Methodological Report on Kaul and Wolf's Working Papers on the Effect of Plain Packaging on Smoking Prevalence in Australia and the Criticism Raised by OxyRomandie

Prof. Dr. Ben Jann

University of Bern, Institute of Sociology, Fabrikstrasse 8, CH-3012 Bern

ben.jann@soz.unibe.ch

March 10, 2015

Contents

1 Introduction

On February 16, 2015 I was asked by Vice President Prof. Schwarzenegger of University of Zurich to provide a methodological assessment of two working papers by Prof. Kaul and Prof. Wolf on the effect of plain packaging on smoking prevalence in Australia and the criticism raised against these working papers by OxyRomandie.

The materials on which I base my assessment include:

- Working paper no. 149 on "The (Possible) Effect of Plain Packaging on the Smoking Prevalence of Minors in Australia: A Trend Analysis" by Ashok Kaul and Michael Wolf (Kaul and Wolf 2014b).
- Working paper no. 165 on "The (Possible) Effect of Plain Packaging on Smoking Prevalence in Australia: A Trend Analysis" by Ashok Kaul and Michael Wolf (Kaul and Wolf 2014a).
- Letter by Pascal A. Diethelm on behalf of OxyRomandie to the President of University of Zurich, including the annex "Errors and issues with Kaul and Wolf's two working papers on tobacco plain packaging in Australia", dated January 29, 2015 (provided by Prof. Schwarzenegger).
- Public reply to the letter of Pascal A. Diethelm, including a reply to the annex of the letter of Pascal A. Diethelm, by Ashok Kaul and Michael Wolf, dated February 11, 2015 (provided by Prof. Schwarzenegger).
- Letter by Pascal A. Diethelm on behalf of OxyRomandie to the President of University of Zurich, including the document "Comments on Kaul and Wolf's reply to our Annex", dated February 19, 2015 (provided by Prof. Schwarzenegger).
- Forthcoming comment on the "Use and abuse of statistics in tobacco industry-funded research on standardised packaging" by Laverty, Diethelm, Hopkins, Watt and Mckee (Laverty et al. forthcoming) (provided by Prof. Schwarzenegger).

- Monthly data on sample sizes and smoking prevalences, January 2001 to December 2013, for minors and adults, as displayed in Figures 1 and 2 in Kaul and Wolf (2014a,b) (provided by Prof. Schwarzenegger).

Prof. Schwarzenegger offered reimbursement of my services at standard rates by my university for external services (capped at a total of CHF 8000.-), which I accepted. Furthermore, I agreed with Prof. Schwarzenegger that my report will be made public. I hereby confirm that I have no commitments to tobacco industry, nor do I have commitments to anti-tobacco institutions such as OxyRomandie. Moreover, apart from this report, I have no commitments to the University of Zurich.

Below I will first comment on the potential of the data used by Kaul and Wolf (2014a,b) for identifying causal effects. I will then provide a reanalysis of the data. Based on this reanalysis and my reading of the above documents, I will then comment on the criticism raised by OxyRomandie against the working papers by Kaul and Wolf. I will conclude my report with some remarks on whether I think the working papers should be retracted or not.

2 General remarks on the potential of the given data to identify a causal effect of plain packaging

In their working papers, Kaul and Wolf analyze monthly population survey data on smoking prevalence of adults and minors in Australia.[1] The time span covers 13 years from January 2001 to December 2013. Plain packaging, according to Kaul an Wolf, was introduced in December 2012, so that there are 143 months of pre-treatment observations and 13 months of treatment-period observations (assuming that plain packaging, the treatment, was introduced on December 1).

In terms of experimental-design language this is called an interrupted time-series design without control group. It is a quasi-experimental design as there is no randomization of the treatment. In general, it is difficult to draw causal conclusions from such a design, as it remains

[1]The data appear to stem from weekly surveys, but Kaul and Wolf base their analyses on monthly aggregates. It is not known to me whether Kaul and Wolf had access to the individual level weekly data or only to the monthly aggregates.

unknown how the counter-factual time trend would have looked. Kaul and Wolf assume a linear time trend and hence base their analyses on a linear fit to the pre-treatment data.[2] Deviations from the extrapolation of the linear fit into the treatment period are then used to identify the effect of the treatment.[3]

The assumption behind such an approach is that the time trend would have continued in the same linear fashion as in the pre-treatment period if there had been no treatment. The problem is that it is hard to find truly convincing arguments for why this should be the case (no such arguments are offered by Kaul and Wolf). As argued in the paper by Laverty et al. (forthcoming) it may be equally plausible that the trend would level off (e.g. because the trend has to level off naturally once we get close to zero or because the pre-treatment declines were caused by a series of other tobacco control treatments), or that the trend would accelerate (e.g. due to business cycles or other factors that might influence tobacco consumption). The point is: we simply do not know how the trend would have been like without the treatment.

A more meaningful design would be an interrupted time-series with control group or difference-in-differences. For example, such a design could be realized if the treatment were implemented only in certain states or districts, but not in others, so that the states or districts without treatment could be used to identify the baseline trend (the treatment effect is then given as the difference between the trend in the control group and the trend in the treatment group). Even though such a design would still be quasi-experimental (i.e. no randomization), one could certainly make more credible causal inferences with such a design than using a simple time-series. Such a pseudo-control group could be considered a reasonable counterfactual if the pre-treatment trends and other significant factors (e.g. business cycles) were similar between the treatment and pseudo-control groups.

---

[2] In the paper on minors, Kaul and Wolf use a linear fit based on all data, including the treatment-period observations. This is problematic because in this case the linear fit will be biased by the treatment effect, resulting in treatment-period residuals that are biased towards zero. The "robustness check" in Section 3.4 of their paper, however, suggests that using only pre-treatment observations for the linear fit does not change their conclusions. In the paper on adults, Kaul and Wolf consistently base the linear fit only on pre-treatment observations.

[3] Kaul and Wolf also compare the mean of the 12 residuals after December 2012 to the mean of the last 12 residuals before December 2012. A minor issue with this approach is that the pre-treatment residuals will tend to be underestimated due to the leverage of the pre-treatment observations with respect to the linear fit.

Of course, the gold standard would be a true randomized experiment with plain packaging introduced in some regions but not in others (though causal inference can still be limited in such a design, for example, due to spillover between regions). What I am trying to say is that causal inference has high demands on research design (and implementation and data quality) and that the design on which the working papers by Kaul and Wolf are based on is not particularly strong. Kaul and Wolf cannot be blamed for this as there might have been no better data, but they could have been more careful in pointing out the weaknesses of their design.

A second aspect, also mentioned in paper by Laverty et al. (forthcoming), is that given the nature of the treatment and the outcome of interest, a treatment period of one year might be too short for the effect to fully unfold. Smoking habits are hard to change, especially with "soft" measures such as plain packaging, and it would be surprising to see a strong and immediate effect. Such an effect would only be expected if accessibility were suddenly restricted (e.g. restaurant bans) or if prices suddenly increased dramatically. The argument, I think, is equally true for existing smokers and those taking up smoking. The idea of plain packaging, as far as I can see, is to influence consumption behavior by changing the "image" of tobacco brands and smoking in general. Such an approach probably has a very slow and subtle effect that might not be observed in just one year. Moreover, although in the meantime we could probably extend the time-series with additional months, increasing the treatment observation period does not really help, as the basic design problem discussed above becomes worse the longer the treatment period. That is, the longer the follow-up, the less convincingly we can argue that the extrapolation of the pre-treatment trend provides a valid estimate of the counterfactual development had there been no treatment.

This argument about the treatment period being too short is specific to the topic at hand. It is not a general design issue. Hence, my argument is not based on theoretical reasoning but on common sense and background knowledge about addiction and human behavior. The argument appears plausible to me, but others might disagree or might even provide scientific evidence proving me wrong. I do not claim that my argument is right. But I do think that it is an issue that might have deserved some discussion in the papers by Kaul and Wolf.

Given that the estimates are based on survey data, other problems might be present. For example, samples might be non-representative (no specific information on sampling is given by Kaul and Wolf) and non-response or social-desirability bias or other measurement errors

might distort the data. Furthermore, the data analyzed by Kaul and Wolf has been aggregated from individual-level measurements and errors might have been introduced during this process (e.g. inadequate treatment of missing values).[4] From this point on, however, I ignore these potential problems, assuming that the data reflect unbiased estimates.

Finally, one could potentially verify the study running an identical time trend analysis using an alternative data set, such as sales figures from tobacco companies or monthly tax revenues from tobacco sales in addition to survey data.

3 A reanalysis of the data

I use Stata/MP 13.1, revision 19 Dec 2014, for all analyses below.[5] Preparation of the data is as follows:

```
. import delimited ../Data/prevMinors.txt, delim(" ") varnames(1) clear
(3 vars, 156 obs)
. generate byte sample = 1
. save tobacco, replace
file tobacco.dta saved
. import delimited ../Data/prevAdults.txt, delim(" ") varnames(1) clear
(3 vars, 156 obs)
. generate byte sample = 2
. append using tobacco
. lab def sample 1 "Minors" 2 "Adults"
. lab val sample sample
. lab var sample "Sample (1=minors, 2=adults)"
. lab var month "Month (1=January 2001)"
. forv i = 1/156 {
  2.     lab def month `i´ "`:di %tmMonCCYY tm(2001m1) + `i´ - 1´", add
  3. }
. lab val month month
. lab var observations "Sample size"
. lab var prevalence "Smoking prevalence"
. order sample month
. sort sample month
. save tobacco, replace
file tobacco.dta saved
```

[4] It is not entirely clear whether Kaul and Wolf received individual-level data and did the aggregation themselves or whether they received pre-aggregated data. If they did have access to the individual-level data, it is unclear to me why they used WLS on aggregate data instead of directly analyzing the individual-level data. Analyzing the individual-level data would be interesting as changing sample compositions could be controlled for or subgroup analyses could be performed.

[5] User packages "coefplot" (Jann 2014), "estout" (Jann 2005b), and "moremata" (Jann 2005a) are required to run the code below. In Stata, type "ssc install coefplot", "ssc install estout", and "ssc install moremata" to install the packages.

```
. describe
Contains data from tobacco.dta
  obs:           312
 vars:             4                          6 Mar 2015 17:17
 size:         4,056
-------------------------------------------------------------------------------
              storage   display    value
variable name   type    format     label      variable label
-------------------------------------------------------------------------------
sample          byte    %8.0g      sample     Sample (1=minors, 2=adults)
month           int     %8.0g      month      Month (1=January 2001)
observations    int     %8.0g                 Sample size
prevalence      double  %10.0g                Smoking prevalence
-------------------------------------------------------------------------------
Sorted by:  sample  month
```

3.1 Choice of baseline model

As mentioned above, Kaul and Wolf (2014a,b) use a two-step approach to analyze the data by first fitting a linear model to the pre-treatment data[6] and then investigating the (out of sample) residuals for the treatment period. This is acceptable, but in my opinion a simpler and more straightforward approach would be to directly estimate the treatment effect by including additional parameters in the model. Irrespective of whether we use a two-step or a one-step approach, however, we first have to choose a suitable baseline model. Kaul and Wolf use a weighted least-squares model (WLS) based on the aggregate data (where the weights are the sample sizes). Using WLS instead of ordinary least-squares (OLS) is appropriate because this yields essentially the same results as applying OLS to the individual data. This is illustrated by the following analysis using the data on minors (the results of the WLS model are identical to the results in Section 3.4 in Kaul and Wolf 2014b):

```
. use tobacco
. quietly keep if sample==1
. regress prevalence month if month<144 [aw=observations]
(sum of wgt is   3.8564e+04)
      Source |       SS       df       MS              Number of obs =     143
-------------+------------------------------           F(  1,   141) =   90.14
       Model |  .030136654     1  .030136654           Prob > F      =  0.0000
    Residual |  .047142285   141  .000334342           R-squared     =  0.3900
-------------+------------------------------           Adj R-squared =  0.3856
       Total |   .07727894   142  .000544218           Root MSE      =  .01829

------------------------------------------------------------------------------
  prevalence |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       month |  -.0003559   .0000375    -9.49   0.000      -.00043   -.0002818
       _cons |    .114086   .0028709    39.74   0.000     .1084105    .1197615
------------------------------------------------------------------------------
. expand observations
(41282 observations created)
```

[6] See also footnote 2.

```
. sort sample month
. by sample month: gen byte smokes = (_n<= round(observations*prevalence))
. regress smokes month if month<144
```

Source	SS	df	MS		
Model	8.12720289	1	8.12720289	Number of obs	= 38564
Residual	3182.40127	38562	.082526873	F(1, 38562)	= 98.48
				Prob > F	= 0.0000
Total	3190.52847	38563	.082735484	R-squared	= 0.0025
				Adj R-squared	= 0.0025
				Root MSE	= .28727

smokes	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
month	-.0003559	.0000359	-9.92	0.000	-.0004262	-.0002856
_cons	.114086	.0027466	41.54	0.000	.1087026	.1194694

We can see that WLS based on aggregated data and OLS based on the expanded individual-level data (which can be reconstructed here because the dependent variable is binary) yield identical point estimates and only differ trivially in standard errors. Given that the dependent variable is dichotomous, however, a more appropriate model for the data might be logistic regression (or Probit regression, which yields almost identical results as logistic regression, apart from scaling). Logistic regression, for example, has the advantage that effects level off once getting close to zero or one by construction, so that predictions outside 0 to 1 are not possible. Logistic regression can be estimated directly from the aggregate data (logistic regression for grouped data), yielding identical results as a standard individual-level Logit model. The output below shows the results and also provides a graph comparing the Logit fit and the WLS fit.

```
. use tobacco
. quietly keep if sample==1
. generate smokers = round(observations*prevalence)
. blogit smokers observations month if month<144
Logistic regression for grouped data              Number of obs   =      38564
                                                  LR chi2(1)      =      99.77
                                                  Prob > chi2     =     0.0000
Log likelihood = -11707.726                       Pseudo R2       =     0.0042
```

