STAT 778 Final Project: Igeta, Takahashi, and Matsui 2018

Tom Wallace

April 23, 2018

1 Introduction

1.1 Overview

This paper documents a simulation study based on Igeta, Takahashi, and Matsui 2018. It gives some background on their method; presents a software program coded in C; and conducts a simulation study using this program.

1.2 Background

Overdispersion refers to a situation in which the variance of a dataset exceeds that expected under the assumed statistical distribution. Overdispersion is common in count data. As a motivating example, consider the Poisson distribution. A single parameter determines both the mean and variance: i.e., for $X \sim \text{Poisson}(\lambda)$, $\mu = \sigma^2 = \lambda$. As a consequence, if the mean and variance differ in a dataset, the researcher cannot adjust parametric assumptions for without also affecting the other. If not addressed, overdispersion results in distorted test statistics and estimated standard errors. Clinical trials often feature count data: e.g., a trial of an anti-epsilepsy treatment may use *number of seizures over study period* as the outcome variable. The serious consequences of statistical error in such settings demands a rigorous method for dealing with overdispersion. Igeta, Takahashi, and Matsui 2018 is a new entry to the large literature on this topic. In particular, it presents methods for calculating statistical power and sample size in the presence of misspecified variance.

2 Methods

2.1 Overview

The following is a brief summary of the methods proposed in Igeta, Takahashi, and Matsui 2018. The reader is encouraged to consult their paper for the necessary details. All notation here follows that used in the original paper.

Consider a randomized control trial featuring n subjects. Subjects are randomly assigned to a treatment or control group. Let n_i be the sample size of the ith group, with $i \in \{1, 2\}$. Let X_{ij} be an indicator variable of subject j's group assignment. $X_{ij} = 0$ indicates assignment to the control group. Let Y_ij be a count variable for patient j in group i over the follow-up period $[0, T_{ij}]$. The expected value of Y_{ij} is affected by a rate parameter λ_i. The goal is to estimate the effect of treatment via a Poisson model

$$\lambda_i = \exp\left(\beta_0 + \beta_1 X_{ij}\right)$$

but λ is overdispersed.

Igeta, Takahashi, and Matsui 2018 propose a procedure for determining sample size and power in the presence of overdispersion. The basic idea is that we assume there is some true variance function V but this function is unknown. In practice we use working variance function $\tilde{V}$. We don't actually know if we have properly specified the true variance function and so would like sample size and power calculations that are robust to misspecification.

Igeta, Takahashi, and Matsui 2018 employ a Wald-type test statistic using the sandwich-type robust variance estimator under the null hypothesis:

$$Z = \frac{\hat{\beta}_1}{\sqrt{n^{-1}\hat{W}_0}} \tag{1}$$

They propose that the asymptotic power of the test using Z with two-sided significance level α is:

$$\Pr\left(Z > z_{1-\alpha/2}\right) = 1 - \Phi\left(z_{1-\alpha/2}\sqrt{\frac{W_0}{W_1}} - \sqrt{n}\frac{\beta_1}{\sqrt{W_1}}\right) \tag{2}$$

The sample size that provides power greater than or equal to $1-\beta$ is

$$n \geq \frac{(z_{1-\alpha/2}\sqrt{W_0} + z_{1-\beta}\sqrt{W_1})^2}{\left(\log(\lambda_2/\lambda_1)\right)^2} \tag{3}$$

The reader should consult the paper to understand these equations, as space constraints disallow a full explanation here. The chief claim is that these methods are robust to misspecification of variance.

2.2 Extensions

Generating appropriately dispersed data, and calculating the Z-statistic given in (1), are non-trivial. Here, we expand upon the definitions given in the paper to obtain a workable formulation.

2.2.1 Generating Overdispersed Data

2.2.2 Calculating $\hat{\beta}_1$

The maximum quasi-likelihood estimate of $\boldsymbol{\beta} = (\beta_0, \beta_1)$ is given as:

$$\sum_{i=1}^{2}\sum_{j=1}^{n_i} \boldsymbol{D}'_{ij}\tilde{V}_{ij}^{-1}(Y_{ij} - \mu_{ij}) = \mathbf{0} \tag{4}$$

where

$$\boldsymbol{D}'_{ij} = \frac{\partial \mu_{ij}}{\partial \boldsymbol{\beta}}$$

Although not directly stated, we will trust that these functions are convex (if not, the resultant estimates are not unique). This system of equations can be numerically solved using the Newton-Raphson algorithm but first we must obtain a more tractable form. Since $\mu_{ij} = T_{ij}\lambda_i$,

$$\frac{\partial \mu_{ij}}{\partial \beta_0} = T_{ij}\exp\left(\beta_0 + \beta_1 X_{ij}\right)$$

$$\frac{\partial \mu_{ij}}{\partial \beta_1} = T_{ij}X_{ij}\exp\left(\beta_0 + \beta_1 X_{ij}\right)$$

