The Remarkable Robustness of Ordinary Least Squares in Randomized Clinical Trials

David R. Judkins[1] and Kristin E. Porter[2]
[1]Abt Associates, 4550 Montgomery Ave, Suite 800 North, Bethesda, MD 20814-3343
[2]MDRC, 475 14th Street, Suite 750, Oakland, CA 94612

Abstract
There has been a series of occasional papers about robust covariate control in the analysis of clinical trials in *Statistics in Medicine* and other journals. The robust semiparametric and nonparametric methods for statistical inference of estimated effects are fairly easy to apply with 21st century computers, but many prefer to continue using t-tests and confidence intervals based on ordinary least squares (OLS) for outcomes that clearly do not follow normal distributions. Presumably, issues of tradition and communication make it very hard to deflect this inertia. In addition, recent papers have demonstrated that the tests are asymptotically equivalent, and the more complex but less parametric procedures make little difference in practice. However, in the literature, there is not sufficient examination of whether the tests and confidence intervals based on OLS are robust to substantial excess kurtosis, particularly in small sample sizes. This paper indicates through simulation where the boundaries lie for two types of strongly nonnormal outcomes: binary outcomes and compound binary/gamma outcomes. We found that traditional OLS methods work very well down to very small sample sizes for these outcomes.

Key Words: OLS, semiparametric ANOVA, nonparametric ANCOVA

1. Introduction

During the 1980s and 1990s Gary Koch and collaborators developed a body of work on nonparametric ANCOVA for categorical outcomes in clinical trials (Koch et al., 1998). Subsequently, Lesaffre and Senn (2003) published a fairly harsh review of the approach. It is important to note, however that the simulations by Lesaffre and Senn only concerned one of two methods proposed by Koch et al. Lesaffre and Senn conducted Monte Carlo simulations of the large sample approximate procedure rather than the small-sample permutation-based procedure. This was unfortunate because the small-sample permutation-based procedure is the more useful of the two. More recently, Tsiatis et al. (2008) discussed the asymptotic equivalence of linear models and Koch's methods as well as other popular robust methods. They also presented simulation studies that showed little benefit of Koch's method. A year later, Tang et al. (2009) re-analyzed some data using semiparametric procedures and also found little change from earlier published simple models.

In retrospect, these findings of little benefit from nonparametric or semi-parametric permutation tests in the contexts of large trials with iid errors should have been anticipated. Scheffé in chapter 10 (The effects of departures from the underlying assumptions) of his famous 1959 textbook states that, "The conclusion is that the effect of violations of the normality assumption is slight on inferences about means …" Table 10.3.2 of that text reproduced from Box and Anderson (1955) shows good preservation of p-values for sample sizes as small as 5 per arm in a one-way ANOVA with skew varying between 0 and 1 and excess kurtosis varying from -1 to 1. An outstanding question is whether ANOVA gives invalid inference for estimating risk difference when the excess kurtosis is more extreme and when the sample size is small. The first condition is met by binary variables with means close to 0 or 1 and by compound binary-gamma distributions where most of the mass is on 0 and there are large outliers.

In this study, we examine the performance of nominal 95 percent confidence intervals based on the OLS estimator of the standard error of a risk difference treatment effect on such outcomes in the context of randomized clinical trials. We examine both perfectly balanced trials in which the probability of being randomly assigned to treatment is 0.5 and substantially unbalanced trials in which the probability of being assigned to treatment is 0.3. This work should apply to vaccine trials for rare diseases, studies of rare side effects and to the blood concentration of substances such as virus antibodies that are rarely detected in healthy patients but may be detected in large quantities in diseased patients. We do not supply any theory. We merely demonstrate the robustness of the OLS method under different circumstances. Our work is very similar in spirit to Lumley et al. (2002), who also used a Monte Carlo to study the robustness of OLS to departures from normality. Randomization-based purists may question the worth of these endeavors saying that one ought to always use permutation tests if the outcomes are nonnormal. In some ways we agree with them, but there seems to be little evidence of the tide turning their way despite a very nice review article by Rosenbaum (2002) and two nice new textbooks (Pesarin and Salmaso, 2010, and Good, 2010). Let us note though that this paper only covers independent errors. The situation may be different in cluster randomized trials. See, for example, La Vange, Durham and Koch (2005), Fan and Judkins (2010), and Pacheco et al. (2009).

