sociological science

Supplement to:

Gowen, Ohjae, Ethan Fosse, and Christopher Winship. 2023. "Cross-Group Differences in Age, Period, and Cohort Effects: A Bounding Approach to the Gender Wage Gap." Sociological Science 10: 731-768.

Online Supplement to:

Gowen, Ohjae, Ethan Fosse, and Christopher Winship. "Cross-Group Differences in Age, Period, and Cohort Effects: A Bounding Approach to the Gender Wage Gap."

Appendix A: Details on the CSL-APC Model

The presentation of the models in the main text are quite general in that we do not specify exactly how the linear and nonlinear components have been constructed. Following previous research (e.g., Holford 1983; Fosse and Winship 2019), we will use orthogonal polynomials such that, for example, a_L denotes the linear component, a_2 denotes the quadratic component, a_3 the cubic component, and so forth.[1] This implies the following re-expression of the L-APC model:

$$\begin{aligned} Y_{ijk} &= \mu + \alpha(i - i^*) + \pi(j - j^*) + \gamma(k - k^*) + \widetilde{\alpha}_i + \widetilde{\pi}_j + \widetilde{\gamma}_k \\ &= \mu + \alpha a_L + \pi p_L + \gamma c_L + \sum_{i+1}^{I-1} \alpha_i a_i + \sum_{j+1}^{J-1} \pi_i p_j + \sum_{k+1}^{K-1} \gamma_k c_k + \eta_{ijk}, \end{aligned} \tag{1}$$

where a_L, p_L, and c_L are the age, period, and cohort linear components with corresponding linear effects α, π, and γ; $a_2, \ldots, a_{I-1}$ are the age nonlinear components with corresponding nonlinear effects $\alpha_2, \ldots, \alpha_{I-1}$; $p_2, \ldots, p_{J-1}$ are the period nonlinear components with corresponding nonlinear effects $\pi_2, \ldots, \pi_{J-1}$; $c_2, \ldots, c_{K-1}$ are the cohort nonlinear components with corresponding nonlinear effects $\gamma_2, \ldots, \gamma_{K-1}$; and η_{ijk} denote the cell-specific error terms.

Likewise, the CSL-APC model is also quite general and there are various ways of parameterizing cross-strata differences in APC effects. While one can allow parameters to vary across any number of levels of a strata variable in principle, we have two levels for gender in our case. Let an indicator (dummy) gender variable G coded as $g = 1$ for women and $g = 0$ for men. Including and interacting this strata variable with the linear and nonlinear components for age, period, and cohort results in the following model, which is analogous to the SL-APC model in Equation 3 in the main text:[2]

$$\begin{aligned} Y_{ijk} = \mu + \alpha a_L + \pi p_L + \gamma c_L + \sum_{i+1}^{I-1} \alpha_i a_i + \sum_{j+1}^{J-1} \pi_i p_j + \sum_{k+1}^{K-1} \gamma_k c_k + \mu_G G + \\ \alpha_G(a_L G) + \pi_G(p_L G) + \gamma_G(c_L G) + \sum_{i+1}^{I-1} \alpha_{Gi}(a_i G) + \sum_{j+1}^{J-1} \pi_{Gj}(p_j G) + \sum_{k+1}^{K-1} \gamma_{Gk}(c_k G) + \eta_{ijk}, \end{aligned} \tag{2}$$

[1] Note that the identity in the below equation may be only approximate depending on how the design matrix is constructed. In our analyses we use QR decomposition to construct the nonlinear components. Because the elements of the design matrix can still be quite large, for the purposes of numerical stability we include the additional step of norming each of the columns of the design matrix representing the nonlinear components. We use a weighted version of orthogonal polynomials to make sure polynomial terms are perpendicular to lower-order terms in our empirical data (see Elbers 2020).

[2] If we had more than two strata levels, then we would specify an expanded set of interaction terms on the right-hand side of the equation with group indicator variables G_1, G_2, and so on.

where α_G, π_G, and γ_G are interaction effects between the gender variable and the age, period, and cohort linear components, respectively; α_{Gi}, π_{Gj}, and γ_{Gk} are interaction effects between the gender variable and the nonlinear components for age, period, and cohort, respectively.

