A simulation-based comparison of two methods for determining the treatment effect in children diagnosed with hyperkinetic disorder

Iachina, Maria; Morling, Peter

Published in:
Communications in Statistics: Simulation and Computation

DOI:
10.1080/03610918.2014.960090

Publication date:
2020

Document version:
Accepted manuscript

Citation for pulished version (APA):
Iachina, M., & Morling, P. (2020). A simulation-based comparison of two methods for determining the treatment effect in children diagnosed with hyperkinetic disorder. *Communications in Statistics: Simulation and Computation*, *49*(6), 1385-1396. https://doi.org/10.1080/03610918.2014.960090

A simulation-based comparison of two methods for determining the treatment effect in children diagnosed with Hyperkinetic Disorder

Maria Iachina and Peter Morling
Center for Clinical Epidemiologi,
Odense University Hospital

corresponding author: Maria Iachina

affiliation: Center for Clinical Epidemiologi,

Odense University Hospital,

Sdr. Boulevard 29, DK-5000 Odense C, Denmark

maria.iachina@ouh.regionsydanmark.dk

short title: Estimation of treatment effect in HKD children

Abstract.

In order to show the effect of treatment, the change between two repeated psychometric measurements at the individual level should be estimated. The simplest method is to calculate the absolute difference between two measurements. However, measurements obtained in a clinical setting are often influenced by other changes not related to the treatment. One of the typical sources of error is regression to the mean (RTM). Iachina and Bilenberg (2012) propose a new method to adjust to the RTM effect. In this work, we will evaluate the performance of absolute difference method and adjusted method in a simulation study.

1. Introduction

In medical science, studies are frequently designed to investigate changes in a specific parameter before and after treatment. The measurements in a clinical setting are often influenced by measurement errors, which leads to regression to the mean (RTM). RTM effect occurs whenever a non-random sample is selected from a population and two variables that are imperfectly correlated are measured. The less correlated the two variables the larger the effect of RTM and the more extreme the value is from the population mean, the more room there is to regress to the mean. This occurs when an extreme group is selected from one variable and then another variable is measured (Bland and Altman, 1994). Thus, as Campbell and Kenny (Campbell and Kenny, 1999) pointed out, RTM is the difference between perfect correlation and observed correlation. The practical problem caused by RTM is the need to distinguish a real change due to treatment effect from this expected change due to the natural variation.

In the literature, several different methods are used to estimate a real change due to treatment. In practice, in most medical papers a change is estimated in the most simple way, as a difference between the measurement before and after the treatment (Twisk, 2003), i.e. ignoring the RTM effect. The goal of this work is to evaluate the performance of the simplest estimating method and compare it to the performance of a more advanced method using a simulation study. The simulations will be constructed in orden to roughly approximate the real data from the Danish CAMHS-database. The Danish CAMHS database is described in details in (Bilenberg et al., 2001)

2. Regression to the mean

This artefact was first mentioned by Galton (Galton, 1886), he called this phenomenon "regression towards mediocrity", and we now call it "regression to the mean".

Let X_1 and X_2 be random variables with joint distribution function F. Assume that X_1 and X_2 have the same marginal distribution and let μ denote their common mean. The distribution F

exhibits regression to the mean if for all $c > \mu$,

$$\mu \leqslant E[X_2|X_1 = c] < c,$$

and for all $c < \mu$,

$$c < E[X_2|X_1 = c] \leqslant \mu.$$

For simplicity, we assume that the measurements have a bivariate normal distribution with a common variance σ^2 and correlation ρ. Then for a given $X_1 = x$ the expected value of X_2 will not be equal to

$$E[X_2|X_1 = x] = x - (1 - \rho)(x - \mu)$$

This equation shows the presence of RTM effect. It depends of both values of $x - \mu$ and of correlation, ρ.

