census_1940[higrade %in% 0:11, ed := "Less than High School"]
census_1940[higrade %in% 12:15, ed := "Some College"]
census_1940[higrade %in% 16, ed := "4-Year Degree"]
census_1940[higrade %in% 17:80, ed := "Post-Bachelors"]
census_1940[, ed := factor(ed, levels = c("Less than High School","Some College","4-Year Degree", "Post
)]
#  create another function to loop over data and
#  calculate occ, ind, and occ + ind var

dem_var_gettr_ed <- function(c.dat2,
                             c.age_cat,
                             c.sex,
                             c.ed){
  occ_only <- get_vars(c.dat2[age_cat %in% c.age_cat &
                                sex %in% c.sex &
                                ed %in% c.ed],
                       c.by_vars = "occ",
                       c.var_interest = "log_incwage")
  ind_only <- get_vars(c.dat2[age_cat %in% c.age_cat &
                                sex %in% c.sex&
                                 ed %in% c.ed],
                       c.by_vars = "ind",
                       c.var_interest = "log_incwage")
  occ_ind <- get_vars(c.dat2[age_cat %in% c.age_cat &
                               sex %in% c.sex&
                                ed %in% c.ed],
                      c.by_vars = c("occ", "ind"),
                      c.var_interest = "log_incwage")

  out_dt <- rbindlist(list(occ_only, ind_only, occ_ind ))
  out_dt[,grouping := c("Occupation", "Industry", "Ind. + Occ.")]
  out_dt[, age_cat := paste(c.age_cat, collapse = ", ")]
  out_dt[, sex :=  paste(c.sex, collapse = ", ")]
  out_dt[, ed :=  paste(c.ed, collapse = ", ")]

  return(out_dt)
}

ed_gettr <- function(c.ed2){
  temp <- lapply(unique(census_1940$age_cat), dem_var_gettr_ed,
                 c.dat = census_1940,
                 c.sex = "male",
                 c.ed = c.ed2)
  temp2 <- lapply(unique(census_1940$age_cat), dem_var_gettr_ed,
                 c.dat = census_1940,
                 c.sex = "female",
                 c.ed = c.ed2)
  temp <- c(temp, temp2)
  return(temp)
```

```
}


lapply(
  unique(census_1940[!is.na(ed)]$ed),
                              ed_gettr) %>%
  lapply(., rbindlist) %>%
  rbindlist() -> plot_dt2

plot_dt2[, f_stat := (avg_between_var/(k-1))/((avg_within_var)/(N-k))]
plot_dt2[, age_start := as.numeric(substr(age_cat,1,2))]

plot_dt2[, ed := factor(ed, levels = c("Less than High School","Some College","4-Year Degree", "Post-Ba
)]

plot_dt2[, r_squared := 1 - avg_within_var/(avg_within_var + avg_between_var)]

plot_dt2[, between_perc :=
           avg_between_var/
           (avg_within_var+avg_between_var)]

ggplot(plot_dt2)+
  geom_line(aes(x = age_start, y = between_perc, color = grouping)) +
  facet_grid(ed~sex) +
    labs(x = "Age", y = "Percent Total Variance Explained")
```

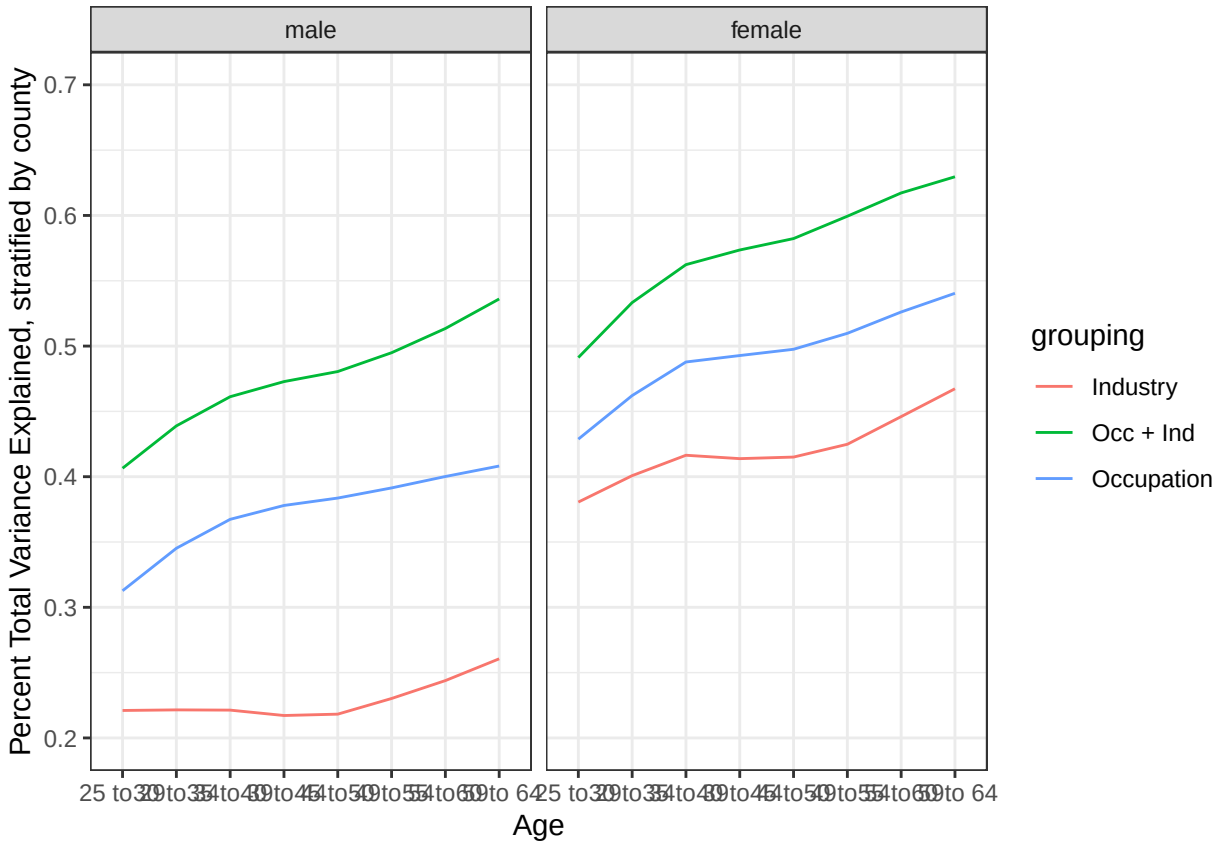