There are two related ways of interpreting the interaction terms in Equation 2. On the one hand, the interaction effects can be interpreted as representing differences in age, period, and cohort effects between the strata. For example, α_G can be interpreted as the difference between the age linear effect for women and the age linear effect for men. On the other hand, the interaction effects might also be interpreted as representing the cross-strata outcome disparity for varying values of age, period, and cohort. For example, the parameters for α_G could be interpreted as the gender "effect" on the outcome (i.e., between-gender wage disparity) for varying levels of age. However, in general we focus on interpreting the parameters as cross-strata differences in age, period, and cohort effects, because such an interpretation more closely aligns with our view that age, period, and cohort are observed proxies for underlying latent causal factors that are not bound by the natural relationship period = age + cohort.[3]

We are now ready to show how the CSL-APC model is derived from a variant of the model in Equation 2, or the SL-APC model in Equation 3 in the main text. After substituting for $G = 1$ and $G = 0$, we can express the CSL-APC model as follows:

$$
\begin{aligned}
&Y_{ijk[G=1]} - Y_{ijk[G=0]} = \\
&\Big(\mu_G \times 1 + \alpha_G(a_L \times 1) + \pi_G(p_L \times 1) + \gamma_G(c_L \times 1) + \sum_{i+1}^{I-1} \alpha_{Gi}(a_i \times 1) + \sum_{j+1}^{J-1} \pi_{Gj}(p_j \times 1) + \sum_{k+1}^{K-1} \gamma_{Gk}(c_k \times 1) + \epsilon_{ijk[G=1]}\Big) \\
&\quad - \Big(\mu_G \times 0 + \alpha_G(a_L \times 0) + \pi_G(p_L \times 0) + \gamma_G(c_L \times 0) + \sum_{i+1}^{I-1} \alpha_{Gi}(a_i \times 0) + \sum_{j+1}^{J-1} \pi_{Gj}(p_j \times 0) + \sum_{k+1}^{K-1} \gamma_{Gk}(c_k \times 0) + \epsilon_{ijk[G=0]}\Big) \\
&\quad = \mu_G + \alpha_G a_L + \pi_G p_L + \gamma_G c_L + \sum_{i+1}^{I-1} \alpha_{Gi} a_i + \sum_{j+1}^{J-1} \pi_{Gj} p_j + \sum_{k+1}^{K-1} \gamma_{Gk} c_k + (\eta_{ijk[G=1]} - \eta_{ijk[G=0]}),
\end{aligned}
$$

or, in a more compact general form:

$$
\Delta Y_{ijk} = \Delta\mu + \Delta\alpha(i - i^*) + \Delta\pi(j - j^*) + \Delta\gamma(k - k^*) + \Delta\widetilde{\alpha}_i + \Delta\widetilde{\pi}_j + \Delta\widetilde{\gamma}_k + \Delta\eta_{ijk}, \quad (3)
$$

which is equivalent to Equation 4 in the main text.[4]

[3]Our case is a specific example of the general conceptual issue that appears when interpreting interaction effects. Suppose there is an interactive effect of a continuous variable *X* and a binary group indicator *G* on an outcome *Y*, namely, $Y = \mu + \beta_1 X + \beta_2 G + \beta_3(X \times G) + \epsilon$. One interpretation of β_3, which aligns with the first interpretation above, is that the effect of *X* is different between the two groups indicated by *G*. Another interpretation, more consistent with the second interpretation above, is that the between-group difference in the outcome indicated by *G* varies depending on the level of *X* (see Fox 2016:140-150)

[4]The estimated nonlinear effects for each age, period, and cohort category as presented in Figure 6 are predicted values based on the parameter estimates for an intercept and orthogonal polynomials.

