The effect of retirement on subjective health: a comparison of different panel data estimators*

Rainer Winkelmann
University of Zurich

October 2019

Abstract

The paper estimates the effect of retirement on health, using a difference-in-differences (DID) framework for older workers in Germany between 1992 and 2017. Outcomes such as self-assessed health, satisfaction with health, or the number of doctor visits require the use of non-linear panel data models. The paper reviews recent developments in the related econometric literature, and considers nine different models in a unified framework and for a single dataset and parameter of interest, allowing to assess the robustness of results to specific modelling assumptions and to highlight the relative advantages of different approaches. The DiD results suggest a positive effect of retirement on subjective health and a negative effect on visits.

Keywords: nonlinear difference-in-differences, correlated random effects, fractional logit, Poisson regression

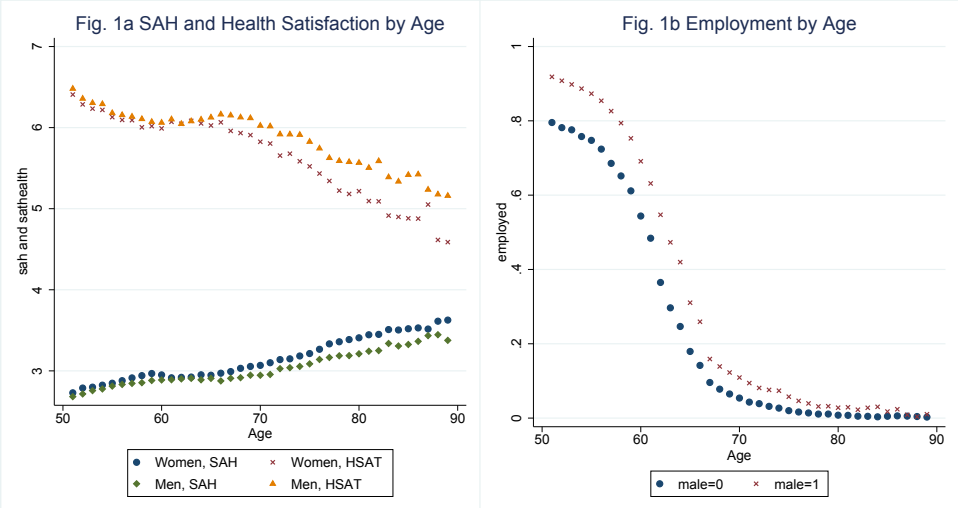
JEL classification: C23, I10, J26.

* Department of Economics, University of Zurich, E-mail: rainer.winkelmann@econ.uzh.ch. The data have been provided by DIW Berlin. I gratefully acknowledge financial support by the Swiss National Science Foundation. Part of the paper was written while I enjoyed the hospitality of the Department of Economics at the University of Melbourne. I received useful comments by Johannes Kunz as well as seminar participants at University of Queensland, University of Melbourne, Monash University and University of Bolzano.

1 Introduction

This paper provides new evidence on the question whether retirement leads to improvements in self-assessed health. The question is approached in a non-linear difference-in-differences (DiD) framework, estimating logit, binomial logit and ordered logit panel data models for women and men in Germany between 1992 and 2017, to determine the effect of retirement on health after accounting for year and individual specific fixed effects.

The paper has both an empirical and a methodological motivation. On the empirical side, consider evidence in Figures 1a and 1b. Figure 1a shows the average life satisfaction and self assessed health for respondents in the German Socio-Economic Panel (SOEP), for the age range from 50 to 90, separately for men and women. The graph is based on 203,652 person-year observations between 1992 and 2017. Satisfaction with health (HSAT) is measured on a 0-10 scale, self-assessed health (SAH) on a five-point scale, where 5 is the worst possible outcome and 1 is the best possible outcome (later, the scale will be reversed to allow for a simple comparison of the HSAT and SAH results).



Source: German Socio-Economic Panel, 1992-2017, own calculations.

There is a remarkable pattern here: although subjective health tends to decrease with age, the decline is temporarily suspended, in particular for men, for a roughly 10-year period between the ages of 60 and 70. As Figure 1b demonstrates, this is exactly the time when most of retirement takes place.[1] Of course, the positive correlation between retirement and subjective health might be a coincidence, due to third factors, or a consequence of sample selection, including for instance

[1]The developing gap between men and women at higher ages is most likely due to the higher mortality of men –

higher mortality for those with the worst subjective health.[2] But it appears equally possible that subjective health is stabilized *because* of retirement. Regression models can help answering this question, by controling for confounders, implementing a within-subject design for subjective health before and after retirement, using instrumental variables, or exploiting regression discontinuities.

On the methodological side, the previous empirical literature estimating models of subjective health as an outcome has oscillated between two extremes: either advocating the ordered probit or ordered logit models (mostly for SAH, as in Jones and Schurer, 2011), where it is difficult to quantify results and determine magnitudes of interest, or else the linear regression model (mostly for HSAT, as in Ferrer-i-Carbonell and Frijters, 2004), that is easy to interpret but stands in contradiction to the data generating process. This paper proposes a middle ground between the two extremes: use a cardinal regression model, but one that is non-linear in order the enforce upper and lower bounds, and consequently non-constant average partial effects required for this type of outcome (both HSAT and SAH).

Effectively, this amounts to estimating fractional response models for suitably scaled dependent subjective health variables. The question then becomes how to account for correlated unobserved heterogeneity, and thereby implement a difference-in-differences type estimator in such non-linear panel data models. Nine potential approaches will be discussed and compared in Section 3, among them the pooled pseudo ML estimator (Wooldridge, 2002), the correlated random effects approach (Papke and Wooldridge, 2008) and a number of fixed effects estimators, including one proposed by Baetschmann et al. (2015) for the ordered logit and one proposed by Machado (2004) for the binomial logit, respectively.

The estimators differ in the assumptions required for consistency, as well as in the population parameters they identify. To paraphrase Tolstoy in his novel Anna Karenina, "linear models are all alike, whereas every nonlinear model is nonlinear in its own way".[3] But there are some common

---

while the gender ratio is close to 50% at the age of 65, it drops to under 40% at the age of 85. Also, not all indicators of health point in the same direction. Mortality rates increase steadily from age 50 to age 85, as does the age-specific per-capita volume of prescriptions (in terms of defined daily dose); see Figure A1 in the appendix. The profile of the self-reported number of doctor visits (from the SOEP) has again a flat part in about the same age range (see Appendix A2).

[2]Unlike medicare eligibility in the US, there are no age-related changes in access to health insurance in Germany.

[3]Originally: "Happy families are all alike; every unhappy family is unhappy in its own way." As an example from econometrics, there is a consistent conditional maximum likelihood estimator for the logit fixed effects model but not for the probit model with fixed effects.

themes: small sample bias, few existing analytical results on consequences of misspecification, the incidental parameters problem, and the non-trivial task of interpreting parameters. The paper reviews recent developments in the related econometric literature, and applies them in a unified framework and for a single dataset. The comparison allows on one hand assessing the robustness of results to specific modeling assumptions, and on the other highlighting the pros and cons of the different approaches.

For a short preview of results, I find that different identification strategies lead to opposite conclusions regarding the effect of retirement on subjective health. For a given age and year, those retired report worse subjective health than those still working. However, when identification comes from the longitudinal dimension, i.e., differential time trends in health for the retired and non-retired, I find that the effect of retirement is positive although not large, e.g. a 2 percent increase in HSAT for men, and often not statistically significant. Within the class of models that allow for correlated unobserved heterogeneity, results are remarkable robust: conditional maximum likelihood, dummy variables as well as correlated random effects approaches yield very similar effect sizes.

2 Previous studies

People living in developed countries can expect to spent about one third of their adult life in retirement, and the share of retirees steadily increases as baby boomers leave the labor market. Understandably then, many studies in economics, epidemiology and related fields have been conducted to assess the consequences of these developments for health and wellbeing.

Past studies have often used one of two large-scale dedicated datasets, the Health and Retirement Study (HRS; Charles, 2004, Insler, 2014) or the Survey of Health and Retirement in Europe (SHARE; Mazzonna and Peracchi, 2017). Other studies focusing on mortality have mostly relied on administrative data (e.g. Hernaes et al., 2013). Beyond studying the effect of retirement per se, there has been some interest in the causal pathways, as the effect of retirement on health seems to work primarily through changes in retirees' health-related behaviors (Insler, 2014, Eibich, 2015). Also, a number of meta studies have been conducted (Van der Heide et al., 2013, Shim et al., 2013, Kuhn, 2018), emphasizing the heterogeneity of findings and the lack of a single, conclusive answer.