For $\tilde{V}_{ij}$, we will only concern ourselves with the working variance function of form $\tilde{V}_{ij} = \tilde{\phi}\mu_{ij} = \tilde{\phi}T_{ij}\exp\left(\beta_0 + \beta_1 X_{ij}\right)$. Plugging everything back into (4), we have

$$\sum_{i=1}^{2}\sum_{j=1}^{n_i} T_{ij}\exp\left(\beta_0 + \beta_1 X_{ij}\right)\frac{1}{\tilde{\phi}T_{ij}\exp\left(\beta_0 + \beta_1 X_{ij}\right)}(Y_{ij} - T_{ij}\exp\left(\beta_0 + \beta_1 X_{ij}\right)) = 0$$

$$\sum_{i=1}^{2}\sum_{j=1}^{n_i} T_{ij}X_{ij}\exp\left(\beta_0 + \beta_1 X_{ij}\right)\frac{1}{\tilde{\phi}T_{ij}\exp\left(\beta_0 + \beta_1 X_{ij}\right)}(Y_{ij} - T_{ij}\exp\left(\beta_0 + \beta_1 X_{ij}\right)) = 0$$

Which simplifies to

$$\sum_{i=1}^{2}\sum_{j=1}^{n_i}(Y_{ij} - T_{ij}\exp(\beta_0 + \beta_1 X_{ij})) = 0$$

$$\sum_{i=1}^{2}\sum_{j=1}^{n_i} X_{ij}(Y_{ij} - T_{ij}\exp(\beta_0 + \beta_1 X_{ij})) = 0 \tag{5}$$

as the exponential expressions cancel out and $\tilde{\phi}$ is an arbitrary constant that does not affect the solution. We now must take the derivatives of the left-hand side of (5):

$$\frac{\partial}{\partial\beta_0}\sum_{i=1}^{2}\sum_{j=1}^{n_i}(Y_{ij} - T_{ij}\exp(\beta_0 + \beta_1 X_{ij})) = -\sum_{i=1}^{2}\sum_{j=1}^{n_i} T_{ij}\exp(\beta_0 + \beta_1 X_{ij})$$

$$\frac{\partial}{\partial\beta_1}\sum_{i=1}^{2}\sum_{j=1}^{n_i}(Y_{ij} - T_{ij}\exp(\beta_0 + \beta_1 X_{ij})) = -\sum_{i=1}^{2}\sum_{j=1}^{n_i} T_{ij}X_{ij}\exp(\beta_0 + \beta_1 X_{ij})$$

$$\frac{\partial}{\partial\beta_0}\sum_{i=1}^{2}\sum_{j=1}^{n_i} X_{ij}(Y_{ij} - T_{ij}\exp(\beta_0 + \beta_1 X_{ij})) = -\sum_{i=1}^{2}\sum_{j=1}^{n_i} T_{ij}X_{ij}\exp(\beta_0 + \beta_1 X_{ij})$$

$$\frac{\partial}{\partial\beta_1}\sum_{i=1}^{2}\sum_{j=1}^{n_i} X_{ij}(Y_{ij} - T_{ij}\exp(\beta_0 + \beta_1 X_{ij})) = -\sum_{i=1}^{2}\sum_{j=1}^{n_i} T_{ij}X_{ij}^2\exp(\beta_0 + \beta_1 X_{ij})$$

Thus, the Newton-Raphson algorithm for finding $\hat{\boldsymbol{\beta}}$ is (omitting the summation symbols for brevity):

