analysis_ii_time_series

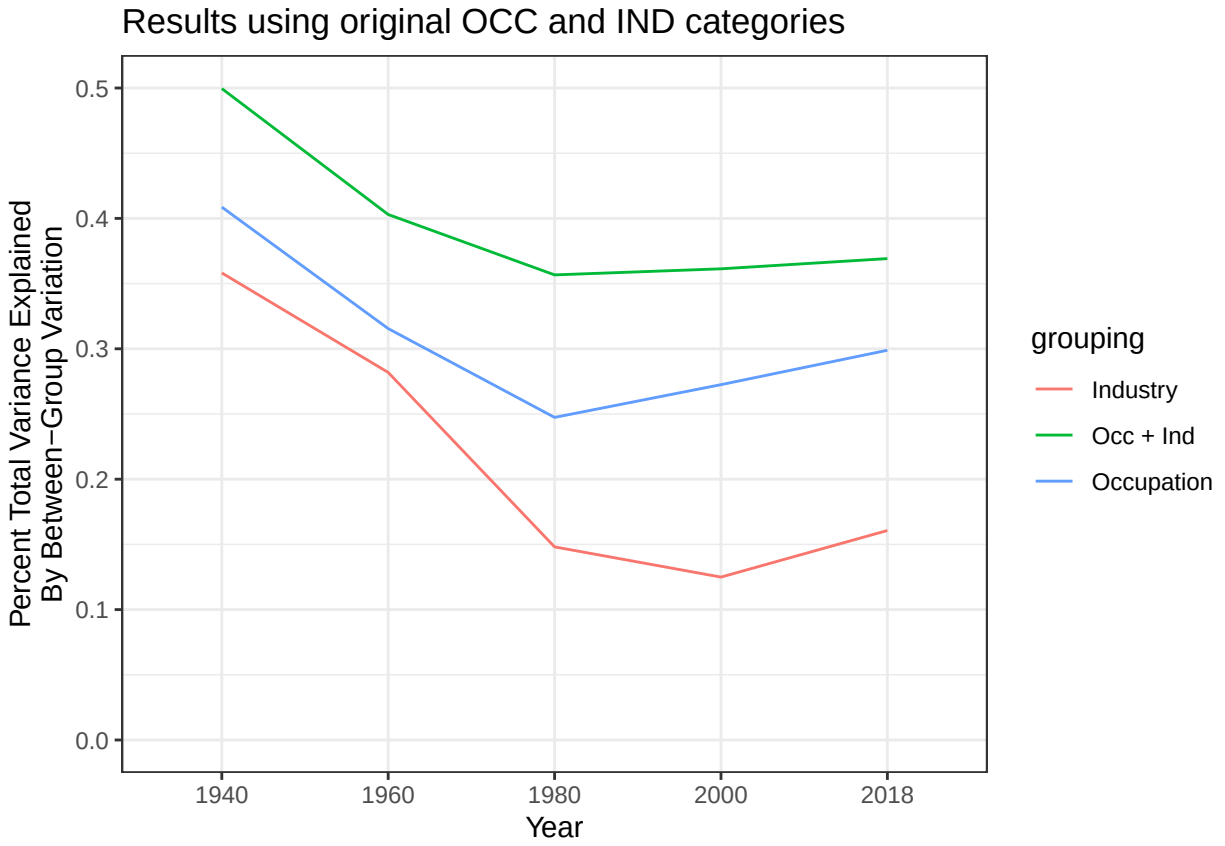
Hunter York

10/10/2020

Time Trends: Decomposition of Variation in Log Earnings across 80 Years

```
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_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

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)
}

temp <- dem_var_gettr(census_1940, c.by_vars_2 = c("year"))


plot_dt <- temp

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[, 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
plot_dt_long <- melt(plot_dt, id.vars = c("year",
                                          "grouping"),
                     measure.vars = c("within_perc",
                                      "between_perc",
                                      "bw_wi_perc_ratio",
                                      "ms_wi",
                                      "ms_bw"))

gg1 <- ggplot(plot_dt_long[variable %like% "between_perc"]) +
  geom_line(aes(x = year, y = value,
                color = grouping, group = grouping))+
  labs(x = "Year", y = "Percent Total Variance Explained\nBy Between-Group Variation",
       title = "Results using original OCC and IND categories") +
  ylim(0, .5)
print(gg1)
```




```
#
```