Appendix B: Lexis Table for Cross-Group APC Analysis

Table B.1: Lexis Table of the Gender Wage Gap

	Period								
Age	1975-79	1980-84	1985-89	1990-94	1995-99	2000-04	2005-09	2010-14	2015-19
25-29	−0.35	−0.30	−0.23	−0.15	−0.14	−0.09	−0.10	−0.09	−0.12
	(0.004)	(0.009)	(0.006)	(0.008)	(0.009)	(0.009)	(0.009)	(0.006)	(0.011)
30-34	−0.49	−0.39	−0.31	−0.26	−0.24	−0.20	−0.16	−0.15	−0.20
	(0.005)	(0.007)	(0.008)	(0.009)	(0.005)	(0.005)	(0.006)	(0.010)	(0.007)
35-39	−0.59	−0.50	−0.44	−0.34	−0.34	−0.29	−0.27	−0.22	−0.19
	(0.009)	(0.009)	(0.008)	(0.010)	(0.009)	(0.005)	(0.011)	(0.007)	(0.007)
40-44	−0.61	−0.58	−0.49	−0.39	−0.36	−0.32	−0.29	−0.26	−0.25
	(0.013)	(0.006)	(0.010)	(0.002)	(0.011)	(0.009)	(0.007)	(0.007)	(0.010)
45-49	−0.63	−0.60	−0.53	−0.45	−0.37	−0.35	−0.33	−0.32	−0.28
	(0.013)	(0.010)	(0.009)	(0.009)	(0.013)	(0.010)	(0.010)	(0.010)	(0.010)
50-54	−0.60	−0.58	−0.53	−0.48	−0.44	−0.36	−0.28	−0.33	−0.27
	(0.011)	(0.014)	(0.015)	(0.012)	(0.011)	(0.012)	(0.013)	(0.010)	(0.009)
55-59	−0.57	−0.61	−0.56	−0.49	−0.44	−0.40	−0.31	−0.30	−0.27
	(0.015)	(0.013)	(0.012)	(0.016)	(0.012)	(0.010)	(0.011)	(0.014)	(0.012)
60-64	−0.52	−0.51	−0.52	−0.44	−0.42	−0.36	−0.39	−0.30	−0.26
	(0.009)	(0.023)	(0.012)	(0.023)	(0.019)	(0.019)	(0.012)	(0.014)	(0.012)

Notes: The rows indicate age categories, and the columns indicate period categories. The input in each cell denotes the gender difference (female−male) in log median annual earnings in the respective age-period category. Standard errors for each difference are presented in parentheses.

In presenting our CSL-APC model, we have assumed that we have only aggregate data (see note 8 in the main text). However, researchers may have individual-level sample data to construct a Lexis table of cross-strata differences, as we do in our empirical example using the CPS data. In these cases, the Lexis table must be estimated from the individual-level sample data. To do this, we ran a regression of log earnings on age, period, gender, and all two-way and three-way interactions between these variables. Because we are modeling a *median* difference between men's and women's earnings conditional on age and period, we relied on a (conditional) quantile regression at the median. The CPS ASEC survey sampling weights are applied to the regression so that the estimated median differences are representative of those in the population. Predicted marginal "effects" of gender, which are allowed to vary by age and period, can be calculated from the fitted median regression and are used as cell values in the Lexis table above. If researchers are interested in modeling a cross-strata difference in *mean* values, they can instead conduct an OLS regression to estimate the corresponding Lexis table.

A potential advantage of using individual-level data as opposed to aggregated data is that researchers may have a better understanding of the precision with which each cell value is estimated. For example, the gender-specific variance of log earnings, which is unlikely to be available in an aggregated data set, affects the precision of the estimated cell values. Accounting for differential precision can provide an efficiency gain over OLS, which assumes constant cell variance, particularly when using a small sample such as a set of summaries from a Lexis table (N=72). We therefore include the estimated standard errors of the cell values (shown in parentheses in Table B.1 above) when fitting a weighted least squares regression of the CSL-APC model. The weight of each cell is calculated as $1/se^2$. However, the point estimates are very similar to the OLS estimates of the CSL-APC model, and the substantive conclusions of the bounding analysis remain the same.

Appendix C: Additional Tables

Table C.1: Summary Statistics of the CPS Sample

	Men				Women			
Variable	Mean	Std. Dev.	Min.	Max.	Mean	Std. Dev.	Min.	Max.
Age	41.7	10.3	25	64	41.88	10.4	25	64
Period	1999.3	12.5	1976	2019	2000.8	12.0	1976	2019
Cohort	1957.7	15.5	1912	1994	1958.9	15.0	1912	1994
Earnings	69,437.2	60,473.6	1.6	1,958,398.9	48,314.5	40,898.2	1.3	1,401,398.9
Obs.	1,121,562				830,856			

Notes: Mean values and their standard deviations are computed for the weighted sample using the CPS ASEC survey sampling weights. Earnings refer to respondents' annual earnings in the last calendar year (in 2018 dollars).

