Difference in differences estimates by levels of an interaction: Case studies from the Power of the Pill literature

Randy Cragun*

*Department of Economics and Political Science, Birmingham-Southern College. Contact: rc@rcragun.net

3 June 2020

First version: 2020

Abstract

This is an abstract. Isn't it cool? I bet you will stop reading it soon and move on to the paper. I further propose that if you are still reading, it is because you think that this informal tone is somehow a rhetorical device that will help illustrate the point of the paper. It isn't. I just wanted to fill in some text here so that I would not lose the formatting.

Introduction

Difference in difference (DD) estimates are a common way to look at the effect of a policy change, but some research has lost sight of the "difference" nature of DD and instead treats the method as equivalent to including group and time dummies in a regression with an indicator for when treatment turns on. Such regression models differ from a standard DD model when the treatment effect is analyzed at multiple levels of some covariate, and this paper analyzes cases where the estimates will be inconsistent as a result. The results are analogous to findings by Yzerbyt, Muller, and Judd (2004) that if the effect of a treatment is allowed to vary with some individual difference, then controls need to be interacted with that individual difference as well.

Difference in differences overview

DD is employed when the researcher believes that two time series would have evolved in the same way absent some intervention. For instance, Card and Krueger estimate employment levels in New Jersey and Pennsylvania before and after a minimum wage change in NJ and argue that the change in employment in PA is a good counterfactual for the employment change in PA. For the two-group, two-period case where only one group is ever treated, a standard estimator for the treatment effect is $\bar{Y}_{i=1,t=2} - \bar{Y}_{i=1,t=1} - (\bar{Y}_{i=2,t=2} - \bar{Y}_{i=2,t=1})$, where $i = 1$ refers to the treatment group, t indexes time, and $\bar{Y}$ is the average of the outcome variable. The estimate can be obtained either through direct calculation of the means or by a regression on dummies for being in the treatment group and being in the post-treatment time.

The DD methodology is popular because it controls for a broad set of potential confounds by differencing out time-invariant group-specific factors and group-invariant time-specific factors with simple group and time dummies. The basic assumption required is that the changes in groups that do not change their treatment status at a given time are good counterfactuals for the changes in groups that do change their treatment status at that time. This is often called the "common trends" condition. Some violations of the common trends assumption can be remedied by modelling the trends in the outcome variable using more than one pre- and post-treatment observation. For instance, if NJ's employment was rising before they lowered the minimum wage while PA's was staying constant, it would be foolish to simply use the before and after average employment rates. We would want to also difference out the upward trend in NJ.

The DD model can be expanded to multiple groups. The most common way to do so is probably through a regression. We start with a policy treatment that turns on at randomly-selected times for different groups that would otherwise behave similarly. For concreteness, consider a simplified version of the scenario in Bailey, Hershbein, and Miller

(2012): most states within the US lowered the age of majority (AoM)[1] sometime during the 1960s or 1970s, and these changes may have reduced the cost of obtaining and using oral contraceptives for the women who reached the affected ages after the policy change, which might have changed the incentives for early human capital accumulation and changed the shape of those women's age-earnings profiles. Suppose that we wanted to know the effect of living under an AoM of 18 when you were age 19 on your earnings at age 30. If the timing of the policy changes is random[2], we can estimate the treatment effect with the difference-in-differences model of the average earnings at age 30 of women from birth cohort c who lived in state s at age 19:

$$Y_{s,c} = \alpha + \delta Policy_{s,c} + \sum_{C} \beta_C 1[c = C] + \sum_{S} \gamma_S 1[s = S] + \eta_{s,c} \quad (1)$$

where

- s and index state,
- c indexes birth cohort,
- Y is the average earnings at age 30 of women from birth cohort c who lived in state s at age 19,
- $Policy_{s,c}$ is 1 if cohort c in state s was subject to the new policy at age 19 and 0 otherwise,
- $1[\cdot]$ is an indicator function that equals 1 if the condition is true and 0 otherwise,

and $\eta_{s,c} \sim N(0, \sigma_c)$.[3] The summations are sets of dummy variables for state and birth cohort. δ is the average treatment effect.