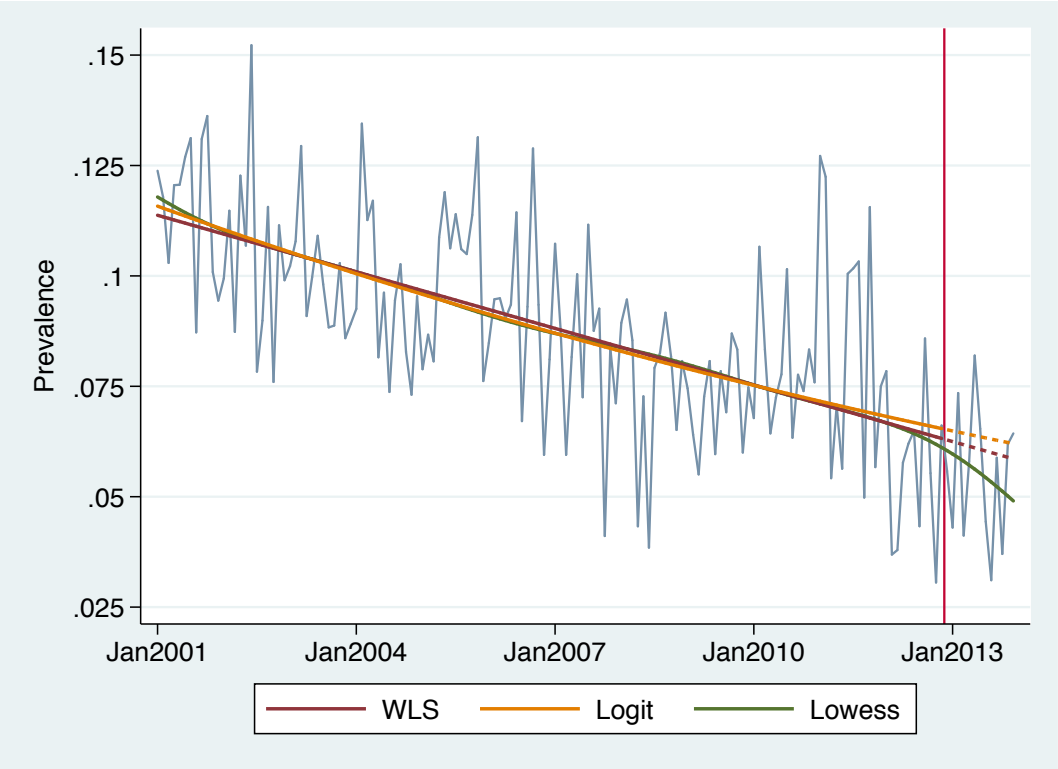
_outcome	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
month	-.0044098	.0004462	-9.88	0.000	-.0052844	-.0035352
_cons	-2.028496	.0317162	-63.96	0.000	-2.090659	-1.966334

```
. predict y_logit, pr
. generate r2_logit = (prevalence - y_logit)^2
. qui regress prevalence month if month<144 [aw=observations]
. predict y_WLS
(option xb assumed; fitted values)
. generate r2_WLS = (prevalence - y_WLS)^2
. summarize r2_* if month<144 [aw=observations]
```

Variable	Obs	Weight	Mean	Std. Dev.	Min	Max
r2_logit	143	38564	.0003292	.0005008	3.87e-09	.0030856
r2_WLS	143	38564	.0003297	.0005065	1.24e-09	.003152

```
. two (line prev    month, lc(*.6)) ///
```

```
>     (lowess prev  month, lw(*1.5) psty(p3)) ///
>     (line y_WLS   month if month<144,  lw(*1.5) psty(p2))  ///
>     (line y_logit month if month<144,  lw(*1.5) psty(p4))  ///
>     (line y_WLS   month if month>=144, lw(*1.5) psty(p2) lp(shortdash)) ///
>     (line y_logit month if month>=144, lw(*1.5) psty(p4) lp(shortdash)) ///
>     , legend(order(3 "WLS" 4 "Logit" 2 "Lowess") rows(1)) ///
>       xti("") yti(Prevalence) xline(143.5) ylab(.025(.025).15, angle(hor)) ///
>       xlab(1(36)145, valuelabel)
```



The two fits are almost identical, although the Logit fit is slightly curved. Also in terms of average squared residuals (weighted by sample size) the two fits are very similar (with slightly smaller squared residuals from the Logit model; see the second table in the output above). For comparison, a standard Lowess fit (unweighted[7]) is also included in the graph. It can be seen that the WLS fit and the Logit fit closely resemble the Lowess fit across most of the pre-treatment period.[8] From these results I conclude that both WLS and Logit with a simple linear

[7]Lowess in Stata does not support weights. A weighted local polynomial fit of degree 1 (linear) yields a very similar result if using a comparable bandwidth (not shown)

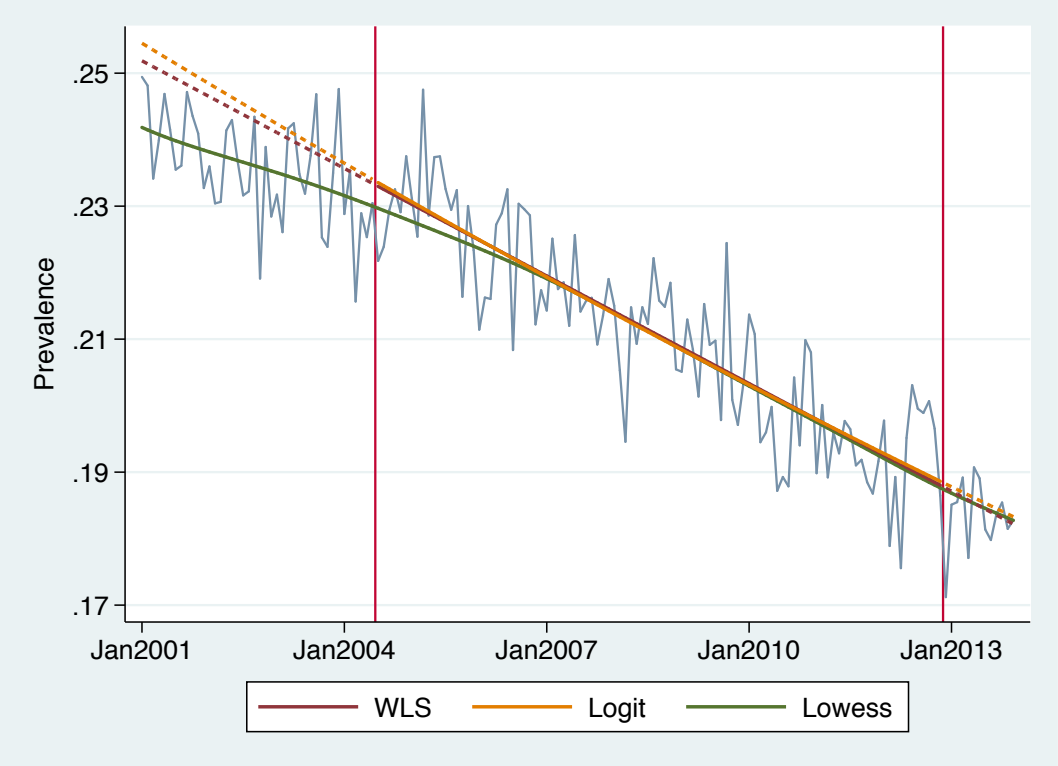
[8]Note that the Lowess fit uses all data including the treatment-period observations and that such nonparametric estimators are affected by boundary problems, hence the deviation at the beginning of the observation period and especially in the treatment period. Whether the drop of the curve in the treatment period is systematic will be evaluated below.

time-trend parameter provide a good approximation of the baseline trend in the pre-treatment period.

The next output and graph show a similar exercise for the adult data. An issue with the adult data is that a linear model does not fit the pre-treatment period very well. For example, a quadratic model indicates curvature (significant coefficient of month squared in the first table of the output below). Based on graphical inspection of a nonparametric smooth, Kaul and Wolf (2014a) decided to use only observations from July 2004 on to estimate the baseline trend in the pre-treatment period. For now I follow this approach, though later I consider how this decision impacted results.

```
. use tobacco
. quietly keep if sample==2
. generate monthsq = month^2
. regress prevalence month monthsq if month<144 [aw=observations]
(sum of wgt is   6.5227e+05)
      Source |       SS       df       MS              Number of obs =     143
-------------+------------------------------           F(  2,   140) =  317.43
       Model |  .036153612     2  .018076806           Prob > F      =  0.0000
    Residual |  .007972556   140  .000056947           R-squared     =  0.8193
-------------+------------------------------           Adj R-squared =  0.8167
       Total |  .044126168   142  .000310748           Root MSE      =  .00755

------------------------------------------------------------------------------
  prevalence |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       month |  -.0002346   .0000611    -3.84   0.000    -.0003555   -.0001138
     monthsq |  -1.04e-06   4.13e-07    -2.52   0.013    -1.86e-06   -2.26e-07
       _cons |   .2419662   .0018901   128.02   0.000     .2382293    .245703
------------------------------------------------------------------------------

. regress prevalence month if inrange(month,43,143) [aw=observations]
(sum of wgt is   4.5396e+05)
      Source |       SS       df       MS              Number of obs =     101
-------------+------------------------------           F(  1,    99) =  301.72
       Model |  .017559625     1  .017559625           Prob > F      =  0.0000
    Residual |  .005761653    99  .000058199           R-squared     =  0.7529
-------------+------------------------------           Adj R-squared =  0.7504
       Total |  .023321278   100  .000233213           Root MSE      =  .00763

------------------------------------------------------------------------------
  prevalence |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       month |  -.0004494   .0000259   -17.37   0.000    -.0005008   -.0003981
       _cons |   .2523039   .0024979   101.01   0.000     .2473475    .2572602
------------------------------------------------------------------------------

. predict y_WLS
(option xb assumed; fitted values)
. generate r2_WLS = (prevalence - y_WLS)^2
. generate smokers = round(observations*prevalence)
. blogit smokers observations month if inrange(month,43,143)
Logistic regression for grouped data              Number of obs   =     453961
                                                  LR chi2(1)      =     474.74
                                                  Prob > chi2     =     0.0000
Log likelihood = -233659.53                       Pseudo R2       =     0.0010

------------------------------------------------------------------------------
    _outcome |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       month |  -.0027059   .0001243   -21.76   0.000    -.0029496   -.0024622
       _cons |  -1.072061   .0118411   -90.54   0.000    -1.095269   -1.048853
------------------------------------------------------------------------------
```

```
. predict y_logit, pr
. generate r2_logit = (prevalence - y_logit)^2
. summarize r2_* if inrange(month,43,143) [aw=observations]
    Variable |      Obs      Weight        Mean   Std. Dev.       Min        Max
-------------+---------------------------------------------------------------------
       r2_WLS |      101      453961     .000057   .0000735   8.90e-08    .000374
    r2_logit |      101      453961    .0000569   .0000732   6.08e-09   .0003838
. two (line prev    month, lc(*.6)) ///
>     (lowess prev  month, lw(*1.5) psty(p3)) ///
>     (line y_WLS   month if inrange(month,43,143), lw(*1.5) psty(p2))    ///
>     (line y_logit month if inrange(month,43,143), lw(*1.5) psty(p4))    ///
>     (line y_WLS   month if month>=144, lw(*1.5) psty(p2) lp(shortdash)) ///
>     (line y_logit month if month>=144, lw(*1.5) psty(p4) lp(shortdash)) ///
>     (line y_WLS   month if month<43,   lw(*1.5) psty(p2) lp(shortdash)) ///
>     (line y_logit month if month<43,   lw(*1.5) psty(p4) lp(shortdash)) ///
>     , legend(order(3 "WLS" 4 "Logit" 2 "Lowess") rows(1)) ///
>       xti("") yti(Prevalence) xline(42.5 143.5) ylab(.17(.02).25, angle(hor)) ///
>       xlab(1(36)145, valuelabel)
```



Again, both WLS and Logit provide a very good approximation of the time trend, at least in the second part of the pre-treatment observation period (from around 2006). In terms of squared residuals both fits perform equally well (again with a tiny advantage for the Logit model; see fourth table in the output above). The WLS results (second table in the output above) are identical to the results reported by Kaul and Wolf (2014a, Equation 3.3).

3.2 Treatment effect estimation

3.2.1 Immediate (time-constant) treatment effect

The most straightforward approach to estimate the treatment effect of plain packaging is to apply the above models to all observations and include an indicator variable for the treatment. The treatment indicator is 0 for observations prior to December 2012 and 1 for observations from December 2012 on. The coefficient of the treatment indicator provides an estimate of the treatment effect, modeled as a parallel shift in the trend from December 2012 on (i.e. an immediate and time-constant treatment effect).[9] The results from such a model for minors and adults are as follows:

```
. use tobacco
. generate byte treat = month>=144
. generate smokers = round(observations*prevalence)
. preserve
. quietly keep if sample==1                          // => minors
. regress prevalence month treat [aw=observations]
(sum of wgt is   4.1438e+04)
```

Source	SS	df	MS		
Model	.043186903	2	.021593452		
Residual	.050089722	153	.000327384		
Total	.093276625	155	.000601785		

```
Number of obs =     156
F(  2,   153) =   65.96
Prob > F      =  0.0000
R-squared     =  0.4630
Adj R-squared =  0.4560
Root MSE      =  .01809
```

prevalence	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
month	-.0003558	.0000368	-9.67	0.000	-.0004285	-.0002831
treat	-.0051258	.0065024	-0.79	0.432	-.017972	.0077203
_cons	.1140834	.0028188	40.47	0.000	.1085145	.1196522

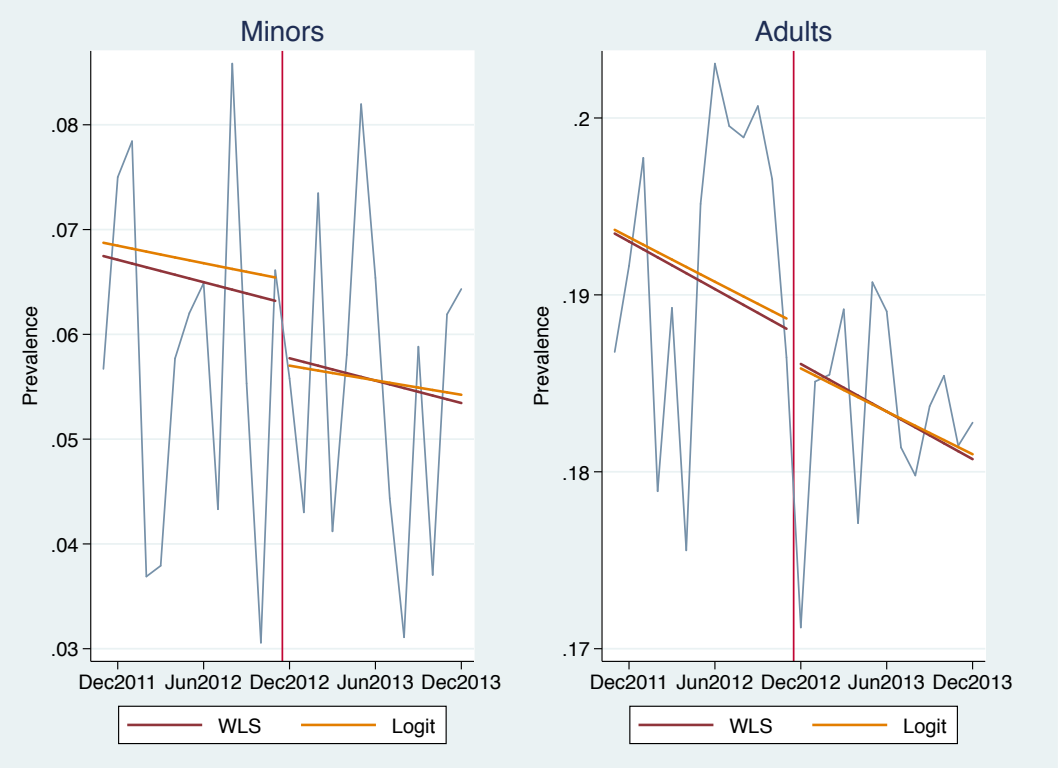
```
. predict y_WLS
(option xb assumed; fitted values)
. blogit  smokers observations month treat
Logistic regression for grouped data              Number of obs   =      41438
                                                  LR chi2(2)      =     146.56
                                                  Prob > chi2     =     0.0000
Log likelihood = -12325.285                       Pseudo R2       =     0.0059
```