Table C.2: Bounding Formulas for Cross-Strata Differences in APC Slopes

Age Bounds:	$\alpha_{\min} \leq \Delta\alpha \leq \alpha_{\max}$
	$\theta_1 - \alpha_{\max} \leq \Delta\pi \leq \theta_1 - \alpha_{\min}$
	$(\theta_2 - \theta_1) + \alpha_{\min} \leq \Delta\gamma \leq (\theta_2 - \theta_1) + \alpha_{\max}$
Period Bounds:	$\theta_1 - \pi_{\max} \leq \Delta\alpha \leq \theta_1 - \pi_{\min}$
	$\pi_{\min} \leq \Delta\pi \leq \pi_{\max}$
	$\theta_2 - \pi_{\max} \leq \Delta\gamma < \theta_2 - \pi_{\min}$
Cohort Bounds:	$(\theta_1 - \theta_2) + \gamma_{\min} \leq \Delta\alpha \leq (\theta_1 - \theta_2) + \gamma_{\max}$
	$\theta_2 - \gamma_{\max} \leq \Delta\pi \leq \theta_2 - \gamma_{\min}$
	$\gamma_{\min} \leq \Delta\gamma \leq \gamma_{\max}$

Notes: Age, period, and cohort slopes are α, π, and γ, respectively, with $(.)_{\min}$ and $(.)_{\max}$ denoting minimum and maximum values imposed by theoretical assumptions. We denote $\theta_1 = \Delta\alpha + \Delta\pi$, $\theta_2 = \Delta\gamma + \Delta\pi$, $\theta_1 - \theta_2 = \Delta\alpha - \Delta\gamma$, and $\theta_2 - \theta_1 = \Delta\gamma - \Delta\alpha$.

Table C.3: Estimated CSL-APC Parameters on the Gender Wage Gap

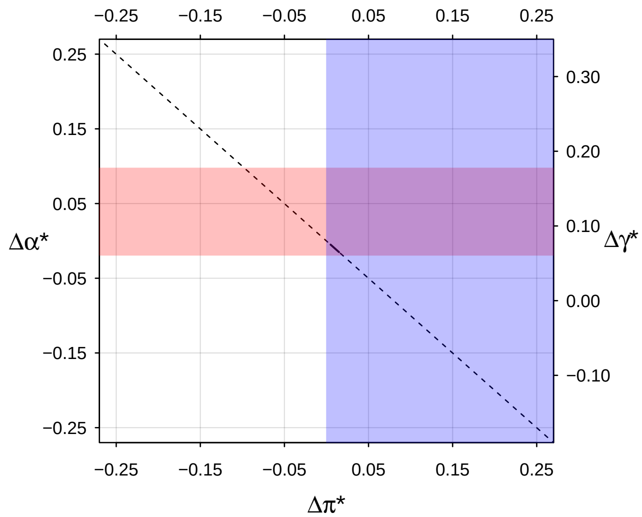
Parameter	Coef.	Std. Error	95% CI Lower	95% CI Upper
$\Delta\mu$	−0.339	0.002	−0.342	−0.335
$\Delta\theta_1$	−0.002	0.002	−0.007	0.002
$\Delta\theta_2$	0.084	0.001	0.081	0.086
$\Delta\alpha_2$	0.152	0.007	0.138	0.166
$\Delta\alpha_3$	−0.026	0.006	−0.039	−0.013
$\Delta\alpha_4$	−0.006	0.006	−0.019	0.007
$\Delta\alpha_5$	0.011	0.006	−0.001	0.023
$\Delta\alpha_6$	−0.003	0.006	−0.015	0.009
$\Delta\alpha_7$	0.009	0.006	−0.003	0.021
$\Delta\pi_2$	−0.063	0.006	−0.075	−0.052
$\Delta\pi_3$	0.009	0.006	−0.003	0.020
$\Delta\pi_4$	0.014	0.006	0.002	0.025
$\Delta\pi_5$	−0.014	0.006	−0.025	−0.002
$\Delta\pi_6$	0.003	0.006	−0.008	0.014
$\Delta\pi_7$	0.008	0.006	−0.004	0.019
$\Delta\pi_8$	−0.011	0.006	−0.023	0.000
$\Delta\gamma_2$	−0.038	0.015	−0.069	−0.007
$\Delta\gamma_3$	−0.150	0.014	−0.178	−0.123
$\Delta\gamma_4$	0.046	0.014	0.018	0.074
$\Delta\gamma_5$	−0.031	0.014	−0.059	−0.004
$\Delta\gamma_6$	−0.046	0.013	−0.073	−0.018
$\Delta\gamma_7$	0.031	0.013	0.005	0.058
$\Delta\gamma_8$	0.049	0.014	0.022	0.077
$\Delta\gamma_9$	0.012	0.013	−0.014	0.038
$\Delta\gamma_{10}$	−0.020	0.012	−0.044	0.003
$\Delta\gamma_{11}$	0.013	0.010	−0.007	0.032
$\Delta\gamma_{12}$	0.012	0.009	−0.006	0.030
$\Delta\gamma_{13}$	−0.006	0.008	−0.023	0.010
$\Delta\gamma_{14}$	−0.009	0.008	−0.024	0.007
$\Delta\gamma_{15}$	−0.003	0.006	−0.016	0.010
Adj. R-squared	0.98			
Number of Cells	72			