In a clinical study a sample of n patients is taken from a population and compares the average of the pretreatment measurement with the average of posttreatment measurement in order to estimate the treatment effect τ. If the sample is random sample from a population the difference between averages will be an unbiased estimate of the treatment effect. However, most clinical studies use the samples which consist of people who were hospitalised or who needs treatment. Clealy, this subset of population is not a random sample. In this case the difference between averages does not estimate the true treatment effect, but instead

$$E[\bar{X_2} - \bar{X_1}|X_1 \geqslant c] = \tau - (1 - \rho)\frac{\sigma\phi(\frac{c-\mu}{\sigma})}{1 - \Phi(\frac{c-\mu}{\sigma})},$$

where ϕ and Φ are density and cumulative density function for normal distribution. Then more extreme observations defined at the first time point will, in the next point of time be closer to the mean even if no real changes occurred (Galton, 1886; Yudkin and Stratton, 1996; Bland and Altman, 1994; Campbell and Kenny, 1999). This RTM confounds the interpretation of score change (Allison et al., 2009).

3. Estimating models

Suppose there are N hospitalised persons, M of those have one measurement of interest before treatment and one measurement after the treatment. Other $N - M$ persons have at least two pre-treatment measurements. Let Y_{it} denote response variabel or measurement of interest for subject i at time t. Using pre-and post-treatament measurements of the response variable the treatment effect can be estimated in different approaches. In the following we will describe two of those.

absolute difference

The simplest approach to estimate the treatment effect, and also mostly used in the clinical trials, is to calculate the absolute difference between two measurements overtime,

$$\tau = E[Y_{i1} - Y_{i2}],$$

for all $i = 1, ..., M$ i.e. ignoring the RTM effect.

adjusted difference

Using the patients who have two pre-treatment observations estimate the RTM effect fitting the following linear regression:

$$Y_{i2} = const + aY_{i1} + \epsilon_i,$$

where $\epsilon < N(0, 1)$

Then we can adjust for this effect constructing a new corrected follow-up score:

$$Y_{i2} = \frac{1}{a}(Y_{i2} - const),$$

where *a* and *const* are estimatetes from the previouse equation.

4. The real data

In order to show the effect of treatment in the naturalistic setting of child and adolescent mental health services the change of Health of the Nation Outcome Scales for Children and Adolescents (HoNOSCA) score at the individual level should be estimated. HoNOSCA is a brief scale for measuring emotional and behavioural problems in children and adolescents, 4-18 years of age, and is to be completed by multidisciplinary staff. HoNOSCA comprises 13 subscales or items measuring disruptive/aggressive behaviours, overactivity/concentration problems, self injury, substance misuse, scholastic skills, physical illness, hallucinations/delusions, non organic somatic symptoms, emotional symptoms, peer relationships, self care, family relationships, and school attendance. All items are scored on a 0-4 point, scale ranging from "no problems" (0) to "severe problems" (4). Thus the total score ranges from 0 to 52. HoNOSCA is an established measure of outcome in child and adolescent mental health services (CAMHS) (Gowers et al., 1999). The score has been evaluated in several clinical studies (Bilenberg, 2003) and (Lesinskiene et al., 2007). This database was established in 2003 in order to improve and evaluate treatment outcome. For every single patient referred, a number of variables are registered at referral, at the first meeting with the patient and family, and at the end of treatment. Changes between post and pre treatment measurements of HoNOSCA scores will reflect the treatment effect (Burgess et al., 2009). At present the Danish CAMHS-database consists of a total of 4,961 persons who had a verified Hyperkinetic Disorder (HKD) diagnosis. The following study will be based only on persons who participated in two interviews and who have a complete assessment at both interviews, which leaves a total of 4,249 persons. The distribution of the initial HoNOSCA score is illustrated in Figure 1.