Now do it with a standardized industry variable and with standardized occupation variable

This uses census-to-census crosswalks. ACS 2018 xwalk values seem to be off. (50 should be 51 or 52 for managers, for instance). https://usa.ipums.org/usa/volii/occ2018.shtml

```
# load in xwalk
xwalk <- data.table(read_excel("../ref/Census_integrated_occ_crosswalks.xlsx"))
xwalk_long <- melt(xwalk, id.vars = c("OCC1950", "Occupation category description"))
```

```
## Warning in melt.data.table(xwalk, id.vars = c("OCC1950", "Occupation category
## description")): 'measure.vars' [1900, 1910, 1920, 1940, ...] are not all of the
## same type. By order of hierarchy, the molten data value column will be of type
## 'character'. All measure variables not of type 'character' will be coerced too.
## Check DETAILS in ?melt.data.table for more on coercion.
```

```
setnames(xwalk_long, c("OCC1950", "OCC1950_desc", "year", "orig_occ"))
xwalk_long[as.character(year) =="ACS 2000-02", year := "2000ACS"]
xwalk_long[as.character(year)  == "ACS 2003-", year := "2018"]

# copy 1950 vals to 1940 for nowxwalk_long
xwalk_long[, year := as.character(year)]
xwalk_long[year == 1950] %>%
  .[, year := 1940] %>%
  rbind(., xwalk_long) -> temp
```

```
xwalk_long[, orig_occ := as.numeric(orig_occ)]
```

```
## Warning in eval(jsub, SDenv, parent.frame()): NAs introduced by coercion
```

```
census_1940[, occ := as.numeric(occ)]
# merge on census
census_1940 <- merge(census_1940, xwalk_long, by.y = c("year", "orig_occ"), by.x = c("year", "occ"), al
```

