#\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

# PROJECT: KILLEWALD- GENDER WAGE DISTRIBUTION

# ANALYSIS FILE: COUNTERFACTUAL DECOMPOSITION ANALYSES

# AUTHOR: NINO CRICCO

# LAST UPDATED: 09/06/2020 (dmy)

# RUNTIME: 1131.871 sec (~19 min)

#\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

# NOTE: THIS CODE FILE DRAWS ON IMPUTS PRODUCED IN THE ANALYSIS\_MAIN.R FILE AND SHOULD BE RUN AFTER

#\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

# FIGURE 4: COUNTERFACTUAL SCENARIOS FOR CHANGES IN THE GENDER WAGE GAP

#\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

# We first "ungroup" the means dataframe (grouped by sex and year)

means.year <- ungroup(means.year)

# These vectors are later used for variable selection

fam\_vars <- c("(Intercept)", "married", "prev.married", "numkids.3plus",

"afb.cat\_21minus", "afb.cat\_22to27", "afb.cat\_27plus")

full\_vars <- c("(Intercept)", "Northeast", "Northcentral", "South", "Black", "Hispanic", "Other",

"married", "prev.married", "housework", "numkids.3plus",

"afb.cat\_21minus", "afb.cat\_22to27", "afb.cat\_27plus", "HighSchool",

"SomeCollege", "ba.advdeg", "union", "govt.job", "log.expf", "hrswrk.dummy\_50plus",

"emp.tenure", "occ.pct.female", "occ.managers", "manuf")

# First, we compute the observed gender wage gap by year

observed\_gap1980 <- exp(means.year$lnhrlywage[means.year$female == 1 & means.year$year == 1981])/

exp(means.year$lnhrlywage[means.year$female == 0 & means.year$year == 1981])

observed\_gap1991 <- exp(means.year$lnhrlywage[means.year$female == 1 & means.year$year == 1991]) /

exp(means.year$lnhrlywage[means.year$female == 0 & means.year$year == 1991])

observed\_gap2001 <- exp(means.year$lnhrlywage[means.year$female == 1 & means.year$year == 2001]) /

exp(means.year$lnhrlywage[means.year$female == 0 & means.year$year == 2001])

observed\_gap2011 <- exp(means.year$lnhrlywage[means.year$female == 1 & means.year$year == 2011]) /

exp(means.year$lnhrlywage[means.year$female == 0 & means.year$year == 2011])

observed\_gap2019 <- exp(means.year$lnhrlywage[means.year$female == 1 & means.year$year == 2019]) /

exp(means.year$lnhrlywage[means.year$female == 0 & means.year$year == 2019])

# Across counterfactual scenarios, we use the sex-specific coefficients for each model for the start of the period

wcoef\_fam1980 <- fam.coefs %>% filter(female == 1, estimate == "coef", year == 1981) %>%

dplyr::select(-c(estimate, female, year))

mcoef\_fam1980 <- fam.coefs %>% filter(female == 0, estimate == "coef", year == 1981) %>%

dplyr::select(-c(estimate, female, year))

wcoef\_full1980 <- full.coefs %>% filter(female == 1, estimate == "coef", year == 1981) %>%

dplyr::select(-c(estimate, female, year))

mcoef\_full1980 <- full.coefs %>% filter(female == 0, estimate == "coef", year == 1981) %>%

dplyr::select(-c(estimate, female, year))

# We then compute the predicted gender wage gap in each model by multiplying the

# year-sex-specific means to the appropriate regression coefficients

# for each model and year-sex combination

# We use these intermediate objects later in the counterfactuaal scenarios

woutcome\_fam <- as.matrix(means.year %>% filter(female == 1, year == 1981) %>%

dplyr::select(fam\_vars)) %\*% t(wcoef\_fam1980)

moutcome\_fam <- as.matrix(means.year %>% filter(female == 0, year == 1981) %>%

dplyr::select(fam\_vars)) %\*% t(mcoef\_fam1980)

fam1980 <- exp(woutcome\_fam)/exp(moutcome\_fam)

full1980 <- exp(as.matrix(means.year %>% filter(female == 1, year == 1981) %>%

dplyr::select(full\_vars)) %\*% t(wcoef\_full1980)) /

exp(as.matrix(means.year %>% filter(female == 0, year == 1981) %>%

dplyr::select(full\_vars)) %\*% t(mcoef\_full1980))

fam1991 <- exp((as.matrix(means.year %>% filter(female == 1, year == 1991) %>%

dplyr::select(fam\_vars)) %\*% t(wcoef\_fam1980))) /

exp((as.matrix(means.year %>% filter(female == 0, year == 1991) %>%

dplyr::select(fam\_vars)) %\*% t(mcoef\_fam1980)))

full1991 <- exp((as.matrix(means.year %>% filter(female == 1, year == 1991) %>%

dplyr::select(full\_vars)) %\*% t(wcoef\_full1980))) /

exp((as.matrix(means.year %>% filter(female == 0, year == 1991) %>%

dplyr::select(full\_vars)) %\*% t(mcoef\_full1980)))

fam2001 <- exp((as.matrix(means.year %>% filter(female == 1, year == 2001) %>%

dplyr::select(fam\_vars)) %\*% t(wcoef\_fam1980))) /

exp((as.matrix(means.year %>% filter(female == 0, year == 2001) %>%

dplyr::select(fam\_vars)) %\*% t(mcoef\_fam1980)))

full2001 <- exp((as.matrix(means.year %>% filter(female == 1, year == 2001) %>%

dplyr::select(full\_vars)) %\*% t(wcoef\_full1980))) /

exp((as.matrix(means.year %>% filter(female == 0, year == 2001) %>%

dplyr::select(full\_vars)) %\*% t(mcoef\_full1980)))

fam2011 <- exp((as.matrix(means.year %>% filter(female == 1, year == 2011) %>%

dplyr::select(fam\_vars)) %\*% t(wcoef\_fam1980))) /

exp((as.matrix(means.year %>% filter(female == 0, year == 2011) %>%

dplyr::select(fam\_vars)) %\*% t(mcoef\_fam1980)))