Approximately 10% of the study population does not getting any treatment, due to children/parents not being interested in pharmacologic intervention and therefore ending services, families seeking treatment elsewhere, or children from families moving to another district etc. There are no signifi-

cant differences in the prognostic variables between this group of children and all other children in the study sample at the time of the first assessment. So we assume that population of children who does not receive any treatment is a random subpopulation of all children with HKD.

5. Design of simulation study

We are interested in evaluating the behaviour of estimates from two different approaches. For this purpose we performed a simulation study based on a Danish CAMHS-database. We would like to simulate three different data scenarios. Firstly, we assume that all children with HoNOSCA score above 13.6 are diagnosed with hyperkinetic disorder. Secondly, we assume that all children with HoNOSCA score above 13.6 is sick with the diagnosis, and additionally 80% of children with HoNOSCA scores between 13.6 and 9, 50% of children with HoNOSCA scores between 9 and 6, and 10% of children with HoNOSCA scores below 6 are diagnosed with hyperkinetic disorder as well. These scenarios may be a little more realistic than the first one because clinicians usually use other criteria than the HoNOSCA score when they diagnose children with hyperkinetic disorder. We simulated the third scenario so that the simulated score distribution similar as possible to the observed. The histograms of the simulated score distributions can be seen in figure 2.

In the following simulation study, we focus on settings where treatment effect $b = 0, 1, 2, 3$ and number of individuals $n = 5000$. Simulations were done using 1000 replications. In every replication sampling of response variable at the time 1 was generated in the following way

$$Y_{i1} = X_i + r_i$$

where $X_i \sim N(10.5, 8.5)$ and r_i denote a personal variability, $r_i \sim N(0, s)$, her $s \in (0, 5)$.

The conditional sampling was generated from the following scenarios,

Scenario I

$$p = 1 \text{ if } Y_{i1} \geqslant 13.6$$

$$p = 0 \text{ if } Y_{i1} < 13.6$$

Scenario II

$$p = 1 \text{ if } Y_{i1} \geqslant 13.6$$

$$p = 0.8 \text{ if } Y_{i1} < 13.6 \text{ and } X_{i1} \geqslant 9$$

$$p = 0.5 \text{ if } Y_{i1} < 9 \text{ and } X_{i1} \geqslant 6$$

$$p = 0.1 \text{ if } Y_{i1} < 6 \text{ and } X_{i1} \geqslant 0$$

$$p = 0 \text{ if } Y_{i1} < 0$$

Scenario III

$$p = 1 \text{ if } Y_{i1} \geqslant 13.6$$

$$p = 0.99 \text{ if } Y_{i1} < 13.6 \text{ and } X_{i1} \geqslant 11$$

$$p = 0.9 \text{ if } Y_{i1} < 11 \text{ and } X_{i1} \geqslant 10$$

$$p = 0.8 \text{ if } Y_{i1} < 10 \text{ and } X_{i1} \geqslant 9$$

$$p = 0.5 \text{ if } Y_{i1} < 9 \text{ and } X_{i1} \geqslant 6$$

$$p = 0.1 \text{ if } Y_{i1} < 6 \text{ and } X_{i1} \geqslant 3$$

$$p = 0.01 \text{ if } Y_{i1} < 3 \text{ and } X_{i1} \geqslant 0$$

$$p = 0 \text{ if } Y_{i1} < 0$$

The response variable at the time 2 was defined in the following way

$$Y_{i2} = Y_{i1} + r_i \text{ if } k < -1.3$$

$$Y_{i2} = Y_{i1} + r_i + b \text{ if } k > -1.3$$

where b represents the treatment effect and $b \in (0, 1, 2, 3)$, $k \sim N(0, 1)$. On that way we have simulated that approximately 10% individuals will be in the non treatment group.

There were simulated 1000 realizations for scenario I, II and III for $b = 0, 1, 2, 3$ and for $s \in (0, 5)$ in 0.5 intervals, then using absolute and adjusted approaches we estimated the treatment effect. The estimated averages and 95% interfractile intervals for the treatment effect are graphically depicted i figures 3-5. Moreover, we simulated 1000 realizatoins of data set for scenario III $b = 3$ where $N = 500$ and $N = 50000$. The results of the simulation are shown in figure 6.