In section 2, we lay out the simulation plan, including some discussion of a problem we ran into with trials in which all participants took on the same outcome value. In section 3, we present the results. In section 4, we close with some remarks.

2. Simulation Plan

Binary Outcome

For the binary outcome, we intentionally invented a distribution for which both linear and traditional logistic models would be incorrect.[1]

$$\begin{aligned} &Y \sim \mathrm{B}\left(1, \mathrm{logit}^{-1}\left[\mathrm{logit}(\mu + T\delta) + \beta X\right]\right) \\ &T \sim \mathrm{B}\left(1, 0.5\right) \text{ or } \mathrm{B}\left(1, 0.3\right) \\ &X \sim \mathrm{N}(0,1) \end{aligned} \qquad (1)$$

In this model, T indicates randomized treatment status, X is a covariate, μ is the prevalence of the characteristic of interest, Y, in the absence of the intervention with X=0, and δ is the effect of treatment. Note that Y is strongly nonnormal and that its conditional variance given X is not constant. So it violates both the assumption of normal

[1] Analysts who rely on parametric methods often fit logistic regression models instead of linear models when outcomes are binary. As pointed out by Moore and van der Laan (2009), even if the logistical model is correct (which in practice, like the linear model, would not be the case), statistical inference for the risk difference parameter, our parameter of interest, requires nonparametric methods.

errors and the assumption of homoscedastic errors. Of the traditional ANOVA assumptions, only the assumption of independent errors is true. Also note that δ is on a scale that is pretty much impossible to communicate meaningfully. However, one can still estimate

$$\varphi \equiv \mathrm{E}\left(Y_1\right) - \mathrm{E}\left(Y_0\right) \tag{2}$$

where Y_1 is the potential outcome under treatment and Y_0 is the potential outcome under no treatment, drawing on the Neyman-Rubin causal model framework (Neyman, 1923/1990, and Rubin (1974). A first-order Taylor series approximation is

$$\varphi \approx \delta \tag{3}$$

A somewhat better approximation using a second-order Taylor series is

$$\varphi \approx \delta + \frac{\beta^2}{2}\left[\left(\mu+\delta\right)\left(1-\mu-\delta\right)\left(1-2\mu-2\delta\right) - \mu\left(1-\mu\right)\left(1-2\mu\right)\right] \tag{4}$$

We simulated four different effect magnitudes, using

$$\delta = \min\left[0.999 - \mu, c\sqrt{4\mu(1-\mu)/n}\right] \tag{5}$$

where *n* is the size of the experiment (across both arms) and *c* took a value from the set {0,1.4,2.8,3.24}, corresponding to the null hypothesis, a weak alternate hypothesis, a strong alternate hypothesis, and a very strong alternate hypothesis. We used $\beta = 0.3$. We examined coverage probabilities for the treatment effect for values of baseline prevalence from 0.5% to 50% and for sample sizes from 50 to 2000. By adapting the effect size to the sample size, we were able to keep the simulation under alternate hypotheses within the realm of reasonably powered studies. The idea was to simultaneously evaluate type I error preservation under the null hypothesis as well as behavior of the confidence interval under the alternate hypotheses. Many of the robust alternatives to linear models only provide hypothesis tests – they do not provide confidence intervals, which are fundamentally more parameter-driven. Thus, our primary interest was in $\varphi = \delta = 0$, but it seemed very easy and potentially interesting to the field to also include the performance of confidence intervals under alternative hypotheses, and so we did so. The resulting kurtosis induced is shown in Table 1.

For each of 5000 samples from the simulated population, we fit a simple linear model

$$\begin{aligned} Y_i &= \theta + T_i\varphi + X_i\beta + e_i \\ e_i &\overset{iid}{\sim} N\left(0,\sigma^2\right) \end{aligned} \tag{6}$$

We then kept track of how often the nominal 95% confidence intervals for φ produced by standard software for this simple linear model (i.e. produced by an OLS estimator of the standard error) across 5000 replications included the second order approximation to the true value of φ. With this number of replications, if the true coverage is equal to nominal coverage, then a 95% prediction interval for the empirical coverage rate runs from 94.4 to 95.6 percent.