Interaction of treatment in DD

As economists, we are probably interested in the entire age-earnings profile—not just earnings at age 30. If we wanted to estimate effects at many ages, we could estimate Equation 1 separately for each age or we could pool the data and (more efficiently) estimate one regression with *every* right-hand-side term in Equation 1 interacted with age group dummies. This methodology maintains the DiD structure of the single-age version. The pooled (across age) version of Equation 1 is

$$Y_{s,c,a} = \sum_{a} \alpha_a D_{s,c,a} + \sum_{a} \delta_a Policy_{s,c} D_{s,c,a} + \sum_{a} \sum_{C} \tilde{\beta}_{a,c} 1[c = C] D_{i,a} + \sum_{a} \sum_{S} \tilde{\gamma}_{a,S} 1[s = S] D_{s,c,a} + \tilde{\eta}_{s,c,a} \quad (2)$$

where $D_{i,a}$ is an indicator equal to 1 if person i is in age group a (in 5-year groups) and 0 otherwise.[4] I will refer to this model as the "full interaction" model.

It is common in practice to instead estimate

$$Y_{s,c,a} = \sum_{a} \tilde{\tilde{\alpha}}_a D_{s,c,a} + \sum_{a} \tilde{\tilde{\delta}}_a Policy_{s,c} D_{s,c,a} + \sum_{C} \tilde{\tilde{\beta}}_C 1[c = C] + \sum_{S} \tilde{\tilde{\gamma}}_S 1[s = S] + \tilde{\tilde{\eta}}_{s,c,a} \quad (3)$$

---

[1] The age at which a person was a legal adult.

[2] Conditional on observable or time-invariant characteristics of states.

[3] There is likely heteroskedasticity and error correlation over time, but addressing that is outside the scope of this paper (see, e.g., Bertrand, Duflo, and Mullainathan 2004; Conley and Taber 2011; Ferman and Pinto 2018; MacKinnon and Webb 2017, 2018; Stephen G. Donald and Kevin Lang 2007). I use state averages to avoid dealing with error correlation across individuals within states.

[4] The notation is a bit loose with intercepts and omitted categories. The age dummies do not need to enter separately if the other summations are over all ages, states, and birth years.

which I will refer to as the "constrained model". This is the method employed by Bailey, Hershbein, and Miller (2012). Notice that this does not condition on the full set of dummies that the age-specific DiD model (Eq. 2) does, as it is missing the interaction terms between age and state and age and birth cohort. Although many readers will be familiar with and comfortable with many DiD estimates with models analogous to Equation 1, I have seen no systematic treatment of the conditions under which consistency fails in models like Equation 3. Suppose that Y is weekly earnings (as in Bailey, Hershbein, and Miller 2012; Cragun and Chatterjee 2020). The first summation term in Equation 3 is a baseline age-earnings profile, and the model allows additive state-specific and birth-cohort-specific deviations from that baseline, but any state-specific changes in the shape of the age-earnings profile over time is a threat to this specification.

Think about how we should expect earnings for women to change over time. I reproduce below Figure III from Bailey (2006) to illustrate. Each line is the female labor for participation rate for a different US birth cohort. Employment is an important determinant of earnings, which tend to follow a similar pattern. To construct a simplified model of the patterns, assume that the early age-earnings profiles are completely flat and that later and later cohorts have steeper and steeper profiles.