_outcome	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
month	-.0044102	.0004461	-9.89	0.000	-.0052846	-.0035358
treat	-.1422188	.0925872	-1.54	0.125	-.3236864	.0392488
_cons	-2.028472	.0317117	-63.97	0.000	-2.090626	-1.966318

```
. predict y_logit, pr
. two (line prev    month, lc(*.6)) ///
>     (line y_WLS   month if month<144 , lw(*1.5) psty(p2))  ///
>     (line y_logit month if month<144 , lw(*1.5) psty(p4)) ///
```

[9] The estimated pre-treatment baseline trend in such a model can be affected by treatment-period observations because the model is not fully flexible. However, this effect is only minimal in the present case (compare the models below with the results in Section 3.1).

```
>     (line y_WLS   month if month>=144, lw(*1.5) psty(p2))  ///
>     (line y_logit month if month>=144, lw(*1.5) psty(p4)) ///
>     if month>=131, legend(order(2 "WLS" 3 "Logit") rows(1)) ///
>     xti("") yti(Prevalence) xline(143.5) ylab(.03(.01).08, angle(hor)) ///
>     xlab(132(6)156, valuelabel) xoverhangs ti(Minors) nodraw name(minors)

. restore, preserve

. quietly keep if sample==2 & month>=43               // => adults

. regress prevalence month treat [aw=observations]
(sum of wgt is   5.0666e+05)

      Source |       SS       df       MS              Number of obs =     114
-------------+------------------------------           F(  2,   111) =  230.95
       Model |  .025753273     2  .012876636           Prob > F      =  0.0000
    Residual |  .006188756   111  .000055755           R-squared     =  0.8063
-------------+------------------------------           Adj R-squared =  0.8028
       Total |  .031942028   113  .000282673           Root MSE      =  .00747

------------------------------------------------------------------------------
  prevalence |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       month |  -.0004484   .0000252   -17.82   0.000    -.0004983   -.0003985
       treat |  -.0015422   .0027153    -0.57   0.571    -.0069227    .0038383
       _cons |   .2522081   .0024292   103.82   0.000     .2473946    .2570217
------------------------------------------------------------------------------

. predict y_WLS
(option xb assumed; fitted values)

. blogit  smokers observations month treat

Logistic regression for grouped data              Number of obs   =     506657
                                                  LR chi2(2)      =     696.42
                                                  Prob > chi2     =     0.0000
Log likelihood = -258774.15                       Pseudo R2       =     0.0013

------------------------------------------------------------------------------
    _outcome |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       month |  -.0027001   .0001242   -21.73   0.000    -.0029435   -.0024566
       treat |  -.0158519   .0139325    -1.14   0.255    -.0431591    .0114553
       _cons |  -1.072587   .0118326   -90.65   0.000    -1.095779   -1.049396
------------------------------------------------------------------------------

. predict y_logit, pr

. two (line prev    month, lc(*.6)) ///
>     (line y_WLS   month if month<144 , lw(*1.5) psty(p2))  ///
>     (line y_logit month if month<144 , lw(*1.5) psty(p4)) ///
>     (line y_WLS   month if month>=144, lw(*1.5) psty(p2))  ///
>     (line y_logit month if month>=144, lw(*1.5) psty(p4)) ///
>     if month>=131, legend(order(2 "WLS" 3 "Logit") rows(1)) ///
>     xti("") yti(Prevalence) xline(143.5) ylab(.17(.01).20, angle(hor)) ///
>     xlab(132(6)156, valuelabel) xoverhangs ti(Adults) nodraw name(adults)

. restore

. graph combine minors adults, imargin(zero)
```



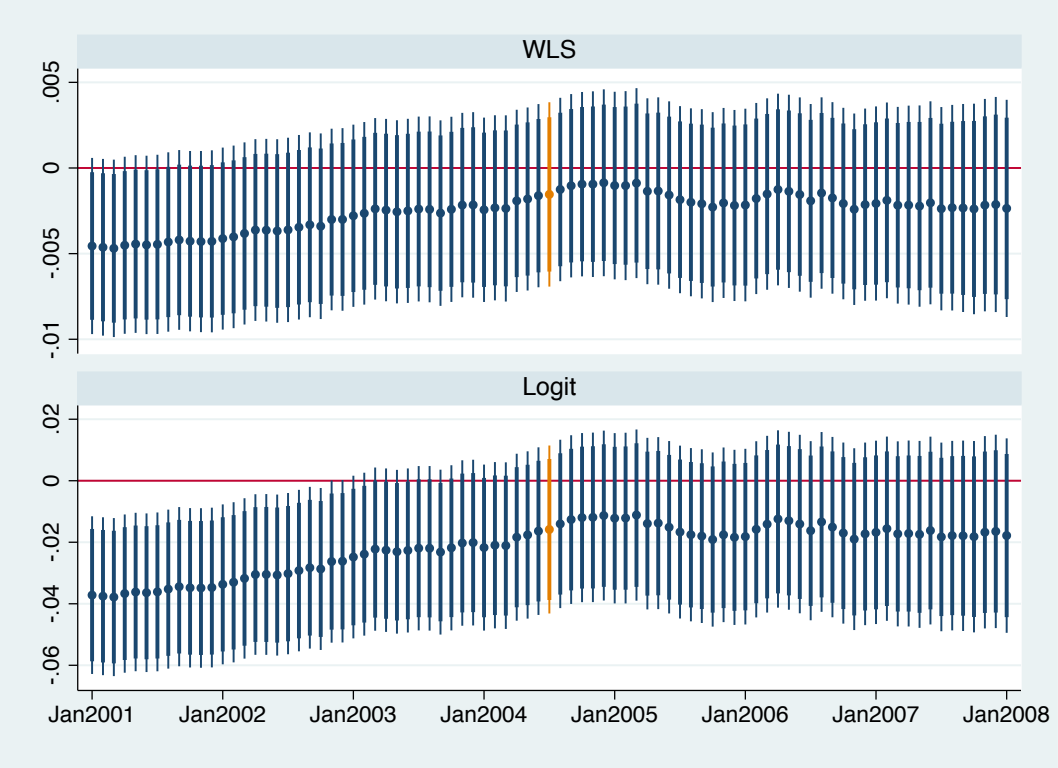
For minors the estimated treatment effect is about 0.5 percentage points (the first table in the output above), for adults the effect is about 0.15 percentage points (the third table in the output above), but none of these treatment effects are significant ($p = 0.432$ and $p = 0.125$ from WLS and Logit for minors; $p = 0.571$ and $p = 0.255$ from WLS and Logit for adults). The graph, zooming in on the last two years of the observation window, illustrates the effect as a parallel shift of the curves between November and December 2012.

Bound to a strict interpretation of significance tests (employing a usual 5% significance level), we would conclude from these results that there is no convincing evidence for an effect of plain packaging on smoking prevalence, neither for minors nor for adults and irrespective of whether we use two-sided tests or one-sided tests. However, if we employ a more gradual interpretation of statistical results without resorting to strict (and somewhat arbitrary) cutoffs, we can acknowledge that the effects at least point in the expected direction.[10] For example, using a one-sided test, the p-value from the logistic regression for minors is $p = 0.062$, which is not far from the conventional 5% level. To be fair, results from WLS, and results for adults, where statistical power is higher due to the larger sample sizes, are considerably less convincing.

[10]Expected in the sense that the purpose of introducing plain packaging was to reduce smoking prevalence.

As mentioned above, an issue with the results for adults is that the pre-treatment observations before July 2004 were excluded due to lack of fit of the linear baseline model. Using July 2004 as cutoff is an arbitrary decision that might favor result in one direction or the other. To evaluate whether the precise location of the cutoff affects our conclusions, we can run a series of models using varying cutoffs. The following graph shows how the effect evolves if we increase the cutoff in monthly steps from January 2001 (i.e. using all data) to January 2008 (where the WLS and Logit fits are essentially indistinguishable from the Lowess fit; see Section 3.1):

```
. use tobacco
. generate byte treat = month>=144
. generate smokers = round(observations*prevalence)
. quietly keep if sample==2
. forv i = 1/85 {
  2.     qui regress prevalence month treat if month>=`i´ [aw=observations]
  3.     mat tmp = _b[treat] \ _se[treat] \ e(df_r)
  4.     qui blogit  smokers observations month treat if month>=`i´
  5.     mat tmp = tmp \ _b[treat] \ _se[treat]
  6.     mat coln tmp = `i´
  7.     mat res = nullmat(res), tmp
  8. }
. coefplot (mat(res[1]), se(res[2]) drop(43) df(res[3]) ) ///
>          (mat(res[1]), se(res[2]) keep(43) df(res[3]) pstyle(p4)), bylabel(WLS) ///
>       || (mat(res[4]), se(res[5]) drop(43)) ///
>          (mat(res[4]), se(res[5]) keep(43)), bylabel(Logit) ///
>       || , at(_coef) ms(o) nooffset levels(95 90) yline(0) ///
>            byopts(cols(1) yrescale legend(off)) ///
>            xlab(1 "`:lab month 1´"  13 "`:lab month 13´" ///
>                25 "`:lab month 25´" 37 "`:lab month 37´" ///
>                49 "`:lab month 49´" 61 "`:lab month 61´" ///
>                73 "`:lab month 73´" 85 "`:lab month 85´")
```



From this graph I conclude that the precise location of the cutoff is rather irrelevant. From July 2004 (highlighted) on there is not much change and all effects are clearly insignificant (the thin and thick lines depict the 95% and 90% confidence intervals, respectively; if the thin line does not cross the red reference line, then the effect is significantly different from zero at the 5% level using a two-sided test; if the thick line does not cross the red reference line, then the effect is significantly negative at the 5% level using a one-sided test). To the left of July 2004 the effect systematically grows and eventually becomes significant. However, in this region there is considerable misfit of the linear model (see Section 3.1 above), which inflates the treatment effect estimate.

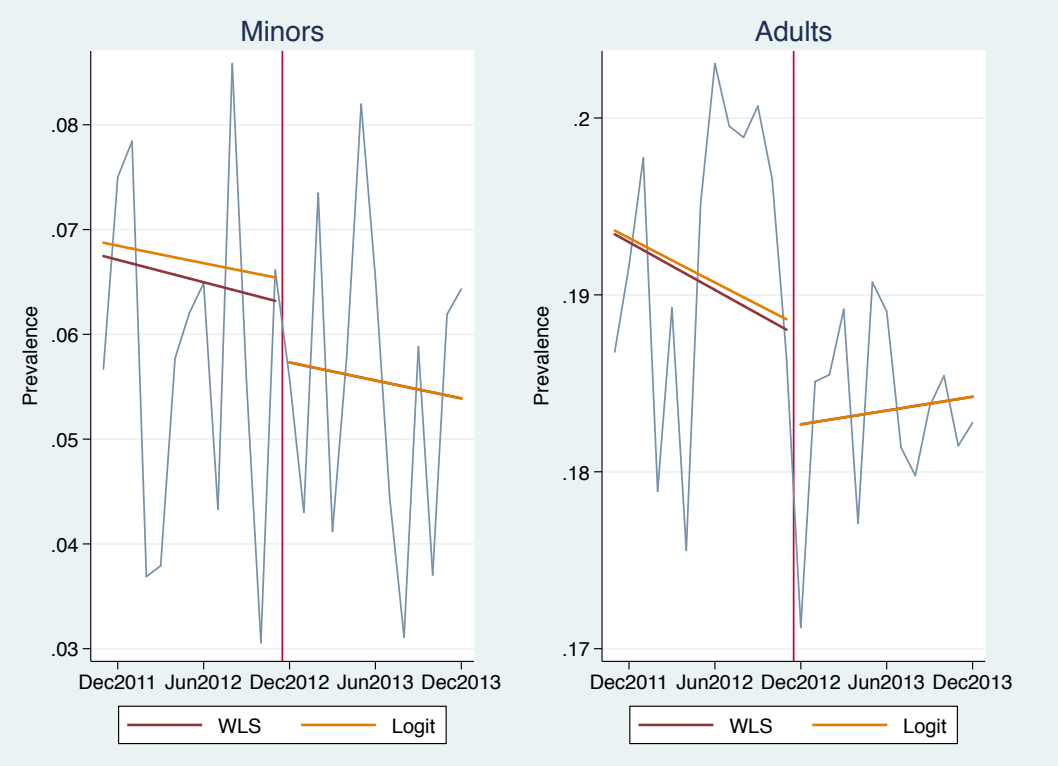
3.2.2 Time-varying treatment effect

In the last section I used a model that assumes an immediate treatment effect that is constant across months. The assumption might not be particularly realistic, but with respect to statistical power the assumption is favorable because it only introduces one additional parameter. A more flexible approach would be to use two parameters so that location and slope of the trend can to change with treatment. This model allows a time-varying treatment effect, with a possible