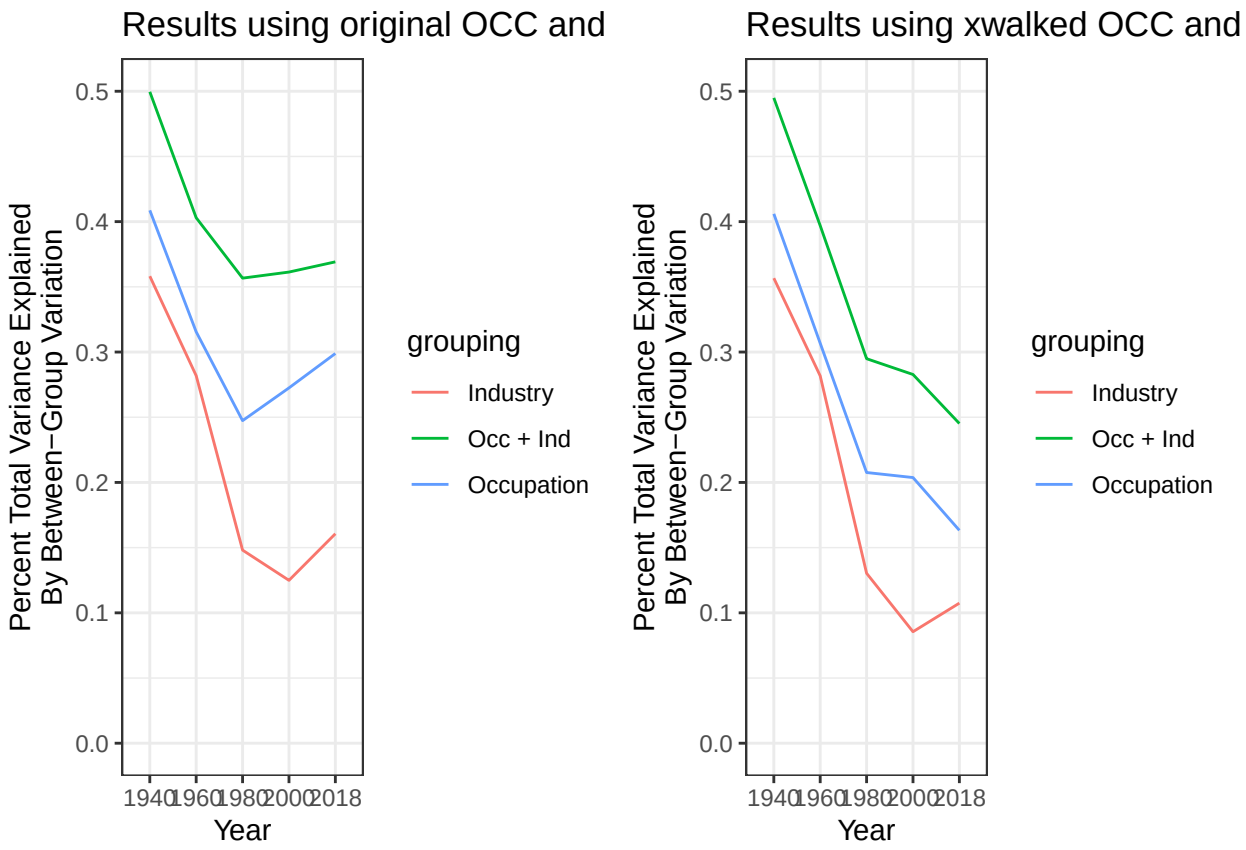
```
dem_var_gettr2 <- function(c.dat2, c.by_vars_2){
  occ_only <- get_vars(c.dat2,
                       c.by_vars = c("OCC1950"),
                       c.by_vars_2 = c.by_vars_2,
                       c.var_interest = "log_incwage")
  occ_only[, grouping := "Occupation"]
  ind_only <- get_vars(c.dat2,
                       c.by_vars = c("ind1950"),
                       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("OCC1950", "ind1950"),
                      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)
}

temp <- dem_var_gettr2(census_1940, c.by_vars_2 = c("year"))


plot_dt2 <- temp

plot_dt2[, f_stat := (avg_between_var/(k-1))/((avg_within_var)/(N-k))]
plot_dt2[, ms_bw := (avg_between_var/(k-1))]
plot_dt2[, ms_wi := (avg_within_var/(N-k))]

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


# cast long
plot_dt2_long <- melt(plot_dt2, id.vars = c("year",
```

```
                                          "grouping"),
                    measure.vars = c("within_perc",
                                     "between_perc",
                                     "bw_wi_perc_ratio",
                                     "ms_wi",
                                     "ms_bw"))

gg2 <- ggplot(plot_dt2_long[variable %like% "between_perc"]) +
  geom_line(aes(x = year, y = value,
                color = grouping, group = grouping))+
  labs(x = "Year", y = "Percent Total Variance Explained\nBy Between-Group Variation",
       title = "Results using xwalked OCC and 1950IND"
       ) +
  ylim(0, .5)

library(gridExtra)
```

```
##
## Attaching package: 'gridExtra'
```

```
## The following object is masked from 'package:dplyr':
##
##     combine
```

```
grid.arrange(gg1, gg2, nrow = 1)
```