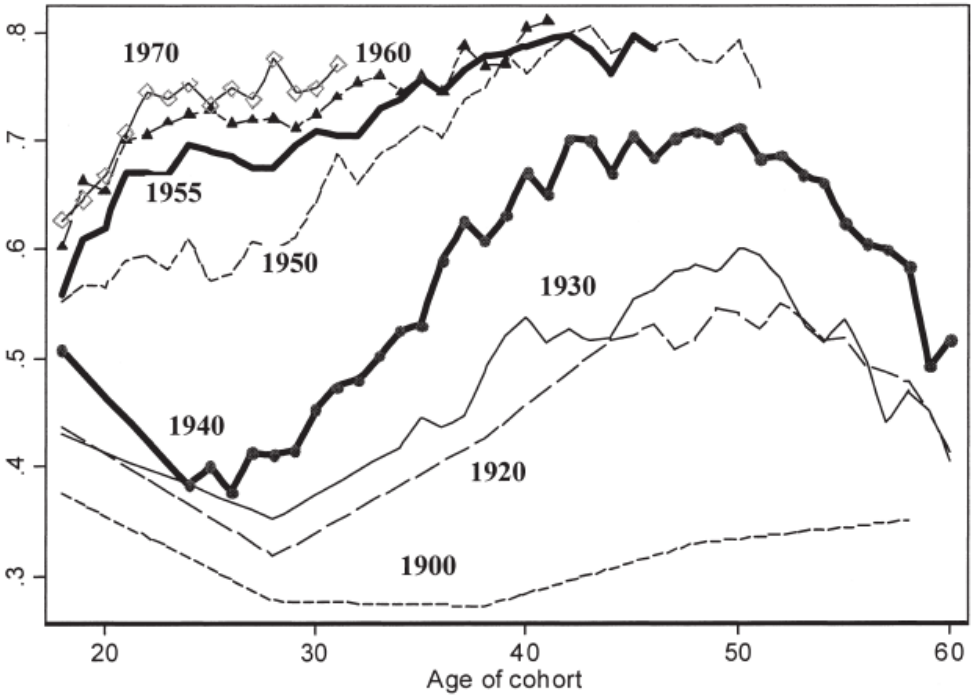


FIGURE III

Age-Specific Labor-Force Participation Rates, by Cohort and Age 1900–1970

Pre-1964 data are averaged over cohorts as in Smith and Ward [1985, Table 1]. For instance, the participation rate for women ages 14 to 19 in 1950 is plotted in this figure as the cohort of 1930 at those ages. Data after 1963 represent participation rates for a single year of birth cohort at the reported age. Synthetic birth cohorts are computed by subtracting the reported age from the year of the survey. Bold lines depict the 1940 and 1955 cohorts. The March sample includes all women not in the military or inmates ages 16 to 60.

Source: 1964–2001 March *CPS;* for years before 1964, data are from Smith and Ward [1985, Table 1].

Suppose that the true data generating process is

$$\bar{Y}_{a,c,s} = a \times \mu_s \times c + Policy_{s,c} \quad (4)$$

where μ_s is an iid positive state-specific random shock. For cohort 0, the average age-earnings profile (for untreated cohorts) is the same across states and has a slope of 0. The slope then increases with each successive birth cohort in a state-specific manner determined by μ_s. The policy treatment variable has a uniform effect of 1 at all ages.

Suppose we estimate Equation 1 for age group $a = 1$ with data generated from this process. Then $\bar{Y}_{1,c,s} = \mu_s \times c + Policy_{s,c}$. When state r changes its policy, the first difference between cohorts within that state is $\mu_r + 1$, and the counterfactual difference is the average of the μ_s for all the other states, which will equal μ_r on average. Thus, we can get unbiased estimates of the treatment effect of the policy (which is 1 in this case). The counterfactual comparisons with Equation 3 are exactly the same as for Equation 1.

Note that μ_s does not change over time, but this does not mean that its effects are controlled for with the state dummies in Equation 3 because it is not an additive effect.

Numerical examples

I generate a series of 3000 data sets with the DGP in Equation 4 with 50 states, 12 birth cohorts, and 3 ages and estimate Equations 2 and 3 with OLS and then plot kernel density estimates for the 3000 coefficients for each age from the two models. I draw μ_s first from a uniform density on [0,1] and then on [-1,0] to illustrate the direction of bias in estimates of Equation 3.

Estimates of δ are in Figure 1. As expected, with the full interaction model, the estimates are centered at 0, the actual treatment effect. Estimates of Equation 3, the constrained model, are biased, and the direction of bias depends on the correlation between age and $\mu_s \times c$. Only the effect at the average age is estimated without bias.

One particularly troubling fact about these estimates is that the bias at each age depends on how ages are normalized.