initial shock and then a linear increase or decrease of the treatment effect over time. Using such a model yields the following results:

```
. use tobacco
. generate byte treat = month>=144
. generate treatmonth = treat * (month - 144)
. generate smokers = round(observations*prevalence)
. preserve
. quietly keep if sample==1                          // => minors
. regress prevalence month treat treatmonth [aw=observations]
(sum of wgt is   4.1438e+04)
      Source |       SS       df       MS              Number of obs =     156
-------------+------------------------------           F(  3,   152) =   43.69
       Model |  .043187597     3  .014395866            Prob > F      =  0.0000
    Residual |  .050089028   152  .000329533            R-squared     =  0.4630
-------------+------------------------------           Adj R-squared =  0.4524
       Total |  .093276625   155  .000601785            Root MSE      =  .01815

------------------------------------------------------------------------------
  prevalence |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       month |  -.0003559   .0000369    -9.64   0.000    -.0004288   -.0002829
       treat |  -.0055249   .0108718    -0.51   0.612    -.0270042    .0159544
  treatmonth |     .00007   .0015261     0.05   0.963    -.0029451    .0030852
       _cons |    .114086   .0028287    40.33   0.000     .1084974    .1196746
------------------------------------------------------------------------------

. testparm treat treatmonth
 ( 1)  treat = 0
 ( 2)  treatmonth = 0
       F(  2,   152) =    0.31
            Prob > F =    0.7341
. predict y_WLS
(option xb assumed; fitted values)
. blogit  smokers observations month treat treatmonth
Logistic regression for grouped data              Number of obs   =      41438
                                                  LR chi2(3)      =     146.56
                                                  Prob > chi2     =     0.0000
Log likelihood = -12325.284                       Pseudo R2       =     0.0059

------------------------------------------------------------------------------
    _outcome |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       month |  -.0044098   .0004462    -9.88   0.000    -.0052844   -.0035352
       treat |  -.1363834   .1573283    -0.87   0.386    -.4447413    .1719745
  treatmonth |  -.0010318   .0225153    -0.05   0.963    -.0451609    .0430974
       _cons |  -2.028496   .0317162   -63.96   0.000    -2.090659   -1.966334
------------------------------------------------------------------------------

. testparm treat treatmonth
 ( 1)  [_outcome]treat = 0
 ( 2)  [_outcome]treatmonth = 0
           chi2(  2) =    2.36
         Prob > chi2 =    0.3071
. predict y_logit, pr
. two (line prev    month, lc(*.6)) ///
>     (line y_WLS   month if month<144 , lw(*1.5) psty(p2)) ///
>     (line y_logit month if month<144 , lw(*1.5) psty(p4)) ///
>     (line y_WLS   month if month>=144, lw(*1.5) psty(p2)) ///
>     (line y_logit month if month>=144, lw(*1.5) psty(p4)) ///
>     if month>=131, legend(order(2 "WLS" 3 "Logit") rows(1)) ///
>     xti("") yti(Prevalence) xline(143.5) ylab(.03(.01).08, angle(hor)) ///
>     xlab(132(6)156, valuelabel) xoverhangs ti(Minors) nodraw name(minors)
. restore, preserve
. quietly keep if sample==2 & month>=43              // => adults
. regress prevalence month treat treatmonth [aw=observations]
(sum of wgt is   5.0666e+05)
      Source |       SS       df       MS              Number of obs =     114
```

```
-------------+------------------------------           F(  3,   110) =  154.22
       Model |  .025806291     3  .008602097           Prob > F      =  0.0000
    Residual |  .006135737   110  .000055779           R-squared     =  0.8079
-------------+------------------------------           Adj R-squared =  0.8027
       Total |  .031942028   113  .000282673           Root MSE      =  .00747

------------------------------------------------------------------------------
  prevalence |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       month |  -.0004494   .0000252   -17.84   0.000    -.0004994   -.0003995
       treat |  -.0049024   .0043881    -1.12   0.266    -.0135985    .0037937
  treatmonth |   .0005794   .0005943     0.97   0.332    -.0005984    .0017572
       _cons |   .2523039   .0024317   103.76   0.000     .2474848    .2571229
------------------------------------------------------------------------------

. testparm treat treatmonth
 ( 1)  treat = 0
 ( 2)  treatmonth = 0
       F(  2,   110) =    0.64
            Prob > F =    0.5311

. predict y_WLS
(option xb assumed; fitted values)

. blogit  smokers observations month treat treatmonth
Logistic regression for grouped data              Number of obs   =     506657
                                                  LR chi2(3)      =     697.76
                                                  Prob > chi2     =     0.0000
Log likelihood = -258773.48                       Pseudo R2       =     0.0013

------------------------------------------------------------------------------
    _outcome |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       month |  -.0027059   .0001243   -21.76   0.000    -.0029496   -.0024622
       treat |  -.0365765    .022708    -1.61   0.107    -.0810834    .0079304
  treatmonth |   .0035737   .0030838     1.16   0.247    -.0024704    .0096179
       _cons |  -1.072061   .0118411   -90.54   0.000    -1.095269   -1.048853
------------------------------------------------------------------------------

. testparm treat treatmonth
 ( 1)  [_outcome]treat = 0
 ( 2)  [_outcome]treatmonth = 0
           chi2(  2) =    2.63
         Prob > chi2 =    0.2687

. predict y_logit, pr

. two (line prev    month, lc(*.6)) ///
>     (line y_WLS   month if month<144 , lw(*1.5) psty(p2)) ///
>     (line y_logit month if month<144 , lw(*1.5) psty(p4)) ///
>     (line y_WLS   month if month>=144, lw(*1.5) psty(p2)) ///
>     (line y_logit month if month>=144, lw(*1.5) psty(p4)) ///
>     if month>=131, legend(order(2 "WLS" 3 "Logit") rows(1)) ///
>     xti("") yti(Prevalence) xline(143.5) ylab(.17(.01).20, angle(hor)) ///
>     xlab(132(6)156, valuelabel) xoverhangs ti(Adults) nodraw name(adults)

. restore

. graph combine minors adults, imargin(zero)
```



The parametrization of the models is such that the main effect of the treatment variable (the second coefficient in the models) reflects the size of the initial shock in December 2012. The results for minors are qualitatively similar to the results from the simpler model above (immediate shift of the curve of about 0.5 percentage points without much change in slope). Results for adults are such that we have an initial shock of about 0.5 percentage points and then the effect declines (positive interaction effect). Since the interaction effect is larger in size than the baseline trend effect, the slope of the trend even turns positive after December 2012. This is certainly not what Australian authorities would have hoped for. However, note that none of these effects are significant, neither the initial shock, nor the change in slope, nor both together using a joint test (see the "testparm" commands in the output).

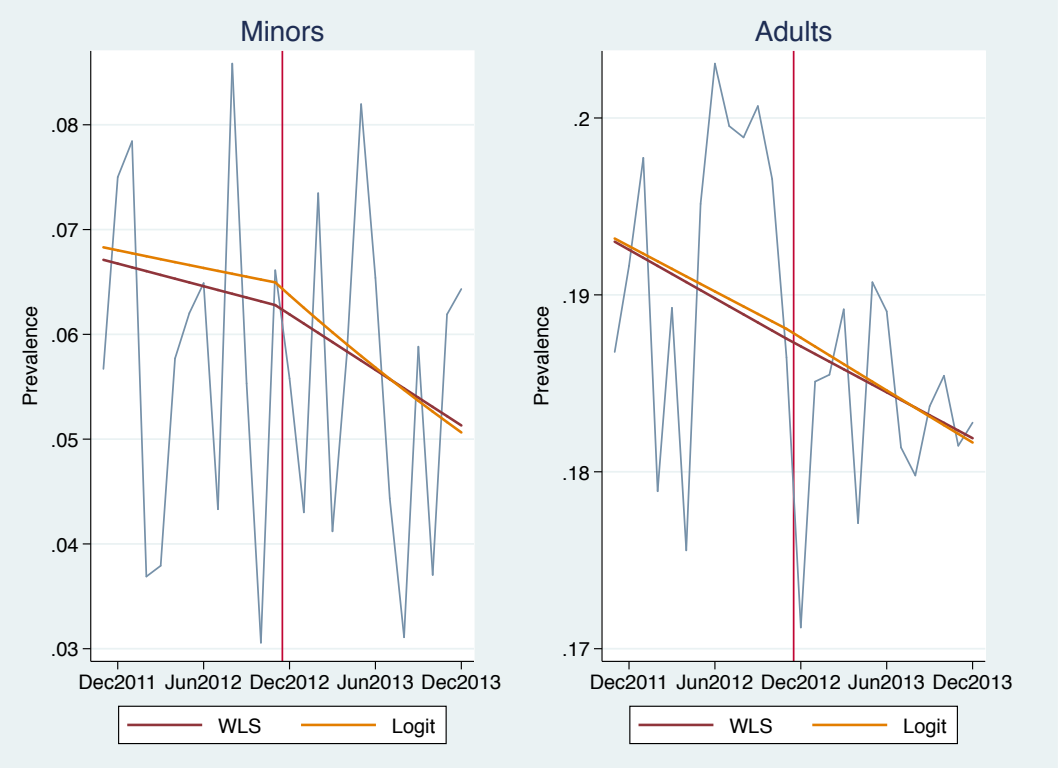
Overall, the results in this section do not seem to add much additional insight.

3.2.3 Gradual treatment effect

A further option to model the treatment effect is to assume that there is no specific initial shock, but that the effect gradually builds up over time. This can be implemented, for example, using a model with linear splines. The following output and graph show the results:

```
. use tobacco
. generate treatmonth = cond(month>143, month-143, 0)
. generate smokers = round(observations*prevalence)
. preserve
. quietly keep if sample==1                          // => minors
. regress prevalence month treatmonth [aw=observations]
(sum of wgt is   4.1438e+04)
      Source |       SS       df       MS              Number of obs =     156
-------------+------------------------------           F(  2,   153) =   65.76
       Model |  .043117457     2  .021558729           Prob > F      =  0.0000
    Residual |  .050159168   153  .000327838           R-squared     =  0.4623
-------------+------------------------------           Adj R-squared =  0.4552
       Total |  .093276625   155  .000601785           Root MSE      =  .01811

------------------------------------------------------------------------------
  prevalence |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       month |  -.0003599   .0000358   -10.06   0.000    -.0004306   -.0002893
  treatmonth |  -.0005235   .0008189    -0.64   0.524    -.0021412    .0010942
       _cons |   .1142626   .0027954    40.87   0.000       .10874    .1197852
------------------------------------------------------------------------------

. lincom month + treatmonth
 ( 1)  month + treatmonth = 0

------------------------------------------------------------------------------
  prevalence |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         (1) |  -.0008834   .0008041    -1.10   0.274     -.002472    .0007051
------------------------------------------------------------------------------

. predict y_WLS
(option xb assumed; fitted values)
. blogit  smokers observations month treatmonth
Logistic regression for grouped data              Number of obs   =      41438
                                                  LR chi2(2)      =     145.96
                                                  Prob > chi2     =     0.0000
Log likelihood = -12325.585                       Pseudo R2       =     0.0059

------------------------------------------------------------------------------
    _outcome |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       month |  -.0044852   .0004358   -10.29   0.000    -.0053394   -.0036309
  treatmonth |  -.0158411   .0119505    -1.33   0.185    -.0392637    .0075816
       _cons |  -2.025471   .0314661   -64.37   0.000    -2.087143   -1.963798
------------------------------------------------------------------------------

. lincom month + treatmonth
 ( 1)  [_outcome]month + [_outcome]treatmonth = 0

------------------------------------------------------------------------------
             |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         (1) |  -.0203262   .0117859    -1.72   0.085    -.0434263    .0027738
------------------------------------------------------------------------------

. predict y_logit, pr
. two (line prev    month, lc(*.6)) ///
>     (line y_WLS   month, lw(*1.5) psty(p2))  ///
>     (line y_logit month, lw(*1.5) psty(p4)) ///
>     if month>=131, legend(order(2 "WLS" 3 "Logit") rows(1)) ///
>     xti("") yti(Prevalence) xline(143.5) ylab(.03(.01).08, angle(hor)) ///
>     xlab(132(6)156, valuelabel) xoverhangs ti(Minors) nodraw name(minors)
. restore, preserve
. quietly keep if sample==2 & month>=43              // => adults
. regress prevalence month treatmonth [aw=observations]
(sum of wgt is   5.0666e+05)
      Source |       SS       df       MS              Number of obs =     114
-------------+------------------------------           F(  2,   111) =  230.14
       Model |  .025735581     2   .01286779           Prob > F      =  0.0000
    Residual |  .006206448   111  .000055914           R-squared     =  0.8057
-------------+------------------------------           Adj R-squared =  0.8022
       Total |  .031942028   113  .000282673           Root MSE      =  .00748
```

```
------------------------------------------------------------------------------
  prevalence |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       month |  -.0004569   .0000243   -18.78   0.000    -.0005051   -.0004087
  treatmonth |   .0000241   .0003319     0.07   0.942    -.0006337    .0006818
       _cons |   .2528662   .0023827   106.12   0.000     .2481447    .2575877
------------------------------------------------------------------------------

. lincom month + treatmonth
 ( 1)  month + treatmonth = 0

------------------------------------------------------------------------------
  prevalence |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         (1) |  -.0004329   .0003208    -1.35   0.180    -.0010686    .0002028
------------------------------------------------------------------------------

. predict y_WLS
(option xb assumed; fitted values)

. blogit  smokers observations month treatmonth
Logistic regression for grouped data              Number of obs   =     506657
                                                  LR chi2(2)      =     695.22
                                                  Prob > chi2     =     0.0000
Log likelihood = -258774.75                       Pseudo R2       =     0.0013

------------------------------------------------------------------------------
    _outcome |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       month |  -.0027573   .0001201   -22.96   0.000    -.0029927   -.0025219
  treatmonth |  -.0005219   .0017073    -0.31   0.760    -.0038682    .0028244
       _cons |   -1.06824   .0115959   -92.12   0.000    -1.090967   -1.045512
------------------------------------------------------------------------------

. lincom month + treatmonth
 ( 1)  [_outcome]month + [_outcome]treatmonth = 0

------------------------------------------------------------------------------
             |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         (1) |  -.0032793   .0016532    -1.98   0.047    -.0065194   -.0000391
------------------------------------------------------------------------------

. predict y_logit, pr

. two (line prev    month, lc(*.6)) ///
>     (line y_WLS   month, lw(*1.5) psty(p2))  ///
>     (line y_logit month, lw(*1.5) psty(p4)) ///
>     if month>=131, legend(order(2 "WLS" 3 "Logit") rows(1)) ///
>     xti("") yti(Prevalence) xline(143.5) ylab(.17(.01).20, angle(hor)) ///
>     xlab(132(6)156, valuelabel) xoverhangs ti(Adults) nodraw name(adults)

. restore

. graph combine minors adults, imargin(zero)
```



The results suggest a change in trend for minors with a treatment effect in December 2012 of about 0.05 percentage points that builds up to about 0.6 percentage points until December 2013. The effect, however, is not significant, with p-values of 0.524 and 0.185 for WLS and the Logit model, respectively (using two-sided tests). Using a one-sided test the change in slope would be significant in the Logit model at the 10% level. For adults, there is hardly any change in slope (with p-values of 0.942 and 0.760). In sum, similar to the models with an immediate treatment effect above, we find some mild evidence for an effect on minors if we are willing to resort to a loose interpretation of statistical tests.

The results also provide tests against a flat trend (the "lincom" results in the output above). Here the null hypothesis is that the smoking prevalence remains constant from November 2012 on. For adults, using the Logit model, we can conclude that there was a further significant decrease in smoking prevalence after November 2012 (using a two-sided test at the 5% level). For minors, the test based on the Logit model is significant at the 5% level only if we are willing to employ a one-sided test. The results from the WLS models are less clear with two-sided p-values of 0.274 and 0.180 for minors and adults, respectively. The fact that we cannot uniformly reject the hypothesis that there was no further decline in smoking prevalence after November 2012 raises concerns about statistical power. Based on the amount of treatment-

period data available it seems to be difficult to reject *any* reasonable null hypothesis about the development of smoking prevalence after November 2012.