```
ggplot(plot_dt2[age_start %in% c(25, 35,45,55)])+
  geom_line(aes(x = ed, y = between_perc, color = grouping, group = grouping)) +
  facet_grid(age_cat~sex) +
  theme(axis.text.x =  element_text(angle = 90))+
    labs(x = "Education", y = "Percent Total Variance Explained")
```

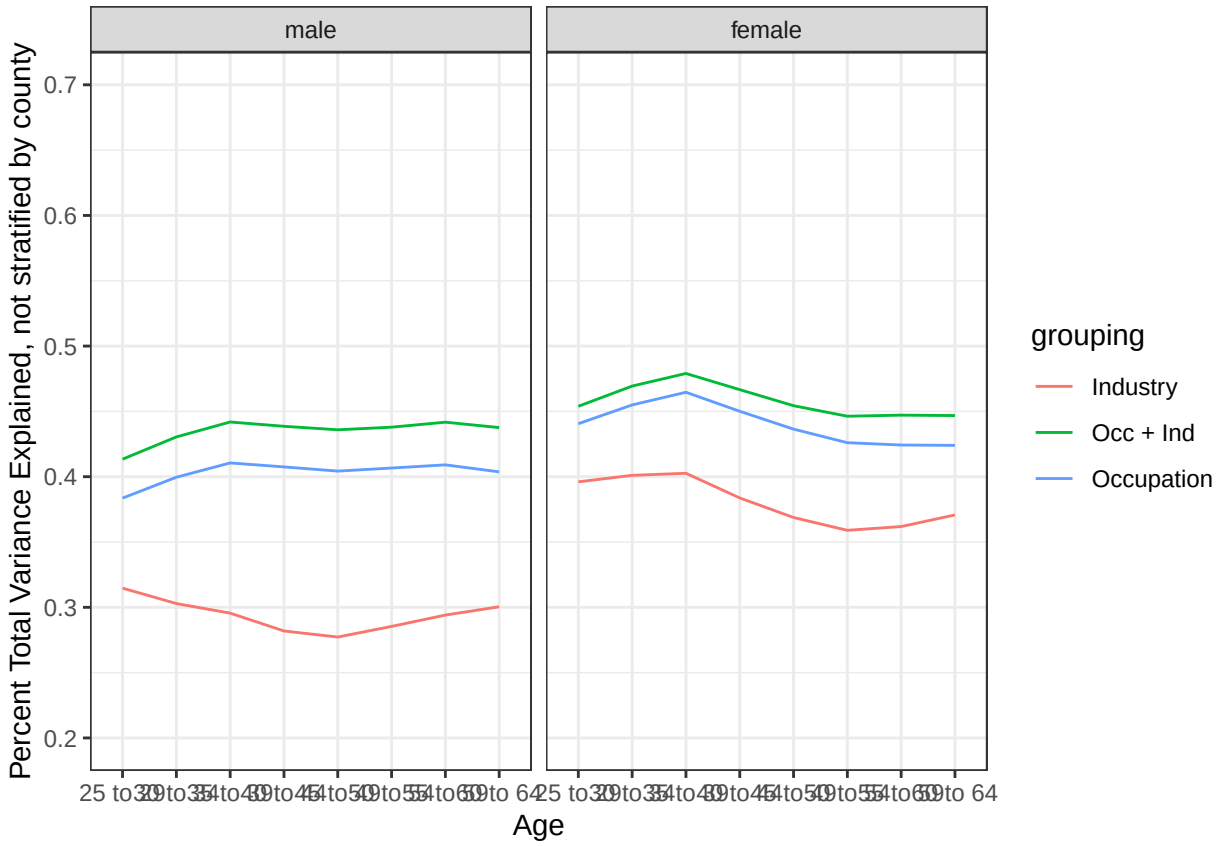