Compound Binary-Gamma Outcome

For the binary-gamma compound outcome, we simulated

$$\begin{aligned} & Y = Z\left(G + T\delta_G 20000/\sqrt{\alpha}\right) \\ & Z \sim B\left(1, \operatorname{logit}^{-1}\left[\mu + T\delta_Z\right]\right) \\ & G \sim \Gamma\left(\alpha, 20000/\alpha\right) \\ & G \perp Z, T \end{aligned} \tag{7}$$

Here *Z* is a binary variable indicating that the person has a condition of interest, making them eligible for exact measurement of some strictly positive interval-valued attribute. In medicine, Z might indicate HIV infection and *Y* represent a CD4 count or viral load, while in labor economics, *Z* might indicate employment and Y represent an hourly wage rate. Another common medical example involves *Z* = survival, and *Y* = quality of life. Returning to other symbols in (1.7), δ_Z is the effect of treatment on the log odds of belonging to the population where the more detailed measurement is taken, and δ_G is the conditional additive effect of treatment on the more detailed measurement given that the measurement is possible. Treatment indicator, *T*, is defined as above, simulated with a mean of either 0.5 or 0.3. Note that the effect of intervention on this compound outcome is easy to calculate despite the complexity of the random variable.[2]

$$\begin{aligned} \varphi &= E\left(Y \middle| T = 1\right) - E\left(Y \middle| T = 0\right) \\ &= 20000\left[\operatorname{logit}^{-1}\left(\mu + \delta_Z\right)\left(1 + \delta_G/\sqrt{\alpha}\right) - \operatorname{logit}^{-1}\left(\mu\right)\right] \end{aligned} \tag{8}$$

Also note that

$$\begin{aligned} & E\left(Y \middle| T = 1, Z = 1\right) - E\left(Y \middle| T = 0, Z = 1\right) \\ & \quad = 20000\delta_G/\sqrt{\alpha} \end{aligned} \tag{9}$$

and that

$$\operatorname{var}\left(Y \middle| T = 0, Z = 1\right) = 20000^2/\alpha \tag{10}$$

So that the effect size from treatment given disease is

$$\frac{E\left(Y \middle| T = 1, Z = 1\right) - E\left(Y \middle| T = 0, Z = 1\right)}{\sqrt{\operatorname{var}\left(Y \middle| T = 0, Z = 1\right)}} = \delta_G \tag{11}$$

We simulated a large set of populations, allowing $\operatorname{logit}^{-1}\left(\mu\right)$ to vary from 0.05 to 0.7, α to vary from 1 to 6, *n* to vary from 10 to 2000, and the pair $\left(\delta_Z, \delta_G\right)$ to take values from the set {(0,0), (0.5,0.2), (0.75,0.3),(1,0.4)}. While we initially chose these values

[2] The simplicity is due to the lack of covariates. If we had included covariates, it would have become difficult to evaluate.

thinking of Z as employment status and G as annual earnings among employed persons, these values should also be interesting thinking about risk factors that simultaneously raise both the probability and the severity of disease. Disease could be interpreted broadly here to include alcohol and tobacco consumption, both of which involve large point masses on zero. The resulting kurtosis induced is shown in Table 2.

For each of 5000 samples from simulated population, we fit a simple linear model

$$\begin{aligned} Y_i &= \theta + T_i\varphi + e_i \\ e_i &\overset{iid}{\sim} N\left(0,\sigma^2\right) \end{aligned} \tag{12}$$

We then kept track of how often the nominal 95% confidence intervals for φ produced by standard software for this simple linear model included the true value of φ.

Failed Trials

For both the binary and the binary-gamma compound outcome, we ran into samples with Y=0 for every member of the sample. If the protocol were to accept the null hypothesis in this eventuality, then these samples could be used for limited inference. One could conclude that there was no effect of treatment, but it would not be possible to construct a sensible confidence interval for the effect of treatment. However, it is not clear that most protocols even deal with this possibility. For well-designed trials, the probability of obtaining such a sample is near zero. One consultant suggested that one response might be to extend the trial, to allow more time for events such as deaths to occur. It seems just as likely to us that trials that resulted in no events would be simply abandoned.