3.2.4 Monthly treatment effects

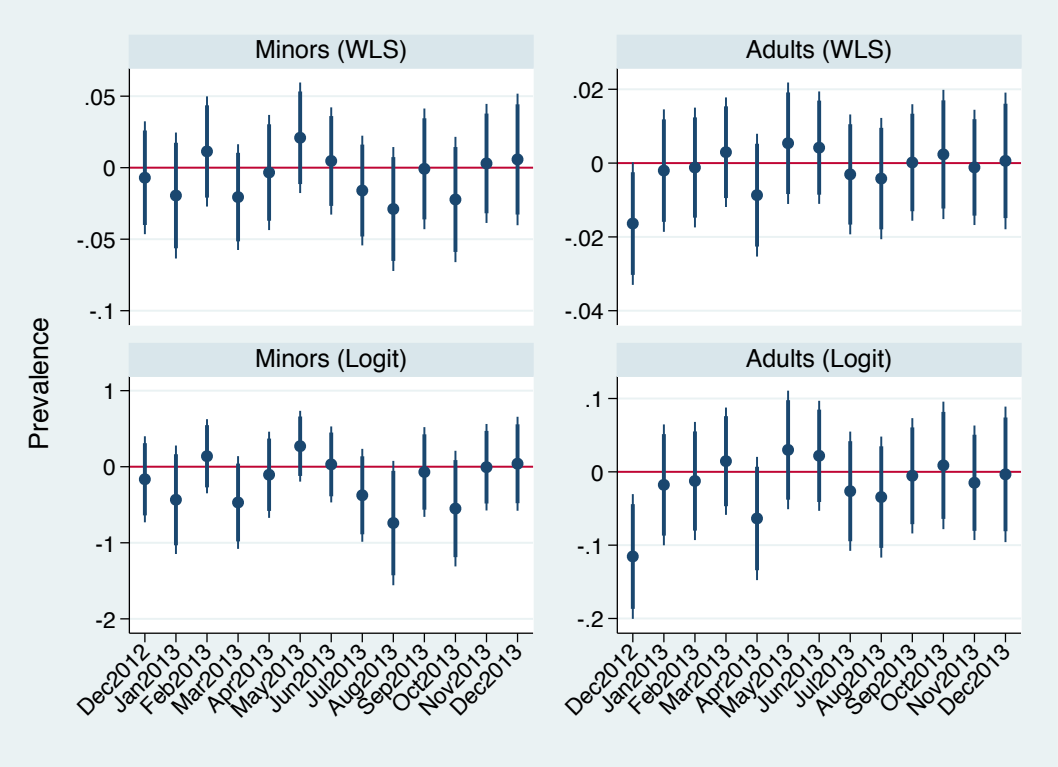
More flexible approaches exist to model the treatment effect, but they all need additional parameters and hence sacrifice statistical power. The most flexible model is one that includes an additional parameter for each treatment-period month, which is analogous to the two-step approach followed by Kaul and Wolf (2014a,b). Using such a model I get the following results:

```
. use tobacco
. clonevar treatmonth = month
. replace  treatmonth = 0 if treatmonth<144
(286 real changes made)
. generate smokers = round(observations*prevalence)
. preserve
. quietly keep if sample==1                          // => minors
. eststo mWLS: regress prevalence month i.treatmonth [aw=observations]
(sum of wgt is   4.1438e+04)
      Source |       SS       df       MS              Number of obs =     156
-------------+------------------------------           F( 14,   141) =    9.56
       Model |  .045415546    14  .003243968           Prob > F      =  0.0000
    Residual |  .047861079   141   .00033944           R-squared     =  0.4869
-------------+------------------------------           Adj R-squared =  0.4359
       Total |  .093276625   155  .000601785           Root MSE      =  .01842

------------------------------------------------------------------------------
  prevalence |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       month |  -.0003559   .0000375    -9.49   0.000      -.00043   -.0002818
             |
  treatmonth |
     Dec2012 |  -.0070456    .019953    -0.35   0.725    -.0464913       .0324
     Jan2013 |   -.019473   .0222739    -0.87   0.383    -.0635071     .024561
     Feb2013 |   .0113415   .0194838     0.58   0.561    -.0271766    .0498596
     Mar2013 |  -.0205735   .0186955    -1.10   0.273    -.0575333    .0163863
     Apr2013 |  -.0033804   .0203613    -0.17   0.868    -.0436332    .0368724
     May2013 |    .020907   .0195403     1.07   0.286    -.0177228    .0595368
     Jun2013 |   .0046803   .0189558     0.25   0.805     -.032794    .0421545
     Jul2013 |  -.0159937   .0193995    -0.82   0.411    -.0543452    .0223579
     Aug2013 |  -.0289045   .0219133    -1.32   0.189    -.0722256    .0144165
     Sep2013 |  -.0008132   .0213365    -0.04   0.970    -.0429941    .0413677
     Oct2013 |  -.0222438   .0221489    -1.00   0.317    -.0660307    .0215431
     Nov2013 |   .0029798   .0210504     0.14   0.888    -.0386355    .0445951
     Dec2013 |   .0057584   .0232658     0.25   0.805    -.0402366    .0517534
             |
       _cons |    .114086   .0028709    39.74   0.000     .1084105    .1197615
------------------------------------------------------------------------------
. testparm i.treatmonth
 ( 1)  144.treatmonth = 0
 ( 2)  145.treatmonth = 0
 ( 3)  146.treatmonth = 0
 ( 4)  147.treatmonth = 0
 ( 5)  148.treatmonth = 0
 ( 6)  149.treatmonth = 0
 ( 7)  150.treatmonth = 0
 ( 8)  151.treatmonth = 0
 ( 9)  152.treatmonth = 0
 (10)  153.treatmonth = 0
 (11)  154.treatmonth = 0
 (12)  155.treatmonth = 0
```

```
 (13)  156.treatmonth = 0
       F( 13,   141) =    0.55
            Prob > F =    0.8884
. eststo mLOG: blogit  smokers observations month i.treatmonth
Logistic regression for grouped data              Number of obs   =      41438
                                                  LR chi2(14)     =     158.01
                                                  Prob > chi2     =     0.0000
Log likelihood = -12319.558                       Pseudo R2       =     0.0064

------------------------------------------------------------------------------
    _outcome |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       month |  -.0044098   .0004462    -9.88   0.000    -.0052844   -.0035352
             |
  treatmonth |
    Dec2012  |  -.1651711   .2884849    -0.57   0.567    -.7305912     .400249
    Jan2013  |  -.4344251   .3638778    -1.19   0.233    -1.147613    .2787623
    Feb2013  |   .1377484   .2485647     0.55   0.579    -.3494295    .6249263
    Mar2013  |  -.4705457   .3109241    -1.51   0.130    -1.079946    .1388543
    Apr2013  |  -.1057625   .2890608    -0.37   0.714    -.6723112    .4607862
    May2013  |   .2696423   .2374735     1.14   0.256    -.1957973    .7350818
    Jun2013  |   .0301178   .2547631     0.12   0.906    -.4692086    .5294442
    Jul2013  |  -.3757892   .3116575    -1.21   0.228    -.9866268    .2350483
    Aug2013  |  -.7405637   .4171966    -1.78   0.076    -1.558254    .0771267
    Sep2013  |  -.0693935   .3010278    -0.23   0.818     -.659397    .5206101
    Oct2013  |  -.5504915   .3878986    -1.42   0.156    -1.310759    .2097759
    Nov2013  |  -.0062395   .2900879    -0.02   0.983    -.5748013    .5623223
    Dec2013  |   .0391461   .3151973     0.12   0.901    -.5786292    .6569214
             |
       _cons |  -2.028496   .0317162   -63.96   0.000    -2.090659   -1.966334
------------------------------------------------------------------------------

. testparm i.treatmonth
 ( 1)  [_outcome]144.treatmonth = 0
 ( 2)  [_outcome]145.treatmonth = 0
 ( 3)  [_outcome]146.treatmonth = 0
 ( 4)  [_outcome]147.treatmonth = 0
 ( 5)  [_outcome]148.treatmonth = 0
 ( 6)  [_outcome]149.treatmonth = 0
 ( 7)  [_outcome]150.treatmonth = 0
 ( 8)  [_outcome]151.treatmonth = 0
 ( 9)  [_outcome]152.treatmonth = 0
 (10)  [_outcome]153.treatmonth = 0
 (11)  [_outcome]154.treatmonth = 0
 (12)  [_outcome]155.treatmonth = 0
 (13)  [_outcome]156.treatmonth = 0
           chi2( 13) =   12.30
         Prob > chi2 =    0.5030
. restore, preserve
. quietly keep if sample==2 & month>=43               // => adults
. eststo aWLS: regress prevalence month i.treatmonth [aw=observations]
(sum of wgt is   5.0666e+05)

      Source |       SS       df       MS              Number of obs =     114
-------------+------------------------------           F( 14,    99) =   31.69
       Model |  .026115162    14  .001865369            Prob > F      =  0.0000
    Residual |  .005826866    99  .000058857            R-squared     =  0.8176
-------------+------------------------------           Adj R-squared =  0.7918
       Total |  .031942028   113  .000282673            Root MSE      =  .00767

------------------------------------------------------------------------------
  prevalence |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       month |  -.0004494   .0000259   -17.37   0.000    -.0005008   -.0003981
             |
  treatmonth |
    Dec2012  |  -.0163843   .0083749    -1.96   0.053    -.0330019    .0002333
    Jan2013  |  -.0020348   .0083603    -0.24   0.808    -.0186234    .0145539
    Feb2013  |  -.0011987   .0081781    -0.15   0.884    -.0174258    .0150283
    Mar2013  |   .0029485   .0074825     0.39   0.694    -.0118984    .0177954
    Apr2013  |  -.0086916   .0083848    -1.04   0.302    -.0253289    .0079457
    May2013  |   .0053801   .0082912     0.65   0.518    -.0110713    .0218316
    Jun2013  |   .0041807   .0076749     0.54   0.587     -.011048    .0194093
```

```
    Jul2013 |  -.0030649   .0081867    -0.37   0.709    -.0193091    .0131793
    Aug2013 |  -.0041985   .0082814    -0.51   0.613    -.0206306    .0122335
    Sep2013 |   .0001632   .0079553     0.02   0.984    -.0156218    .0159482
    Oct2013 |   .0023396   .0088255     0.27   0.791    -.0151722    .0198513
    Nov2013 |  -.0011681   .0078685    -0.15   0.882    -.0167809    .0144447
    Dec2013 |   .0005878   .0093303     0.06   0.950    -.0179255    .0191012

      _cons |   .2523039   .0024979   101.01   0.000     .2473475    .2572602
------------------------------------------------------------------------------

. testparm i.treatmonth
 ( 1)  144.treatmonth = 0
 ( 2)  145.treatmonth = 0
 ( 3)  146.treatmonth = 0
 ( 4)  147.treatmonth = 0
 ( 5)  148.treatmonth = 0
 ( 6)  149.treatmonth = 0
 ( 7)  150.treatmonth = 0
 ( 8)  151.treatmonth = 0
 ( 9)  152.treatmonth = 0
 (10)  153.treatmonth = 0
 (11)  154.treatmonth = 0
 (12)  155.treatmonth = 0
 (13)  156.treatmonth = 0
       F( 13,    99) =    0.50
            Prob > F =    0.9219

. eststo aLOG: blogit  smokers observations month i.treatmonth
Logistic regression for grouped data              Number of obs   =     506657
                                                  LR chi2(14)     =     706.98
                                                  Prob > chi2     =     0.0000
Log likelihood = -258768.87                       Pseudo R2       =     0.0014

------------------------------------------------------------------------------
    _outcome |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       month |  -.0027059   .0001243   -21.76   0.000    -.0029496   -.0024622

  treatmonth |
    Dec2012 |  -.1154462   .0433894    -2.66   0.008    -.2004879   -.0304044
    Jan2013 |  -.0177622   .0420483    -0.42   0.673    -.1001754    .064651
    Feb2013 |  -.0124954   .0410966    -0.30   0.761    -.0930432    .0680525
    Mar2013 |   .0145005   .0373205     0.39   0.698    -.0586464    .0876474
    Apr2013 |   -.063631   .0428737    -1.48   0.138     -.147662    .0203999
    May2013 |   .0298738   .0412358     0.72   0.469    -.0509469    .1106946
    Jun2013 |   .0218607   .0382867     0.57   0.568    -.0531798    .0969012
    Jul2013 |  -.0264287    .041477    -0.64   0.524    -.1077223    .0548648
    Aug2013 |  -.0344222   .0420938    -0.82   0.414    -.1169246    .0480803
    Sep2013 |   -.005408   .0401065    -0.13   0.893    -.0840152    .0731992
    Oct2013 |   .0087724   .0443499     0.20   0.843    -.0781519    .0956967
    Nov2013 |  -.0149389   .0398428    -0.37   0.708    -.0930293    .0631516
    Dec2013 |  -.0034587   .0471415    -0.07   0.942    -.0958544     .088937

      _cons |  -1.072061   .0118411   -90.54   0.000    -1.095269   -1.048853
------------------------------------------------------------------------------

. testparm i.treatmonth
 ( 1)  [_outcome]144.treatmonth = 0
 ( 2)  [_outcome]145.treatmonth = 0
 ( 3)  [_outcome]146.treatmonth = 0
 ( 4)  [_outcome]147.treatmonth = 0
 ( 5)  [_outcome]148.treatmonth = 0
 ( 6)  [_outcome]149.treatmonth = 0
 ( 7)  [_outcome]150.treatmonth = 0
 ( 8)  [_outcome]151.treatmonth = 0
 ( 9)  [_outcome]152.treatmonth = 0
 (10)  [_outcome]153.treatmonth = 0
 (11)  [_outcome]154.treatmonth = 0
 (12)  [_outcome]155.treatmonth = 0
 (13)  [_outcome]156.treatmonth = 0
           chi2( 13) =   11.67
         Prob > chi2 =    0.5549

. restore
. coefplot (mWLS), bylabel(Minors (WLS))    ///
```

```
>       || (aWLS), bylabel(Adults (WLS))    ///
>       || (mLOG), bylabel(Minors (Logit))  ///
>       || (aLOG), bylabel(Adults (Logit))  ///
>       ||, keep(*.treatmonth) levels(95 90) vertical xlabel(,angle(45)) ///
>     yline(0) ylab(, angle(hor)) yti(Prevalence) ///
>     byopts(yrescale rows(2))
```



The graph depicts the monthly treatment effects by sample and estimation technique, including 95% and 90% confidence intervals as thin and thick lines, respectively. The WLS estimates closely resemble the results shown in Figure 5 in Kaul and Wolf (2014b) and Figures 6 and 8 in Kaul and Wolf (2014a). For minors we do not see any clear deviations from the baseline trend,[11] except maybe the negative deviation in August 2013 in the Logit model, which is significant using a one-sided test at the 5% level. Consequently, the overall tests (the "testparm" results) do not provide much evidence for systematic deviations (with p-values of 0.888 and 0.503). The overall tests for adults are similarly weak (with two-sided p-values of 0.922 and 0.555), although there is a significant effect in December 2012 (with p-values of .053 and .008). Yet, as indicated by the overall tests, it is not very surprising to see a significant effect among 13 monthly estimates (for example, using a 5% level, we would expect one

[11] The picture does not change much if a flat trend from November 2012 on is assumed, because the variability in monthly estimates is much stronger than the slope of the trend (not shown).

significant result per twenty test). It is thus hard to say whether the December 2012 effect is systematic. Such a conclusion could only be drawn if there had been a strong and well justified prior expectation of a specific effect in December 2012 only, but not in other months.

3.3 Power

As mentioned above, statistical power to detect a treatment effect with the present data is likely to be low. Below are some results of a power analysis. To generate the simulated data I follow the procedure suggested by Kaul and Wolf (2014a,b). That is, the data in a single replication are generated as

$$y_t = \beta_0 + \beta_1 \cdot t + \Delta_t + \frac{\epsilon_t}{\sqrt{n_t}}$$

where $t = 1, \ldots, 156$ is the index of the month ($t = 43, \ldots, 156$ in case of adults), n_t is the sample size in month t, and β_0 and β_1 are the coefficient estimates from the pre-treatment WLS fit (see Section 3.1). ϵ_t is a random error following a normal distribution with mean 0 and variance $\sigma^2 \cdot \overline{n}$, where $\overline{n}$ is the average sample size in the pre-treatment months and σ^2 is the mean squared error (i.e. the error variance estimate) of the pre-treatment WLS fit.[12] Δ_t is the treatment effect defined as $-\delta \cdot I[t > 143]$, where $I[]$ is the index function (i.e. 0 in pre-treatment months, 1 in treatment-period months) and δ is varied from 0 to .02 in steps of .0025 (corresponding to a reduction in smoking rates due to plain packaging between 0 and 2 percentage points). Using such a definition the treatment effect is immediate and time-constant. To alternatively model a situation in which the treatment effect builds up gradually I also use $\Delta_t = -(\delta/13) \cdot (t - 143) \cdot I[t > 143]$, again varying δ from 0 to .02 in steps of .0025. Using this definition the treatment effect unfolds linearly, from $-\delta/13$ in December 2012 to $-\delta$ in December 2013.