full2011 <- exp((as.matrix(means.year %>% filter(female == 1, year == 2011) %>%

dplyr::select(full\_vars)) %\*% t(wcoef\_full1980))) /

exp((as.matrix(means.year %>% filter(female == 0, year == 2011) %>%

dplyr::select(full\_vars)) %\*% t(mcoef\_full1980)))

fam2019 <- exp((as.matrix(means.year %>% filter(female == 1, year == 2019) %>%

dplyr::select(fam\_vars)) %\*% t(wcoef\_fam1980))) /

exp((as.matrix(means.year %>% filter(female == 0, year == 2019) %>%

dplyr::select(fam\_vars)) %\*% t(mcoef\_fam1980)))

full2019 <- exp((as.matrix(means.year %>% filter(female == 1, year == 2019) %>%

dplyr::select(full\_vars)) %\*% t(wcoef\_full1980))) /

exp((as.matrix(means.year %>% filter(female == 0, year == 2019) %>%

dplyr::select(full\_vars)) %\*% t(mcoef\_full1980)))

# The following section computes the counterfactual wage gaps for each yeaar

# Across all counterfactual scenarios, we hold the "returns" to men and women's characteristics

# at the level observed in 1980, so we multiply the hypothetical changes in men and women's

# characteristics by the sex-specific regression coefficients in 1980 to get predicted wages

# under different counterfactual scenarios of characteristics change

# First we compute the scenarios highlighting the composition process

# In this scenario, the "levels" of men and women's family characteristics change over time,

# but sex gap in characteristics stays constant at 1980 levels

# When "women" are the reference group, women's family characteristics are anchored to

# changes in men's family characteristics: that is, men's family characteristics change as observed,

# while women's family characteristics are manipulated to change in ways that maintain the

# characteristics gap with men observed in 1980

# Creating the counterfactual means for women under the composition scenario

q\_comp\_1991\_women <- means.year %>% filter(female == 0, year == 1991) %>%

dplyr::select(fam\_vars) +

(means.year %>% filter(female == 1, year == 1981) %>%

dplyr::select(fam\_vars) -

means.year %>% filter(female == 0, year == 1981) %>%

dplyr::select(fam\_vars))

# Computing the numerator (women's average pay under the composition scenario \*

# women's coefficients for the family model in 1980

num\_comp\_1991\_refwomen <- as.matrix(wcoef\_fam1980) %\*% t(q\_comp\_1991\_women)

# Computing the denominator (men's average pay changes as observed \*

# men's coefficients for the family model in 1980

denom\_comp\_1991\_refwomen <- as.matrix(mcoef\_fam1980) %\*%

t(means.year %>% filter(female == 0, year == 1991) %>% dplyr::select(fam\_vars))

# Computing the counterfactual pay gap

comp\_1991\_refwomen <- exp(num\_comp\_1991\_refwomen)/exp(denom\_comp\_1991\_refwomen)

# When "men" are the reference group, men's family characteristics are anchored to

# changes in women's family characteristics: that is, women's family characteristics change as observed,

# while men's family characteristics are manipulated to change in ways that maintain the

# characteristics gap with women observed in 1980

# Creating the counterfactual means for men under the composition scenario

q\_comp\_1991\_men <- means.year %>% filter(female == 1, year == 1991) %>%

dplyr::select(fam\_vars) +

(means.year %>% filter(female == 0, year == 1981) %>%

dplyr::select(fam\_vars) -

means.year %>% filter(female == 1, year == 1981) %>%

dplyr::select(fam\_vars))

# Computing the numerator (women's average pay changes as observed \*

# women's coefficients for the family model in 1980)

num\_comp\_1991\_refmen <- as.matrix(wcoef\_fam1980) %\*%

t(means.year %>% filter(female == 1, year == 1991) %>% dplyr::select(fam\_vars))

denom\_comp\_1991\_refmen <- as.matrix(mcoef\_fam1980) %\*% t(q\_comp\_1991\_men)

comp\_1991\_refmen <- exp(num\_comp\_1991\_refmen)/exp(denom\_comp\_1991\_refmen)

# Composition: 2001

q\_comp\_2001\_women <- means.year %>% filter(female == 0, year == 2001) %>%

dplyr::select(fam\_vars) +

(means.year %>% filter(female == 1, year == 1981) %>%

dplyr::select(fam\_vars) -

means.year %>% filter(female == 0, year == 1981) %>%

dplyr::select(fam\_vars))

q\_comp\_2001\_men <- means.year %>% filter(female == 1, year == 2001) %>%

dplyr::select(fam\_vars) +

(means.year %>% filter(female == 0, year == 1981) %>%

dplyr::select(fam\_vars) -

means.year %>% filter(female == 1, year == 1981) %>%

dplyr::select(fam\_vars))

num\_comp\_2001\_refwomen <- as.matrix(wcoef\_fam1980) %\*% t(q\_comp\_2001\_women)

denom\_comp\_2001\_refwomen <- as.matrix(mcoef\_fam1980) %\*%

t(means.year %>% filter(female == 0, year == 2001) %>% dplyr::select(fam\_vars))

comp\_2001\_refwomen <- exp(num\_comp\_2001\_refwomen)/exp(denom\_comp\_2001\_refwomen)

num\_comp\_2001\_refmen <- as.matrix(wcoef\_fam1980) %\*%

t(means.year %>% filter(female == 1, year == 2001) %>% dplyr::select(fam\_vars))

denom\_comp\_2001\_refmen <- as.matrix(mcoef\_fam1980) %\*% t(q\_comp\_2001\_men)

comp\_2001\_refmen <- exp(num\_comp\_2001\_refmen)/exp(denom\_comp\_2001\_refmen)