6. Random number generator

We used STATA version 12 to develop our simulation and although STATA runs in an interpreted environment, which is somewhat slower than optimized FORTRAN or C code, it provides routines for graphing and the running of simulations which is suitable for a simulations study.

Random number generators (RNG) are used everywhere in today's software, e.g. in computer games, stock exchange algorithms, heuristic optimization algorithms, simulation studies etc. Clearly RNG has an important role in software development. However, history has shown several potential pitfalls in generating random numbers and care should be taken upon the selection of a good RNG. For this selection the sequence of randomly generated numbers should at least satisfy the following properties: the sequence should be uniform, it should be independent and reproducible, and it should have a long period so that the sequence does not start to repeat itself (Knuth, 1997; Klimasauskas, 2002; Jones, 2010).

To meet the above criteria and to ensure our results would not rely on a single RNG, we run experiments using three different RNG: 1) In STATA the built-in RNG is "The KISS generator" (Marsaglia and Zaman, 1993; StataCorp, 2011) that has a period length that lies above 2^{123}. Additionally, we used 2) "Mersenne Twister" (Matsumoto and Nishimura, 1998) that has a very long period of $2^{19937} – 1$, and 3) "Mother-of-all" (Marsaglia, 1994) that has a period length that lies above 2^{59}. For both of these we made re-implementations of a C++ class library of random num-

ber generators provided by A. Fog (Fog, 2010). This was done using GCC/C++ to create shared objects, also known as Stata Plugins, to link with STATA. These 1-3) are all commonly known RNG that passes numerous tests for statistical randomness, including the Diehard tests (Marsaglia, 1997).

In a single simulation we use RNG to generate random numbers from the uniform (0,1] distribution to create the population samples of interest. In the worst case we draw $4 \times 10^8 < 2^{29}$ random numbers in a sequence. That is, for each of 50.000 observations, we draw 8 different random numbers for the creation of population samples, finally this is repeated 1000 times.

When we run experiments, the variables shown in table 1 and described in above sections are permuted in a nested loop giving us $3 \times 3 \times 4 \times 11 = 396$ combinations of arguments which is set to simulate. At the end of each simulation mean values are calculated and appended to an output file together with the current set of arguments, following the guidelines given in (Burton et al., 2006).

To initialize the RNG we tried with different random seed values which had no influence on our results and we have run all simulations with $seed = 123$.

We simulated also 1000 realizatoins of data set for scenario III $b = 3$ where $N = 5000$ using three different Random Number Generators. The results of the simulation are shown in figure 7.

7. Results of the simulatin study

Figures 3-5 shows the estimated treatment effect and 95% interfractile intervals estimated by absolute and adjusted methods for the different scenarios and $b = 0, 1, 2, 3$, $s \in (0, 5)$.

First we will underline,that all figures show that both treatment estimates and the 95% interfractile intervals estimated by absolute and adjusted methods are equal and unbiased for $s = 1$, i.e. then there are no persons variability in the measurement of HoNOSCA score. Figures shows that estimational results for the both absolute and adjusted method for all three scenalios are very similar.

Scenario I: For b equal to zero, i.e. no treatment effect, the absolute method estimates the

highest bias for s equal to 5 (-2.65 with 95% interfractile interval (-2.9;-2.4)) and the adjusted method estimates the highest bias also for s equal to 5 (-0.02 (-1.1; 1.2)). For b equal to one, the absolute method estimates the highest bias for s equal to 5 (-2.6 (-2.9;-2.3)) and the adjusted method estimates the highest bias also for s equal to 5 (0.3 (-0.2; 1.7)). For b equal to two and three again both methods estimate the highest bias for s equal to 5. The absolute method estimates bias equal to -2.6 (-2.9; -2.3) for b =2 and corresponding -2.6 (-2.9; -2.4) for b =3. The adjusted method estimates bias equal to 0.7 (-0.5; 2.1) for b =2 and corresponding 1.7 (-0.2; 2.6) for b =3.