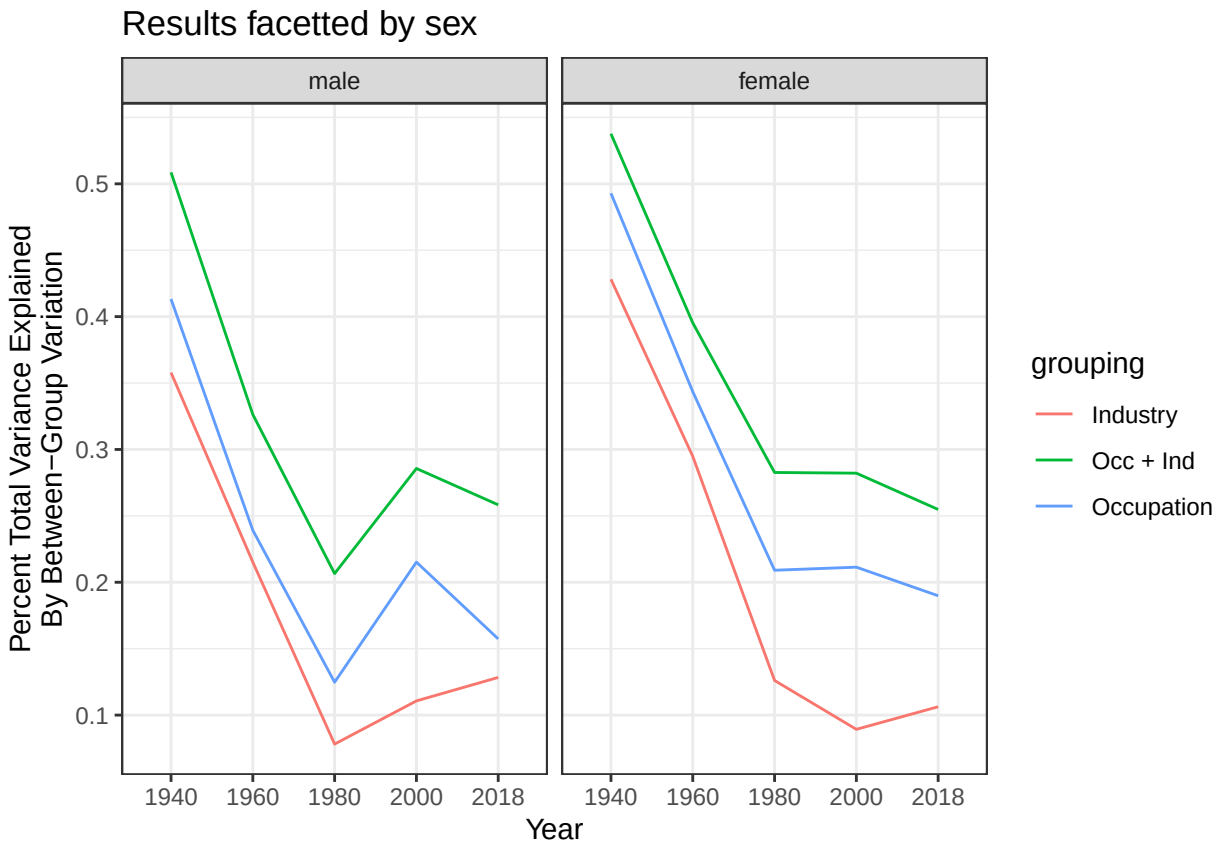

```
#
```

Tables of numbers of occ and ind code by decade

```
census_1940[,.(occ = length(unique(occ)),
               ind = length(unique(ind)),
               OCC1950 = length(unique(OCC1950)),
               ind1950 = length(unique(ind1950))), by = .(year)]
```