# Composition: 2011

q\_comp\_2011\_women <- means.year %>% filter(female == 0, year == 2011) %>%

dplyr::select(fam\_vars) +

(means.year %>% filter(female == 1, year == 1981) %>%

dplyr::select(fam\_vars) -

means.year %>% filter(female == 0, year == 1981) %>%

dplyr::select(fam\_vars))

q\_comp\_2011\_men <- means.year %>% filter(female == 1, year == 2011) %>%

dplyr::select(fam\_vars) +

(means.year %>% filter(female == 0, year == 1981) %>%

dplyr::select(fam\_vars) -

means.year %>% filter(female == 1, year == 1981) %>%

dplyr::select(fam\_vars))

num\_comp\_2011\_refwomen <- as.matrix(wcoef\_fam1980) %\*% t(q\_comp\_2011\_women)

denom\_comp\_2011\_refwomen <- as.matrix(mcoef\_fam1980) %\*%

t(means.year %>% filter(female == 0, year == 2011) %>% dplyr::select(fam\_vars))

comp\_2011\_refwomen <- exp(num\_comp\_2011\_refwomen)/exp(denom\_comp\_2011\_refwomen)

num\_comp\_2011\_refmen <- as.matrix(wcoef\_fam1980) %\*%

t(means.year %>% filter(female == 1, year == 2011) %>% dplyr::select(fam\_vars))

denom\_comp\_2011\_refmen <- as.matrix(mcoef\_fam1980) %\*% t(q\_comp\_2011\_men)

comp\_2011\_refmen <- exp(num\_comp\_2011\_refmen)/exp(denom\_comp\_2011\_refmen)

# Composition: 2019

q\_comp\_2019\_women <- means.year %>% filter(female == 0, year == 2019) %>%

dplyr::select(fam\_vars) +

(means.year %>% filter(female == 1, year == 1981) %>%

dplyr::select(fam\_vars) -

means.year %>% filter(female == 0, year == 1981) %>%

dplyr::select(fam\_vars))

q\_comp\_2019\_men <- means.year %>% filter(female == 1, year == 2019) %>%

dplyr::select(fam\_vars) +

(means.year %>% filter(female == 0, year == 1981) %>%

dplyr::select(fam\_vars) -

means.year %>% filter(female == 1, year == 1981) %>%

dplyr::select(fam\_vars))

num\_comp\_2019\_refwomen <- as.matrix(wcoef\_fam1980) %\*% t(q\_comp\_2019\_women)

denom\_comp\_2019\_refwomen <- as.matrix(mcoef\_fam1980) %\*%

t(means.year %>% filter(female == 0, year == 2019) %>% dplyr::select(fam\_vars))

comp\_2019\_refwomen <- exp(num\_comp\_2019\_refwomen)/exp(denom\_comp\_2019\_refwomen)

num\_comp\_2019\_refmen <-as.matrix(wcoef\_fam1980) %\*%

t(means.year %>% filter(female == 1, year == 2019) %>% dplyr::select(fam\_vars))

denom\_comp\_2019\_refmen <- as.matrix(mcoef\_fam1980) %\*% t(q\_comp\_2019\_men)

comp\_2019\_refmen <- exp(num\_comp\_2019\_refmen)/exp(denom\_comp\_2019\_refmen)

# We then compute the scenarios highlighting the convergence process

# In this scenario, the sex gap in family characteristics changes over time, but

# remain anchored to the "levels" of men and women's family characteristics at their 1980 levels.

# For these scenarios, the gap between men and women's family characteristics changes as observed,

# but one group remains anchored at their 1980 levels. When women are the reference group, menâ€™s traits

# are anchored at their 1980 levels and womenâ€™s traits converge to / diverge from that level as observed.

conv\_1991\_women <- exp(as.matrix(wcoef\_fam1980) %\*% # We multiply women's 1980 family coefficients by

# counterfactual traits for women, where their traits change as they did relative

# to men's over time but remain anchored to men's 1980 levels

t((means.year %>% filter(female == 0, year == 1981) %>%

dplyr::select(fam\_vars) +

(means.year %>% filter(female == 1, year == 1991) %>%

dplyr::select(fam\_vars) -

means.year %>% filter(female == 0, year == 1991) %>%

dplyr::select(fam\_vars)))))/

# Here, we already computed the predicted outcome for men with traits at 1980 levels and

# with coefficients from the 1980 family model for men- men's traits are anchored at their 1980 levels

exp(moutcome\_fam)

# When men are the reference groups, women's traits are anchored at their 1980 levels

# We use the already computed predicted outcome for women with traits at their 1980 levels and

# with coefficients from the 1980 family model for women

conv\_1991\_men <- exp(woutcome\_fam) /

# We then multiply men's 1980 family coefficients by counterfactual traits for men,

# where men's traits change as they did over time relative to women's, but remain anchored at women's 1980 levels

exp(as.matrix(mcoef\_fam1980) %\*% t(means.year %>%

filter(female == 1, year == 1981) %>%

dplyr::select(fam\_vars) +

(means.year %>%

filter(female == 0, year == 1991) %>%

dplyr::select(fam\_vars) -

means.year %>%

filter(female == 1, year == 1991) %>%

dplyr::select(fam\_vars))))

# Repeating the same procedure as above for subsequent years

conv\_2001\_women <- exp(as.matrix(wcoef\_fam1980) %\*% t((means.year %>%

filter(female == 0, year == 1981) %>%

dplyr::select(fam\_vars)+

(means.year %>%

filter(female == 1, year == 2001) %>% dplyr::select(fam\_vars) -

means.year %>%

filter(female == 0, year == 2001) %>% dplyr::select(fam\_vars)))))/

exp(moutcome\_fam)

conv\_2001\_men <- exp(woutcome\_fam) /

exp(as.matrix(mcoef\_fam1980) %\*% t(means.year %>%

filter(female == 1, year == 1981) %>%

dplyr::select(fam\_vars) +

(means.year %>%

filter(female == 0, year == 2001) %>%

dplyr::select(fam\_vars) -

means.year %>%

filter(female == 1, year == 2001) %>%

dplyr::select(fam\_vars))))

conv\_2011\_women <- exp(as.matrix(wcoef\_fam1980) %\*% t((means.year %>%

filter(female == 0, year == 1981) %>%

dplyr::select(fam\_vars)+

(means.year %>%

filter(female == 1, year == 2011) %>% dplyr::select(fam\_vars) -

means.year %>%

filter(female == 0, year == 2011) %>% dplyr::select(fam\_vars)))))/

exp(moutcome\_fam)