Scenario II: For b equal to zero, i.e. no treatment effect, the absolute method estimates the highest bias for s equal to 5 (-1.6 with 95% interfractile interval (-1.8;-1.4)) and the adjusted method estimates the highest bias also for s equal to 5 (-0.9 (-1.4; 0.9)). For b equal to one, the absolute method estimates the highest bias for s equal to 5 (-1.6 (-1.8;-1.4)) and the adjusted method estimates the highest bias also for s equal to 5 (0.3 (0.4; 1.3)). For b equal to two and three again both methods estimate the highest bias for s equal to 5. The absolute method estimates bias equal to -1.6 (-1.8; -1.4) for b =2 and corresponding -1.6 (-1.8; -1.4) for b =3. The adjusted method estimates bias equal to 0.7 (-0.5; 2.2) for b =2 and corresponding 1.1 (0.1; 2.2) for b =3.

Scenario III: Estimating results for the scenario III are almost identical to estimating results from the scenario II. For b equal to zero, i.e. no treatment effect, the absolute method estimates the highest bias for s equal to 5 (-2.65 with 95% interfractile interval (-2.8;-2.4)) and the adjusted method estimates the highest bias also for s equal to 5 (-0.01 (-0.9; 0.9)). For b equal to one, the absolute method estimates the highest bias for s equal to 5 (-2.6 (-2.8;-2.4)) and the adjusted method estimates the highest bias also for s equal to 5 (0.3 (-0.6; 1.3)). For b equal to two and three again both methods estimate the highest bias for s equal to 5. The absolute method estimates bias equal to -2.6 (-2.8; -2.4) for b =2 and corresponding -2.6 (-2.8; -2.4) for b =3. The adjusted method estimates bias equal to 0.7 (-0.3; 1.7) for b =2 and corresponding 1.1 (0.1; 2.1) for b =3.

For all three scenarios and for all b absolute approach estimates are biased and bias grows for $s > 0$.

Figure 6 shows the results of a simulation study for $n = 500$, $n = 5000$ and $n = 50000$, figure 7 shows the results of a simulation study for $n = 5000$ using three different Random Number Generators. According to these figures, the Random Number Generator and the sample size does not seems to have any influence on the treatment estimate and 95% interfractile intervals became less for bigger sample size, as expected.

8. Conclusion

From the results of the simulation study we can see that bias estimated for scenario I is bigger than for scenarion II and III. The treatment effect estimated by adjusted method is unbiased for $b = 0$, and slightly biased for $b = 1, 2, 3$ bias grows for s bigger than 1. Moreover, the 95% interfractile intervals estimated by adjusted method became bigger for bigger s.

We can conclude that the treatment effect estimated by the absolute difference approach will be biased when personal variability is different from zero. The parameters estimated by the adjusted approach are unbiased only when there is no treatment effect and bias grows when the treatment effect grows. Unfortunately, the variability of these estimates is very big for a big personal variability.

Overall, if the measurement has a large personal variability any estimate of the difference will have a big variability and can not be credible.

Using this simulation study we will not recommend to use absolute difference approach to estimate the difference between two measurements in any cases. We can recommend, to use the adjusted approach instead of absolute to estimate the difference when the personal variability of measurement is small.

References

Allison, D., A. Loebel, I. Lombardo, S. Romano, and C. Siu (2009). Understanding the relationship between baseline bmi and subsequent weight change in antipsychotic trials: Effect modification or regression to the mean? *Psychiatry Research 170*, 172–176.

Bilenberg, N. (2003). Health of the nation outcome scales for children and adolescents (honosca) - results of a danish field trial. *European Child (and) Adolescence Psychiatry 2(6)*, 298–302.