Notes: The nonlinearity parameters are estimated for different degrees of orthogonal polynomials, as indicated by their subscripts (see Appendix A and Appendix B for more details on the estimation process). The estimated nonlinear effects shown in Figure 6 are predicted values based on the above nonlinearity and intercept estimates.

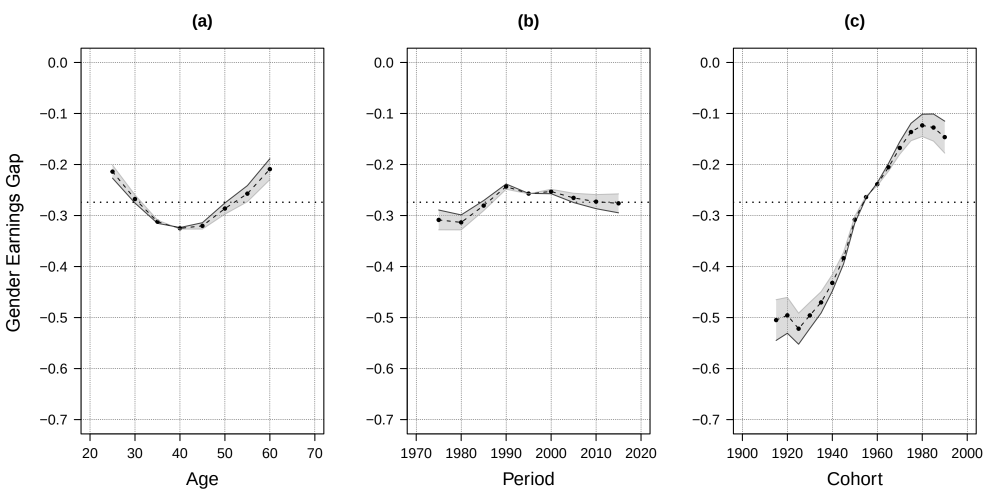
Appendix D: Additional Figures

Figure D.1: Bounding Analysis Results of Gender Differences in Median Hourly Wages

I. Upper and Lower Bounds of Cross-Strata APC Linear Effects on the Gender Wage Gap



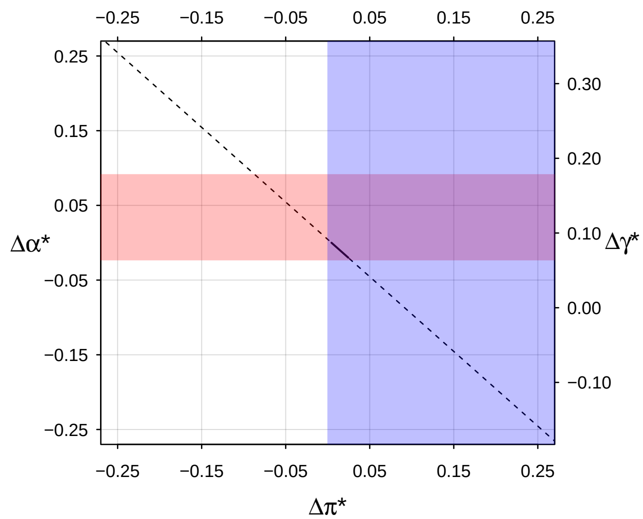
II. Bounded Cross-Strata Age, Period, Cohort Effects on the Gender Wage Gap