conv\_2011\_men <- exp(woutcome\_fam) /

exp(as.matrix(mcoef\_fam1980) %\*% t(means.year %>%

filter(female == 1, year == 1981) %>%

dplyr::select(fam\_vars) +

(means.year %>%

filter(female == 0, year == 2011) %>%

dplyr::select(fam\_vars) -

means.year %>%

filter(female == 1, year == 2011) %>%

dplyr::select(fam\_vars))))

conv\_2019\_women <- exp(as.matrix(wcoef\_fam1980) %\*% t((means.year %>%

filter(female == 0, year == 1981) %>%

dplyr::select(fam\_vars)+

(means.year %>%

filter(female == 1, year == 2019) %>% dplyr::select(fam\_vars) -

means.year %>%

filter(female == 0, year == 2019) %>% dplyr::select(fam\_vars)))))/

exp(moutcome\_fam)

conv\_2019\_men <- exp(woutcome\_fam) /

exp(as.matrix(mcoef\_fam1980) %\*% t(means.year %>%

filter(female == 1, year == 1981) %>%

dplyr::select(fam\_vars) +

(means.year %>%

filter(female == 0, year == 2019) %>%

dplyr::select(fam\_vars) -

means.year %>%

filter(female == 1, year == 2019) %>%

dplyr::select(fam\_vars))))

# For the full models, we use the same procedure for both counterfactual scenarios

# but the level of men and women's characteristics change as observed for

# non-family characteristics

full\_vars\_exc <- c("Northeast", "Northcentral", "South", "Black", "Hispanic", "Other",

"housework", "HighSchool",

"SomeCollege", "ba.advdeg", "union", "govt.job", "log.expf", "hrswrk.dummy\_50plus",

"emp.tenure", "occ.pct.female", "occ.managers", "manuf")

# In the composition scenario, we use the already computed counterfactual family traits for women by year

q\_comp\_1991\_women\_full <- bind\_cols(q\_comp\_1991\_women,

# and combine them with the observed non-family characteristics

means.year %>% filter(female == 1, year == 1991) %>%

dplyr::select(full\_vars\_exc)) %>%

dplyr::select(full\_vars) # this line re-orders the variables in the appropriate order

# Multiplying coefficients from women's full 1980 regression to women's counterfactual

# characteristics in 1991 (non-family variables change as observed)

num\_comp\_1991\_refwomen\_full <- as.matrix(wcoef\_full1980) %\*% t(q\_comp\_1991\_women\_full)

# Multiplying coefficients from men's full 1980 regression to men's observed 1991 characteristics

denom\_comp\_1991\_refwomen\_full <- as.matrix(mcoef\_full1980) %\*%

t(means.year %>% filter(female == 0, year == 1991) %>% dplyr::select(full\_vars))

# Computing wage gap under the composition counterfactual scenario with the full model and women as ref group

comp\_1991\_refwomen\_full <- exp(num\_comp\_1991\_refwomen\_full)/exp(denom\_comp\_1991\_refwomen\_full)

# Repeating the same procedure as above, but using the already computed counterfactual family traits for men by year

q\_comp\_1991\_men\_full <- bind\_cols(q\_comp\_1991\_men,

means.year %>% filter(female == 0, year == 1991) %>%

dplyr::select(full\_vars\_exc)) %>%

dplyr::select(full\_vars)

# When men are the reference group, women's characteristics change as observed

num\_comp\_1991\_refmen\_full <- as.matrix(wcoef\_full1980) %\*%

t(means.year %>%

filter(female == 1, year == 1991) %>%

dplyr::select(full\_vars))

# Men's family characteristics change as per the composition counterfactual scenario,

# non-family characteristics change as observed. Multiplying by men's full coefficients in 1980

denom\_comp\_1991\_refmen\_full <- as.matrix(mcoef\_full1980) %\*% t(q\_comp\_1991\_men\_full)

# Computing wage gap under the composition counterfactual scenario with the full model and men as ref group

comp\_1991\_refmen\_full <- exp(num\_comp\_1991\_refmen\_full)/exp(denom\_comp\_1991\_refmen\_full)

# Composition, full, 2001, women as reference group

q\_comp\_2001\_women\_full <- bind\_cols(q\_comp\_2001\_women,

means.year %>% filter(female == 1, year == 2001) %>%

dplyr::select(full\_vars\_exc)) %>%

dplyr::select(full\_vars)

num\_comp\_2001\_refwomen\_full <- as.matrix(wcoef\_full1980) %\*% t(q\_comp\_2001\_women\_full)

denom\_comp\_2001\_refwomen\_full <- as.matrix(mcoef\_full1980) %\*%

t(means.year %>% filter(female == 0, year == 2001) %>% dplyr::select(full\_vars))

comp\_2001\_refwomen\_full <- exp(num\_comp\_2001\_refwomen\_full)/exp(denom\_comp\_2001\_refwomen\_full)

# Composition, full, 2001, men as reference group

q\_comp\_2001\_men\_full <- bind\_cols(q\_comp\_2001\_men,

means.year %>% filter(female == 0, year == 2001) %>%

dplyr::select(full\_vars\_exc)) %>%

dplyr::select(full\_vars)

num\_comp\_2001\_refmen\_full <- as.matrix(wcoef\_full1980) %\*%

t(means.year %>% filter(female == 1, year == 2001) %>% dplyr::select(full\_vars))

denom\_comp\_2001\_refmen\_full <- as.matrix(mcoef\_full1980) %\*% t(q\_comp\_2001\_men\_full)

comp\_2001\_refmen\_full <- exp(num\_comp\_2001\_refmen\_full)/exp(denom\_comp\_2001\_refmen\_full)

# Composition, full, 2011, women as reference group

q\_comp\_2011\_women\_full <- bind\_cols(q\_comp\_2011\_women,

means.year %>% filter(female == 1, year == 2011) %>%

dplyr::select(full\_vars\_exc)) %>%

dplyr::select(full\_vars)

num\_comp\_2011\_refwomen\_full <- as.matrix(wcoef\_full1980) %\*% t(q\_comp\_2011\_women\_full)

denom\_comp\_2011\_refwomen\_full <- as.matrix(mcoef\_full1980) %\*%

t(means.year %>% filter(female == 0, year == 2011) %>% dplyr::select(full\_vars))

comp\_2011\_refwomen\_full <- exp(num\_comp\_2011\_refwomen\_full)/exp(denom\_comp\_2011\_refwomen\_full)