Three major factors are responsible for the differences in results: measurement, method, and context. Measurement relates to the definition of retirement and health. For instance, Mosca

and Barrett (2016) find that voluntary retirement has no effect on mental health, while involuntary retirement has a negative effect. Many studies focus on early, rather than regular retirement (Kerkhofs and Lindeboom, 2007, Hernaes et al., 2013). Health can relate to mental or physical health. Focusing on longitudinal studies, Van der Heide et al. (2013) conclude that the effects on physical health are unclear, while there seem to be beneficial effects on mental health. Also, health can be measured using subjective or objective indicators, with findings being often positive for the former and non-significant or negative for the latter (Nichimura et al., 2018).

Regarding methods, any study needs to deal with the potential endogeneity of the retirement decision. A given correlation can indicate reversed causation, as people are forced to retire because their health deteriorates, or omitted third factors (such as insurance status, e.g. Medicare coverage in the US). The study by Kerkhofs and Lindeboom (1997) is one of the first using a difference-in-differences approach to address the endogenous decision linking retirement and health. They find for the Netherlands that health improves after early retirement.

Alternatively, instrumental variables strategies have been applied, using exogenous variation in retirement provided by quasi-natural experiments, for instance changes in mandatory retirement age and monetary incentives for early retirement (see e.g., Bloemen, Hochguertel and Zweerink, 2017). These estimators naturally lead to variation in effect sizes, either because the exclusion restriction does not hold exactly, or because the sub-population of compliers changes.

Lately, regression discontinuity designs (e.g. Eibich, 2015) have somewhat displaced DiD designs. However, fuzzy RDD is predicated on jumps of retirement probabilities at certain age thresholds (such as 65) which can in turn be relatively minor in practice, as exemplified by Figure 1b, that depicts entry into retirement as a smooth, gradual process. The DiD approaches, in contrast, uses information from all individuals that enter into retirement at some stage during the observation period.

Finally, it is clear that context matters. For example, it is often found that retirement of women has smaller or no effects, while the positive effects of retirement are bigger for physically demanding occupations (Kuhn, 2018). Also, there are differences by countries, possibly relating to the financial generosity of the pension system. As another example, Picchio and van Ours (2018) find a difference depending on whether the retiring person is single or living with a partner. Single men experience a drop in self-assessed health upon retirement, married men an increase.

It is fair to say that not much, if any, attention has been paid so far to the discrete and bounded

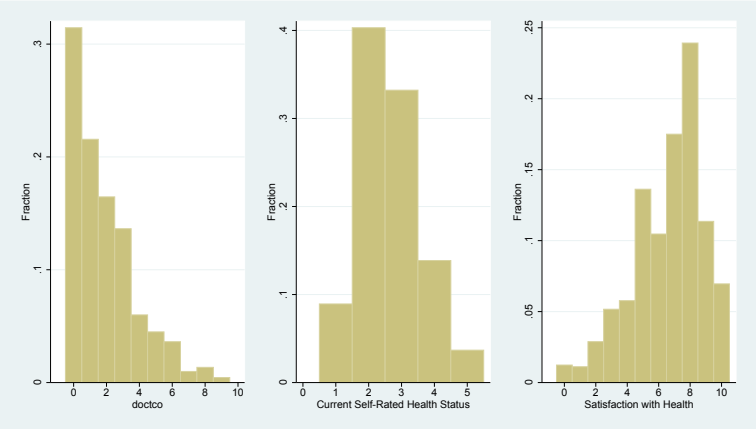
nature of the employed outcome measures, often indicators of subjective health, activities of daily living scales, or health services utilization counts. Obviously, non-linear models *per se* cannot solve the underlying problems of reversed causation, dynamic selection into retirement etc. But ignoring the inherent non-linearities can generate problems of its own, like wrong average treatment effects, nonsense predictions and the like. So, arguably, it is worthwhile to have a closer look at this aspect as well, which is what this paper does.

3 Data and methods

3.1 Measuring health and defining retirement

Data are from the German Socio-Economic Panel, a representative longitudinal household survey initiated in 1984. Because the questions on self-assessed health (SAH) and satisfaction with health (HSAT) were consecutively asked only from 1994 onwards, our data cover the period 1994-2017. The study focuses on SAH and HSAT in the three years preceding and following retirement, for individuals between the ages of 50 and 80.[4] Retirement in t is defined here as working in $t-3$, $t-2$ and $t-1$, and not working in t, $t+1$ and $t+2$. This rules out people who retire from unemployment. Also, no distinction is made between retiring from full-time work or from part-time work. A few individuals with more than one such transition in the data are dropped from the sample. The result is a balanced panel, with 6 years of data for a total of 2,274 individuals.

The distributions of outcomes are given in Figure 2.



[4]Results are robust to changing to a shorter two-years or a longer four-years window around retirement. For a recent discussion of the advantages of using a symmetric design, see Sylvain Chabé-Ferret, 2015.

The exact wording of the questions deserves some comment: for SAH, it is "How would you describe your current health?" with labeled categorical answers "Very good", "Good", "Satisfactory", "Poor" and "Bad". For HSAT, the question is "How satisfied are you with your health?". Answers are labeled numerically, taking values 0, 1,..., 10. Next to the lowest value "0" appears the text "completely dissatisfied", and next to the highest value "10" the text "completely satisfied". This is interesting for two reasons. First, respondents are aware of the numerical values of the scale when they respond. Second, the scale is not open: one cannot be more satisfied than "completely". This is a sharp bound. While it might be a strange notion for an economist, where there always is a "more", and more is always better, the question thus relates to the concept of satisficing, which seems quite intuitive and answerable to people.

In terms of modeling, this has two consequences. First, since there is a "perfect state", one can define a fractional variable, say 0%, 10%, ..., 100%, that measures the "degree of complete satisfaction" in percentage terms.[5] Second, neither should the top and bottom categories be viewed as censored, nor does the linear index represent "true" unbounded satisfaction. This removes any ambiguity in the interpretation of mean differences in such outcomes, an issue recently raised by Bond and Lang (2019).

The argument is less compelling for self-assessed health. Still, we take the same approach and assign numerical values to labels, which allows then using regression models for the conditional expectation function. The degree of very good health is thus 100% for highest response "Very good", 75% for the next response "good" and so forth, assuming a linear numerical cardinalization of the categorical responses. As for HSAT, the fractional SAH is bound between 0 and 1, but it is not binary.[6]

Table 1 provides selected descriptive statistics, separately for retired and non-retired observations. By construction, the two groups have the same size, and the average age in the retired sample is exactly three years above the average age in the non-retired sample. For the retired observations, the statistics do not refer to the moment of retirement, but rather to the average characteristics

[5]See also Studer and Winkelmann (2017). A related argument applies when considering data on the frequency of an emotion, such as happiness or calmness, with answer categories "never", "seldom", "sometimes", "often", "mostly", "continuously", as in the Dutch LISS panel (Picchio and van Ours, 2018). Again, there are sharp bounds of the outcome variables. There cannot be more than "continuously" or less than "never".

[6]It has been quite common in the literature to dichotomize SAH by having a single category for "poor health", and another one for the rest (see e.g, Dave et al., 2008). While this allows for a convenient application of standard binary response models, it entails a loss of information, as not all variation in the dependent variable is used.

during the first three years of retirement.

Overall, I find considerable variation in both age and time of retirement. The average age in the last year of work is 61.5 (and the average age of the employed and retired accordingly 60.5 and 63.5 years, respectively, since all individuals are observed for three years prior to retirement and three years during retirement). Since the minimum age of the employed sub-sample is 51, and the maximum age is 76, the retirement age in this sample varies between 53 and 76, and there are retired people observed between the ages of 54 and 79.

The time invariant variables "educ" (years of schooling) and "male" are identical in the two sub-samples by construction. As the later focus will be on fixed effects estimators, educ, like all other potential time-invariant confounders, will be absorbed by the fixed effect, whereas gender will be used as a splitting variable, to allow the retirement effects to differ between men and women.

Retirement is associated with a fall (worsening) of the two subjective health indicators (that are defined as fractions in Table 1). In contrast, the number of doctor visits falls as well. In the period leading up to retirement, more individuals work part-time (54%) than full-time. Of course, these rates differ between men and women. Retirement is associated with a reduction in household income by about 20 percent, something to account for in the following regression analyses. Finally, about 16% of all retirement observations are associated with "early retirement", defined here as being retired and younger than 60 years.

3.2 Difference-in-differences in non-linear models

All models considered in the following analysis are varieties of the basic non-linear regression

$$\mathrm{E}(y_{it}|retired_{it}, age_{it}, z_{it}, \theta_t, \alpha_i) = G(\alpha_i + \theta_t + \pi \text{ retired}_{it} + \delta \text{ age}_{it} + z'_{it}\gamma) \tag{1}$$

where G is some appropriately defined non-linear, continuous and monotone function that accounts for the boundedness of y_{it}, α_i is an individual specific fixed effect, θ_t is a time effect, z_{it} are control variables including income and marital status, and π is the main coefficient of interest. Estimation of π in the sample exploits variation in retirement age and year of retirement.[7]

In a linear DiD model, the counterfactual outcome, i.e. the potential subjective health of the retired had they not retired, is established using the parallel trends assumption: the pre-post retirement

[7]The age-period-cohort indeterminacy is addressed by excluding the first two time dummies, i.e., assuming that there was no underlying time trend in subjective health between 1994 and 1995.

change in health of those who did retire at age a in year t, had they not retired, is assumed to be the same as the change in health for those of the same age who did not retire in that year.