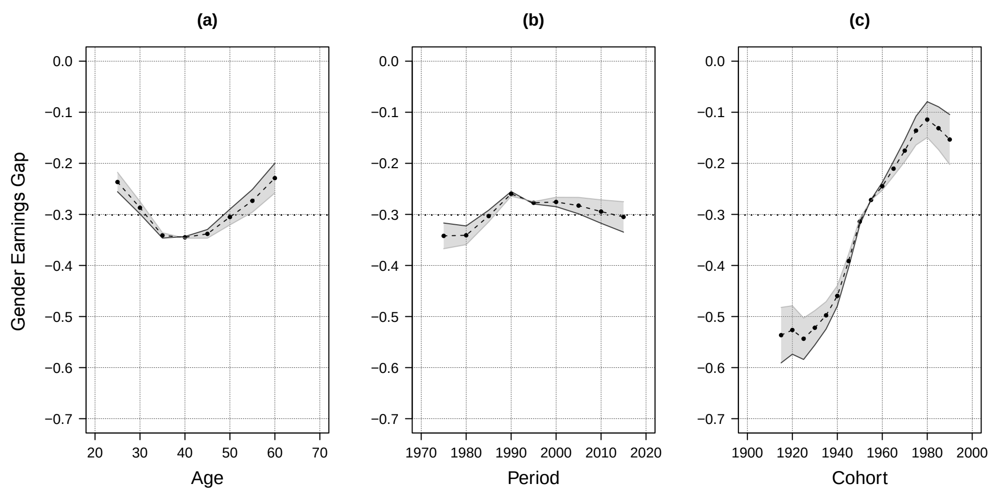
Notes: Hourly wages are analyzed instead of annual earnings. In the top panel (I.), the left y-axis indicates a range of cross-strata age linear effects, the x-axis indicates a range of cross-strata period linear effects, and the right y-axis indicates a range of cross-strata cohort linear effects. The dashed line indicates all possible cross-strata linear effects consistent with the data. The shaded red region denotes the set of values consistent with the age-related assumptions, while the shaded blue region indicates the set of values consistent with the period-related assumptions. The solid line in the overlapping shaded regions refers to the feasible set of cross-strata linear effects given that the assumptions about the cross-strata age and period effects are satisfied. In the bottom panel (II.), the shaded areas represent the bounded cross-strata effects of age (a), period (b), and cohort (c) on the gender wage gap based on the three assumptions about the age and period effects. The dotted lines follow the midpoints in each shaded area. The dark bold lines along one edge of the shaded areas indicate when the cross-strata age linear effect is most positive, the cross-strata period linear effect is most negative, and the cross-strata cohort linear effect is most positive within the bounds.

Figure D.2: Bounding Analysis Results of Gender Differences in Median Hourly Wages, Including Half-Year-Round Workers

I. Upper and Lower Bounds of Cross-Strata APC Linear Effects on the Gender Wage Gap



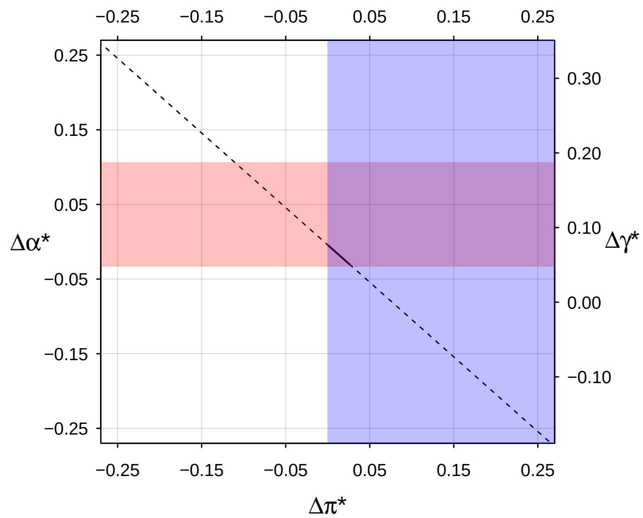
II. Bounded Cross-Strata Age, Period, Cohort Effects on the Gender Wage Gap