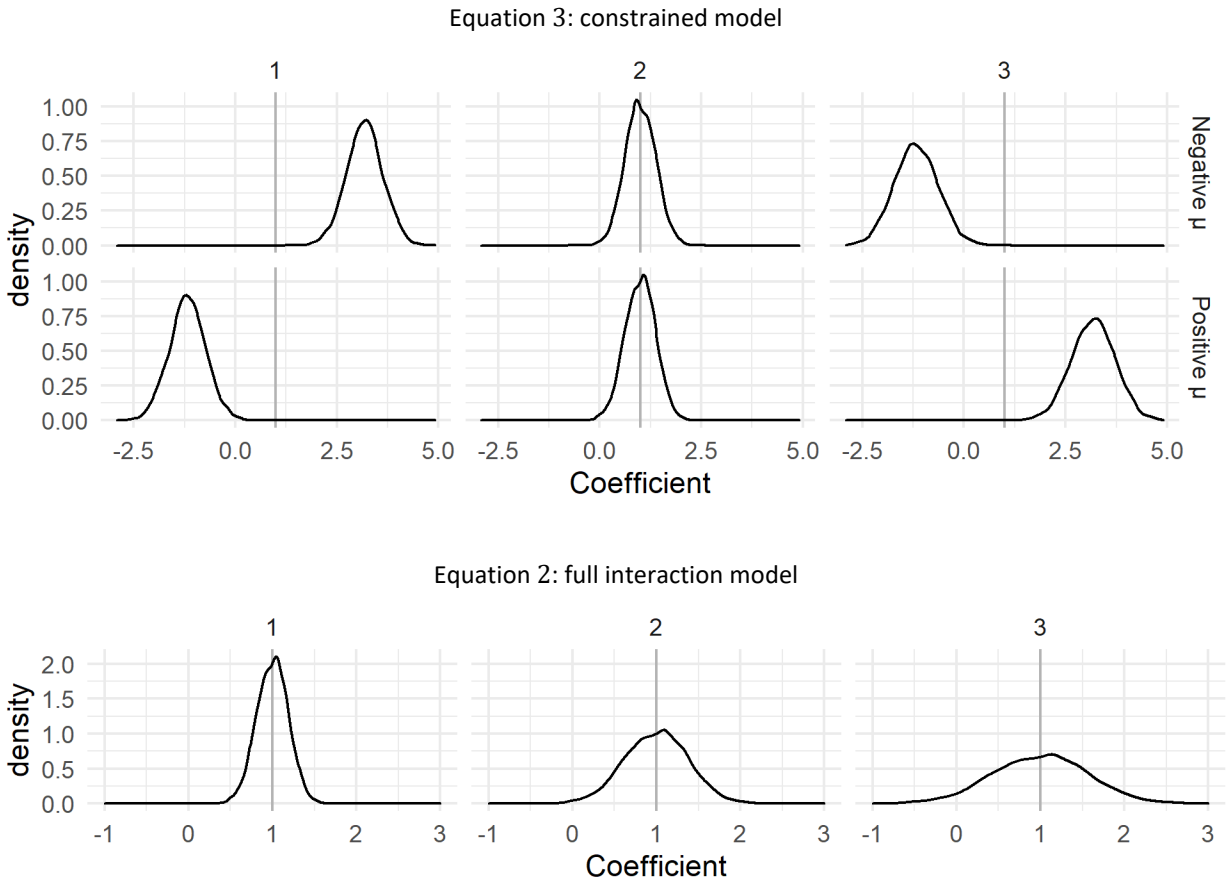


FIGURE 1: ESTIMATES OF δ FROM EQUATIONS 2 AND 3 WITH DATA GENERATED BY EQUATION 4

Real world example: Power of the Pill and ELA

In this section, I attempt to reproduce estimates from Bailey, Hershbein, and Miller (2012) (hereafter BHM). Although the simplified example above used only age of majority laws to identify changes in treatment status, BHM use a variety of types of policies in addition to AoM changes: repeal of "Comstock" laws that banned the sale of the pill; legislative, judicial, and administrative "mature minor" doctrines that allowed minors to give medical consent if they were mature enough to understand the choice; family planning policies that explicitly granted the right to provide contraceptives to minors (and sometimes mandated provision of such services on request); and medical consent statutes that set legal minimum ages for giving consent for medical treatment that differed from the age of majority. A birth cohort is treated if there was some law that would have allowed young women under age 21 to obtain and use the pill legally without the consent of a parent or guardian before the year in which that cohort turned 21.

To estimate Equation 2, we would ideally want longitudinal data on individuals that included their birth date, state of residence at ages 18 – 20, and earnings at each age. BHM use restricted-access data from the National Longitudinal Study of Young Women (NLYW). The NLSYW reports time series of earnings for each woman and her birth state, which BHM use instead of state of residence in early adulthood. I do not have access to the NLSYW, but there are a few alternatives: the Current Population Survey (CPS) and the decennial Censuses. Public-use microdata samples from the decennial Censuses report state of birth, but the earnings data are only observed every 10 years, so few cohorts are observed at any given age. The CPS is monthly but does not contain information on birth state. Within the CPS, there are multiple measures of earnings, and not all are available every month. The Annual Social and Economic Supplement (ASEC, sometimes just called the "March CPS") to the CPS asks more detailed questions about

employment and earnings, so the main analysis uses only the ASEC, but I check the results against alternatives with every CPS month and with the decennial Censuses.

Although there is no statistical test for bias, there are ways to get suggestive evidence. Randomization inference (RI) simulates the distribution of a statistic under a null hypothesis with random reassignment of the treatment condition. If the assignment mechanism is known, it can be reapplied arbitrarily-many times.[5] If the treatment has no effect and the estimator is unbiased, the empirical distribution of the estimates should be centered at 0. A hypothesis of no effect in this case is a joint hypothesis that all the α_a are 0.

Randomization inference procedure

I use the following RI procedure to estimate the distribution of the OLS estimates of α from Equations 2 and 3. The strategy is similar to Ferman and Pinto (2018).