In the power analysis I include the WLS models from Section 3.2.1 (immediate treatment-effect model; power is the proportion of replications in which the treatment dummy is negative and significant), Section 3.2.3 (gradual treatment-effect model; power is the proportion of replications in which the spline parameter is negative and significant), and Section 3.2.4 (monthly

[12] I slightly deviate from Kaul and Wolf (2014a,b) in the computation of the error variance in that I (implicitly) divide by $T - 2$ (where T is the number of months on which the pre-treatment WLS fit is based) to account for the degree of freedom consumed by the slope parameter. The numbers provided by Kaul and Wolf (2014a,b) suggest that they divided by $T - 1$. The rest of the procedure is identical, despite the difference in notation.

treatment-effects model; power is (a) the proportion of replications in which any of the monthly dummies is negative and significant and (b) the proportion of replications in which the overall test across all dummies is significant). All tests use a nominal significance level of 5%. Each run of the simulation is repeated twice, once using two-sided tests and once using one-sided tests (the overall test for monthly dummies is only included in the two-sided test setting). The results, based on 10,000 replications, are as follows:

```
. set seed 5218716
. local reps 10000
. program define wlssim, rclass
  1.         syntax [, cons(real 0) slope(real 0) sigma(real 1) ///
>            delta(real 0) twosided onesided gradual immediate ]
  2.         if "`onesided´"!="" local level .10
  3.         else                local level .05
  4.         if "`gradual´"==""  local teff `delta´*treat
  5.         else                local teff (`delta´/13)*(month-143)*treat
  6.         tempvar y
  7.         generate `y´ = `cons´ + `slope´*month - `teff´ ///
>             + rnormal(0, `sigma´)/sqrt(observations)
  8.         regress `y´ month treat [aw=observations]
  9.         return scalar b1 = _b[treat]
 10.         test treat
 11.         return scalar p1 = (r(p)<`level´) & (_b[treat]<0)
 12.         regress `y´ month pmonth [aw=observations]
 13.         return scalar b2 = _b[pmonth]
 14.         test pmonth
 15.         return scalar p2 = (r(p)<`level´) & (_b[pmonth]<0)
 16.         regress `y´ month i.imonth [aw=observations]
 17.         local any 0
 18.         forv i = 144/156 {
 19.             test `i´.imonth
 20.             if (r(p)<`level´) & (_b[`i´.imonth]<0) local any 1
 21.         }
 22.         return scalar p3 = `any´
 23.         if "`onesided´"=="" {
 24.             testparm i.imonth
 25.             return scalar p4 = (r(p)<`level´)
 26.         }
 27. end
. use tobacco
. drop if sample==2 & month<43
(42 observations deleted)
. generate byte treat = month>=144
. generate pmonth = cond(month>143, month-143, 0)
. generate imonth = cond(month>143, month, 0)
. capt mat drop wlssim
. forv s =1/2 {
  2.     local sti: word `s´ of minors adults
  3.     qui regress prevalence month if sample==`s´ & treat==0 [aw=observations]
  4.     su observation if sample==`s´ & treat==0, mean
  5.     local cons  = _b[_cons]
  6.     local slope = _b[month]
  7.     local sigma = sqrt(r(mean)) * e(rmse)
  8.     preserve
  9.     foreach type in "immediate" "gradual" {
 10.         foreach test in twosided onesided {
 11.             local p4
 12.             if "`test´"=="twosided" local p4 "p4=r(p4)"
 13.             capt mat drop tmp
 14.             forv delta = 0(0.0025).02 {
 15.                 qui keep if sample==`s´
 16.                 qui simulate b1=r(b1) p1=r(p1) b2=r(b2) p2=r(p2) p3=r(p3) ///
>                    `p4´, reps(`reps´) : wlssim, cons(`cons´) slope(`slope´) ///
```

```
>                     sigma(`sigma´) delta(`delta´) `test´ `type´
 17.                 qui mean *
 18.                 mat tmp = nullmat(tmp) \ ((`delta´), e(b))
 19.                 restore, preserve
 20.             }
 21.             mat coleq tmp = "`sti´ - `type´ effect - `test´ tests"
 22.             mat wlssim = nullmat(wlssim) \ tmp´
 23.         }
 24.     }
 25.     restore, not
 26. }
. estout matrix(wlssim, fmt(4)), coll(none) nolz varw(26) modelw(7) del("") noabbrev ml(none) ///
>     drop(b*) varlab(c1 "effect size" p1 "- immediate effect model" p2 "- gradual effect model" ///
>     p3 "- monthly effects: any" p4 "- monthly effects: overall") refcat(p1 "power", nolabel)
```

minors - immediate effect - twosided tests

effect size	.0000	.0025	.0050	.0075	.0100	.0125	.0150	.0175	.0200
power									
- immediate effect model	.0250	.0581	.1110	.2113	.3232	.4659	.6139	.7546	.8516
- gradual effect model	.0270	.0502	.0920	.1629	.2449	.3502	.4671	.5954	.7123
- monthly effects: any	.2634	.3386	.4220	.4983	.5861	.6690	.7446	.8219	.8688
- monthly effects: overall	.0507	.0549	.0668	.0857	.1157	.1572	.2331	.3111	.4115

minors - immediate effect - onesided tests

effect size	.0000	.0025	.0050	.0075	.0100	.0125	.0150	.0175	.0200
power									
- immediate effect model	.0511	.1047	.1811	.2980	.4443	.5898	.7313	.8370	.9159
- gradual effect model	.0491	.0921	.1537	.2455	.3536	.4750	.5998	.7044	.8133
- monthly effects: any	.4703	.5524	.6323	.7165	.7907	.8438	.9035	.9351	.9659

minors - gradual effect - twosided tests

effect size	.0000	.0025	.0050	.0075	.0100	.0125	.0150	.0175	.0200
power									
- immediate effect model	.0273	.0385	.0620	.0868	.1128	.1650	.2099	.2702	.3272
- gradual effect model	.0291	.0435	.0700	.1075	.1456	.2142	.2854	.3573	.4340
- monthly effects: any	.2703	.3120	.3423	.3872	.4331	.4892	.5395	.5896	.6347
- monthly effects: overall	.0530	.0526	.0567	.0634	.0687	.0804	.1016	.1293	.1536

minors - gradual effect - onesided tests

effect size	.0000	.0025	.0050	.0075	.0100	.0125	.0150	.0175	.0200
power									
- immediate effect model	.0525	.0728	.1078	.1411	.1959	.2461	.3196	.3921	.4652
- gradual effect model	.0486	.0789	.1198	.1711	.2423	.3028	.4038	.4917	.5817
- monthly effects: any	.4656	.5221	.5590	.6041	.6547	.7059	.7537	.7938	.8296

adults - immediate effect - twosided tests

effect size	.0000	.0025	.0050	.0075	.0100	.0125	.0150	.0175	.0200
power									
- immediate effect model	.0253	.1460	.4288	.7703	.9463	.9939	.9998	1.0000	1.0000
- gradual effect model	.0241	.1132	.3129	.6043	.8334	.9523	.9906	.9993	1.0000
- monthly effects: any	.2617	.4559	.6601	.8348	.9394	.9827	.9966	.9998	1.0000
- monthly effects: overall	.0511	.0726	.1450	.3157	.5608	.8006	.9461	.9923	.9989

adults - immediate effect - onesided tests

effect size	.0000	.0025	.0050	.0075	.0100	.0125	.0150	.0175	.0200
power									
- immediate effect model	.0512	.2244	.5569	.8490	.9709	.9980	.9999	1.0000	1.0000
- gradual effect model	.0485	.1818	.4303	.7154	.9049	.9786	.9959	.9997	.9999
- monthly effects: any	.4659	.6694	.8352	.9470	.9843	.9985	.9997	.9999	1.0000

adults - gradual effect - twosided tests

effect size	.0000	.0025	.0050	.0075	.0100	.0125	.0150	.0175	.0200
power									
- immediate effect model	.0248	.0686	.1474	.2863	.4572	.6337	.7833	.8956	.9544
- gradual effect model	.0260	.0817	.1950	.3883	.6096	.8027	.9170	.9729	.9942
- monthly effects: any	.2661	.3668	.4717	.6134	.7309	.8320	.9176	.9608	.9856

```
- monthly effects: overall  .0490  .0565  .0775  .1321  .2228  .3329  .5045  .6671  .8090
-------------------------------------------------------------------------------------------

adults - gradual effect - onesided tests
>
effect size                .0000  .0025  .0050  .0075  .0100  .0125  .0150  .0175  .0200
power
- immediate effect model    .0512  .1227  .2431  .3995  .5869  .7490  .8666  .9424  .9781
- gradual effect model      .0518  .1443  .3003  .5119  .7238  .8750  .9581  .9891  .9982
- monthly effects: any      .4613  .5806  .6944  .7989  .8865  .9473  .9758  .9914  .9979
-------------------------------------------------------------------------------------------
```

For the case with a zero treatment effect we see that the type I error rates more or less match the nominal rate for the immediate effect model and the gradual effect model, about 2.5% for the two-sided tests (since only negative effects are counted) and about 5% for the one-sided tests (see first column of the tables in the output above). Also the overall test for the monthly dummies model has the expected type I error rate of about 5%. The type I error rates of the test whether any monthly dummy is significantly negative, however, are inflated due to multiple testing (about 27% in the two-sided setting, and about 47% in the one-sided setting, instead of the nominal 2.5% and 5%). That is, even in situations where there is no treatment effect, we would expect some monthly deviations that are significantly negative with considerable likelihood. Consequently, such a test procedure also has high power for detecting existing treatment effects. For example, in case of an immediate one percentage point treatment effect, we would expect to see at least one significantly negative monthly deviation with a probability of 59% (two-sided) or 78% (one-sided) for minors and 94% (two-sided) or 99% (one-sided) for adults. Also for a gradual treatment effect that reaches one percentage point by the end of the observation period (which is harder to detect), power is relatively high, at least for adults (44% and 66% for minors, 74% and 89% for adults, for the two-sided and one-sided tests, respectively). Of course power is high because the error rate is inflated. Nonetheless we can say that if there was a strong treatment effect, such a test procedure would be quite likely to pick it up.[13]

Naturally, the power of the tests that retain the nominal error rate is lower. If there is an immediate treatment effect of one percentage point, the power of the constant effect model is 33% (two-sided) or 44% (one-sided) for minors and 94% (two-sided) or 97% (one-sided) for adults. That is, power is still great for adults, but for minors we now already have a probability

[13] I consider a one percentage point difference a strong treatment effect (for minors, this is a 17% decrease in smoking prevalence compared to the linear baseline trend in December 2013; for adults the corresponding decline is 5.5%), but others might disagree. See the tables in the output above for power values for other effect sizes.

of a type II error (i.e. not do detect an existing effect) of more than 50%. Moreover, power reduces further if the effect is gradual. In the setting with a gradual effect that builds up to one percentage point by December 2013, the gradual effect model has a power of only 15% (two-sided) or 24% (one-sided) for minors and 60% (two-sided) or 73% (one-sided) for adults. That is, for minors it would be quite unlikely that such an effect is detected, for adults the power values are also below the conventional 80% level. Using a one-sided test with a conventional 5% significance level, an immediate treatment effect would have to be close to 1.75 percentage points for minors, and about 0.75 percentage points for adults, to reach the conventional 80% power level; a gradual effect would have to build up to more than 2 percentage points in December 2013 for minors, and to about 1.25 percentage points for adults.[14]

The power simulations above might be criticized because they do not take into account that error variances tend to decline the closer the prevalence gets to zero (a natural property of 0/1 variables). In the present context, this leads to underestimation of power. A data-generating process that better matches the nature of the data should be based on the binomial distribution. In the additional simulations below, I therefore generate the number of smokers in the monthly samples as draws from binomial distribution $B(n_t, p_t)$, where n_t is the sample size in month t and p_t is the smoking probability. The smoking probability p_t is modeled using the logistic function

$$p_t = \frac{1}{1 + e^{-z_t}} \quad \text{with} \quad z_t = \gamma_0 + \gamma_1 \cdot t + \Delta_t + u_t$$

where γ_0 and γ_1 are the Logit coefficients from the baseline trend and u_t is is a random error following a normal distribution with mean 0 and variance σ_u^2. The values for γ_0, γ_1 and σ_u are taken from a random-effects logistic regression fitted to the pre-treatment observations, where σ_u reflects the extra between-month variance in the data.[15] Δ_t is defined as above, with the value of δ chosen such that a certain target percentage-point effect is reached in December 2013 (again varying from 0 to 2 percentage points in steps of 0.25 percentage points). The

[14]Power at 2 percentage points is 58% for minors. Finding out how much stronger the effect has to be to reach the 80% level would require additional simulations.