But this assumption is unreasonable in a non-linear model. To understand why, consider the simplest possible setup, where $G(\ )$ is the cumulative density function of the logistic distribution, denoted as Λ, and there are two regressors only, $T \in \{0, 1\}$ for intervention period (pre- and post) and $D \in \{0, 1\}$ for treatment or control group. In this case

$$\mathrm{E}(Y|T, D) = \Lambda(\beta_0 + \beta_1 T + \beta_2 D + \beta_3 D \times T)\,,$$

The pre-treatment outcome for the treated is $\mathrm{E}(Y|T = 0, D = 1) = \Lambda(\beta_0 + \beta_2)$. The change in the control group is $\Delta = \Lambda(\beta_0 + \beta_1) - \Lambda(\beta_0)$. Applying the change Δ to obtain a counterfactual for the treated, we thus obtain a "treatment effect" (or cross difference) given by

$$\Lambda(\beta_0 + \beta_1 + \beta_2 + \beta_3) - \Lambda(\beta_0 + \beta_2) - \Lambda(\beta_0 + \beta_1) + \Lambda(\beta_0)$$

We see that $\beta_3 = 0$ is neither necessary nor sufficient for a zero "treatment effect" (Ai and Norton, 2003). The reason is that unless $\beta_2 = 0$ (the pre-treatment outcomes for treated and controls are identical) changes in the control and treated groups are evaluated at different curvatures of G, and thus cannot be the same in the absence of a treatment effect.[8] The parallel trends assumption cannot hold.

Hence, the counterfactual outcome should be established in a different way, by imposing the common trends assumption on the linear index function (see also Blundell et al., 2004, Puhani, 2012): Start at the pre-treatment outcome of the treated group, and ask, how the latent trend estimated for the control group would have changed the outcome of the treated in the absence of treatment (namely by $\Lambda(\beta_0 + \beta_1 + \beta_2) - \Lambda(\beta_0 + \beta_2)$. In this case, we obtain the causal effect

$$\mathrm{CE} = \Lambda(\beta_0 + \beta_1 + \beta_2 + \beta_3) - \Lambda(\beta_0 + \beta_1 + \beta_2)$$

This effect is zero iff $\beta_3 = 0$ and the sign of CE and β_3 are the same. Usually, the model also includes control variables and one can compute the average partial effect by averaging over the distribution of x.[9] For small β_3, the discrete change can be approximated by the partial effect

$$\mathrm{PE} = \Lambda'(\beta_0 + \beta_1 + \beta_2)\beta_3 = \Lambda(\beta_0 + \beta_1 + \beta_2)[1 - \Lambda(\beta_0 + \beta_1 + \beta_2)]\beta_3$$

[8] Another shortcoming of this approach is that the implied counterfactual can take negative values or values greater than one, which is logically impossible.

[9] Without control variables, the model is saturated and there is full equivalence to the linear DiD model.

The same arguments apply to the ordered and binomial logit models, and to the panel data version of the DiD estimator, where one needs estimates of the individual specific effects α_i in order to compute the average causal effect. For the exponential regression model, it is meaningful to make a "proportional trends" assumption, i.e. the relative change in the outcome for the treated group in the absence of treatment is the same as the relative change observed for the control. In this case, the treatment odds-ratio is simply the exponentiated coefficient on the interaction term.[10]

3.3 Limitations

While fixed effects methods address the possible confounding from pre-existing health conditions, they might not be able to fully capture unobservable transitory shocks affecting both retirement decisions and health. For example, if a purely temporary health shock causes a person to leave the labor force, the full "recovery" within a year will be erroneously counted as a positive effect of retirement when it truly would have taken place anyway. But this scenario (the "Ashenfelter Dip"), while perhaps plausible in some job-training situations, seems less plausible for retirement: If the health shock is temporary, why not stay in work? The more likely scenario for a negative health shock is that it may cause a person to retire if the shock is perceived to be persistent. But then, the change in (average) health pre-/ post-retirement will tend to be negative. This is a problem in studies that find a negative effect using the longitudinal design (such as Dave et al., 2008, Charles 2004), but makes a positive finding even stronger.

The severity of this issue also depends on the time window. If the data cover only a short interval (such as one year on each side of retirement), the panel estimator is more heavily influenced by transitory changes of health. While this paper does not estimate a fully dynamic model of retirement and health, and in fact assumes a constant effects model, the longer time window allows to reduce the aforementioned concern regarding temporary shocks. Also, health induced retirement is more likely to occur among those retiring before the normal retirement ages. A simple way to capture this effect is by way of including a separate early-retirement dummy.

Finally, a further potential concern is selective attrition, be it from panel non-response or mortality.

[10]The changes-in-changes model provides an alternative approach to estimating treatment effects when observing treated and control units over time (CiC, Athey and Imbens, 2006). However, that estimator does not exploit the presence of real panel data, where the same individual is observed repeatedly pre-and post treatment. Also, model (1) identifies the treatment effect from changes across sub-populations in mean outcomes only, whereas CiC uses changes in the entire distributions and thus requires stronger independence assumptions.

If driven by time-invariant unobserved confounders, the DiD strategy will be unaffected. Differential trends, however, for example because those whose health worsens most over time are more likely to drop out from the sample, may become a problem. It is unclear whether in such a case attenuation bias prevails, or a general upward bias, as such adverse health events are likely correlated with retirement. In any case, mortality for people aged around 60-65 is low, so this may not matter much in practice.

3.4 Models and Estimators

Estimation of the parameters of (1) depends on the assumptions one is willing to make regarding the distribution of time-invariant unobserved heterogeneity α_i. To simplify notation, define $x_{it} = \{\mathbb{1}(\text{year} = t)_{it},\ \text{retired}_{it},\ \text{age}_{it}, z_{it}\}$. Conditional models use $\mathrm{E}(y_{it}|x_{it}, \alpha_i) = G(x'_{it}\beta + \alpha_i)$, marginal models $\mathrm{E}(y_{it}|x_{it}) = G(x'_{it}\tilde{\beta})$, or, in order to control for time-invariant individual heterogeneity, $\mathrm{E}(y_{it}|x_{it}, \bar{x}_i) = G(x'_{it}\tilde{\beta} + \bar{x}'_i\xi)$. The two models are in general not mutually consistent. For instance, if G is of logit form, there is no known non-degenerate distribution of α_i such that $\mathrm{E}(y_{it}|x_{it}) = \mathrm{E}_\alpha\mathrm{E}(y_{it}|x_{it}, \alpha_i)$. Still, there is a remarkable empirical observation that if either α_i is independent of the regressors, or else if α_i can be decomposed into a linear function of $\bar{x}_i$ and another term independent of the regressors, the marginal model is nearly unbiased for the average partial effects. This will be discussed in more detail in the next section.

Alternatively, if no assumptions are made on the distribution of α_i, it is possible to implement fixed effects type estimators. Treating the $\alpha'_i s$ as parameters to be estimated and including $N - 1$ dummy variables leads in general to the incidental parameters problem if T is small and fixed. Instead, one can assume that (a) the fractions follow a binomial logit distribution, or (b) that the G-function is of an ordered logit type, and then estimate the parameters (but not the α_i's) consistently by conditional maximum likelihood. An implicit assumption in this case is that observations are independent over time conditional on α_i. This assumption is problematic in the context of difference-in-differences. It for instance contradicts using cluster-robust standard errors, something that may be critical in order to obtain correct standard errors for the estimated treatment effects, as pointed out by Bertrand, Duflo, Mullainathan (2004).

All in all, the following nine models will be compared in the empirical analysis:

Marginal models	*Conditional models*
Pooled logit (PML, GEE)	Logit dummy variables (≡Binomial logit DV)
Logit correlated random effects (CRE)	Fixed effects (FE) binomial logit (CML)
Pooled ordered logit	Fixed effects OLS
Ologit correlated random effects	Fixed effects Ologit (CML)
Pooled OLS	

where PML stands for pseudo maximum likelihood, GEE for generalized estimating equations, CRE for correlated random effects, and CML for conditional maximum likelihood.

3.4.1 Marginal modeling of ratings and fractional responses

The pooled logit pseudo maximum likelihood estimator can be used both for the uncorrelated and the correlated random effects models. This is arguably the simplest and most straightforward way to estimate average partial effects. In particular, it does not assume independence conditional on α_i, and clustered standard errors can be used for inference.

This estimator uses a misspecified likelihood function, both because the dependent variable is a fraction and not binary, and because observations over time for the same individual are necessarily correlated because they depend on a common α_i. Also, it does not identify the "structural" parameters β and the G-function is misspecified if the conditional model $\mathrm{E}(y_{it}|x_{it},\alpha_i) = G(x_{it}'\beta+\alpha_i)$ is logit.