# Composition, full, 2011, men as reference group

q\_comp\_2011\_men\_full <- bind\_cols(q\_comp\_2011\_men,

means.year %>% filter(female == 0, year == 2011) %>%

dplyr::select(full\_vars\_exc)) %>%

dplyr::select(full\_vars)

num\_comp\_2011\_refmen\_full <- as.matrix(wcoef\_full1980) %\*%

t(means.year %>% filter(female == 1, year == 2011) %>% dplyr::select(full\_vars))

denom\_comp\_2011\_refmen\_full <- as.matrix(mcoef\_full1980) %\*% t(q\_comp\_2011\_men\_full)

comp\_2011\_refmen\_full <- exp(num\_comp\_2011\_refmen\_full)/exp(denom\_comp\_2011\_refmen\_full)

# Composition, full, 2019, women as reference group

q\_comp\_2019\_women\_full <- bind\_cols(q\_comp\_2019\_women,

means.year %>% filter(female == 1, year == 2019) %>%

dplyr::select(full\_vars\_exc)) %>%

dplyr::select(full\_vars)

num\_comp\_2019\_refwomen\_full <- as.matrix(wcoef\_full1980) %\*% t(q\_comp\_2019\_women\_full)

denom\_comp\_2019\_refwomen\_full <- as.matrix(mcoef\_full1980) %\*%

t(means.year %>% filter(female == 0, year == 2019) %>% dplyr::select(full\_vars))

comp\_2019\_refwomen\_full <- exp(num\_comp\_2019\_refwomen\_full)/exp(denom\_comp\_2019\_refwomen\_full)

# Composition, full, 2019, men as reference group

q\_comp\_2019\_men\_full <- bind\_cols(q\_comp\_2019\_men,

means.year %>% filter(female == 0, year == 2019) %>%

dplyr::select(full\_vars\_exc)) %>%

dplyr::select(full\_vars)

num\_comp\_2019\_refmen\_full <- as.matrix(wcoef\_full1980) %\*%

t(means.year %>% filter(female == 1, year == 2019) %>% dplyr::select(full\_vars))

denom\_comp\_2019\_refmen\_full <- as.matrix(mcoef\_full1980) %\*% t(q\_comp\_2019\_men\_full)

comp\_2019\_refmen\_full <- exp(num\_comp\_2019\_refmen\_full)/exp(denom\_comp\_2019\_refmen\_full)

# For the convergence scenario, we once again allow non-family characteristics to change

# as observed. When women are the reference group, women's characteristics converge with

# men's characteristics as they did, but changes are anchored at men's 1980 levels:

# thus, when women are the reference group, men's family characteristics also stay at 1981 levels

# We first create the counterfactual characteristics for women

q\_conv\_1991\_women\_full <- (means.year %>%

filter(female == 0, year == 1981) %>%

dplyr::select(fam\_vars)+

(means.year %>%

filter(female == 1, year == 1991) %>% dplyr::select(fam\_vars) -

means.year %>%

filter(female == 0, year == 1991) %>% dplyr::select(fam\_vars))) %>%

bind\_cols(., means.year %>% filter(female == 1, year == 1991) %>%

dplyr::select(full\_vars\_exc)) %>%

dplyr::select(full\_vars)

# Then, we compute the counterfactual wage gap under this scenario

conv\_1991\_women\_full <- exp(as.matrix(wcoef\_full1980) %\*% t(q\_conv\_1991\_women\_full))/

exp(as.matrix(mcoef\_full1980) %\*% t(bind\_cols(means.year %>%

filter(female == 0, year == 1981) %>%

dplyr::select(fam\_vars),

means.year %>%filter(female == 0, year == 1991) %>%

dplyr::select(full\_vars\_exc)) %>%

dplyr::select(full\_vars)))

# Repeating the same process as above, first creating counterfactual characteristics for men

# (family traits converge with women's traits as observed, but are anchored at women's 1980 levels)

q\_conv\_1991\_men\_full <- (means.year %>%

filter(female == 1, year == 1981) %>%

dplyr::select(fam\_vars) +

(means.year %>%

filter(female == 0, year == 1991) %>%

dplyr::select(fam\_vars) -

means.year %>%

filter(female == 1, year == 1991) %>%

dplyr::select(fam\_vars))) %>%

bind\_cols(., means.year %>% filter(female == 0, year == 1991) %>%

dplyr::select(full\_vars\_exc)) %>%

dplyr::select(full\_vars)

# Then, we compute the counterfactual wage gap under this scenario

conv\_1991\_men\_full <- exp(as.matrix(wcoef\_full1980) %\*% t(bind\_cols(means.year %>%

filter(female == 1, year == 1981) %>%

dplyr::select(fam\_vars),

means.year %>%filter(female == 1, year == 1991) %>%

dplyr::select(full\_vars\_exc)) %>%

dplyr::select(full\_vars))) /

exp(as.matrix(mcoef\_full1980) %\*% t(q\_conv\_1991\_men\_full))