[15]The variability of the prevalence estimates across months, net of the linear trend, is slightly higher than one would expect from the binomial distribution alone. Power would be somewhat overestimated if this extra variability was not taken into account

same types of simulations and tests are performed as above, but now all tests are based on the Logit model instead of the WLS model. The results are as follows:

```
. set seed 477350
. local reps 10000
. mata:
------------------------------------------------- mata (type end to exit) -----------------------------
: real scalar getdelta(p, cons, slope)
> {
>     rc = mm_root(x=., &myfun(), 0, 1, 0, 1000, p, cons, slope)
>     if (rc) _error(499)
>     return(x)
> }
: real scalar myfun(x, p, cons, slope)
> {
>     return(p - (invlogit(cons + slope*156 - x) - invlogit(cons + slope * 156)))
> }
: end
------------------------------------------------------------------------------------------------------

. program define logitsim, rclass
  1.         syntax [, cons(real 0) slope(real 0) sigma(real 1) ///
>             delta(real 0) twosided onesided gradual immediate ]
  2.         if "`onesided´"!="" local level .10
  3.         else                local level .05
  4.         if "`gradual´"==""  local teff `delta´*treat
  5.         else                local teff (`delta´/13)*(month-143)*treat
  6.         tempvar y
  7.         generate `y´ = rbinomial(observations, ///
>             invlogit(`cons´ + `slope´*month - `teff´ + rnormal(0, `sigma´)))
  8.         blogit `y´ observations month treat
  9.         return scalar b1 = _b[treat]
 10.         test treat
 11.         return scalar p1 = (r(p)<`level´) & (_b[treat]<0)
 12.         blogit `y´ observations month pmonth
 13.         return scalar b2 = _b[pmonth]
 14.         test pmonth
 15.         return scalar p2 = (r(p)<`level´) & (_b[pmonth]<0)
 16.         blogit `y´ observations month i.imonth
 17.         local any 0
 18.         forv i = 144/156 {
 19.             test `i´.imonth
 20.             if (r(p)<`level´) & (_b[`i´.imonth]<0) local any 1
 21.         }
 22.         return scalar p3 = `any´
 23.         if "`onesided´"=="" {
 24.             testparm i.imonth
 25.             return scalar p4 = (r(p)<`level´)
 26.         }
 27. end
. use tobacco
. drop if sample==2 & month<43
(42 observations deleted)
. generate byte treat = month>=144
. generate pmonth = cond(month>143, month-143, 0)
. generate imonth = cond(month>143, month, 0)
. capt mat drop logitsim
. forv s =1/2 {
  2.     local sti: word `s´ of minors adults
  3.     preserve
  4.     qui keep if treat==0 & sample==`s´
  5.     keep month observations prevalence
  6.     qui expand observations
  7.     sort month
  8.     by month: gen byte smokes = (_n<= round(observations*prevalence))
  9.     xtlogit smokes month, i(month)
 10.     local cons  = _b[_cons]
 11.     local slope = _b[month]
 12.     local sigma = e(sigma_u)
```

```
 13.     restore, preserve
 14.     foreach type in "immediate" "gradual" {
 15.         foreach test in twosided onesided {
 16.             local p4
 17.             if "`test´"=="twosided" local p4 "p4=r(p4)"
 18.             capt mat drop tmp
 19.             forv delta = 0(0.0025).02 {
 20.                 mata: st_local("d", strofreal(getdelta(-`delta´, `cons´, `slope´)))
 21.                 qui keep if sample==`s´
 22.                 qui simulate b1=r(b1) p1=r(p1) b2=r(b2) p2=r(p2) p3=r(p3) ///
>                     `p4´, reps(`reps´) : logitsim, cons(`cons´) slope(`slope´) ///
>                     sigma(`sigma´) delta(`d´) `test´ `type´
 23.                 qui mean *
 24.                 mat tmp = nullmat(tmp) \ ((`delta´), e(b))
 25.                 restore, preserve
 26.             }
 27.             mat coleq tmp = "`sti´ - `type´ effect - `test´ tests"
 28.             mat logitsim = nullmat(logitsim) \ tmp´
 29.         }
 30.     }
 31.     restore, not
 32. }
Fitting comparison model:

Iteration 0:   log likelihood = -11757.613
Iteration 1:   log likelihood = -11708.002
Iteration 2:   log likelihood = -11707.726
Iteration 3:   log likelihood = -11707.726

Fitting full model:

tau =  0.0     log likelihood = -11707.726
tau =  0.1     log likelihood = -11739.458

Iteration 0:   log likelihood = -11739.461
Iteration 1:   log likelihood = -11710.224
Iteration 2:   log likelihood = -11707.679
Iteration 3:   log likelihood = -11707.527
Iteration 4:   log likelihood =  -11707.52
Iteration 5:   log likelihood =  -11707.52

Random-effects logistic regression              Number of obs      =     38564
Group variable: month                           Number of groups   =       143

Random effects u_i ~ Gaussian                   Obs per group: min =       161
                                                               avg =     269.7
                                                               max =       381

Integration method: mvaghermite                 Integration points =        12

                                                Wald chi2(1)       =     90.64
Log likelihood  =  -11707.52                    Prob > chi2        =    0.0000

------------------------------------------------------------------------------
      smokes |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       month |   -.004408    .000463    -9.52   0.000    -.0053154   -.0035005
       _cons |  -2.029978   .0333685   -60.84   0.000    -2.095379   -1.964577
-------------+----------------------------------------------------------------
    /lnsig2u |  -5.671567   1.649077                      -8.9037   -2.439435
-------------+----------------------------------------------------------------
     sigma_u |   .0586725   .0483778                      .011657    .2953136
         rho |   .0010453    .001722                     .0000413    .0258241
------------------------------------------------------------------------------

Likelihood-ratio test of rho=0: chibar2(01) =     0.41 Prob >= chibar2 = 0.260

Fitting comparison model:

Iteration 0:   log likelihood =  -233896.9
Iteration 1:   log likelihood = -233659.66
Iteration 2:   log likelihood = -233659.53
Iteration 3:   log likelihood = -233659.53

Fitting full model:

tau =  0.0     log likelihood = -233659.53
tau =  0.1     log likelihood = -234180.57

Iteration 0:   log likelihood = -233805.85
Iteration 1:   log likelihood = -233686.44  (not concave)
Iteration 2:   log likelihood = -233663.19  (not concave)
Iteration 3:   log likelihood = -233658.26  (not concave)
Iteration 4:   log likelihood = -233655.37
```

```
Iteration 5:   log likelihood = -233654.19
Iteration 6:   log likelihood = -233654.02
Iteration 7:   log likelihood = -233654.02

Random-effects logistic regression              Number of obs      =    453961
Group variable: month                           Number of groups   =       101

Random effects u_i ~ Gaussian                   Obs per group: min =      4005
                                                               avg =    4494.7
                                                               max =      5126

Integration method: mvaghermite                 Integration points =        12

                                                Wald chi2(1)       =    305.26
Log likelihood  = -233654.02                    Prob > chi2        =    0.0000

------------------------------------------------------------------------------
      smokes |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       month |  -.0027056   .0001549   -17.47   0.000    -.0030091   -.0024021
       _cons |   -1.07224   .0148719   -72.10   0.000    -1.101388   -1.043092
-------------+----------------------------------------------------------------
    /lnsig2u |  -7.224383   .4008665                     -8.010067   -6.438699
-------------+----------------------------------------------------------------
     sigma_u |   .0269926   .0054102                      .0182237    .0399811
         rho |   .0002214   .0000887                      .0001009    .0004856
------------------------------------------------------------------------------

Likelihood-ratio test of rho=0: chibar2(01) =    11.03 Prob >= chibar2 = 0.000

. estout matrix(logitsim, fmt(4)), coll(none) nolz varw(26) modelw(7) del("") noabbrev ml(none) ///
>     drop(b*) varlab(c1 "effect size" p1 "- immediate effect model" p2 "- gradual effect model" ///
>     p3 "- monthly effects: any" p4 "- monthly effects: overall") refcat(p1 "power", nolabel)

--------------------------------------------------------------------------------
minors - immediate effect - twosided tests
>
effect size                 .0000  .0025  .0050  .0075  .0100  .0125  .0150  .0175  .0200
power
- immediate effect model    .0264  .0731  .1645  .3228  .5089  .7211  .8651  .9547  .9854
- gradual effect model      .0276  .0618  .1280  .2395  .3934  .5720  .7307  .8750  .9460
- monthly effects: any      .1992  .2635  .3452  .4435  .5481  .6592  .7513  .8583  .9166
- monthly effects: overall  .0522  .0419  .0430  .0585  .0941  .1640  .2711  .4418  .6170
--------------------------------------------------------------------------------
minors - immediate effect - onesided tests
>
effect size                 .0000  .0025  .0050  .0075  .0100  .0125  .0150  .0175  .0200
power
- immediate effect model    .0518  .1339  .2546  .4325  .6312  .8125  .9257  .9756  .9944
- gradual effect model      .0504  .1154  .2027  .3480  .5192  .6966  .8341  .9261  .9744
- monthly effects: any      .4184  .5306  .6300  .7310  .8111  .8903  .9432  .9721  .9896
--------------------------------------------------------------------------------
minors - gradual effect - twosided tests
>
effect size                 .0000  .0025  .0050  .0075  .0100  .0125  .0150  .0175  .0200
power
- immediate effect model    .0269  .0461  .0780  .1204  .1820  .2535  .3518  .4757  .5951
- gradual effect model      .0254  .0475  .0852  .1415  .2199  .3245  .4522  .6005  .7244
- monthly effects: any      .2015  .2315  .2685  .3155  .3710  .4392  .4873  .5772  .6429
- monthly effects: overall  .0547  .0485  .0414  .0471  .0531  .0685  .0830  .1198  .1745
--------------------------------------------------------------------------------
minors - gradual effect - onesided tests
>
effect size                 .0000  .0025  .0050  .0075  .0100  .0125  .0150  .0175  .0200
power
- immediate effect model    .0497  .0844  .1348  .1964  .2706  .3716  .4846  .5898  .7096
- gradual effect model      .0488  .0907  .1525  .2299  .3323  .4554  .5887  .7145  .8251
- monthly effects: any      .4214  .4763  .5436  .5936  .6534  .7148  .7736  .8398  .8894
--------------------------------------------------------------------------------
adults - immediate effect - twosided tests
>
effect size                 .0000  .0025  .0050  .0075  .0100  .0125  .0150  .0175  .0200
power
- immediate effect model    .0572  .2645  .6476  .9138  .9908  .9995 1.0000 1.0000 1.0000
- gradual effect model      .0534  .2168  .5248  .8016  .9582  .9948 1.0000 1.0000 1.0000
- monthly effects: any      .4731  .6981  .8838  .9706  .9950  .9997  .9999 1.0000 1.0000
- monthly effects: overall  .2799  .3387  .5444  .7880  .9436  .9932  .9992 1.0000 1.0000
--------------------------------------------------------------------------------
adults - immediate effect - onesided tests
```

```
>
effect size                 .0000  .0025  .0050  .0075  .0100  .0125  .0150  .0175  .0200
power
- immediate effect model    .0915  .3629  .7370  .9492  .9969 1.0000 1.0000 1.0000 1.0000
- gradual effect model      .0918  .3022  .6276  .8737  .9774  .9979 1.0000 1.0000 1.0000
- monthly effects: any      .6679  .8496  .9567  .9911  .9995 1.0000 1.0000 1.0000 1.0000
------------------------------------------------------------------------------------------

adults - gradual effect - twosided tests
>
effect size                 .0000  .0025  .0050  .0075  .0100  .0125  .0150  .0175  .0200
power
- immediate effect model    .0549  .1360  .2762  .4772  .6890  .8404  .9379  .9815  .9960
- gradual effect model      .0527  .1652  .3514  .6046  .8224  .9396  .9862  .9977  .9998
- monthly effects: any      .4687  .6089  .7299  .8439  .9253  .9728  .9913  .9984  .9999
- monthly effects: overall  .2756  .3054  .3748  .5149  .6712  .8155  .9167  .9761  .9956
------------------------------------------------------------------------------------------

adults - gradual effect - onesided tests
>
effect size                 .0000  .0025  .0050  .0075  .0100  .0125  .0150  .0175  .0200
power
- immediate effect model    .0857  .2046  .3718  .5782  .7688  .8994  .9635  .9896  .9974
- gradual effect model      .0857  .2344  .4616  .7029  .8789  .9668  .9933  .9988 1.0000
- monthly effects: any      .6543  .7739  .8699  .9366  .9730  .9941  .9984  .9999 1.0000
------------------------------------------------------------------------------------------
```

We see that the power values are generally somewhat higher than in the WLS simulation, but the overall picture is similar.[16] I therefore refrain from a detailed discussion of these results. However, Table 1 and Table 2 present a summary from both power simulations for minors and adults, respectively, and also contain the results from Kaul and Wolf (2014a,b) and Laverty et al. (forthcoming) for the purpose of comparison. From my simulations I include results presuming an immediate effect (columns labeled "Immediate") and results presuming a gradual effect (columns labeled "Gradual"). For both cases results from the WLS model (based on the first set of simulations) and from the Logit model (based on the second set of simulations) are listed. Furthermore, I report the results for whether any monthly effect is significantly negative (column labeled "Any monthly effect") and for whether the treatment indicator from the immediate effect model (for the simulations presuming an immediate effect) or the spline parameter from the gradual effect model (for the simulation presuming a gradual effect) is significantly negative (column "Treatment effect model"). All results are based on one-sided tests at a nominal significance level of 5 percent. I report the results from one-sided tests

[16] Curiously, the type I error rates in the adult sample are somewhat inflated. This seems to be due to peculiarities of the data at hand, in particular, to the combined effect of the extra variance in monthly prevalence estimates (reflected in the u_t error term) and the high variance in monthly sample sizes in the treatment period (see Figure 1 in Kaul and Wolf 2014a). If the extra variance term u_t is omitted or if the simulation for adults is based on the minors' sample sizes, the simulated error rates match the nominal rates (results not shown). Probably, treating the sample sizes as random instead of fixed in the simulations or using a random-effects Logit for the tests would also fix the problem.

Table 1: Summary of power analyses for minors

Effect size in pct points	Kaul&Wolf	Any monthly effect				Treatment effect model			
		Immediate		Gradual		Immediate		Gradual	
		WLS	Logit	WLS	Logit	WLS	Logit	WLS	Logit
0.00		47	42	47	42	5	5	5	5
0.25	58	55	53	52	48	10	13	8	9
0.50	67	63	63	56	54	18	25	12	15
0.75	75	72	73	60	59	30	43	17	23
1.00	**83**	79	**81**	65	65	44	63	24	33
1.25	**88**	**84**	**89**	71	71	59	**81**	30	46
1.50	**93**	**90**	**94**	75	77	73	**93**	40	59
1.75		**94**	**97**	79	**84**	**84**	**98**	49	71
2.00		**97**	**99**	**83**	**89**	**92**	**99**	58	**83**

Power in percent from one-sided tests at the 5% level; power values over 80% highlighted.

because we are interested in whether plain packaging *decreases* smoking prevalence. After all, this is why plain packaging has been introduced. The choice of the test is not a matter of whether an opposite effect is possible or not. It rests on an a-priori decision to look only for evidence for an effect in one direction, which seems appropriate in the present context.

In Table 1 for minors we see that the power values computed by Kaul and Wolf (2014b, Table 3), assuming an immediate effect, could be replicated using the tests for whether any monthly dummy is significantly negative. Power values reach 80% at about a treatment effect of 1 percentage point. Assuming a gradual effect, however, the treatment effect would have to be about 1.75 percentage points by December 2013 for the tests to reach a power of 80%. As mentioned before, these tests are problematic because they have an inflated error rate of around 45% due to multiple testing. Using tests that have the correct size (column "Treatment effect model") we see that the treatment effect would have to be larger than one percentage point in any case for power to reach 80%, or considerably larger depending on whether the effect is immediate or gradual. From these results I conclude that, for minors, only a very strong treatment effect could be detected with these data with reasonable power.

For adults, results look more favorable (Table 2). Again, assuming an immediate effect, the simulation results from Kaul and Wolf (2014a) could be qualitatively replicated.[17] My results from the tests whether any monthly effect is significantly negative approximately match the IM-2 and IM-3 results from Kaul and Wolf (2014a).[18] Also if a gradual effect is assumed, power values from these tests are hight, reaching 80% at an effect of 0.75 percentage points (WLS) or 0.5 percentage points (Logit). But again, all these tests have an inflated type I error rate of about 45 to 65 percent. That is, these tests are also quite likely to indicate an effect if there is, in fact, none. Looking at the tests that have approximately correct size (column labeled "Treatment effect model") reveals that power values reach 80% at about a treatment effect of 0.75 percentage points if the effect is immediate, and at about one percentage point if the effect is gradual. Hence, due to the larger sample size, smaller treatment effects can be detected with acceptable power for adults than for minors. Also note that a given percentage-point treatment effect corresponds to a smaller effect for adults in relative terms than for minors, because the smoking prevalence among adults is higher than among minors. From these results I conclude that the chance of detecting an effect for adults based on the given data would have been reasonably high, if the effect was around 1 percentage point or larger.

4 Remarks on the errors and issues raised by OxyRomandie

In the Annex to the letter of Mr. Diethelm from January 29, 2015, seven errors and seven issues are listed. I will briefly address these errors and issues below. The titles of the errors and issues are taken from the Annex.