For an intuition why this approach works for estimating average partial effects, consider the probit model with a normally distributed random effect, $\alpha_i \sim normal(0,\sigma_\alpha^2)$. In this case (see Wooldridge, 2002)

$$\Pr(y_{it}=1|x_{it}) = \Pr(\varepsilon_{it}+\alpha_i > -x_{it}'\beta) = \Phi(x_{it}'\beta_\alpha)\,,$$

where $\beta_\alpha = \beta/\sqrt{1+\sigma_\alpha^2}$. Hence, there is "attenuation bias" of the structural parameters, essentially due to variance re-scaling. However, the average partial effects, conditional on x, are given by $\mathrm{APE}(x) = \beta_\alpha\phi(x'\beta_\alpha)$. Hence, they are estimated consistently by the marginal model. The same results apply for estimating the parameters of $\mathrm{E}(y_{it}|x_{it})$ when y_{it} is a rating or a fraction. To introduce correlation between the random effect and x_{it}, replace α_i by c_i and assume that $c_i = \bar{x}_i'\gamma + \alpha_i$. In this case, $\mathrm{E}(y_{it}|x_{it},c_i) = \Phi(x_{it}'\beta + \bar{x}_i'\xi + \alpha_i)$ and $\mathrm{E}(y_{it}|x_{it},\bar{x}_i) = \Phi(x_{it}'\beta_\alpha + \bar{x}_i'\xi_\alpha)$, i.e., the same arguments hold, only that the means $\bar{x}_i$ need to be included among the regressors.

Unfortunately, marginalizing over α_i does not result in a closed-form for the logit case, regardless of the distribution of $\alpha_i|x_i$. The functional form of the true marginal CEF is thus unknown. If we nevertheless assume that it is logit, this is necessarily a misspecification. But it is known from Monte Carlo studies that the logit CRE tends to estimate APEs very well. Examples are Cramer (2007) and Ramalho and Ramalho (2010) for the cross section logit model with omitted variables, and Kwak, Martin and Wooldridge (2018) for logit CRE panel data models. Similar results are obtained here, see Figures 3a and 3b below, that show the mean relative difference between the estimated and the true average partial effects, for the logit CRE and OLS.

The figures include linear fixed effects results for comparison, because it is often claimed that simple OLS provides a good approximation to the APEs if the true DGP is logit (e.g., Angrist and Pischke, 2008). But this is not a general result, as the simulation results in Figures 3a and 3b illustrate. In Figure 3a (DGP I), 5000 repeated sample have been generated from a binary logit DGP, for $T = 2$ and $N = 1000$. The binary responses satisfy

$$\mathrm{E}(y_{it}|x_{it}, \alpha_i) = 1/(1 + \exp(-(x_{it} \times 2 + \sqrt{T} \times \bar{x}_i + \alpha_i)))$$

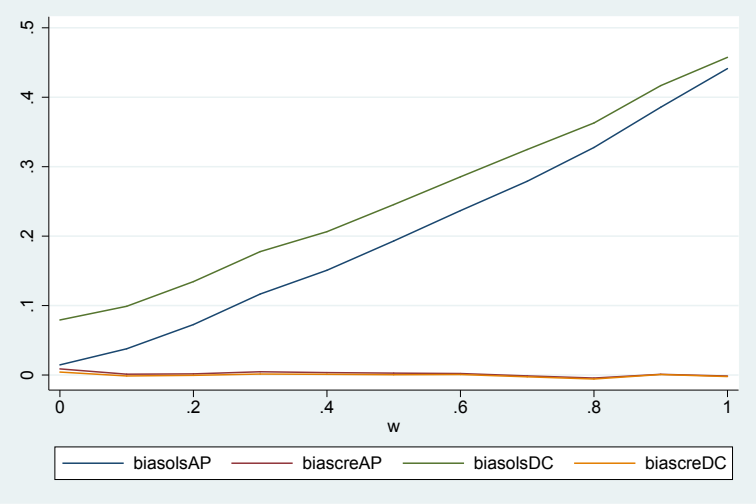
where $\alpha_i \sim 0.1 \times N(0, 1)$, and x_{it} is a mixture of a standard normal and a uniform(-1,1) distribution. The x-axis in Figure 3a states the mixture probability, w, that varies between 0 (pure normal) and 1 (pure uniform). Two models were estimated for each sample, a linear fixed effects model and a marginal logit model with mean $\bar{x}_i$ included as a regressor. For the linear model, the estimated APE is simply the slope coefficient $\hat{\beta}$. For the logit model, the APE can be estimated by

$$\widehat{\mathrm{APE}} = \sum_{i=1}^{N} \sum_{t=1}^{T} \hat{\beta} \frac{\exp(x_{it}\hat{\beta} + \bar{x}_i\hat{\xi})}{(1 + \exp(x_{it}\hat{\beta} + \bar{x}_i\hat{\xi}))^2}$$

The y-axis gives the average bias of the estimated APE relative to true APE. `biasolsDC` is the bias for the average discrete change associated with a one-standard deviation increase in x.

There are two key takeaways from Figure 3a: First, the simulation results confirm the earlier findings by Cramer (2007) and Kwak, Martin and Wooldridge (2018) that the marginal logit CRE model performs well in the presence of unobserved heterogeneity. Secondly, OLS can also be accurate, but this depends on the distribution of the regressor. With neglected non-linearity of the true CEF, the OLS APE becomes a function of the distribution of x. In particular, it has been known since Stoker (1986) that the OLS slope estimand equals the average partial effect, $\mathrm{E}[\partial \mathrm{E}(y|x)/\partial x] = \mathrm{E}[\partial G(x)/\partial x]$ if x is normally distributed. But for non-normal x, OLS slope and APE differ.

Figure 3a. *Relative APE bias of OLS and Logit CRE, DGP I*



This is very evident from Figure 3a. As the weight w shifts away from the normal distribution towards the uniform distribution, the relative bias increases from virtual none to about 45%. Hence, an unqualified use of the linear model to estimate marginal effects for binary or fractional responses cannot be recommended.

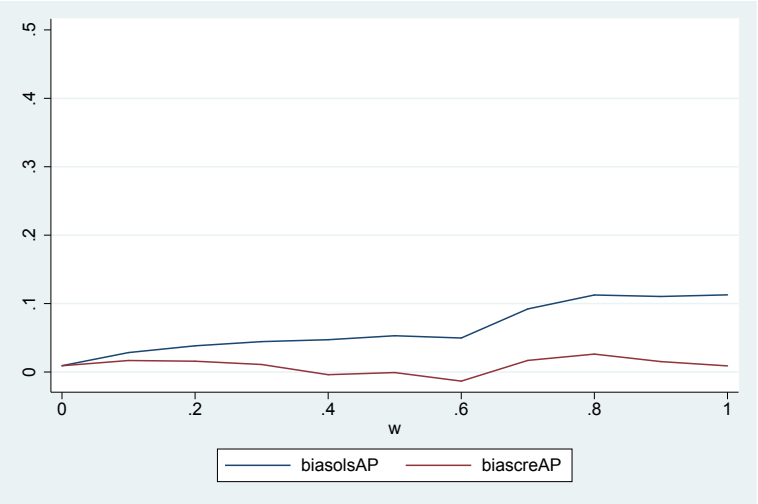
Results are very similar if α_i has a larger variance, say $N(0,1)$, or if it is uniform rather than normal. One may object that the simulation scenario is unrealistic, since it postulates a single regressor, and because the coefficient $\beta = 2$ and hence also the true APE, is quite large. In applied work, it is often the case that there are many regressors with a small partial effect each. Also, the previous DGP specified the correlated random effect as a linear function of $\bar{x}_i$, which meant that the logit CRE model was correctly specified, up to random effect α_i. Hence, one could argue that the good performance of the logit CRE average partial effect is not surprising.

To address both concerns, Figure 3b shows simulation results for an alternative data generating process (DGP II), where three modifications have been made. First an independent Normal(0,1) regressor x_2 is added; second, coefficients are smaller ($\beta_1 = -\beta_2 = 0.1$); and third, the correlated random effect is modeled as $c_i = \sqrt{T} \times (\bar{x}_{1i} + \bar{x}_{2i} + \bar{x}_{1i} \times \bar{x}_{2i}) + \alpha_i$, including the nonlinear interaction between $\bar{x}_{1i}$ and $\bar{x}_{2i}$. Hence, the linear-in-means logit CRE model is misspecified in this case.

From Figure 3b, we see that the relative bias of the logit CRE estimator remains quite minor, despite of misspecification. Although the linear fixed effects model has a smaller upward bias in this DGP, it still amounts to above 10 percent if the regressor of interest has a uniform distribution. On the other hand, it is not surprising that researchers reporting both OLS and logit APEs find "similar" effects, because as a practical matter, a 10 percent difference among often small APEs

will hardly be noticed, not matter much in terms of substantive conclusions, and in most cases not be "statistically significant".