# Repeating the process for subsequent years

q\_conv\_2001\_women\_full <- (means.year %>%

filter(female == 0, year == 1981) %>%

dplyr::select(fam\_vars)+

(means.year %>%

filter(female == 1, year == 2001) %>% dplyr::select(fam\_vars) -

means.year %>%

filter(female == 0, year == 2001) %>% dplyr::select(fam\_vars))) %>%

bind\_cols(., means.year %>% filter(female == 1, year == 2001) %>%

dplyr::select(full\_vars\_exc)) %>%

dplyr::select(full\_vars)

conv\_2001\_women\_full <- exp(as.matrix(wcoef\_full1980) %\*% t(q\_conv\_2001\_women\_full))/

exp(as.matrix(mcoef\_full1980) %\*% t(bind\_cols(means.year %>%

filter(female == 0, year == 1981) %>%

dplyr::select(fam\_vars),

means.year %>%filter(female == 0, year == 2001) %>%

dplyr::select(full\_vars\_exc)) %>%

dplyr::select(full\_vars)))

q\_conv\_2001\_men\_full <- (means.year %>%

filter(female == 1, year == 1981) %>%

dplyr::select(fam\_vars) +

(means.year %>%

filter(female == 0, year == 2001) %>%

dplyr::select(fam\_vars) -

means.year %>%

filter(female == 1, year == 2001) %>%

dplyr::select(fam\_vars))) %>%

bind\_cols(., means.year %>% filter(female == 0, year == 2001) %>%

dplyr::select(full\_vars\_exc)) %>%

dplyr::select(full\_vars)

conv\_2001\_men\_full <- exp(as.matrix(wcoef\_full1980) %\*% t(bind\_cols(means.year %>%

filter(female == 1, year == 1981) %>%

dplyr::select(fam\_vars),

means.year %>%filter(female == 1, year == 2001) %>%

dplyr::select(full\_vars\_exc)) %>%

dplyr::select(full\_vars))) /

exp(as.matrix(mcoef\_full1980) %\*% t(q\_conv\_2001\_men\_full))

q\_conv\_2011\_women\_full <- (means.year %>%

filter(female == 0, year == 1981) %>%

dplyr::select(fam\_vars)+

(means.year %>%

filter(female == 1, year == 2011) %>% dplyr::select(fam\_vars) -

means.year %>%

filter(female == 0, year == 2011) %>% dplyr::select(fam\_vars))) %>%

bind\_cols(., means.year %>% filter(female == 1, year == 2011) %>%

dplyr::select(full\_vars\_exc)) %>%

dplyr::select(full\_vars)

conv\_2011\_women\_full <- exp(as.matrix(wcoef\_full1980) %\*% t(q\_conv\_2011\_women\_full))/

exp(as.matrix(mcoef\_full1980) %\*% t(bind\_cols(means.year %>%

filter(female == 0, year == 1981) %>%

dplyr::select(fam\_vars),

means.year %>%filter(female == 0, year == 2011) %>%

dplyr::select(full\_vars\_exc)) %>%

dplyr::select(full\_vars)))

q\_conv\_2011\_men\_full <- (means.year %>%

filter(female == 1, year == 1981) %>%

dplyr::select(fam\_vars) +

(means.year %>%

filter(female == 0, year == 2011) %>%

dplyr::select(fam\_vars) -

means.year %>%

filter(female == 1, year == 2011) %>%

dplyr::select(fam\_vars))) %>%

bind\_cols(., means.year %>% filter(female == 0, year == 2011) %>%

dplyr::select(full\_vars\_exc)) %>%

dplyr::select(full\_vars)

conv\_2011\_men\_full <- exp(as.matrix(wcoef\_full1980) %\*% t(bind\_cols(means.year %>%

filter(female == 1, year == 1981) %>%

dplyr::select(fam\_vars),

means.year %>%filter(female == 1, year == 2011) %>%

dplyr::select(full\_vars\_exc)) %>%

dplyr::select(full\_vars))) /

exp(as.matrix(mcoef\_full1980) %\*% t(q\_conv\_2011\_men\_full))

q\_conv\_2019\_women\_full <- (means.year %>%

filter(female == 0, year == 1981) %>%

dplyr::select(fam\_vars)+

(means.year %>%

filter(female == 1, year == 2019) %>% dplyr::select(fam\_vars) -

means.year %>%

filter(female == 0, year == 2019) %>% dplyr::select(fam\_vars))) %>%

bind\_cols(., means.year %>% filter(female == 1, year == 2019) %>%

dplyr::select(full\_vars\_exc)) %>%

dplyr::select(full\_vars)

conv\_2019\_women\_full <- exp(as.matrix(wcoef\_full1980) %\*% t(q\_conv\_2019\_women\_full))/

exp(as.matrix(mcoef\_full1980) %\*% t(bind\_cols(means.year %>%

filter(female == 0, year == 1981) %>%

dplyr::select(fam\_vars),

means.year %>%filter(female == 0, year == 2019) %>%

dplyr::select(full\_vars\_exc)) %>%

dplyr::select(full\_vars)))

q\_conv\_2019\_men\_full <- (means.year %>%

filter(female == 1, year == 1981) %>%

dplyr::select(fam\_vars) +

(means.year %>%

filter(female == 0, year == 2019) %>%

dplyr::select(fam\_vars) -

means.year %>%

filter(female == 1, year == 2019) %>%

dplyr::select(fam\_vars))) %>%

bind\_cols(., means.year %>% filter(female == 0, year == 2019) %>%

dplyr::select(full\_vars\_exc)) %>%

dplyr::select(full\_vars)

conv\_2019\_men\_full <- exp(as.matrix(wcoef\_full1980) %\*% t(bind\_cols(means.year %>%

filter(female == 1, year == 1981) %>%

dplyr::select(fam\_vars),

means.year %>%filter(female == 1, year == 2019) %>%

dplyr::select(full\_vars\_exc)) %>%

dplyr::select(full\_vars))) /

exp(as.matrix(mcoef\_full1980) %\*% t(q\_conv\_2019\_men\_full))