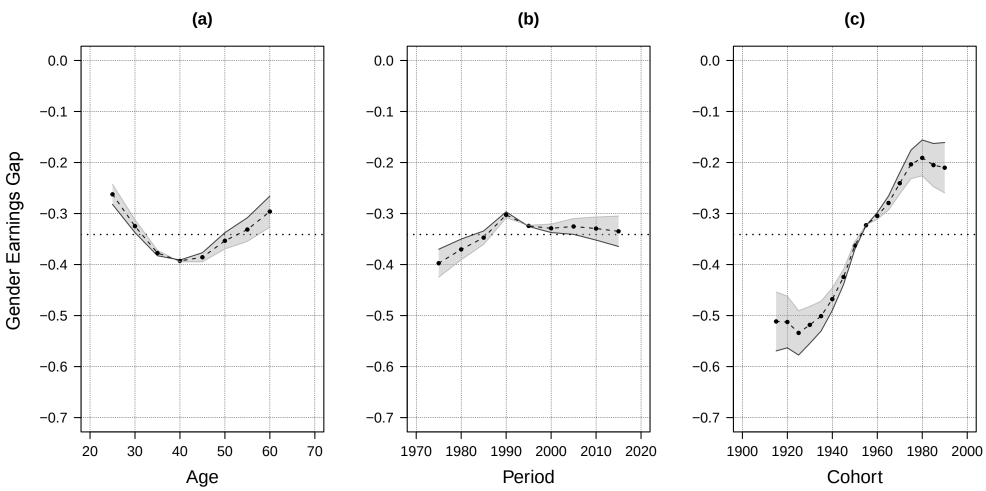
Notes: Hourly wages are analyzed instead of annual earnings, and full-time wage/salary workers who worked at least 26 weeks in the last calendar year are included. In the top panel (I.), the left y-axis indicates a range of cross-strata age linear effects, the x-axis indicates a range of cross-strata period linear effects, and the right y-axis indicates a range of cross-strata cohort linear effects. The dashed line indicates all possible cross-strata linear effects consistent with the data. The dashed line indicates all possible cross-strata linear effects consistent with the data. The shaded red region denotes the set of values consistent with the age-related assumptions, while the shaded blue region indicates the set of values consistent with the period-related assumptions. The solid line in the overlapping shaded regions refers to the feasible set of cross-strata linear effects given that the assumptions about the cross-strata age and period effects are satisfied. In the bottom panel (II.), the shaded areas represent the bounded cross-strata effects of age (a), period (b), and cohort (c) on the gender wage gap based on the three assumptions about the age and period effects. The dotted lines follow the midpoints in each shaded area. The dark bold lines along one edge of the shaded areas indicate when the cross-strata age linear effect is most positive, the cross-strata period linear effect is most negative, and the cross-strata cohort linear effect is most positive within the bounds.

Figure D.3: Bounding Analysis Results of Gender Differences in Mean Annual Earnings

I. Upper and Lower Bounds of Cross-Strata APC Linear Effects on the Gender Wage Gap



II. Bounded Cross-Strata Age, Period, Cohort Effects on the Gender Wage Gap



Notes: Differences in log mean annual earnings between men and women are analyzed instead of median differences. In the top panel (I.), the left y-axis indicates a range of cross-strata age linear effects, the x-axis indicates a range of cross-strata period linear effects, and the right y-axis indicates a range of cross-strata cohort linear effects. The dashed line indicates all possible cross-strata linear effects consistent with the data. The dashed line indicates all possible cross-strata linear effects consistent with the data. The shaded red region denotes the set of values consistent with the age-related assumptions, while the shaded blue region indicates the set of values consistent with the period-related assumptions. The solid line in the overlapping shaded regions refers to the feasible set of cross-strata linear effects given that the assumptions about the cross-strata age and period effects are satisfied. In the bottom panel (II.), the shaded areas represent the bounded cross-strata effects of age (a), period (b), and cohort (c) on the gender wage gap based on the three assumptions about the age and period effects. The dotted lines follow the midpoints in each shaded area. The dark bold lines along one edge of the shaded areas indicate when the cross-strata age linear effect is most positive, the cross-strata period linear effect is most negative, and the cross-strata cohort linear effect is most positive within the bounds.

References

Elbers, Benjamin. 2020. "Orthogonal Polynomial Contrasts and Applications to Age-Period-Cohort Models." Retrieved (osf.io/preprints/socarxiv/xrbgv). doi:10.31235/osf.io/xrbgv.

Fosse, Ethan and Christopher Winship. 2019. "Analyzing Age-Period-Cohort Data: A Review and Critique." *Annual Review of Sociology* 45:467–492. doi:10.1146/annurev-soc-073018-022616.

Fox, John. 2016. *Applied Regression Analysis and Generalized Linear Models*. 3rd ed. Los Angeles, CA: SAGE Publications.

Holford, Theodore R. 1983. "The Estimation of Age, Period and Cohort Effects for Vital Rates." *Biometrics* 39:311–324. doi:10.2307/2531004.