We considered three options. The first was to construct degenerate confidence intervals on the treatment effect of [0,0] for samples with Y≡0, i.e., no events of interest in either arm. The second was to discard samples with Y≡0. The third was to only simulate populations for which the probability of Y≡0 was vanishingly small. We decided on the second option. We also ran some simulations using the first option, but we decided that no semi-parametric technique would be better able to deal with Y≡0 than a standard ANOVA, so there was no point in reporting on inferential problems with this technique. We rejected the third option because badly designed trials are a fact of life and people do try to analyze them.

3. RESULTS

All of the tables below present results for simulations in which the treatment effect is zero and in which the treatment effect is the strongest considered ($c = 3.24$ in Equation (1.5) for binary simulations and $(\delta_Z, \delta_G) = (1{,}0.4)$ in Equation (1.8) for binary-gamma simulations). For each of these effect sizes, the tables present, for each combination of event rarity under the null (μ) and sample size (n), the percent of the 5000 trials that did not fail and the coverage probabilities obtained from using traditional ANCOVA (i.e. OLS estimator of the standard errors). For simulations with the strongest effect, the tables also present robust coverage probabilities – that is, coverage probabilities obtained from robust standard errors, with the small sample size correction HC3 due to Davidson and MacKinnon (1993).

The results for those combinations of μ and sample size n that result in 5 or more failed trials per thousand, are highlighted in gray. Among the others, we highlighted in bold and red those coverage probabilities and robust coverage probabilities that are less than 0.940.

Binary Outcome

Table 3 presents results for simulations with a binary outcome when the design is balanced (i.e. $P(T = 1) = 0.5$). The table shows that whether or not the sample size is large enough to avoid failed trials, traditional ANCOVA yields valid inference under the null hypothesis. The smallest empirical coverage rate we found under the null was 93.9 percent. On the other hand, there can be undercoverage when there are strong treatment effects even if the trial is large enough to have a minimal probability of failure. Using the robust HC3 errors does not solve those problems.

Table 4 presents results for simulations with a binary outcome when the design is unbalanced (i.e. $P(T = 1) = 0.3$). As for the balanced design, we found no problem with inference under the null. However, under the assumption of a strong effect coverage probabilities tend to fall well short of 95 percent. Using robust standard errors helps and approximately 95 percent coverage is achieved far more frequently. However, despite the small sample size correction, the robust standard errors are not sufficient to obtain correct coverage for smaller sample sizes paired with low prevalence.

Compound Binary-Gamma Outcome

Table 5 shows parallel results for a compound binary-gamma outcome for a balanced design (i.e. $P(T = 0.5)$). Again, we found no evidence of invalid inference under the null. Coverage rates were close to nominal or conservative. In contrast, when there was a strong treatment effect (equivalent to a 10 percentage point boost in the probability of positive response and an increase of 0.04 in the mean response when positive), the coverage probabilities instead fall short of nominal promises for the smaller sample sizes for each of the event rarity levels. Robust coverage does not provide an improvement. Basically, it seems that a standard ANCOVA works fine for a corner variable like this unless over 90 percent of the mass is on the zero point, the gamma shape leads to extreme outliers, and the sample size is small.

Table 6 shows parallel results for a compound binary-gamma outcome for an unbalanced design (i.e. $P(T = 0.3)$). For the simulations with no effect, there are no Type I error/coverage problems. However, as we saw with the simulations with a binary outcome, the combination of a strong treatment effect and unbalance in the design leads to under-coverage – in this case for all sample sizes considered. Also as we saw with the simulations with a binary outcome, the HC3 robust standard errors improve the coverage, but only reach the approximate level of 95 percent for the larger sample sizes shown.