Bilenberg, N., T. Isager, and J. Buchhave (2001). Bupbase - en klinisk kvalitetsdatabase i børne- og ungdomspsykiatri. *Ugeskr Laeger 163(43)*, 6002–6004.

Bland, J. M. and D. G. Altman (1994). *Statistic Notes: Regression towards the mean*, Volume 308. British Medical Journal.

Burgess, P., T. Trauer, T. Coombs, R. McKay, and J. P. J. (2009). What does 'clinical significance' mean in the context of the health of the nation outcome scales? *Australasian Psychiatry 17(2)*, 141–148.

Burton, A., D. Altman, P. Royston, and R. Holder (2006). The design of simulation studies in medical statistics. *Statistics in Medicine 25*, 4279–4292.

Campbell, D. T. and D. A. Kenny (1999). *A Primer on Regression Artifacts*. Guilford.

Fog, A. (2010). Uniform random number generators in C++. `http://www.agner.org/random/`. [Online; accessed 7-Marts-2013].

Galton, F. (1886). Regression towards mediocrity in hereditary stature. *The Journal of the Anthropological Institute of Great Britain and Ireland 15*, 246–263.

Gowers, S., R. Harrington, A. Whitton, P. Lelliott, A. Beevor, J. Wing, and R.Jezzard (1999). Brief scale for measuring the outcomes of emotional and behavioural disorders in children. health of the nation outcome scales for children and adolescents (honosca). *Br.J.Psychiatry 174*, 413–416.

Iachina, M. and N. Bilenberg (2012). Measuring reliable change of emotional and behavioural problems in children. *Psychiatry Research 200(2-3)*, 867–871.

Jones, D. (2010). Good Practice in (Pseudo) Random Number Generation for Bioinformatics Applications. `http://www0.cs.ucl.ac.uk/staff/D.Jones/GoodPracticeRNG.pdf`. [Online; accessed 7-Marts-2013].

Klimasauskas, C. (2002). Not knowing your random number generator could be costly: Random generators - why are they important. *PC Artificial Intelligence Magazine 16*, 52–58.

Knuth, D. (1997). *The Art of Computer Programming, volume 2 - Seminumerical Algorithms, third edition*. Addison Wesley.

Lesinskiene, S., J. Senina, and N. Ranceva (2007). Use of the honosca scale in the teamwork of inpatient child psychiatry unit. *Journal of Psychiatrical Mental Health Nurs 14(8)*, 727–733.

Marsaglia, G. (1994). Yet another rng. Posted to sci.stat.math August 1, 1994.

Marsaglia, G. (1997). DIEHARD: A battery of tests of randomness. `http://stat.fsu.edu/pub/diehard/`. [Online; accessed 7-Marts-2013].

Marsaglia, G. and A. Zaman (1993). The KISS generator. Technical report, Dept. of Statistics, Florida State University.

Matsumoto, M. and T. Nishimura (1998). Mersenne twister: A 623-dimensionally equidistributed uniform pseudorandom number generator. *ACM Transactions on Modelling and Computer Simulation 7(1)*, 3–30.

StataCorp (2011). *Stata: Release 12. Statistical Software.* College Station, TX: StataCorp LP.

Twisk, J. W. R. (2003). *Applied Longitudinal Data Analysis for Epidemiology: A Practical Guide*. Cambridge.

Yudkin, P. L. and I. M. Stratton (1996). How to deal with regression to the mean in intervention studies. *Lancet 347*, 241–243.

Table 1: Description of variables and constants used running simulations in STATA 12.