Figure 3b. *Relative APE bias of OLS and Logit CRE, DGP II*



3.4.2 The binomial logit fixed effects estimator

Consistently estimating β in the conditional CEF (1) directly, without integrating over α_i, requires embedding it in a distributional model, for which a conditional maximum likelihood estimator (CMLE) exists, so that the incidental parameters problem (IPP) can be avoided. The logit CMLE (Chamberlain, 1980) cannot be applied directly, since the dependent variable is not binary. However, one can use an "expansion trick" (as in Baetschmann, Staub and Winkelmann, 2015, for the ordered logit model): consider $Y_{it} \in \{0, 1, \ldots, K\}$ as a realisation of K independent Bernoulli trials, with Y_{it} "successes" and $K - Y_{it}$ "failures". This is equivalent to assuming a binomial data generating process for the fractions:

$$Ky_{it}|p_{it} \sim binomial(K, p_{it}), \qquad i = 1, \ldots, N; \qquad t = 1, \ldots, T \tag{2}$$

where K is a known integer, and

$$y_{it} \in \left\{0, \frac{1}{K}, \frac{2}{K}, \ldots, 1\right\}$$

where y_{it} is the proportion of successes for observation unit i in period t and as before,

$$\mathrm{E}(y_{it}|x_{it}, \alpha_i) = p_{it} = \frac{\exp(x_{it}'\beta + \alpha_i)}{1 + \exp(x_{it}'\beta + \alpha_i)} \tag{3}$$

Machado (2004) introduced a consistent estimator of β that uses conditional maximum likelihood estimation to avoid the incidental parameters problem. Winkelmann and Xu (2019) propose a

simple way to implement the Machado estimator using off-the-shelf econometric packages. The method is based on an expanded dataset where every y_{it} is replaced by K binary variables d_{ijt}, such that $\sum_j d_{ijt} = Ky_{it} = Y_{it}$. Conditional on $\sum_t \sum_j d_{ijt} = \sum_t Ky_{it}$,$i = 1, 2, \ldots, N$, the conditional density function for each individual i can be calculated as follows

$$f\left(\{d_{ijt}\}|\sum_t\sum_j d_{ijt}\right) = \frac{\Pi_t\Pi_j p_{it}^{d_{ijt}}(1-p_{it})^{1-d_{ijt}}}{\sum_{\mathbf{s}\in S_i}\Pi_t\Pi_j p_{it}^{s_{jt}}(1-p_{it})^{K-s_{jt}}} = \frac{\exp(\sum_t\sum_j d_{ijt}x_{it}'\beta)}{\sum_{\mathbf{s}\in S_i}\exp(\sum_t\sum_j s_{jt}x_{it}'\beta)} \tag{4}$$

where $S_i = \{\mathbf{s} = (s_{11}, s_{12}, \ldots, s_{1T}, s_{21}...s_{KT})|s_{jt} \in \{0,1\}, \sum_t\sum_j s_{jt} = \sum_t\sum_j d_{ijt})\}$. It does not depend on α_i. Maximizing this conditional log-likelihood function yield a consistent estimator for β.

Winkelmann and Xu (2019) explore the properties of the estimator under failure of the binomial assumption, for example due to overdispersion. They find in simulation studies that the conditional maximum likelihood estimator maintains a rather good performance even if the binomial model is misspecified as long as at least one of three conditions is met: either the degree of overdispersion is modest, or else the length of the panel T or the number of Bernoulli trials K must be large (greater or equal to five, say). A main shortcoming of the CML approach is that the fixed effects are not estimated. Therefore, quantities of interest that are functions of α_i cannot be estimated either. This includes predicted means as well as average partial effects. We show below that an estimator for the average *relative* partial effect is available in the binomial logit fixed effects model.

3.4.3 Logit with unit dummy variables

An alternative to CML treats α_i's as parameters to be estimated. For instance, one can use a logit pseudo-likelihood approach and add $N-1$ group dummies to the set of regressors. For fixed T, every increase in sample size N leads also to an increase in the number of parameters, giving rise in principal to the aforementioned incidental parameters problem. However, no analytical results on the magnitude of this bias are available for the fractional response models. In simulations, Machado (2004) and Winkelmann and Xu (2019) find, that the bias of the logit dummy variables PML estimator disappears quite quickly with increasing T and, perhaps surprisingly, with increasing K as well. This is relevant in the present application in particular for HSAT, since $K = 10$ is quite large. Importantly, the difference between CML and logit DV remains small even if the conditional likelihood is misspecified due to overdispersion. For example, in the SAH model for men, the DV coefficient for `retired` is 0.0337; this is very close to the Blogit CML coefficient of 0.0323.

Alternatively, there exists by now a sizeable literature that studies bias reduction methods applicable when T grows as well, albeit at a slower rate than N. Several implementations are available for binary response models (including Kunz, Staub and Winkelmann, 2019), but none so far explicitly targets the fractional response case. Expanding the dataset to K copies with binary dependent variable is possible, but the resulting maximization problems can become computationally demanding.

3.4.4 Cardinalized ordered logit model

The standard ordered response model does not fit directly into the declared strategy of this paper, a focus on difference-in-differences effects for conditional expectations rather than probabilities. However, it can be "cardinalized", by using the same category labels as those generated earlier for the fractional responses. In this case, defining

$$\mathrm{E}(y_{it}|x_{it},\alpha_i) = \sum_{k=1}^{K} k/K \Pr(y_{it} = k/K|x_{it},\alpha_i) \tag{5}$$

ensures that $0 \leq \mathrm{E}(y_{it}|x_{it},\alpha_i) \leq 1$. For the ordered logit

$$\Pr(y_{it} = k/K|x_{it},\alpha_i) = \Lambda(\tau_{k+1} - x_{it}'\beta - \alpha_i) - \Lambda(\tau_k - x_{it}'\beta - \alpha_i)$$

where the τ's are additional parameters such that $\tau_0 = -\infty$, $\tau_1 = 0$ (a necessary normalization) and $\tau_K = +\infty$.[11] Fixed effects estimators for this model that avoid the incidental parameters problem are discussed in Das and Van Soest (1999) and Baetschmann, Staub and Winkelmann (2015). However, they don't estimate individual α_i's which means that estimates for the conditional expectation (5) are not available, and they assume independence of outcomes over time conditional on x_i and α_i, which is likely violated in DiD settings.

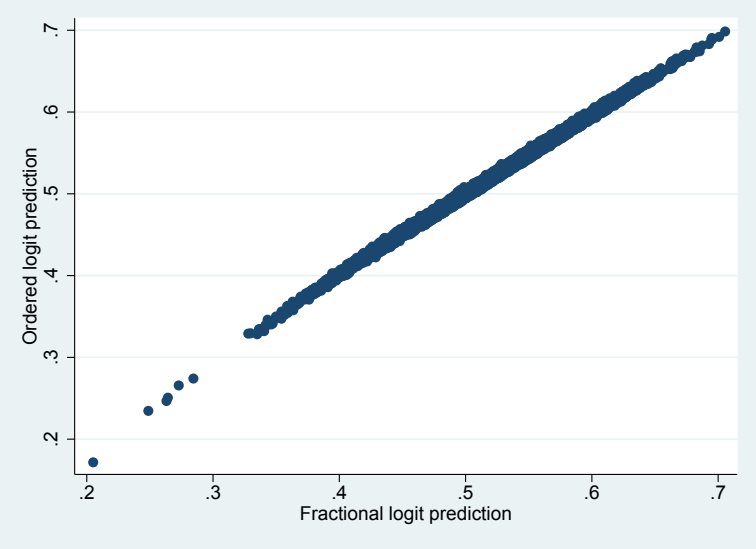
Because of this paper's focus on average partial effects, a more suitable version of the cardinalized ordered probit adopts the correlated random effects strategy used before.[12] This means implementing a pseudo-likelihood estimator for a pooled dataset with x_{it} and $\bar{x}_i$ as regressors. Figure 4 plots the fitted CEF values of the ordered logit CRE model against those of the logit CRE model, using

[11] van Praag and Ferrer-i-Carbonell (2008) derive expressions for $\mathrm{E}(y_{it}^*|y_{it} = k)$, assuming a normal latent model and thresholds derived such that the average predicted probabilities for each category fit the marginal distribution of the ordered outcome. Regressing these terms on x_{it} provides then another type of cardinalization.

[12] In fact, the ordered logit model is a restricted version of a sequence of binary logit models for cumulative outcomes such as $\Pr(y_{it} \geq k/K|x_{it},\alpha_i)$).

SAH coefficients for the male subsample. Clearly, the two are quite similar; a linear fit has an R-squared of 99.7%. This speaks against using the ordered logit in this case, as it requires estimation of additional threshold parameters, whereas the fractional logit model is more parsimonious.