4. DISCUSSION

Our simulations have confirmed that there is a large region of experimental designs for the study of binary and compound binary-gamma distributed outcomes for which OLS provides valid inference. Larger sample sizes are required for OLS to provide valid inference when the research groups are unbalanced and as the effect size increases (and robust standard errors also help).[3] Lumley et al. (2002) found similar results for other distributions. Simulations like ours can help guide researchers to find the thresholds – as combinations of event rarity and sample size - for valid inference, either with OLS estimators of standard errors or with robust standard errors as defined above. Researchers need to turn to more robust methods only if the sample sizes are very small and kurtosis is extreme, particularly if the design is unbalanced. If an analyst has an experimental design and outcome combination that lies within the region where linear models will not yield valid inferences, then there are good textbooks on semi-parametric methods available to the analyst (e.g., Pesarin and Salmaso, 2010, Good, 2010, and van der Laan, 2011).

Our simulations also point to event rarity/sample size thresholds for minimizing the probability of failed trials when the outcome is binary our binary-gamma.

REFERENCES

Box, G.E.P. and Anderson, S.L. (1955). Permutation theory in the derivation of robust criteria and the study of departures from assumption. *Journal of the Royal Statistical Society*, series B; **17**:1-34.

Cheung, Y.B. (2007). A modified least-squares regression approach to the estimation of risk difference. *American Journal of Epidemiology*; **166(11)**: 1337-1344.

Davidson, R. and MacKinnon, J.G. (1993). *Estimation and Inference in Econometrics*. Oxford University Press: New York.

Fan, J. and Judkins D. (2010). Robust covariate control in cluster-randomized trials with MPLUS and WinBUGS. *Proceedings of the Joint Statistical Meetings* [CD-ROM] pp. 951-962. American Statistical Association: Alexandria, VA.

Good, P.I .(2010). *Permutation, Parametric and Bootstrap Tests of Hypotheses* (third ed.). Springer Science +Business Media: New York.

Koch, G.G., Tangen, C.M., Jung, J.W., and Amara, I.A. (1998). Issues for covariance analysis of dichotomous and ordered categorical data from randomized clinical trials and non-parametric strategies for addressing them. *Statistics in Medicine*; **17**:1863-1892.

LaVange, L.M, Durham, T.A., and Koch, G.G. (2005). Randomization-based nonparametric methods for the analysis of multicentre trials. *Statistical Methods in Medical Research*; **14**:281-301.

Lesaffre, E. and Senn, S. (2003). A note on non-parametric ANCOVA for covariate adjustment in randomized clinical trials. *Statistics in Medicine*; **22**:3586-3596.

Long, J.S. and Ervin, L.H. (2000). Using heteroscedasticity consistent standard errors in the linear regression model. *The American Statistician*; **54**:217–24.

Lumley, T., Diehr, P., Emerson, S., and Chen, L. (2002). The importance of the normality assumption in large public health data sets. *Annual Review of Public Health*, **23**, 151-169.

Neyman, J, Trans. Dorota M. Dabrowska and Terence P. Speed. (1990). On the application of probability theory to agricultural experiments: essay on principles (1923). *Statistical Science*; **5(4)**:465–472.

Pacheco, G.D., Hattendorf, J., Colford, J.M., Mäusezahl, D., and Smith, T. Performance of analytical methods for overdispersed counts in cluster randomized trials; Sample size, degree of clustering and imbalance. *Statistics in Medicine* 2009; **28**:2989-3011.

[3] These results are similar to simulation findings by Cheung (2007), who also examined coverage probabilities of OLS and HC3 confidence intervals, but not in the case of low prevalence probabilities.

Pesarin, F. and Salmaso, L. (2002). *Permutation Tests for Complex Data*. Wiley: Chichester.

Rosenbaum, P.R. Covariance adjustment in randomized experiments and observational studies (with discussion). *Statistical Science* 2002; **17**:286-327.

Rubin, D.B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*; **66(5)**:688–701.

Scheffé, H. (1959). *The Analysis of Variance*. Wiley: New York.

Tang, L., Duan, N., Klap, R., Asarnow, J.R. and Belin, T.R. (2009). Applying permutation tests with adjustment for covariates and attrition weights to randomized trials of health-services interventions. *Statistics in Medicine*; **28**; 65-74.

Tsiatis, A.A., Davidian, M., Zhang, M. and Lu, X. (2008). Covariate adjustment for two-sample treatment comparison in randomized clinical trials; A principled yet flexible approach. *Statistics in Medicine*; **27**:4658-4677.

van der Laan, M.J. and Rose, S. (2011). *Targeted Learning: Causal Inference for Observational and Experimental Data*. Springer Science +Business Media: New York.