# We create an additional counterfactual scenario for comparison to the full model, where

# family traits stay at their 1980 levels for both groups but non-family traits change as observed

nofamchange\_1991 <- exp(as.matrix(wcoef\_full1980) %\*% t(bind\_cols(means.year %>%

filter(female == 1, year == 1981) %>%

dplyr::select(fam\_vars),

means.year %>%filter(female == 1, year == 1991) %>%

dplyr::select(full\_vars\_exc)) %>%

dplyr::select(full\_vars))) /

exp(as.matrix(mcoef\_full1980) %\*% t(bind\_cols(means.year %>%

filter(female == 0, year == 1981) %>%

dplyr::select(fam\_vars),

means.year %>%filter(female == 0, year == 1991) %>%

dplyr::select(full\_vars\_exc)) %>%

dplyr::select(full\_vars)))

nofamchange\_2001 <- exp(as.matrix(wcoef\_full1980) %\*% t(bind\_cols(means.year %>%

filter(female == 1, year == 1981) %>%

dplyr::select(fam\_vars),

means.year %>%filter(female == 1, year == 2001) %>%

dplyr::select(full\_vars\_exc)) %>%

dplyr::select(full\_vars))) /

exp(as.matrix(mcoef\_full1980) %\*% t(bind\_cols(means.year %>%

filter(female == 0, year == 1981) %>%

dplyr::select(fam\_vars),

means.year %>%filter(female == 0, year == 2001) %>%

dplyr::select(full\_vars\_exc)) %>%

dplyr::select(full\_vars)))

nofamchange\_2011 <- exp(as.matrix(wcoef\_full1980) %\*% t(bind\_cols(means.year %>%

filter(female == 1, year == 1981) %>%

dplyr::select(fam\_vars),

means.year %>%filter(female == 1, year == 2011) %>%

dplyr::select(full\_vars\_exc)) %>%

dplyr::select(full\_vars))) /

exp(as.matrix(mcoef\_full1980) %\*% t(bind\_cols(means.year %>%

filter(female == 0, year == 1981) %>%

dplyr::select(fam\_vars),

means.year %>%filter(female == 0, year == 2011) %>%

dplyr::select(full\_vars\_exc)) %>%

dplyr::select(full\_vars)))

nofamchange\_2019 <- exp(as.matrix(wcoef\_full1980) %\*% t(bind\_cols(means.year %>%

filter(female == 1, year == 1981) %>%

dplyr::select(fam\_vars),

means.year %>%filter(female == 1, year == 2019) %>%

dplyr::select(full\_vars\_exc)) %>%

dplyr::select(full\_vars))) /

exp(as.matrix(mcoef\_full1980) %\*% t(bind\_cols(means.year %>%

filter(female == 0, year == 1981) %>%

dplyr::select(fam\_vars),

means.year %>%filter(female == 0, year == 2019) %>%

dplyr::select(full\_vars\_exc)) %>%

dplyr::select(full\_vars)))

# Combining the observed wage gap into one vector

observed <- data.frame(year = c(1980, 1990, 2000, 2010, 2019),

Observed = c(observed\_gap1980, observed\_gap1991, observed\_gap2001,

observed\_gap2011, observed\_gap2019))

# Combining the wage gap predicted under each model into one vector

predicted <- data.frame(year = c(1980, 1990, 2000, 2010, 2019),

Family = c(fam1980[1], fam1991[1], fam2001[1], fam2011[1], fam2019[1]),

Full = c(full1980[1], full1991[1], full2001[1], full2011[1], full2019[1]))

# Combining the wage gap predicted under the scenario of no family change into one vector

nofamchange <- data.frame(year = c(1980, 1990, 2000, 2010, 2019),

Family = rep(NA, 5),

Full = c(full1980[1], nofamchange\_1991[1], nofamchange\_2001[1], nofamchange\_2011[1], nofamchange\_2019[1]))

# We then combine these elements into a single dataframe

fig4\_data <- data.frame(year = c(1980, 1990, 2000, 2010, 2019),

"Convergence\_Women.Family" = c(

fam1980[1], conv\_1991\_women[1], conv\_2001\_women[1],

conv\_2011\_women[1], conv\_2019\_women[1]),

"Convergence\_Men.Family" = c(

fam1980[1], conv\_1991\_men[1], conv\_2001\_men[1],

conv\_2011\_men[1], conv\_2019\_men[1]),

"Convergence\_Women.Full" = c(

full1980[1], conv\_1991\_women\_full[1], conv\_2001\_women\_full[1],

conv\_2011\_women\_full[1], conv\_2019\_women\_full[1]),

"Convergence\_Men.Full" = c(

full1980[1], conv\_1991\_men\_full[1], conv\_2001\_men\_full[1],

conv\_2011\_men\_full[1], conv\_2019\_men\_full[1]),

"Composition\_Women.Family" = c(

fam1980[1], comp\_1991\_refwomen[1], comp\_2001\_refwomen[1],

comp\_2011\_refwomen[1], comp\_2019\_refwomen[1]),

"Composition\_Men.Family" = c(

fam1980[1], comp\_1991\_refmen[1], comp\_2001\_refmen[1],

comp\_2011\_refmen[1], comp\_2019\_refmen[1]),

"Composition\_Women.Full" = c(

fam1980[1], comp\_1991\_refwomen\_full[1], comp\_2001\_refwomen\_full[1],

comp\_2011\_refwomen\_full[1], comp\_2019\_refwomen\_full[1]),

"Composition\_Men.Full" = c(

fam1980[1], comp\_1991\_refmen\_full[1], comp\_2001\_refmen\_full[1],

comp\_2011\_refmen\_full[1], comp\_2019\_refmen\_full[1])) %>%

gather(model, value, -year) %>%

mutate(pathway = ifelse(grepl("Convergence", model), "Convergence", "Composition"),

reference = ifelse(grepl("Women", model), "Women", "Men"),

model = ifelse(grepl("Family", model), "Family", "Full")) %>%

left\_join(., predicted %>% gather(model, predicted, - c(year))) %>%

left\_join(., nofamchange %>% gather(model, nofamchange, - c(year))) %>%

left\_join(., observed) %>%

gather(linetype, value, -c(year, model, pathway, reference)) %>%

mutate(linetype = case\_when(linetype == "Observed" ~ "Observed",

linetype == "predicted" ~ "Covariate Change",

linetype == "value" ~ "Hypothetical",

linetype == "nofamchange" ~ "No Family Change"),

linetype = ifelse(linetype == "Hypothetical", paste(linetype, pathway, sep = ", "), linetype))