Figure 4. *Comparison of ordered logit and fractional logit predictions for SAH, male subsample.*



3.4.5 Summary

Table 2 summarizes the features of the nine different estimators for SAH and HSAT. All of these estimators have their pros and cons ex-ante, and it is a matter of judgement which weight one should put on the various factors.

For example, if the true mean conditional on x_{it} and α_i is logit (or cardinalized ordered logit), it is in general not the case that the marginal models remain in the logit class. However, assuming logit type marginal models leads to predictions inside the unit interval, allows to compute average partial effects and does not require conditional independence for estimation. The non-linear fixed effects estimators do not restrict the distribution of the unobserved heterogeneity. While they estimate the structural parameters of the true conditional mean consistently under the assumption of conditional independence, they do not allow to estimate average partial effects.

OLS gives predictions outside of the unit interval. For example, in the linear fixed effects model, consider individual units such that $\bar{y}_i = 0$. In this case, we know that $\bar{x}_i'\hat{\beta} + \hat{\alpha}_i = 0$, and unless x_{it} is constant for i, there must be at least one time period with $x_{it}'\hat{\beta} + \hat{\alpha}_i < 0$. Negative predicted values preclude the computation of average semi-elasticities, or odds ratios, something that might

be of interest in some applications.

Table 2. Estimators for $\mathrm{E}(y_{it}|x_{it},\alpha_i) = \Lambda(x_{it}'\beta + \alpha_i)$

	$G(\,) \in (0,1)$	$f(\alpha_i\|x_i)$ unrestricted	β	APEs	IPP	Allows corr. t,s
Marginal models						
Pooled logit	yes	no	no	yes	no	yes
Logit CRE	yes	no	no	yes	no	yes
Pooled ordered logit	yes[1]	no	no	no[3]	no	yes
Ologit CRE	yes	no	no	no[3]	no	yes
Pooled OLS	no	no	no	yes[4]	no	yes
Conditional models						
Logit DV (≡Blogit DV)	yes	yes	yes[2]	yes[2]	yes	yes
FE Blogit (CML)	yes	yes	yes	yes[5]	no	no
FE OLS	no	yes	no	yes[4]	no	yes
FE Ologit (CML)	yes[1]	yes	yes	no	no	no

[1] Assuming an appropriate cardinalization as $0, 1/K, \ldots, K$.

[2] Subject to the incidental parameters problem.

[3] So far lack of analytical results and/or simulation evidence.

[4] Best linear approximation; Slopes equal APE if regressors are normally distributed.

[5] Only average *relative* partial effects.

Based on Table 2, the logit DV model has a number of attractive properties if it were not for the incidental parameters problem. As described in Section 3.4.3 above, there is some indication that the IPP is not as severe for the fractional response model, in particular as the number of categories increases. Also, previous simulation studies have shown that bias tends to be smaller for APE estimates than for slopes (e.g., Kwak, Martin and Wooldridge, 2018). These considerations notwithstanding, it requires by far the longest computating time among all models considered in Table 2.

4 Results

4.1 Overview

Tables A1-A3 in the Appendix contain all the estimated coefficients for the two subjective health outcomes HSAT and SAH. They all consist of four panels, with results for men in the upper part and results for women in the lower part, and HSAT on the left and SAH on the right. Table A1 shows coefficients for the the fractional logit models, Table A2 those for the linear models and Table

A3 those for the ordered logit models. In each case, the first column has results for the pooled model that assumes that individual specific unobserved heterogeneity is independent of regressors, and is inconsistent otherwise. The models following on the right adjust for correlated individual effects in one way or the other.

The unadjusted, pooled coefficients of "retired" are negative and statistically significant in all estimated models, for men and women and both SAH and HSAT. Hence, for a given age and year, those having retired report worse subjective health than those still working. This negative retirement "effect" is reinforced in the case of early retirement (defined here a being retired and younger than 60).

The sign of the "retired" coefficient is reversed in models that control for individual specific heterogeneity. Taking the point estimates at face value, there is a positive effect of retirement on subjective health throughout. The effect of early retirement, being the sum of the retirement coefficient and the "early" coefficient, is estimated to be close to zero in most cases. In the fixed effects type models, income has no effect on subjective health. Being married has a positive effect on health for women but not for men.

Clearly, it is not possible to compare magnitudes of coefficients across the three models, logit, linear and ordered logit. It is only in the linear model that coefficients provide provide direct estimates of average partial effects, albeit potentially biased ones. Similarly, among the non-linear models, the conditional estimation results identify structural parameters, whereas the marginal models do not. On way, to make results comparable is to state them in terms of relative coefficients, or trade-off ratios. Another one is to convert coefficients into average partial effects.

4.2 Partial effects

In the context of models for non-negative outcomes, such as fractional response or exponential regression models for counts, it is important to distinguish between average partial effects (APE) and average relative partial effects (ARPE). For fractional logit and Poisson models, it is particularly simple to obtain ARPEs (or expected semi-elasticities). In the log-linear Poisson model, $G(x_{it}'\beta + \alpha_i) = \exp(x_{it}'\beta + \alpha_i)$ and

$$\mathrm{E}\left[\frac{\partial \ln G(x_{it}'\beta + \alpha_i)}{\partial x_{it}}\right] = \beta$$

In the logit case (see Kitazawa, 2012),

$$\mathrm{E}\left[\frac{\partial \ln \Lambda(x_{it}'\beta + \alpha_i)}{\partial x_{it}}\right] = \beta(1 - \mathrm{E}(\Lambda_{it}))$$

Hence the logit ARPE is simply a fraction of β. Importantly, to estimate $\mathrm{E}(\Lambda_{it})$ one can use the grand mean of y_{it}, $\bar{y}$, and does not require estimates of α_i, which would not be available after conditional maximum likelihood. Also, if the regressors x are measured in log scales, the above quantities provide average response elasticities.

Unfortunately, no such simple expressions exist for the cardinalized ordered logit, the linear regression or the marginal logit models. For OLS, considering

$$ARPE = \mathrm{E}\left[\frac{\beta}{x_{it}'\beta + \alpha_i}\right]$$

is ill-defined, because the denominator can be negative. From the marginal logit, we obtain estimates $\tilde{\beta} \approx \beta_\alpha$, which have been shown above to be useful for obtaining APEs, but average partial effects $ARPE = \beta(1 - \mathrm{E}(\Lambda_{it})$ require the structural parameters β, and typically, $|\tilde{\beta}| < |\beta|$.

Hence, there is no entirely satisfying solution to comparing results from all models on a common scale. Since estimated APEs are readily available for a larger number of models, these are given, together with 95% confidence intervals, in Table 3. The Table also includes CML logit results, where the ARPE has been scaled by the mean outcome, $\bar{y}$. This is not the same as the APE, but corresponds to what Wooldridge (2002) refers to as "partial effect at the average", or PEA. Since the Table is intended for illustrative purposes, results are given for a single coefficient, that of being `Retired`, and a single outcome variable, Satisfaction with health, and for the male subsample only.

Table 3. Average Partial Effect of `Retired` on Satisfaction with Health[1]

	APE	95% CI
OLS	-0.164	(-0.294 , -0.034)
OLS with fixed effects	0.119	(-0.015 , 0.255)
Logit[2]	-0.165	(-0.363 , 0.011)
CRE logit[2]	0.123	(-0.100 , 0.354)
Logit DV[2]	0.113	(-0.022 , 0.248)
Logit CML[3]	0.134	(-0.018 , 0.285)
Cardinalized ologit[2]	-0.145	(-0.273 , -0.017)
Cardinalized CRE ologit[2]	0.098	(-0.038 , 0.235)

[1] Satisfaction with health is measured on a 0-10 scale; results are for the male subsample.

[2] Confidence intervals computed using Ham and Woutersen (2019) method.

[3] Partial effect at the average.

The main takeaway from Table 3 is that APEs tend to mirror results seen for the coefficients themselves: it matters a lot whether the model allows for correlated unobserved heterogeneity or not, but once it does, quite different approaches lead to qualitatively similar conclusions. In this particular application, the similarity of findings also includes models where departures could be expected on theoretical grounds, e.g. for the logit DV because of the incidental parameters problem, for the logit CML because APEs and PEAs usually differ, and for the marginal (and cardinalized) ologit models, where neither simulation studies nor prior theoretical results on APEs are available so far. In any case, the DiD effects are small overall, with APEs of around 0.12 on the 0-10 HSAT scale, which is roughly compatible with the naive trend-comparisons of Figure 1A. The coefficients have p-values between 0.05 and 0.1, and this is also reflected in the 95% confidence intervals for the APEs, which include zero throughout.

4.3 Number of doctor visits

Results for doctor visits are given in Table A4. The dependent variable is the self-reported number of visits during the three months period preceding the day of the survey. Since this is a count without a well-defined upper bound, it would be inappropriate to convert it into a fractional reponse. Instead, exponential regression models are implemented by means of Poisson pseudo maximum likelihood estimation, using pooled and CRE marginal models and a fixed effects model, respectively.[13] Standard errors are clustered at the individual level. The number of observations is slightly lower in the fixed effects models (7074 as compared to 7326 for the male subsample; 6216 as compared to 6318 for the female subsample) because they drop perfectly predicted observations, in this case individuals for which the number of doctor visits is zero in each of the six years.