1. For each state, randomly sample a year between the minimum and maximum (inclusive) years for ELA from BHM.[6] Call the resulting variable ELAYear.
2. For each individual, construct imaginary ELA based on ELAYear for her state of residence (ELA = 1 if sample year – age + 20 > ELAYear and 0 otherwise)
3. Estimate the relevant regressions (with or without the additional interactions)
4. Repeat 1 – 3 many times (I did 1000 for the smaller regression and 100 for the larger)

The RI estimates for Equation 2 are in Figure 2, while estimates of Equation 3 are in Figure 3. Each panel gives a kernel density estimate for the age group indicated above it. The densities in Figure 2 are centered at 0, but the densities in Figure 3 are not. In fact, the estimates for younger ages are centered many thousands of dollars below zero and the estimates for older ages are centered many thousands of dollars above zero. An analysis that uses Equation 2 and simply calculates clustered standard errors will severely overreject the null of no effect for the oldest and youngest women. We could still estimate Equation 3 and use the empirical RI distributions for p-values, but the point estimates are probably not informative because they will not be at the center of relevant confidence intervals (which could also be estimated with RI). One solution is to subtract the means of the RI distributions from the estimates with the actual laws as a bias correction. Another solution is to return to the DD model in Equation 2. It is also important to note that adding the full set of interactions (Figure 2) does not inflate the spread of the densities, so we are not losing much by including them.

One concern about using the full interaction model is that the treatment variable includes measurement error that might be severe: a woman's current state of residence is not the same as her state of residence at age 19 or 20. This will tend to attenuate estimates of the treatment effect more with the full interaction model than with the constrained model. It is possible to put bounds on this source of bias (Griliches and Hausman 1984), but I have not done the calculations yet. Thus, I report estimates of both Equation 2 and Equation 3 with RI p-values for the hypothesis that all the α_a are 0.

[5] The assignment mechanism in this case is not strictly known. I attempt multiple plausible assignment mechanisms and get nearly-identical results.

[6] I also tried randomly permuting the years for ELA from BHM. This is similar to Fisher's Exact Test.

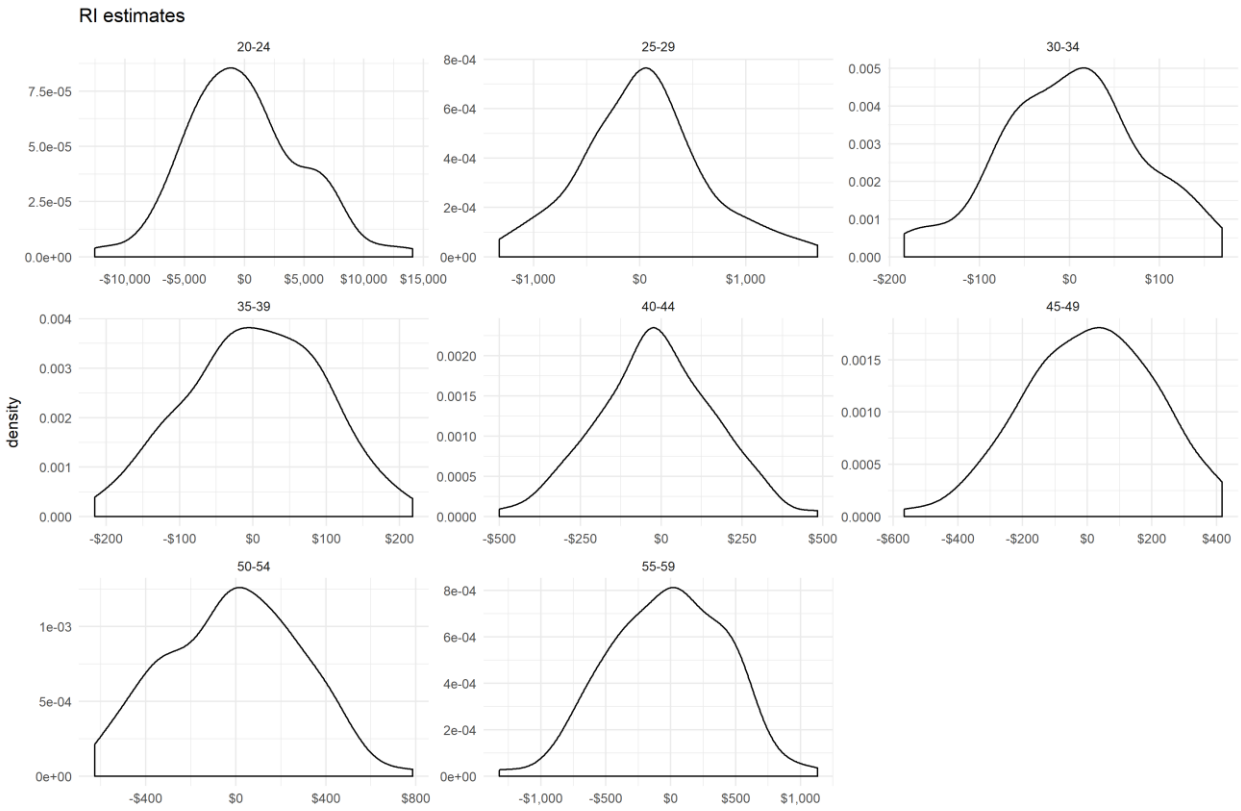


FIGURE 2: RI ESTIMATES OF EQUATION 2 (FULL INTERACTION MODEL)