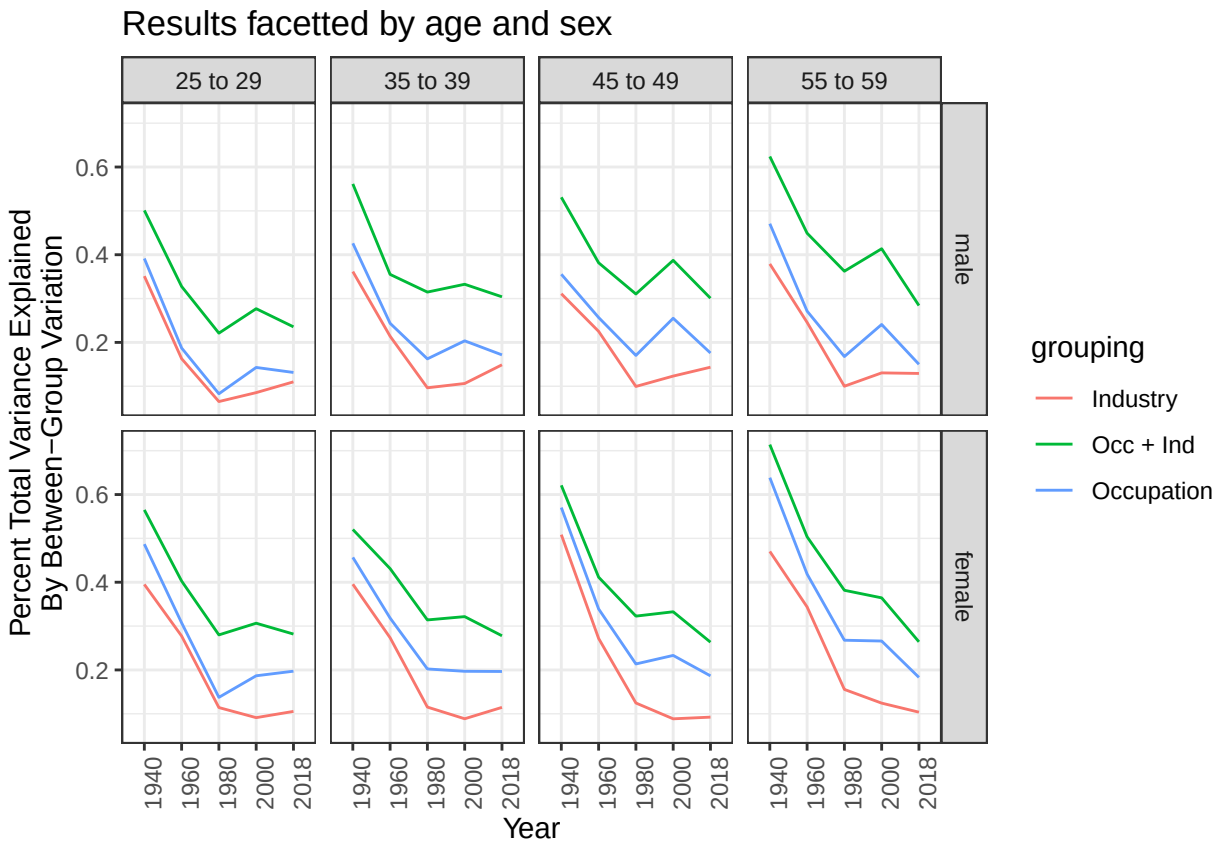
```
##    year occ ind OCC1950 ind1950
## 1: 1940 226 133     214     121
## 2: 1960 291 151     264     145
## 3: 1980 499 231     220     143
## 4: 2000 506 264     186     134
## 5: 2018 529 269     156     131
```

```
temp <- dem_var_gettr2(census_1940, c.by_vars_2 = c("year", "sex"))


plot_dt2 <- temp

plot_dt2[, f_stat := (avg_between_var/(k-1))/((avg_within_var)/(N-k))]
plot_dt2[, ms_bw := (avg_between_var/(k-1))]
plot_dt2[, ms_wi := (avg_within_var/(N-k))]

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


# cast long
plot_dt2_long <- melt(plot_dt2, id.vars = c("year",
                                            "sex",
                                           "grouping"),
                      measure.vars = c("within_perc",
                                       "between_perc",
                                       "bw_wi_perc_ratio",
                                       "ms_wi",
                                       "ms_bw"))

ggplot(plot_dt2_long[variable %like% "between_perc" ]) +
  geom_line(aes(x = year, y = value,
                color = grouping, group = grouping))+
  facet_wrap(~sex)+
  labs(x = "Year", y = "Percent Total Variance Explained\nBy Between-Group Variation",
       title = "Results facetted by sex")
```