Again, results from the CRE model and from the fixed effects model are very similar, both in terms of point estimates and estimated precision. For men, retirement leads to a reduction in the number of visits by 12 percent, for women by 17 percent. These effects are statistically significant. For men, there is an additional negative effect of early retirement, but not so for women. Also, in each case, pooled models that do not allow for correlated unobserved heterogeneity in the underlying propensity of doctor visits lead to quite different results than the DiD approach, in this case essentially zero effects. Thus there is evidence that retirees are negatively selected with regards to health care utilization.

[13] In the Poisson model, conditional maximum likelihood and the dummy variables approach give the same estimator, and there is therefore no incidental parameters problem.

5 Concluding remarks

When modelling health care utilization, e.g., the number doctor visits, there is a broad consensus in the literature that count data models, such as Poisson quasi-likelihood estimators, should be used. In the case of subjective measures of health, there is much less agreement on how to proceed. The literature has oscillated between advocating the use of the linear regression model on one side, and ordered logit or probit models on the other. This paper puts forward the idea of a middle ground, fractional response models, that account for the bounded nature of the outcomes while providing straightforward estimators of average partial effects. Importantly, simple panel data extensions for models with correlated time-invariant unobserved heterogeneity are available, allowing for the application of these estimators in the context of difference-in-differences models.

Regarding the substantive conclusions of the paper, different identification strategies led to opposite conclusions regarding the effect of retirement on subjective health. When identification comes from differential time trends in health, the effect of retirement is positive (although statistically significant only for men and HSat, not for SAH). When levels are compared instead, like in simple pooled, marginal models, the estimated effect of retirement, conditional on age and year, is negative, suggesting that there is negative selection into retirement.

Given the longitudinal, DiD research design employed here, results are remarkably robust to the specific model choice (under the proviso of allowing for correlated individual effects). The logit pooled likelihood estimator with means included is simple and performs well in this application.

Compliance with Ethical Standards

Funding: This study was funded by the Swiss National Science Foundation
(grant number 100018_178874/1)

The author declares that he has no conflict of interest.

References

Ai, C. and E. Norton (2003) Interaction terms in logit and probit models, *Economics Letters*, 80, 123129.

Angrist, J. D. and S. Pischke (2008) *Mostly Harmless Econometrics: An Empiricists Companion*, Princeton, University Press.

Athey, S. and G. Imbens (2006), Identification and inference in nonlinear difference-in-differences models, *Econometrica*, 74, 431-497.

Baetschmann, G., K.E. Staub and R. Winkelmann (2015) Consistent estimation of the fixed effects ordered logit model, *Journal of the Royal Statistical Society* A, 178, 685-703.

Bertrand, M., E. Duflo and S. Mullainathan (2004) How much should we trust differences-in-differences estimates?, *Quarterly Journal of Economics*, 119(1), 249-275.

Bloemen, H., S. Hochguertel and J. Zweerink (2017) The causal effect of retirement on mortality: Evidence from targeted incentives to retire early, *Health Economics* 26, 204 - 218.

Blundell, R., M. Costa Dias, C. Meghir and J. van Reenen (2004) Evaluating the employment impact of a mandatory job search program, *Journal of the European Economic Association*, 2, 569-606.

Bond, T. and K. Lang (2019) The sad truth about happiness scales, *Journal of Political Economy*, 127, 1629-1640.

Chabé-Ferret, S. (2015) Analysis of the bias of matching and difference-in-differences under alternative earnings and selection processes, *Journal of Econometrics*, 185, 110-123.

Chamberlain, G. (1980) Analysis of covariance with qualitative data, *Review of Economic Studies* 47, 225-238.

Charles, K. (2004) Is retirement depressing? Labor force inactivity and psychological well-being in later life, *Research in Labor Economics*, 23, 26999.

Coca V. and K. Nink (2011) Arzneimittelverordnungen nach Alter und Geschlecht. In: Schwabe U., Paffrath D. (eds) *Arzneiverordnungs-Report 2011*, Springer, Berlin, Heidelberg.

Cramer, J.S. (2007) Robustness of logit analysis: unobserved heterogeneity and misspecified disturbances, *Oxford Bulletin of Economics and Statistics*, 69, 545-555.

Das, M. and A. van Soest (1999), A panel data model for subjective information on household income growth, *Journal of Economic Behavior & Organization*, 40, 409-426.

Dave, D., I. Rashad and J. Spasojevic (2008) The effects of retirement on physical and mental health outcomes, *Southern Economic Journal* 75, 497-523.

Eibich, P. (2015) Understanding the effect of retirement on health: Mechanisms and heterogeneity, *Journal of Health Economics* 43, 112.

Ferrer-i-Carbonell, A. and Frijters,P. (2004) How important is methodology for the estimates of the determinants of happiness? *Economic Journal*, 114, 641-659.

Hernaes, E., S. Markussen, J. Piggott, and O. L. Vestad (2013) Does retirement age impact mortality? *Journal of Health Economics* 32, 586598.

Insler, M. (2014) The health consequences of retirement, *Journal of Human Resources* 49, 195233.

Jones, A. and S. Schurer (2011) How does heterogeneity shape the socioeconomic gradient in health satisfaction? *Journal of Applied Econometrics* 26, 549-579.

Kerkhofs, M. and M. Lindeboom (1997) Age related health dynamics and changes in labour market status, *Health Economics* 6(4), 407423.

Kitazawa, Y. (2012) Hyperbolic transformation and average elasticity in the framework of the fixed effects logit model, *Theoretical Economics Letters*, 2, 192-199.

Kuhn, A. (2018) The complex effects of retirement on health, *IZA World of Labor* 2018: 430.

Kunz, J. S., K. E. Staub and R. Winkelmann (2019) Predicting fixed effects in panel probit models, Monash University Discussion Paper 1019.

Kwak, D., R.S. Martin, and J.M. Wooldridge (2018) The robustness of conditional logit for binary response panel data models with serial correlation, BLS Working Paper No 502.

Mazzonna, F. and F. Peracchi (2017) Unhealthy retirement? *Journal of Human Resources*, 52(1), 128-151.

Mosca I. and A. Barrett (2016) The impact of voluntary and involuntary retirement on mental health: evidence from older Irish adults, *Journal of Mental Health Policy and Economics*, 19, 33-44.

Nishimura, Y., M. Oikawa and H. Motegi (2018) What explains the difference in the effect of retirement on health? Evidence from global aging data, *Journal of Economic Surveys*, 32, 792-847.

Papke, L. and J. Wooldridge (2008) Panel data methods for fractional response variables with an application to test pass rates, *Journal of Econometrics*, 145, 121-133.

Picchio, M. and J. van Ours (2018) The causal effect of retirement on health and happiness, Tinbergen Institute Discussion Paper, Rotterdam.

Puhani, P. (2012) The treatment effect, the cross difference, and the interaction term in nonlinear "difference-in-differences" models, *Economics Letters*, 115, 85-87.

Ramalho, E.A. and J.J.S. Ramalho (2010) Is neglected heterogeneity really an issue in binary and fractional regression models? A simulation exercise for logit, probit and loglog models, *Computational Statistics & Data Analysis* 54(4): 987-1001.

Shim, M. J., Gimeno, D., Pruitt, S. L., McLeod, C. B., Foster, M. J., and Amick III, B. C. (2013) A systematic review of retirement as a risk factor for mortality. In Hoque, Nazrul, McGehee, Mary A., Bradshaw, Benjamin S. (Eds.) *Applied Demography and Public Health* (pp. 277-309). Springer Netherlands.

Stoker, T.M. (1986) Consistent estimation of scaled coefficients, *Econometrica* 54, 1461-1481.

Studer, R. and R. Winkelmann (2014) Reported happiness, fast and slow, *Social Indicators Research*, 117, 1055-1067.

Van der Heide, I., R. van Rijn, S. Robroek, A. Burdorf, and K. Propper (2013) Is retirement good for your health? A systematic review of longitudinal studies, *BMC Public Health*, 13, 1180.

van Praag, B.M.S. and A. Ferrer-i-Carbonell (2008) *Happiness Quantified: A Satisfaction Calculus Approach*, Oxford University Press.

Woutersen, T. and J. Ham (2019) Confidence sets for continuous and discontinuous functions of parameters, mimeo., University of Arizona.