Table 1. Excess kurtosis for simulated binary outcome as function of mu

μ	Excess kurtosis (under null)
.005	186
.01	91
.05	14
.1	5
.25	-0.7
.5	-2

Table 2. Excess kurtosis for simulated binary-gamma composite outcome as function of alpha and inverse logit mu

	$\text{logit}^{-1}(\mu)$				
α	**0.05**	**0.1**	**0.2**	**0.4**	**0.7**
1	131.9	65.1	28.6	12.9	6.9
1.5	81.1	42.3	18.0	7.2	4.1
2	66.1	27.7	13.7	5.7	2.4
4	36.3	17.4	6.3	1.4	0.2

Table 3. Coverage rates of nominal 95% confidence intervals constructed by OLS estimator of the standard error of an effect on a binary outcome, balanced design

		No treatment effect (c=0)		Very Strong Treatment Effect (c=3.24)		
μ	n	Percent of trials not failed	Coverage	Percent of trials not failed	Coverage	Robust coverage
0.005	50	0.233	0.999	0.868	0.969	0.968
0.005	100	0.411	0.997	0.951	0.971	0.967
0.005	200	0.641	0.998	0.989	0.894	0.899
0.005	400	0.874	0.982	1.000	**0.929**	**0.932**
0.005	500	0.932	0.978	1.000	**0.935**	**0.935**
0.005	750	0.980	0.957	1.000	0.941	0.942
0.005	1000	0.995	0.953	1.000	0.945	0.945
0.005	2000	1.000	0.953	1.000	0.948	0.948
0.01	50	0.403	0.999	0.953	0.967	0.963
0.01	100	0.647	0.997	0.990	0.911	0.929
0.01	200	0.882	0.988	0.999	**0.931**	**0.933**
0.01	400	0.984	0.962	1.000	0.943	0.943
0.01	500	0.996	0.955	1.000	0.943	0.945
0.01	750	1.000	0.957	1.000	0.945	0.946
0.01	1000	1.000	0.959	1.000	0.944	0.944
0.01	2000	1.000	0.950	1.000	0.949	0.948
0.05	50	0.927	0.970	1.000	**0.935**	0.945
0.05	100	0.994	0.948	1.000	0.942	0.945
0.05	200	1.000	0.953	1.000	0.945	0.947
0.05	400	1.000	0.951	1.000	0.948	0.949
0.05	500	1.000	0.947	1.000	0.944	0.946
0.05	750	1.000	0.956	1.000	0.950	0.952
0.05	1000	1.000	0.952	1.000	0.957	0.957
0.05	2000	1.000	0.945	1.000	0.953	0.953
0.1	50	0.996	0.944	1.000	0.938	0.947
0.1	100	1.000	0.944	1.000	0.946	0.948
0.1	200	1.000	0.951	1.000	0.949	0.951
0.1	400	1.000	0.947	1.000	0.954	0.955
0.1	500	1.000	0.952	1.000	0.947	0.948
0.1	750	1.000	0.948	1.000	0.951	0.952
0.1	1000	1.000	0.952	1.000	0.948	0.947
0.25	50	1.000	0.945	1.000	0.941	0.946
0.25	100	1.000	0.954	1.000	0.944	0.949
0.25	200	1.000	0.943	1.000	0.947	0.949
0.25	400	1.000	0.950	1.000	0.948	0.950
0.5	50	1.000	**0.939**	1.000	0.947	0.951
0.5	100	1.000	0.949	1.000	0.947	0.949
0.5	200	1.000	0.944	1.000	0.951	0.953

Table 4. Coverage rates of nominal 95% confidence intervals constructed by OLS estimator of the standard error of an effect on a binary outcome, unbalanced design