$$\begin{bmatrix}\beta_0^{(n+1)}\\ \beta_1^{(n+1)}\end{bmatrix} = \begin{bmatrix}\beta_0^{(n)}\\ \beta_1^{(n)}\end{bmatrix} - \begin{bmatrix}-T_{ij}\exp(\hat{\beta}_0^{(n)} + \hat{\beta}_1^{(n)}X_{ij}) & -T_{ij}X_{ij}\exp(\hat{\beta}_0^{(n)} + \hat{\beta}_1^{(n)}X_{ij})\\ -T_{ij}X_{ij}\exp(\hat{\beta}_0^{(n)} + \hat{\beta}_1^{(n)}X_{ij}) & -T_{ij}X_{ij}^2\exp(\hat{\beta}_0^{(n)} + \hat{\beta}_1^{(n)}X_{ij})\end{bmatrix}^{-1}\begin{bmatrix}Y_{ij} - T_{ij}\exp(\hat{\beta}_0^{(n)} + \hat{\beta}_1^{(n)}X_{ij})\\ X_{ij}(Y_{ij} - T_{ij}\exp(\hat{\beta}_0^{(n)} + \hat{\beta}_1^{(n)}X_{ij}))\end{bmatrix} \tag{6}$$

2.3 Calculating $\hat{W}_0$

It turns out that calculating $\hat{W}_0$ requires $\hat{\tilde{\phi}}^*$, the estimate of the dispersion parameter of the working variance function under the null hypothesis. This paper only deals with the working variance function $V = \phi\mu$; Igeta, Takahashi, and Matsui 2018 only provide a method for calculating $V = \mu + \phi\mu^2$.

3 Software Program

3.1 Overview

We can empirically test these methods by the following procedure.

- Randomly generate data that is overdispersed according to some true variance function; that has true treatment effect $\exp(\beta_1)$; and that has sample size equal to that recommended by (3) for the desired power $1 - \beta$ calculated using (2).
- Conduct a test of $H_0 : \beta_1 = 0$ vs. $H_1 : \beta_1 \neq 0$ using (1). Record whether a type II error occurred.
- Conduct many iterations of the previous two steps. Calculate the proportion of type II errors:

$$\bar{\beta} = \frac{1}{N}\sum_{t=1}^{N}\beta_{(t)}$$

  where $\beta_{(t)}$ is an indicator of whether a type II error occured for trial t and N is the total number of iterations.

- Compare $1 - \bar{\beta}$ to the asymptotic power calculated using (2). We hope that $1 - \bar{\beta} \approx 1 - \beta$.

We conduct this procedure for multiple different true variance functions and working variance functions and effect sizes. If the empirical power $1 - \bar{\beta}$ matches the asymptotic power across all specifications, we conclude that the claims of Igeta, Takahashi, and Matsui 2018 are correct and that their methods are robust.

3.2 Technical Details

The software program carries out the testing strategy outlined above. Source code is contained in `final.c`. The program requires the GNU Scientific Library (GSL), an open-source numerical library. It can be obtained from `www.gnu.org/software/gsl`; or, it can installed from most standard Linux package managers. An example command to achieve the latter is:

```
sudo apt-get install gsl-bin libgsl-dev
```

Compilation of `final.c` is best achieved in two steps. First, use the below command to compile the program but not link it. You may need to change the argument passed to the `-I` flag to wherever the `gsl` header files live on your computer.

```
gcc -I/usr/include -c final.c
```

This command should create an object file `final.o`. Link this object file to relevant libraries with the following command. You may need to change the argument passed to the `-L` flag to wherever `libgsl` lives on your computer.

```
gcc -L/usr/lib final.o -o final -lgsl -lgslcblas -lm
```

Once successfully compiled, the program can be executed. It does not require any arguments. Output is comma-separated text printed to `stdout`. You likely want to pipe this output to a text file, as per the following command:

```
./final > output.csv
```

4 Simulation Study

4.1 Settings

The rate of event incidence in the control group, λ_1, is 1.25. Equivalently, this corresponds to $\beta_0 = 0.22314$ in $\exp(\beta_0)$. The allocation ratio is equal across groups (i.e. $q_1 = q_2 = 0.5$). The follow-up period τ is 1 year. Some subjects drop out early. Time to dropout follows an exponential distribution with parameter value 0.356, such that about 30% of subjects drop out early. So in practice the observed follow-up time T is the miniimum of time to dropout and τ.

The tested values for true relative risk (effect size) $\exp(\beta_1)$ are $\exp(\beta_1) = \frac{\lambda_2}{\lambda_1} = 0.4|0.6|0.8$. The desired power is 90% across all specifications. The power for a particular true effect size (per above) is calculated using (2). The number of observations is computed using (3).

The true variance function is

5 Conclusion

6 References

[1] Masataka Igeta, Kunihiko Takahashi, and Shigeyuki Matsui. "Power and sample size calculation incorporating misspecifications of the variance function in comparative clinical trials with over-dispersed count data". In: *Biometrics* (2018). DOI: 10.111/biom.12878.