Table 1: Descriptive Statistics for Estimation Sample

-> retired = 0 (NT=6,822)				-> retired = 1 (NT=6,822)			
Variable	Mean	SE	Min Max	Mean	SE	Min	Max
year	2004.0	(0.063)	1994 2014	2007.0	(0.063)	1997	2017
age	60.53	(0.050)	51 76	63.53	(0.050)	54	79
educ	12.07	(0.033)	7 18	12.06	(0.033)	7	18
fHSAT	0.613	(0.003)	0 1	0.606	(0.003)	0	1
fSAH	0.531	(0.003)	0 1	0.518	(0.003)	0	1
male	0.536	(0.006)	0 1	0.536	(0.006)	0	1
#doctorco	3.511	(0.064)	0 80	3.219	(0.050)	0	90
full-time	0.456	(0.006)	0 1				
part-time	0.543	(0.006)	0 1				
married	0.799	(0.004)	0 1	0.782	(0.004)	0	1
loghhinc	10.41	(0.006)	7.82 12.93	10.21	(0.007)	-0.693	13.24
early				0.163	(0.004)	0	1

(N = 2,274, T = 6) Source: GSOEP 1994-2017

Appendix

Table A1. Binomial Logit Models for Health Satisfaction and Self-assessed Health

Outcome:	Satisfaction with health				Self-assessed Health			
	(1) Pooled	(2) Means	(3) CML	(4) DV	(5) Pooled	(6) Means	(7) CML	(8) DV
Panel a): Men								
retired	-0.071**	0.051*	0.056*	0.057*	-0.095***	0.030	0.032	0.033
	(0.028)	(0.029)	(0.032)	(0.033)	(0.027)	(0.028)	(0.030)	(0.032)
early	-0.097*	-0.051	-0.050	-0.051	-0.141**	-0.031	-0.035	-0.036
	(0.058)	(0.041)	(0.045)	(0.046)	(0.061)	(0.045)	(0.050)	(0.052)
loghhinc	0.248***	0.030	0.037	0.037	0.230***	0.037	0.045	0.047
	(0.035)	(0.040)	(0.047)	(0.048)	(0.035)	(0.031)	(0.035)	(0.036)
married	-0.127**	0.174*	0.188*	0.191*	-0.149**	0.156*	0.161*	0.168*
	(0.059)	(0.095)	(0.105)	(0.107)	(0.059)	(0.089)	(0.094)	(0.098)
N	7326	7326	72960	7326	7326	7326	29208	7326
Panel b) : Women								
retired	-0.080***	0.032	0.027	0.028	-0.052*	0.053*	0.053	0.055
	(0.030)	(0.031)	(0.034)	(0.035)	(0.030)	(0.030)	(0.032)	(0.034)
early	-0.110*	-0.026	-0.018	-0.018	-0.149**	-0.037	-0.040	-0.042
	(0.057)	(0.041)	(0.046)	(0.046)	(0.058)	(0.040)	(0.044)	(0.046)
loghhinc	0.237***	0.041	0.047	0.048	0.259***	0.044	0.052	0.055
	(0.039)	(0.038)	(0.042)	(0.043)	(0.040)	(0.038)	(0.042)	(0.044)
married	0.062	0.048	0.046	0.046	0.047	0.119	0.125	0.131
	(0.055)	(0.081)	(0.086)	(0.087)	(0.056)	(0.096)	(0.101)	(0.105)
N	6318	6318	62940	6318	6318	6318	25056	6318

All regressions include year dummies * p<0.10, ** p<0.05, *** p<0.01

Table A2. Linear Fixed Effects Models for Health Satisfaction and Self-assessed Health

Outcome:	Satisfaction with health		Self-assessed Health	
	(1)	(2)	(3)	(4)
	Pooled	Fixed Effects	Pooled	Fixed Effects
Panel a): Men				
retired	-0.165**	0.120*	-0.0958***	0.0281
	(0.0661)	(0.0691)	(0.0265)	(0.0279)
early	-0.263*	-0.109	-0.140**	-0.0266
	(0.142)	(0.0998)	(0.0599)	(0.0440)
loghhinc	0.573***	0.0764	0.220***	0.0384
	(0.0817)	(0.0958)	(0.0336)	(0.0300)
married	-0.293**	0.400*	-0.144**	0.153*
	(0.141)	(0.222)	(0.0587)	(0.0869)
N	7326	7326	7326	7326
Panel b): Women				
retired	-0.184***	0.0585	-0.0514*	0.0475
	(0.0695)	(0.0730)	(0.0295)	(0.0298)
early	-0.284**	-0.0409	-0.148***	-0.0354
	(0.139)	(0.0992)	(0.0571)	(0.0391)
loghhinc	0.556***	0.0997	0.254***	0.0478
	(0.0910)	(0.0904)	(0.0393)	(0.0377)
married	0.151	0.105	0.0478	0.118
	(0.132)	(0.192)	(0.0562)	(0.0944)
N	6318	6318	6318	6318

Table A3. Ordered Logit Models for Health Satisfaction and Self-assessed Health

Outcome:	Satisfaction with health			Self-assessed Health		
	(1)	(2)	(3)	(4)	(5)	(6)
	Pooled	Means	CML	Pooled	Means	CML
Panel a): Men						
retired	-0.126**	0.0854	0.174*	-0.190***	0.0252	0.0972
	(0.0564)	(0.0592)	(0.102)	(0.0576)	(0.0602)	(0.110)
early	-0.210*	-0.106	-0.140	-0.296**	-0.0703	-0.0552
	(0.121)	(0.0844)	(0.131)	(0.132)	(0.0983)	(0.150)
loghhinc	0.521***	0.0563	0.104	0.485***	0.0840	0.133
	(0.0722)	(0.0916)	(0.137)	(0.0770)	(0.0645)	(0.101)
married	-0.271**	0.254	0.585*	-0.324**	0.319*	0.721*
	(0.121)	(0.197)	(0.309)	(0.128)	(0.184)	(0.368)
N	7326	7326	22536	7326	7326	8514
Panel b) : Women						
retired	-0.135**	0.0435	0.0795	-0.115*	0.1000	0.175
	(0.0593)	(0.0634)	(0.106)	(0.0633)	(0.0642)	(0.114)
early	-0.252**	-0.0591	-0.0416	-0.313**	-0.0783	-0.109
	(0.117)	(0.0830)	(0.133)	(0.122)	(0.0841)	(0.139)
loghhinc	0.509***	0.0462	0.120	0.534***	0.0588	0.163
	(0.0805)	(0.0753)	(0.119)	(0.0854)	(0.0817)	(0.131)
married	0.0886	0.120	0.124	0.0912	0.341*	0.381
	(0.112)	(0.158)	(0.248)	(0.119)	(0.198)	(0.327)
N	6318	6318	19674	6318	6318	7476

All regressions include year dummies; * p<0.10, ** p<0.05, *** p<0.01

Table A4. Poisson Models for Number of Doctor Visits

Outcome:	Number of Doctor Visits		
	(1)	(2)	(3)
	Pooled	Means	Fixed Effects
Panel a): Men			
retired	0.0215	-0.123**	-0.123**
	(0.0440)	(0.0553)	(0.0539)
early	0.0249	-0.143**	-0.134**
	(0.0694)	(0.0669)	(0.0681)
loghhinc	-0.0293	-0.0485	-0.0412
	(0.0458)	(0.0760)	(0.0700)
married	0.162**	-0.122	-0.125
	(0.0793)	(0.1480)	(0.1690)
N	7326	7326	7074
Panel b): Women			
retired	-0.0295	-0.172***	-0.172***
	(0.0430)	(0.0609)	(0.0605)
early	0.138**	-0.0065	-0.0134
	(0.0686)	(0.0648)	(0.0659)
loghhinc	-0.0868*	-0.117*	-0.109
	(0.0510)	(0.0661)	(0.0673)
married	-0.111	-0.0431	-0.0399
	(0.0692)	(0.1140)	(0.1240)
N	6318	6318	6216

Figure A1

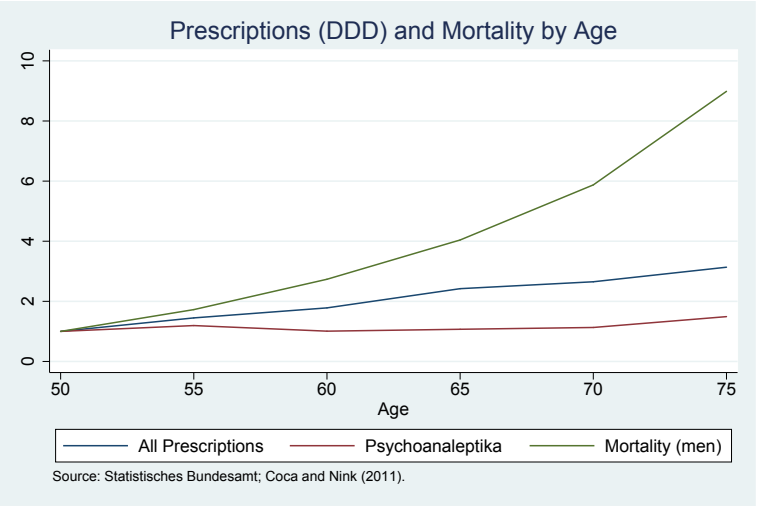


Figure A2