		No treatment effect (c=0)		Very Strong Treatment Effect (c=3.24)		
μ	n	Percent of trials not failed	Coverage	Percent of trials not failed	Coverage	Robust coverage
0.005	50	0.230	0.974	0.726	0.894	0.924
0.005	100	0.413	0.980	0.862	0.878	0.928
0.005	200	0.658	0.974	0.952	0.892	0.941
0.005	400	0.881	0.960	0.994	0.813	0.898
0.005	500	0.923	0.965	0.997	**0.858**	**0.906**
0.005	750	0.981	0.958	1.000	**0.840**	**0.916**
0.005	1000	0.995	0.960	1.000	**0.877**	**0.932**
0.005	2000	1.000	0.950	1.000	**0.890**	0.944
0.01	50	0.420	0.975	0.853	0.872	0.923
0.01	100	0.645	0.970	0.954	0.885	0.936
0.01	200	0.878	0.965	0.994	0.811	0.899
0.01	400	0.985	0.959	1.000	**0.851**	**0.926**
0.01	500	0.995	0.963	1.000	**0.867**	**0.921**
0.01	750	1.000	0.956	1.000	**0.880**	**0.935**
0.01	1000	1.000	0.952	1.000	**0.888**	**0.939**
0.01	2000	1.000	0.954	1.000	**0.896**	**0.938**
0.05	50	0.929	0.957	0.998	**0.847**	**0.923**
0.05	100	0.997	0.958	1.000	**0.872**	**0.935**
0.05	200	1.000	0.951	1.000	**0.888**	0.941
0.05	400	1.000	0.953	1.000	**0.905**	0.940
0.05	500	1.000	0.945	1.000	**0.910**	0.945
0.05	750	1.000	0.952	1.000	**0.909**	0.943
0.05	1000	1.000	0.954	1.000	**0.921**	0.950
0.05	2000	1.000	0.946	1.000	**0.934**	0.954
0.1	50	0.995	0.957	1.000	**0.881**	**0.933**
0.1	100	1.000	0.946	1.000	**0.901**	0.940
0.1	200	1.000	0.945	1.000	**0.913**	0.946
0.1	400	1.000	0.955	1.000	**0.919**	0.947
0.1	500	1.000	0.954	1.000	**0.924**	0.950
0.1	750	1.000	0.944	1.000	**0.930**	0.951
0.1	1000	1.000	0.952	1.000	**0.926**	0.945
0.25	50	1.000	0.946	1.000	**0.933**	0.940
0.25	100	1.000	0.953	1.000	**0.933**	0.948
0.25	200	1.000	0.948	1.000	**0.929**	0.944
0.25	400	1.000	0.946	1.000	**0.934**	0.945
0.5	50	1.000	0.940	1.000	0.987	0.951
0.5	100	1.000	0.940	1.000	0.969	0.943
0.5	200	1.000	0.948	1.000	0.958	0.950

Table 5. Coverage rates of nominal 95% confidence intervals constructed by OLS estimator of the standard error of an effect on a binary-gamma outcome, balanced design

			No treatment effect (δZ,δG)=(0,0)		Very strong treatment effect (δZ,δG)=(1,0.4)		
Probability of positive response $\text{logit}^{-1}(\mu)$	α	n	Percent of trials not failed	Coverage	Percent of trials not failed	Coverage	Robust coverage
0.05	1	50	0.924	0.993	0.990	0.886	0.888
0.05	1	100	0.993	0.982	1.000	**0.913**	**0.913**
0.05	1	200	1.000	0.970	1.000	**0.937**	0.940
0.05	1	2000	1.000	0.952	1.000	0.950	0.949
0.05	1.5	50	0.926	0.989	0.991	0.887	0.892
0.05	1.5	100	0.993	0.978	1.000	**0.928**	**0.929**
0.05	1.5	200	1.000	0.963	1.000	**0.935**	**0.938**
0.05	1.5	2000	1.000	0.951	1.000	0.950	0.949
0.05	2	50	0.922	0.985	0.991	0.901	0.907
0.05	2	100	0.995	0.976	1.000	**0.926**	**0.929**
0.05	2	200	1.000	0.959	1.000	**0.937**	0.941
0.05	2	2000	1.000	0.954	1.000	0.947	0.947
0.05	4	50	0.924	0.986	0.990	0.920	0.925
0.05	4	100	0.991	0.964	1.000	**0.938**	0.942
0.05	4	200	1.000	0.955	1.000	**0.936**	**0.939**
0.05	4	2000	1.000	0.953	1.000	0.946	0.948
0.1	1	50	0.995	0.976	1.000	**0.911**	**0.917**
0.1	1	100	1.000	0.963	1.000	**0.934**	**0.937**
0.1	1	200	1.000	0.963	1.000	**0.939**	0.944
0.1	1	2000	1.000	0.950	1.000	0.954	0.955
0.1	1.5	50	0.996	0.978	1.000	**0.922**	**0.927**
0.1	1.5	100	1.000	0.958	1.000	0.942	0.944
0.1	1.5	200	1.000	0.951	1.000	0.946	0.949
0.1	1.5	2000	1.000	0.951	1.000	0.949	0.948
0.1	2	50	0.995	0.970	1.000	**0.933**	**0.939**
0.1	2	100	1.000	0.959	1.000	0.944	0.948
0.1	2	200	1.000	0.946	1.000	0.947	0.948
0.1	2	2000	1.000	0.952	1.000	0.950	0.951
0.1	4	50	0.994	0.962	1.000	**0.928**	**0.931**
0.1	4	100	1.000	0.947	1.000	0.944	0.949
0.1	4	200	1.000	0.950	1.000	0.946	0.947
0.1	4	2000	1.000	0.948	1.000	0.946	0.946