Variables	Description	Values
seed	random seed	123
reps	no of repititions	1000
n	no of observations	500, 5.000, 50.000
s	personal variability	0, 0.5, 1, 1.5, 2, 2.5, 3, 3.5, 4, 4.5, 5
b	treatment effect	0, 1, 2, 3
sc	scenario	1, 2, 3

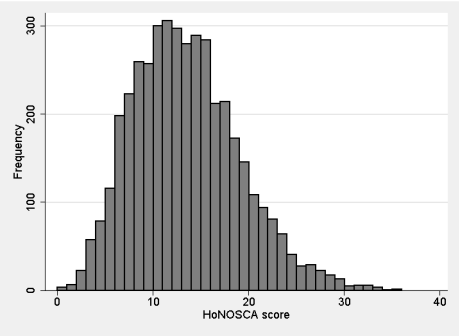

Figure 1: The distribution of HoNOSCA score measured children with HKD disorder

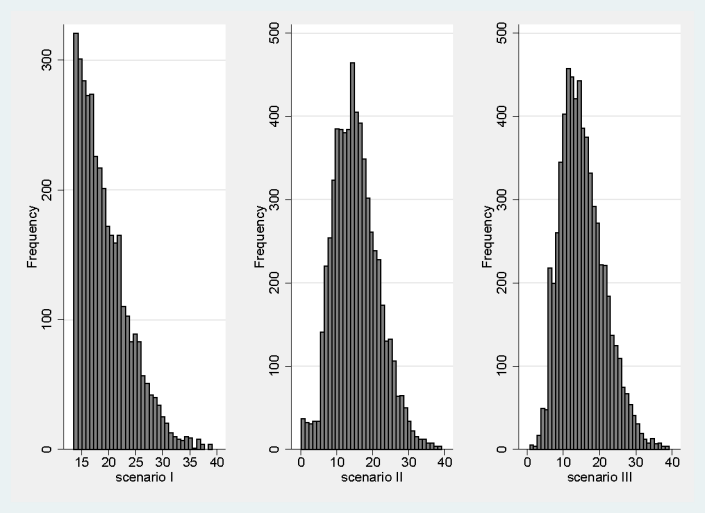

Figure 2: The simulated distributions of HoNOSCA score after scenarios

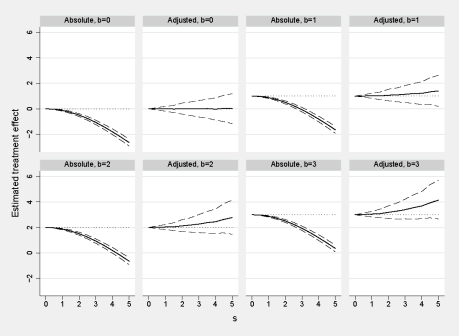