```
get_vars <- function(c.data, c.by_vars,c.by_vars_2, c.var_interest){
  out_dt_1 <- c.data[,.(w_i_ss = weighted.var(get(c.var_interest), perwt) * .N,
                        N = .N,
                        k = max(.GRP)),
                     by = c(c.by_vars, c.by_vars_2)]
  out_dt_1[N==1, w_i_ss := 0]
  out_2 <- c.data[,.(tot_ss=weighted.var(get(c.var_interest), perwt) * .N), by = c.by_vars_2]
  out_dt_1 <- merge(out_dt_1, out_2, by = c.by_vars_2)
  out_dt_1 <- out_dt_1[!is.na(tot_ss)& !is.na(w_i_ss) & !is.nan(tot_ss)& !is.nan(w_i_ss) &
                         !is.infinite(tot_ss)& !is.infinite(w_i_ss),
                       .(avg_within_var = sum(w_i_ss),
                         avg_total_var = mean(tot_ss),
                         avg_between_var = mean(tot_ss) -sum(w_i_ss),
                         N = sum(N),
                         k = length(unique(N[!is.na(w_i_ss)]))),
                       by = c.by_vars_2]
  return(out_dt_1)
}

#  create another function to loop over data and
#  calculate occ, ind, and occ + ind var
```

```r
dem_var_gettr <- function(c.dat2, c.by_vars_2){
  occ_only <- get_vars(c.dat2,
                       c.by_vars = c("occ"),
                       c.by_vars_2 = c.by_vars_2,
                       c.var_interest = "log_incwage")
  occ_only[, grouping := "Occupation"]
  ind_only <- get_vars(c.dat2,
                       c.by_vars = c("ind"),
                       c.by_vars_2 = c.by_vars_2,

                       c.var_interest = "log_incwage")
    ind_only[, grouping := "Industry"]

  occ_ind <- get_vars(c.dat2,
                      c.by_vars = c("occ", "ind"),
                      c.by_vars_2 = c.by_vars_2,
                      c.var_interest = "log_incwage")
    occ_ind[, grouping := "Occ + Ind"]


  out_dt <- rbindlist(list(occ_only, ind_only, occ_ind ))
  return(out_dt)
}

census_1940[, perwt := 1]
dem_var_gettr(census_1940, c.by_vars_2 = c("statefip", "countyicp", "sex", "age_cat")) -> plot_dt3


plot_dt3[, f_stat := (avg_between_var/(k-1))/((avg_within_var)/(N-k))]



plot_dt3[, between_perc :=
          avg_between_var/
          (avg_within_var+avg_between_var)]
plot_dt3 <- plot_dt3[!is.nan(between_perc) & !is.na(between_perc) & !is.infinite(between_perc),.(betwee

ggplot(plot_dt3)+
  geom_line(aes(x = age_cat, y = between_perc, color = grouping, group = grouping)) +
  facet_wrap(~sex) +
    labs(x = "Age", y = "Percent Total Variance Explained, stratified by county") +
  ylim(.2, .7)
```




```
census_1940[, perwt := 1]
dem_var_gettr(census_1940, c.by_vars_2 = c( "sex", "age_cat")) -> plot_dt3


plot_dt3[, f_stat := (avg_between_var/(k-1))/((avg_within_var)/(N-k))]



plot_dt3[, between_perc :=
           avg_between_var/
           (avg_within_var+avg_between_var)]
plot_dt3 <- plot_dt3[!is.nan(between_perc) & !is.na(between_perc) & !is.infinite(between_perc),.(betwee

ggplot(plot_dt3)+
  geom_line(aes(x = age_cat, y = between_perc, color = grouping, group = grouping)) +
  facet_wrap(~sex) +
    labs(x = "Age", y = "Percent Total Variance Explained, not stratified by county")+
  ylim(.2, .7)
```


male
female
Percent Total Variance Explained, not stratified by county
0.7
0.6
0.5
0.4
0.3
0.2
25 to 29
30 to 34
35 to 39
40 to 44
45 to 49
50 to 54
55 to 59
60 to 64
Age
grouping
Industry
Occ + Ind
Occupation