4.1 Error #1: Erroneous and misleading reporting of study results

The argument is that there is a difference between "no evidence for an effect" and "evidence for no effect". The distinction is due to the fact that a statistical test tries to find evidence against a

[17] I could not replicate the results from Laverty et al. (forthcoming). It is unclear to me how these power values were computed and on which kind of test they are based.

[18] IM-1 is for the test whether the mean of the post-treatment residuals is significantly lower than the mean of the last 12 pre-treatment residuals. IM-2 is for the test whether any monthly deviation is significantly negative. IM-3 records whether IM-1 or IM-2 is significant. Since IM-2 is very likely to be significant if IM-1 is, the power values of IM-2 and IM-3 are very similar.

Table 2: Summary of power analyses for minors

Effect size in pct points	Kaul&Wolf			Laverty et al.	Any monthly effect				Treatment effect model			
	IM-1	IM-2	IM-3		Immediate		Gradual		Immediate		Gradual	
					WLS	Logit	WLS	Logit	WLS	Logit	WLS	Logit
0.00					47	67	46	65	5	9	5	9
0.25	20	64	67	11	67	**85**	58	77	22	36	14	23
0.50	45	**82**	**85**	22	**84**	**96**	69	**87**	56	74	30	46
0.75	72	**93**	**96**	37	**95**	**99**	**80**	**94**	**85**	**95**	51	70
1.00	**91**	**98**	**99**	55	**98**	>**99**	**89**	**97**	**97**	>**99**	72	**88**
1.25					>**99**	>**99**	**95**	**99**	>**99**	>**99**	**88**	**97**

Power in percent from one-sided tests at the 5% level; power values over 80% highlighted.

null hypothesis (the null hypothesis in the present context is that plain packaging has no effect, or a positive effect, on smoking prevalence). If the test fails, this does not imply that the null hypothesis is true. It only means that there was not sufficient evidence to reject it (sufficient in the sense that there is a maximum a-priori probability, typically set to 5%, of falsely rejecting the null hypothesis). It may well be that there was simply not enough data to be able to detect an existing effect. OxyRomandie does not claim that Kaul and Wolf misrepresented the results in their working papers. They accuse Kaul and Wolf of not having put enough effort into policing the use of their results by others (i.e. tobacco industry). I agree that some of the quotes provided by OxyRomandie read as "evidence for no effect" instead of "no evidence for an effect". However: (1) The difference between the two formulations is subtle and my experience is—based on teaching statistics—that people without statistical training are typically not aware of the difference. Of course, it is better to always use the correct formulation, but I do not think that it really makes a big difference (statistically trained people will appreciate the correct meaning either way, while others will probably not understand the difference). (2) I am very skeptical of whether researchers can be held responsible for monitoring the use and interpretation of their results by others. This would be an obligation that is impossible to fulfill and it would strongly discourage researchers from publishing anything. Of course, we can expect researchers to pay attention to a correct representation of their results in press releases or similar materials, if they are given the chance to do so. But we cannot make them responsible

for what is published by others and we cannot expect them to actively watch out for material misinterpreting their results.

4.2 Error #2: Power is obtained by sacrificing significance

OxyRomandie claims that the high power for detecting an existing effect reported by Kaul and Wolf (2014a,b) is due to an inflated type I error rate. This is certainly true, as the power analyses above illustrate. However, this does not change the fact that the statistical procedure used by Kaul and Wolf does have the power they claim it has (given their assumptions about the nature and size of the effect). It only means that their procedure also has a high probability of falsely finding evidence for an effect if there is none.

4.3 Error #3: Inadequate model for calculating power which introduces a bias towards exceedingly large power values

OxyRomandie claims that the assumption of an immediate effect, on which the power analyses by Kaul and Wolf (2014a,b) are based, is unrealistic, and that it would be more realistic to assume a gradual effect. I do agree that a gradual effect appears more realistic and that it is more difficult to detect a gradual effect. However, Kaul and Wolf (2014a,b) are very clear about what their assumptions are and anyone is free to provide alternative analyses based on other assumptions.

4.4 Error #4: Ignorance of the fact that disjunctive grouping of two tests results in a significance level higher than the significance level of the individual tests

In their power analyses, Kaul and Wolf (2014a,b) combine two tests (whether the mean of the treatment-period residuals is significantly lower than the mean of an an equal number of pre-treatment residuals, and whether any monthly treatment-period residual is significantly negative). OxyRomandie claims that this combination of tests increases the significance level (i.e. the type I error rate). This is true, although the effect is not particularly strong in the present case (because the two tests are not independent: the second test is very likely to be significant if the first test is). However, again, this only means that there is an increased risk of finding an

effect if there is none. It does not render the power values reported by Kaul and Wolf (2014a,b) invalid.

4.5 Error #5: Failure to take into account the difference between pointwise and uniform confidence intervals

This point addresses the problem that testing 13 (or 12) monthly deviations leads to high power at the expense of an inflated type I error rate and that a global test that maintains the nominal error rate (e.g. using uniform confidence intervals instead of pointwise confidence intervals) has less power. Essentially, this is the same argument as in "Error #2" and "Error #4". The argument is undoubtedly true, but, again, it does not render the power values reported by Kaul and Wolf (2014a,b) invalid.

4.6 Error #6: Invalid significance level due to confusion about one-tail vs. two- tail test

Kaul and Wolf (2014a) report tests for monthly deviations based on 90% confidence intervals and based on 95% confidence intervals. OxyRomandie argues that watching for a 90% confidence interval to be entirely negative is equivalent to a one-sided test at the nominal 5% significance level. I agree. Kaul and Wolf (2014a), however, interpret their 90% confidence interval approach as a two-sided test at the nominal 10% significance level. This is also correct, as long as both cases, an entirely negative confidence interval and an entirely positive confidence interval are watched for. Therefore, the dispute is over whether a two-sided test or a one-sided test is appropriate. Like OxyRomandie I consider a one-sided test appropriate in the present context (see above), but it is essential to realize that there is no right or wrong here. Whether to use a one-sided test or a two-sided test is an a-priori decision made by the researchers. However, if the researchers decide to use a one-sided test they need to consistently adhere to this decision regardless of the analysis' results. That is, by deciding to do a one-sided test for a negative effect we restrict ourselves to find a negative effect and forgo the possibility to do detect a positive effect. I think in the present context such a restriction is justifiable, but others might disagree. In fact, Kaul and Wolf, in their reply to the Annex by OxyRomandie, argue that we should not forgo this possibility.

However, note that the analyses performed by Kaul and Wolf are (mostly) in line with a one-sided setting, not a two-sided setting. In particular, in their power simulations only

significantly negative deviations are counted. Hence, the computed power values are valid for one-sided tests at the 5% level, not for two-sided tests at the 10% level. The point is illustrated by the fact that in my "two-sided" simulations above the type I error rates are around 2.5% and not 5% ("immediate effect model" and "gradual effect model"). That is, the reported power values are really power values for one-sided tests at the 2.5% and not power values for two-sided tests at the 5% levels, as the tests only look for negative effects. Only the results reported for the overall test for monthly deviations are for two-sided tests at the 5% level. This may all be somewhat confusing, but the bottomline is that indeed there is an inconsistency in Kaul and Wolf (2014a): The power analyses they perform use one-sided tests, but their interpretation of the December 2012 effect is in terms of a two-sided test. For reasons of consistency, it would be more appropriate to write something along the lines of "if we are willing to use a one-sided test" instead of "if we are willing to accept a relatively low level of statistical significance" (cf. abstract and conclusions in Kaul and Wolf 2014a). However, I would not consider this a "fundamental flaw".

4.7 Error #7: Invalid assumption of long term linearity

OxyRomandie argues that the assumption of a linear trend in the pre-treatment period is invalid and evidence that there are some significant deviations from the linear trend is provided. From my own analysis I cannot support the claim that a linear fit is inappropriate. There is some extra variation in the monthly estimates in addition to what we would expect from a binomial distribution, but the deviations appear unsystematic (i.e. not suggesting a specific alternative longtime trend) and it would be difficult to come up with a convincing alternative trend model.

4.8 Issue #1: Avoiding evidence by post-hoc change to the method

OxyRomandie wonders why December 2012 was excluded from the set of treatment-period months in the power simulations in Kaul and Wolf (2014a), while it was included in the power simulations in Kaul and Wolf (2014b). I agree that it is odd to exclude December 2012, as plain packaging came into effect in December 2012. Excluding December 2012 from the power simulations, however, is not critical (as can be seen in Table 2 above, comparing Kaul and Wolf's results with my results that include December 2012). I guess what Kaul and Wolf try to say is: yes, we do find some evidence for an effect in December 2012, but if there was a lasting

effect it should also show up in the remaining 12 months. The approach followed by Kaul and Wolf (2014a) does appear a bit post-hoc, though.

4.9 Issue #2: Unnecessary technicality of the method, hiding the methodological flaws of the papers

OxyRomandie claims that Algorithm 3.1 in Kaul and Wolf (2014a) could be simplified (evaluating whether a monthly estimate lies below the confidence interval around the trend line instead of constructing a confidence interval around the monthly residual and evaluating whether the upper bound is below zero). Obviously, the approach by Kaul and Wolf and the approach by OxyRomandie are formally equivalent. Alternatively, we could also directly compute a confidence interval around zero and see whether the residual lies below the lower bound. In my analysis above I used yet a another (slightly different) approach, testing monthly dummies against zero in a linear model. I do not consider the approach proposed by Kaul and Wolf (2014a) particularly complicated. It is quite straightforward, as is the approach proposed by OxyRomandie.

4.10 Issue #3: Very ineffective and crude analytic method

The Annex presents a graph from a power analysis showing that a "simple *t*-test" outperforms Kaul and Wolf's test strategy based on confidence intervals. It is entirely unclear to me what kind of *t*-test the authors employ here, nor can I link the presented power values for Kaul and Wolf's approach back to the results published in the working papers by Kaul and Wolf.

Although I am puzzled about what OxyRomandie tries to illustrate here, I do agree that better test approaches exist than the one used by Kaul and Wolf. However, it is not an issue of lack of power, it is an issue of inflated type I error (see above).

4.11 Issue #4: Non standard, ad-hoc method

OxyRomandie accuses Kaul and Wolf of using analytic techniques "that are far from standard" and that their power analyses require writing a "computer program". I do not see the problem. All steps taken by Kaul and Wolf are simple and transparent and their analyses are easy to

replicate. No fancy tools are needed. Power analyses will always require a bit of programming except for very basic problems for which canned solutions exist.

4.12 Issue #5: Contradiction and lack of transparency about the way data was obtained

Here OxyRomandie accuses Kaul and Wolf of providing unclear and inconsistent information about how they acquired their data, and of falsely claiming that the data are "publicly available". It is not really clear to me what exactly OxyRomandie's critique is. In their papers, Kaul and Wolf write that the data are from Roy Morgan, that PMI (Philip Morris International) provided funding, and that the data for adults were received through PMI, but not the data for minors. My reading of this is that the data for minors were received directly from Roy Morgan.

OxyRomandie claims that the data are not "publicly available" because they have to be paid for. My observation is that there are many "publicly available" dataset for which one has to sign a data contract and pay a fee. Maybe, however, it is a question of how high the costs are, which is unknown to me.

At a superficial level, if one wants to save the costs of buying the data, the analyses by Kaul and Wolf can be replicated by reverse engineering the data from Figures 1 and 2 in their papers. This will provide the raw data points that were used in their analyses. However, as mentioned in Section 2 there is no information about how these data points were created from the original, individual-level survey data. A full replication, ideally, would use these original datasets.

4.13 Issue #6: Conflict of interest not fully declared

OxyRomandie accuses Kaul and Wolf of not having been fully transparent about the role of PMI. Kaul and Wolf did declare that PMI provided funding. In my opinion, this clearly identifies the papers as industry sponsored research. Whether PMI, by contract, had the possibility to comment on the papers prior to planned publication or not, does not really appear essential to me. Things might be different if the contract gave PMI a right of veto against publication (which does not seem to be the case according to the quote from the contract presented by OxyRomandie).

In general, the problem with industry sponsored research might not be so much that single studies are biased or flawed. A much bigger problem, in my opinion, is that industry funding biases the selection of studies that are conducted and that unfavorable results are often withheld,

leading to publication bias. In the present case it does not seem that PMI could have withheld publication if results would have been unfavorable, but we do not really know.

On a different note, one might be tempted to argue that tobacco industry fighting plain packaging is an indirect proof for the efficacy of plain packaging to reduce smoking prevalence (why bother, if it doesn't do anything?). However, this is not necessarily true. First and foremost, branding is an instrument to secure or increase market shares in a competitive market. It may well be that certain measures do not affect overall prevalence, but that they affect which brands people consume. Hence tobacco companies have an interest in being able to use the instrument even if it has no effect on overall prevalence. On the other hand, of course, declining overall prevalence reduces the size of the company's potential market.

4.14 Issue #7: Lack of peer review

OxyRomandie criticize that the working papers were published without having gone through peer review. All I can say is that it is standard practice in economics to make results available as a working paper before submission to a peer reviewed journal.

Whether Kaul and Wolf assisted in an orchestrated campaign of tobacco industry to promote the papers I cannot judge.

5 Conclusions

OxyRomandie requested the retraction of the two working papers by Kaul and Wolf from the website of the University of Zurich. Generally speaking, I do not recommend retraction of research papers that seem to be flawed. Science is based on information, this includes information on things that went wrong. Retracting a paper deemed as flawed would remove an essential piece of information. This is particularly relevant after a public debate has taken place.

Regarding the papers by Kaul and Wolf examined in this report, I come to the following conclusions:

- Although I am not happy with all aspects of the papers (see, e.g., Section 2), I do not think that the papers are fundamentally flawed from a methodological point of view. I do not suggest their retraction. There is some space for improvement and some of the interpretations by Kaul and Wolf might be challenged, but these are all issues that can

be resolved through usual scientific discourse. Moreover, the studies provide enough information about data, methods, and procedures to be replicable by other researchers.

- Due to the high relevance of the topic, however, it might be reasonable to add a note (either on the website providing the working papers or directly within the working papers) stating that these studies have been discussed controversially (including reverences to relevant documents).

References

Jann, B. 2005a. moremata: Stata module (Mata) to provide various functions. Statistical Software Components S455001 (available from https://ideas.repec.org/c/boc/bocode/s455001.html).

———. 2005b. Making regression tables from stored estimates. *The Stata Journal* 5(3): 288–308.

———. 2014. Plotting regression coefficients and other estimates. *The Stata Journal* 14(1): 708–737.

Kaul, A., and M. Wolf. 2014a. The (Possible) Effect of Plain Packaging on Smoking Prevalence in Australia: A Trend Analysis. University of Zurich Department of Economics Working Paper No. 165 (available from http://www.econ.uzh.ch/static/workingpapers.php).

———. 2014b. The (Possible) Effect of Plain Packaging on the Smoking Prevalence of Minors in Australia: A Trend Analysis. University of Zurich Department of Economics Working Paper No. 149 (available from http://www.econ.uzh.ch/static/workingpapers.php).

Laverty, A. A., P. Diethelm, N. S. Hopkinson, H. C. Watt, and M. McKee. forthcoming. Use and abuse of statistics in tobacco industry-funded research on standardised packaging. *Tobacco Control* .