Table 6. Coverage rates of nominal 95% confidence intervals constructed by OLS estimator of the standard error of an effect on a binary-gamma outcome, unbalanced design

			No treatment effect (δZ,δG)=(0,0)		Very strong treatment effect (δZ,δG)=(1,0.4)		
Probability of positive response $\text{logit}^{-1}(\mu)$	α	n	Percent of trials not failed	Coverage	Percent of trials not failed	Coverage	Robust coverage
0.05	1	50	0.917	0.965	0.978	0.793	0.808
0.05	1	100	0.995	0.957	0.999	**0.840**	**0.864**
0.05	1	200	1.000	0.962	1.000	**0.857**	**0.904**
0.05	1	2000	1.000	0.952	1.000	**0.888**	0.952
0.05	1.5	50	0.926	0.966	0.974	0.816	0.823
0.05	1.5	100	0.995	0.965	1.000	**0.841**	**0.872**
0.05	1.5	200	1.000	0.961	1.000	**0.857**	**0.905**
0.05	1.5	2000	1.000	0.952	1.000	**0.879**	0.948
0.05	2	50	0.925	0.971	0.974	0.817	0.828
0.05	2	100	0.994	0.960	1.000	**0.848**	**0.883**
0.05	2	200	1.000	0.956	1.000	**0.872**	**0.919**
0.05	2	2000	1.000	0.951	1.000	**0.878**	0.945
0.05	4	50	0.923	0.960	0.979	0.836	0.848
0.05	4	100	0.995	0.959	0.999	**0.849**	**0.893**
0.05	4	200	1.000	0.957	1.000	**0.866**	**0.917**
0.05	4	2000	1.000	0.950	1.000	**0.883**	0.947
0.1	1	50	0.995	0.957	0.999	**0.835**	**0.860**
0.1	1	100	1.000	0.958	1.000	**0.867**	**0.902**
0.1	1	200	1.000	0.953	1.000	**0.871**	**0.924**
0.1	1	2000	1.000	0.946	1.000	**0.891**	0.949
0.1	1.5	50	0.995	0.960	1.000	**0.853**	**0.881**
0.1	1.5	100	1.000	0.954	1.000	**0.873**	**0.915**
0.1	1.5	200	1.000	0.953	1.000	**0.879**	**0.931**
0.1	1.5	2000	1.000	0.951	1.000	**0.897**	0.952
0.1	2	50	0.995	0.954	0.999	**0.854**	**0.886**
0.1	2	100	1.000	0.959	1.000	**0.879**	**0.924**
0.1	2	200	1.000	0.950	1.000	**0.888**	**0.936**
0.1	2	2000	1.000	0.950	1.000	**0.893**	0.950
0.1	4	50	0.996	0.962	0.999	**0.864**	**0.900**
0.1	4	100	1.000	0.953	1.000	**0.884**	**0.929**
0.1	4	200	1.000	0.954	1.000	**0.894**	0.942
0.1	4	2000	1.000	0.952	1.000	**0.894**	0.947