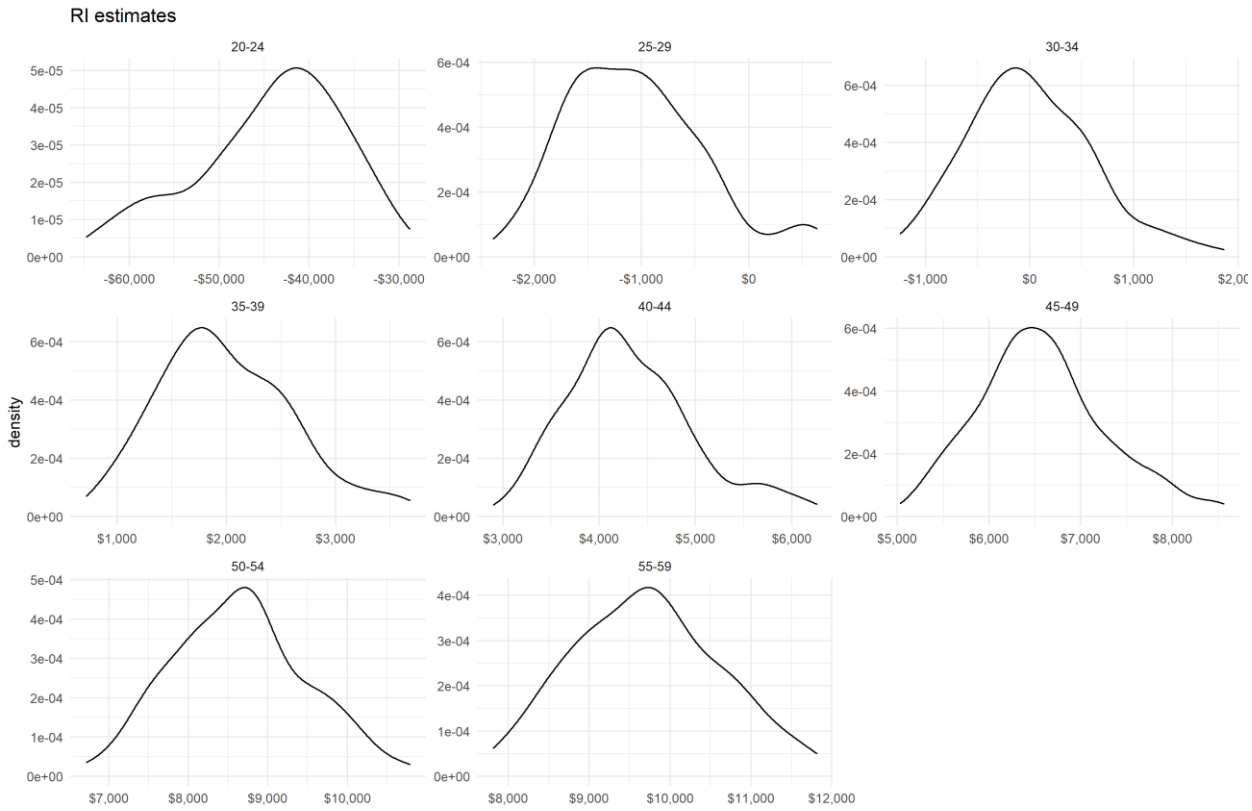


FIGURE 3: RI ESTIMATES OF EQUATION 3 (CONSTRAINED MODEL)

Now that we know the distributions of the candidate estimators, I estimate Equations 2 and 3 for Australia and the US using the BHM policy coding for ELA (instead of imaginary, randomly-assigned policy changes). I focus on estimates with the level of annual earnings including zeros, but leaving out zeros or using weekly earnings or the log of earnings has little effect. The policy coding in BHM is not based on the most recently-available information (primarily from Bailey et al. 2011), but updating the legal coding does not seem to make much difference, so I report only the estimates using the BHM coding for simplicity.