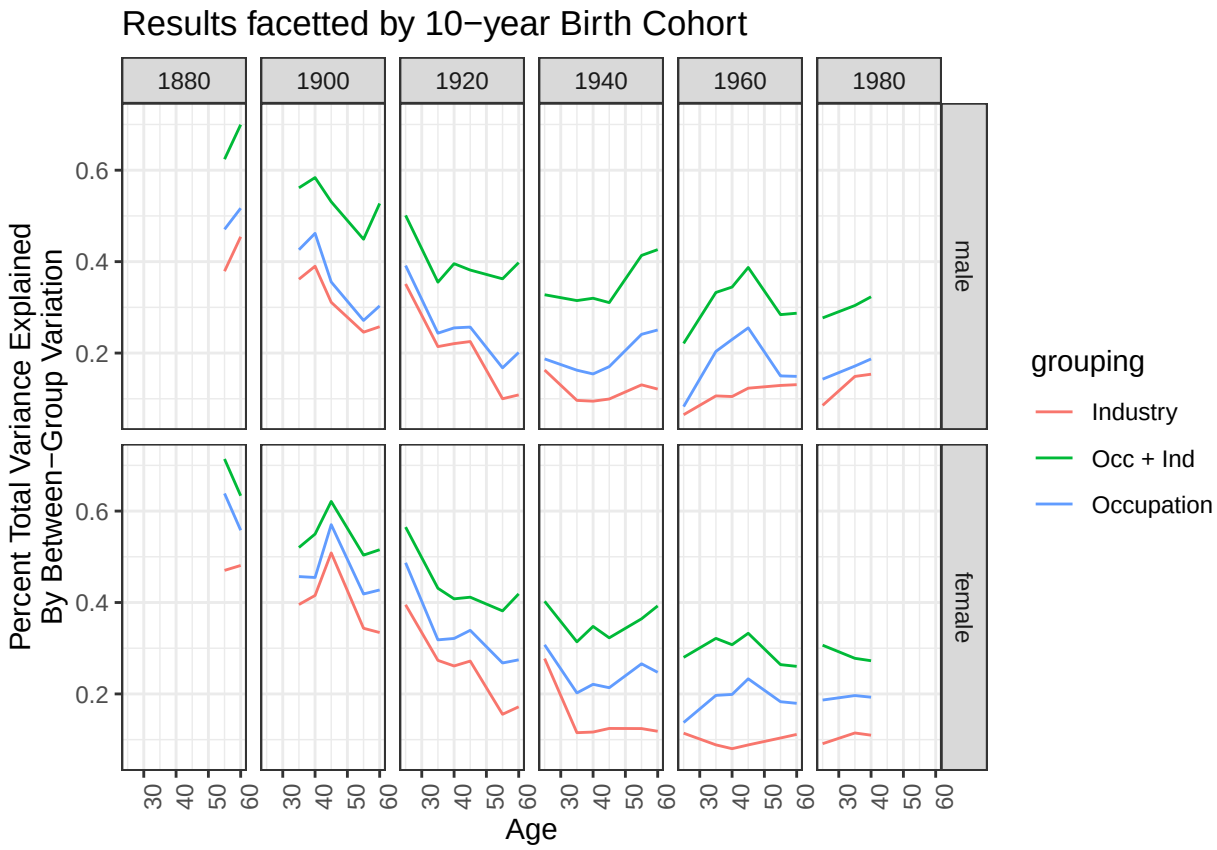



```
temp <- dem_var_gettr2(census_1940, c.by_vars_2 = c("year", "sex", "age_cat"))


plot_dt2 <- temp

plot_dt2[, f_stat := (avg_between_var/(k-1))/((avg_within_var)/(N-k))]
plot_dt2[, ms_bw := (avg_between_var/(k-1))]
plot_dt2[, ms_wi := (avg_within_var/(N-k))]

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


# cast long
plot_dt2_long <- melt(plot_dt2, id.vars = c("year",
                                            "sex",
                                            "age_cat",
                                           "grouping"),
                      measure.vars = c("within_perc",
                                       "between_perc",
                                       "bw_wi_perc_ratio",
                                       "ms_wi",
```

```
                                        "ms_bw"))

ggplot(plot_dt2_long[variable %like% "between_perc"& age_cat %like% "5 to"]) +
  geom_line(aes(x = year, y = value,
                color = grouping, group = grouping))+
  facet_grid(sex~age_cat)+
  labs(x = "Year", y = "Percent Total Variance Explained\nBy Between-Group Variation",
       title = "Results facetted by age and sex") +
  theme(axis.text.x = element_text(angle = 90))
```




```
plot_dt2_long[, age_start := as.numeric(substr(age_cat,1,2))]
plot_dt2_long[, cohort := as.numeric(year) - age_start]
plot_dt2_long[, cohort := round(as.numeric(cohort)/10)*10]
ggplot(plot_dt2_long[variable %like% "between_perc" & cohort %in% seq(1880,1980,20)]) +
  geom_line(aes(x = age_start, y = value,
                color = grouping, group = grouping))+
  facet_grid(sex~cohort)+
  labs(x = "Age", y = "Percent Total Variance Explained\nBy Between-Group Variation",
       title = "Results facetted by 10-year Birth Cohort") +
  theme(axis.text.x = element_text(angle = 90))
```


Results facetted by 10−year Birth Cohort
1880
1900
1920
1940
1960
1980
male
female
Percent Total Variance Explained
By Between−Group Variation
0.2
0.4
0.6
30
40
50
60
Age
grouping
Industry
Occ + Ind
Occupation