Figure 3: Simulation results for scenario I, n=5000

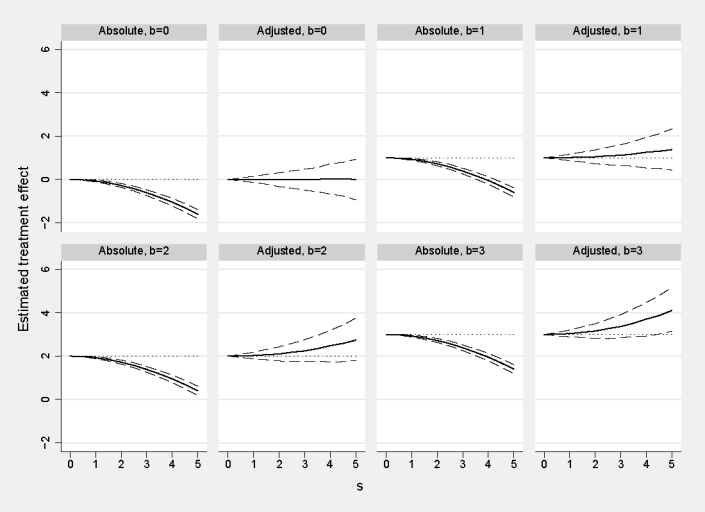

Figure 4: Simulation results for scenario II, n=5000

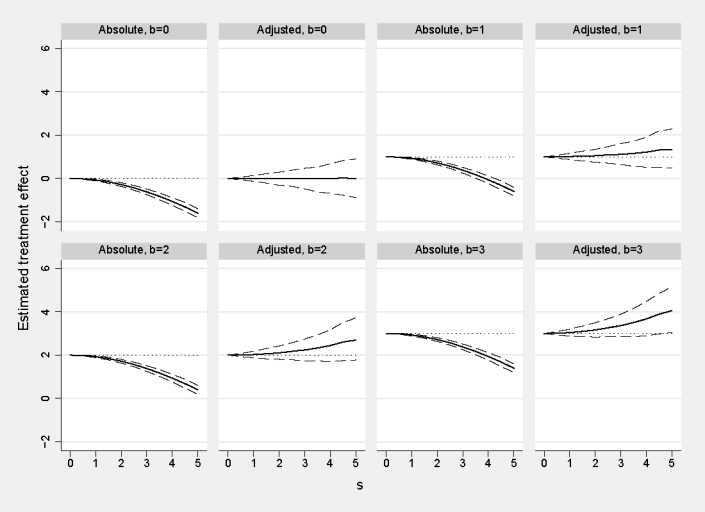

Figure 5: Simulation results for scenario III, n=5000

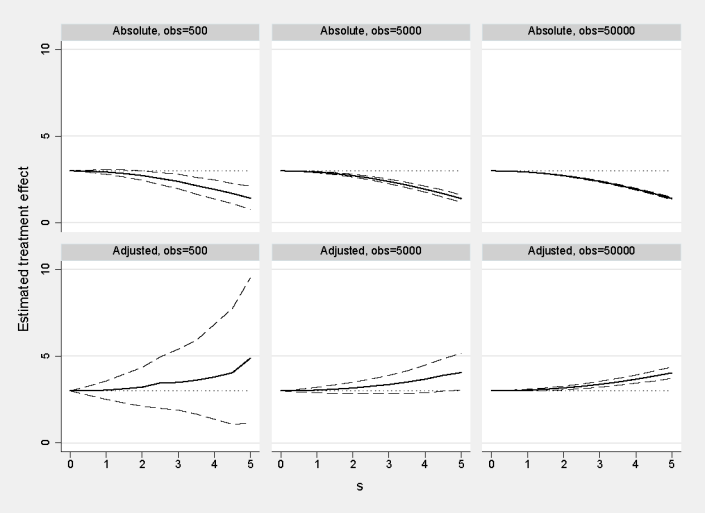

Figure 6: Simulation results for scenario III, b=3

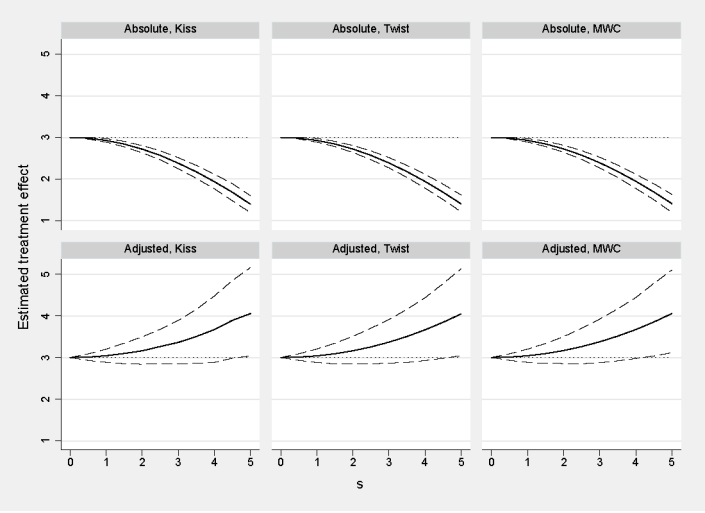

Figure 7: Simulation results for scenario III, b=3, n=5000 using three different Random Number Generators