ASEC data

Here, I rely on the 1977–2018 Annual Social and Economic supplement to the Current Population Surveys from the US Census. This sample gives us annual wage and salary earnings for a broad set of women. A major weakness of these data is that we observe only what state the respondent was surveyed in rather than where she lived at ages 18–20. I check the estimates against the decennial Censuses, which report state of birth, and the results are not substantively different. A strength of these data is that we can rely on variation in policies in a larger set of states than in the National Longitudinal Survey of Young Women (NLSYW) used by BHM. The NLSYW surveyed women who were ages 14 to 24 in 1968, and the youngest would have turned 21 by the mid-70s, years before some of the age of majority changes.

I construct weekly earnings by dividing wage and salary income over the previous year by the number of weeks the person reported working in the previous year.[7] Because of the difficulty in determining labor earnings for the self-

[7] From 1962 through 1975, weeks worked are only reported in categories (e.g. 1–13 weeks, 14–26 weeks, etc.), so we calculate the mean number of weeks worked in each category for 1976–1979 and assign that value to anyone in the category before 1976.

employed, I ignore self-employment earnings and include in the sample only people who report that earnings at the job they worked at the longest over the previous year were wages or salary.

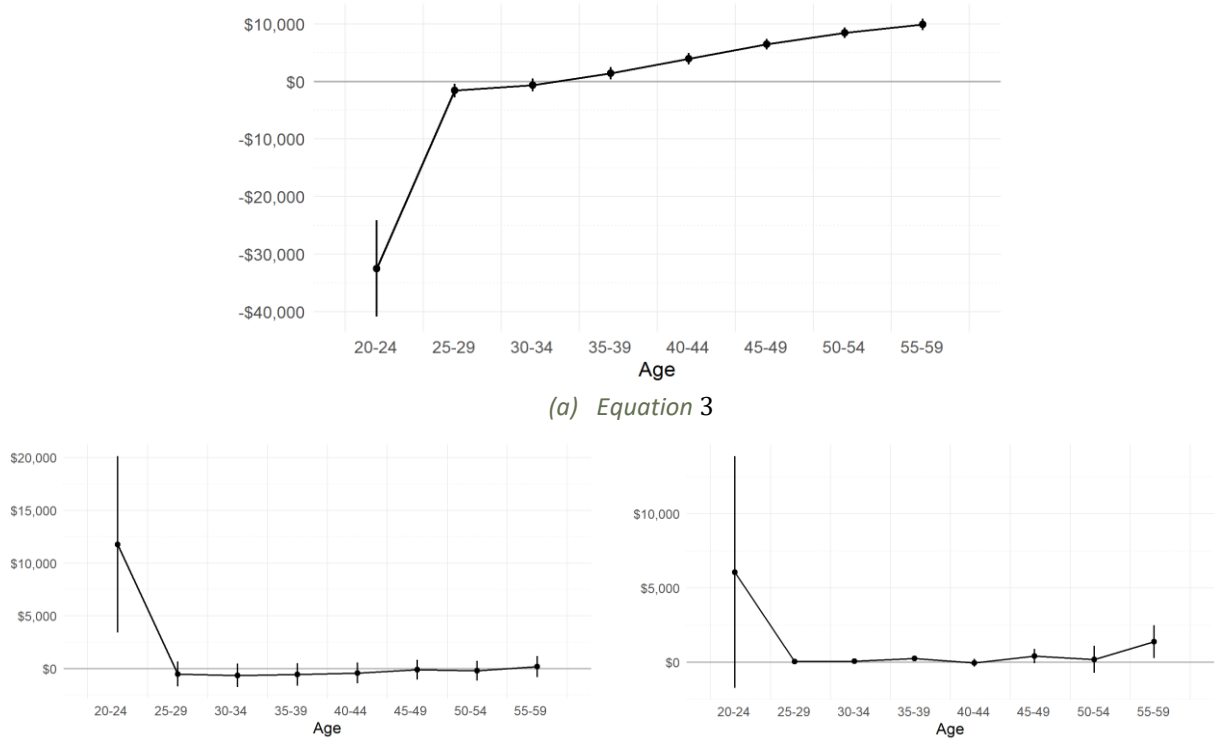
Some respondents reported not working but having positive wage and salary income. If their annual earnings were unusually low or high, we would expect that these were cases of measurement error in weeks worked (e.g. someone earned a small amount of income for a small amount of work but reported not working because they viewed their work time as negligible) or in wages (e.g. misreporting other income sources as wages).

I check results for annual earnings and check the estimates against the decennial Censuses as well.

I do not include zeros whenever the dependent variable is the log of earnings and always exclude infinite apparent weekly earnings (cases where annual wage and salary earnings were positive but the respondent reported working zero weeks in the last year).