# We then create figures for these counterfactual trajectories for men and women

fig4\_women <- fig4\_data %>%

filter(reference == "Women") %>%

distinct() %>%

ggplot(aes(y = value, x = year, linetype = linetype, color = linetype, shape = linetype)) +

geom\_line() +

facet\_wrap(~model) +

theme\_bw() +

geom\_point() +

scale\_linetype\_manual(values=c(5, 4, 3, 2, 1)) +

scale\_color\_manual(values = c("grey31", "snow4", "grey55", "grey69", "grey1")) +

scale\_shape\_manual(values=c(15, 17, 3, 8, 19)) +

labs(title = "(Reference: Women)",

y = "Gender Pay Gap",

x = "") +

geom\_point() +

geom\_line() +

theme(plot.title = element\_text(hjust = 0.5), legend.position = "bottom") +

scale\_x\_continuous(breaks=c(1980, 1990, 2000, 2010, 2018)) +

ylim(.5, 1)

fig4\_men <- fig4\_data %>%

filter(reference == "Men") %>%

distinct() %>%

ggplot(aes(y = value, x = year, linetype = linetype, color = linetype, shape = linetype)) +

geom\_line() +

facet\_wrap(~model) +

theme\_bw() +

geom\_point() +

scale\_linetype\_manual(values=c(5, 4, 3, 2, 1)) +

scale\_color\_manual(values = c("grey31", "snow4", "grey55", "grey69", "grey1")) +

scale\_shape\_manual(values=c(15, 17, 3, 8, 19)) +

labs(title = "(Reference: Men)",

y = "Gender Pay Gap",

x = "") +

geom\_point() +

geom\_line() +

theme(plot.title = element\_text(hjust = 0.5), legend.position = "bottom") +

scale\_x\_continuous(breaks=c(1980, 1990, 2000, 2010, 2018)) +

ylim(.5, 1)

# Combining these figures

fig4 <- grid.arrange(fig4\_women, fig4\_men, ncol = 1,

top = "Changing Pay Gap under Different Counterfactual Scenarios")

# Saving figure

ggsave(plot = fig4, "/Users/ninocricco/Desktop/fig4.jpg",

width = 10, height = 10, units = "in", device='jpeg', dpi=700)

# Outputting figure values

View(fig4\_data %>%

pivot\_wider(names\_from = c(linetype, pathway), values\_from = value) %>%

dplyr::select(-c("Observed\_Composition", "Covariate Change\_Composition", "No Family Change\_Composition")) %>%

rename("Observed" = "Observed\_Convergence",

"Covariate Change" = "Covariate Change\_Convergence",

"No Family Change" = "No Family Change\_Convergence"))

# We also create an appendix Table to show the counterfactual means in each scenario

app\_table\_q <- bind\_rows(

q\_conv\_1991\_men\_full %>% mutate(year = 1991, ref = "men", q = "convergence"),

q\_conv\_2001\_men\_full %>% mutate(year = 2001, ref = "men", q = "convergence"),

q\_conv\_2011\_men\_full %>% mutate(year = 2011, ref = "men", q = "convergence"),

q\_conv\_2019\_men\_full %>% mutate(year = 2019, ref = "men", q = "convergence"),

q\_conv\_1991\_women\_full %>% mutate(year = 1991, ref = "women", q = "convergence"),

q\_conv\_2001\_women\_full %>% mutate(year = 2001, ref = "women", q = "convergence"),

q\_conv\_2011\_women\_full %>% mutate(year = 2011, ref = "women", q = "convergence"),

q\_conv\_2019\_women\_full %>% mutate(year = 2019, ref = "women", q = "convergence"),

q\_comp\_1991\_men\_full %>% mutate(year = 1991, ref = "men", q = "composition"),

q\_comp\_2001\_men\_full %>% mutate(year = 2001, ref = "men", q = "composition"),

q\_comp\_2011\_men\_full %>% mutate(year = 2011, ref = "men", q = "composition"),

q\_comp\_2019\_men\_full %>% mutate(year = 2019, ref = "men", q = "composition"),

q\_comp\_1991\_women\_full %>% mutate(year = 1991, ref = "women", q = "composition"),

q\_comp\_2001\_women\_full %>% mutate(year = 2001, ref = "women", q = "composition"),

q\_comp\_2011\_women\_full %>% mutate(year = 2011, ref = "women", q = "composition"),

q\_comp\_2019\_women\_full %>% mutate(year = 2019, ref = "women", q = "composition")) %>%

dplyr::select(year, ref, q, everything()) %>%

gather(key, value, -c(year, ref, q)) %>%

filter(key != "(Intercept)") %>%

pivot\_wider(names\_from = c(year, ref, q), values\_from = value) %>%

dplyr::select(-key)

colnames(app\_table\_q) <- (c("1991", "2001", "2011", "2019",

"1991 ", "2001 ", "2011 ", "2019 ",

"1991 ", "2001 ", "2011 ", "2019 ",

"1991 ", "2001 ", "2011 ", "2019 "))

rownames(app\_table\_q) <- c("Northeast", "Northcentral",

"South", "Black", "Latinx", "Other", "Married",

"Previously Married", "Housework", "Number of Kids: 3+",

"<= 21", "22 to 27", "28 +",

"High School", "Some College", "BA/Advanced Degree",

"Unionized Job", "Government Job",

"Log, Full-Time Experience", #"Part-Time",

#"35 to 39", "40 to 44", "45 to 59",

"Overwork",

#"Less than 50", "50+",

"Employer Tenure",

#"Percent in Male Occupations",

#"Percent in Female Occupations", "Percent in Integrated Occupations",

"Percent Female in Occupation", "Management Occupation", "Manufacturing")

knitr::kable(app\_table\_q %>% mutate\_if(is.numeric, round, digits = 2),

booktabs = T, format = "latex",

caption = "Counterfactual Characteristics Change by Decade and Pathway") %>%

add\_header\_above(c(" ", "Male" = 4, "Female" = 4, "Male" = 4, "Female" = 4)) %>%

add\_header\_above(c(" ", "Convergence" = 8, "Composition" = 8)) %>%

pack\_rows("Region", 1, 3, bold = F) %>% pack\_rows("Race", 4, 6, bold = F) %>%

pack\_rows("Marital Status", 7, 8, bold = F) %>% pack\_rows("Age at First Birth", 11, 13, bold = F) %>%

pack\_rows("Education", 14, 16, bold = F)