Effect of ELA on age-earnings profiles for the US

Estimates of Equations 2 and 3 are in Figure 4. Using the BHM methodology (panel a), I find a much later age at which women earn positive returns than they did, but the broad pattern is similar. However, when I make the RI bias correction (panel b) or condition on the full set of interactions (panel c), the effects disappear. The vertical bars are 90% confidence intervals based on standard errors allowing heteroskedasticity and error correlation within states, but they are almost certainly much too small. There is much uncertainty regarding estimates for ages 20–24 because the surveys start in 1977, after most states had changed their policies. With the bias-corrected estimates of the constrained Equation 3, the RI-based p-value on the hypothesis that all the coefficients are 0 is 0.0417. Clearly, the low p-value is due almost entirely to ages 20–24. With the estimates of the full interaction Equation 2, the RI-based p-value on the hypothesis that all the coefficients are 0 is .



(a) Equation 3

(b) Bias-corrected Equation 3

(c) Equation 2 (full interactions)

FIGURE 4: ESTIMATES OF THE EFFECT OF ELA ON ANNUAL EARNINGS BY AGE WITH ASEC DATA. VERTICAL BARS ARE 90% CONFIDENCE INTERVALS BASED ON STANDARD ERRORS CLUSTERED BY STATE.

Life-cycle incomes for women in the US (monthly CPS)

Leaving this here for possible later inclusion

I check my estimates with the annual and weekly earnings against a measure of hourly earnings with the 1982–2018 Current Population Surveys from the US Census. Before 1990, I use the monthly surveys, while from 1990 on, I use only the March Annual Social and Economic Supplement.

References

Bailey, Martha. 2006. “More Power to the Pill: The Impact of Contraceptive Freedom on Women’s Life Cycle Labor Supply.” *The Quarterly journal of Economics* 121(1).

———. 2010. “‘Momma’s Got the Pill’: How Anthony Comstock and Griswold v. Connecticut Shaped US Childbearing.” *American Economic Review* 10: 98–129.

Bailey, Martha, Melanie Guldi, Allison Davido, and Erin Buzuvis. 2011. “Early Legal Access: Laws and Policies Governing Contraceptive Access, 1960–1980.”

Bailey, Martha, Brad Hershbein, and Amalia R. Miller. 2012. “The Opt-In Revolution? Contraception and the Gender Gap in Wages.” *American Economic journal: Applied Economics* 4(3).

Bertrand, Marianne, Esther Duflo, and Sendhil Mullainathan. 2004. “How Much Should We Trust Differences-In-Differences Estimates?” *The Quarterly Journal of Economics* 119(1): 249–75.

Conley, Timothy G., and Christopher R. Taber. 2011. “Inference with ‘Difference in Differences’ with a Small Number of Policy Changes.” *http://dx.doi.org/10.1162/REST_a_00049*. https://www.mitpressjournals.org/doix/abs/10.1162/REST_a_00049 (August 2, 2019).

Cragun, Randy. 2019. “Effects of Lower Ages of Majority on Oral Contraceptive Use: Evidence on the Validity of The Power of the Pill.”

Cragun, Randy, and Ishita Chatterjee. 2020. “Age of Majority and Women’s Early Human Capital Accumulation in Australia.” *ResearchGate*. https://www.researchgate.net/publication/341626454_Age_of_Majority_and_Women's_Early_Human_Capital_Accumulation_in_Australia (May 26, 2020).

Ferman, Bruno, and Cristine Pinto. 2018. “Inference in Differences-in-Differences with Few Treated Groups and Heteroskedasticity.” *The Review of Economics and Statistics*: 1–16.

Griliches, Zvi, and Jerry A Hausman. 1984. *Errors in Variables in Panel Data*. National Bureau of Economic Research. Working Paper. http://www.nber.org/papers/t0037 (May 26, 2020).

MacKinnon, James G., and Matthew D. Webb. 2017. “Wild Bootstrap Inference for Wildly Different Cluster Sizes.” *Journal of Applied Econometrics* 32(2): 233–54.

———. 2018. “The Wild Bootstrap for Few (Treated) Clusters.” *The Econometrics Journal* 21(2): 114–35.

Myers, Caitlin Knowles. 2017. “Confidential and Legal Access to Abortion and Contraception, 1960–2017.”

Stephen G. Donald, and Kevin Lang. 2007. “Inference with Difference-in-Differences and Other Panel Data.” *The Review of Economics and Statistics* 89(2): 221.

Yzerbyt, Vincent Y., Dominique Muller, and Charles M. Judd. 2004. “Adjusting Researchers’ Approach to Adjustment: On the Use of Covariates When Testing Interactions.” *Journal of Experimental Social Psychology* 40(3): 424–31.